EPFL

Thèse n° 9329

Advancing Computational Chemistry with Stochastic and Artificial Intelligence Approaches

Présentée le 10 août 2023

Faculté des sciences de base
Laboratoire de chimie et biochimie computationnelles
Programme doctoral en physique

pour l'obtention du grade de Docteur ès Sciences

par

Justin VILLARD

Acceptée sur proposition du jury

Prof. F. Mila, président du jury
Prof. U. Röthlisberger, directrice de thèse
Prof. M. D. Coutinho Neto, rapporteur
Prof. D. Truhlar, rapporteur
Dr R. Laplaza Solanas, rapporteur

École polytechnique fédérale de Lausanne

2023

À ma famille

Acknowledgements

As I come to the end of my time at EPFL, I would like to express my sincere gratitude to Prof. *Ursula Röthlisberger* for providing me with the opportunity to pursue a PhD in the Laboratory of Computational Chemistry and Biochemistry (LCBC). Beyond all the knowledge and valuable advice she shared with me, her kindness and empathy significantly eased the pressure and stress that often accompany academic research in uncharted territories. Alongside her, the senior postdogs *Quanta*, *Mocca*, and the late *Ruby* were wonderful companions, helping me take breaks from the screen and gaining perspective on both scientific and non-scientific challenges.

Before moving on, I would like to express my sincere thanks to the members of my thesis examination jury: Prof. *Mauricio D. Coutinho Neto*, Prof. *Donald Truhlar*, and Dr. *Rubén Laplaza Solanas*. Their thorough review of my work and thoughtful comments were truly valuable and greatly appreciated. I would also like to thank Prof. *Frédéric Mila* for chairing my jury and offering support throughout the process.

During my time at LCBC, I had the privilege of interacting with numerous brilliant and friendly people, which made science feel more human and alive. I would therefore like to thank both the current and former members of the laboratory whom I had the pleasure of meeting and enjoying during my stay, listed in alphabetical order: *Lorenzo Agosta, Paramvir Ahlawat, Andrej Antalik, Pablo Baudin, Swarnendu Bhattacharyya, Martin Bircher, Viacheslav Bolnykh, Ariadni Boziki, Esra Bozkurt, Nicholas Browning, Virginia Carnevali, Mathias Dankl, Polydefkis Diamantis, Elisa Donati, Simon Dürr, Guido Frisari, Farzaneh Jahanbakhshi, Sophia Johnson, Murat Kılıç, Nikolaos Lempesis, Andrea Levy, Maria Letizia Merlini, Marko Mladenovic, Irea Mosquera, François Mouvet, Daniele Narzi, Karin Pasche, Vladislav Slama, Alicia Solano, Olga Syzgantseva, Cecilia Vona*, and *Thibaud von Erlach*.

Among them, I want to give special recognition to our benevolent secretary, *Karin Pasche*. She has been consistently dedicated to keeping not only the administrative but also the emotional foundation of the laboratory strong and steady. Alongside her, as well as *Nicholas Browning*, *Murat Kılıç*, and *Ursula Röthlisberger*, I had the opportunity to organize the "Computational Chemistry meets Artificial Intelligence" (CC2AI) conference in 2018, an experience that has left me with unforgettable memories. Furthermore, I am incredibly grateful for the collaborative efforts and publications

with my office mates and friends: *Martin Bircher*, *Murat Kılıç*, and *François Mouvet*. *Martin*, thank you for your unwavering knowledge, patience, and assistance during our countless hours spent in CPMD. Your guidance has been invaluable, and I truly appreciate the lively debates and candid discussions we had during the compilation process, as well as your adept management of coffee breaks. Moreover, many thanks for taking the time to translate my abstract into German. *Murat*, I would like to extend a special thank you to you for being the experienced and supportive maaaaan in the office, always showing us youngsters the right way to handle things. Your expertise and sense of humor have been invaluable assets. Merci beaucoup *François*, fidèle camarade de fortune depuis le début, pour le meilleur et pour le pire. Merci d'avoir su encaisser mes plaintes et inquiétudes quotidiennes et contribué à les apaiser pendant ces nombreuses années. Nous avons vaincu, bravo!

Cette thèse clôt plus d'une décennie passée à l'EPFL dont la survie est également due au soutien mental, physique, et surtout hautement intellectuel de mes confrères *Cyril Alispach, Sylvain Hauser, Gaël Lederrey, Guillaume Meyrat, Joachim Muth, Luc Testa* et *Loïc Urio*. A ceux-ci s'ajoute l'équipe des docteurs en physique *Samy Adjam, Virgile Favre, Samuel Gozel, Francisco Kim* et *Luc Testa*, toujours motivés par des pauses de midi et cafés prolongés à parler de tout sauf de science, politique et avenir de la condition humaine. Merci à toutes ces personnes sympathiques! Je ne peux bien sûr pas parler de l'EPFL sans mentionner mes fidèles et réconfortants amis *Mathias Dorier* et *Luc Testa*, présents déjà bien avant cette aventure, et qui s'y sont aussi frottés. Merci les gus, pop à vous, vous êtes grands, beaux, vaillants, des kings!

Pour conclure, mes sincères remerciements vont à mon entourage, famille et amis. Un merci spécial aux Bratislaboys, *Vincent, Romain, Yoann* et *Samuel*, pour m'avoir fait continuer à vivre dans le monde "extérieur". À *Valérie* ♡, merci infiniment pour m'avoir encouragé et épaulé au quotidien, contribuant grandement à cette réussite. Finalement, je remercie bien entendu ma soeur *Caroline* et son mari *Benoit*, mon frère *Matthieu*, et mes chers parents *Marylène* et *Jean-François* pour leur soutien et présence indispensables durant toutes ces années d'études et de doctorat.

À toutes et tous, encore un sincère merci!

Lausanne, July 28, 2023 Justin

Abstract

Computational chemistry aims to simulate reactions and molecular properties at the atomic scale, advancing the design of novel compounds and materials with economic, environmental, and societal implications. However, the field relies on approximate quantum chemical methods that balance cost and accuracy. This trade-off hinders effective configuration sampling when combining ab initio methods with molecular dynamics (MD), limiting thermodynamic examination to systems with a few hundred atoms and temporal sampling of hundreds of picoseconds.

This thesis focuses on leveraging unconventional approaches based on stochastic sampling and artificial intelligence (AI) to address the three-fold challenge of attaining high accuracy, accommodating large system sizes, and enhancing the efficiency of configurational sampling for specific problems.

It starts with the implementation of second-order Møller-Plesset perturbation theory (MP2) in a plane wave (PW) basis set, that allows to systematically converge reference energies to the complete basis set (CBS) limit, devoid of basis set superposition errors, and enables the application of MP2 to periodic systems. A comparison of PW MP2 interaction energies with computationally more expedient correlation-consistent basis sets reveals the limitations of the latter in capturing full correlation energies at the CBS limit and for larger systems. Secondly, a PW Monte Carlo MP2 method is introduced, which stochastically samples virtual space contributions to the correlation energy, and reduces execution times up to a thousand-fold while maintaining low statistical errors. The PW MP2 implementation is not only valuable independently but also in the context of density functional theory (DFT), where it enables the development of the most accurate double-hybrid DFT functionals to date. Despite this, the accuracy of DFT results still depends on the specific system being studied. Thus, as a third step, the accuracy of popular Minnesota DFT functionals in describing the properties of liquid water is assessed, thanks to the acceleration and transferability of a machine learning (ML) multiple time step MD scheme. Comparisons with other DFT approximations and experimental data highlight the importance of a judicious amount of exact exchange for capturing hydrogen bonding. The M06-2X(-D3) functionals are identified as the top-performing candidates, demonstrating good performance for both structural and dynamical properties. This spotlights their potential for further validation when combined with an explicit treatment of nuclear quantum effects. The fourth topic

of the thesis addresses configurational sampling using genetic algorithms (GAs) to sample low-energy configurations of peptides as observed in ultracold spectroscopy experiments. By utilizing a surrogate energy model and subsequent refinement with DFT, the GA approach enables efficient exploration of the potential energy surface (PES) in a matter of hours, significantly faster than traditional search methods that require weeks of trials. Remarkably, the newly developed GA approach successfully retrieves lowest-energy structures that align with experimentally-resolved infrared spectra. Finally, an alternative GA combined with unsupervised learning is introduced, improving PES coverage in low-energy regions. In summary, this thesis contributes to enhancing the PES accuracy and sampling by combining quantum chemical methods with stochastic and AI approaches.

Keywords: correlation energy, non-covalent interactions, second-order Møller Plesset perturbation theory, plane waves, correlation-consistent basis sets, Monte Carlo integration, liquid water, density functional theory, exchange-correlation functionals, ab initio molecular dynamics, machine learning, peptide structures, genetic algorithms

Résumé

La chimie computationnelle vise à simuler les réactions chimiques et les propriétés moléculaires à l'échelle atomique dans le but de concevoir de nouveaux composés et matériaux chimiques ayant de vastes implications économiques, environnementales et sociétales. Néanmoins, la chimie computationnelle possède tout un ensemble d'approximations quantiques fournissant divers compromis entre coût de calcul et précision. De tels compromis compliquent l'échantillonnage rapide de configurations lorsque les méthodes ab initio sont combinées avec la dynamique moléculaire (MD), limitant ainsi l'analyse thermodynamique de systèmes comptant plus de quelques centaines d'atomes pour un échantillonage temporel s'étalant sur quelques centaines de picosecondes.

Cette thèse porte sur l'exploitation d'approches non conventionnelles stochastiques ou basées sur l'intelligence artificielle pour résoudre les trois problèmes que sont l'atteinte d'une grande précision, la considération de systèmes de grande taille, et la minimisation du temps d'échantillonnage pour des problèmes spécifiques.

Elle commence par l'implémentation de la théorie des perturbations de Møller-Plesset au second ordre (MP2) dans les ondes planes (PW), qui permet la convergence systématique d'énergies de référence à la limite de la base complète, exemptes d'erreur due à la superposition de la base, et ouvrant la possibilité de traiter des systèmes périodiques. Une comparaison d'énergies d'interaction MP2 dans les ondes planes à celles de bases cohérentes avec la corrélation, plus rapides, révèle les limites de ces dernières à capturer totalement l'énergie de corrélation dans la limite de la base complète, ainsi que pour des systèmes plus grands. Deuxièmement, une méthode PW Monte Carlo MP2 est présentée. Cette approche échantillonne stochastiquement les contributions de l'espace virtuel à l'énergie de corrélation, réduisant les temps d'exécution d'un facteur pouvant atteindre trois ordres de grandeur tout en maintenant les erreurs stochastiques à des niveaux négligeables. L'implémentation de la PW MP2 n'est pas seulement avantageuse en soi, mais aussi dans le contexte de la théorie de la fonctionnelle de la densité (DFT) où elle permet le dévelopment des fonctionnelles doubles hybrides les plus précises à ce jour. Malgré cela, la précision des résultats obtenus avec la DFT dépend du système étudié. Ainsi, comme troisième axe de recherche, la précision des fonctionnelles DFT du Minnesota pour décrire les propriétés de l'eau liquide est analysée, grâce à l'accélération et à la transférabilité d'une méthode d'apprentissage

automatique couplée à un schéma MD à pas de temps multiples. La comparaison entre les fonctionnelles du Minnesota, d'autres approximations DFT, et les données expérimentales soulignent l'importance d'un choix judicieux de la quantité d'échange exact dans la description des liaisons hydrogène. Les fonctionnelles M06-2X(-D3) apparaissent comme des candidats précis pour les propriétés structurelles et dynamiques de l'eau, et donc prometteurs pour une future validation avec la considération explicite des effets quantiques nucléaires. Le quatrième axe s'intéresse à l'échantillonage de l'espace des configurations grâce aux algorithmes génétiques (GAs), afin d'explorer les configurations de basse énergie de peptides telles qu'observées dans la spectroscopie ultrafroide. En utilisant un modèle d'énergie de substitution avant réoptimisation par la DFT, l'approche des GAs permet d'explorer efficacement la surface d'énergie potentielle (PES) en quelques heures, alors que les méthodes de recherche traditionnelles nécessitent des semaines d'essais. De façon remarquable, les GAs sont capables de retrouver des structures correspondant aux spectres infrarouges expérimentaux. Enfin, un GA alternatif combiné à de l'apprentissage non supervisé est présenté, et permet une couverture encore meilleure de la PES dans les régions de basses énergies. Ainsi, cette thèse contribue à l'amélioration de la précision et de l'échantillonnage de la PES grâce à la combinaison des méthodes quantiques traditionnelles avec des approches stochastiques et issues de l'intelligence artificielle.

Mots-clés : énergie de corrélation, intéractions non covalentes, théorie de la perturbation de Møller-Plesset au second ordre, ondes planes, fonctions de base cohérentes avec la corrélation, intégration Monte Carlo, eau liquide, théorie de la fonctionnelle de la densité, fonctionelles d'échange-corrélation, dynamique moléculaire ab initio, apprentissage automatique, structures de peptides, algorithmes génétiques

Zusammenfassung

Die computergestützte Chemie zielt mit der atomistischen Simulation von chemischen Reaktionen und molekularen Eigenschaften darauf ab, die Entwicklung neuer Stoffe und Materialen voranzutreiben und damit positive Effekte auf ökonomischer, ökologischer und gesellschaftlicher Ebene zu erzielen. Die Disziplin der computergestützten Chemie basiert auf genäherten quantenmechanischen Methoden, die ein geeignetes, wenngleich kompromissbehaftetes, Gleichgewicht zwischen Ressourcenbedarf und Genauigkeit erzielen. Sollen solche ab initio Methoden mit Molekulardynamik (MD) verbunden werden, so verhindert ebendieser Kompromiss das effiziente Abtasten aller möglichen Konfigurationen, womit thermodynamische Untersuchungen dieser Art auf Systeme von wenigen hunderten Atomen und Zeiträume von einigen hundert Pikosekunden beschränkt bleiben.

Diese Arbeit befasst sich mit der Nutzung unkonventioneller Lösungsansätze; namentlich stochastischem Sampling und künstlicher Intelligenz (KI). Im Falle dreier Beispiele wird gezeigt, wie es diese Ansätze ermöglichen, die dreifache Herausforderung zu meistern, die sich aus dem Bedürfnis nach hoher Genauigkeit, genügend grossen Modellsystemen und effizientem Abtasten des Konfigurationsraums ergibt.

Zunächst wird die Implementierung von Møller-Plesset-Störungstheorie zweiter Ordnung (MP2) bei Verwendung von ebenen Wellen (engl. plane waves, PW) als Basisfunktionen vorgestellt. Dieser Ansatz erlaubt ein systematisches Konvergieren von Referenzenergien zum Grenzfall eines kompletten Basissatzes (engl. complete basis set (CBS) limit) ohne Basissatzüberlappfehler und ermöglicht zudem die Behandlung periodischer Systeme. Ein Vergleich von PW MP2-Interaktionsenergien mit korrelationskonsistenten Basissätzen, die einen geringeren Rechenbedarf nach sich ziehen, zeigt die Einschränkungen Letzterer auf, falls die komplette Interaktionsenergie im CBS bzw. für grössere Systeme berechnet werden soll. Im Anschluss wird ein PW-Monte-Carlo-MP2-Ansatz eingeführt, in dem der Raum, der durch diejenigen virtuellen Orbitale, die zur Korrelationsenergie beitragen, aufgespannt wird, stochastisch gesampelt wird. Dadurch kann die Laufzeit solcher Rechnungen um bis zu drei Grössenordnungen reduziert werden, wobei der statistische Fehlerbereich klein bleibt. Diese Implementierung kann ausser für MP2 zudem im Bereich der Dichtefunktionaltheorie (DFT) verwendet werden, wo sie die Entwicklung genauerer Doppelhybridfunktionale ermöglicht, die Stand heute zu den genauesten Dichtefunktionalen gehören. Trotz

der hohen Genauigkeit dieser Funktionalfamilie hängt die Gesamtgenauigkeit von DFT-Rechnungen nach wie vor vom System ab, das betrachtet wird. Wir betrachten folglich die Genauigkeit der bekannten und beliebten Minnesota-Funktionale für die Beschreibung von Wasser in der flüssigen Phase unter Nutzung eines flexiblen und transferierbaren Molekulardynamikansatzes, der sich auf maschinelles Lernen (ML) in Verbindung mit Multiplen-Zeitschritt-Integratoren stützt. Der Vergleich mit anderen DFT-Näherungen und experimentellen Daten betont die Wichtigkeit eines ausgeglichenen Beitrags der Austauschintegrale (exact exchange) zur Gesamtenergie, sofern Wasserstoffbrücken beschrieben werden sollen. Die M06-2X(-D3)-Funktionale zeigen hier die besten Resultate, sowohl im Hinblick auf strukturelle als auch auf dynamische Eigenschaften. Dies macht diese Funktionale zu vielversprechenden Kandidaten für weitere Validierungen, die zusätzlich das quantenmechanische Verhalten der Nuklei berücksichtigen. Zum Schluss wird das Sampling des Konfigurationsraums mittels genetischen Algorithmen (GA) diskutiert; dieser Ansatz wird zum Abtasten von energetisch tiefliegenden Konfigurationen von Peptiden, wie sie bei ultrakalten spektroskopischen Experimenten auftreten, angewandt. Durch Verwendung eines Ersatzenergiemodells und einer nachgeschalteten Verfeinerung mittels DFT ermöglicht es der GA-Ansatz, die Potentialhyperfläche (PES) effizient in wenigen Stunden zu sampeln. Dies stellt im Vergleich zu traditionellen Suchalgorithmen, die mehrere Wochen in Anspruch nehmen, eine signifikante Beschleunigung dar. Bemerkenswerterweise findet dieser neu entwickelte GA-Algorithmus erfolgreich Tiefstenergiestrukturen, die mit experimentell zugänglichen Infrarotspektren übereinstimmen. Zuletzt führen wir einen alternativen GA mit unüberwachtem Lernen ein, der die Abdeckung des PES in Tiefenergiedomänen verbessert. Im Ganzen trägt diese Dissertation durch die Kombination quantenmechanischer Methoden und stochastischer bzw. KI-Ansätze zur Verbesserung der Genauigkeit des PES sowie zu dessen effizienterem Sampling bei.

Schlüsselwörter: Korrelationsenergie, Nichtkovalente Bindungen, Møller-Plesset-Störungstheorie zweiter Ordnung, ebene Wellen, korrelationskonsistente Basissätze, Monte-Carlo-Integration, flüssiges Wasser, Dichtefunktionaltheorie, Austausch-Korrelations-Funktionale, ab-initio-Molekulardynamik, Maschinelles Lernen, Peptidstrukturen, Genetische Algorithmen

Contents

Acknowledgements i

Abstract (English/Français/Deutsch) iii

1 Introduction 1
- 1.1 Computational chemistry . . . 2
  - 1.1.1 The ultimate goal . . . 2
  - 1.1.2 The cost-accuracy trade-off . . . 3
  - 1.1.3 The multiple minima problem and effective sampling . . . 5
- 1.2 Ways of improvement . . . 7
- 1.3 Aim and outline of this thesis . . . 8
  - 1.3.1 Improving the accuracy of the potential energy surface . . . 8
  - 1.3.2 Navigating the potential energy surface with artificial intelligence approaches . . . 9

I Describing nuclei and electrons in matter 13

2 Electronic structure theory 15
- 2.1 The interacting many-body problem . . . 15
  - 2.1.1 Born-Oppenheimer approximation . . . 16
  - 2.1.2 Variational principle . . . 18
- 2.2 Wavefunction-based methods . . . 19
  - 2.2.1 Hartree-Fock method . . . 19
  - 2.2.2 Electron correlation . . . 23
  - 2.2.3 Rayleigh–Schrödinger perturbation theory . . . 24
  - 2.2.4 Post-Hartree-Fock methods . . . 26
- 2.3 Density functional theory . . . 33
  - 2.3.1 Kohn-Sham density functional theory . . . 35
  - 2.3.2 Density functional approximations . . . 37

3 Computational approaches 49
- 3.1 Basis sets . . . 50
  - 3.1.1 Atom-centered basis . . . 51

3.1.2 Plane wave basis 52

II Improving the accuracy of the potential energy surface 59

4 Plane waves versus correlation-consistent basis sets: A comparison of MP2 non-covalent interaction energies in the complete basis set limit 61
4.1 Abstract 62
4.2 Introduction 62
4.3 Methods 66
4.3.1 Second-order Møller-Plesset perturbation theory 66
4.3.2 Plane wave basis set 66
4.3.3 Correlation-consistent GTO basis sets 71
4.4 Computational details 75
4.4.1 Plane wave basis set 76
4.4.2 Correlation-consistent GTO basis sets 77
4.5 Results and discussion 77
4.5.1 Converging accurate MP2 energies with plane waves 77
4.5.2 HF/MP2 energies in PWs versus GTO bases 81
4.5.3 HF/MP2 energies in PWs versus extrapolated GTO bases 83
4.5.4 Other GTO extrapolations 87
4.5.5 System size dependency 89
4.6 Conclusions and outlook 90

5 Efficient treatment of correlation energies at the basis-set limit by Monte Carlo summation of continuum states 93
5.1 Abstract 94
5.2 Introduction 94
5.2.1 Møller-Plesset perturbation theory 96
5.3 Distribution of continuum states and stochastic sampling 98
5.4 Computational methods 101
5.4.1 General setup 101
5.4.2 Extrapolation of correlation energies 101
5.5 Results and discussion 102
5.5.1 Accuracy of stochastic summation 102
5.5.2 Performance and speedups 104
5.5.3 Generalization to the random phase approximation 106
5.6 Conclusions and outlook 108

III Making the nuclei move 111

6 Ab initio molecular dynamics 113

6.1 Time versus ensemble averages and ergodicity . . . 113
6.2 Equations of motion . . . 115
6.3 Born-Oppenheimer molecular dynamics . . . 118
6.4 Multiple time step algorithms . . . 119
6.5 Machine learning-aided multiple time step algorithms . . . 122
6.6 Car-Parrinello molecular dynamics . . . 127

IV Navigating the potential energy surface with artificial intelligence approaches 129

7 Structure and dynamics of liquid water from ab initio simulations: Adding Minnesota density functionals to Jacob's ladder 131
7.1 Abstract . . . 132
7.2 Introduction . . . 132
7.3 Theory and methods . . . 136
7.3.1 Minnesota density functionals . . . 136
7.3.2 Simulations . . . 138
7.3.3 Analysis . . . 140
7.4 Results and discussion . . . 143
7.4.1 Structural properties . . . 143
7.4.2 Dynamical properties . . . 153
7.5 Conclusions . . . 156

8 Surrogate based genetic algorithm method for efficient identification of low-energy peptide structures 159
8.1 Abstract . . . 160
8.2 Introduction . . . 160
8.3 Methods . . . 163
8.3.1 Reference data . . . 163
8.3.2 Genetic algorithms . . . 164
8.3.3 Optimization of peptide conformations with EVOLVE . . . 166
8.4 Computational details . . . 171
8.4.1 GA parameters . . . 171
8.4.2 Surrogate fitness function . . . 171
8.5 Results and discussion . . . 172
8.5.1 Performance of different surrogate fitness functions . . . 172
8.5.2 GA optimization of GPGG . . . 178
8.5.3 GA optimization of Gramicidin . . . 183
8.5.4 Computational performance . . . 188
8.6 Conclusions and outlook . . . 190
8.7 Additional details . . . 191
8.7.1 Simulated binary crossover . . . 191

8.7.2 Radius of gyration and number of hydrogen bonds 193

9 Enhanced screening of low-energy peptide structures with genetic algorithms and clustering-based novelty search 195
9.1 Introduction 195
9.2 Clustering-enhanced genetic algorithms 196
9.2.1 Agglomerative hierarchical clustering 199
9.3 Results and discussion 202
9.3.1 Number of local minima 202
9.3.2 Global minimum detection 205
9.4 Conclusions and outlook 206

V Conclusions and outlook 207

A Appendix of chapter 4: Plane-wave vs correlation-consistent MP2 217

B Appendix of chapter 5: Plane-wave Monte Carlo MP2 231

C Appendix of chapter 7: Minnesota functionals on liquid water 245

D Appendix of chapter 8: Genetic algorithms for peptide structures 253

Bibliography 267

Curriculum Vitae 297

1 Introduction

```
Sitting in the middle of the class, I watched as the teacher showed us
how the algebra we had learned that morning could be used to describe
the movement of a marble on a path of hills and valleys.
Eight years later, I found myself in an auditorium where the ball was
replaced by an electron that was no longer observable to the naked eye,
and the field of bumps was substituted with an electromagnetic field of
nanometric wavelength. I realized that my faithful notepad and pens were
no longer sufficient to solve such an "easy" problem...
so how could we even understand the properties and behavior of molecules,
materials and chemical reactions from the fundamental laws of physics?
```

The emergence of computers since the mid-20th century has revolutionized scientific research. From the very first calculations of ballistic trajectories[1] to recent digital reconstructions and simulations of the brain,[2] computational studies have become the third pillar of the scientific toolkit, complementing traditional experimental and theoretical approaches. Their ability to synthesize data, theory, and numerical results has enabled researchers to gain a more complete understanding of the natural world.

By utilizing theoretical models and computational simulations, researchers can interpret and analyze experimental data, make predictions, and test hypotheses. In turn, these models and simulations rely on experimental data to validate the predictions, improving their accuracy and reliability. The feedback loop between experiments, theory, and computations allows scientists to refine their understanding of complex systems and develop new theories and models that can explain and predict phenomena at different scales and levels of complexity. From simulating the very elementary particles of matter to predicting the behavior of the universe, computers have opened up new possibilities for scientific discovery and enabled us to explore the natural world in unprecedented ways.

1.1 Computational chemistry

Computational chemistry is a field that combines theoretical methods, numerical implementations, and computer performance to study the properties and behavior of molecules and materials at the atomic and molecular level.[3–5] Given the vast combinatorial size of the chemical space, the potential applications of computational chemistry span across numerous areas of science and technology, including drug discovery, materials science, catalysis, and atmospheric science, among others. Atomistic computer simulations offer a fundamental advantage over experimental techniques in providing in-depth understanding of various molecular processes. Some phenomena may prove elusive to experimental observation due to the lack of sufficiently accurate analytical techniques. Moreover, computational simulations make properties accessible that are otherwise impossible to measure, either because they involve perturbation of the system under study, require complex experimental setups, or arise in dangerous or practically unreachable environments.

Most calculations in computational chemistry focus on determining the structure and total (free) energy of a target system, or the interaction energy between complexes. Beyond structural properties, computations provide valuable information on dynamical and electronic properties, such as characteristic reaction rates, spectroscopic quantities, effective cross sections for collisions, band gaps, and electrostatic features like charges, dipoles, and multipole moments. Overall, the breadth of computational results enables a detailed description of either static properties or reaction mechanisms, with identification of key intermediates and transition states involved in chemical processes. Owing to its ability to scrutinize matter at the atomic scale, computational chemistry not only enhances experimental efforts with predictive and exploratory capabilities but also aids in clarifying and rationalizing experimental observations.

1.1.1 The ultimate goal

The ultimate goal of computational chemistry is to achieve a complete and accurate representation of atomic arrangements and natural phenomena purely through in silico methods, under various thermodynamic conditions. Such comprehensive characterization would enable boundless explorations of chemical space, leading to the discovery of novel molecules and materials with desirable properties. These findings could then be directly exploited by experimentalists and industrial professionals, benefiting crucial sectors such as health, energy, mobility, and housing.

Currently, the inverse design that attempts to synthesize molecules and materials with given properties faces two primary challenges related to the size and complexity of the systems; these are the trade-off between the cost and accuracy of computational approaches, and the ability to even explore all relevant configurations of a specific

system[6] before addressing the vastness of chemical space.[7,8] As of now, computational chemistry plays instead a significant role in direct design, where properties are calculated for a given system, providing guidance on the modifications needed to achieve a desired outcome. With ongoing advancements in computational power and methodological developments, the field is poised to make even more substantial contributions to scientific discovery and technological innovation in the future.

1.1.2 The cost-accuracy trade-off

Figure 1.1 illustrates the hierarchy of some of the models available for the calculation of molecular properties and their respective cost estimates in terms of elapsed time to solution. In the ongoing pursuit of simulating increasingly larger systems ($\geq$100-100 000 atoms) within manageable time frames, modern force fields can reasonably predict the properties of solvated biological macromolecules (e.g., proteins, DNA, RNA) under equilibrium conditions.[3,4,9,10] However, they often lack explicit inclusion of polarization and tend to serve as qualitative tools for systems for which they have not been particularly parameterized. Furthermore, because they rely on classical mechanics, force fields are intrinsically incapable of describing bond breaking, thus chemical reactions, and quantum phenomena resulting from the significant spatial rearrangements of the electron distribution.

This thesis focuses on the description of atoms at the *first-principles* (ab initio) level of theory, where electrons are indeed treated as what they are, quantum particles, and their motion treated separately from the much heavier nuclei within *Born-Oppenheimer approximation*.[11] In this picture, quantum-mechanical methods are generally more accurate than force fields because they describe the rearrangement of the electrons as found ubiquitously in chemical reactions and molecular processes. However, they imply solving of the non-relativistic time-independent Schrödinger equation that unfortunately has exact analytical solutions only for the simplest one-electron systems, such as the hydrogen atom or hydrogen-like ions.

Very accurate numerical methods exist in principle, like full configuration interaction (FCI),[12] but do not allow the simulation of more than a dozen atoms due to the rapid growth of computational cost with the number of electrons. Solving the many-body problem to determine the electronic structure therefore requires approximations to the exact quantum theory.[13–16] A starting point is usually to consider the Coulombic electron-electron repulsion in an average or mean-field approximation, which yields the so called Hartree-Fock (HF) theory (Figure 1.1).[17–19] Semiempirical methods can be seen as a simplification of HF, where the two-electron part of the Hamiltonian is not explicitly evaluated.[4,20] These are much faster than HF but their performance strongly relies on empirical parameters, so that their accuracy drops down for systems not resembling the set used for parametrization. Though being relatively fast, HF neglects

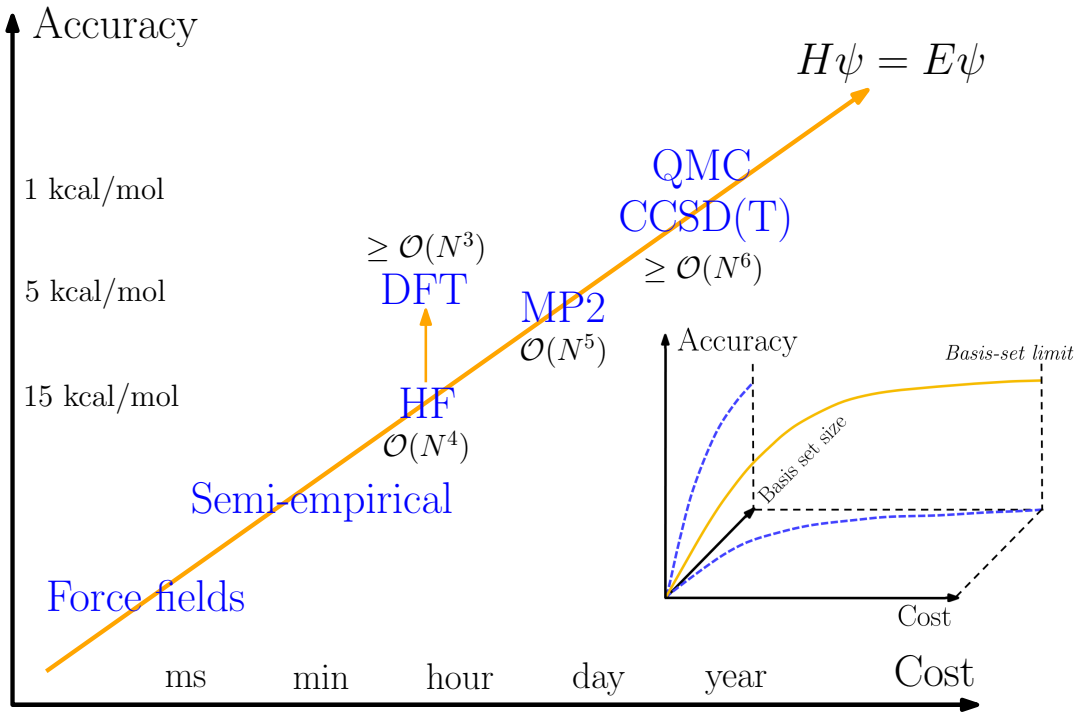


Figure 1.1: Illustration of the trade-off between cost and accuracy of computational chemistry approaches. The units of cost provide an estimate of the time required to evaluate the energy of a middle-sized system (∼100 atoms). Formal scalings of conventional implementations as a function of the number N of electrons in the system are also shown. For DFT, only pure density functionals behave like $\mathcal{O}(N^3)$. For coupled cluster, the scaling of CCSD is $\mathcal{O}(N^6)$, while it is $\mathcal{O}(N^7)$ for CCSD(T). Inset: Representation of the basis set cost-accuracy trade-off for a given quantum chemical method.

a crucial quantum contribution to the electronic problem: the *correlation* between electronic degrees of freedom. To remedy this, other methods build upon HF and account for a major fraction (≥80%) of the electron correlation like the Møller-Plesset perturbation theory at order n (MPn).[21]

Accounting for electron correlation (in the general definition) plays a crucial role in the understanding of complex phenomena like dispersion interactions, long-range order, collective behaviors, exotic phases of matter, high-temperature superconductors, Mott insulators and many more.[14,15,22,23] To that end, correlated wavefunction-based approaches like coupled cluster with single, double, and perturbative triple excitations from the HF determinant (CCSD(T))[24,25] or quantum Monte Carlo (QMC),[26] which deals with an explicit correlated wavefunction and stochastically applies the variational principle, are considered as state-of-the-art gold standards for reaching chemical accuracy (≤1 kcal/mol).[27] In practice, however, the computational cost of these latter remains substantial, hindering their routine application to many or large systems.

Striking a balance between cost and accuracy, density functional theory (DFT) emerges

as a more efficient approach that incorporates both quantum exchange and correlation effects.[28–30] Instead of relying on the electronic wavefunction, the exact formalism of DFT expresses the total energy as a functional of the electron density. However, in its widely-used Kohn-Sham (KS-DFT) variant, the interacting many-electron system is substituted by a fictitious system of non-interacting electrons (represented by Kohn-Sham orbitals) that share the same ground-state electron density as the original system. As a result, KS-DFT does not entirely eliminate the mathematical representation of electron wavefunctions.

In order to convert the continuous problem of determining the exact many-electron wavefunction into a discrete problem that can be addressed using linear algebra and numerical techniques, quantum chemical methods employ basis sets, which are selected prior to computation.[14,15] Basis sets offer a finite collection of mathematical functions that are linearly combined to represent molecular orbitals, which in turn can be used to express the many-electron wavefunction and electron density. As shown in the inset of Figure 1.1, the selection of basis functions and their linear combinations also govern the accuracy and computational cost of quantum chemical calculations. By systematically improving the basis set, such as increasing the number of basis functions, adding polarization functions, or incorporating diffuse functions, the accuracy of the molecular orbital representation can be enhanced. This results in a more accurate description of the electronic structure and, consequently, better predictions of various molecular properties. Consequently, the highest level of accuracy achievable by any quantum method is represented by its performance at the *complete basis set* (CBS) limit. Although intractable in most cases, the exact solution of the many-electron Schrödinger equation in this context would be attained by the FCI approach in the CBS limit.

As shown, the spectrum of quantum chemistry methods and basis sets offers a variety of accuracy-to-computational cost ratios, often necessitating a trade-off between accuracy and speed to obtain desired properties within a reasonable time frame. In this regard, this thesis focuses on simulating ground-state, closed-shell systems of the order of $\sim$100 first-row atoms using MP2 and DFT approaches, where the errors arising from non-relativistic and Born–Oppenheimer approximations are of minor significance.[11,31] The consideration of relativistic effects, multireference character, or excited states is consequently beyond the scope of the current work.

1.1.3 The multiple minima problem and effective sampling

In the Born-Oppenheimer approximation, the total many-body wavefunction can be expressed as the product of an electron wavefunction and a nuclear wavefunction. This decoupling allows for the solution of the electronic Schrödinger equation before addressing variations in the nuclear degrees of freedom. In the electronic Schrödinger

equation, nuclear positions are fixed and appear as constant parameters that influence the electronic wavefunction and energy eigenvalues. The nth electronic eigenvalue, in turn, defines the potential in the nth electronic state that acts on the decoupled nuclear wavefunction. Therefore, the eigenvalue of the electronic Schrödinger equation as a function of nuclear positions determines what is known as the *potential energy surface* (PES), on which the nuclei ultimately evolve.[6,32]

From a static or absolute zero-temperature perspective, the equilibrium structure of an atomic system is characterized by the global minimum of the PES. As a result, computational chemistry extensively relies on optimization algorithms capable of finding optimal or locally optimal geometries.[4,6] While such information is crucial for determining structural properties, atoms in real systems actually vibrate thermally around their equilibrium positions. Moreover, the highly-multidimensional PES is often so complex and irregular that sometimes a distinct single equilibrium structure cannot be distinguished in the low-energy range. This results from the numerous and flexible nuclear degrees of freedom that define the PES. At finite temperatures, thermodynamic (entropic) effects come into play, and the systems can visit different local minima in a probabilistic manner. The accurate description of real systems therefore relies on statistical mechanics, which calculates thermodynamic properties as statistical averages in accordance with experimental conditions. Computationally, the sampling of probabilistic structures that contribute to equilibrium properties is commonly achieved through *molecular dynamics* (MD) or Monte Carlo (MC) simulations.[9,33,34]

In both MD and MC sampling, the composition of the system under investigation is known, with the objective of studying its reaction mechanisms or equilibrium properties. The critical challenge is to ensure that the computational lens is placed onto the correct regions of the PES. In theory, MD and MC simulations should broadly sample the PES so that preferential regions are discovered and reliable statistics collected. However, this process can demand substantial or infeasible computational time in the presence of high energy barriers or structures trapped in metastable states with lifetimes exceeding the simulation duration. If resources permit, conducting longer simulations or multiple replicas with different starting structures can help address this multiple minima problem. However, as the number of atoms in the system increases, so does the number of degrees of freedom of the PES resulting in an exponentially growing number of local minima. For example, obtaining a properly folded “ab initio" protein structure from its amino acid sequence has been a tremendous ongoing challenge over the past decades.[35,36] These time scale limitations necessitate the use of so-called enhanced sampling techniques;[37–43] for instance, simulated annealing improves the interconversion between minima by heating the system to high temperatures and gradually cooling it down to the target temperature, after which statistics are collected.[41,44] Metadynamics introduces a biasing potential to progressively fill in and visit most minima, ultimately reconstructing the (free) energy surface based on the history of the filling.[39,40]

In MD, atoms are propagated on the PES using classical Newtonian equations of motion, which offer dynamic insights into a specific phenomenon or property. The accuracy of a value obtained as a time-average over an MD trajectory is directly correlated with the time scale of the fluctuations for that value. This necessitates sufficiently long simulation dynamics to ensure adequate and predictive sampling. As discussed in the previous section, quantum chemical methods enable accurate atomic-scale representations, and when combined with MD, they result in ab initio MD (AIMD) simulations where the PES is derived directly from the electronic Schrödinger equation.[32,45] Consequently, the cost-accuracy trade-off of quantum chemical methods exacerbates the challenges of effectively sampling the PES, thereby limiting the system size and time scales attainable with current (super)computers. In practice, the trade-off often favors reduced execution time over pursuing the highest possible accuracy in the underlying electronic structure method. The advent of multiscale mixed quantum mechanical/molecular mechanical (QM/MM) simulations, combined with a variety of potent enhanced sampling techniques, has expanded the spatial and temporal scales of AIMD.[45,46] With the growth in computational power, it is now feasible to carry out AIMD for systems containing several hundred to thousands of atoms over time frames of 10-100 picoseconds. Nevertheless, the three-fold challenge of system size, sampling time, and high accuracy remains substantial. Ideally, researchers would like to conduct MD simulations with the system size and sampling times typical of force field-based MD while achieving the accuracy of high-level ab initio methods.

1.2 Ways of improvement

Computational chemistry continues to evolve thanks to the ever-increasing computational power provided by advancements in processors, memory, and storage. Presently, handling systems with hundreds or thousands of electrons using post-HF or advanced DFT methods, even for non-dynamical calculations, demands significant computing power far beyond that of a conventional machine but was nearly impossible just a decade ago. Highly parallelized implementations of ab initio codes that run on high-performance supercomputers have played a major role in the success of quantum-mechanical calculations.[47–49] More recently, the emergence of GPUs has further transformed the field by notably enhancing computational power and efficiency.[50–53] This enables researchers to tackle larger and more complex systems and longer time scales while reducing calculation times, making hardware developments crucial for progress in the field. In this context, waiting can be part of the solution, as processor performance is expected to roughly double every 18 months as long as Moore's law remains valid.[54]

On the other hand, the mission of computational chemists is to advance simultaneously on theoretical development, implementation of methods, and design of innovative algorithms. Such efforts involve either providing or improving accuracy at a given

computational cost or accelerating calculations while keeping errors under control. These two aspects are at the heart of the work presented in this thesis.

1.3 Aim and outline of this thesis

The objective of this thesis is to tackle the three-fold challenge of achieving high accuracy, accommodating large system sizes, and reducing sampling time in computational chemistry, utilizing non-conventional approaches from stochastic sampling and artificial intelligence.

1.3.1 Improving the accuracy of the potential energy surface

Part I begins by discussing the key theoretical concepts involved in resolving the electronic structure of matter using wavefunction-based and DFT-based quantum mechanical methods in **Chapter 2**. It then explains how these theories are applied in numerical computations in **Chapter 3**.

Part II focuses on the calculation of the key ingredient for the implementation of double-hybrid (DH) functionals in KS-DFT, namely the second-order Møller-Plesset (MP2) correlation energy.[21] DHs are a recent development in DFT that offers improved performance over local, semilocal functionals, and hybrids by providing a more accurate description of the correlation contribution.[55–57] This is achieved by including a certain percentage of the MP2-like correlation energy evaluated using Kohn-Sham orbitals.[21,58,59] For many years, MP2 was considered the standard method for estimating non-covalent interactions,[60,61] commonly replaced by coupled-cluster (CCSD(T)) when affordable. Additionally, the first extrapolations of CCSD(T) energies to the CBS limit were made possible by composite delta-level extrapolations built on top of MP2 interaction energies.[61,62] The development of DHs can be thus seen as combining the best of both wavefunction-based and DFT approaches. Anecdotally, to highlight the significance of this combination, it is worth noting that among the 20 most cited papers in journals of the American Physical Society, 17 publications are related to DFT, along with the only wavefunction-based paper by Møller and Plesset[a].

As a prerequisite, the routine usage of DHs relies on addressing the computational challenges associated with MP2. In this sense, **Chapter 4** discusses the implementation and validation of the MP2 energy within a plane wave (PW) basis set in the CPMD molecular dynamics package,[47] that has the advantage of eventually running MD simulations or local geometry relaxations with this level of accuracy, or coupling it to more expedient sampling techniques. The advantage of PWs is their orthonormal nature, that avoids the so called basis set superposition error. In addition, PWs exhibit

[a] This list was compiled by Prof. Nicola Marzari in his lectures on DFT presented in 2020.[63]

monotonic convergence and reach completeness with the systematic increase of a single parameter, the kinetic energy cutoff, regardless of the level of theory employed. Their periodic character also makes them a basis of choice when simulating condensed phase systems at the complete basis set limit. However, only a limited number of PW codes have currently focused on providing access to the MP2 energy due to certain inherent complexities.[64,65] The first resides in the enormous amount of basis functions that is required to converge properties accurately (typically of the order of 10^5 PWs). Also, when calculating the MP2 energy, the delocalized nature of PWs makes most high-lying virtual orbitals correspond to free states (continuum-like) that are close in energy and, due to their negligible overlap with the occupied space, contribute little to the MP2 energy. Consequently, the MP2 correlation energy is obtained by summing an astronomically large number of small contributions which substantially hampers convergence. The prohibitive quintic scaling of the method with respect to system size also makes the approach very expensive for all but the smallest systems, even when substantial computational resources are available. Nonetheless, thanks to a systematic convergence to the basis-set limit,[66] a careful analysis of the parameters affecting accuracy, and our highly parallelized implementation, the calculation of non-covalent interactions of 100-electron systems has become feasible on large memory supercomputers. In **Chapter 4**, MP2 energies with PWs are thus compared to results from atom-centered bases, allowing for an assessment of the accuracy of correlated wavefunction-based methods depending on the type of basis set employed.

The investigation of stochastic approaches for quantum computations has gained momentum in the present century.[26,67–79] These approaches exhibit promising potential in expediting computational processes while effectively mitigating statistical errors. **Chapter 5** proposes an accelerated version of the conventional MP2 method by introducing a random sampling of the contributions of high-lying (continuum) virtual states to the MP2 energy. By circumventing the explicit calculation and summation of an extensive number of terms, the utilization of Monte Carlo integration enables significant acceleration of up to three orders of magnitude with only minimal compromises on accuracy. As a result, this advancement enables MP2 computations for larger systems comprising several hundreds of electrons with PWs, which were previously unattainable.

1.3.2 Navigating the potential energy surface with artificial intelligence approaches

Part III presents an overview of the techniques employed in this work to incorporate nuclear movement and navigate the PES. It contains the theoretical **Chapter 6** that outlines the principles of AIMD, which enables statistical sampling of the PES.

The impact of artificial intelligence (AI) on our daily lives is profound, transforming

various aspects ranging from speech and facial recognition to autonomous vehicles, smart robots, adapted search engines, and predictive systems for consumer behaviors and medical diagnoses. A parallel revolution is underway in the field of computational sciences, where the integration of machine learning (ML) is gradually reshaping the landscape. This transition also presents new opportunities for computational chemistry, encompassing quantum calculations[80–83] and molecular simulations.[8,84–87] The increasing availability of vast data sets, coupled with advancements in computing power and algorithms, necessitates and empowers the application of ML techniques in the field.[88–90] However, AI extends beyond ML and incorporates other but connected approaches such as Bayesian inference[91–93] and optimization methods based on evolutionary algorithms (EAs).[94–97] As such, Part IV explores the transformative potential of AI in computational chemistry and examines various approaches, ranging from kernel-based ML, EA-based optimization techniques, and unsupervised learning to enhance the sampling of the PES.

Chapter 7 showcases the advantages of employing ML to significantly accelerate the simulation of structural and dynamical properties obtained from AIMD. More precisely, when coupled to a multiple-time-step (MTS) propagation of the nuclei, ML effectively provides fast force components with sufficient accuracy, allowing for their decoupling from slower DFT-based components evaluated at larger time steps only.[45,98,99] By construction, the MTS scheme ensures that the dynamics are captured at the DFT level of accuracy employed during larger time steps. In a study focused on liquid water, this ML-MTS approach achieves speedups of 6 to 15 compared to standard Born-Oppenheimer AIMD, enabling the use of higher-level DFT functionals within feasible time frames. Thanks to this, **Chapter 7** benchmarks the accuracy of the Minnesota hybrid functionals[61,100,101] on liquid water. The findings are placed in the context of current knowledge on DFT for characterizing bulk water under ambient conditions and are compared to experimental data. Subsequently, such comparisons indicate the possible strengths and limitations of DFT for the simulation of biological systems in aqueous media.

The final chapters concern the implementation of genetic algorithms (GAs) to efficiently screen the PES. GAs, which belong to the broader class of EAs, mimic evolutionary principles to find optimal solutions for complex multi-variable problems, including PES optimization.[94,102–104] In **Chapter 8**, GAs are introduced as a reliable alternative for generating low-lying peptide structures observed in ultracold infrared spectroscopy.[105–111] By employing a fast yet reasonably accurate PES model, low-energy geometries can be identified within a few hours, compared to existing techniques relying on ab initio PES that take weeks or months. In all three test case systems, refinement of GA-generated structures using DFT reproduces geometries that align with experimental spectra, validating the time-saving advantages of this approach.

Lastly, **Chapter 9** deals with possible improvements of GAs in the exploration of

low-energy regions within the PES. By integrating GAs with an unsupervised ML technique,[104,112] notable improvements in performance are achieved. Utilizing a lightweight feature space derived from the algorithm's complete history, ML clustering not only enhances the exploration of low-lying regions but also significantly increases the probability of identifying the most probable global minimum of the PES. Importantly, these enhancements are accomplished without compromising execution times, rendering this integrated approach a powerful tool for efficient and effective PES exploration from the algorithmic point of view.

To conclude, a summary of the findings and potential future directions are finally drawn in Part V.

I wish a good reading of (parts of...) my thesis to any interested reader in the hope that it may contribute to their knowledge and inspire new advances to address the fundamental and ongoing challenge of computational chemistry; that is achieving high accuracy, accommodating large systems, and reducing sampling time.

Part I

Describing nuclei and electrons in matter

2 Electronic structure theory

`Going quantum - When the description of the marble no longer reflects reality.`

This chapter draws heavily on refs [4, 14–16, 20], which I recommend to the reader for further knowledge and details.

2.1 The interacting many-body problem

Thanks to the seminal work of quantum mechanics pioneers about a century ago, we now know that matter can be described as a collection of interacting atomic nuclei and electrons. Remarkably, the same mathematical apparatus allows to describe systems in the gas phase[113] as well as the condensed phase.[114] Indeed, be it for molecules, clusters, wires, bulk solids, liquids, surfaces, and all other possible kinds of atomic assemblies, the underlying physics at the atomic scale consists of Coulombic interactions between electrons and nuclei. As an exciting consequence, all equilibrium properties of non-relativistic systems can be, in theory, deduced by solving the non-relativistic time-independent Schrödinger equation[115]

$$\hat{\mathcal{H}}\Psi(\{\mathbf{R}\},\{\mathbf{r}\}) = \mathcal{E}\Psi(\{\mathbf{R}\},\{\mathbf{r}\}) \tag{2.1}$$

where $\{\mathbf{R}\}$ is a set of P nuclear coordinates $\{\mathbf{R}_I, I = 1,...,P\}$ and $\{\mathbf{r}\}$ is the set of N electronic coordinates $\{\mathbf{r}_i, i = 1,...,N\}$. $\mathcal{E}$ is the total energy of the system described by Ψ, the overall many-body wavefunction which incorporates intrinsic quantum information, thus properties, of electrons and nuclei. $|\Psi(\{\bar{\mathbf{R}}\},\{\bar{\mathbf{r}}\})|^2$ represents the probability density of finding P nuclei at respective positions $\{\bar{\mathbf{R}}\}$ and the N electrons located in $\{\bar{\mathbf{r}}\}$, respectively. Solutions $(\Psi,\mathcal{E})$ of eq 2.1 only exist for specific values of energy imposed by the quantization of bound states in quantum mechanics. While electrons are fermions of half-integer spin, nuclear species can be either of fermionic

or bosonic nature depending on the nuclear spin. Thus, the total wavefunction Ψ is antisymmetric with respect to the exchange of coordinates of two electrons and either antisymmetric or symmetric in the exchange of nuclear positions.

In the absence of external fields, and neglecting magnetic interactions in the system, the non-relativistic time-independent many-body Hamiltonian takes the form[a]

$$\hat{\mathcal{H}} = \underbrace{-\frac{1}{2}\sum_{I=1}^{P}\frac{1}{M_I}\nabla_I^2}_{\hat{T}_N} \underbrace{-\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2}_{\hat{T}_e} \underbrace{-\sum_{I,i=1}^{P,N}\frac{Z_I}{|\mathbf{R}_I-\mathbf{r}_i|}}_{\hat{V}_{eN}} + \underbrace{\sum_{i<j}^{N}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|}}_{\hat{V}_{ee}} + \underbrace{\sum_{I<J}^{P}\frac{Z_I Z_J}{|\mathbf{R}_I-\mathbf{R}_J|}}_{\hat{V}_{NN}} \tag{2.2}$$

that is composed of the nuclear $\hat{T}_N$, and electronic $\hat{T}_e$, kinetic contributions as well as Coulomb pairwise attractive interactions $\hat{V}_{eN}$ between the electrons and the nuclei. The Coulomb repulsion between electrons is given by $\hat{V}_{ee}$, with $\hat{V}_{NN}$ its analogue between nuclei. M_I and Z_I are, respectively, the nuclear masses and charges.

2.1.1 Born-Oppenheimer approximation

Getting an accurate form of the many-body wavefunction Ψ is in practice extremely difficult for realistic systems, since pairwise operators in the Hamiltonian (2.2) imply that particle degrees of freedom are interdependent, or *correlated* in a quantum-statistical language. A simplification of the problem is achieved within the Born-Oppenheimer approximation. This latter relies on the fact that the movement of the (heavy) nuclei is seen as very slow in the time frame of the lighter electrons, so that electrons and nuclei decouple into separate wavefunctions. By neglecting the coupling between electronic and nuclear degrees of freedom, it is supposed that electrons instantaneously arrange themselves in their ground state for any change in the nuclear geometry. In most cases, this approximation remains valid because the energy difference between electronic states is big enough so that the coupling between electronic states remains negligible upon small displacements of the nuclei. When electrons and nuclei are seen as evolving on decoupled time scales, the many-particle wavefunction is mathematically separable into nuclei and electronic wavefunctions

$$\Psi(\{\mathbf{R}\},\{\mathbf{r}\}) = \Theta_n(\{\mathbf{R}\})\Phi_e(\{\mathbf{r}\};\{\mathbf{R}\}) \tag{2.3}$$

where $\Theta_n(\{\mathbf{R}\})$ is the nuclear wavefunction, and the electronic wavefunction $\Phi_e(\{\mathbf{r}\};\{\mathbf{R}\})$ therefore accounts for the electron distribution at fixed nuclei in $\{\mathbf{R}\}$, whose inclusion, importantly, acts as a parametric dependency.

In the Born-Oppenheimer approximation, putting aside the treatment of the nuclei for a moment, the interacting many-body problem reduces first into the following

[a] Hartree atomic units are used throughout this thesis, unless otherwise specified.

many-electron problem:

$$\hat{\mathcal{H}}_e \Phi_e(\{\mathbf{r}\};\{\mathbf{R}\}) = E(\{\mathbf{R}\})\Phi_e(\{\mathbf{r}\};\{\mathbf{R}\}) \tag{2.4}$$

with the electronic Hamiltonian being

$$\hat{\mathcal{H}}_e = \sum_{i=1}^{N}\left[-\frac{1}{2}\nabla_i^2 - \sum_{I=1}^{P}\frac{Z_I}{|\mathbf{R}_I - \mathbf{r}_i|}\right] + \sum_{i<j}^{N}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{2.5}$$

Solving eq 2.4 and thus getting the ground state $E(\{\mathbf{R}\})$ for any set of fixed nuclear positions $\{\mathbf{R}\}$ defines what is called the potential energy surface (PES). This hypersurface defines the potential that enters the ultimate quantum equation to solve for the nuclei. Indeed, according to the Born-Oppenheimer approximation, once the many-electron problem is solved, the total Hamiltonian becomes

$$\langle \Phi_e(\{\mathbf{r}\};\{\mathbf{R}\})|\hat{\mathcal{H}}|\Phi_e(\{\mathbf{r}\};\{\mathbf{R}\})\rangle \simeq \hat{h}_n + E(\{\mathbf{R}\}) \tag{2.6}$$

with the remaining nuclear Hamiltonian being

$$\hat{h}_n = -\frac{1}{2}\sum_{I=1}^{P}\frac{1}{M_I}\nabla_I^2 + \sum_{I<J}^{P}\frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} \tag{2.7}$$

This allows finally to solve the many-body Schrödinger equation (eq 2.1) via the solution of the nuclear Schrödinger-like equation

$$\left[\hat{h}_n + E(\{\mathbf{R}\})\right]\Theta_n(\{\mathbf{R}\}) = \mathcal{E}_{\mathrm{BO}}\Theta_n(\{\mathbf{R}\}) \tag{2.8}$$

where $\mathcal{E}_{\mathrm{BO}}$ is the total (approximated) energy of the many-body system.

Decoupling the nuclear and electronic motions is a common and justified approximation in quantum chemistry, as the significant difference in mass between protons/neutrons and electrons (a factor of about 1800) allows for the simplification of calculations. This separation is essential for the concept of PES and the parametric description of equilibrium and transition structures in terms of nuclear coordinates. However, there are cases where the Born-Oppenheimer approximation fails, and the quantum mechanical nature of the nuclei cannot be neglected. This thesis focuses on developing methods to address the cost-accuracy trade-off in evaluating the PES, specifically by solving the electronic Schrödinger equation (eq 2.4). The electronic wavefunction is described quantum mechanically, accounting for the Coulombic interactions between electrons. Due to the many-body nature of the problem, finding exact solutions to the electronic Schrödinger equation is computationally demanding, as it requires solving a partial differential equation with $3N$ degrees of freedom. Exact analytical solutions are only tractable for two-body systems. The electronic structure of many-electron atoms, molecules, solids or electron gases is only obtainable through approximate nu-

merical methods. To illustrate the complexity, consider an iron atom with 26 electrons described on a cubic grid of 10 points per dimension. Storing the entire wavefunction for this simple case would require the saving of $10^{3 \cdot 26}$ real numbers. Considering that each floating-point number requires 8 bytes, the total memory requirements for handling such a wavefunction would finally necessitate $1.25 \cdot 10^{66}$ terabytes, which is unimaginable from the hardware point of view, rendering exact calculations practically impossible. Therefore, approximations are necessary to address the scaling issues associated with the number of electrons, leading to the development of numerous methods to tackle the electronic structure problem. The following sections will provide insights into some of the ab initio methods that aim to solve the electronic problem.

2.1.2 Variational principle

The variational principle is a fundamental concept in quantum chemistry that serves as a cornerstone for many computational methods. It provides a powerful framework for finding approximate solutions to the Schrödinger equation by minimizing the expectation value of the energy with respect to a trial wavefunction Φ_{trial}. Mathematically, the variational principle can be expressed as

$$E_{\text{var}} \geq E_0 \tag{2.9}$$

where E_{var} is the variational energy obtained from the trial wavefunction and E_0 is the exact ground state energy of the problem. For the resolution of the electronic structure in the Born-Oppenheimer approximation (eq 2.4), the variational energy is expressed as

$$E_{\text{var}} = \frac{\langle \Phi_{\text{trial}} | \hat{\mathcal{H}}_e | \Phi_{\text{trial}} \rangle}{\langle \Phi_{\text{trial}} | \Phi_{\text{trial}} \rangle} \tag{2.10}$$

with $\hat{\mathcal{H}}_e$ representing the electronic Hamiltonian operator, which describes the total energy of the electron system in the parametric field of the nuclei. The variational principle states that the variational energy will always be greater than or equal to the exact ground state energy, ensuring that the variational approach provides an upper bound to the true energy.

By selecting an appropriate trial wavefunction and optimizing its parameters, one can systematically improve the accuracy of the energy approximation. Various computational methods, such as Hartree-Fock (HF) theory (Section 2.2.1), configuration interaction (CI) (Section 2.2.4), and density functional theory (DFT) (Section 2.3), are built upon the variational principle and employ different strategies for constructing and optimizing the trial wavefunction (respectively, the electron density). The variational principle provides a versatile tool for developing efficient and accurate quantum chemical methods that enable the study of complex molecular systems and their properties.

2.2 Wavefunction-based methods

Wavefunction methods, traditionally embraced by quantum chemists,[13,14] have been distinguished from approaches that focus on the electronic density as the primary variable, such as DFT, which has gained popularity among physicists[28,116] due to its ability to significantly reduce the number of degrees of freedom. However, in recent years, wavefunction-based approaches have emerged as viable options in condensed matter physics as well.[117–122] These methods have found utility not only as standalone techniques but also in conjunction with density-based approaches,[58,59,123–125] allowing for the exploration of complex systems and capturing a broader range of electronic properties.

Dictated by the fundamental postulates of quantum mechanics, the electronic wavefunction must adhere to certain criteria. One crucial aspect is the consideration of spin, a quantum number that plays a central role in numerous electronic phenomena. Working with an electronic wavefunction that is an eigenstate of the spin operator is therefore highly advantageous in this regard:

$$\hat{S}^2 \left|\Phi_e\right\rangle = S(S+1) \left|\Phi_e\right\rangle \tag{2.11}$$

$$\hat{S}_z \left|\Phi_e\right\rangle = S_z \left|\Phi_e\right\rangle \tag{2.12}$$

Furthermore, electrons are fermions, meaning they are indistinguishable spin-1/2 particles. As a result, the many-electron wavefunction must exhibit antisymmetry upon the exchange of particles. This requirement arises from Pauli's exclusion principle, which states that two electrons cannot occupy the same quantum state. The influential Hartree-Fock theory tackles this property by incorporating the necessary antisymmetrization of the wavefunction, as described below.

2.2.1 Hartree-Fock method

In order to encode Pauli's principle, the Hartree-Fock (HF) method[17–19] introduces an antisymmetrized many-electron (trial/ansatz) wavefunction in the form of a single Slater determinant

$$\Phi_{\mathrm{HF}}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_N(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_N(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \cdots & \phi_N(\mathbf{x}_N) \end{vmatrix} \tag{2.13}$$

where $\phi_i(\mathbf{x}_j)$ is the i-th one-electron spin orbital of respective spatial and spin components $\mathbf{x}_j = (\mathbf{r}_j, \sigma_j)$. Note how the problem of keeping track of the entire electronic wavefunction is reduced by the Hartree-Fock ansatz. Indeed, it requires storing only

orbitals whose spatial variables are extended over a space mesh. For our previous example on the iron atom (Section 2.1.1), this leads to the storage of $26 \cdot 10^3$ floating-point numbers, against $10^{3 \cdot 26}$ floating-point numbers that were needed to account for a fully numerical wavefunction.

The ground state energy of the system is obtained from the variational principle on the total energy with the ansatz wavefunction:

$$E_{\text{HF}} = \frac{\langle \Phi_{\text{HF}} | \, \hat{\mathcal{H}}_e \, | \Phi_{\text{HF}} \rangle}{\langle \Phi_{\text{HF}} | \Phi_{\text{HF}} \rangle} \tag{2.14}$$

In practice, Φ_{HF} has to be projected onto a finite basis, that allows to workout linear algebra for the computational minimization of E_{HF}. Details about the choice of basis will be provided later (Section 3.1). For the time being, it is sufficient to remember that the spin orbitals $\phi_i(\mathbf{x}_j)$ can be further expanded as linear combinations in an auxiliary basis. The coefficients of such linear combinations therefore consists in the ultimate parameters to optimize to find the ground state wavefunction Φ_{HF} that minimizes E_{HF}. To facilitate the formalism, let us rewrite the electronic Hamiltonian (eq 2.5) in terms of one and two-body operators

$$\hat{\mathcal{H}}_e = \sum_{i=1}^{N} \hat{h}_1(i) + \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \hat{v}_2(i,j) \tag{2.15}$$

with

$$\hat{h}_1(i) = -\frac{1}{2}\nabla_i^2 \underbrace{- \sum_{I=1}^{P} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}_i|}}_{v_{ext}(\mathbf{r}_i)} \tag{2.16}$$

which is the one-electron operator acting on the electron i immersed in the external potential v_{ext} due to the fixed nuclei. The two-body operator comes from the Coulomb electron-electron interaction in the absence of spin-orbit coupling:

$$\hat{v}_2(i,j) = \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{2.17}$$

Injecting eqs 2.13 and 2.15 in eq 2.14, and working out the algebra a bit,[4] gives the following expression for the HF variational energy:

$$E_{\text{HF}} = \sum_{i=1}^{N} \langle i | \, \hat{h}_1 \, | i \rangle + \underbrace{\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\langle ij | \, \hat{v}_2 \, | ij \rangle - \langle ij | \, \hat{v}_2 \, | ji \rangle \right)}_{\langle V_{ee} \rangle_{\text{HF}}} \tag{2.18}$$

where $\langle \mathbf{x} | i \rangle$ denotes the spin orbital $\phi_i(\mathbf{x})$, in association with the one-electron integrals

defined by

$$\langle i|\,\hat{h}_1\,|j\rangle = \int d\mathbf{x}_1\ \phi_i^*(\mathbf{x}_1)\hat{h}_1(\mathbf{r}_1)\phi_j(\mathbf{x}_1) \tag{2.19}$$

and the two-electron integrals given by

$$\langle ij|\,\hat{v}_2\,|kl\rangle = \iint d\mathbf{x}_1 d\mathbf{x}_2\ \phi_i^*(\mathbf{x}_1)\phi_j^*(\mathbf{x}_2)\frac{1}{|\mathbf{r}_1-\mathbf{r}_2|}\phi_k(\mathbf{x}_1)\phi_l(\mathbf{x}_2) \tag{2.20}$$

Looking closer at the two-body terms in eq 2.18, it is noticed that the sums over the first integrals $\langle ij|\,\hat{v}_2\,|ij\rangle$ correspond to the classical Coulomb electrostatic energy between two charge distributions $|\phi_i(\mathbf{x})|^2$ and $|\phi_j(\mathbf{x}')|^2$. However, the second term $\langle ij|\,\hat{v}_2\,|ji\rangle$ does not match any classical equivalence and represents a quantum stabilizing effect coming from the antisymmetrized nature of the wavefunction (Pauli's principle). Similar to the Coulomb term apart from the exchange between two spin orbitals, the quantity

$$E_x^{\mathrm{HF}} = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\langle ij|\,\hat{v}_2\,|ji\rangle \tag{2.21}$$

is therefore universally defined as the *exchange energy* among both the wavefunctions and DFT computational communities.

The HF ground state can now be obtained from the minimization of the variational energy E_{HF} with respect to the spin orbitals ϕ_i, with the constraint that such orbitals are orthogonal due to the principles of quantum mechanics. The constraint optimization can be realized with the Lagrange's method and the introduction of the unknown multipliers λ_{ij} that take the dimension of energy. Such an operation gives the HF equations that have to be satisfied by the HF ground state orbitals

$$\hat{F}\phi_i(\mathbf{x}) = \sum_{j=1}^{N}\lambda_{ij}\phi_j(\mathbf{x}) \tag{2.22}$$

where $\hat{F}$ is the Fock operator defined as

$$\hat{F} = \hat{h}_1 + \sum_{j=1}^{N}\left(\hat{J}_j - \hat{K}_j\right) \tag{2.23}$$

that includes a combination of Coulomb $\hat{J}_j$ and exchange $\hat{K}_j$ operators, respectively, that act on the orbitals like

$$\hat{J}_j\phi_i(\mathbf{x}) = \left(\int d\mathbf{x}'\ \frac{\phi_j^*(\mathbf{x}')\phi_j(\mathbf{x}')}{|\mathbf{r}'-\mathbf{r}|}\right)\phi_i(\mathbf{x}) \tag{2.24}$$

$$\hat{K}_j\phi_i(\mathbf{x}) = \left(\int d\mathbf{x}'\ \frac{\phi_j^*(\mathbf{x}')\phi_i(\mathbf{x}')}{|\mathbf{r}'-\mathbf{r}|}\right)\phi_j(\mathbf{x}) \tag{2.25}$$

When acting on the same spin orbital i, it is observed that the Coulomb $\hat{J}_i$ and exchange $\hat{K}_i$ contributions to the Fock operator cancel each other, which emphasizes that HF is free from self-interaction error. In HF theory, therefore, one electron does not "feel" its own presence, but interacts with all other electrons. However, particles are independent in the HF approximation because each electron moves in an effective potential, represented by the attraction of the nuclei (in $\hat{h}_1$), and the average repulsive effect of surrounding electrons screened by the exchange operator. This becomes evident from the Coulomb part of the Fock operator applied on the spin orbital $\phi_i(\mathbf{x})$

$$\sum_{j\neq i}^{N} \hat{J}_j \phi_i(\mathbf{x}) = \sum_{j\neq i}^{N} \left(\int d\mathbf{x}' \, \frac{|\phi_j(\mathbf{x}')|^2}{|\mathbf{r}' - \mathbf{r}|} \right) \phi_i(\mathbf{x}) \tag{2.26}$$

that is the Coulomb interaction of the i-th electron with the average charge distribution of the other electrons.

Solving the HF equations does not give a unique set of solutions for the one-electron spin orbitals. Indeed, it can be checked that any unitary transformation U applied on the spin orbitals does not alter the Slater determinant, nor the Fock operator.[13] It is then always possible to find new orbitals $\phi'_i(\mathbf{x})$ such that

$$\phi'_i(\mathbf{x}) = \sum_j U_{ij} \phi_j(\mathbf{x}) \qquad \text{with} \qquad \sum_k U^*_{ik} U_{kj} = \delta_{ij} \tag{2.27}$$

Thus, the most convenient and common choice of orbitals is the canonical representation, in which the matrix of Lagrange multipliers becomes diagonal ($\lambda_{ij} := \varepsilon_i \delta_{ij}$) such that the HF problem (eq 2.22) translates into

$$\hat{F} \phi'_i(\mathbf{x}) = \varepsilon_i \phi'_i(\mathbf{x}) \tag{2.28}$$

This representation has the particularity of accessing the HF orbitals from an eigenvalue (Schrödinger-like) equation. The eigenvalues ε_i can consequently be interpreted as excitation energies of the canonical orbitals $\phi'_i(\mathbf{x})$. Moreover, the Koopmans's theorem states that the ionization energy coming from the removal of an electron of orbital $\phi'_i(\mathbf{x})$ is equal to the negative of its eigenvalue[126]

$$I = E^{i}_{\mathrm{HF}}(N-1) - E_{\mathrm{HF}}(N) = -\varepsilon_i \tag{2.29}$$

This observation highlights the significance of occupied energies in the HF approximation. Ionization energies obtained within HF are typically accurate, although they neglect the relaxation of the remaining orbitals when an electron is removed. In contrast, virtual (empty) orbitals lack physical interpretation and are generally unbound. They arise as a consequence of the additional dimensions in the eigenvalue problem. However, these additional wavefunctions can still be employed as a (complete) basis for perturbation theory, as will be demonstrated in the subsequent discussion.

Let us go back to the expression of the HF energy depending on the eigenvalues of the Fock operator. Playing around with the previous definitions, one obtains

$$\varepsilon_i = \langle i| \hat{F} |i\rangle = \langle i| \hat{h}_1 |i\rangle + \sum_{j=1}^{N} \left(\langle ij| \hat{v}_2 |ij\rangle - \langle ij| \hat{v}_2 |ji\rangle \right) \tag{2.30}$$

which, when compared to eq 2.18, leads to

$$E_{\text{HF}} = \sum_{i=1}^{N} \varepsilon_i - \underbrace{\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\langle ij| \hat{v}_2 |ij\rangle - \langle ij| \hat{v}_2 |ji\rangle \right)}_{\langle V_{ee} \rangle_{\text{HF}}} \tag{2.31}$$

that highlights the important fact that the HF energy is not equal to the sum of the orbital energies. An additional term should indeed correct for the double-counting of the mean-field interaction $\langle V_{ee} \rangle_{\text{HF}}$ in the one-electron eigenvalues.

Formally, the HF method scales as $\mathcal{O}(N^4)$ with the number of electrons due to the evaluation of the two-electron integrals. The main approaches for solving the HF equations in eq 2.28 are either purely numerical (exact HF) or rely on a set of basis functions that define the algebraic framework in which operations on matrices will operate (Hartree-Fock-Roothaan equations).[13] In both cases, direct diagonalization of the Fock operator cannot be performed, as it depends itself on the orbitals. Instead, an iterative process is employed. Initially, a guess for the orbitals is made, and these are then used to construct the Fock operator. Equations 2.28 are solved for this guess, providing new solution orbitals to be compared with the guess. This process is iteratively repeated until solution orbitals are consistent with those defining the Fock operator, which gives the HF method its name as a self-consistent-field (SCF) approach.

2.2.2 Electron correlation

In the previous section, we saw how the HF ansatz for the wavefunction, coupled to the variational principle, leads to an upper bound of the exact ground state energy E_0. In practice, it appears that the HF theory gives insufficient flexibility in the wavefunction to recover exact energies accurately. From the fact that HF is a mean-field approximation, it fails to explicitly account for *electron correlation* effects, which arise from the complex and dynamic interactions between electrons. As a consequence, the electron correlation energy E_c is thus formally defined as being the energy missing between the exact energy and the HF energy

$$E_c = E_0 - E_{\text{HF}} \tag{2.32}$$

The consideration of electron correlation is crucial to accurately describe the electronic structure of molecules and solids. Although the HF approximation accounts for the

exact exchange, it completely ignores the correlation energy due to the form of its many-body ansatz for the wavefunction: a single Slater determinant. To go beyond the HF approximation and incorporate electron correlation effects, various post-Hartree-Fock methods have been developed (Section 2.2.4) and employ more sophisticated trial wavefunctions. Examples include configuration interaction (CI) methods, coupled cluster (CC) methods, and multi-reference methods, among others.[13] In these, the electronic wavefunction is expanded as a linear combination of Slater determinants, or as an exponential series, that enables more flexibility for the description of electron correlation. The CI method, for example, includes multiple determinants to capture the contributions of different electron configurations. The CC method incorporates a hierarchy of excitations from the HF reference determinant. These methods aim to systematically improve the accuracy of the wavefunction by including higher-order correlation contributions, that may address different types of correlation effects. Indeed, correlation can be conceptually divided into static and dynamic origins. Static correlation arises when multiple electronic configurations (Slater determinants) contribute significantly to the ground state wavefunction. This typically occurs in cases involving near-degeneracies or strong electron-electron correlation effects that require so called multi-reference methods. Static correlation relates to the improvement in the energy accuracy when expanding the wavefunction in the configuration (determinant) space. Dynamic correlation, on the other hand, refers to the dynamical motion of electrons and their response to changes in the electronic environment. It involves the correlation effects arising from electron-electron reorganization during molecular excitations and reactions and is therefore essential for accurately describing molecular properties such as bond dissociation, reaction barriers, and spectroscopic properties.

There are three main lines of research to achieve the holy grail of correlation energy calculation in wavefunction-based methods. The first resides in finding the optimal combination of Slater determinants (CI-like). The second introduces explicit terms in *correlated wavefunction* ansatz to recover most of the correlation energy (e.g., variational quantum Monte Carlo).[26] And the last treats the HF theory as a starting point for perturbation theory. This thesis focuses precisely on the latter. Finally, it should be remembered that the numerical solution of the electronic problem relies on basis functions (Section 3.1) which themselves have an impact on the accuracy of the energy if they do not cover the Hilbert space automatically and completely. Therefore, to obtain the exact absolute ground state energy E_0 of a system, the energy E_{HF} and the correlation energy E_c in eq 2.32 would have to be calculated in the *complete basis set* limit.

2.2.3 Rayleigh–Schrödinger perturbation theory

Generally, the HF approximation already covers a major part of the total electron energy that makes it a good starting point for a perturbative approach.[127] Let us define

$\hat{H}_0$ which is the reference Hamiltonian of a problem for which the exact solution is known and for which the wavefunction is non-degenerate:

$$\hat{H}_0\Phi_i^{(0)} = E_i^{(0)}\Phi_i^{(0)} \qquad i = 0, 1, ..., \infty \tag{2.33}$$

Let us assume now that we want to solve a more complex problem whose Hamiltonian $\hat{H}$ differs slightly from $\hat{H}_0$ so that

$$\hat{H} = \hat{H}_0 + \lambda\Delta\hat{H} \tag{2.34}$$

where the parameter λ has been introduced in order to control the time-independent perturbation $\Delta\hat{H}$. For $\lambda = 0$, the problem is the unperturbed reference while the full perturbed Hamiltonian stands for $\lambda = 1$.

The solution of the perturbed Hamiltonian is then given by

$$\hat{H}\Phi_i = E_i\Phi_i \qquad i = 0, 1, ..., \infty \tag{2.35}$$

With this, the new energy and wavefunction should change continuously as the perturbation is increased, that is when λ varies from 0 to 1. This consequently allows the following Maclaurin expansions:

$$E_i = \lambda^0 E_i^{(0)} + \lambda^1 E_i^{(1)} + \lambda^2 E_i^{(2)} + \lambda^3 E_i^{(3)} + ... \tag{2.36}$$

$$\Phi_i = \lambda^0 \Phi_i^{(0)} + \lambda^1 \Phi_i^{(1)} + \lambda^2 \Phi_i^{(2)} + \lambda^3 \Phi_i^{(3)} + ... \tag{2.37}$$

Since these expansions hold for any value of λ, one can insert them in eq 2.35 and equating the various powers of λ. Since the overall phase of the wavefunction is undetermined in quantum mechanics, it is convenient to choose the perturbed wavefunction so that $\langle\Phi_i|\Phi_i^{(0)}\rangle = 1$. This has the consequence that all correction terms are orthogonal to the reference wavefunction $\langle\Phi_i^{(n\neq 0)}|\Phi_i^{(0)}\rangle = 0$. The zero-order logically gives the unperturbed wavefunction and energy:

$$E_i^{(0)} = \langle\Phi_i^{(0)}|\,\hat{H}_0\,|\Phi_i^{(0)}\rangle \tag{2.38}$$

At first-order, the correction to the energies is the expectation value of the perturbation in the unperturbed eigenstates:

$$E_i^{(1)} = \langle\Phi_i^{(0)}|\,\Delta\hat{H}\,|\Phi_i^{(0)}\rangle \tag{2.39}$$

while the correction to the second-order starts to involve the first-order correction to

the eigenstates:

$$E_i^{(2)} = \langle \Phi_i^{(0)} | \, \Delta\hat{H} \, | \Phi_i^{(1)} \rangle = \sum_{j \neq i}^{\infty} \frac{| \, \langle \Phi_i^{(0)} | \, \Delta\hat{H} \, | \Phi_j^{(0)} \rangle \, |^2}{E_i^{(0)} - E_j^{(0)}} \tag{2.40}$$

In order to get an expression for $|\Phi_i^{(1)}\rangle$, this latter was expanded in the complete set of orthogonal functions given by the reference solution $\{\Phi_i^{(0)}\}_{i=0}^{\infty}$ to finally get

$$|\Phi_i^{(1)}\rangle = \sum_{j \neq i}^{\infty} \frac{\langle \Phi_j^{(0)} | \, \Delta\hat{H} \, | \Phi_i^{(0)} \rangle}{E_i^{(0)} - E_j^{(0)}} \, |\Phi_j^{(0)}\rangle \tag{2.41}$$

from the first-order equations in λ. This is known as the Rayleigh-Schrödinger perturbation theory. The same recursive logic can be continued in order to get higher-order corrections to the eigenstates and energies but analytical expressions become horribly complex. It must be emphasized that the solution to the perturbed system will depend a lot on the perturbation, the quality of the reference problem, and the convergence properties of the expansion.

2.2.4 Post-Hartree-Fock methods

Second-order Møller-Plesset perturbation theory

When aiming to faithfully recover the electron-electron correlation, a reasonable guess would be to consider the HF problem as a non-interacting reference system for a perturbation theory. While possible in principle, there is no guarantee that the method converges fast or even monotonically to the exact energy. However, the advantage of the HF approximation is that it can generally account for between 80% to 99% of the exact energy which therefore makes it a good starting point that shows small deviations with respect to the perturbed Hamiltonian.

Based on the Rayleigh-Schrödinger perturbation theory, the Møller-Plesset perturbation theory[21] takes as reference Hamiltonian $\hat{H}_0$ the sum over single-particle Fock operators (eq 2.23) that writes

$$\hat{H}_0 = \sum_{i=1}^{N} \hat{F}(i) = \sum_{i=1}^{N} \hat{h}_1(i) + \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\hat{J}_j(i) - \hat{K}_j(i) \right) \tag{2.42}$$

so that the perturbation is given by the difference between the full electronic Hamiltonian (eq 2.15) and such HF terms:

$$\Delta\hat{H} = \hat{\mathcal{H}}_e - \hat{H}_0 = \sum_{i<j}^{N} \hat{v}_2(i,j) - \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\hat{J}_j(i) - \hat{K}_j(i) \right) \tag{2.43}$$

Resorting to eq 2.38, the zeroth-order wavefunction is nothing else than the HF determinant, that we choose in the canonical basis (eq 2.28) in order to write the zeroth-order energy as a sum over eigenvalues of the orbitals

$$E_{\mathrm{MP}}^{(0)} = \sum_{i=1}^{N} \langle \Phi_{\mathrm{HF}} | \, \hat{F}(i) \, | \Phi_{\mathrm{HF}} \rangle = \sum_{i=1}^{N} \langle \phi_i | \, \hat{F}(i) \, | \phi_i \rangle = \sum_{i=1}^{N} \varepsilon_i \tag{2.44}$$

The first-order energy correction is given by eq 2.39 which, with some small additional effort,[4] translates in the Møller-Plesset approach into

$$E_{\mathrm{MP}}^{(1)} = \langle \Phi_{\mathrm{HF}} | \, \Delta\hat{H} \, | \Phi_{\mathrm{HF}} \rangle = \langle V_{ee} \rangle_{\mathrm{HF}} - 2 \langle V_{ee} \rangle_{\mathrm{HF}} = - \langle V_{ee} \rangle_{\mathrm{HF}} \tag{2.45}$$

where we had defined

$$\langle V_{ee} \rangle_{\mathrm{HF}} = \sum_{i<j}^{N} \langle \Phi_{\mathrm{HF}} | \, \hat{v}_2(i,j) \, | \Phi_{\mathrm{HF}} \rangle = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\langle ij | \, \hat{v}_2 \, | ij \rangle - \langle ij | \, \hat{v}_2 \, | ji \rangle \right) \tag{2.46}$$

as the expectation value of the electron-electron interaction in the HF ground state. Having that said, it is obtained by comparaison with eq 2.31 that the total energy at first-order is nothing else than the HF energy

$$E_{\mathrm{MP1}} = E_{\mathrm{MP}}^{(0)} + E_{\mathrm{MP}}^{(1)} = \sum_{i=1}^{N} \varepsilon_i - \langle V_{ee} \rangle_{\mathrm{HF}} = E_{\mathrm{HF}} \tag{2.47}$$

This implies that the correlation energy is only retrieved when getting to higher orders. The first non-trivial correction to the HF problem is provided by the second-order term (MP2) which, according to eq 2.40, involves excited Slater determinants of the HF ansatz like those appearing in a CI expansion. Fortunately, thanks to Brillouin's theorem and the fact that the many-body perturbation $\Delta\hat{H}$ possesses only two-body operators, its matrix elements are only non-zero for excited determinants that differ by no more than two excitations from the HF reference ground state. Those determinants, that we write $|\Phi_{ij}^{ab}\rangle$, are generated by exciting two electrons from the occupied orbital i,j to the virtual orbitals a, b. With these, the second-order Møller-Plesset perturbation energy yields

$$E_{\mathrm{MP}}^{(2)} = \sum_{i<j}^{N_{\mathrm{occ}}} \sum_{a<b}^{N_{\mathrm{vir}}} \frac{| \langle \Phi_{\mathrm{HF}} | \, \Delta\hat{H} \, | \Phi_{ij}^{ab} \rangle |^2}{E_{\mathrm{HF}} - E_{ij}^{ab}} = \sum_{i<j}^{N_{\mathrm{occ}}} \sum_{a<b}^{N_{\mathrm{vir}}} \frac{| \langle ij | \, \hat{v}_2 \, | ab \rangle - \langle ij | \, \hat{v}_2 \, | ba \rangle |^2}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{2.48}$$

where N_{occ} and N_{vir} are respectively the number of occupied and virtual orbitals included in the solution of the HF problem. For closed-shell systems, one can adopt the restricted case where spin orbitals are occupied by two electrons of opposite spins,

which transforms the MP2 expression into sums over spatial orbitals

$$E^{(2)}_{\text{MP, rest}} = \sum_{i=1}^{N_{\text{occ}}} \sum_{j=1}^{N_{\text{occ}}} \sum_{a=1}^{N_{\text{vir}}} \sum_{b=1}^{N_{\text{vir}}} \frac{\langle ij|\,\hat{v}_2\,|ab\rangle\,(2\,\langle ab|\,\hat{v}_2\,|ij\rangle - \langle ab|\,\hat{v}_2\,|ji\rangle)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{2.49}$$

As a result, the second-order Møller-Plesset perturbation theory approximates the total energy as the sum of the HF energy and the contribution of the MP2 correlation energy, that finally writes

$$E_{\text{MP2}} = E_{\text{HF}} + E^{(2)}_{\text{MP}} \tag{2.50}$$

MP2 is a common method choice in computational chemistry as it already provides 80-90% of the correlation energy while being the cheapest correlated wavefunction-based method.[16] Geometries obtained with MP2 are also usually close to higher-level methods or experiments. Also non-negligibly, MP2 correctly reproduces the $-C_6/R^6$ behavior of dissociation curves of van der Waals systems at large distance R, meaning that it reliably accounts for long-range dispersion effects. This is partly responsible for the good accuracy of double-hybrid DFT functionals that incorporate a fraction of the Kohn-Sham MP2 correlation (Section 2.3.2). Such a description of dispersion is typically absent in HF and in most lower-level purely local DFT approaches (LDA, GGA),[128] as illustrated in Figure 2.1. For supposedly more accuracy, one could of course go to higher orders in the perturbation expansion (MP3, MP4, ..., MPn). However, the MPn approach is not variational and has no monotonic convergence to the exact energy when going to higher orders. Another possible limitation of the Møller-Plesset perturbation theory is the foundation on the HF problem. In case the HF solution is too far from the exact solution of the perturbed Hamiltonian, the perturbation expansion will require more terms to be accurate, or may not converge at all. Furthermore, if the systems investigated require more than a single Slater determinant in the wavefuntion (e.g., in transition metals, superconductors or transition metal oxides), the HF reference will fail at providing a reliable starting point. In this situation, multi-reference MP techniques should be used (e.g., CASPT2).

Resolution of the identity and density fitting

The most time consuming part of the MP2 method is the evaluation of two-electron four-index integrals of the type $\langle ij|\,\hat{v}_2\,|ab\rangle$ entering the MP2 correlation energy expression. A solution to speed up this step consists of reducing these integrals to three indices with the help of an auxiliary orthonormal basis set $\{\chi'_\alpha\}_{\alpha=1}^{M'}$ and the resolution of identity (RI)

$$\mathbb{1} \simeq \sum_{\alpha=1}^{M'} |\chi'_\alpha\rangle\langle\chi'_\alpha| \tag{2.51}$$

that remains an approximation for a finite basis set of size M'. Introduced in front of

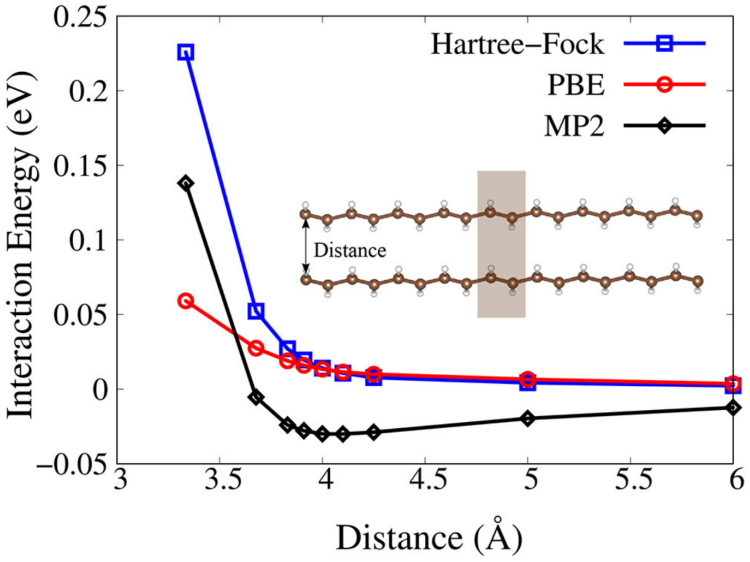


Figure 2.1: Interaction energy as a function of the distance between two trans-polyacetylene chains as predicted by the Hartree-Fock, PBE DFT functional (GGA), and MP2 method. Reproduced from ref [129] under the terms of the CC-BY 4.0 License.

the two-electron operator, this yields the new expression

$$\langle ij|\,\hat{v}_2\,|ab\rangle = \sum_{\alpha=1}^{M'} \langle ij|\chi'_\alpha\rangle\langle\chi'_\alpha|\,\hat{v}_2\,|ab\rangle \tag{2.52}$$

that reduces the scaling with respect to the number of basis functions by an order since it involves the product of three-index integrals only. This can also be viewed as the projection of the (pair) density distribution onto the auxiliary basis set that can be carefully chosen for this purpose, the reason why computational chemists refer to this approach as RI or density fitting (DF).[130–132] The approximation made with the finite basis of the RI is usually compensated when calculating relative energies and permits appreciable speedups.

Laplace transformation and stochastic orbitals

The difference of eigenvalues in the denominator of the MP2 correlation energy is an issue when it comes to using other than canonical orbitals (because those would require a non-canonical formulation of MP2). Almlöf had the idea to circumvent this with a Laplace transform (LT) technique.[133] In this way, the denominator of eq 2.48 can be rewritten as

$$\frac{1}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} = -\int_0^\infty d\tau\; e^{(\varepsilon_i+\varepsilon_j-\varepsilon_a-\varepsilon_b)\tau} \tag{2.53}$$

which, by redefining "time-dependent" orbitals $\phi_i(\tau) = \phi_i \exp(\frac{1}{2}\varepsilon_i\tau)$ and $\phi_a(\tau) =$

$\phi_a \exp(-\frac{1}{2}\varepsilon_a \tau)$, yields a new expression for the MP2 correction

$$E_{\mathrm{MP}}^{(2)} = \int_0^\infty d\tau \sum_{i<j}^{N_{\mathrm{occ}}} \sum_{a<b}^{N_{\mathrm{vir}}} |\langle i(\tau)j(\tau)|\hat{v}_2|a(\tau)b(\tau)\rangle - \langle i(\tau)j(\tau)|\hat{v}_2|b(\tau)a(\tau)\rangle|^2 \tag{2.54}$$

This allows the computation of the MP2 energy with orbitals obtained from appropriate rotations since the new integrand in eq 2.54 is now invariant under unitary transformations. In practice, the integral over τ is achieved through numerical quadrature, that requires a small number of grid points of the order of 10 τ-evaluations. Thanks to this, the scaling can be reduced from $\mathcal{O}(N^5)$ to $\mathcal{O}(N^4)$ at the price of a prefactor with reasonable computational overhead.[121,131]

A further extension of the concept used in the LT-MP2 method is the idea of introducing stochastic orbitals. These are defined as linear combinations of the canonical orbitals with random expansion coefficients that are iteratively optimized with Monte Carlo-like methods. Chemical accuracy with respect to the exact MP2 can be reached after a sufficient number of stochastic samples. Neuhauser et al. developed a stochastic method that is faster than the conventional approach for large systems with close to linear scaling.[134] More recently, developers of the VASP code combined stochastic orbitals with LT-MP2 to get close to cubic scaling for a fixed absolute statistical error, while linear scaling was reached for a fixed relative error per valence orbital.[76] However, if high MP2 accuracy is required, the sample variance prevents the improvement in performance. Indeed, the sampling error decreases as $1/\sqrt{N_s}$ with the number of samples N_s, so that a very large amount of samples might be required to achieve accurate statistical estimates.

Configuration interaction

Configuration interaction (CI) is conceptually the simplest post-HF method. It relies on configurations, that are different occupational distributions of the electrons among molecular orbitals. In CI, the molecular orbitals are taken from a HF calculation and held fixed (except for multi-configurational methods like CASSCF and MRCI). The different configurations are represented by Slater determinants, that when expanded as a linear combination define the CI wavefunction. Thus, let $|\Phi_{\mathrm{HF}}\rangle$ be the HF reference determinant, the CI wavefunction then writes

$$|\Phi_{\mathrm{CI}}\rangle = c_0|\Phi_{\mathrm{HF}}\rangle + \sum_{i,a} c_i^a |\Phi_i^a\rangle + \sum_{i<j,a<b} c_{ij}^{ab} |\Phi_{ij}^{ab}\rangle + \sum_{i<j<k,a<b<c} c_{ijk}^{abc} |\Phi_{ijk}^{abc}\rangle + \ldots \tag{2.55}$$

where the c coefficients are called the CI coefficients that correspond to the various determinants. $|\Phi_i^a\rangle$ represents the configuration obtained from $|\Phi_{\mathrm{HF}}\rangle$ with the excitation of an electron from orbital i to orbital a, $|\Phi_{ij}^{ab}\rangle$ represents double excitations from i to a and j to b, and so on. The dots indicate higher excitations. In practice, the maximum

size of the CI expansion is determined by the number of molecular orbitals resulting from the HF calculation (that in turn depends on the size of the basis used to solve the HF problem). In case the HF problem with N electrons results in M spatial molecular orbitals, a maximum number of $\binom{2M}{N}$ configurations can be generated from $|\Phi_{\text{HF}}\rangle$. The CI method is called full-CI (FCI) if all excited determinants are considered in the CI wavefunction. Truncated CI refers to the truncation of the expansion in practical use, at for example singly (CIS), doubly (CISD), or triply (CISDT) excited configurations.

The goal of CI is to find the coefficients c that minimize the variational energy

$$E_{\text{CI}} = \frac{\langle \Phi_{\text{CI}} | \hat{\mathcal{H}}_e | \Phi_{\text{CI}} \rangle}{\langle \Phi_{\text{CI}} | \Phi_{\text{CI}} \rangle} \tag{2.56}$$

Minimizing E_{CI} in a manner analogous to the HF energy (eq 2.22) leads to the CI Schrödinger equations, that can be reformulated as the matrix problem

$$\sum_J H_{IJ} c_J = E_{\text{CI}} c_J \tag{2.57}$$

where $H_{IJ} = \langle \Phi_I | \hat{\mathcal{H}}_e | \Phi_J \rangle$ are the respective Hamiltonian matrix elements, c_J are the CI coefficients, and E_{CI} is the CI ground state energy. The Hamiltonian matrix takes a specific form, with many elements being zero due to the Brillouin's theorem and the Slater-Condon rules.[13] However, the Hamiltonian matrix in the CI basis becomes very large due to the factorial increase in the number of possible Slater determinants. This applies to all but the smallest systems, truncation orders or basis set sizes, that makes CI calculations very time-consuming in practice.

Coupled-cluster method

As an alternative post-HF approach, the coupled-cluster (CC) method provides a systematic and rigorous way to include electron correlation effects of many-body systems. In the CC method, the electronic wavefunction is expressed as an exponential ansatz that again builts upon the HF wavefunction:

$$|\Phi_{\text{CC}}\rangle = e^{\hat{T}} |\Phi_{\text{HF}}\rangle \tag{2.58}$$

where $|\Phi_{\text{HF}}\rangle$ represents the reference determinant, and $\hat{T}$ is the excitation operator. The exponent of the operator $\hat{T}$ is defined by the Taylor series

$$e^{\hat{T}} = \sum_{n=0}^{\infty} \frac{1}{n!} \left(\hat{T}\right)^n \tag{2.59}$$

and the total excitation operator can be further divided into

$$\hat{T} = \sum_{n=1}^{\infty} \hat{T}_n \tag{2.60}$$

where each $\hat{T}_n$ represents the n-th order of electron excitations. In practical calculations, truncation is made at a certain excitation level to approximate the full exponential ansatz. For example, at first order, the operator

$$\hat{T}_1 = \sum_{i,a} t_i^a \hat{T}_i^a \tag{2.61}$$

represents all single excitations with $\hat{T}_i^a$ acting as the electron-wise excitation operator, $\hat{T}_i^a |\Phi_{\text{HF}}\rangle = |\Phi_i^a\rangle$. Similarly, for double excitations, one has

$$\hat{T}_2 = \sum_{i<j,a<b} t_{ij}^{ab} \hat{T}_{ij}^{ab} \tag{2.62}$$

with $\hat{T}_{ij}^{ab} |\Phi_{\text{HF}}\rangle = |\Phi_{ij}^{ab}\rangle$. In CC, the truncation only refers to the total excitation operator $\hat{T}$ (eq 2.60) but not to the exponential expansion in eq 2.59. The amplitudes t_i^a, t_{ij}^{ab},... appear as the CC coefficients to optimize with respect to the energy. The Schrödinger equation can be rewritten using the CC ansatz:

$$\hat{\mathcal{H}}_e e^{\hat{T}} |\Phi_{\text{HF}}\rangle = E_{\text{CC}} e^{\hat{T}} |\Phi_{\text{HF}}\rangle \tag{2.63}$$

where $\hat{\mathcal{H}}_e$ is the electronic Hamiltonian operator and E_{CC} is the energy eigenvalue. The ground state energy can be obtained by projecting the Schrödinger equation onto the left with the bra $\langle\Phi_{\text{HF}}|$:

$$\langle\Phi_{\text{HF}}|\hat{\mathcal{H}}_e e^{\hat{T}}|\Phi\rangle = E_{\text{CC}} \langle\Phi_{\text{HF}}|e^{\hat{T}}|\Phi_{\text{HF}}\rangle \tag{2.64}$$

Since $e^{\hat{T}}|\Phi_{\text{HF}}\rangle = (1 + \hat{T} + \frac{1}{2}\hat{T}^2 + \dots)|\Phi_{\text{HF}}\rangle = |\Phi_{\text{HF}}\rangle + \hat{T}|\Phi_{\text{HF}}\rangle + \frac{1}{2}\hat{T}^2|\Phi_{\text{HF}}\rangle + \dots$, all terms except for the first one cancel out due to their orthogonality with the Φ_{HF} reference. Consequently, the E_{CC} variational energy reads

$$E_{\text{CC}} = \langle\Phi_{\text{HF}}|\hat{\mathcal{H}}_e e^{\hat{T}}|\Phi\rangle \tag{2.65}$$

The CC method therefore provides a systematic description of electron correlation effects by expanding the electronic wavefunction as an exponential ansatz and solving the resulting Schrödinger equation. In principle, once the ansatz is decided, the ground state energy can be evaluated by optimizing the amplitudes t in a non-variational way by resorting to iterative or projection methods. The CCSD(T) method that includes all single and double excitations and triples in a perturbative manner is generally considered as one of the gold standard methods for reaching chemical accuracy in electronic structure theory. However, its cost makes it impractical for systems containing more

than a few tens, maximum a hundred, electrons.

2.3 Density functional theory

Over the last 60 years, density functional theory (DFT) has become a very powerful and widely used approach in first-principles calculations, be it for properties of materials or molecules,[15,16,23,28,30] due to the ability to provide a good trade-off between computational efficiency and accuracy. Thanks to DFT, the theoretical simulations of systems with hundreds or even thousands of atoms have been made possible. At the same time, DFT is in general fairly accurate (not as good as CI or CC) for e.g. bond lengths (1-2%) and energies (several kcal/mol). It performs also well with pronounced electron correlation effects such as in transition metals. Historically, DFT has been developed in parallel to the ongoing efforts on wavefunction-based methods, with the idea that the fundamental quantity to treat the many-body problem is instead the electron density

$$\rho(\mathbf{r}) = \mathcal{N} \int \Phi_e^*(\mathbf{r}, \mathbf{r}_2, \mathbf{r}_3, ..., \mathbf{r}_N)\Phi_e(\mathbf{r}, \mathbf{r}_2, \mathbf{r}_3, ..., \mathbf{r}_N)\ d\mathbf{r}_2 ... d\mathbf{r}_N \tag{2.66}$$

where $\mathcal{N}$ is a normalization constant that ensures the correct number of electrons $N = \int \rho(\mathbf{r}) d\mathbf{r}$ in the system. Major computational advantages of DFT reside in the fact that it deals with the 3-dimensional real density rather than the $3N$-dimensional many-electron wavefunction.

In 1964, Hohenberg and Kohn established two fundamental theorems concerning the ground state of an interacting electron gas in an external potential $v_{ext}(\mathbf{r})$.[135] The first theorem states that the total electron density $\rho(\mathbf{r})$, up to a constant, uniquely defines the external potential

$$\rho(\mathbf{r}) \rightarrow v_{ext}(\mathbf{r}) \tag{2.67}$$

Consequently, as the external potential in turn completely determines the Hamiltonian, this implies that also the many-electron ground state wavefunction is a unique functional of the electron density. Therefore, there exists a unique functional $F[\rho]$ such that the energy E of the electron system, subject to the potential $v_{ext}(\mathbf{r})$, is given by

$$E[\rho] = \langle \Phi_e[\rho] | \hat{T}_e + \hat{V}_{ee} | \Phi_e[\rho] \rangle + \int \rho(\mathbf{r}) v_{ext}(\mathbf{r}) d\mathbf{r} = F[\rho] + \int \rho(\mathbf{r}) v_{ext}(\mathbf{r}) d\mathbf{r} \tag{2.68}$$

where $\Phi_e[\rho]$ denotes the ground state wavefunction of the N-electron interacting system with the external potential determined by $\rho(\mathbf{r})$.

The second theorem establishes the variational principle for the energy functional. It states that any trial density ρ_{trial} that is not the true ground state density ρ_0 will lead to an energy expectation value $E[\rho_{\text{trial}}]$ that is greater than the true ground state energy

$E[\rho_0]$. That is, the density minimizing the total energy corresponds to the ground state density

$$E_0 = \min_{\rho \to N}(E[\rho]) = E[\rho_0] \leq E[\rho_{\text{trial}}] \tag{2.69}$$

The Hohenberg-Kohn theorems therefore state the existence of a functional

$$F[\rho] = \langle \Phi_e[\rho] | \hat{T}_e + \hat{V}_{ee} | \Phi_e[\rho] \rangle = T_e[\rho] + V_{ee}[\rho] \tag{2.70}$$

that can be separated into kinetic $T_e[\rho]$ and potential $V_{ee}[\rho]$ contributions. Since $F[\rho]$ is independent of the external potential, it is universal over all electron systems. Having at hand the expression for the variational energy in eq 2.68, the DFT problem can be solved "à la HF" following the Euler-Lagrange equations to minimize the energy, leading to

$$\frac{\delta F[\rho]}{\delta \rho(\mathbf{r})} + v_{ext}(\mathbf{r}) = \mu \tag{2.71}$$

where μ appears from the Lagrange multipliers constraining the number of electrons, and is identified as the chemical potential of the interacting system.[28] As a consequence, eq 2.71 demonstrates that the external potential $v_{ext}(\mathbf{r})$ and the number of electrons N are the unique ingredients needed to completely define the quantum problem in exact DFT. Unfortunately, the curse of exact DFT lies in the fact that the universal functional $F[\rho]$ is unknown and no reliable scheme has so far provided conclusive and transferable accuracy.

Earliest efforts in the 1920s, even before the formulation of the Hohenberg-Kohn theorems, aimed to develop an expression for the energy as a functional of the electron density, which included contributions from kinetic, external, and electron-electron interactions. Among these terms, the most challenging aspect is the determination of the kinetic energy contribution, that can be seen as obtaining the second derivative of the wavefunction from the charge density. Consider a plane wave solution, such as that of a free electron, where the density is uniform in space ($\rho = 1/V$, with V representing the volume). In this case, the kinetic energy is $\frac{1}{2}\mathbf{k}^2$. This example highlights the difficulty in establishing a direct relationship between the density and kinetic energy, as the constant density ρ appears to have lost any information regarding the kinetic energy. Despite the fact that electronic wavefunctions can possess both long and short wavelengths, such variations do not manifest in the density alone. Thomas and Fermi proposed an approximation to address this challenge by assigning the kinetic energy density at each point in space to be equivalent to the kinetic energy density of a non-interacting uniform electron gas (UEG).[136,137] This approximation was based on the assumption that the electron density varies slowly in space, and marked the first inception of a local density approximation. Generalizing to inhomogeneous systems, and treating the electron-electron interactions classically, the Thomas-Fermi model

results in

$$E_{\mathrm{TF}}[\rho] = \int \tau^{\mathrm{UEG}}(\rho(\mathbf{r}))d\mathbf{r} + \frac{1}{2}\iint \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' + \int \rho(\mathbf{r})v_{ext}(\mathbf{r})d\mathbf{r} \tag{2.72}$$

where

$$\tau^{\mathrm{UEG}}(\rho(\mathbf{r})) = \frac{3}{10}(3\pi^2)^{2/3}\rho(\mathbf{r})^{5/3} \tag{2.73}$$

is the kinetic energy density of the UEG. Few years later, Dirac proposed to add some consideration of electron exchange effects into eq 2.72, derived again for the UEG, that led to the Thomas-Fermi-Dirac (TFD) model[138]

$$E_{\mathrm{TFD}}[\rho] = E_{\mathrm{TF}}[\rho] - \frac{3}{4}(3/\pi)^{1/3}\int \rho(\mathbf{r})^{4/3}d\mathbf{r} \tag{2.74}$$

While providing first functional expressions for the energy, the Thomas-Fermi model and its extensions are not accurate enough to describe the electronic structure of inhomogeneous systems. Particularly, the Thomas-Fermi model is incapable of describing chemical bonding. As a bridge towards the construction of more accurate functionals to approximate the exact $F[\rho]$, the Kohn-Sham formalism of DFT has emerged as the mainstream approach. This formalism employs orbitals to represent the electron density. It is worth mentioning that research efforts are still ongoing in the development of density functionals that remain orbital-free within the framework of orbital-free DFT.[139] However, the discussion of orbital-free DFT approaches is beyond the scope of this thesis.

2.3.1 Kohn-Sham density functional theory

While the Hohenberg-Kohn theorems established the existence of a density functional, they did not provide its universal expression, nor a practical way to handle it. In 1965, Kohn and Sham (KS) addressed this issue by introducing a fictitious system of N non-interacting electrons (described by KS orbitals) that reproduces the exact ground state density $\rho(\mathbf{r})$ of the physical interacting system.[140] Thanks to this fictitious system, the KS energy ansatz takes the following form

$$E_{\mathrm{KS}}[\rho] = T_S[\rho] + \frac{1}{2}\iint \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' + \int \rho(\mathbf{r})v_{ext}(\mathbf{r})d\mathbf{r} + E_{xc}[\rho] \tag{2.75}$$

with $T_S[\rho]$ being the kinetic energy of the KS system. For such a non-interacting system, the KS many-electron wavefunction can be expressed as a Slater determinant Φ_S composed of molecular orbitals $\phi_i(\mathbf{r})$ such that

$$T_S[\rho] = \langle\Phi_S|\sum_{i=1}^{N} -\frac{1}{2}\nabla_i^2|\Phi_S\rangle = \sum_{i=1}^{N}\int \phi_i^*(\mathbf{r})\left(-\frac{1}{2}\nabla_i^2\right)\phi_i(\mathbf{r})d\mathbf{r} \tag{2.76}$$

where the functional dependence in ρ is realized by the link between the KS orbitals and the density

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2 \tag{2.77}$$

The last term in eq 2.75 is the *exchange-correlation* functional and accounts for the energy difference between the exact functional $F[\rho]$ and the new constituents of $E_{\text{KS}}[\rho]$:

$$E_{xc}[\rho] := F[\rho] - T_S[\rho] - \frac{1}{2}\iint \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}' = T_e[\rho] - T_S[\rho] + V_{ee}[\rho] - \frac{1}{2}\iint \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}' \tag{2.78}$$

Therefore, $E_{xc}[\rho]$ includes both the kinetic energy difference between the interacting and fictitious systems, and the difference between the exact electron-electron interaction and their classical Coulomb counterpart. By analogy with the HF problem, the exchange-correlation energy is perceived as covering two quantum components of the interacting system

$$E_{xc}[\rho] = E_x[\rho] + E_c[\rho] \tag{2.79}$$

that are respectively the exchange (x) and correlation (c) effects. In wavefunction-based theories, the correlation energy is defined as the difference between the exact and the HF energy (Section 2.2.2). In KS-DFT, in contrast, the correlation energy E_c is given by the difference between the total energy E_{KS} and the sum of kinetic, direct and exchange Coulomb terms. Since the exchange operator is usually local in KS-DFT, and the exchange-correlation functional also accounts for the fictitious system, the concept of correlation energy in KS-DFT differs from the one in wavefunction-based methods.

Having set the new KS functional expression (eq 2.75), the variational principle can be applied in order to determine the orbitals, thus the density, that minimize the energy. The Euler-Lagrange minimization of E_{KS} with respect to the KS orbitals $\phi_i(\mathbf{r})$ provides the KS equations to be ultimately solved:

$$\left\{-\frac{1}{2}\nabla_i^2 + v_{ext}(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' + \frac{\delta E_{xc}[\rho]}{\delta\rho(\mathbf{r})}\right\}\phi_i(\mathbf{r}) = \varepsilon_i\phi_i(\mathbf{r}) \tag{2.80}$$

The KS formalism of DFT thus maps the electronic Schrödinger equation for the ground state energy to a set of N single-electron Schrödinger equations that is exact in principle. Having said that, however, the main challenge in KS-DFT calculations finally resides in the consideration of the exchange-correlation energy $E_{xc}[\rho]$ that appears in eq 2.80, for which exact formulations remain unknown. The next section discusses the approximations available to treat $E_{xc}[\rho]$ in practice.

2.3.2 Density functional approximations

The development of density functional approximations (DFAs) has been at the forefront of research over the last decades, highlighted by the proposal of more than several hundreds of DFAs since the 1990s.[141] Such a large number of DFAs not only demonstrates the challenges in establishing the correct universal exchange-correlation functional but also the system and property-dependent character of suitable approximations. As a categorization of the plethora of DFAs, Perdew proposed the concept of Jacob's ladder illustrated in Figure 2.2, which defines a hierarchy among functionals in terms of the complexity of their ingredients.[142] As a very general formulation, the exchange-correlation functional can be written

$$E_{xc}[\rho = \rho_\alpha + \rho_\beta] = \int \rho(\mathbf{r})\epsilon_{xc}[\rho_\alpha, \rho_\beta](\mathbf{r})d\mathbf{r} \tag{2.81}$$

where $\epsilon_{xc}(\mathbf{r})$ denotes the exchange-correlation energy density at position $\mathbf{r}$, that is a functional of the spin-separated densities $\rho_{\sigma=\alpha,\beta}$, and therefore encodes all kinds of mathematical forms appearing in DFAs.

"On the ground level", i.e. still below the first rung of the ladder, the exchange-correlation energy is entirely absent in the Kohn-Sham equations (eq 2.80) which is equivalent to the Hartree problem, i.e. HF (eq 2.28) without exchange. Then, when climbing up Jacob's ladder, the functional form of $\epsilon_{xc}(\mathbf{r})$ gradually becomes more complex. Up to the third rung (meta-GGA), DFAs are often referred to as semilocal because their $\epsilon_{xc}(\mathbf{r})$ are functions (not functionals) of local quantities such as the density ρ, the density gradient $\nabla\rho$, or the KS kinetic energy density τ. Semilocal functionals are computionally very efficient and therefore very popular in electronic structure calculations of large

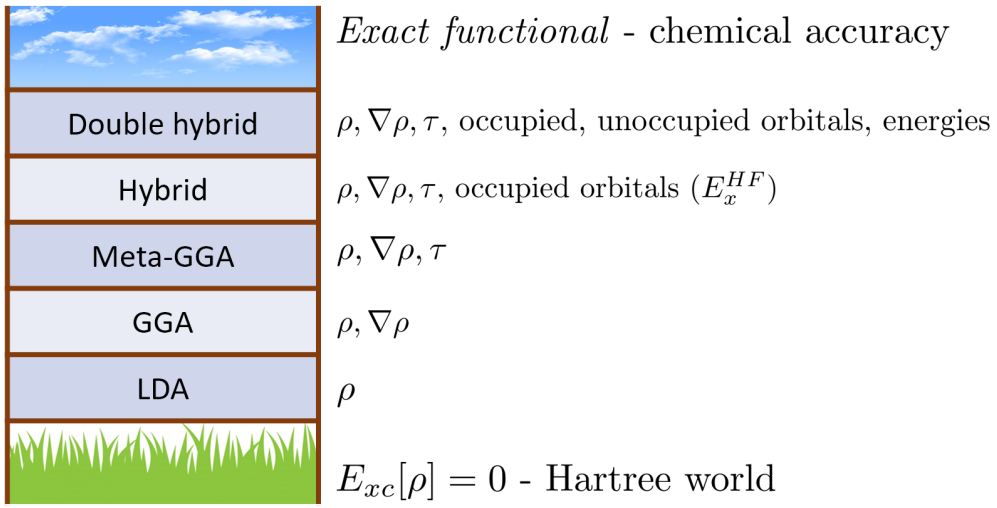


Figure 2.2: Illustration of Perdew's Jacob's ladder that categorizes density functional approximations.

molecules, clusters, liquids, and solids. However, their accuracy remains elusive when complex electron interactions drive the quantum phenomena in the systems under investiation.

A class of functionals that could potentially lead to higher accuracy includes a fraction of exact non-local exchange in E_{xc}. Such approximations are called hybrid functionals and have become standard approach in quantum-chemistry applications. With these, the realm of the original KS framework (that assumes the existence of a local exchange-correlation potential) is left, and hybrid functionals are rather formulated within a generalized KS approach. In the context of molecular systems, these functionals demonstrate enhanced accuracy in capturing not just atomization energies, but also ionization potentials and electron affinities. However, certain limitations persist, such as the occasional failure to achieve the desired level of chemical accuracy and the absence of a general solution for accounting for all van der Waals interactions. Nonetheless, the incorporation of exact exchange unequivocally represents an improvement, holding significant potential for addressing these challenges and proving beneficial in various scenarios. When employed in semiconductors and insulators, hybrid functionals offer notable advantages over semilocal functionals. They yield improved descriptions of structural parameters, resulting in values that align closer to experimental observations. Additionally, hybrid functionals consistently provide electronic band gaps that are systematically larger than those obtained using semilocal functionals, leading to enhanced agreement with experimental data.[143] However, incorporating exact exchange into calculations incurs a significantly higher computational cost that exceeds that of semilocal functionals by up to an order of magnitude with atom-centered basis sets,[144] and up to two orders of magnitude with plane waves.[45,145] Nevertheless, there is growing interest in such calculations as computational resources continue to increase.

At the last rung of Jacob's ladder are so called rung-5 functionals which explicitly take into account the virtual orbitals, contrary to hybrids which use only the occupied orbitals. Rung-5 functionals combine elements of hybrid functionals, which include a fraction of exact exchange, and correlation functionals coming either from post-HF considerations or the random phase approximation (RPA), with the aim to improve the accuracy of electronic structure calculations. These have been shown to provide improved accuracy for a wide range of molecular properties, including thermochemistry, reaction energies, and noncovalent interactions. However, due to their increased complexity, they are computationally even more demanding than the hybrid functionals.

In 2011, Goerigk and Grimme conducted a thorough energy benchmark study of various density functionals.[146] They used an extensive test set that contains data for assessing performance on general main group thermochemistry, kinetics and noncovalent interactions. As a quantitative confirmation of the concept of Perdew,

their results, reproduced in Figure 2.3, demonstrated that the average accuracy of functionals is indeed increasing when climbing up Jacob's ladder, with double-hybrid functionals being among the best performing functionals to date.

Next, I will delve into more details about the various rungs of Jacob's ladder. However, it is important to reiterate the ongoing challenges associated with the usage and develop-

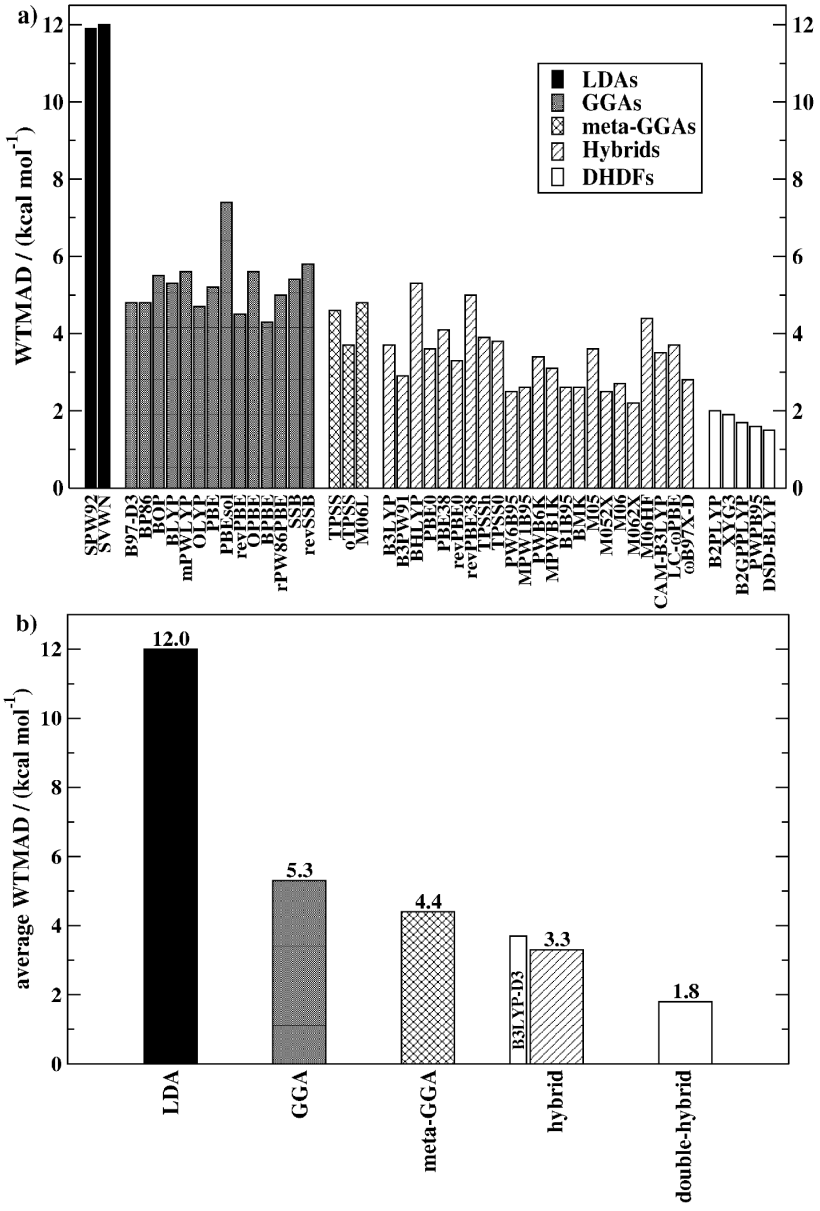


Figure 2.3: Quantitative illustration of the improvement in accuracy when climbing up Jacob's ladder. a) Weighted total mean absolute deviation (WTMAD) achieved against test data, shown per functional. b) Average WTMAD achieved per rung of the Jacob's ladder. All functionals were combined with the DFT-D3 correction.[147] Used with permission of Royal Society of Chemistry, from ref [146]; permission conveyed through Copyright Clearance Center, Inc.

ment of DFAs. First, one of the challenges in KS-DFT is the vast number of functionals available, each with its own set of approximations and (sometimes empirical) parameters. This makes it difficult to choose the most appropriate functional for a given system. Additionally, there is a trade-off between performance and generality, as functionals that perform particularly well for one type of system may not generalize well to others. Ideally, each level of approximation on Jacob's ladder should provide a general accuracy that consistently outperforms lower rungs. However, there is currently no systematic way to derive such approximations or universal rules to assess their reliability. In fact, from simple considerations up to highly sophisticated mathematical formulations, formal properties of the exact E_{xc} are known such as e.g., coordinate scaling relations, exact conditions for one-electron systems to remove the self-interaction error, piecewise linearity, spin scaling relation, and lower bounds.[20,30,148] In this context, a promising strategy to improve the reliability and accuracy of DFAs therefore relies on satisfying exact constraints and norms when developing new exchange-correlation functionals.[149]

Local density approximation

In a generalization of the Thomas-Fermi model, the local density approximation (LDA) relies on the assumption that contributions to the total energy can be divided into local volume elements.[140] Supposing that the electron density is slowly varying in space, this latter can be seen as uniform over infinitesimal volume elements. The energy is thus obtained by integrating over local contributions for which the energy density is provided by that of the UEG. In LDA, therefore, the approximate exchange-correlation energy takes the form

$$E_{xc}^{\mathrm{LDA}}[\rho] = \int \rho(\mathbf{r})\epsilon_x^{\mathrm{UEG}}(\rho(\mathbf{r}))d\mathbf{r} + \int \rho(\mathbf{r})\epsilon_c^{\mathrm{UEG}}(\rho(\mathbf{r}))d\mathbf{r} \tag{2.82}$$

where the exchange energy density $\epsilon_x^{\mathrm{UEG}}(\rho(\mathbf{r})) = -\frac{3}{4}(3/\pi)^{1/3}\rho(\mathbf{r})^{1/3}$ is known analytically (eq 2.74)[138] and the correlation energy density is typically approximated by parametrized fits to highly accurate quantum Monte Carlo (QMC) results for the UEG at different densities.[150] Several fittings of QMC data were proposed, leading to different LDA approximations. Among these, the analytical form of Vosko, Wilk and Nusair (VWN),[151] and the one of Perdew and Wang (PW92)[152] are the most widely used. The LDA only depends on the value of the density at each point, making it the simplest and least computationally expensive approximation. Being in general a crude approximation for real systems, it appears to be surprisingly successful for solids, which can partly be attributed to the fact that $E_{xc}^{\mathrm{LDA}}[\rho]$ satisfies several formal properties of the exact exchange-correlation functional. In the case of dimers consisting of closed-shell atoms and molecules, such as rare gases, the LDA yields an attractive interaction that resembles the dispersion interaction but this apparent attraction is actually an artifact arising from the approximation made for the exchange term.[153] In general, the

LDA often fails to accurately describe systems where the electron density is highly inhomogeneous.

Generalized gradient approximations

The generalized gradient approximations (GGAs) appear on the second rung of Jacob's ladder and take the formal form

$$E_{xc}^{\mathrm{GGA}}[\rho] = \int \rho(\mathbf{r})\epsilon_{xc}^{\mathrm{GGA}}(\rho_\alpha(\mathbf{r}), \rho_\beta(\mathbf{r}), \nabla\rho_\alpha(\mathbf{r}), \nabla\rho_\beta(\mathbf{r}))d\mathbf{r} \tag{2.83}$$

that makes $E_{xc}^{\mathrm{GGA}}[\rho]$ depend locally on the density and its gradient.[154,155] As a working quantity, the magnitude of the gradient can be included in form of the dimensionless reduced gradient defined as

$$x_\sigma(\mathbf{r}) = \frac{|\nabla\rho_\sigma(\mathbf{r})|}{\rho_\sigma(\mathbf{r})^{4/3}} \tag{2.84}$$

or, further rescaled for convenience as

$$s_\sigma(\mathbf{r}) = \frac{x_\sigma(\mathbf{r})}{2(3\pi^2)^{1/3}} \tag{2.85}$$

such that $s_\sigma(\mathbf{r})$ is the only gradient-based quantity remaining in the functional:

$$E_{xc}^{\mathrm{GGA}}[\rho] = \int f_x(\rho_\alpha(\mathbf{r}), \rho_\beta(\mathbf{r}), s_\alpha(\mathbf{r}), s_\beta(\mathbf{r}))d\mathbf{r} + \int f_c(\rho_\alpha(\mathbf{r}), \rho_\beta(\mathbf{r}), s_\alpha(\mathbf{r}), s_\beta(\mathbf{r}))d\mathbf{r} \tag{2.86}$$

where f_x and f_c describe the exchange and correlation contributions. Due to the exact condition that the exchange part must obey under scaling of the density (the UEG limit), this latter can be re-expressed as

$$E_x^{\mathrm{GGA}}[\rho] = \int \rho(\mathbf{r})\epsilon_x^{\mathrm{UEG}}(\rho_\alpha(\mathbf{r}), \rho_\beta(\mathbf{r}))F_x(s_\alpha(\mathbf{r}), s_\beta(\mathbf{r}))d\mathbf{r} \tag{2.87}$$

where F_x is the so-called exchange enhancement factor. From this general form, multiple GGA approximations have been designed, differing in the enhancement factor F_x and the correlation approximation $f_c(\rho, s)$.

For example, the BLYP GGA functional was built by merging the B88 exchange functional that has the form[155]

$$F_x^{\mathrm{B88}}(s_\alpha(\mathbf{r}), s_\beta(\mathbf{r})) = 1 - \frac{\beta}{\epsilon_x^{\mathrm{UEG}}(\rho(\mathbf{r}))\rho(\mathbf{r})} \sum_{\sigma=\alpha,\beta} (\rho_\sigma)^{4/3} \frac{(2(3\pi^2)^{1/3})^2 s_\sigma^2}{1 + 6\beta(2(3\pi^2)^{1/3})s_\sigma \sinh^{-1}((2(3\pi^2)^{1/3})s_\sigma)} \tag{2.88}$$

with the LYP analytical correlation functional based on a correlated wavefunction expression for the helium atom, developed earlier by Colle and Salvetti in 1975.[156] GGAs rely on empirical parameters, like for instance the parameter $\beta = 0.0042$ a.u. in

eq 2.88 that was determined by fitting the exchange energy of six rare gas atoms.

The advancement of reliable GGA functionals marked a significant milestone in the progress of KS-DFT. With GGA, reasonable results could be achieved for various molecular systems, and structural predictions based on GGA are typically highly accurate. However, the performance of GGA functionals for energy-related quanities, such as thermochemical properties, can be unreliable or even subpar, although they do show significant improvements compared to LDA. Predictions of reaction enthalpies are often unreliable, and barrier heights are consistently underestimated, sometimes to a significant degree. Nevertheless, there are cases where GGA predictions demonstrate unexpected accuracy, possibly benefiting from error compensation.

Meta-GGA

A rather straightforward extension of the GGA formalism consists of including knowledge about the second density derivatives into the functional, leading to the development of meta-GGAs. Those functionals at the third rung of Jacob's ladder include the second derivative information via the kinetic energy density

$$\tau(\mathbf{r}) = \frac{1}{2}\sum_{i=1}^{N} \phi_i^*(\mathbf{r})\nabla^2\phi_i(\mathbf{r}) \tag{2.89}$$

and thus take the general form (with implicit consideration of spins)

$$E_{xc}^{\mathrm{mGGA}}[\rho] = \int \rho(\mathbf{r})\epsilon_{xc}^{\mathrm{mGGA}}(\rho(\mathbf{r}), \nabla\rho(\mathbf{r}), \tau(\mathbf{r}))d\mathbf{r} \tag{2.90}$$

Thanks to the presence of the higher-order derivatives, meta-GGA possibly lead to better consideration of the chemical environment. In fact, the kinetic energy density carries more chemical information that makes meta-GGA approximations able to satisfy more DFT exact constraints than GGAs.[157] Over the past two decades, numerous meta-GGAs have been proposed using both non-empirical and empirical approaches. Some of the most popular are those of the Minnesota family,[101] that despite being largely empirical provide impressive "accross-the-board" accuracy (see Section 7.3.1 for a description of Minnesota density functionals). In the opposite trend, the SCAN functional was developed in the spirit of constraining most formal properties of the exact exchange-correlation functional.[149] SCAN satisfies the 17 constraints that are applicable to meta-GGAs and its remaining parameters were fitted against exact reference data of some prototypical systems. SCAN was found to outperform most of the GGA functionals, especially the widely-used PBE, which is quite an outstanding performance for a functional built entirely on non-empirical criteria.

Hybrid functionals

The DFAs discussed so far produce a fully local, orbital-independent exchange-correlation potential, aligning with the original proposal by Kohn and Sham. However, Becke proposed a significant paradigm shift in 1993,[158] introducing an empirical approach based on the adiabatic connection theorem.[159,160] The adiabatic connection theorem establishes a connection between the exact exchange-correlation functional and the electron-electron interaction. It introduces a parameter λ that represents the coupling strength between an exactly solvable reference system and the actual interacting system. By varying λ, one can explore different levels of exchange and correlation effects, ranging from pure Hartree-Fock exchange (λ = 0) to the fully interacting system (λ = 1). The adiabatic connection theorem provides a theoretical foundation for developing DFAs that gradually incorporate the correct exchange-correlation effects as λ varies. As a tuning of the adiabatic connection towards most accurate models, Becke suggested incorporating a fraction of orbital-dependent, non-local exchange from Hartree-Fock theory (eq 2.21) into KS-DFT. This resulted in the development of generalized KS-DFT with hybrid functionals of the form

$$E_{xc}^{\text{hybrid}}[\rho] = \lambda E_x^{\text{HF}}[\rho] + (1-\lambda)E_x^{\text{(m)GGA}}[\rho] + E_c^{\text{(m)GGA}}[\rho] \tag{2.91}$$

with the exact exchange given by

$$E_x^{\text{HF}}[\rho] = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\iint d\mathbf{x}_1 d\mathbf{x}_2 \, \frac{\phi_i^*(\mathbf{x}_1)\phi_j^*(\mathbf{x}_2)\phi_j(\mathbf{x}_1)\phi_i(\mathbf{x}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{2.92}$$

where $\mathbf{x} = (\mathbf{r}, \sigma)$ accounts now for KS spin orbitals.

An illustration of a popular hybrid DFA is the PBE0 functional that takes the form[161]

$$E_{xc}^{\text{PBE0}}[\rho] = \frac{1}{4}E_x^{\text{HF}} + \frac{3}{4}E_x^{\text{PBE}} + E_c^{\text{PBE}} \tag{2.93}$$

and mixes a portion of exact exchange with the PBE GGA functional. Finally, the B3LYP functional, widely-used for molecular systems, is based on a somewhat more complex mixing scheme, namely[162]

$$E_{xc}^{\text{B3LYP}}[\rho] = a_0 E_x^{\text{HF}} + a_x E_x^{\text{B88}} + (1-a_0-a_x)E_x^{\text{LDA}} + a_c E_c^{\text{LYP}} + (1-a_c)E_c^{\text{LDA}} \tag{2.94}$$

where the numerical coefficients were optimized to be $a_0 = 0.20$, $a_x = 0.72$ and $a_c = 0.81$.

Range separation

For molecular systems, the erroneous long-range decay of the LDA and GGA approximations has been identified as a common cause for failures in many cases. The

Coulomb operator indeed imposes an asymptotic $1/r$ behavior at large distance that is not retrieved in semilocal functionals. Notably, the exact long-range decay is beneficial for describing valence, Rydberg, and charge-transfer excitations in a balanced and accurate way. As a remedy to the wrong asymptotic behavior, it was proposed to split the Coulomb interaction between long (lr) and short range (sr), with the long-range part being treated in an HF manner, akin hybrids. The most common approach to separate between ranges consists of splitting the Coulomb operator with the error function[163]

$$\hat{v}_{12} = \frac{1}{r_{12}} := \hat{v}_{12}^{\mathrm{lr},\mu} + \hat{v}_{12}^{\mathrm{sr},\mu} = \frac{\mathrm{erf}(\mu r_{12})}{r_{12}} + \frac{\mathrm{erfc}(\mu r_{12})}{r_{12}} \tag{2.95}$$

where $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$ and μ is introduced as the range-separation parameter. This separation enables the adjustment of the percentage of HF exchange as a function of interelectronic separation. This flexibility allows for the inclusion of 100% exchange at large interelectronic distances r_{12}, resulting in accurate descriptions of long-range charge-transfer excitations. At the same time, smaller values of exact exchange can be incorporated at small and intermediate interelectronic separations, leading to improved performance for valence and Rydberg excitations. The long-range corrected (LC) form of a hybrid functional can be therefore written as

$$E_{xc}^{\mathrm{LC}}[\rho] = E_x^{\mathrm{HF,lr}} - E_x^{\mathrm{SL,lr}} + E_{xc}^{\mathrm{SL}} = E_x^{\mathrm{HF,lr}} + E_x^{\mathrm{SL,sr}} + E_c^{\mathrm{SL}} \tag{2.96}$$

where SL denotes a semilocal functional such as a GGA or meta-GGA functional form, evaluated with the corresponding long-range (short-range) Coulomb interaction. Resorting to range-separated hybrids usually improves the results for ground-state anions and excited states in the time-dependent DFT formulation.[164]

Fifth-rung functionals

Even at the hybrid level, chemical accuracy (1 kcal/mol error) is usually not reached. As a further improvement, the highest rung of Jacob's ladder contains more sophisticated functionals that depend not only on the occupied but also on the unoccupied (virtual) KS orbitals. Two main fifth-rung, or post-HF-based approaches exist, that rely either on perturbation theory[21,56,165–170] or on the adiabatic-connection-fluctuation-dissipation theorem (ACDFT).[130,171–174]

Double-hybrid functionals

The double-hybrid functionals extend the hybrids by further mixing the semilocal correlation functional with the post-HF treatment of the KS reference system. These latter thus consider the unoccupied KS orbitals to accurately describe the wavefunction correlation, in addition to occupied orbitals used for exact exchange.[57] In the approach of Görling-Levy (GL), correlation is obtained from second-order Rayleigh-Schrödinger

perturbation theory applied to the zeroth-order KS non-interacting system[58,59] that is

$$E_c^{\mathrm{GL2}} = \sum_{i<j}^{N_{\mathrm{occ}}} \sum_{a<b}^{N_{\mathrm{vir}}} \frac{|\langle ij|\,\hat{v}_2\,|ab\rangle_{\mathrm{KS}} - \langle ij|\,\hat{v}_2\,|ba\rangle_{\mathrm{KS}}|^2}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} + \sum_{i=1}^{N_{\mathrm{occ}}} \sum_{a=1}^{N_{\mathrm{vir}}} \frac{|\langle i|\,\hat{v}_x^{\mathrm{HF}} - v_x^{\mathrm{KS}}\,|a\rangle_{\mathrm{KS}}|^2}{\varepsilon_i - \varepsilon_a} \quad (2.97)$$

where the first term corresponds to the evaluation of the MP2 energy (eq 2.48) with the KS orbitals ϕ_i and eigenvalues ε_i.[21] The second term reflects the second-order contribution to the correlation energy due to the difference between the non-local HF exchange potential $\hat{v}_x^{\mathrm{HF}}$ and the local, thus multiplicative, v_x^{KS} Kohn-Sham exchange potential. As such, this second term is much smaller and is most of the time neglected. In a similar way to the introduction of the exact exchange in hybrid functionals, the MP2 correlation energy can be included in the correlation functional, leading to the following functional form for double hybrids

$$E_{xc}^{\mathrm{DH}} = c_x E_x^{\mathrm{HF}} + (1 - c_x) E_x^{\mathrm{SL}} + c_c E_c^{\mathrm{MP2}} + (1 - c_c) E_c^{\mathrm{SL}} \quad (2.98)$$

where SL denotes a semilocal functional such as a GGA or meta-GGA. The coefficients c_x and c_c are obtained from fitting to training datasets. For instance, the B2PLYP functional by Grimme has $c_x = 0.53$, $c_c = 0.27$ with the B88 exchange functional and LYP correlation functional.[56] Since the appearance of B2PLYP, over 75 double-hybrid functionals have been developed with the aim to further improve accuracy.[57] In the spirit of getting the best out of both worlds wavefunction-based methods and DFT,

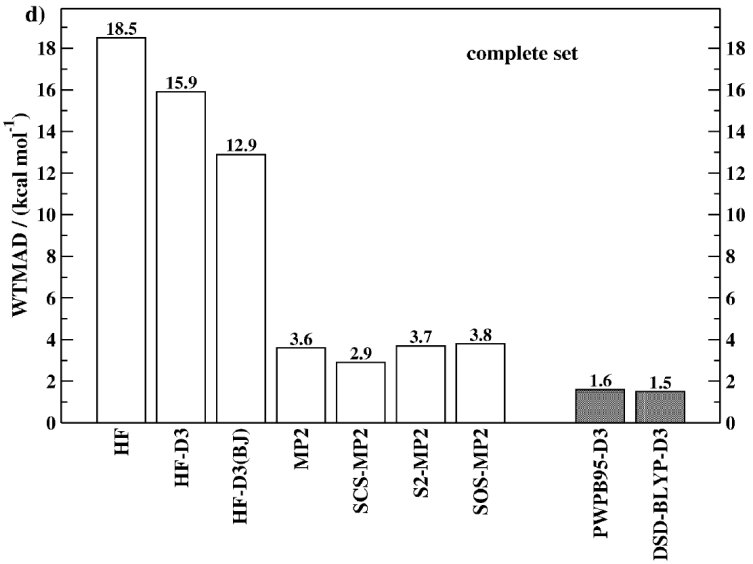


Figure 2.4: Quantitative illustration of the improvement in accuracy of double-hybrid functionals (PWPB95-D3[175] and DSD-BLYP-D3[176]) over HF and MP2-based approaches. Shown is the weighted total mean absolute deviation (WTMAD) achieved by the methods when tested on Grimme's database.[146] Used with permission of Royal Society of Chemistry, from ref [146]; permission conveyed through Copyright Clearance Center, Inc.

the double-hybrid functionals usually outperform straightforward wavefunction-only based MP2 as shown in Figure 2.4.

The success of double-hybrid functionals can be attributed to their strong theoretical foundation in the form of the GL2 framework, as well as their ability to capture new physical effects. Just as the exact HF exchange incorporates nonlocal exchange contributions, the correlation energy should also possess inherently nonlocal components. This is evident in the significant role played by the long-range component for London dispersion interactions, that standard functionals up to hybrids fail to fully capture. By incorporating orbital-dependent correlation terms, double-hybrid functionals address this issue and improve the treatment of long-range correlation effects in molecules.[170]

Random phase approximation

The random phase approximation (RPA) is another approximation that belongs to the fifth rung. It derives the correlation energy through the adiabatic-connection-fluctuation-dissipation theorem (ACDFT)[130,171–174] in the field of DFT. Specifically, the RPA correlation energy is obtained from the dynamical response function of the non-interacting Kohn-Sham system. The RPA correlation energy E_c^{RPA} is usually combined with the total exact exchange as provided by a baseline DFT calculation, such that the total RPA energy is[130,174]

$$E^{\text{EXX+RPA}} = E_x^{\text{HF@SL}} + E_c^{\text{RPA@SL}} \tag{2.99}$$

where the quantities are computed employing the orbitals obtained through a given semilocal (SL) functional. The EXX and RPA combination is thus often referred as EXX@SL and RPA@SL with the most common choice for the DFT functional being PBE (i.e. RPA@PBE). I refer to Section 5.5.3 or ref [120] for the mathematical form of the E_c^{RPA} correlation functional. Like for double hybrids, the exact exchange plus RPA (EXX+RPA) approach has emerged as a promising method capable of more accurately capturing dispersion interactions for achieving better accuracy when predicting van der Waals binding energies, adsorption energies on surfaces, water properties, or lattice constants in molecular solids.[66,131,177–180]

Dispersion corrections

Dispersion interactions, also known as van der Waals or London forces, arise from the attraction between atoms or molecules due to correlated quantum fluctuations in their electron densities. In the context of DFT, these interactions are often referred to as van der Waals (vdW) forces. At long distances, where the overlap between the electron densities is negligible, the dispersion energy decreases at leading order with the inverse 6th power of the intermolecular distance R. Standard DFT functionals, such as local, semilocal, and hybrid functionals, primarily account for electrostatic

and polarization interactions and cannot accurately capture the $1/R^6$ behavior of dispersion. To address this limitation, various approaches have been developed to incorporate dispersion effects into DFT like resorting to rung-5 functionals as seen earlier. However, rung-5 functionals are generally so expensive that their application remains limited. To overcome the difficulty of conventional semilocal and hybrid functionals without increasing too much computational cost, lots of efforts have focused in developing vdW corrections on top of semilocal or hybrid DFAs. These are e.g., the addition of dispersion-corrected atom-centered potentials (DCACPs),[181–183] empirical dispersion corrections (-D) (e.g., Grimme's D2[184] and D3[147]), or non-local correlation (NLC) terms (e.g., (r)VV10,[185–187] vdW-DF,[188] TS-vdW[189,190]) in conjunction with semilocal functional forms. Among these, the most widely used is probably the so called DFT-D method,[147,184] that is generally given by:

$$E^{\text{DFT-D}} = E^{\text{DFT}} - \sum_{I=1}^{P} \sum_{J>i}^{P} \frac{C_{6,IJ}}{R_{IJ}^6} f_{\text{damp}}(R_{IJ}; I, J) \tag{2.100}$$

where the second term represents the dispersion energy, P is the total number of atoms, $C_{6,IJ}$ is the pairwise dispersion coefficient between atoms I and J, R_{IJ} is the intermolecular distance between atoms I and J, and $f_{\text{damp}}(R_{IJ}; I, J)$ is a damping function that ensures that the correction vanishes between any pair of atoms I and J when their densities overlap. In more recent developments (-D3 correction[147]), the coefficients $C_{6,IJ}$ depend on the chemical environments of the I-J pair, and are therefore updated on-the-fly against molecular reference parameters. In addition, terms at 8th order are included ($C_{8,IJ} f_{\text{damp}}(R_{IJ}; I, J)/R_{IJ}^8$), along with three-body correction terms that take the form of Axilrod–Teller–Muto potentials. It follows that two empirical parameters remain to be fitted depending on the initial DFA chosen (three for double hybrids).[146] Overall, dispersion corrected DFAs have allowed many successful applications with reasonable or even negligible calculation overhead. However, one should keep in mind that the performance of such corrections ultimately rely on the original DFA and an ad-hoc addition of dispersion corrections may not always improve, but may even deteriorate properties.[183,191]

Through the last sections, I have exposed the variety of ingredients entering in the definition of DFAs in KS-DFT. As a final summary, I find that Figure 2.5 illustrates well the different components to remember when choosing, using, or assessing the performance of a DFA in practice.

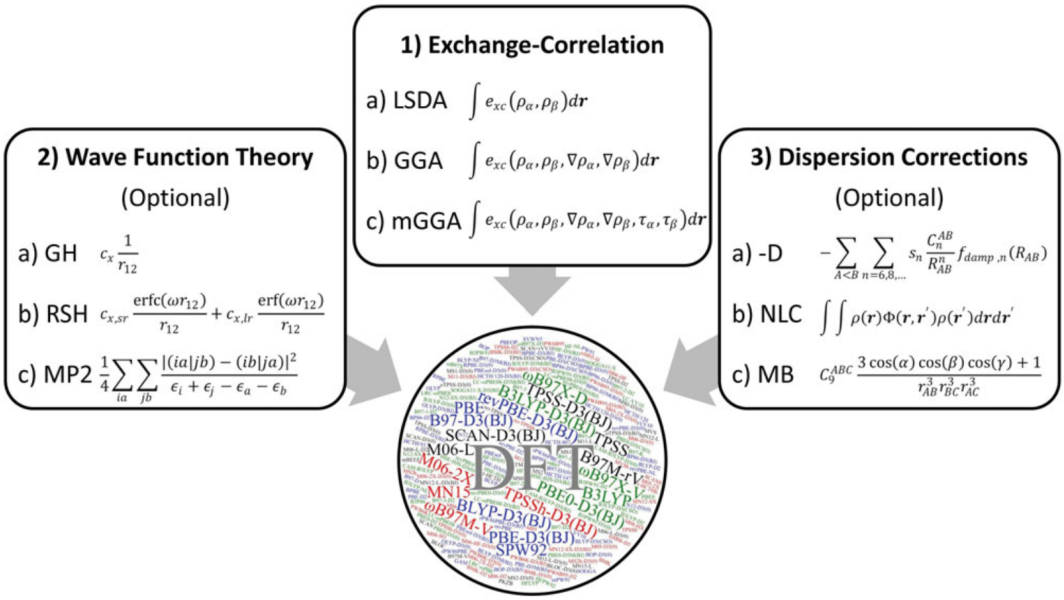


Figure 2.5: Graphical representation of ingredients entering in the definition of most current density functional approximations. The circle contains the names of density functionals available in most computational chemistry software. GH stands for global hybrid and RSH for range-separated hybrids. MB corresponds to different many-body corrections. Reproduced from ref [141] under the terms of the CC-BY-NC-ND 4.0 License.

3 Computational approaches

`Going numerical - Replacing paper and pen.`

In the previous chapter, we presented how the central many-body problem of resolving the electronic distribution around fixed nuclei can be simplified in a system of N coupled partial differential equations involving simpler one-electron orbitals, rather than treating the complete many-body wavefunction. Be it for Hartree-Fock (HF, eq 2.28) in the canonical representation or Kohn-Sham density functional theory (KS-DFT, eq 2.80), the mathematical problem at hand consequently reduces to the following eigenvalue equation:

$$\left\{-\frac{1}{2}\nabla_i^2 + v_{\text{eff}}(\mathbf{r})\right\}\phi_i(\mathbf{r}) = \varepsilon_i\phi_i(\mathbf{r}) \tag{3.1}$$

with either

$$v_{\text{eff}}(\mathbf{r}) = v_{ext}(\mathbf{r}) + \sum_{j=1}^{N}\left(\hat{J}_j - \hat{K}_j\right) \qquad \text{(HF)} \tag{3.2}$$

$$v_{\text{eff}}(\mathbf{r}) = v_{ext}(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + \frac{\delta E_{xc}[\rho]}{\delta\rho(\mathbf{r})} \qquad \text{(KS-DFT)} \tag{3.3}$$

The HF and DFT equations are therefore very similar, apart from the explicit inclusion of the non-local exchange operator $\hat{K}$ and the absence of correlation in HF. Both theories include the classical Coulomb potential $v_J(\mathbf{r})$ although expressed differently. Indeed, from eq 2.24, one has

$$v_J(\mathbf{r}) := \sum_{j=1}^{N}\hat{J}_j = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}'. \tag{3.4}$$

Furthermore, the external potential $v_{ext}(\mathbf{r})$, in the context of this thesis, is nothing else than the interaction between the electrons and the fixed nuclei in the Born-

Oppenheimer approximation. This consequently yields

$$v_{ext}(\mathbf{r}) = -\sum_{I=1}^{P} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}|}. \tag{3.5}$$

with the respective atomic number Z_I and position $\mathbf{R}_I$ of nuclei I.

As described below, the translation of such equations into a numerical problem is achieved by resorting to basis sets, ensembles of convenient mathematical functions that define the basis of the Hilbert space in which single-particle orbitals are projected.

3.1 Basis sets

Basis sets are the numerical pillars for the computational solution of the Schrödinger equation in the Born-Oppenheimer approximation. Up to now, nothing has been said about the representation of the one-electron molecular orbitals although they completely define the problem to be solved, be it HF, KS-DFT or post-HF methods. By analogy with the linear combination of atomic orbitals (LCAO), or a vectorial space representation, the spin-dependent $\phi_i(\mathbf{x}_j)$ orbitals can be expanded in a mathematical basis chosen prior to any calculation:

$$\phi_i(\mathbf{r}, \sigma) := \sum_{\alpha=1}^{M} c_{i\alpha}\chi_\alpha(\mathbf{r}) \otimes \langle\sigma|S\rangle \tag{3.6}$$

where the spatial functions $\chi_\alpha(\mathbf{r})$ define the absolute reference frame to work out the linear algebra machinery that will solve the electronic structure problem. Commonly, the spin component is considered explicitly in mathematical expressions containing the orbitals. The spin components are especially tracked when dealing with spin-polarized systems (e.g. open-shell, unrestricted, magnetic). In the closed shell approximation, electron pairs occupy the same spatial orbital $\phi_i(\mathbf{r})$.

As a choice of a basis set, there exist as many options with their respective advantages and drawbacks. Among the most widely-used today, those are either based on plane waves,[15,32] atom-centered,[13,14] augmented[16] or numerical functions.[192]

For a specific basis set $\{\chi_\alpha\}_{\alpha=1}^{M}$ composed of M basis functions, the insertion of eq 3.6 into eq 3.1 gives rise to the *generalized linear eigenvalue problem* in the basis that is

$$\sum_{\beta=1}^{M} \left(H_{\alpha\beta} - \varepsilon_i S_{\alpha\beta}\right) c_{i\beta} = 0 \iff \mathbf{Hc} = \mathcal{E}\mathbf{Sc} \tag{3.7}$$

where $\mathcal{E}$ contains the eigenvalues on its diagonal and the columns of $\mathbf{c}$ are the corresponding eigenvectors (coefficients). $\mathbf{H}$ is the effective Hamiltonian matrix expressed

in the basis and **S** is the overlap matrix that includes integral overlaps between basis functions:

$$H_{\alpha\beta} = \langle\chi_\alpha| -\frac{1}{2}\nabla^2 + \hat{v}_{\text{eff}}\,|\chi_\beta\rangle = \int \chi^*_\alpha(\mathbf{r}) \left\{-\frac{1}{2}\nabla^2 + v_{\text{eff}}(\mathbf{r})\right\} \chi_\beta(\mathbf{r}) d\mathbf{r} \tag{3.8}$$

$$S_{\alpha\beta} = \langle\chi_\alpha|\chi_\beta\rangle = \int \chi^*_\alpha(\mathbf{r})\chi_\beta(\mathbf{r}) d\mathbf{r} \tag{3.9}$$

Note that **S** is the identity matrix in case the basis is orthonormal. In HF, eq 3.7 takes the name of *Roothaan-Hall*[193,194] equations, who were the first to formulate them for closed-shell systems. Thanks to this matrix formulation of the eigenvalue problem, the solution of the electronic Schrödinger equation can be solved numerically, finally giving access to the orbitals ($c_{i\alpha}$ coefficients) and eigenvalues ε_i that determine the (post-)HF/KS-DFT ground state energies.

Various algorithms exist to solve eq 3.7. First let us note that the Hamiltonian matrix **H** depends explicitly on the molecular orbitals, thus on coefficients **c**, via the effective potential v_{eff}. This consequently prevents the use of a single diagonalization of the matrix problem and calls for more elaborate techniques capable of converging the solution **c** towards *self-consistency*. Self-consistency is reached when input orbitals **c** used to compute **H** equal the output ones **c**′ within a given threshold. I refer the reader to refs [16] and [195] for a detailed overview of the methods targeting self-consistency of eq 3.7.

Whatever the method, it relies on the numerical components of the Hamiltonian and overlap matrices, which are projected onto specifically chosen basis functions. To this end, I present below more details on atom-centered and plane-wave bases that are most relevant to this thesis.

3.1.1 Atom-centered basis

Atom-centered basis sets are based on functions that are local in real space, and located on the nuclei. These can be expressed in a spherical coordinate system as

$$\bar{\chi}_\alpha(\mathbf{r}) \to \bar{\chi}_{nlm}(\mathbf{r} - \mathbf{R}_I) \tag{3.10}$$

with

$$\bar{\chi}_{nlm}(\mathbf{r}) := \bar{\chi}_{nlm}(r,\theta,\phi) = R_{nl}(r)\mathcal{Y}_{lm}(\theta,\phi) \tag{3.11}$$

where $\mathbf{R}_I$ is the position of atom I, and n, l, m are respectively the radial and angular indices of the atomic-like functions. $\mathcal{Y}_{lm}(\theta,\phi)$ are the spherical harmonics.[11] The radial function $R_{nl}(r)$ depends on the atom-type and can take several forms that determine the class of the atom-centered basis set. In particular, if the radial function takes the form $\exp(-\zeta r)$, the basis functions are defined as Slater-type orbitals (STOs). If it rather

takes the form $\exp(-\zeta r^2)$, these are called Gaussian-type orbitals (GTOs). In practice, the basis functions χ_α are constructed as contractions of atomic functions, such that

$$\chi_\alpha(\mathbf{r}) = \sum_{\beta=1}^{K} d_\beta \bar{\chi_\beta}(\mathbf{r}) \tag{3.12}$$

where d_β are fixed (optimized) contraction coefficients and $\bar{\chi}_\beta$ are primitive (Gaussian or Slater) atomic functions.

STOs possess the benefit of satisfying the cusp condition at $r = 0$, and also demonstrate the same exponential decay as the exact atomic orbitals as $r \to \infty$. Those features are missing in GTOs. However, the main advantage of GTOs over STOs is that the product of two Gaussians centered at different points is still a Gaussian, which simplifies the calculation of two-electron integrals significantly in implementations based on GTOs. This allows for considerable speedups when considering post-HF methods, or hybrid and rung-5 DFT functionals, since these latter involve the calculation of numerous multi-electron integrals (Chapter 2).

In order to achieve the same level of accuracy as STOs for the wavefunction characteristics, several Gaussians are needed because these drop off more rapidly than the exponential function in STOs. Examples of Gaussian-type basis sets include Pople's 6-31G(*,**) and the (aug-)cc-pVXZ family of Dunning that is investigated in Chapter 4. Each of these has different numbers and shapes of Gaussians (i.e., different ζ values and contraction coefficients d_β) to best represent atomic orbitals for different purposes, offering a balance between computational cost and accuracy.[3,192] Despite this, Gaussian basis sets are inherently non-orthogonal and imply overlap integrals between the basis functions (eq 3.9). As a consequence, such bases suffer from the so called basis set superposition error when computing binding energies of complexes. Finally, although usually constructed in a systematic and strategic manner, the gradual enlargement of atom-centered bases does not necessarily guarantee a similar systematic convergence for all properties. Chapter 4 discusses these aspects in more detail for the (aug-)cc-pVXZ bases where extrapolation schemes are necessary to converge accurate energies to the complete basis set limit. For the interested reader, I recommend refs [13] and [14] which intensively discuss the use of Gaussian basis sets in computational chemistry.

3.1.2 Plane wave basis

Plane waves (PWs) are extended basis functions, intrinsically delocalized and independent of atomic positions. They are more intuitively defined in the context of condensed phases with periodic boundary conditions, characterized by an infinite replication of unit (primitive or conventional) cells of atomic positions in all directions represented

by unit vectors $\mathbf{a}_i$ ($i = 1, 2, 3, ...$, depending on the dimension) of the corresponding Bravais lattice.[196]

For such systems in the presence of a periodic external potential $v_{ext}(\mathbf{r}) = v_{ext}(\mathbf{r} + \mathbf{a}_i)$, the Bloch's theorem tells that the electronic wavefunction follows[114]

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}) \tag{3.13}$$

and is thus composed of a phase factor defined by a quantization vector $\mathbf{k}$ and a function $u_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r} + \mathbf{a}_i)$ that has the same periodicity as the external potential. Consequently, the $\mathbf{k}$ wavefunctions are periodic from cell to cell according to a phase change:

$$\psi_{\mathbf{k}}(\mathbf{r} + \mathbf{a}_i) = e^{i\mathbf{k}\cdot\mathbf{a}_i} \psi_{\mathbf{k}}(\mathbf{r}) \tag{3.14}$$

which ensures that the probability density retains the lattice periodicity $|\psi_{\mathbf{k}}(\mathbf{r})|^2 = |\psi_{\mathbf{k}}(\mathbf{r} + \mathbf{a}_i)|^2$. Bloch's theorem thus indicates that all the mathematics necessary to describe periodic systems can be reduced to the consideration of the sole real space unit cell, since wavefunctions are identical over the entire space up to a phase $e^{i\mathbf{k}\cdot\mathbf{a}_i}$. Furthermore, according to eq 3.14, vectors $\mathbf{k}$ can be chosen following $e^{i\mathbf{k}\cdot\mathbf{a}_i} = 1$ to impose the periodicity at the level of the wavefunction.

This defines a set of vectors $\mathbf{k}$ whose values should respect $\mathbf{k} \cdot \mathbf{a}_i = n \cdot 2\pi$ ($n \in \mathbb{N}$), and unit vectors $\mathbf{b}_i$ that satisfy $\mathbf{b}_j \cdot \mathbf{a}_i = 2\pi\delta_{ij}$ and span the $\mathbf{k}$-*reciprocal space* of the Bravais lattice. From this condition, it is straightforward to express the reciprocal unit vectors in terms of their real space counterparts. In the three-dimensional case,

$$\mathbf{b}_1 = 2\pi\frac{\mathbf{a}_2 \times \mathbf{a}_3}{\Omega};\ \mathbf{b}_2 = 2\pi\frac{\mathbf{a}_3 \times \mathbf{a}_1}{\Omega};\ \mathbf{b}_3 = 2\pi\frac{\mathbf{a}_1 \times \mathbf{a}_2}{\Omega} \tag{3.15}$$

with $\Omega = \mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)$ being the volume of the real space unit cell. The reciprocal space (Wigner-Seitz) primitive cell of volume $\Omega_{BZ} = \mathbf{b}_1 \cdot (\mathbf{b}_2 \times \mathbf{b}_3) = (2\pi)^3/\Omega$ is called the first Brillouin zone (BZ).

In addition, the periodicity of $u_{\mathbf{k}}(\mathbf{r})$ enables expansions in terms of Fourier series such that

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}=0}^{\infty} \tilde{\psi}_{\mathbf{k}}(\mathbf{k} + \mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \tag{3.16}$$

$$\tilde{\psi}_{\mathbf{k}}(\mathbf{k} + \mathbf{G}) = \frac{1}{\Omega}\int_{\Omega} \psi_{\mathbf{k}}(\mathbf{r}) e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} d\mathbf{r} \tag{3.17}$$

along with the condition that $e^{i\mathbf{G}\cdot\mathbf{a}_i} = 1$. This latter imposes the vectors $\mathbf{G}$ to be also located in the reciprocal space, i.e. $\mathbf{G} = n_1\mathbf{b}_1 + n_2\mathbf{b}_2 + n_3\mathbf{b}_3$ ($n_i \in \mathbb{N}$), by analogy to the previous reasoning for vectors $\mathbf{k}$.

Under periodic boundary conditions, infinite solids are viewed as $N = N_1 \times N_2 \times N_3$

three-dimensional replicas of the Bravais lattice unit cells, referred to as *supercell*. Therefore, $\psi_{\mathbf{k}}(\mathbf{r} + N_i\mathbf{a}_i) = \psi_{\mathbf{k}}(\mathbf{r})$, which reduces the spacing between points of the reciprocal space according to

$$\mathbf{k} = \frac{m_1}{N_1}\mathbf{b}_1 + \frac{m_2}{N_2}\mathbf{b}_2 + \frac{m_3}{N_3}\mathbf{b}_3 \qquad m_i = 0, 1, ..., N_i - 1 \text{ for } i = 1, 2, 3 \tag{3.18}$$

The number of $\mathbf{k}$-points contained in the BZ therefore increases, and is equal to the number N of unit cells in the Bravais lattice. Vectors $\mathbf{k}'$ out of the first BZ can always be translated back with the help of a reciprocal vector $\mathbf{G}_0$, such that $\mathbf{k} = \mathbf{k}' + \mathbf{G}_0$ belongs to the BZ. It follows that a wavefunction in $\mathbf{k}'$ out of the BZ is exactly equivalent to a function lying in the BZ:

$$\begin{aligned}\psi_{\mathbf{k}'}(\mathbf{r}) &= \sum_{\mathbf{G}=0}^{\infty} \tilde{\psi}_{\mathbf{k}'}(\mathbf{k}' + \mathbf{G}) e^{i(\mathbf{k}'+\mathbf{G})\cdot\mathbf{r}} = \sum_{\mathbf{G}=0}^{\infty} \tilde{\psi}_{\mathbf{k}-\mathbf{G}_0}(\mathbf{k} - \mathbf{G}_0 + \mathbf{G}) e^{i(\mathbf{k}-\mathbf{G}_0+\mathbf{G})\cdot\mathbf{r}} \\ &= \sum_{\mathbf{G}'=\mathbf{G}-\mathbf{G}_0}^{\infty} \tilde{\psi}_{\mathbf{k}}(\mathbf{k} + \mathbf{G}') e^{i(\mathbf{k}+\mathbf{G}')\cdot\mathbf{r}} = \psi_{\mathbf{k}}(\mathbf{r})\end{aligned} \tag{3.19}$$

The calculation of the all-electron wavefunction in an infinite periodic system has therefore been mapped onto the calculation of a finite number of electron wavefunction $\psi_{\mathbf{k}}(\mathbf{r})$ in the supercell. In the thermodynamic limit ($\Omega, N \to \infty$), the spacing between $\mathbf{k}$-points vanishes (eq 3.15 and 3.18) and an infinite number of $\mathbf{k}$-points (Bloch states) lie in the first BZ: the reciprocal space becomes continuous. In numerical practice, the continuous spacing between $\mathbf{k}$ reciprocal vectors has to be replaced by a discrete set of points that must be carefully sampled accross the BZ. A discussion of methods used for BZ sampling of periodic systems is out of the scope of this thesis but these are intensively discussed in the common literature on ab initio simulations of materials, for which the eigenvalue problem of eqs 3.1 and 3.7 has to be solved for each $\mathbf{k}$-point in the BZ ($\phi_i, \varepsilon_i \to \phi_{i,\mathbf{k}}, \varepsilon_{i,\mathbf{k}}$).[15,23]

As a consequence of the translational invariance, the basis functions in periodic systems must satisfy the Bloch's theorem, which slightly modifies the one-electron expansion encountered so far (eq 3.6) that becomes $\mathbf{k}$-dependent,

$$\phi_{i,\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} \sum_{\alpha=1}^{M} c_{i\alpha,\mathbf{k}}\chi_{\alpha}(\mathbf{r}) = \sum_{\alpha=1}^{M} c_{i\alpha,\mathbf{k}}\chi_{\alpha,\mathbf{k}}(\mathbf{r}) \tag{3.20}$$

PWs are the solutions of the Schrödinger equation of an electron in a constant external potential and respect the invariance over translation. This makes them ideal for representing basis functions under three-dimensional periodicity. PWs are defined by

$$\chi_{\mathbf{G}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} e^{i\mathbf{G}\cdot\mathbf{r}} \tag{3.21}$$

where $\mathbf{G}$ lie in the reciprocal space defined by the unit vectors $\mathbf{b}_i$ and the cell volume Ω (eq 3.15). Each plane-wave state (eq 3.21) has an energy value of $\frac{1}{2}\mathbf{G}^2$. By construction, all $\mathbf{G}$ vectors are located outside of the BZ except for $\mathbf{G} = 0$, as opposed to $\mathbf{k}$-points that all belong to the BZ. It follows that eq 3.20 can be expressed as the inverse Fourier transform

$$\phi_{i,\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} \sum_{\mathbf{G}=0}^{\mathbf{G}_{cut}} \tilde{\phi}_{i,\mathbf{k}}(\mathbf{G})\chi_{\mathbf{G}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}=0}^{\mathbf{G}_{cut}} \tilde{\phi}_{i,\mathbf{k}}(\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} = \mathrm{FT}^{-1}[e^{i\mathbf{k}\cdot\mathbf{r}}\tilde{\phi}_{i,\mathbf{k}}(\mathbf{G})](\mathbf{r}) \tag{3.22}$$

where linear coefficients are given by the direct Fourier transform (FT)

$$\begin{aligned} \tilde{\phi}_{i,\mathbf{k}}(\mathbf{G}) &= \frac{1}{\sqrt{\Omega}} \int_{\Omega} \phi_{i,\mathbf{k}}(\mathbf{r}) e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} d\mathbf{r} = \mathrm{FT}[e^{-i\mathbf{k}\cdot\mathbf{r}}\phi_{i,\mathbf{k}}(\mathbf{r})](\mathbf{G}) \\ &\simeq \frac{1}{N_R} \sum_{n=1}^{N_R} \phi_{i,\mathbf{k}}(\mathbf{r}_n) e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}_n} \end{aligned} \tag{3.23}$$

that is numerically evaluated by summing over N_R real-space grid points located at $\mathbf{r}_n$.

In principle, the Fourier series should run over an infinite number of reciprocal $\mathbf{G}$ vectors, but has to be cut in practice with respect to a user-defined *energy cutoff* E_{cut} that fixes the maximal boundary $\mathbf{G}_{cut}$ of $\mathbf{G}$ vectors according to

$$E_{cut} = \frac{1}{2} |\mathbf{k} + \mathbf{G}_{cut}|^2 \tag{3.24}$$

This truncation is justified by the fact that $\tilde{\phi}_{i,\mathbf{k}}(\mathbf{G})$ coefficients decrease when frequencies $|\mathbf{k} + \mathbf{G}|$ increase. However, the decrease rate depends on the system/wavefunction. E_{cut} should thus be chosen carefully in order to converge the electron orbitals with sufficient accuracy. The truncation must indeed maintain the *completeness* of the PW basis set to maintain the accuracy of the calculated physical quantities. The PW basis is orthonormal according to

$$\langle \chi_{\mathbf{G}} | \chi_{\mathbf{G}'} \rangle = \frac{1}{\Omega} \int_{\Omega} e^{i(\mathbf{G}'-\mathbf{G})\cdot\mathbf{r}} d\mathbf{r} = \frac{1}{\Omega} (\Omega \delta_{\mathbf{G}\mathbf{G}'}) = \delta_{\mathbf{G}\mathbf{G}'} \tag{3.25}$$

such that the overlap matrix $\mathbf{S}$ (eq 3.7) reduces to the identity matrix. This remains true when including explicitly the phase factor into the basis functions, i.e. $\chi^{\mathbf{k}}_{\mathbf{G}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}\chi_{\mathbf{G}}(\mathbf{r})$.

Having said that, one can finally arrive to the expression of the Hamiltonian matrix in the PW basis set. From eq 3.8, it appears its terms are easily evaluated in this basis. Indeed, the kinetic energy operator takes the diagonal form

$$\langle \chi^{\mathbf{k}}_{\mathbf{G}} | -\frac{1}{2}\nabla^2 | \chi^{\mathbf{k}}_{\mathbf{G}'} \rangle = \frac{1}{2} |\mathbf{k} + \mathbf{G}|^2 \delta_{\mathbf{G}\mathbf{G}'} \tag{3.26}$$

and the potential term is nothing else than the Fourier transform of the effective potential

$$\langle \chi^{\mathbf{k}}_{\mathbf{G}} | \, \hat{v}_{\text{eff}} \, | \chi^{\mathbf{k}}_{\mathbf{G}'} \rangle = \frac{1}{\Omega} \int_{\Omega} v_{\text{eff}}(\mathbf{r}) \; e^{-i(\mathbf{G}-\mathbf{G}')\cdot\mathbf{r}} d\mathbf{r} = \text{FT}[v_{\text{eff}}(\mathbf{r})](\mathbf{G}-\mathbf{G}') = \tilde{v}_{\text{eff}}(\mathbf{G}-\mathbf{G}') \quad (3.27)$$

such that the HF or KS-DFT eigenvalue problem takes a simple form in reciprocal space:

$$\sum_{\mathbf{G}'=0}^{\mathbf{G}_{cut}} \left(\frac{1}{2} |\mathbf{k}+\mathbf{G}|^2 \, \delta_{\mathbf{G}\mathbf{G}'} + \tilde{v}_{\text{eff}}(\mathbf{G}-\mathbf{G}') \right) \tilde{\phi}_{i,\mathbf{k}}(\mathbf{G}') = \varepsilon_{i,\mathbf{k}} \tilde{\phi}_{i,\mathbf{k}}(\mathbf{G}) \quad (3.28)$$

Solving the set of eqs 3.28 in the PW basis set is at the heart of the CPMD software.[47] In practice, the number of PWs ($\mathbf{G}$ vectors) contained in the basis set is of the order of 10'000-100'000 per $\mathbf{k}$-point. To handle such dimensions, CPMD includes a suite of highly-parallelized routines capable of running on supercomputing infrastructures. It also computes the terms that define $\tilde{v}_{\text{eff}}$ very efficiently, by strategically switching from reciprocal space to real space and vice versa.[16,195]

Γ-point sampling

Numerically, the reciprocal $\mathbf{k}$ and $\mathbf{G}$ spaces have to be discretized and truncated, which requires a convergence analysis of the properties of interest with respect to the $\mathbf{k}$-point sampling and the cutoff energy E_{cut}, respectively. For extended systems, with fixed volume of the (super)cell, results like total energy, energy eigenvalues and forces reach convergence when increasing the number of $\mathbf{k}$-points sampled. Alternatively, convergence is also achieved at fixed $\mathbf{k}$-point sampling of finite density if the supercell gains volume. It is evident that considerable speedups are realized if computations are run at lower $\mathbf{k}$-point density, smaller supercell size, or smaller energy cutoff. For that reason, the (approximate) sampling of the sole Γ-point ($\mathbf{k}$=0) contributions of the BZ is often chosen in association with large simulation cells. This approach is also relevant to minimize (decouple) the Coulomb interactions between the inherent periodic cells, in particular when calculating properties of interfaces, defects or surfaces, where vacuum regions must be properly extended. The Γ-point approximation is also key in making ab initio molecular dynamics simulations tractable (Chapter 6). The advantage of this simplification resides in the neglect of Bloch states, that makes the orbitals (eq 3.22) have the following expression

$$\phi_i(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}=0}^{\mathbf{G}_{cut}} \tilde{\phi}_i(\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}} = \text{FT}^{-1}[\tilde{\phi}_i(\mathbf{G})](\mathbf{r}) \quad (3.29)$$

The orbitals in real space can then be taken as real so that $\tilde{\phi}_i(-\mathbf{G}) = \tilde{\phi}_i^*(\mathbf{G})$. With this, only half of the plane-wave components have to be formally manipulated and stored in memory, and special tricks can be employed to accelerate the numerical fast FTs

(FFTs).[195] At the Γ-point, the calculation of the density reduces to

$$\rho(\mathbf{r}) = \frac{1}{\Omega} \sum_{i,j=1}^{N_{\text{occ}}} \sum_{\mathbf{G},\mathbf{G}'}^{\mathbf{G}_{cut}} \tilde{\phi}_i^*(\mathbf{G}')\tilde{\phi}_j(\mathbf{G}) e^{i(\mathbf{G}-\mathbf{G}')\cdot\mathbf{r}} = \sum_{\mathbf{G}=0}^{2\mathbf{G}_{cut}} \tilde{\rho}(\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}} \tag{3.30}$$

where the expansion of the density in reciprocal space is made consistent with the resolution on orbitals, that is the truncation of the density is generally achieved at a cutoff energy $E_{cut}^{\rho} = 4E_{cut}^{\phi}$. At the Γ-point, the wavefunction cutoff energy defines a sphere in reciprocal space, allowing to estimate the number of basis functions as

$$N_{\mathbf{G}}(\mathbf{k}=0) \simeq \frac{4\pi}{3}\left(\frac{\bar{N}_{\mathbf{a}}}{2}\right)^3 = \frac{1}{2\pi^2}\,\Omega[\text{Bohr}^3]\;E_{cut}[\text{a.u.}]^{3/2} \tag{3.31}$$

versus the average number of grid points $\bar{N}_{\mathbf{a}}$ along each real-space direction. This relation is important since it indicates that the number of PWs increases linearly with the volume of the simulation cell and slightly more with the cutoff energy. This intrinsically relates to the calculation cost, since it ultimately fixes the discretization of the Hilbert space in which the effective Hamiltonian has to be diagonalized.

A particular benefit of PW implementations relies in the fact that the Coulomb potential $v_J(\mathbf{r})$ (eq 3.4) is diagonal in reciprocal space, which facilitates the calculations of such terms. This latter is obtained from the Poisson equation

$$\nabla^2 v_J(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{3.32}$$

that relates the potential to the electron density (eq 3.30). With an expansion of the Coulomb potential in reciprocal space

$$v_J(\mathbf{r}) = \sum_{\mathbf{G}=0}^{\mathbf{G}_{cut}} v_J(\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}} \tag{3.33}$$

the solution of the Poisson equation is given by

$$v_J(\mathbf{G}) = \frac{4\pi}{\mathbf{G}^2}\rho(\mathbf{G}) \tag{3.34}$$

which therefore only depends on the $\mathbf{G}$ component in reciprocal space of the periodic setup.

When treating isolated systems (molecules), the periodicity defining the Bloch states no longer stands, and the BZ folds naturally into the Γ-point, thus avoiding $\mathbf{k}$-point sampling. Although this reduces the computational expense, additional overheads are generated due to the required decoupling scheme that cancels the Coulombic interactions between the periodic images of isolated (or semi-periodic) systems. Indeed, long-range electrostatics makes properties converge very slowly with respect to the

supercell size, often preventing the use of a simple increase of volume to decouple the periodic images. Even for polar neutral systems, the R^{-3} decay of dipole-dipole interactions, and the R^{-5} dependency between quadrupole moments sometimes necessitate a very large cell, which consequently enlarges the basis set size according to eq 3.31. Thus, as the size of the supercell increases, the number of $\mathbf{G}$ vectors increases as well, as well as their density in reciprocal space (eq 3.15). Therefore, the smallest $\mathbf{G}$-vectors become closer and closer to $\Gamma = \mathbf{0}$. In this limit, a small deviation in $\rho(\mathbf{G})$ can lead to a large error in the potential given by eq 3.34 because $\mathbf{G}$ is very small. This explains why the convergence of properties with respect to large systems/supercells becomes delicate and necessitate a careful treatment of the $\mathbf{G} \simeq 0$ instability (discussed in Section 4.3.2). Moreover, the electrostatic energy associated with charges that are periodically repeated diverges, so that rather ad-hoc and approximate approaches are employed to neutralize the system charge with a uniform background of opposite charge.[15,16] When treating semi-periodic and isolated systems, a formal solution to address those issues is the utilization of *Poisson solvers*, which offer appropriate corrected/screened formulations of the Coulomb potential by modifying the Poisson equation.[197–201]

Pseudopotentials

The main drawback of PWs is the relatively large number of basis functions needed to achieve convergence of the wavefunction. This is particularly obvious when one has to take into account fast oscillations in the wavefunction (corresponding to high $\mathbf{G}$ vectors), as they appear e.g., in the behavior of one-electron orbitals near the nuclei. Nevertheless, from a chemical point of view, core electrons are generally not involved in the bonding of molecules or atoms, which allows to neglect active contributions associated with core orbitals. The core electrons can therefore be considered as belonging to an ionic entity that they form with the nuclei. In this picture, the presence of the nuclei can be modelled by *pseudopotentials* which are effective potentials due to the ionic nuclei of charge $Z_V = Z - Z_{\text{core}}$, where Z represents the total nuclear charge and Z_{core} is the charge associated with the core electrons.

The use of pseudopotentials is essential for making PW calculations tractable, because it allows a reduction in the number of explicit electrons considered as well as a decrease in the number of basis functions necessary to converge the *pseudo*-wavefunction. Technically, the introduction of pseudopotentials introduces several modifications in the mathematical expression of the PW eigenvalue problem (eq 3.28). I will not elaborate further on the discussion of pseudopotentials, which would quickly become very technical. However, I let the interested reader consult the references [16] and [195] for a complete presentation of how the eigenvalue problem is modified when "pseudization" is used in the context of a plane wave basis set.

Part II

Improving the accuracy of the potential energy surface

4 Plane waves versus correlation-consistent basis sets: A comparison of MP2 non-covalent interaction energies in the complete basis set limit

Still, it is always wise to be prudent with any extrapolation to the infinite limit since the assumptions [...] may not always be satisfied.
— Trygve Helgaker[14]

Chapter 4 is a preprint version of an article entitled:

Villard, J.; Bircher, M. P.; Rothlisberger, U. Plane waves versus correlation-consistent basis sets: A comparison of MP2 non-covalent interaction energies in the complete basis set limit. *ChemRxiv* 2023, 10.26434/chemrxiv-2023-203z9.

Reproduced under the terms of the CC-BY 4.0 License.

4.1 Abstract

Second-order Møller Plesset perturbation theory (MP2) is the most expedient wavefunction-based method for considering electron correlation in quantum chemical calculations and as such provides a cost-effective framework to assess the effects of basis sets on correlation energies, for which the complete basis set (CBS) limit can commonly only be obtained via extrapolation techniques. Software packages providing MP2 energies are commonly based on atom-centered bases with innate issues related to possible basis set superposition errors (BSSE), especially in the case of weakly-bonded systems. Here, we present non-covalent interaction energies in the CBS limit, free of BSSE, for 20 dimer systems of the S22 dataset obtained via a highly-parallelized MP2 implementation in the plane-wave pseudopotential molecular dynamics package CPMD. The specificities related to plane waves for accurate and efficient calculations of gas-phase energies are discussed, and results compared to the localized (aug-)cc-pV[D,T,Q,5]Z correlation-consistent bases as well as their extrapolated CBS estimates. We find that the BSSE-corrected aug-cc-pV5Z basis can provide MP2 energies highly consistent with the CBS plane wave values with a minimum mean absolute deviation of $\sim$0.05 kcal/mol without the application of any extrapolation scheme. In addition, we tested the performance of 13 different extrapolation schemes and found that the X^{-3} expression applied to the (aug-)cc-pVXZ bases provides the smallest deviations against CBS plane wave values if the extrapolation sequence is composed of points D and T, while $(X+\frac{1}{2})^{-4}$ performs slightly better for TQ and Q5 extrapolations. Also, we propose $A(X-\frac{1}{2})^{-3}+B(X+\frac{1}{2})^{-4}$ as a reliable alternative to extrapolate total energies from the DTQ, TQ5 or DTQ5 data points. In spite of the general good agreement between the values obtained from the two types of basis set, it is noticed that differences between plane waves and (aug-)cc-pVXZ basis sets, extrapolated or not, tend to increase with the number of electrons, thus raising the question whether these discrepancies could indeed limit the attainable accuracy for localized bases in the limit of large systems.

4.2 Introduction

Basis functions used in any ab initio calculation, whether they concern solids or molecules, are the algebraic pillars of the electronic wavefunction whenever the Schrödinger equation has to be solved numerically. Gaussian-type orbitals (GTOs) are by far the most popular basis functions in quantum chemistry due to their atomically localized analytical forms that allow for an efficient evaluation of the multi-electron integrals appearing in wavefunction-based methods[14,192] and, to a lesser extent, in Kohn-Sham density functional theory (DFT).[135,140]

Despite their numerical advantages, GTO basis sets are inherently non-orthogonal and prone to linear dependencies that become more pronounced as the size of the basis increases. In addition, the calculation of relative energies with atom-centered functions

suffers from the basis set superposition error (BSSE)[14] because of the completeness mismatch between systems of different sizes, which tends to overstabilize bound clusters relative to single fragments, thus overestimating binding energies. Schemes for estimating the BSSE and correcting for the basis set imbalance are therefore commonly employed, with the most standard approximation being the counterpoise (CP) correction.[202,203] Nevertheless, it has been argued that such an ad hoc rectification can lead to spurious effects on the final accuracy,[203–211] for instance due to an unequal amount of correction between the entire system and its constituents.[211] As a result, the CP correction applied to the calculation of interaction energies is sometimes viewed more as an estimate. Such an estimate can be considered neither an upper nor a lower bound for the actual BSSE.[205,208,212]

An additional complication with GTO bases is that their intrinsic construction is not necessarily systematic as their size increases, which possibly leads to difficulties in converging properties in a smooth and monotonic way, which is a known problem for Hartree-Fock (HF) or correlated methods such as e.g., second-order Møller Plesset perturbation (MP2)[21] or coupled cluster (CC)[24] energies. For this reason, Dunning's cc-pVXZ correlation-consistent polarized bases[213] (where X=D,T,Q,5,6 is the *cardinal number*), as well as their augmented aug-cc-pVXZ analogues containing diffuse functions,[214,215] are among the most popular for post-HF approaches because of their meticulous design which makes it possible to gradually recover a maximum of electron correlation by increasing the basis' cardinality.

Development of such basis sets subsequently led to the proposal of numerous and mostly empirical extrapolation schemes to estimate values in the complete basis set (CBS) limit,[31,216,217] some of which are detailed later in Section 4.3.3. Although the BSSE vanishes in the CBS limit, these extrapolation procedures are not intended to directly correct for it, but rather attempt to eliminate the error due to the finite basis that we call here the basis set incompleteness error (BSIE).[203,211,218] While most extrapolation schemes are of a fully empirical nature, some are based on theoretical motivations about the leading behavior of approximate atomic wavefunctions, so that their applicability to molecular systems requires that the correlation energy be dominated by the electron-electron (Coulomb) cusp and that assumptions are transferable to polyatomic systems.[14,31,219] In addition, it is generally assumed that extrapolations are applicable from one correlated method to another. For example, the usual X^{-3} scheme introduced by Helgaker[14,220] relies on the finding that the *principal expansion* of the helium configuration-interaction (CI) energy holds for energies obtained with correlation-consistent bases, because both converge according to the principal quantum number n for this two-electron atom. Transferred to molecules, this one would then assume that lower order terms as well as chemical bonding effects are negligible, and that the expression applies equally to all system sizes and quantum chemical methods.[14]

As an alternative to cope with the BSIE, explicitly correlated methods[14,31,221,222] (e.g., MP2-R12/F12) are particularly suited for fast convergence of correlation energies due to their correlating n-electron basis functions that depend explicitly on the interelectronic coordinates r_{12}. R12 or F12 refers to whether the explicit two-electron functions (geminals) are given by linear or Gaussian-type expressions, respectively. Apart from the wavefunction Ansatz, R12/F12 methods are similar to their standard counterparts and thus converge in theory to values very close to the CBS limit. However, such calculations are neither free of BSSE for smaller bases, nor of linear dependencies for larger molecules, and they can suffer from numerical instabilities.[212,222] In addition, while recovering most of the correlation energy for a much smaller number of basis functions, they may still have difficulty incorporating the very last bits required to reach very high accuracy ($\leq$ 0.1 kcal/mol) due to the choice of the geminal and integral approximations, such as the resolution of identity or neglect of terms, that are necessary to maintain reasonable calculation costs.[14,31,220,222,223] For instance, a deviation of about 0.5 kcal/mol was observed between H_2O total energies coming from different MP2-R12 approximations and a 1 kcal/mol difference was identified between different R12 basis sets.[220] Deviations of the order of 0.1 kcal/mol were also noticed when it comes to interaction energies.[222] Nevertheless, MP2(CCSD(T))-R12/F12 calculations have so far been the only CBS references available to assess the reliability of GTO extrapolations from the (aug-)cc-pVXZ bases.[216,222,224]

In this chapter, we follow a different route to obtain converged values in the CBS limit by evaluating the MP2 correlation energy in a converged plane wave basis set. Plane waves (PWs) have the advantage of forming an orthogonal basis, the completeness of which is established regularly and monotonically with the increase of a single parameter, the kinetic energy cutoff, regardless of the level of theory employed. Since the basis functions are fixed in space rather than being located at atomic centers, there is no BSSE from the outset, and the BSIE is systematically and progressively reduced to reach the ultimate intrinsic level of precision achievable by the quantum chemical method itself. In contrast, PWs generally describe explicitly only valence electrons in order to reduce the number of basis functions and keep computational cost in store, with pseudopotentials replacing the effects of the core electrons in accommodating the variations of the wavefunction near the nuclei that would require the inclusion of rapidly varying basis functions, i.e. high energy cutoffs leading to computationally unfeasible basis set sizes. Even if pseudopotentials are employed, a PW calculation can necessitate a number of basis functions of up to a few hundred times that of GTOs (typically of the order of 10^5 PWs) for a similar level of convergence with respect to the basis set limit, thus requiring a highly-optimized parallel implementation.[225,226]

In principle, whether obtained with GTOs or PWs, CBS-converged energies must be identical. However, fundamental differences exist between those two types of basis functions and have never been thoroughly compared. We hence attempt to fill this gap with the present work. More specifically, due to the CP correction and extrapolation

schemes required for GTOs or due to some peculiarities of PWs in the treatment of isolated systems (where the interaction between periodic replica intrinsic to plane waves have to be explicitly removed), it is fundamental to clarify how the results obtained with these two different approaches may differ in practice. To answer this question, non-covalent interaction energies of dimer systems provide a sensitive test case, because the description of weak dispersion interactions often requires an accuracy of the order of 0.05-0.5 kcal/mol.[14,62,227] Such systems thus challenge the ability of a basis set to best capture the short-range components of the correlation energy around the Coulomb cusp while at the same time incorporate the long-range features of the intermolecular interactions. *A priori*, the delocalized and balanced nature of PWs seems more appropriate for such a treatment of e.g., hydrogen-bonded and van der Waals complexes, but at the cost of a much larger number of basis functions. On the other hand, polarization and delocalization of the electronic wavefunction necessitate larger and/or augmented GTO bases for a better coverage of real space.[31] If these effects are accounted for, the presence of diffuse functions causes the basis set to be more prone to the BSSE, which also makes non-covalent interactions a problem of choice for examining the effect of the CP correction.

In the following, we first give some general information on how to obtain MP2 interaction energies in the CBS limit with PWs (Section 4.3.2) as well as with correlation-consistent GTOs (Section 4.3.3); thereafter, specific computational details are reported (Section 4.4). We then demonstrate how to efficiently and accurately converge MP2 relative energies in PW basis sets as implemented in the CPMD software[47] (Section 4.5.1) and present the results of applying this approach to the calculation of non-covalent interaction energies of 20 systems from the S22 benchmark set,[62,227] which we then compare to their (aug-)cc-pVXZ analogues of different sizes X=D,T,Q,5 (Section 4.5.2). In the following, we will call our test subset S22* for brevity's sake. We then search among 13 GTO extrapolation schemes reported in the literature in order to identify those that agree the best with PWs in the CBS limit (Section 4.5.3) and investigate the capability of new, different extrapolation laws (Section 4.5.4). We find that the CBS limits reached with the best GTO extrapolations and PWs show no significant difference, i.e. they do not deviate by more that 0.2 kcal/mol for all systems studied herein. However, it is observed that some residual deviations increase with the system size (Section 4.5.5). Finally, we conclude with some general recommendations concerning the choice of correlation-consistent basis sets for the calculation of correlated energies in the CBS limit (Section 4.6).

4.3 Methods

4.3.1 Second-order Møller-Plesset perturbation theory

In Møller-Plesset perturbation theory, the dynamic correlation energy is estimated as a series of perturbative terms originating from Rayleigh-Schrödinger perturbation theory around a zero-order Hamiltonian given by the sum of Fock operators.[21] By taking the ground state Slater determinant that solves the Hartree-Fock problem as the unperturbed wavefunction, the total electronic energy E is approximated at second order by the sum of the Hartree-Fock (HF) energy and the second-order Møller-Plesset (MP2c) correlation contribution,

$$E \approx E^{\mathrm{MP2}} = E^{\mathrm{HF}} + E_c^{\mathrm{MP2}} \tag{4.1}$$

As a consequence of Brillouin's theorem, single excitations of the HF reference do not couple to the HF ground state determinant and only doubly excited determinants contribute to the MP2 correlation (MP2c) energy, leading to an expression that includes double sums over occupied and virtual molecular orbitals. For the spin-restricted case where two opposite-spin electrons occupy the same spatial orbital, the expression reads[13,192]

$$E_c^{\mathrm{MP2}} = \sum_{i}^{N_{\mathrm{occ}}} \sum_{j}^{N_{\mathrm{occ}}} \sum_{a}^{N_{\mathrm{vir}}} \sum_{b}^{N_{\mathrm{vir}}} \frac{\langle ij|ab\rangle \left(2\langle ab|ij\rangle - \langle ab|ji\rangle\right)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{4.2}$$

where i, j denote (valence-only) spatial occupied orbitals $\phi_{i,j}$ and a, b their virtual counterparts that are all eigenstates of the Fock operator with respective eigenvalues $\varepsilon_{i,j,a,b}$. The numerator in eq 4.2 accounts for Coulomb-type interactions between occupied-virtual pairs of orbitals in the evaluation of two-electron integrals

$$\langle ij|ab\rangle = \int d\mathbf{r} \int d\mathbf{r}' \frac{\phi_i^*(\mathbf{r})\phi_j^*(\mathbf{r}')\phi_a(\mathbf{r})\phi_b(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \tag{4.3}$$

The evaluation of this term in atom-centered basis-sets is straightforward and documented elsewhere.[14,192]

4.3.2 Plane wave basis set

In this chapter, we address the calculation of the MP2 energy for isolated systems in the PW basis set, considering the Γ-point sampling of the Brillouin zone only; orbitals are therefore expanded in reciprocal space as

$$|i\rangle = \sum_{\mathbf{G}=\mathbf{0}}^{G_{\mathrm{max}}^{\phi}} \tilde{\phi}_i(\mathbf{G}) \, |\mathbf{G}\rangle \tag{4.4}$$

with the reciprocal space ($\mathbf{G}$) coefficients $\tilde{\phi}_i(\mathbf{G})$ that are the Fourier components of the molecular orbitals and a PW $\langle \mathbf{r}|\mathbf{G}\rangle = \Omega^{-1/2} e^{i\mathbf{G}\cdot\mathbf{r}}$ in the simulation supercell of volume Ω. Computationally, the infinite basis is truncated by restricting the maximum norm G^{ϕ}_{max} of $\mathbf{G}$ vectors to respect

$$\frac{1}{2}G^2 \leq \frac{1}{2}\left(G^{\phi}_{\text{max}}\right)^2 := E^{\phi}_{cut} \tag{4.5}$$

The wavefunction cutoff energy E^{ϕ}_{cut}, as well as the volume Ω (both user-defined and system-dependent), act as parameters to ensure the convergence of the energy with respect to the basis size. The number $N_{\mathbf{G}}$ of basis functions is dictated by the following estimate:[195]

$$N_{\mathbf{G}} \approx \frac{1}{2\pi^2}\Omega(E^{\phi}_{cut})^{3/2} \tag{4.6}$$

In reciprocal space, the two-electron integrals (eq 4.3) can be evaluated with linear scaling with respect to $N_{\mathbf{G}}$ since the Coulomb operator takes the diagonal form

$$\langle ij|ab\rangle = \frac{1}{\Omega}\sum_{\mathbf{G}=\mathbf{0}}^{G^{\rho_{ia}}_{\text{max}}} \Phi(\mathbf{G})\rho_{ia}(\mathbf{G})\rho_{jb}(-\mathbf{G}) \tag{4.7}$$

where $\Phi(\mathbf{G})$ is the generalized Coulomb potential that depends on the dimensionality and boundary conditions of the system studied, which we describe in more detail in Section 4.3.2. The overlap pair densities appearing in eq 4.7 are obtained from the Fourier transforms

$$\rho_{ia}(\mathbf{G}) = \mathcal{F}[\phi^*_i\phi_a](\mathbf{G}) = \int d\mathbf{r}\ \phi^*_i(\mathbf{r})\phi_a(\mathbf{r})e^{-i\mathbf{G}\cdot\mathbf{r}} \tag{4.8}$$

which are replaced by Fast Fourier Transforms (FFTs) in the case of discrete representation of the ϕ_i. In principle, because the charge density ρ depends on the square of the orbitals, the maximum radius for the density expansion in the reciprocal space should be as high as $2G^{\phi}_{\text{max}}$, meaning that a density cutoff energy $E^{\rho}_{cut} = 4E^{\phi}_{cut}$ (c.f. eq 4.5) is needed to maintain a consistent resolution between orbitals and densities in reciprocal space.[195] For example, a factor 4 is used here by default in the calculation of the exchange and Coulomb integrals involved in the zero-order HF calculation. For the MP2 correlation energy (eq 4.2), whatever the ratio between $E^{\rho_{ia}}_{cut}$ and E^{ϕ}_{cut}, the canonical scaling in the PW basis set remains quintic and behaves like $\mathcal{O}\left(N^2_{\text{occ}}N^2_{\text{vir}}N^{\rho_{ia}}_{\mathbf{G}}\right)$ if the pair densities $\rho_{ia}(\mathbf{G})$ are precalculated and stored in memory. Consequently, the number $N^{\rho_{ia}}_{\mathbf{G}}$ of $\mathbf{G}$ vectors entering the expansion of ρ_{ia} has a drastic effect on the overall performance of the MP2 energy evaluation as it also affects the prefactor and memory requirements. For this reason, it is imperative to study below to what extent a reduction in $E^{\rho_{ia}}_{cut}$ alters the accuracy of the MP2 energy (Section 4.5.1).

Extrapolation of PWs to the basis set limit

The convergence of post-HF energies with respect to the number of GTO basis functions is markedly slower than it is for HF or DFT. This is attributed to the need of large atom-centered bases to fully accommodate the asymptotic behavior of the wavefunction around the electron-electron cusp.[14] For PWs associated with effective pseudopotentials, we observe that the value of E_{cut}^{ϕ} that converges relative HF energies is in general close that required for recovering most of the MP2c contribution, and both require a fairly large $N_{\mathbf{G}}$. In contrast to atom-centered bases, the sluggish convergence of the MP2 energy in PWs is rather reflected in the progression of the sum

$$E_{c,n}^{\mathrm{MP2}} = \sum_{i,j}^{N_{\mathrm{occ}}} \sum_{a,b}^{n} \frac{\langle ij|ab\rangle \left(2\langle ab|ij\rangle - \langle ab|ji\rangle\right)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{4.9}$$

with respect to the contribution of an additional virtual orbital n ($\leq N_{\mathrm{vir}}$). In theory, the effective Hilbert space defined by the numerical basis has the size $N_{\mathbf{G}} = N_{\mathrm{occ}} + N_{\mathrm{vir}}$ and the virtual space is the algebraic consequence of the basis set being larger than the number of electrons in the system. As an illustration, the MP2 calculation with the largest basis set considered in this work has $N_{\mathbf{G}} = 408\,126$ and $N_{\mathrm{occ}} = 37$, which makes the entire Fock matrix diagonalization and the direct evaluation of eq 4.2 simply intractable ($N_{\mathrm{vir}} = 408\,089$, $\sim 10^{13}$ summands, $\sim 6 \cdot 10^{12}$ $\rho_{ia}(\mathbf{G})$ points to be stored with double precision in 45 TB of RAM). For comparison, the same calculation with the GTO aug-cc-pV5Z basis set requires only 2945 basis functions. Therefore, the enormous size of the basis set coupled with the steep scaling of the methods constitute the main challenges when carrying out correlated calculations with PWs.

Fortunately, as it was established by numerical[76,117,118,121,228] and analytical[66] considerations, the PW correlation energy can be extrapolated to the CBS limit with respect to the virtual orbitals. Relying on the model of the homogeneous electron gas (HEG) in a finite cell, Shepherd et al. showed that the MP2 correlation energy in the large basis set limit ($E_{cut}^{\phi} \to \infty$) behaves like

$$E_{c,\infty}^{\mathrm{MP2}} - E_{c,E_{cut}^{\phi}}^{\mathrm{MP2}} \propto \left(E_{cut}^{\phi}\right)^{-3/2} \propto N_{\mathbf{G}}^{-1} \tag{4.10}$$

By noticing that the eigenstates of the HEG Fock matrix are nothing else than pure PWs $|\mathbf{G}\rangle$, any HF orbital of a many-electron system can be interpreted as the results of a unitary transformation of the HEG HF problem, so that the same extrapolation law generalizes to single-reference quantum chemical methods of solids and molecules (in the limit of a complete and sufficiently large basis).[66] In another interpretation, one can assume that the virtual states of very high (continuum) energy lose their molecular character and become closer to PWs, so that their contributions to the correlation energy resemble those of the HEG. Based on that, the same authors proposed a *single-point extrapolation* of eq 4.9, from intermediate points of a single calculation, that

converges smoothly and reliably to the basis set limit according to

$$E_{c,N_{\text{vir}}}^{\text{MP2}} - E_{c,n}^{\text{MP2}} \propto \varepsilon_n^{-3/2} \propto n^{-1} \tag{4.11}$$

At E_{cut}^{ϕ} ($N_{\mathbf{G}}$) sufficiently large, N_{vir} is large enough to recover the CBS energy and ε_n acts as the cutoff energy of an auxiliary basis which is gradually expanding towards the CBS limit. The fact that the orbitals and eigenvalues of eq 4.9 originate from the complete basis has no effect on the extrapolation in practice. This technique has the advantage of considerably truncating the virtual space required to calculate the MP2 correlation energy since a maximum n of the order of $n_{\text{max}} = 10\,000$-$20\,000$ is satisfactory for extrapolating relative energies. In addition, the Fock operator must be diagonalized only for the n_{max} orbitals of lowest eigenvalues.[229–231]

Despite this, MP2 computations still involve the storage of $N_{\text{occ}} n_{\text{max}} N_{\mathbf{G}}^{\rho_{ia}}$ values of the pair densities (eq 4.8) for calculating the integrals over $N_{\mathbf{G}}^{\rho_{ia}} \sim 10^4$-$10^5$ integrands (eq 4.7) as well as the contribution of $\sim 10^9$-10^{11} *ijab* sums (eq 4.9), so that only a parallel approach can handle such intense RAM and CPU requirements in a reasonable time. The PW/pseudopotential MP2 method used herein has been developed and implemented in the CPMD program[47,232] for which we give the pseudocode of the parallel implementation in Appendix A (Algorithm 1).

Treatment of the Coulomb potential for isolated systems

The PW basis set offers the possibility to evaluate the nonlocal and cumbersome Coulomb potential $\langle \mathbf{r} | \, \hat{v}_{12} \, | \mathbf{r}' \rangle = 1/|\mathbf{r} - \mathbf{r}'|$ in reciprocal space, with its Fourier transform (sampling the Γ-point only) being[195]

$$\tilde{\Phi}(\mathbf{G}) = \langle \mathbf{G} | \, \hat{v}_{12} \, | \mathbf{G}' \rangle = \mathcal{F}[\langle \mathbf{r} | \, \hat{v}_{12} \, | \mathbf{r}' \rangle] = \frac{4\pi}{\mathbf{G}^2} \delta_{\mathbf{G},\mathbf{G}'} \tag{4.12}$$

but the long-range nature of the Coulomb interactions in direct space poses problems for the evaluation of multi-center integrals such as those appearing in HF[143,233–237] or MP2 correlation energy.[117] Indeed, discrete sums of the type of eq 4.7 with $\Phi(\mathbf{G}) = \tilde{\Phi}(\mathbf{G})$ are facing a singularity in $\mathbf{G} = 0$ which is only properly integrable in the thermodynamic limit ($\Omega \to \infty$, $\sum_{\mathbf{G}} \to \Omega/(2\pi)^3 \int d\mathbf{G}$) and is a consequence of the finite simulation cell imposed by the numerical computation. Simply ignoring the problematic component makes the convergence of the integrals very slow and requires either many replicas of the unit cell (large supercell), or a much finer and careful sampling of the eventual k-points mesh. Therefore, schemes have been suggested in order to screen the $\mathbf{G} = 0$ divergence and obtain a faster convergence.[233,234,237]

In the context of hybrid functionals, be it for isolated or periodic systems, Broqvist et al. (BAP)[143] proposed to use an auxiliary function $f(\mathbf{G})$ which acts as a singularity

correction by transforming the summand of eq 4.7 into a regular function with

$$\Phi_{\text{BAP}}(\mathbf{G}) = \begin{cases} \dfrac{4\pi}{\mathbf{G}^2} & \text{for } \mathbf{G} \neq 0 \\ \chi = \dfrac{\Omega}{(2\pi)^3} \displaystyle\int d\mathbf{G}\ f(\mathbf{G}) - \sum_{\mathbf{G}\neq 0} f(\mathbf{G}) & \text{for } \mathbf{G} = 0 \end{cases} \tag{4.13}$$

and the function chosen as

$$f(\mathbf{G}) = \frac{4\pi}{\mathbf{G}^2} e^{-\gamma \mathbf{G}^2} \tag{4.14}$$

such that most of the singularity is retrieved for $f(\mathbf{G}) \to \tilde{\Phi}(\mathbf{G})$, meaning that

$$\chi = \lim_{\gamma \to 0} \left[\frac{\Omega}{\sqrt{\pi\gamma}} - \sum_{\mathbf{G}\neq 0} \frac{4\pi}{\mathbf{G}^2} e^{-\gamma \mathbf{G}^2} \right] \tag{4.15}$$

allows faster convergence of the integrals with respect to the supercell volume. For isolated systems, the symmetry of the "fictitious" supercell and its volume Ω are therefore the adjustable parameters to converge the exchange-like integrals, with the aim of removing the electrostatic interactions between periodic images when the box size increases. In that case, it has been shown that the correction χ of the singularity in the Coulomb potential greatly improves the convergence of the total energy as well as the HOMO-LUMO gap with respect to the supercell size, as opposed to simply neglecting the $\mathbf{G} = 0$ component. Hence, we will focus on the behavior of this correction when applied to the MP2 correlation energy.

An alternative treatment, specific to isolated molecules, is the effective decoupling of the Coulomb interactions between the system and its unphysical periodic replicas. For this purpose, special *Poisson solvers*[197,198,200,238] provide an expression of the potential induced by the cluster charge density when modeled in an infinitely replicated periodic setup. The method of Martyna and Tuckerman (MT)[197] assumes that the density vanishes far enough from the boundaries of the box so that the potential can be seen as having the same periodicity as the simulation domain D. In this first/nearest image picture, the potential in the MT method converges towards the isolated system limit when the supercell is sufficiently expanded. Separating its action at short and long distance with the help of the parameter α ($1/r = [\text{erfc}(\alpha r) + \text{erf}(\alpha r)]/r$), the latter can be recast as

$$\Phi_{\text{MT}}(\mathbf{G}) = \begin{cases} \dfrac{4\pi}{\mathbf{G}^2} + \underbrace{\displaystyle\int_{D(\Omega)} d\mathbf{r}\ \frac{\text{erf}(\alpha r)}{r} e^{-i\mathbf{G}\cdot\mathbf{r}}}_{\widetilde{\Phi}^{\text{long}}} & \text{for } \mathbf{G} \neq 0 \\ \dfrac{\pi}{\alpha^2} - \displaystyle\sum_{\mathbf{G}\neq 0} \underbrace{\frac{4\pi}{\mathbf{G}^2} e^{-\frac{\mathbf{G}^2}{4\alpha^2}}}_{\widetilde{\Phi}^{\text{long}}} & \text{for } \mathbf{G} = 0 \end{cases} \tag{4.16}$$

where the new terms that add up to $\tilde{\Phi}(\mathbf{G}) = 4\pi/\mathbf{G}^2$ come from the difference between the Fourier transform $\overline{\Phi}^{\text{long}}(\mathbf{G})$ and the Fourier series components $\widetilde{\Phi}^{\text{long}}(\mathbf{G})$ of the long distance part, that acts as a screen of the interactions between the isolated system and its infinite periodic images. Note that the singularity of $\tilde{\Phi}(0)$ would be exactly canceled by that of $\widetilde{\Phi}^{\text{long}}(0)$ and both were ignored in eq 4.16, but the non-singular difference $\lim_{\mathbf{G}\to 0}\left[\tilde{\Phi}(\mathbf{G}) - \widetilde{\Phi}^{\text{long}}(\mathbf{G})\right] = \frac{\pi}{\alpha^2}$ must be included. It has been shown in practice that for $\alpha L \sim 7$, where L is the smallest size of the parallelepiped box, $\overline{\Phi}^{\text{long}}(\mathbf{G})$ can be efficiently evaluated via a FFT that converges rapidly with respect to the Cartesian grid. In the framework of PW/pseudopotential DFT, the evaluation of Coulomb-like integrals with the MT Poisson solver provides accurate energies, provided that the integration domain spans about *twice the size of the electron density*. Thus, as noted by MT, increasing the size of the supercell becomes analogous to converging the energy according to the largest width diffuse function included in a Gaussian basis set.

BAP (eq 4.13) and MT (eq 4.16) schemes look very similar for $\mathbf{G} = 0$, where the sum over $\mathbf{G} \neq 0$ vectors actually corresponds to the electrostatic energy of a Gaussian charge distribution interacting with a compensating uniform background in a periodic setup.[143] The first term in eq 4.15 accounts for the electrostatic self-energy of an isolated Gaussian charge in the supercell so that, intuitively, χ corrects the singularity using the difference between the electrostatic energy of an isolated probe charge and its periodically repeated analogues in a compensating background.[143,239] For MT instead, the Gaussian charge distribution is used to construct the screening function $\overline{\Phi}^{\text{long}}(\mathbf{G}) - \widetilde{\Phi}^{\text{long}}(\mathbf{G})$ of the long-range electrostatic interactions after an Ewald-type splitting of the Coulomb potential.[34,240] Very few analyses of the BAP or the MT treatment have been performed on HF calculations with full exact exchange and, to the best of our knowledge, none has been done on the MP2 correlation energy. It is therefore crucial to understand how these act on such energy contributions to ensure the convergence and accuracy of PW results in what follows.

4.3.3 Correlation-consistent GTO basis sets

In the realm of GTOs, the cc-pVXZ[213,215] and aug-cc-pVXZ[214,215] basis families of Dunning, Peterson and coworkers have been designed to recover most of the correlation energy due to the valence electrons. More precisely, these are called correlation-consistent since the basis functions that are added at each level of cardinality X=D,T,Q,... contribute with similar amounts of energy, independently of their type (*s*,*p*,*d*,...) and in a consistent manner even in the presence of polarization or diffuse functions. All these are optimized so as to maximize their contributions to the atomic correlation energy. For example, to balance the set, *s* and *p* functions are added when the polarization space is extended, so that the correlation energy error due to the *s* and *p* functions does not exceed the error from the polarization space. The exponents of such *s* and

p functions are optimized with respect to atomic HF energies, while the correlating polarization functions come from valence energy minimization at the atomic CISD level of theory. The aug-cc-pVXZ bases are derived from the original cc-pVXZ, with the addition of one diffuse function per angular momentum present in the set, and whose (smaller) exponents are determined by minimizing the atomic CISD energy of anions. Although first intended for a better description of electron affinities, the aug-cc-pVXZ basis family has been shown to progressively enhance the convergence of other molecular properties such as proton affinities,[241] dipoles and polarizabilities,[242–244] or energies of weakly bound systems.[208,243,245–247]

The main advantage of the (aug-)cc-pVXZ basis sets is their ability to converge results toward the basis set limit in a (semi)systematic manner, at the cost of increasing the number of contracted basis functions N_b. For first-row atoms, this latter increases with the cardinal number X as[220]

$$N_b^{\text{cc-pVXZ}} = \tfrac{1}{3}(X+1)\left(X+\tfrac{3}{2}\right)(X+2) \tag{4.17}$$

$$N_b^{\text{aug-cc-pVXZ}} = N_b^{\text{cc-pVXZ}} + (X+1)^2 \tag{4.18}$$

such that, for a computer time associated with MP2 that scales as $N^5 N_b^4$ where N is the number of atoms, improving the correlation energy with a larger basis grows as $N^5 X^{12}$. Q, 5 or 6 zeta calculations may therefore be prohibitively expensive for larger systems of interest.[248–250]

This is in contrast to PWs, for which the basis size does not depend explicitly on the number of atoms N but only on the volume of the supercell and the cutoff energy, so that the MP2 energy scales as $N_{\text{occ}}^2 n_{\text{max}}^2 \Omega (E_{cut}^{\rho_{ia}})^{3/2}$ (Section 4.3.2). Assuming that a sufficiently high energy cutoff may be chosen to faithfully describe pair densities over a wide range of systems, and that n_{max} increases less than linearly with N (as we have observed[232]) , the PW basis set then becomes more favorable in the limit of large systems,[192] provided that Ω does not increase significantly for the correct convergence of Coulomb interactions in a periodic setup (Section 4.3.2).

Extrapolations of GTOs to the basis set limit

Although the correlation-consistent basis sets provide a gradual and monotonic progress, their associated computational cost grows faster than the rate of convergence. As a rule of thumb,[31] it is globally said that an improvement of the energy accuracy by a factor of 10 necessitates a computational effort increased by a factor of 10^4, and the convergence of the correlation energy, be it MP2, CCSD, CCSD(T), ..., remains so slow that basis set limit estimates can only be reached by extrapolation.[31,216,217,251] As a consequence, expressions based on the comparison with very large basis calculations or, most commonly, with explicitly correlated methods (e.g., MP2-R12/F12, CCSD(T)-

R12/F12) have been suggested, but the computational overhead for obtaining such accurate references have often restricted the size of the validation systems to a few dozen electrons.

Even though non-exhaustive, we report in List 4.1 some formulas found in the literature to estimate correlated energies in the basis set limit, when extrapolated according to $X = 2(\mathrm{D}), 3(\mathrm{T}), 4(\mathrm{Q}), 5, \ldots$. First proposals by Feller et al. (eq *Feller* (4.19)),[208,216,252–254] Peterson et al. (*Peterson* (4.20))[244,255–257] and Truhlar (*Truhlar* (4.21),(4.22))[249] are all based on empirical interpolations of the total energy, with the difference that Truhlar suggested different powers for the convergence of the HF and MP2c energies. All expressions contain three parameters and therefore require at least three points for extrapolation. However, in the case of the expressions from Truhlar, $\alpha = 3.4$ and $\beta = 2.2$ were found to provide a minimal RMSD with respect to MP2-R12 energies of small systems, so that these can also be used for two-point extrapolations (e.g., DT, TQ, Q5). In some cases, it has been argued that the CBS values calculated directly from relative rather than total energies are more accurate,[216,244] but no clear explanation or justification was provided to support that claim.[224]

$E_X^{MP2} = E_\infty^{MP2}$	$+ Ae^{-\alpha X}$	*Feller* (4.19)
$E_X^{MP2} = E_\infty^{MP2}$	$+ Ae^{-(X-1)} + Be^{-(X-1)^2}$	*Peterson* (4.20)
$E_X^{HF} = E_\infty^{HF}$	$+ AX^{-\alpha}$	*Truhlar* (4.21)
$E_{c,X}^{MP2} = E_{c,\infty}^{MP2}$	$+ BX^{-\beta}$	(4.22)
$E_X^{MP2} = E_\infty^{MP2}$	$+ A\left(X+\frac{1}{2}\right)^{-4}$	*Martin4* (4.23)
$E_X^{MP2} = E_\infty^{MP2}$	$+ A\left(X+\frac{1}{2}\right)^{-4} + B\left(X+\frac{1}{2}\right)^{-6}$	*Martin46* (4.24)
$E_X^{MP2} = E_\infty^{MP2}$	$+ A\left(X+\frac{1}{2}\right)^{-\alpha}$	*Martin*α (4.25)
$E_{c,X}^{MP2} = E_{c,\infty}^{MP2}$	$+ AX^{-3} + CX^{-5}$	*Wilson35* (4.26)
$E_{c,X}^{MP2} = E_{c,\infty}^{MP2}$	$+ A(X+1)^{-4} + B(X+1)^{-5}$	*Wilson45* (4.27)
$E_X^{HF} = E_\infty^{HF}$	$+ Ae^{-\alpha X}$	*Helgaker* (4.28)
$E_{c,X}^{MP2} = E_{c,\infty}^{MP2}$	$+ BX^{-3}$	(4.29)
$E_{c,X}^{MP2} = E_{c,\infty}^{MP2}$	$+ AX^{-3} + BX^{-4}$	*Varandas34* (4.30)
$E_{c,X}^{MP2} = E_{c,\infty}^{MP2}$	$(1 - 2.4X^{-3})$	*Varandas3-fit* (4.31)
$E_{c,X}^{MP2} = E_{c,\infty}^{MP2}$	$(1 + AX^{-3} + A[ae^{bA} + c]X^{-4})$ where $a = 6.1793$, $b = 1.0940$, $c = -0.9766$ are optimized parameters.	*Var.34-fit* (4.32)

LIST 4.1: Extrapolation expressions tested in this work for the (aug-)cc-pVXZ basis sets.

On the other hand, one possible theoretical motivation holds its origin in the partial-wave expansion of a two-electron atom:[14,31,217] By treating the Hamiltonian of the bare nucleus of the helium atom as zero-order, and the electron interaction as a perturbation, Schwartz studied the convergence of the correlated atomic wavefunction in a one-electron basis.[31,258] He established that the partial wave increments $\delta E_l^{(2)}$ to the second-order energy $E^{(2)} = \sum_{l=0}^{\infty} \delta E_l^{(2)}$ follow an asymptotic formula in the limit of large l, that behaves like

$$\delta E_l^{(2)} = A \left(l + \tfrac{1}{2}\right)^{-4} + B \left(l + \tfrac{1}{2}\right)^{-6} + \mathcal{O}\left(l^{-8}\right) \tag{4.33}$$

with l being the degree of Legendre polynomials entering the partial wave expansion of the first-order wavefunction. Interestingly, the l-th component in the partial wave expansion corresponds to a one-electron atomic function with angular momentum l.[259] Translated to many-electron atoms, this implies that $\delta E_l^{(2)}$ is equivalent to the energy increase due to the addition of a saturated shell of basis functions of angular momentum l to the basis set that expands the first-order wavefunction. For standard electronic structure methods (e.g., MPn[259] or CI[260,261]) and n-electron atoms, similar forms were derived with odd terms that may also arise, where one makes the general assumption that the increment of the correlation energy can be expanded as

$$\delta E_l^{(2)} = \sum_{m=4} A_m \left(l + \tfrac{1}{2}\right)^{-m} \tag{4.34}$$

with numerical coefficients A_m. In the limit of large L, with the omission of all basis functions with $l > L$, the error on the correlation energy resulting from the basis set truncation can therefore be estimated as[31,250]

$$E_{c,\infty} - E_{c,L} \approx \sum_{m=4} A_m \int_{L+1/2}^{\infty} \left(l + \tfrac{1}{2}\right)^{-m} dl \tag{4.35}$$

$$= \sum_{m=4} \frac{A_m}{m-1} \left(L+1\right)^{-m+1} \tag{4.36}$$

which consequently describes the asymptotic limit of the energy for consecutive enlargements of the basis set. This stands under the assumption that each increment of the basis set contains all functions covering the atomic angular momentum up to L. However, choosing the atomic angular momentum as the parameter for assessing energy convergence is questionable when generalizing to molecules. Moreover, this quantum number is not consistent with the construction of the (aug-)cc-pVXZ bases that rather involves successive increments of functions with different angular momenta (Section 4.3.3).

In spite of this, expressions inspired by eq 4.36 have demonstrated their potential for the extrapolation of (aug-)cc-pVXZ energies to the basis set limit, also in the case of polyatomic systems. For example, Martin et al. proposed to average between hydrogen,

helium ($L \sim X - 1$) and first-row ($L \sim X$) atoms to replace L by $X - \frac{1}{2}$, yielding eqs *Martin4* (4.23) and *Martin46* (4.24)[219,262] that correspond to the leading orders found by Kutzelnigg et al. for the MP2 energies.[259] He later suggested that the quality of the results can improve if the HF and MP2c energies are processed separately with *Martin*α (4.25).[263]

Furthermore, by comparing with MP2-R12 references of Ne, HF, H_2O and N_2, Wilson and Dunning found that the ansatz

$$E_{c,X}^{MP2} = E_{c,\infty}^{MP2} + \frac{A}{(X+D)^{\alpha}} + \frac{B}{(X+D)^{\alpha+1}} + \frac{C}{(X+D)^{\alpha+2}} \tag{4.37}$$

was giving the best match for ($\alpha = 3$, $B = 0$, $D = 0$) and ($\alpha = 4$, $C = 0$, $D = 1$), consequently proposing eqs *Wilson35* (4.26) and *Wilson45* (4.27) for extrapolating correlated energies.[224]

Alternatively, Helgaker et al. put forward the use of eq *Helgaker* (4.29)[212,220,223,264] which corresponds to the leading order of eq 4.36 and the identification of $L \sim X - 1$ based on a better agreement with R12 results.[220] This allows the extrapolation of the correlation energy from a two-point linear fit and was later motivated by analyzing the energy increments of the *principal expansion* of the ground-state helium atom, yielding in this case a convergence with respect to the principal quantum number n which is *a priori* more in line with the progressive construction of the (aug-)cc-pVXZ bases.[14,31] Since the number of basis functions increases cubically with X (eq 4.17), *Helgaker* (4.29) is equivalent to a convergence of the correlation energy as a function of $1/N_b$. Interestingly, this is analogous to PWs for which the correlation energy converges as $1/N_{\mathbf{G}}$ (eq 4.10). A different rate of convergence was observed for the HF energy so that Helgaker argued for its separate treatment according to *Helgaker* (4.28).[220,264]

Finally, Varandas investigated the universality of the B parameter in *Helgaker* (4.29), and found that eq *Varandas3-fit* (4.31) minimizes the difference with a set of CCSD(T)/MP2 energies of small molecules and different basis sets.[250] However, the match was better if considering a fourth order term with the general form of *Varandas34* (4.30), or by exploiting an empirical interdependence between the parameters that leads to eq *Var.34-fit* (4.32) which has the advantage of requiring only two points for extrapolation.

Subsequently, we will examine which of these extrapolations provide better or worse agreement between GTO and PW MP2 interaction energies in the CBS limit.

4.4 Computational details

The geometries of the test systems were taken from the paper defining the S22 dataset.[62] Like for the original work on the S22 dataset and its revised version,[227] deformation en-

ergies are neglected, monomer structures are kept identical in the dimer configuration and no monomer relaxation is carried out. To save on computational resources, we opted to exclude the two adenine-thymine dimers from our test set, resulting in a set of 20 structures that we refer to as S22*.

4.4.1 Plane wave basis set

A development version of CPMD 4.3 has been used for all MP2 calculations with PWs,[47] in combination with hard normconserving Goedecker-Teter-Hutter (GTH)[265] pseudopotentials specifically parametrized for HF.[119] The HF wavefunction has been optimized with either DIIS[266] or preconditioned conjugate gradient optimization up to a maximum residual component of the gradient on occupied orbitals lower than 10^{-7} a.u., respectively 10^{-5} a.u. for the $n_{\max}$ virtual orbitals obtained via subsequent Davidson diagonalization.[231]

The MP2c contribution to the interaction energy of the AB complex is extrapolated according to eq 4.11, i.e. the extrapolation is performed on $\Delta E_{c,n}^{\mathrm{MP2}} = E_{AB,c,n}^{\mathrm{MP2}} - E_{A,c,n}^{\mathrm{MP2}} - E_{B,c,n}^{\mathrm{MP2}}$ which greatly accelerates the energy convergence compared to individual extrapolations. This allows to set smaller $n_{\max}$ virtual orbitals to be diagonalized and processed in the MP2c double summation, and thus drastically reduces the computational requirements. No significant difference (> 0.001 kcal/mol) was observed if the extrapolation is done as a function of n^{-1} or $\varepsilon_n^{-3/2}$, the latter eigenvalues corresponding to the dimer being therefore used. Extrapolation points are spaced by an increment of 100 virtual orbitals and a better accuracy is obtained with a linear fit according to

$$\Delta E_{c,n}^{\mathrm{MP2}} \cdot \varepsilon_n^{3/2} = \alpha \varepsilon_n^{3/2} + \beta \tag{4.38}$$

where $\alpha = \Delta E_c^{\mathrm{MP2}}$ recovers the PW CBS MP2c energy by ensuring that $n_{\max}$ is chosen large enough in order for eq 4.38, respectively eq 4.11, to be valid. For the systems studied, $n_{\max}$ is between 10000 and 20000. To account for the sensitivity of the extrapolated value with respect to the fitting range, the results from all possible intervals ending at $n_{\max}$ are calculated, and the final $\Delta E_c^{\mathrm{MP2}}$ value averaged among the intervals that respect eq 4.38.

The cutoff energy E_{cut}^{ϕ} of the wavefunction has been set to 150 Ry, and the density cutoff to the usual $E_{cut}^{\rho} = 4E_{cut}^{\phi}$ for all systems and supercell sizes. No change larger than ∼0.01 kcal/mol was observed on the extrapolated MP2 interaction energies at larger cutoffs (cf. Table A1 in Appendix A). The effects of the cutoff energy for the MP2c pair densities, the supercell dimensions as well as the decoupling between periodic images are discussed below in Section 4.5.1.

4.4.2 Correlation-consistent GTO basis sets

The (aug-)cc-pVXZ calculations were performed with Orca 5.0.3,[267,268] or for the larger systems and augmented bases with Turbomole V7.1[269] after checking that both programs give identical results. The HF wavefunction and energies are obtained for all-electron calculations, while the *frozen-core* approximation is used for the MP2 correlation energy, i.e. occupied orbitals corresponding to core electrons are omitted in the MP2c evaluation. The convergence threshold for the SCF wavefunction was set to `VeryTightSCF` for Orca, respectively to 10^{-7} a.u. for the energy gradient in Turbomole.

The CP correction scheme is used to correct for the BSSE.[202,203] Therefore, uncorrected and BSSE-corrected interaction energies (HF or MP2c) are calculated as follows

$$\Delta E_{\text{uncorr.}} = E_{AB}^{\{AB\}} - E_{A}^{\{A\}} - E_{B}^{\{B\}} \tag{4.39}$$

$$\Delta E_{\text{CP-corr.}} = E_{AB}^{\{AB\}} - E_{A}^{\{AB\}} - E_{B}^{\{AB\}} \tag{4.40}$$

$$\Delta E_{\text{half-CP}} = \tfrac{1}{2}\left(\Delta E_{\text{uncorr.}} + \Delta E_{\text{CP-corr.}}\right) \tag{4.41}$$

where $E_{A}^{\{A\}}$ designates the energy of monomer A calculated in its basis $\{A\}$, and $E_{A}^{\{AB\}}$ its corrected energy calculated in the full $\{AB\}$ basis that includes *ghost* functions located on system B. Since it was noticed that $\Delta E_{\text{uncorr.}}$ and $\Delta E_{\text{CP-corr.}}$ may converge to the basis set limit from opposite sides, it is sometimes assumed that the average $\Delta E_{\text{half-CP}}$ energy provides faster convergence,[27,212,270,271] a strategy that we also examine.

4.5 Results and discussion

4.5.1 Converging accurate MP2 energies with plane waves

In this section, we discuss a number of technical details that are essential for making calculations of MP2 interaction energies with PWs tractable. As already mentioned, the leading computational effort for this task scales as $\mathcal{O}\left(N_{\text{occ}}^2 n_{\text{max}}^2 \Omega (E_{cut}^{\rho_{ia}})^{3/2}\right)$ for which n_{max} is reduced by the joint extrapolation of relative energies in the virtual space. Moreover, Ω and $E_{cut}^{\rho_{ia}}$ define the number $N_{\text{occ}} n_{\text{max}} N_{\mathbf{G}}^{\rho_{ia}}$ of $\rho_{ia}(\mathbf{G})$ pair density values (eq 4.6) to be stored for the two-electron integrals (eq 4.7), and therefore have a strong influence on the memory requirements.

The effect of $E_{cut}^{\rho_{ia}}$ on the correlation energy is reported in Table 4.1 which shows that no difference greater than 0.01 kcal/mol results from reducing $E_{cut}^{\rho_{ia}}$ to the wavefunction cutoff energy E_{cut}^{ϕ} (150 Ry) for various systems and supercell volumes. This amounts to projecting the pair densities, which require less high-frequency components, onto an auxiliary basis set for efficient computation of integrals, similar to what is done for example in the resolution of identity with GTO approaches (e.g., RI-MP2).[272] We

Table 4.1: MP2c contribution to the MP2 interaction energy for selected systems from the S22 set and pair density cutoff energies $E_{cut}^{\rho_{ia}}$. Energies are in [kcal/mol], obtained with the MT Poisson solver. $r_{x,y,z}$ are the respective x, y, z ratios of the orthorhombic supercell dimensions with respect to the HF electron density measured at an isosurface of 0.002 a.u., while Ω is the volume of the supercell. σ_c^{MP2} corresponds to the standard deviation of ΔE_c^{MP2} values extrapolated on different fitting ranges in the virtual space.

S22 system	$r_x\ r_y\ r_z$	Ω [Å^3]	$E_{cut}^{\rho_{ia}}$ [Ry]	ΔE_c^{MP2}	σ_c^{MP2}
$(NH_3)_2$	2.0 2.0 2.0	987.84	150	-1.763	0.004
			300	-1.762	0.004
			600	-1.762	0.004
$(H_2O)_2$	1.7 1.9 2.2	648.86	150	-1.354	0.004
			300	-1.347	0.004
			600	-1.347	0.004
Formic acid	1.5 1.6 2.3	953.50	150	-3.093	0.013
			300	-3.095	0.013
Formamide	1.4 1.4 1.4	539.82	150	-3.658	0.004
			300	-3.655	0.005
			600	-3.655	0.006
PD benzene	1.8 1.8 1.8	2913.80	150	-10.470	0.021
			300	-10.461	0.024

strongly emphasize the great benefit of such a reduction; in the case of e.g., the parallel-displaced (PD) benzene dimer, computing the MP2 energies with $E_{cut}^{\rho_{ia}} = 600$ Ry is simply impossible on 25 nodes with 128 GB of memory each, while all test systems reported below could be evaluated with such a setup by fixing $E_{cut}^{\rho_{ia}}$ to 150 Ry from now on.

The last parameter affecting the computational cost is the supercell volume that should be as small as possible while accurately decoupling the interactions between periodic images of the system. As we saw, the choice of Ω is mainly dictated by the treatment of low frequency components of the Coulomb operator acting in exchange-like integrals (Section 4.3.2). Figure 4.1 shows that significant differences exist between the BAP (eq 4.13) and MT (eq 4.16) potentials for converging interaction energies. As already observed,[143] BAP greatly accelerates the convergence of the total HF energy compared to the simple neglect of the $\mathbf{G} = 0$ component (Figure A1). However, both schemes perform identically when it comes to HF interaction energies (Figure 4.1a). This is explained by computing the BAP correction

$$\begin{aligned} E^{\text{HF}}_{\Phi_{\text{BAP}}(\mathbf{G})} - E^{\text{HF}}_{\Phi(\mathbf{G}=0)=0} &= -\frac{1}{2\Omega}\sum_{i,j}^{N_{\text{occ}}} \chi \rho_{ij}(0)\rho_{ji}(0) \\ &= -\frac{1}{2\Omega}\sum_{i,j}^{N_{\text{occ}}} \chi \delta_{ij} = -\frac{\chi}{4\Omega} N_e \end{aligned} \tag{4.42}$$

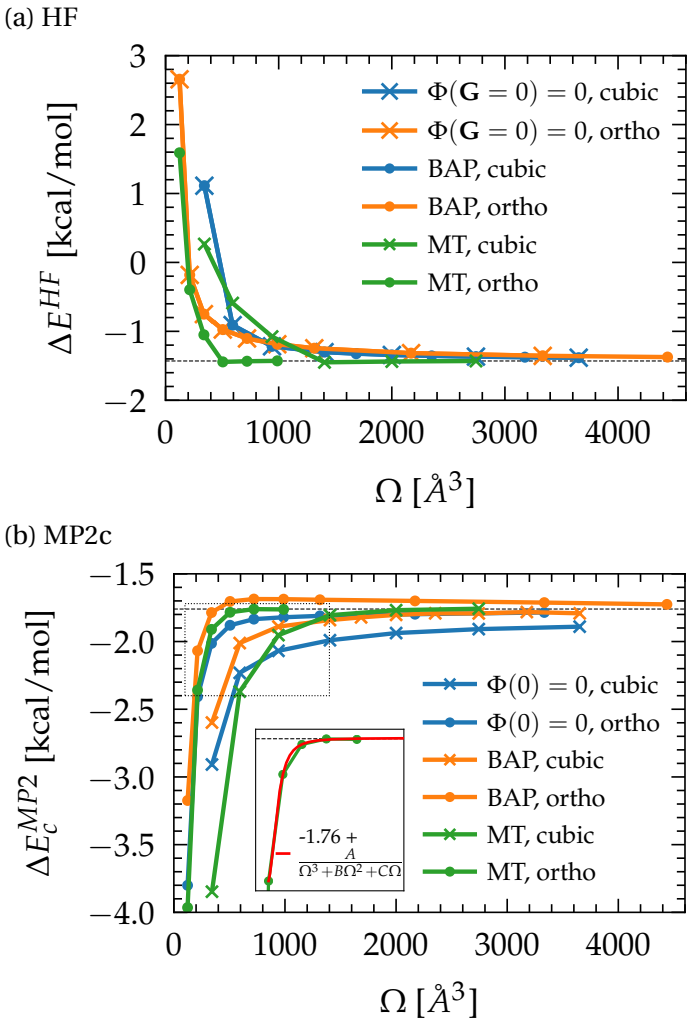


Figure 4.1: (a) HF and (b) MP2c contributions to the MP2 interaction energy of the NH_3 dimer for different exchange (Coulomb) potentials. Ω is the volume of expanding cubic or orthorombic (ortho) supercells around the dimer electron density.

that is proportional to the number of electron N_e in the system and consequently cancels out between the energies of the dimer and monomers. Therefore, although beneficial for total energies, the BAP scheme does not improve the convergence of HF interaction energies for both cubic and orthorhombic boxes, and necessitates large volumes to recover the last fraction of the mean field energy. Moreover, although a cubic box expansion with BAP accelerates the MP2c convergence against $\Phi(0) = 0$ (Figure 4.1b), the BAP correction with an orthorhombic box makes it converge more slowly and non-monotonically. This is a consequence of the BAP singularity correction (eq 4.15) that may switch sign depending on if the repulsion between the repeated Gaussian charge images or their attraction with the compensating background dominates according to the elongation of the cell.[143] This demonstrates that the convergence behavior of the MP2 correlation energy for various supercell sizes and symmetries is

non-trivial when resorting to effective Coulomb potentials.

In comparison, the MT potential performs better and is consistent between HF and MP2c contributions, most presumably thanks to the explicit cancellation of the $\mathbf{G} = 0$ singularity and the directional effects of the screening function $\overline{\Phi}^{\text{long}}(\mathbf{G}) - \widetilde{\Phi}^{\text{long}}(\mathbf{G})$. Expanding an orthorhombic box around the electron density coupled to the MT Poisson solver allows to converge the MP2 interaction energy for a much smaller volume/-computational cost, e.g., with a reduction factor of approximately 3 to 4 against other schemes for the NH_3 dimer test system. In addition, it has been found that the MP2c energies of all systems considered herein can be extrapolated at large Ω according to

$$\Delta E_{c,\Omega}^{\text{MP2}} = \Delta E_{c,\Omega\to\infty}^{\text{MP2}} + \frac{A}{\Omega^3 + B\Omega^2 + C\Omega} \tag{4.43}$$

as illustrated in the inset of Figure 4.1b. While it is well established that the MT potential requires the supercell to span at least twice the size of the density to converge DFT energies,[47,197] our results show for the first time that the same criterion also applies to the MP2 energy. For practical information, all MP2 interaction energies considered in this work are converged to within 0.07 kcal/mol when setting the orthorhombic cell dimensions to $r_{x,y,z} = 1.8$ times the extent of the density, measured at an 0.002 a.u. isosurface. For very high accuracy ($\sim$0.01 kcal/mol), a ratio of 2.2 is recommended instead. Hence, eq 4.43 is of significant help to ensure the recovering of the last fraction of the correlation energy, and becomes indispensable for the treatment of larger systems that would impose a too large box size and intractable computational cost.

Within these settings, we have shown how various factors can push the limits of MP2 calculations with PWs. The first factor consists of truncating the virtual space thanks to an analytical extrapolation (eq 4.11), the second relates to the reduced number of PWs necessary to expand the pair densities and, finally, the last refers to the choice of an efficient Coulomb operator for treating isolated systems and correlation energies. Thanks to these findings, it has been made possible to access the MP2 interaction energies of systems with up to $\sim$100 electrons that are listed in Table A2 of Appendix A. Convergence was achieved by progressively expanding an orthorhombic cell ($r_{x,y,z} = 1.2, 1.4, 1.6, 1.8$ and $2.0, 2.2$ when possible) with the MT Poisson solver. The HF components were retained when no variations larger than 0.01 kcal/mol were measured, while MP2c contributions were extrapolated first via eq 4.38 and then eq 4.43. Standard deviations due to this extrapolation procedure are also reported in Table A2 and do not exceed 0.05 kcal/mol for energies spanning a range from -0.50 to -20.19 kcal/mol.

4.5.2 HF/MP2 energies in PWs versus GTO bases

Figure 4.2 displays the statistics of the differences between GTO and PW contributions to the MP2 interaction energy. For HF, the CP uncorrected or half-corrected energies converge from below and confirm that the BSSE tends to overbind dimer systems at the HF level. Once the BSSE is removed by the CP correction, HF energies converge faster and from above, which is the expected behavior from a gradual decrease of the (sole) BSIE as the size of the basis set increases.[212] The augmented basis converges slightly faster than its standard counterpart, certainly due to its larger size and spatial extent for the same cardinal number. When CP-corrected, the HF energies are already converged within less than 0.2 kcal/mol for the triple (T) zeta bases. Overall, at each cardinal number, the CP-corrected results obtained with the augmented basis sets provide the best agreement with PWs as also reported in Table 4.2. The remarkably small deviations at the HF level between the Q/5 zeta all-electron GTOs and PWs support the fact that the use of pseudopotentials does not cause any spurious differences between the PW (pseudopotential) and the GTO (all-electron) results.

The MP2c correlation, and hence the MP2 energies, are more sensitive to the basis set size and slower to converge than HF, with e.g., MP2 deviations of about 0.7-0.8 kcal/mol for the T zeta bases. This is because the MP2c energy is more prone to the BSIE which is noticeably exacerbated for the smaller cc-pVDZ and cc-pVTZ bases once CP-corrections are applied. Overall, D zeta basis sets do not provide a satisfactory level of convergence, with deviations that might surpass the order of 2 kcal/mol whatever

Table 4.2: Best agreement between GTO and PW interaction energies of the S22* test set at each cardinal number. Mean absolute errors (MAE) and maximum deviations are in [kcal/mol].

Level	Size	Set	BSSE corr.	MAE	Max dev.
HF	D	non-aug	CP	0.16	0.69
		aug	CP	0.07	0.48
	T	non-aug	CP	0.06	0.15
		aug	CP	0.04	0.12
	Q	non-aug	CP	0.02	-0.08
		aug	CP	0.02	0.05
	5	non-aug	CP	0.02	0.04
		aug	CP	0.02	0.05
MP2	D	non-aug	half-CP	0.97	2.71
		aug	half-CP	0.68	-1.88
	T	non-aug	half-CP	0.24	0.70
		aug	CP	0.30	0.82
	Q	non-aug	half-CP	0.11	0.31
		aug	CP	0.11	0.23
	5	non-aug	half-CP	0.07	-0.20
		aug	CP	0.06	0.16

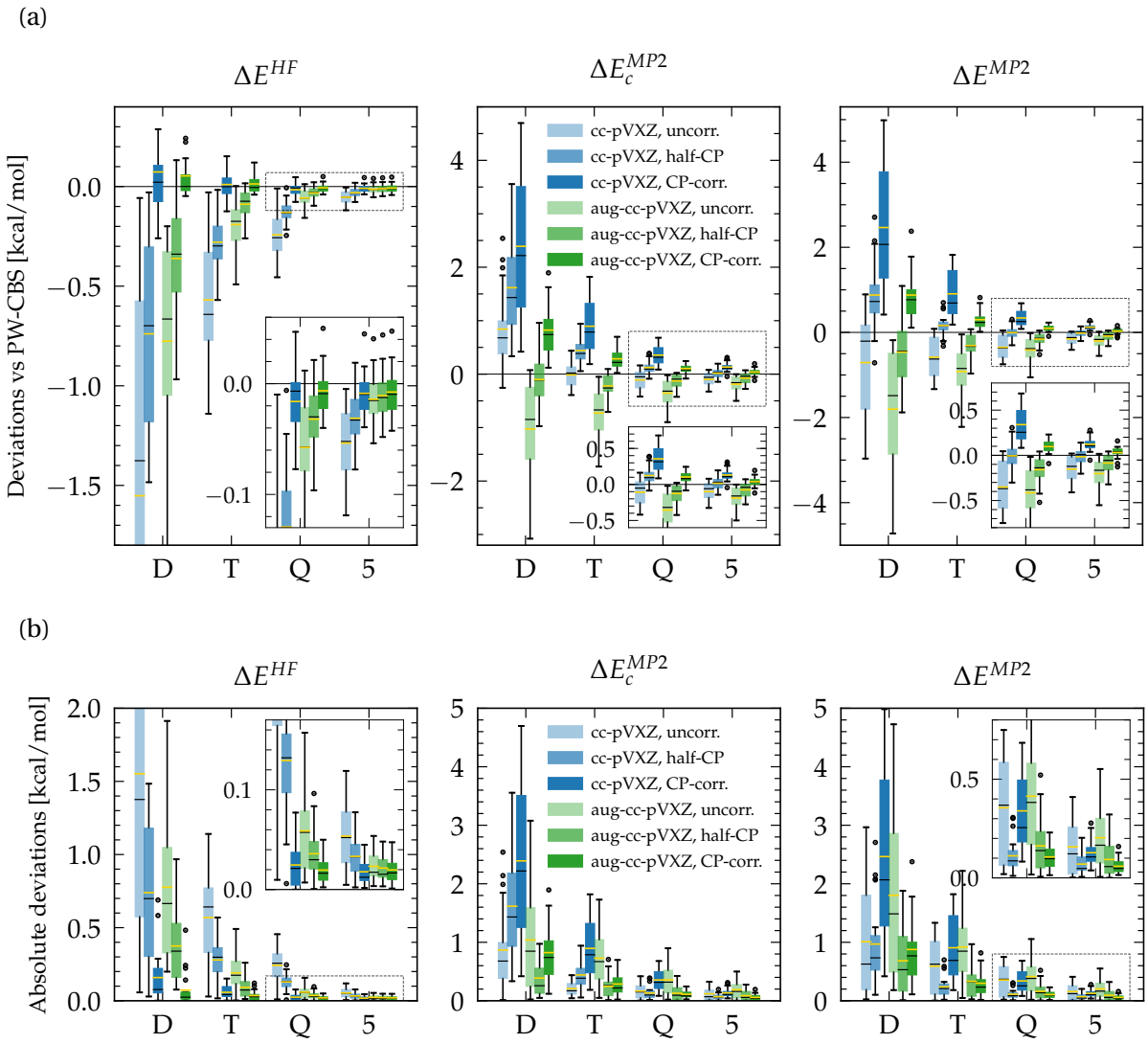


Figure 4.2: Box plots of the differences $\Delta E_{\mathrm{GTO}} - \Delta E_{\mathrm{PW}}$ between GTO and PW interaction energies of the S22* test systems. Separate HF and MP2c contributions to the total MP2 energies are given. Signed differences are given in (a) while (b) reports absolute values. Medians are shown as horizontal black lines and yellow lines stand for the mean signed deviation (MSD) in (a) and the mean absolute error (MAE) in (b), respectively. Dots represent outliers that are located further than 1.5 times the interquartile range from respectively the first and third quartiles (i.e. the limits of the rectangular boxes).

the augmentation or the BSSE correction. At small D and T cardinalities, the best match with PWs is observed when considering only half of the CP correction with the (aug-)cc-pVDZ or cc-pVTZ bases, respectively (Table 4.2), which confirms that the accuracy at such levels is mainly due to a fortuitous cancellation of the BSSE, which lowers the energies, and the BSIE which increases them. Although not formally recommended, this can be exploited for a crude first estimate in case of limited computational budget,

as provided for example by the cc-pVTZ/half-CP combination (with potential ∼1 kcal/mol error). For the same reason, the match between non-augmented Q/5 bases and PWs is better with only half-CP correction. For the augmented bases, however, the same conclusions as for HF hold: the MP2c and MP2 energies are generally too low without a full BSSE correction, and approach the CBS PW values from above when corrected. Thanks to this, the overall best agreement between GTOs and PWs is measured for the aug-cc-pV5Z basis with CP correction, that shows a MAE of only 0.06 kcal/mol.

As a result, the gradual convergence of GTOs toward PW values validates both the MP2 implementation in CPMD as well as the previously proposed extrapolation procedure to compute PW interaction energies in the CBS limit (Section 4.5.1). Furthermore, our results highlight the importance of diffuse basis functions needed to incorporate the long-range components of the electron correlation in weakly bound systems[a], and in this respect provide further confirmation of the use of (aug-)cc-pVXZ bases for converging binding/interaction energies with correlated methods, although they were originally designed for the treatment of anions and electron affinities.[214,215,244]

4.5.3 HF/MP2 energies in PWs versus extrapolated GTO bases

We are now interested in exploring whether the different extrapolations of List 4.1 for consecutive X=D,T,Q,5 cardinal numbers improve or deteriorate the GTO CBS estimates as compared to PW energies. When uncorrected for the BSSE, HF and MP2 interaction energies converge non-systematically and sometimes non-monotonically because of the varying balance between BSSE and BSIE (as illustrated in Figure A2 or ref [212]). For this reason, GTO extrapolations performed on relative energies yield results that are either very similar or worse than those for absolute total energies. Thus, in what follows, the energy of each subsystem will rather be extrapolated individually.

Truhlar (4.21) and *Helgaker* (4.28) treat the HF contribution separately[b] and Figure 4.3 shows how they perform in this regard. Both contain three parameters and require at least three data points for extrapolation. The power expression of *Truhlar* with D, T, and Q (DTQ) data points generally worsens the deviations from PWs compared to simple Q energies. When including the 5 point (TQ5, DTQ5), results are also worse than or comparable to the plain 5 values in terms of MAE and maximum deviation for both non-augmented and augmented bases. For *Helgaker*, DTQ points improve the convergence of uncorrected and half-CP Q zeta energies (Figure 4.2b, ΔE^{HF}) but slightly deteriorate those that are CP corrected. The TQ5 results are essentially similar to the plain 5 zeta energies, and all extrapolated values are either slightly better or

[a] For stronger (e.g., covalent) interactions, however, diffuse augmentation was sometimes found to hamper the basis set convergence of correlation energies.[251]

[b] Note that *Martin*α (4.25) was also proposed for HF energies, and yields identical conclusions to *Truhlar* (4.21) (not reported).

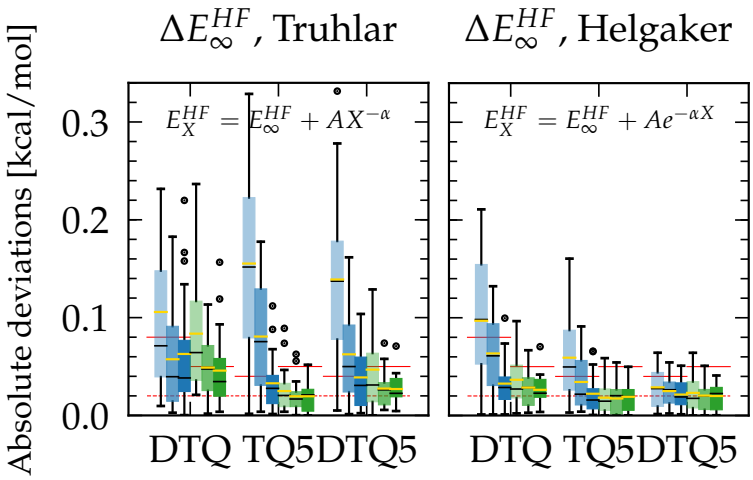


Figure 4.3: Box plots of the differences $|\Delta E_{\mathrm{GTO}}^{\mathrm{HF}} - \Delta E_{\mathrm{PW}}^{\mathrm{HF}}|$ between extrapolated GTO and PW HF interaction energies of the S22* test systems. Medians are shown as horizontal black lines and yellow lines stand for the mean absolute error (MAE). The dashed(solid) red lines correspond to the smallest MAE(maximum deviation) obtained with plain Q and 5 zeta basis sets reported in Table 4.2. The legend is given in Figure 4.2. Signed deviations are also provided in Figure A3.

similar when considering all DTQ5 points. Thus, *Helgaker* (4.28) applied with the CP correction always gives the best match with PW HF energies as summarized in Table 4.3. From this observation, such an exponential expression is the most appropriate for extrapolating HF interaction energies, which are degraded if extrapolated with the scheme of *Truhlar*. Although its usefulness is rather marginal on relative energies that are essentially converged from the Q zeta level (c.f. Table 4.2), the agreement or small improvement over the non-extrapolated results, and the quality of interpolation (Table A3), support the fact that the total (absolute) HF energies can be accurately extrapolated with *Helgaker* (4.28).

With respect to MP2 interaction energies, results show that a majority of GTO extrapolations generally induce larger deviations from PW values than results directly obtained with the highest X point in the fitting sequence. This is particularly the case for CP-corrected energies for which extrapolations are expected to perform well on the reminiscent BSIE effects, and happens for the expressions of *Truhlar* (4.22), *Martin*α (4.25), *Wilson35* (4.26), *Wilson45* (4.27), *Varandas34* (4.30), *Varandas3-fit* (4.31) and *Var.34-fit* (4.32). These schemes can be therefore invalidated outright to accurately estimate GTO energies in the CBS limit and are left in Appendix A for the interested reader's discretion (Figures A5, A6, and A7).

The remaining extrapolations are plotted in Figure 4.4. *Martin4* (4.23) and *Helgaker* (4.29) have the advantage of extrapolating GTO energies from two points only, although for the latter the exponential form of the HF energy requires three parameters, but HF calculations are orders of magnitude cheaper and converge faster than MP2 (as

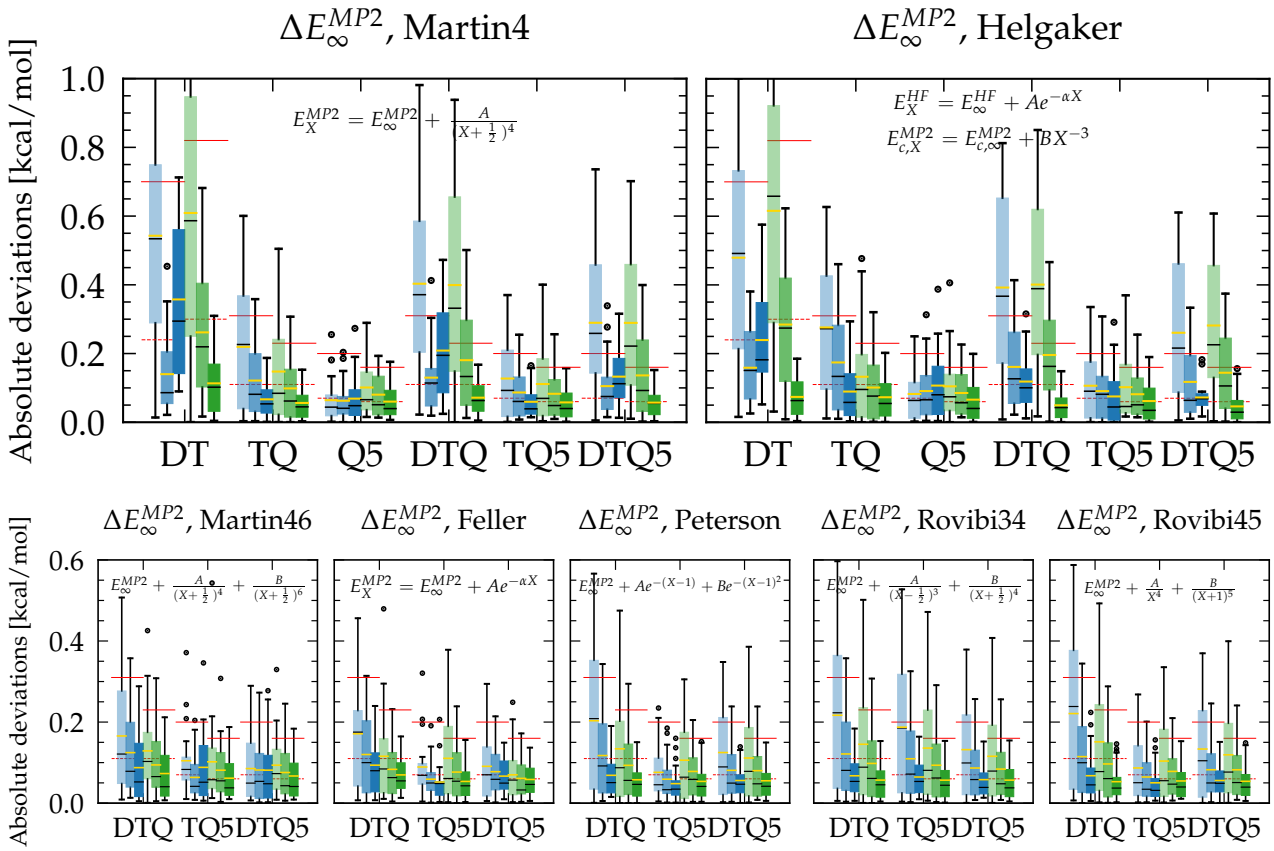


Figure 4.4: Box plots of the differences $|\Delta E_{\mathrm{GTO}}^{\mathrm{MP2}} - \Delta E_{\mathrm{PW}}^{\mathrm{MP2}}|$ between extrapolated GTO and PW MP2 interaction energies of the S22* test systems. Medians are shown as horizontal black lines and yellow lines stand for the mean absolute error (MAE). The dashed(solid) red lines correspond to the smallest MAE(maximum deviation) obtained with plain T, Q or 5 zeta basis sets respectively, as reported in Table 4.2. The legend is given in Figure 4.2. Signed deviations are also provided in Figure A4.

seen previously). Based on DT points, *Helgaker* provides a better agreement with PWs in terms of MAE and maximum deviation, especially for the aug-cc-pVTZ/CP combination with a MAE (max dev.) of 0.3 (0.8) kcal/mol (Table 4.2) that reduces to 0.07 (0.19) kcal/mol when extrapolating with DT data points (Table 4.3). If the available points are TQ instead, the energies obtained by *Martin4* are globally closer to the PW values than those of *Helgaker*, which is also the case for Q5 points albeit in these cases, no extrapolation outperforms the aug-cc-pV5Z/CP calculations and their MAE (max dev.) of 0.06 (0.16) kcal/mol (Table 4.2).

If DTQ points are calculated, *Helgaker* performs better than *Martin4* for both augmented and non-augmented bases coupled with the CP correction, and provides energies that are more converged than if kept at the Q level only. However, when resorting to non-augmented/CP basis sets, the three-parameter schemes *Martin46* (4.24),

Table 4.3: Best agreement between extrapolated GTO and PW interaction energies of the S22* test set for different fitting points. Mean absolute errors (MAE) and maximum deviations are in [kcal/mol]. Worth indicates that the extrapolated energies are statistically closer to the CBS PW values than the corresponding direct results obtained at the highest extrapolation point (c.f. Table 4.2).

Points	Set	Scheme	BSSE	MAE	Max dev.	Worth?
			HF			
DTQ	non-aug	Helgaker	CP	0.03	-0.10	$\simeq$
	aug	Helgaker	CP	0.03	-0.07	$\simeq$
TQ5	non-aug	Helgaker	CP	0.02	-0.07	$\simeq$
	aug	Helgaker	CP	0.02	0.05	$\simeq$
DTQ5	non-aug	Helgaker	CP	0.02	-0.05	$\simeq$
	aug	Helgaker	CP	0.02	0.04	$\simeq$
			MP2			
DT	non-aug	Helgaker	half-CP	0.16	-0.38	✓
	aug	Helgaker	CP	0.07	0.19	✓
TQ	non-aug	Martin4	CP	0.07	0.19	✓
	aug	Martin4	CP	0.06	-0.15	✓
Q5	non-aug	Martin4	half-CP	0.06	-0.20	$\simeq$
	aug	Martin4	CP	0.06	-0.18	$\simeq$
DTQ	non-aug	Peterson	CP	0.07	0.19	✓
	aug	Helg. (Pet.)	CP	0.05	-0.15	✓
TQ5	non-aug	Mart.4 (Pet.)	CP	0.06	-0.16	✓
	aug	Mart.4 (Pet./Fel.)	CP	0.06	-0.16	$\simeq$
DTQ5	non-aug	Helg. (Pet.)	CP	0.07	0.18	$\simeq$
	aug	Helgaker	CP	0.05	-0.16	$\simeq$

Feller (4.19) and *Peterson* (4.20) provide even smaller deviations than *Helgaker*, with *Peterson* surpassing the others. Because *Martin4* and *Helgaker* are based on leading orders at large X (cf. Section 4.3.3), those are likely to deteriorate when considering the smallest cc-pVDZ basis set in the extrapolation sequence.[220] Indeed, *Martin4* and *Helgaker* in general produce slightly lower energies with respect to the PW references, but these deviate more widely from above when D points are considered (Figure A4). Thus, the fact that *Peterson* is the most appropriate for the non-augmented basis with DTQ points appears quite coincidental, and may also result from the lack of diffuse functions that causes additional BSIE not related to the description of the Coulomb cusp, but rather due to long-range effects which cannot be fully corrected for by GTO extrapolations.[212]

Finally, extrapolating from the TQ5 points with *Martin4* and *Peterson* gives similar smallest deviations when the energies are CP-corrected, and the same applies to *Helgaker* and *Peterson* when using all DTQ5 points, but *Helgaker* is performing somewhat better for the augmented sets. Note that the extrapolations that include the (aug-)cc-pV5Z data point do not further improve the interaction energies (Table 4.3), with an average MAE (max dev.) against PWs of 0.06 (0.17) kcal/mol that is comparable to

the 0.07 (0.18) kcal/mol for the non-extrapolated results (Table 4.2). This remaining difference is discussed below in Section 4.5.5.

To summarize, our results demonstrate that the extrapolated GTO energies are always closer to PW reference values when the CP correction is used to tackle the BSSE with sets that are augmented by diffuse functions (Table 4.3). For non-augmented bases, the interplay between the BSIE and some residual BSSE can make the *Helgaker* and *Martin4* two-point extrapolations perform fortuitously better in conjunction with the half-CP correction. The empirical *Peterson* scheme surprisingly provides interaction energies that are very close to the PW results, comparable to *Helgaker* or *Martin4*, but the latter two appear to be more robust candidates because of their theoretical foundations and the fact that they depend on two parameters only. To confirm this, the same analysis has been carried out with omission of four outlier systems (the two uracil, benzene-water and T-shaped indole benzene dimers for which the aug-cc-pV5Z/CP interaction energies are already lower than the CBS PW values). For this smaller test set, *Feller* occasionally beats *Peterson*, but *Helgaker* and *Martin4* perform consistently better, too. Therefore, if the extrapolation sequence includes the D zeta level, it is suggested to use the *Helgaker* (4.28)(4.29) scheme on aug-cc-pVXZ CP-corrected energies in order to obtain the most accurate estimates in the CBS limit. If not, the *Martin4* (4.23) expression is recommended.

4.5.4 Other GTO extrapolations

Motivated by the best agreements found so far, as well as the general expression of eq 4.36, we have tested all possible extrapolations in the form of

$$E_X^{MP2} = E_\infty^{MP2} + A\,(X+a)^{-\alpha} \tag{4.44}$$

and

$$E_X^{MP2} = E_\infty^{MP2} + A\,(X+a)^{-\alpha} + B\,(X+b)^{-\beta} \tag{4.45}$$

with $\alpha, \beta = 3, 4, 5, 6$ and $a, b = -1, -\frac{1}{2}, 0, \frac{1}{2}, 1$, to investigate whether different schemes could universally improve the remaining deviations reported in Table 4.3. For the first eq 4.44, no combination gives overall better results, whether for the augmented or non-augmented bases, reinforcing the recommendation of *Helgaker* (4.29) for DT points only and *Martin4* (4.23) for TQ or Q5 pairs. If three points are available, however, two new expressions stand out as providing very similar or lower deviations than those of *Helgaker* and *Martin4*. We call them from now on *Rovibi34* and *Rovibi45* defined by

$$E_X^{MP2} = E_\infty^{MP2} + A\left(X-\tfrac{1}{2}\right)^{-3} + B\left(X+\tfrac{1}{2}\right)^{-4} \qquad \textit{Rovibi34}\ (4.46)$$

$$E_X^{MP2} = E_\infty^{MP2} + AX^{-4} + B\,(X+1)^{-5} \qquad \textit{Rovibi45}\ (4.47)$$

whose results are also reported in Figure 4.4 and Table 4.4 for comparison.

Table 4.4: Best agreement between *Rovibi* extrapolations and PW interaction energies for the S22* test set. Mean absolute errors (MAE) and maximum deviations are in [kcal/mol]. Worth indicates that the extrapolated energies are statistically closer to the CBS PW values than the corresponding direct results obtained at the highest extrapolation point (c.f. Table 4.2).

Points	Set	Scheme	BSSE corr.	MAE	Max dev.	Worth?
			MP2			
DTQ	non-aug	Rovibi34	CP	0.07	0.18	✓
	aug	Rovibi34	CP	0.06	-0.15	✓
TQ5	non-aug	Rovibi34	CP	0.06	0.16	✓
	aug	Rovibi34	CP	0.06	-0.15	≃
DTQ5	non-aug	Rovibi34	CP	0.06	-0.15	✓
	aug	Rovibi34	CP	0.06	-0.15	≃
DTQ	non-aug	Rovibi45	CP	0.07	0.22	✓
	aug	Rovibi45	CP	0.05	-0.15	✓
TQ5	non-aug	Rovibi45	CP	0.05	-0.16	✓
	aug	Rovibi45	CP	0.05	-0.15	≃
DTQ5	non-aug	Rovibi45	CP	0.06	0.15	✓
	aug	Rovibi45	CP	0.05	-0.15	≃

Such laws indicate that the -3 and -4 orders are indeed good leading candidates, but that terms of higher orders may also be significant. If some rational explanation were to be found, *Rovibi34* (4.46) is compatible with contributions resulting from the principal expansion proposed by Helgaker[14,31] (power -3) and those of the highest angular momentum L present in the basis set from the partial wave expansions put forward by Caroll,[260] Hill[261] and Kutzelnigg.[259] On the other hand, *Rovibi45* (4.47) suggests that an additional order to *Martin4* improves the extrapolation and that the (minus) third order does not dominate. For both, the X-shifts reflect not only the balance between the orders, but also between the basis functions that have $L = X - 1$ for H and $L = X$ for C, N and O atoms in the systems studied here. Based on these considerations, and because the leading order of *Martin4* (-4) was motivated empirically by comparison with experimental atomization energies of small molecules,[219,262,263] the *Rovibi34* scheme seems more formally justified.

Up to this point, only the relative energies extrapolated to the CBS limit have been compared to PW results, but the quality of the GTO extrapolations can also be assessed by how faithfully they reproduce the single data points. Averaged on all systems, fitting curves of *Rovibi34* and *Rovibi45* show a MAE relative to the data points (total energies) of no more than 0.5 kcal/mol (Table A3), while the latter lies between 3.8-7.6 kcal/mol for *Helgaker* and *Martin4*. As a reminder, these errors refer to the total (absolute) energies of the dimers and monomers that have been extrapolated individually. Hence, *Rovibi34* and *Rovibi45* not only provide interaction energies close to the PWs in the CBS limit, but are also capable of interpolating total energies well. In this respect, *Rovibi34* performs best with a MAE of 0.15-0.3 kcal/mol against 0.3-0.5 kcal/mol for

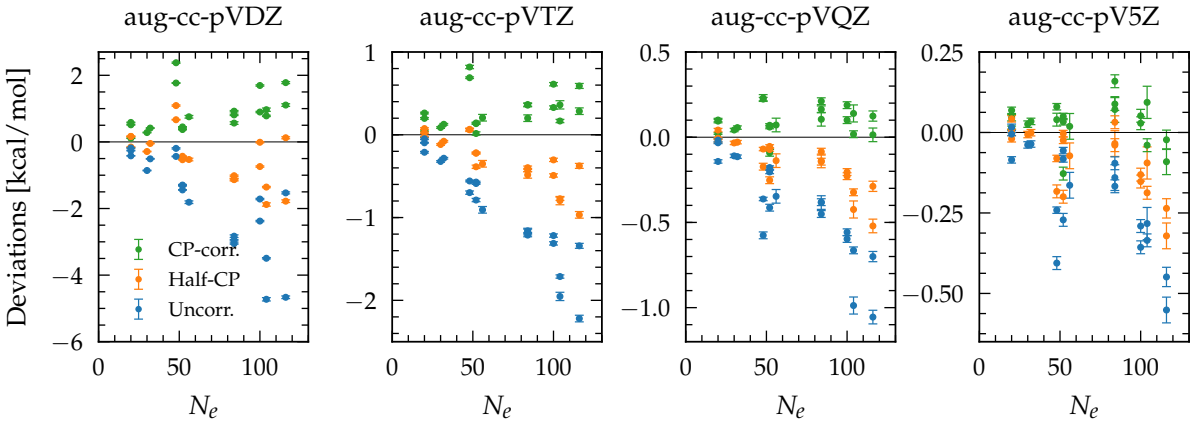
Rovibi45. We also stress that the double exponential scheme of *Peterson*, although purely empirical, shows even smaller fitting MAEs of ∼0.1 kcal/mol (Table A3) and provides deviations against PWs that are similar to those of *Rovibi34* (Figure 4.4). Consequently, our results do not disprove it as a good extrapolation law. However, based not only on agreement with PWs and the ability to interpolate GTO energies, but also on theoretical indications, it appears that *Rovibi34* (4.46) constitutes likely the best global choice when resorting to three/four-point extrapolations (DTQ, TQ5, DTQ5) to the CBS limit.

4.5.5 System size dependency

Finally, whatever the effort in order to match GTOs with PW results, via CP correction, basis set augmentation or extrapolation, one notices that the best MP2 errors do not reach MAEs lower than 0.05 kcal/mol with maximum differences of at least 0.15 kcal/-mol (Tables 4.2, 4.3 and 4.4). While this might at first be taken as an indication that interaction energies are essentially converged at the aug-cc-pV5Z/CP level within an 0.05 kcal/mol numerical accuracy, further analysis instead reveals that residual discrepancies occur because GTO results tend to further deviate from PWs as the system size increases. Figure 4.5a shows indeed that GTO interaction energies that are not corrected for the BSSE are lower than those from PWs, and that the BSSE tends to become larger with an increasing number of electrons in the system. Interestingly, assuming that the CP correction removes the majority of the BSSE, only the BSIE remains and in turn increases with the system size. Hence, although the size of a GTO basis grows with the number of atoms, the incomplete coverage (lack of completeness) of this expansion leads to a smaller and smaller correction of the BSIE with increasing system size. In other words, for a given GTO basis set, its capacity to capture the (correlation) energy decreases as the size of the system increases. Such a size inconsistency is even more pronounced for the smaller cc-pVXZ bases (Figure A9a). In the limit of large bases and systems, however, the occurrence of linear dependencies can further interfere with this behavior.

Once extrapolated, the GTO CBS estimates follow a similar trend with larger (absolute) differences attributed to larger numbers of electrons (N_e) (Figure 4.5b and Figure A9b), thus questioning the agreement between GTOs and PWs in the limit of (very) large systems. Note that such a difference applies to all promising extrapolations found in this work (Figure A10 and A11). The reasons can be multiple, and arise from a combination of the BSSE, the accuracy of the extrapolation scheme, and the intrinsic nature of the basis functions. Nevertheless, as observed earlier (Section 4.5.2), the CP correction seems adequate to eliminate the BSSE so that the last two factors will dominate, which are directly linked to the BSIE. Let us recall that the initial motivation behind the extrapolation schemes is to cope with the electron-electron cusp that hampers the basis set convergence.[14,212] Therefore, although not firmly established

(a) Non-extrapolated



(b) Extrapolated, closest to PWs in the CBS limit

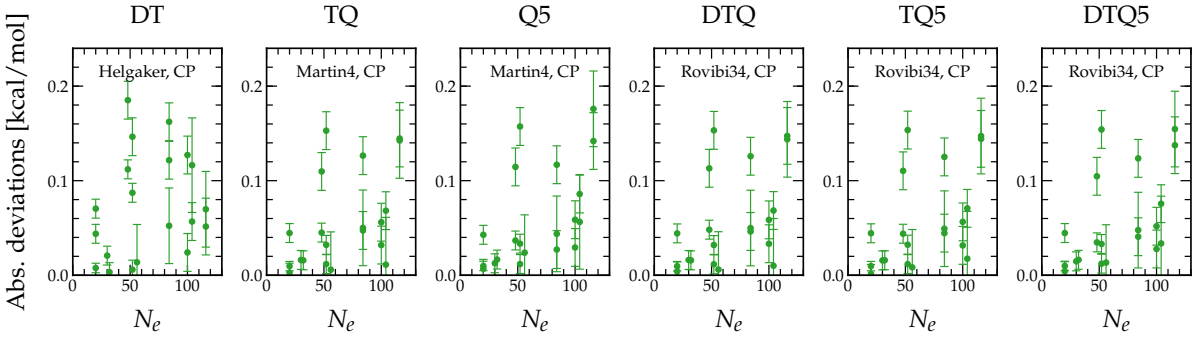


Figure 4.5: Deviations between the aug-cc-pVXZ and PW MP2 interaction energies as a function of the number of electrons N_e in the dimer system. (a) for plain basis set results, (b) for energies extrapolated to the CBS limit with best schemes of Tables 4.3 and 4.4.

by our results, the fact that extrapolated energies deviate from PWs could indicate a residual completeness mismatch between GTOs and PWs in the limit of large systems, originating rather from an in(over)complete description of the wavefunction to capture the long-range (polarization, dispersion) contributions to the correlation energy. This encourages further comparison of atom-centered bases against other basis sets, either explicitly correlated, plane waves or purely numerical, in the calculation of correlation energies of large systems.

4.6 Conclusions and outlook

The main motivation for this work was to analyze the effects of basis sets on correlated energies. To this end, MP2 interaction energies of 20 complexes belonging to the S22 test set were computed, in the most common Gaussian-type correlation-consistent

bases as well as in plane waves, for which we implemented the MP2 method in the CPMD plane-wave pseudopotential package. Although more computationally demanding at such system scales, plane wave calculations have been made accessible on conventional computing architectures through an extrapolation protocol involving both the virtual space orbitals and the supercell volume, ultimately requiring no more than a few days per molecule on several high memory compute nodes.

By comparing atom-centered interaction energies with plane wave results, we established that both basis set types provide consistent values, especially when the CP correction eliminates the BSSE from the former. Indeed, (aug-)cc-pVXZ relative energies generally converge towards plane wave values, free of BSSE, for progressive enlargements of X=D,T,Q,5, and differ by less than 1 kcal/mol for X $\geq$ T. However, the slower convergence of the cc-pVXZ bases makes their agreement with plane waves occasionally better with only half of the CP correction due to a fortuitous error cancellation between the BSSE and their BSIE. Overall, the aug-cc-pV5Z basis set with the CP correction provides the closest interaction energies within 0.16 kcal/mol to the fully converged plane wave results. This demonstrates the benefits of diffuse functions in the description of long-range interactions as occurring in weakly bound systems, although their faster energy convergence may slow down for stronger (covalent) interactions.[251]

Hence, based on the agreement with plane wave results at the CBS limit, theoretical foundations and interpolation capabilities, we can confidently make the following recommendations for the extrapolation of (aug-)cc-pV[D,T,Q,5]Z correlated energy sequences to the CBS limit:

- Use the CP correction for interaction/binding energies.
- Resort to the aug-cc-pVXZ bases if long-range effects are sizable.
- Extrapolate total energies separately according to
  - (D)TQ5 points and $A(X-\frac{1}{2})^{-3}+B(X+\frac{1}{2})^{-4}$
  - DTQ points and $A(X-\frac{1}{2})^{-3}+B(X+\frac{1}{2})^{-4}$
  - Q5 points and $B(X+\frac{1}{2})^{-4}$
  - TQ points and $B(X+\frac{1}{2})^{-4}$
  - DT points and $A\ X^{-3}$

where the choice of the extrapolation points (and basis augmentation) is left to the practitioner since these will depend on the computational budget and the problem/-software at hand. However, note that in principle the higher the point in the sequence, the more accurate the extrapolated value, and the DT scheme should rather be taken as a first rough estimate. In practice, the universal application of such a procedure across correlated wavefunction methods (MP2, CCSD, CCSD(T),...) has been widely accepted.[31,212,217,219,220,223,250,256,263]

Finally, getting the electron correlation in the CBS limit relies on the saturation of the one-electron basis, whose basis functions should adequately span short distances and be flexible enough to incorporate long-range components of the wavefunction. The latter are directly linked to the delocalized nature of the virtual states that contribute to the correlation energy, and therefore necessitate a balanced and complete space coverage. In that sense, plane waves are capable of capturing high-lying (continuum-like) states as well as localized occupied states, at the cost of a sufficiently large cutoff energy.[232] When compared to plane waves, we noticed that the ability of the correlation-consistent bases to cope with the BSIE decreases as the number of electrons increases, therefore questioning the capability of localized basis sets to recover most of the correlation energy as the system size increases. Plane wave, explicitly correlated, or numerical bases calculations on larger systems would confirm (or refute) this statement, but their computational overhead seems to compromise their application for the time being. It is therefore not excluded that, with the improvement of wavefunction-based methods, the precision of correlated energies becomes comparable to the basis set errors; such that the nature of the basis or its extrapolation to the CBS limit ultimately becomes dominant and leads to noticeable deviations that exceed chemical accuracy (1 kcal/mol) for larger molecules.[27]

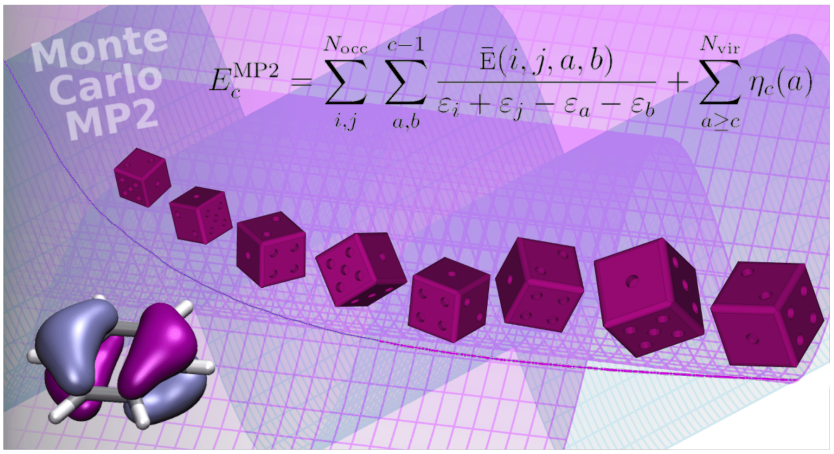
Code and data availability

The development of this work is intended to be available in a future release of CPMD.[47] Data, extrapolation and analysis scripts will be provided on Zenodo at https://doi.org/10.5281/zenodo.7838778.

Appendix

Appendix A provides further implementation details of the PW MP2 correlation energy in CPMD. It also reports information about the energy convergence against the PW wavefunction energy cutoff; energy convergence versus the box volume for different treatments of the electrostatic couplings in the PW periodic setup; PW and (aug-)cc-pV5Z S22* interaction energies; additional deviations between GTO and PW interaction energies against cardinal numbers and the number of electrons in the complexes; and figures of merit for the interpolation quality of various GTO extrapolations.

5 Efficient treatment of correlation energies at the basis-set limit by Monte Carlo summation of continuum states



Chapter 5 is a postprint version of an article published as:

Bircher, M. P.[§]; **Villard, J.**[§]; Rothlisberger, U. Efficient treatment of correlation energies at the basis-set limit by Monte Carlo summation of continuum states. *Journal of Chemical Theory and Computation* 2020, 16, 6550-6559.

[§]M.P.B. and J.V. contributed equally to this work.

Reproduced under the terms of the CC-BY 4.0 License.

5.1 Abstract

The calculation of electron correlation is vital for the description of atomistic phenomena in physics, chemistry and biology. However, accurate wavefunction-based methods exhibit steep scaling and often sluggish convergence with respect to simulation methods and the basis set at hand. Because of their delocalization and ease of extrapolation to the basis-set limit, plane waves would be ideally suited for the calculation of basis-set limit correlation energies. However, the routine use of correlated wavefunction approaches in a plane-wave basis set is hampered by prohibitive scaling due to a large number of virtual continuum states and has not been feasible for all but the smallest systems, even if substantial computational resources are available and methods with comparably beneficial scaling, such as the Møller-Plesset perturbation theory to second order (MP2), are used. Here, we introduce a stochastic sampling of the MP2 integrand based on Monte Carlo summation over continuum orbitals, which allows for speedups of up to a factor of 1000. Given a fixed number of sampling points, the resulting algorithm is dominated by a flat scaling of $\sim\mathcal{O}(N^2)$. Absolute correlation energies are accurate to <0.1 kcal/mol with respect to conventional calculations for several hundreds of electrons. This allows for the calculation of unbiased basis-set limit correlation energies for systems containing hundreds of electrons with unprecedented efficiency gains based on a straightforward treatment of continuum contributions.

5.2 Introduction

Electron correlation lies at the heart of a wide range of fundamental physical and chemical phenomena, which range from the structural diversity and dynamics of water[273] over the dissociation of liquid hydrogen at high pressure[274] and the stability and mobility of point defects in semiconductors[275] to the barrier height of chemical reactions. Wavefunction-based methods allow for a conceptually simple and convenient treatment of electron correlation[13,14,276] and have found widespread and long-lasting use in theoretical chemistry. Correlated wavefunction methods have been widely applied as a benchmarking tool[277,278] in the development of computationally more expedient methods such as Kohn-Sham density functional theory (KS-DFT).[135,140] More recently, their scope has been enlarged by rigorous hybridisation schemes that combine KS-DFT with correlated wavefunction approaches,[56,279–282] giving rise to some of the most accurate density functional approximations available to date.[56,175,280,283–285] In particular, while it has been pointed out that many recently developed density functional approximations fail to yield correct densities and energies at the same time,[286] double hybrid (DH) functionals have been shown to be able to overcome this fundamental problem.[287] Recently, modern machine learning techniques have considerably increased time-scales and system-sizes that can be sampled on conventional infrastructures, but the generation of reliable input data for the training of such methods still

relies on the computational feasibility of reference calculations of sufficient accuracy. In this perspective, the importance and scope of wavefunction-based first-principles techniques applied to condensed matter systems can therefore only be expected to grow further, both as a standalone method and in combination with DFT.

To this day, wavefunction-based correlation methods are hampered by a scaling that is polynomial at best and that is associated with a considerable prefactor. This implies that for larger systems, trade-offs have to be made between the accuracy of the basis set employed and the number of electrons that can be treated with reasonable computational resources. Moreover, correlated wavefunction approaches have only scarcely been applied in the condensed phase, which is due to additional difficulties encountered in periodic systems.[65,117,118,122,272,288–301] These difficulties can be further exacerbated by the large basis sets needed to obtain basis-set limit reference energies.[302] This precludes the routine use of wavefunction-based methods for large condensed phase systems; at the same time, benchmarking possibilities, for example, against newly developed density functionals, remain restricted to comparably few atoms and small supercells[117,118,292,296,300] or have to be based on basis sets which are far from completeness. In benchmarking, this can be particularly problematic in combination with the erroneous convergence behaviour of certain density functionals, which obfuscates any comparison that is not explicitly made at the complete basis-set limit.[303,304] The availability of basis-set limit results is therefore necessary not only for formal reasons, but is also of great importance for the assessment of the physical accuracy of existing models and approximations, representing an important guideline in the development of new techniques and approximations that are able to reach far beyond current system sizes and limitations.

In principle, plane wave (PW) basis sets would constitute an ideal choice for the calculation of basis-set limit correlation energies, since they do not introduce any localisation bias and allow for a controlled, simple and well-defined extrapolation to the complete basis-set limit[66] without the linear dependency issues commonly encountered in large atom-centered bases.[118,119,293,305] In particular, since a single PW is the solution of the Schrödinger equation of a free electron, their use enables the description of continuum states, which have been shown to play a crucial role in a complete description of electron correlation.[306] In the following, we will refer to continuum states in finite systems as those virtual states that resemble a free electron; it is also this resemblance that lies at the heart of simple extrapolation to basis-set limit values.[66] By virtue of their very nature, conventional atom-centered bases such as Gaussian functions or combined Gaussian/PW (GPW)[307] bases are unable to describe such continuum states, which also accounts for the absence of physical models that would allow for a simple extrapolation to the basis-set limit. Instead, they usually rely on specifically constructed basis sets that allow for certain extrapolation models to be applied; this, however, does not commonly hold for density functionals.[192,303] PWs have not been reported to suffer from this drawback.[304] In addition, PWs are

equally suitable for the treatment of both, periodic and non-periodic systems with either wavefunction-based methods, density functional techniques or hybridisations thereof. These advantages, however, come at a price: The presence of a large number of continuum states in PW setups exacerbates the steep scaling of correlated wavefunction methods, making them computationally intractable for all but the smallest systems, which on their own will already require substantial resources on conventional high-performance compute clusters. In the following, we will show that PW-based correlated wavefunction calculations can be sped up by a factor of up to 1000 by stochastically sampling continuum state contributions via Monte-Carlo summation. The error introduced by this stochastic approach remains below 0.01 kcal/mol per electron. This enables correlated wavefunction calculations in PWs for unprecedented system sizes on conventional computational infrastructure, making unbiased basis-set limit values routinely accessible for systems with up to hundreds of electrons. The same reflections hold for hybrid wavefunction/DFT methods, such as DH[279] density functionals.

5.2.1 Møller-Plesset perturbation theory

Among the correlated wavefunction methods, second-order Møller-Plesset perturbation theory (MP2)[21] exhibits a comparably flat scaling of $\mathcal{O}(N^5)$ with number of electrons or basis functions N, making it one of the flattest scaling correlated, wavefunction-based approach available, second only to the random phase approximation (RPA) and the direct RPA (dRPA), respectively.[308,309] In general, MP2 has been found to provide a good first estimate of the dynamic correlation energy.[276,310] Conceptually simple, the MP2 correlation energy E_c^{MP2} is obtained by a perturbative treatment that includes up to doubly excited determinants and summation over pairs of all N_{occ} occupied and N_{vir} virtual orbitals. For the spin-restricted case,[13]

$$E_c^{\mathrm{MP2}} = \sum_{i}^{N_{\mathrm{occ}}} \sum_{j}^{N_{\mathrm{occ}}} \sum_{a}^{N_{\mathrm{vir}}} \sum_{b}^{N_{\mathrm{vir}}} \frac{\bar{\mathrm{E}}(i,j,a,b)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{5.1}$$

where i,j and a,b denote spatial occupied and virtual orbitals ϕ, respectively, that are eigenstates of the Hartree-Fock operator with eigenvalues ε. The MP2 matrix element $\bar{\mathrm{E}}(i,j,a,b)$ is expressed in terms of four-electron Coulomb energies, which can be cast into a positive-definite form

$$\bar{\mathrm{E}}(i,j,a,b) = |\langle ij|ab\rangle|^2 - \langle ij|ab\rangle \langle ba|ij\rangle + |\langle ij|ba\rangle|^2 \tag{5.2}$$

where $\langle ij|ab\rangle$ are two-electron matrix elements. In DH density functionals, the second-order integrals in eq 5.1 are evaluated using a ground-state Kohn-Sham determinant, rather than the Hartree-Fock solution,[279] and will only contribute to a fraction of the total correlation energy, the remainder being treated by pure density-functional

methods. If the advantages of DHs are to be made routinely available in condensed matter applications, they too require an efficient treatment of the terms in eq 5.1, and any improvements in the calculation of the MP2 term will directly benefit DH calculations.

Historically, E_c^{MP2} has been evaluated using atom-centered basis sets or mixed GPW[131] implementations. In particular in periodic setups, use of localised basis functions has been reported to be susceptible to basis-set convergence issues,[118,119,293,305] whereas problematic basis-set superposition effects and possible linear dependencies are absent in a PW representation. Applications in solid state physics have also been scarce, which is in part due to the divergence of the MP2 integrand in zero-band gap systems, but a Thomas-Fermi screening of the MP2 amplitudes can resolve this issue.[311]

The presence of continuum (or continuum-like) states allows for simple extrapolation to the complete basis set limit: Alavi and co-workers have demonstrated that in a PW basis at large N_{vir}, E_c^{MP2} decays as $\propto \varepsilon_a^{-3/2}$.[66] PW-based methods therefore offer the unique advantage of a simple evaluation of basis-set limit values in both periodic and isolated systems, making them a potentially invaluable tool for basis-set bias-free calculations. PW MP2 calculations are few and have only recently been reported.[76,117,118,121] This is in part due to their extensive memory requirements. The presence of a number of virtual states close to the number of PWs themselves (up to 10^9) can further complicate the calculations. However, given sufficient memory, the integrals of eq 5.2 are easily evaluated in reciprocal space: Exchange-like matrix elements are easily obtained by solving the Poisson equation for a set of pair densities $\rho_{ia}(\mathbf{r}) = \phi_i^*(\mathbf{r})\phi_a(\mathbf{r})$ in reciprocal space, where the Coulomb operator is diagonal. The real-space pair densities can simply be subjected to a Fast Fourier Transform (FFT), $\rho_{ia}(\mathbf{G}) = \mathrm{FFT}[\rho_{ia}(\mathbf{r})]$. Then, at the Γ-point, the reciprocal-space equivalent of a matrix element reads[235]

$$\langle ij|ab\rangle = \frac{1}{\Omega}\sum_{\mathbf{G}}^{G_{\max}} \Phi(\mathbf{G})\rho_{ia}(\mathbf{G})\rho_{jb}^{\dagger}(\mathbf{G}) \tag{5.3}$$

where † stands for complex conjugate and index permutation. Ω is the volume of the system, $\mathbf{G}$ are reciprocal-space vectors and $\Phi(\mathbf{G})$ is a suitably generalised form of the reciprocal-space Coulomb potential.[312,313] Attempts to reduce the overall cost of the method have included mapping the virtual states onto a localised subspace,[299,314–318] the use of stochastic orbitals,[75,134,319,320] and (real-space[69,321] or graph-based[67,76]) sampling approaches, respectively, as well as exploiting Laplace transforms[121,133,310,322–324] to enhance parallel efficiency. In an alternative strategy, one seeks to accelerate convergence of the correlation energy by improving the description of the electron-electron cusp. To this end, explicitly correlated basis sets can be used. In an ansatz commonly referred to as F12 theory,[221,325] the description of the electron-electron cusp is improved by combining a conventional, atom-centered Gaussian basis with a set of strongly orthogonal geminals. This approach has recently been extended to PWs.[298]

However, localisation procedures fail for continuum states; previously reported stochastic techniques either introduce system-dependent errors of up to 2 kcal/mol per occupied orbital[75] or carry a prefactor too large for practical applications,[69] and approaches based on the Laplace-transform[121] require repeated calculations at different quadrature points, increasing the overall operation count for the sake of parallel efficiency, with modest reported speedups of around 4 to 5. A combined stochastic/Laplace-transform approach that results in appreciable computational speedups leads to errors of at least 20%.[76] Similarly, due to the presence of non-factorizable many-electron integrals, the cost of evaluating the MP2 integrand in PW-F12 is higher than for a pure PW basis set,[298] and the efficiency of graph-based approach[67] was hindered by the absence of an optimized weighting scheme for graph generation. In a more general scope, a recent diagrammatic decomposition of the coupled cluster pair correlation function has allowed for the introduction of a basis-set correction in PWs that results in speedups of 2 orders of magnitude.[326] Alternatively, in the context of Full Configuration Interaction Quantum Monte Carlo calculations,[68] use of an effective, transcorrelated Hamiltonian[77] has been shown to substantially improve convergence of the correlation energy of the uniform electron gas. In the GW theory,[327] stochastic sampling schemes have successfully been applied with competitive accuracy and favourable timings with respect to deterministic calculations.[328]

In the following, we shall demonstrate how the presence of continuum states can be exploited to drastically reduce the computational cost of MP2 calculations without impacting their accuracy. The approach is based on a simple stochastic sampling of the integrands of eq 5.1 and can be implemented with little effort, representing a sleek and clean approach to tackle the issues arising in the continuum. This opens the path to routine applications of PW MP2-based approaches in both isolated and periodic systems with up to hundreds of occupied orbitals, making it possible to obtain basis-set limit DH or MP2 correlation energies on conventional computational infrastructure within a reasonable time.

5.3 Distribution of continuum states and stochastic sampling

The possibility of introducing stochastic sampling is rooted in the behaviour of the integrand at large ε_a, where continuum states arise. In this regime, E_c^{MP2} grows as $\varepsilon_a^{-3/2}$,[66] and the overlap elements $\langle ij|ab\rangle$ must therefore be of low magnitude. It is obvious from eq 5.3 that overlaps will only be non-negligible whenever the symmetries of the continuum states match, but explicit symmetry determination for all states would be prohibitively expensive. Instead, given two high-lying virtual states, the statistical distribution of non-negligible overlaps is expected to be similar between truncations of eq 5.1 at a and some subsequent $a+\delta$ with arbitrary δ. For a spatially infinite supercell at infinite PW cutoff, one can define a cutoff energy ε_c with orbital index c, from where on all orbitals $a \geq c$ are part of the continuum. Then, the correlation energy can be

separated

$$E_c^{\mathrm{MP2}} = \sum_{i,j}^{N_{\mathrm{occ}}} \sum_{a,b}^{c-1} \frac{\bar{\mathrm{E}}(i,j,a,b)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} + \sum_{a\geq c}^{N_{\mathrm{vir}}} \eta_c(a) \tag{5.4}$$

with the continuum contribution $\eta_c(a)$ accounting for the incremental change in correlation energy when adding an additional continuum orbital to the system:

$$\eta_c(a) := \sum_{i,j}^{N_{\mathrm{occ}}} \left[\frac{\bar{\mathrm{E}}(i,j,a,a)}{\varepsilon_i + \varepsilon_j - 2\varepsilon_a} + 2\sum_{b\geq c}^{a-1} \frac{\bar{\mathrm{E}}(i,j,a,b)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} + 2\sum_{b'=1}^{c-1} \frac{\bar{\mathrm{E}}(i,j,a,b')}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_{b'}} \right] \tag{5.5}$$

where $\sum_{b'}$ is due to pairs of continuum (a) and noncontinuum (b') virtual orbitals and we have used that, at the Γ-point, $\phi_a = \phi_a^*$. Note that the last term in eq 5.4 contains only one explicit sum over virtual orbitals N_{vir}, with the orbital pairs themselves being formed by the triple sum in eq 5.5.

The simplest possible stochastic treatment of eq 5.5 is given by a uniform sampling of the summand, but this calls for a regular distribution of the overlap values obtained over all tuples *ijab*, which are the arguments of $\eta_c(a)$; that is, the high-lying virtual orbitals of a finite system at finite PW cutoff need to reasonably approximate free, continuum electrons. Figure 5.1 shows the distribution of these arguments for a finite periodic box with $\varepsilon_c - \varepsilon_{\mathrm{HOMO}} = 50$ eV and a total of $N_{\mathrm{vir}} = 3000$ virtual orbitals. With the positive-definite definition of $\bar{\mathrm{E}}$ adopted in eq 5.2, the resulting distribution is indeed regular and smooth, indicating that $\eta_c(a)$ could be predestinated to be treated by uniform Monte-Carlo sampling. This is the approach we will privilege in the following. Note that in the limit of an infinite system, this is equal to Monte Carlo integration over the continuum; such integration techniques have been used as early as 1957 to calculate the high-density correlation energy of the uniform electron gas.[329]

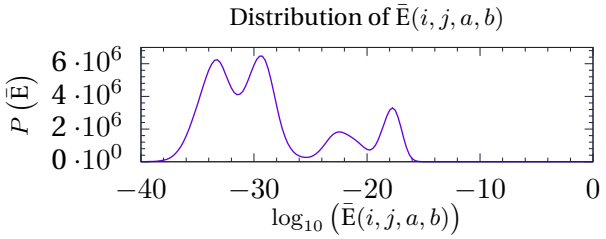


Figure 5.1: Figure showing the absolute occurrence P of orders of magnitude of the summand $\bar{\mathrm{E}}(i,j,a,b)$ of $\eta_c(a)$ of eq 5.5 from $a = c$ to $a = 3000$ for $\varepsilon_c - \varepsilon_{\mathrm{HOMO}} = 50$ eV. The order of magnitude of the matrix elements is small, and its distribution is smooth.

Let $p_s \in [0,1)$ be a predefined sampling probability (i.e., p_s is larger than or equal to zero but always smaller than 1). We define the function p as

$$p(x,p_s) = \begin{cases} 1 & \text{if} \quad x \leq p_s \\ 0 & \text{if} \quad x > p_s \end{cases} \tag{5.6}$$

Symmetry at the Γ-point, which gives rise to a factor of 2 in continuum/continuum and mixed continuum/noncontinuum terms of eq 5.5, can be accounted for by introducing a Kronecker delta. Then, the sums in $\eta_c(a)$ can be simplified, and the stochastic expression for $\eta_c(a)$ is

$$\langle \eta_c(a) \rangle := \frac{1}{p_s} \sum_{i,j}^{N_{\text{occ}}} \left[\sum_{b=1}^{a} (2-\delta_{ab})\, p(x_{ijab},p_s) \frac{\bar{\text{E}}(i,j,a,b)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \right] + \mathcal{O}\left(\frac{1}{\sqrt{N_{\text{MC}}}}\right) \tag{5.7}$$

where $x_{ijab} \sim U([0,1))$ is a sequence of random numbers drawn from a uniform distribution U and b covers both b and b' of eq 5.5. In the limit of a continuum, the error of the sampling is expected to decrease as $1/\sqrt{N_{\text{MC}}}$, with N_{MC} being the number of tuples *ijab* that are explicitly sampled. The combination of eqs 5.4 and 5.7 amounts to a stochastic summation over all orbital pairs that contribute to an increase in E_c^{MP2}, once a continuum orbital a is added to a system already containing $a-1$ virtual orbitals.

For any sampling probability $p_s < 1$, only elements of $\eta_c(a)$ with $p(x_{ijab},p_s) = 1$ have to be evaluated. In a finite cell with a finite PW cutoff and a sufficiently large number of virtual orbitals N_{vir}, estimates for $\eta_c(a)$ are then obtained for every given continuum orbital $a \in [c, N_{\text{vir}}]$ as follows: rather than randomly generating a tuple of indeces for a predetermined amount of Monte Carlo moves, a random number $p(x_{ijab},p_s)$ is drawn for every element of the summand in eq 5.7, determining whether a particular tuple *ijab* will enter into the estimator of $\eta_c(a)$. p_s therefore defines a target value $N_{\text{MC}} \propto p_s$ of tuples that are expected to be sampled for every continuum orbital a. The advantages of such an algorithm are two-fold: On one hand, it avoids under- or oversampling of the subspace associated with orbital a; on the other hand, it enables efficient extrapolation to the basis-set limit in one single calculation, directly providing E_c^{MP2} as a function of the highest virtual orbital included in eq 5.4.

In the following, we will adopt an orbital-dependent sampling probability $p_s(a) = N_{\text{MC}}/N_{\text{card}}(a)$, where $N_{\text{card}}(a) = (2a-1)N_{\text{occ}}(N_{\text{occ}}+1)/2$ is the product of the cardinalities of the sums in eq 5.7. $N_{\text{card}}(a)$ therefore explicitly depends on the virtual orbital a that is added to eq 5.4. With this choice of a continuum-orbital-dependent $p_s(a)$, the density of the sampling decreases as a increases. This allows for N_{MC} to remain a fixed, system-independent input quantity. We shall later show that conservative estimates for N_{MC} and ε_c can be regarded as system-independent.

For orbitals that are part of the continuum, the resulting algorithm scales formally as

$\mathcal{O}(N_{\text{PW}}N_{\text{vir}}N_{\text{MC}})$, where N_{PW} is the number of PWs in the basis set. This follows from eqs 5.4 and 5.7, since the cost of evaluation of the triple sum over $N_{\text{occ}}^2 N_{\text{vir}}$ in eq 5.7 is reduced to elements with $p(x_{ijab}, p_s(a)) = 1$, which in turn is proportionate to N_{MC}. For virtual orbitals with $\varepsilon_a < \varepsilon_c$, the conventional $\mathcal{O}(N_{\text{PW}}N_{\text{occ}}^2 N_{\text{vir}}^2)$ scaling applies (cf. eqs 5.1 and 5.3). Further on in the text, we will show that in practice, the number of terms due to eigenvalues $\varepsilon_a < \varepsilon_c$ does not dominate scaling. Once N_{MC} can be made both independent of the orbital index a and the system at hand, the scaling of the resulting method reduces to $\mathcal{O}(N_{\text{PW}}N_{\text{vir}})$ integral evaluations for all orbitals with eigenvalue $\varepsilon > \varepsilon_c$.

5.4 Computational methods

5.4.1 General setup

Hard pseudopotentials of the Goedecker, Teter and Hutter (GTH) form[330] parametrised for Hartree-Fock calculations[296] have been used for all calculations. PW MP2 and MPs2 energies were calculated using a modified version of the CPMD code.[47] The convergence threshold on the residual gradient on occupied orbitals was set to 10^{-7} a.u., whereas a threshold of 10^{-5} a.u. was used for the virtual space. For isolated systems, periodic images were decoupled using the Poisson solver by Martyna and Tuckerman.[197] The wavefunction cutoff energies E_{cut}^{ϕ} were set to 150 Ry for all systems but the ethylene crystal, where a value of 140 Ry was used. A density cutoff $E_{\text{cut}}^{\rho} = 4E_{\text{cut}}^{\phi}$ was adopted for all systems, while cutoff energies for MP2 pair densities were set to E_{cut}^{ϕ} without impacting accuracy. The calculation of the MP2 term is based on straightforward evaluation of eq 5.3 as in ref [235]. Since no derivatives with respect to E_c^{MP2} have to be evaluated, reciprocal-space orbital pairs $\rho_{ia}(\mathbf{G})$ are stored in memory, allowing for substantial speedups with respect to an on-the-fly evaluation of every pair density. The size of the periodic super- or unit cells, the number of electronic states as well as molecular geometries and PW energy cutoffs are given in Appendix B.

5.4.2 Extrapolation of correlation energies

Correlation energies are obtained from single-point extrapolation using the $\varepsilon_{N_{vir}}^{-3/2}$ dependency (equivalent to $1/N_{vir}$) as described in ref [66]. Such a leading-order approximation requires a sufficient number of high-energy orbitals in order to gather sufficient statistics and reliable extrapolated values. In this work, we adopt a fitting scheme with various windows (ranges of points to fit) moving along the curve. Those vary in size with respect to the number of orbitals included, with a shift of $\sim$200 orbitals between them. The maximum window size is taken as the one that retains a fitting error comparable to smaller windows when finishing fitted ranges at $N_{\text{vir}}^{\text{max}}$, the highest orbital available from a Davidson diagonalization. The smallest window is given by

the smallest possible window size that provides a stable fit. MP(s)2 curves are fitted according to

$$E_c^{\mathrm{MP(s)2}}(N_{\mathrm{vir}}) \cdot \varepsilon_{N_{\mathrm{vir}}}^{3/2} = \alpha \cdot \varepsilon_{N_{\mathrm{vir}}}^{3/2} + \beta \tag{5.8}$$

The series of $\alpha(N_{\mathrm{vir}})$, obtained from all windows along the curve that terminate at N_{vir} up to $N_{\mathrm{vir}}^{\mathrm{max}}$, converge to an estimate of the MP(s)2 energies at the basis-set limit, $E_c^{\mathrm{MP(s)2}}(\varepsilon_{N_{\mathrm{vir}}} \to \infty)$ Averaging over different windows allows to account for sensitivity and variance of extrapolated values, which are given in Appendix B along with figures that illustrate the extrapolation procedure and error determination.

5.5 Results and discussion

5.5.1 Accuracy of stochastic summation

Our stochastic sampling scheme was tested both on isolated systems as well as in periodic, condensed-phase setups. Test systems in the condensed phase include solid-state ethylene and benzene molecular crystals as well as a liquid consisting of a hydronium ion solvated in 32 water molecules. Isolated (gas-phase) systems are represented by a benzene monomer and dimer in sandwich configuration. All calculations were carried out with a modified version of the CPMD code[331] using GTH pseudopotentials generated for Hartree-Fock calculations.[296,330]

We first investigate the dependency of the accuracy of our stochastic sampling of eqs 5.4 and 5.7, called MPs2, on the number of orbitals in the occupied subspace and the choice of the continuum cutoff value $\varepsilon_c^{\mathrm{gap}}$. This quantity is given with respect to the highest occupied molecular orbital (HOMO) in order to be independent of the system setup or reference frame: $\varepsilon_c^{\mathrm{gap}} = \varepsilon_c - \varepsilon_{\mathrm{HOMO}}$. Figure 5.2 shows the absolute difference between MP2 and MPs2 calculations for a benzene crystal and a benzene monomer for $\varepsilon_c^{\mathrm{gap}}$ ranging from 20 to 180 eV. Differences rapidly decrease by increasing $\varepsilon_c^{\mathrm{gap}}$. Errors averaged over independent stochastic runs remain $<$0.01% for continuum cutoffs $\geq$120

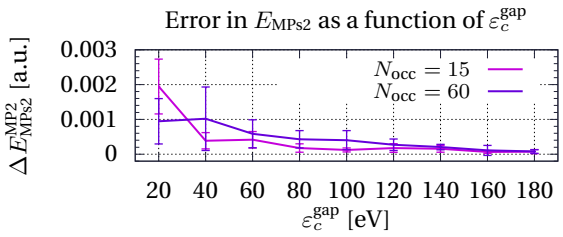


Figure 5.2: Absolute differences between stochastic and nonstochastic MP2 correlation energies, $\Delta E_{\mathrm{MPs2}}^{\mathrm{MP2}}$, as a function of the continuum cutoff energy $\varepsilon_c^{\mathrm{gap}}$ for a benzene crystal ($N_{\mathrm{occ}} = 60$, four molecules per unit cell) and a benzene monomer ($N_{\mathrm{occ}} = 15$) at $N_{\mathrm{MC}} = 12000$. Error bars were obtained by averaging over six independent runs.

eV. Reducing this value by half to 60 eV results in doubling of the relative error, which can still be acceptable. Further lowering of $\varepsilon_c^{\text{gap}}$, however, leads to rapidly increasing errors. The standard deviation of the sampling error depends much more strongly on $\varepsilon_c^{\text{gap}}$ than on N_{occ}. From $\varepsilon_c^{\text{gap}} = 120$ eV on, standard deviations become negligible and virtually identical for both systems. Notably, relative sampling errors are lower for the crystal at $N_{\text{occ}} = 60$ than for the monomer (graphs are provided in Appendix B). This strongly supports that accuracy is mainly influenced by $\varepsilon_c^{\text{gap}}$, whereas at constant N_{MC}, the error is independent of the partitioning of occupied and virtual states in eq 5.7.

In the following, and in line with the values of Figure 5.2, we will adopt $N_{\text{MC}} = 12000$ and $\varepsilon_c^{\text{gap}} = 120$ eV to investigate the accuracy of stochastic sampling for different systems. Figure 5.3 shows the correlation energy as a function of the eigenvalue of the highest virtual orbital included ($\varepsilon_{N_{\text{vir}}}$) for two exemplary systems: one periodic (ethylene crystal) and one isolated (benzene monomer) setup, calculated both with a full summation according to eq 5.1, as well as with our stochastic sampling. The corresponding extrapolated basis-set limit values are shown in Table 5.1. The largest errors of the extrapolated, absolute MP2 correlation energies lie between 0.02 and 0.1 kcal/mol. Errors in the binding energy of the benzene sandwich are of comparable magnitude, which is far below chemical accuracy. Differences per electron, $\Delta E/e^-$, do not exceed 0.1 meV. These values compare well to an expected stochastic error of $\leq$0.01%. Table 5.1 also shows the number of matrix element evaluations for all setups. With our system-independent choice of N_{MC}, stochastic sampling reduces this number by 1-2 orders of magnitude compared to conventional calculations.

For both, absolute energies and energy differences, the observed deviations are several orders of magnitude lower than the reported error of other stochastic or Laplace-transform based schemes.[67,69,75,76,134,319–321] For comparison, values of correlation energies obtained with atom-centered all-electron[332] and GPW codes[333] are listed in

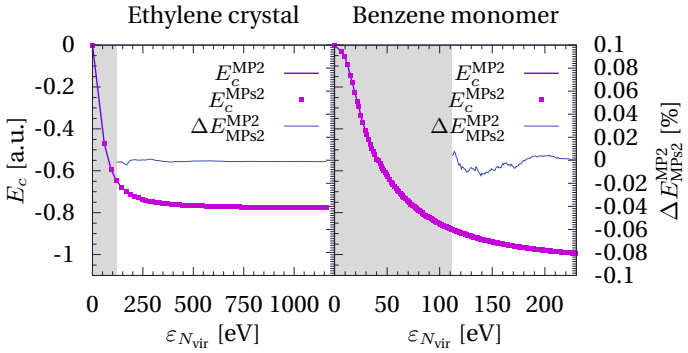


Figure 5.3: E_c^{MP2} as a function of the highest eigenvalue $\varepsilon_{N_{\text{vir}}}$ in the sum of eq 5.1 for both conventional and stochastic MP2 (MPs2) for an ethylene crystal ($N_{\text{vir}}^{\text{max}} = 11158$) and an isolated benzene monomer ($N_{\text{vir}}^{\text{max}} = 14985$). Domains without stochastic sampling are coloured in gray. Differences between the curves, $\Delta E_{\text{MPs2}}^{\text{MP2}}$, are plotted on a secondary y-axis. Extrapolated basis-set limit values for all systems described here are found in Table 5.1.

Table 5.1: MP2 correlation energies E_c^{MP2} obtained from a conventional MP2 calculation and the stochastic approach MPs2. $\Delta E_{\text{MPs2}}^{\text{MP2}}$ and $\Delta E/e^-$ denote absolute and per-electron energy differences between stochastic sampling and conventional calculations, respectively. N_{ijab} denotes the number of matrix elements (*ijab*-tuples) sampled in a conventional calculation, and $p_s N_{\text{ijab}}$ is the number of effectively sampled matrix elements in a stochastic MPs2 calculation, with both numbers rounded to three digits. The threshold for stochastic sampling expressed with respect to the HOMO was identical for all systems, $\varepsilon_c^{\text{gap}} = \varepsilon_c - \varepsilon_{\text{HOMO}} = 120$ eV. The same holds for $N_{\text{MC}} = 12000$, the number of terms sampled per virtual contribution $\eta_c(a)$, as described in eq 5.7. For details on the systems used, cf. Appendix B.

System		E_c^{MP2} [a.u.]	E_c^{MPs2} [a.u.]	$\Delta E_{\text{MPs2}}^{\text{MP2}}$ [a.u.]	$\Delta E/e^-$ [a.u.]	N_{ijab}	$p_s N_{\text{ijab}}$
Ethylene	crystal	-0.78054	-0.78056	$2 \cdot 10^{-5}$	$8 \cdot 10^{-7}$	$5.06 \cdot 10^{9}$	$3.26 \cdot 10^{8}$
Benzene	crystal	-4.69164	-4.69149	$-1.5 \cdot 10^{-4}$	$-1 \cdot 10^{-6}$	$1.30 \cdot 10^{11}$	$3.50 \cdot 10^{9}$
	monomer	-1.05681	-1.05695	$1.4 \cdot 10^{-4}$	$5 \cdot 10^{-6}$	$8.62 \cdot 10^{9}$	$6.26 \cdot 10^{8}$
	dimer	-2.12780	-2.12777	$-3 \cdot 10^{-5}$	$-5 \cdot 10^{-7}$	$3.33 \cdot 10^{10}$	$8.78 \cdot 10^{9}$
	binding	-0.01417	-0.01387	$-3 \cdot 10^{-4}$	$-5 \cdot 10^{-6}$		

Appendix B; all values are consistently higher than those reported for our PW calculations. In particular, we note that differences between different basis sets tend to be substantially larger than the stochastic sampling error, further confirming the viability of uniform, stochastic sampling of the continuum space.

5.5.2 Performance and speedups

With the error of the stochastic sampling scheme being considerably lower than the errors documented for other PW implementations, effective speedups remain to be determined. Figure 5.4 shows the resulting cumulative execution times and speedups of the stochastic sampling compared to the direct implementation of eqs 5.1 and 5.3 for the ethylene crystal. Timings are reported as a function of the highest orbital index included in the expansion, N_{vir}. For the ethylene crystal, at $N_{\text{vir}} = 10000$, speedups of up to 957 can be reached with stochastic summation, making the calculation about 3 orders of magnitude faster compared to a conventional implementation. All the while, the error introduced by uniform stochastic summation with respect to a full calculation is only about 10^{-2} kcal/mol for this system (cf. Table 5.1). This has to be compared to maximum speedups of about 5 documented for Laplace transform-based schemes that allow for similar accuracy to be retained[121] and 20% errors in correlation energies for algorithms that allow for larger speedups[76] versus errors around 0.01% reported here. Figure 5.4 also includes a fit of the CPU time of our stochastic scheme to a function $O(N_{\text{vir}}) = a_0 + a_1 N_{\text{vir}}$, demonstrating that the scaling is described well by $\mathcal{O}(N_{\text{PW}} N_{\text{vir}})$, assuming that $N_{\text{MC}} = \text{cnst}$. $\mathcal{O}(N^2)$ constitutes a formal improvement over the scaling achieved with a stochastic graph-based approach in atom-centered basis sets,[67] which was reported to be of $\mathcal{O}(N^{2.6})$.

Together with the high accuracy of the method, this considerable gain in efficiency

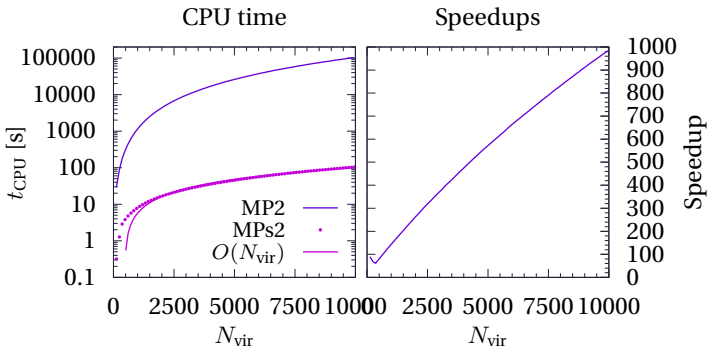


Figure 5.4: Cumulative execution times (left) and speedups (right) compared between conventional algorithm and stochastic sampling for an ethylene crystal as a function of the highest virtual orbital index N_{vir} included in eq 5.4. $O(N_{vir}) = a_0 + a_1 N_{vir}$ denotes a least-squares fit on $N_{vir} \in [5000, 10000]$ with $a_0 = -4.76586$ s, $a_1 = 0.010515$ s. For sufficiently large $N_{vir} > 2000$, formal and practical scaling show excellent agreement. At $N_{vir} = 10000$, the stochastic approach with $N_{MC} = 12000$ is 3 orders of magnitude faster. Timings were obtained by dividing the computational load over 5 OMP (Open Multi-Processing) threads and 24 MPI (Message Passing Interface) tasks.

allows for the treatment of systems that would be intractable when treated with conventional algorithms. One typical usage example of accurate wavefunction-based theories or computationally demanding high-quality DFT methods such as DH functionals lies in postprocessing of simulation data, for example, in *a posteriori* calculations of potential energy surfaces or reaction paths generated using lower-level methods in the context of molecular dynamics or Monte Carlo simulations. Recent developments have even made it possible to directly sample[334] high-quality potential energy surfaces by virtue of multiple time step (MTS) propagators,[335] which allow for an important decrease in computational cost and substantial improvements in tractability by permitting less-frequent evaluation of the full, high-quality Hamiltonian during a first-principles molecular dynamics run.

A liquid constituted of 1 hydronium ion solvated in 32 water molecules (264 electrons) will serve as an example of the performance of stochastic sampling in a typical setup encountered in first-principles (MTS-)molecular dynamics. Using the stochastic summation scheme reported here, calculating the basis-set limit correlation energy of this system, shown in Figure 5.5, is feasible in about 15 h on 25 16-core compute nodes with 128 GB of RAM, using the same, conservative estimates for $\varepsilon_c^{\text{gap}}$ reported in Table 5.1, which yielded accurate results for all test systems considered so far. Additionally, calculations using $\varepsilon_c^{\text{gap}} = 60$ eV and $\varepsilon_c^{\text{gap}} = 90$ eV have been carried out for the sake of comparison. Already at $\varepsilon_c^{\text{gap}} = 90$ eV, the execution time is almost halved to about 8 h, and can be further reduced by using $\varepsilon_c^{\text{gap}} = 60$ eV, at which the basis-set limit correlation energy can be calculated in a mere 5 h. In particular, this drastic reduction in execution time is not accompanied by a considerable loss of accuracy. Extrapolated correlation energies are given in Table 5.2. It should be noted that postprocessing protocols can be applied when training machine learning algorithms with high-quality

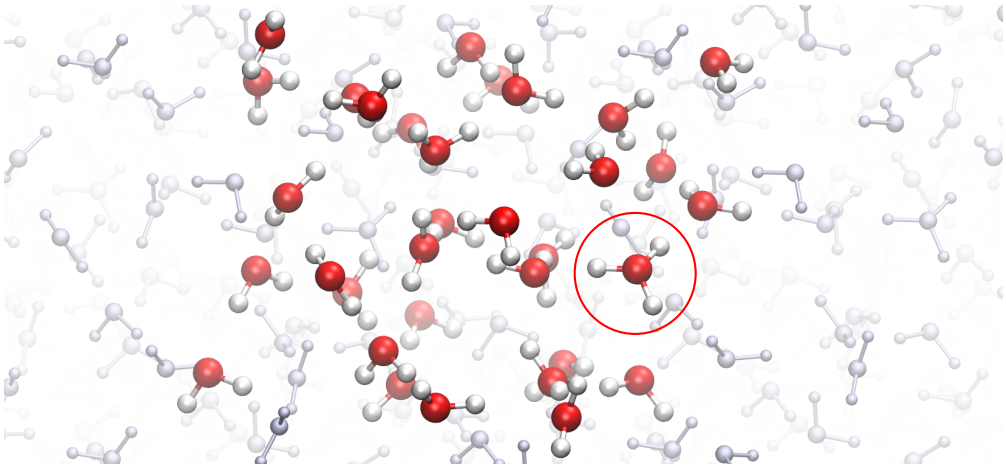


Figure 5.5: Hydronium ion solvated in 32 water molecules (264 electrons explicitly accounted for). Calculation of E_c^{MPs2} takes between 5 and 15 h on 25 16-core compute nodes. Molecules within the periodically repeated supercell are highlighted. Using $\varepsilon_c^{\mathrm{gap}} = 120$ eV and $N_{\mathrm{MC}} = 12000$ yields an extrapolated $E_c^{\mathrm{MPs2}} = -10.22952$ a.u.

Table 5.2: Stochastic MP2 correlation energies E_c^{MPs2} for a hydronium ion solvated in 32 water molecules using different thresholds for continuum energies $\varepsilon_c^{\mathrm{gap}}$. Timings t, rounded to 5 min, are given for the execution time of the MP2 routine on 25 16-core nodes in a hybrid setup (50 MPI tasks, 8 OMP threads).

E_c^{MPs2} [a.u.]	N_{MC}	$\varepsilon_c^{\mathrm{gap}}$ [eV]	t
-10.22771	12000	60	4 h 55
-10.23190	12000	90	8 h 10
-10.22952	12000	120	15 h 25

data, based on a coarse sampling of configuration space with a lower-level method (cf., e.g. ref [273]). Speedups in the calculation of correlation energies with MP2 or DH functionals will therefore be directly reflected in less time-consuming training procedures, thus considerably increasing throughput.

5.5.3 Generalization to the random phase approximation

The promising performance of the stochastic summation scheme described here also opens the possibility of its application to similar approaches in which continuum states can play a role. In DFT, the exact exchange plus RPA (EXX+RPA) approach has emerged as a promising method capable of more accurately predicting van der Waals binding energies, adsorption energies on surfaces, or lattice constants in molecular solids.[66,131,177–179] Based on the similarity of the MP2 energy expression and the RPA, one can expect transferability of the stochastic sampling approach to the evaluation of

the RPA. The reciprocal-space form of the RPA correlation energy E_c^{RPA} is[308,309]

$$E_c^{\mathrm{RPA}} = \int_0^\infty \frac{\mathrm{d}\omega}{2\pi} \frac{1}{N_{\mathbf{q}}} \sum_{\mathbf{q}\in \mathrm{BZ}} \mathrm{Tr}\{\ln\left[1 - \widetilde{\chi}_{\mathbf{G},\mathbf{G}'}(\mathbf{q}, i\omega)\right] + \widetilde{\chi}_{\mathbf{G},\mathbf{G}'}(\mathbf{q}, i\omega)\} \tag{5.9}$$

where $\widetilde{\chi}_{\mathbf{G},\mathbf{G}'}(\mathbf{q}, i\omega)$ are elements of the full density response function, including the Coulomb interaction. At the Γ-point with $\mathbf{q} = 0$, the diagonal elements in a stochastically sampled RPA scheme become

$$\widetilde{\chi}_{\mathbf{G},\mathbf{G}}(0, i\omega) = \frac{1}{\Omega} \sum_{j}^{N_{\mathrm{occ}}} \left[\sum_{a}^{c-1} \chi_{\mathbf{G},\mathbf{G}}(i\omega, j, a) + \frac{1}{p_s} \sum_{b\geq c}^{N_{\mathrm{vir}}} p(x_{jb}, p_s) \chi_{\mathbf{G},\mathbf{G}}(i\omega, j, b) \right] \tag{5.10}$$

with

$$\chi_{\mathbf{G},\mathbf{G}}(i\omega, j, a) = \frac{\Phi(\mathbf{G})\rho_{ja}(\mathbf{G})\rho^*_{ja}(\mathbf{G})}{i\omega + \varepsilon_j - \varepsilon_a} - \frac{\Phi(\mathbf{G})\rho_{aj}(\mathbf{G})\rho^*_{aj}(\mathbf{G})}{i\omega + \varepsilon_a - \varepsilon_j} \tag{5.11}$$

where the variables c, x_{jb} and p_s are used analogously to eq 5.7. In this limit, we investigate the RPA integrand of eq 5.9 for an ethylene crystal at different values of ω. Figure 5.6 shows the values of E_c^{RPA} at varying $i\omega$ for $N_{\mathrm{vir}} = 11040$, as well as an estimate of the overall RPA correlation energy based on trapezoidal integration. Stochastic sampling with $p_s = 1/3$ introduces a maximum error of about 1% in the integrand. Overall, the stochastic sampling introduces a final error of less than 0.03 kcal/mol, which is comparable to the error obtained in MP2 calculations for the same system. These results demonstrate that the stochastic sampling of continuum states can also be applied for methods other than MP2 that include a substantial continuum-state contribution.

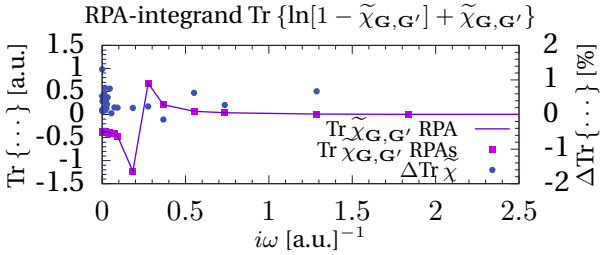


Figure 5.6: Value of the integrand of eq 5.9, abbreviated as Tr $\{\cdots\}$, for an ethylene crystal at $\mathbf{q} = 0$ and $N_{\mathrm{vir}} = 11040$ as a function of $i\omega$ for both conventional and stochastic calculations. The difference between the curves, ΔTr $\{\cdots\}$, rapidly decreases as $i\omega$ is increased; differences are plotted on the secondary y-axis and do not exceed 1.5%. Simple trapezoidal integration leads to $E_c^{\mathrm{RPA}} = -0.03365$ a.u. for a full calculation and $E_c^{\mathrm{RPAs}} = -0.03360$ a.u. for stochastic sampling.

5.6 Conclusions and outlook

Continuum states have been shown to be an important contributor to the overall electron correlation energy.[306] Among the basis sets commonly used in solid-state physics, quantum chemical calculations, and first-principles molecular dynamics, PWs stand out as the only choice that can effectively account for continuum contributions, which are also the base of a simple basis-set limit extrapolation technique.[66] Here, we have introduced a uniform stochastic sampling approach to treat continuum states arising in correlation energy calculations, where contributions due to states with orbital eigenvalues beyond some threshold ε_c are added by stochastic summation. This algorithm has been applied to the calculation of second-order perturbation energies which occur in both MP2 and DH density functionals. We have shown that stochastic summation over the continuum orbitals allows for the calculation of MP2 correlation energies with speedups of up to 3 orders of magnitude at remarkably low errors. This significant increase in efficiency enables calculations with several hundreds of electrons at a relatively low computational cost, making it possible to standardly apply MP2 and DH methods in a PW basis, which has so far been intractably expensive even on high-performance compute clusters. Importantly, the results presented here also enable straightforward DH calculations in the condensed phase, thus extending the availability of one of the most accurate density functional methods available to date to condensed matter systems. We have also shown that stochastic sampling of the continuum orbitals can easily be extended to other approaches, demonstrating the generality of the ansatz employed here. Calculations carried out within a stochastic RPA suggest errors comparable to the stochastic MP2 scheme. The stochastic sampling scheme itself is straightforward, easy to implement, and based on simple physical concepts.

Stochastic sampling of continuum states permits to easily obtain basis-set limit values, be it for periodic or isolated systems, with maximum errors of only 0.1 meV per electron. This accuracy makes basis-set limit values for correlation energies available using reasonable computational resources and execution times. This will allow for thorough benchmarking of new computational methods without basis-set bias, for routine postprocessing of potential energy landscapes generated using lower-level methods, as well as on-the-fly generation of high-accuracy first-principles molecular dynamics trajectories using multiple time stepping schemes. This data, in turn, can be used to feed high-throughput methods based on artifical intelligence. Overall, the techniques presented in this text pave the road to routinely apply accurate MP2 and DH calculations at the basis-set limit in condensed matter systems, ultimately extending the use of a method well-established for isolated systems to the condensed phase.

Appendix

Appendix B contains details on system setups; fitting errors for all basis-set limit values reported in this chapter; graphs of $E_c^{\mathrm{MP(s)2}}(N_{\mathrm{vir}})$ for all systems; E_c^{MP2} obtained in Gaussian and Gaussian/PW basis sets for comparison; as well as relative sampling errors and their standard deviation as a function of $\varepsilon_c^{\mathrm{gap}}$ for the benzene monomer and crystal.

Part III
Making the nuclei move

6 Ab initio molecular dynamics

`Because nature is dynamic and states statistical.`

This chapter provides further theoretical background relevant to the subsequent chapters of the thesis and draws heavily on refs [9, 20, 33, 195], which I recommend to the reader for detailed information. It provides an introduction to ab initio molecular dynamics, presenting the fundamental concepts and methodologies before returning to the results of the thesis in the next part.

6.1 Time versus ensemble averages and ergodicity

Statistical mechanics (in its classical flavour), links the observation of macroscopic properties (observables) to microscopic states. A microscopic state can be described as a point $\gamma = (\{\mathbf{p}\}, \{\mathbf{q}\})$ defined by the generalized momenta $\{\mathbf{p}\}$ and coordinates $\{\mathbf{q}\}$. This point lies in the phase space $\Gamma(\{\gamma\})$ of all generalized momenta and coordinates attainable by the system.[9,33] Furthermore, a collection of states that satisfy the same macroscopic observables (e.g., fixed number of particles N, volume V, and internal energy E) is called an *ensemble*. Thus, a complete ensemble is characterized by all microscopic states that are consistent with specific macroscopic characteristics.

Let us name $\mathcal{A}$ one of the properties corresponding to a defined ensemble. One can now assume that $\mathcal{A}$ is a function of the points γ that traverse the phase space and belong to the corresponding ensemble. At instantaneous time t, the property is given by $\mathcal{A}(\gamma(t))$. The observed property $\mathcal{A}_{\text{obs}}$ at macroscopic scale is rather obtained over the observation time t_{obs}. Thus, the observable property $\mathcal{A}_{\text{obs}}$ can be calculated as the time average

$$\mathcal{A}_{\text{obs}}(t_{\text{obs}}) = \langle \mathcal{A}(\gamma(t)) \rangle_{t_{\text{obs}}} = \frac{1}{t_{\text{obs}}} \int_0^{t_{\text{obs}}} \mathcal{A}(\gamma(t)) dt \tag{6.1}$$

In experiments, the concept of time averaging arises naturally as a series of mea-

surements is conducted over specific time intervals, from which the average value is determined.

In the conventional theory of statistical mechanics, the time average is replaced with an ensemble average. For a given ensemble, the microscopic states can be assigned a specific probability distribution. The partition function, denoted as Z, plays a significant role in characterizing the probability distribution of an equilibrium ensemble since it serves as a normalization factor. Assuming that each state γ has a probability proportional to $f(\gamma)d\gamma$ to occur, the partition function is defined as

$$Z = \int_{\Gamma} f(\gamma)d\gamma \tag{6.2}$$

such that the ensemble average of the property $\mathcal{A}$ reads

$$\langle \mathcal{A} \rangle_{\text{ens}} = \frac{1}{Z} \int_{\Gamma} \mathcal{A}(\gamma) f(\gamma) d\gamma \tag{6.3}$$

However, accounting for all microscopic states in the calculation of the partition function is in general prohibitive due to the incommensurate size of the ensemble (and phase space), and calls for what is called (enhanced) sampling methods. The question arises if it would be possible to similarly track states in the ensemble by observing them over a sufficiently large amount of time. The answer is positive under the assumptions of the *ergodic hypothesis*.

The ergodic hypothesis states that when considering long time intervals, a system will spend time in different regions of the phase space of microstates with the same energy in proportion to the volume of each region. In other words, over a significant period of time, all accessible microstates are equally likely to occur. The ergodic hypothesis thus provides a compelling alternative to the calculation or direct sampling of the partition function by relating it to time averages:

$$\langle \mathcal{A} \rangle_{\text{ens}} = \lim_{t_{\text{obs}} \to \infty} \mathcal{A}_{\text{obs}}(t_{\text{obs}}) \tag{6.4}$$

If a system is observed for a sufficiently long time t_{obs}, the average value of an observable over time will approach its average value over an ensemble. This assumes that the system and the simulation algorithm are ergodic, meaning that no specific region in phase space is excluded, and that the density distribution of the points γ covered by the trajectory reaches a stationary distribution.

While ergodicity is not universally proven and has been disproven in certain cases, there is substantial evidence supporting its validity in many scenarios. This assumption is one of the main motivations for conducting molecular dynamics (MD) simulations. The average behavior of a system can therefore be studied by computing its time evolution: thermodynamical quantities can be calculated as time averages of the observables

of interest over a sufficiently long MD trajectory. In this sense, the next sections will discuss how MD can be run in combination with quantum chemical approaches like those introduced previously in the theoretical sections about wavefunction-based (Section 2.2) and DFT (Section 2.3) methods.

6.2 Equations of motion

When looking at the dynamics of quantum many-body systems, the consideration of the fundamental Schrödinger equation must turn to its time-dependent general form. The time-dependent Schrödinger equation describes the time evolution of the many-body wavefunction, which represents the quantum state of the system over time. The time-dependent Schrödinger equation is

$$i\hbar\frac{\partial}{\partial t}\Psi(\{\mathbf{R}\},\{\mathbf{r}\},t) = \hat{\mathcal{H}}\Psi(\{\mathbf{R}\},\{\mathbf{r}\},t) \tag{6.5}$$

where Ψ represents the many-body wavefunction, $\hbar$ is the reduced Planck's constant, and $\hat{\mathcal{H}}$ is the Hamiltonian operator defined previously in eq 2.2, which describes the total energy of the system. The P nuclear coordinates $\{\mathbf{R}_I, I = 1, ..., P\}$ are represented by $\{\mathbf{R}\}$ as well as the N electronic positions $\{\mathbf{r}_i, i = 1, ..., N\}$ that are contained in the set $\{\mathbf{r}\}$. The time-dependent Schrödinger equation provides a powerful framework for studying the behavior of electrons in atoms, molecules, and solids, enabling the calculation of various electronic properties and spectroscopic observables. Solving this equation is a challenging task due to the complex nature of the many-body problem and the high dimensionality of the wavefunction. Various approximation methods and numerical techniques are employed to tackle this equation and obtain insights into the dynamic behavior of quantum systems.

We will once again rely on the Born-Oppenheimer approximation to approximate quantum dynamics (c.f. Section 2.1.1). This approximation is based on the observation that the time scale associated with the motions of the nuclei is generally slower than that of the electrons. Alternatively, it can be demonstrated that, at leading orders, the effects of the movement of the nuclei do not affect the electronic states.[16] Therefore, under appropriate assumptions, the electrons do not transition between stationary states upon reasonable movement of the nuclei. This is called the *adiabatic approximation*, which assumes that the electrons instantaneously follow the motion of the nuclei, while remaining in the same stationary (ground) state of the electronic Hamiltonian. This stationary state evolves in time because of the electrostatic coupling between the nuclei and the electrons, but the many-electron wavefunction instantaneously adjusts to the nuclear wavefunction. This approximation thus ignores the possibility of having non-radiative or any other transitions between electronic eigenstates. The adiabatic approximation starts by separating the total Hamiltonian into a nuclear component dictating the time-evolution, and an electronic Hamiltonian describing the stationary

states:

$$\hat{\mathcal{H}} = \hat{h}_n(\{\mathbf{R}\}) + \hat{\mathcal{H}}_e(\{\mathbf{R}\},\{\mathbf{r}\}) \tag{6.6}$$

whose terms have been defined in eqs 2.5 and 2.7. Previous chapters of this thesis described the methods to solve the time-independent many-electron problem in the Born-Oppenheimer approximation (eq 2.4) that I recall here

$$\hat{\mathcal{H}}_e\Phi_e(\{\mathbf{r}\};\{\mathbf{R}\}) = E(\{\mathbf{R}\})\Phi_e(\{\mathbf{r}\};\{\mathbf{R}\}) \tag{6.7}$$

In fact, it can be seen that the eigenstates $\Phi_{e,n}$ of this problem form a complete basis

$$\int_{-\infty}^{+\infty} \Phi_{e,n}(\{\mathbf{r}\};\{\mathbf{R}\})\Phi_{e,m}(\{\mathbf{r}\};\{\mathbf{R}\})d\mathbf{r}_1\ldots d\mathbf{r}_N = \delta_{nm} \tag{6.8}$$

such that the many-body wavefunction can be expanded as

$$\Psi(\{\mathbf{R}\},\{\mathbf{r}\},t) = \sum_{n=1}^{\infty} \Theta_n(\{\mathbf{R}\},t)\Phi_{e,n}(\{\mathbf{r}\};\{\mathbf{R}\}) \tag{6.9}$$

where $\Theta_n(\{\mathbf{R}\},t)$ are the time-dependent expansion coefficients that correspond to the wavefunction components of the nuclei in each one of the *adiabatic* electronic eigenstates $\Phi_{e,n}$. Substituting the wavefunction ansatz of eq 6.9 into the many-body time-dependent Schrödinger eq 6.5, followed by multiplying $\Phi^*_{e,n}(\{\mathbf{r}\};\{\mathbf{R}\})$ from the left and subsequent integration over all electronic coordinates, gives a set of coupled differential equations

$$i\hbar\frac{\partial}{\partial t}\Theta_n(\{\mathbf{R}\},t) = \left[\hat{h}_n(\{\mathbf{R}\}) + E_n(\{\mathbf{R}\})\right]\Theta_n(\{\mathbf{R}\},t) + \sum_{m=1}^{\infty}\chi_{nm}(\{\mathbf{R}\})\Theta_m(\{\mathbf{R}\},t) \tag{6.10}$$

that involve the nuclear Hamiltonian $\hat{h}_n(\{\mathbf{R}\})$ for the nuclei evolving on the PES $E_n(\{\mathbf{R}\})$, provided by the stationary state n of the electron system. The coefficients $\chi_{nm}(\{\mathbf{R}\})$ couple the electronic and nuclear degrees of freedom and are termed the non-adiabatic couplings

$$\begin{aligned}\chi_{nm}(\{\mathbf{R}\}) = &\int_{-\infty}^{+\infty}\Phi^*_{e,n}(\{\mathbf{r}\};\{\mathbf{R}\})\left\{-\frac{1}{2}\sum_{I=1}^{P}\frac{1}{M_I}\nabla_I^2\right\}\Phi_{e,m}(\{\mathbf{r}\};\{\mathbf{R}\})d\mathbf{r}_1\ldots d\mathbf{r}_N \\ &-\sum_{I=1}^{P}\frac{1}{M_I}\int_{-\infty}^{+\infty}\Phi^*_{e,n}(\{\mathbf{r}\};\{\mathbf{R}\})\nabla_I\Phi_{e,m}(\{\mathbf{r}\};\{\mathbf{R}\})d\mathbf{r}_1\ldots d\mathbf{r}_N\nabla_I\end{aligned} \tag{6.11}$$

The non-adiabatic couplings have diagonal contributions χ_{nn} that depend only on the adiabatic state $\Phi_{e,n}$, and represent part of the correction to the energy $E_n(\{\mathbf{R}\})$ due to the electron-nuclei coupling. We note that no approximation has been made up to here, since the solution of the electronic problem $\Phi_{e,n}$ was only introduced as an ad-hoc basis to project the total wavefunction Ψ. As a consequence, the adiabatic

approximation to the entire non-adiabatic problem consists of retaining only the diagonal coupling terms ($\chi_{nm} = 0$ if $n \neq m$) in eq 6.10 that transforms into

$$i\hbar\frac{\partial}{\partial t}\Theta_n(\{\mathbf{R}\},t) = \left[\hat{h}_n(\{\mathbf{R}\}) + E_n(\{\mathbf{R}\}) + \chi_{nn}(\{\mathbf{R}\})\right]\Theta_n(\{\mathbf{R}\},t) \tag{6.12}$$

which is completely decoupled as opposed to eq 6.10. For the diagonal elements χ_{nn}, note that the second term in eq 6.11 vanishes.

The latter equation therefore demonstrates that the time-dependent Schrödinger equation can be approximated by the sole equations of motion of the nuclei, given the Born-Oppenheimer PES and states $\Phi_{e,n}$ provided by the electronic time-independent problem. This implies that the nuclear motion occurs without changing the electronic state during time evolution. This is equivalent to a decoupling of the coupled nuclear-electronic many-body wavefunction into a direct product of nuclear and electronic wavefunctions

$$\Psi(\{\mathbf{R}\},\{\mathbf{r}\},t) \simeq \Theta_n(\{\mathbf{R}\},t)\Phi_{e,n}(\{\mathbf{r}\};\{\mathbf{R}\}) \tag{6.13}$$

such that, by comparison with eq 6.9, the adiabatic approximation corresponds to retaining only one component of the many-body wavefunction projected onto the electronic basis.

In practice, the diagonal coupling terms χ_{nn} are of small magnitude.[16] As a result, in the ultimate Born-Oppenheimer approximation of the nuclear dynamics, the diagonal couplings are also neglected and the equations of motion for the nuclei finally simplify to

$$i\hbar\frac{\partial}{\partial t}\Theta_n(\{\mathbf{R}\},t) = \left[\hat{h}_n(\{\mathbf{R}\}) + E_n(\{\mathbf{R}\})\right]\Theta_n(\{\mathbf{R}\},t) \tag{6.14}$$

This equation is nothing but the time-dependent Schrödinger equation for the nuclear wavefunction under the Born-Oppenheimer approximation. It is valid for systems where the kinetic energy of the nuclei is much smaller than the relative energies between neighboring electronic levels. In this case, it is accurate to describe the nuclear dynamics with the PES of the electronic ground state. However, the non-adiabatic coupling terms must be considered for systems that have two or more electronic states contributing to the physical properties, or have crossovers between PES of different electronic states.

From now on, for the sake of notation, we will invariably move the potential operator of the nuclei-nuclei interactions $\hat{V}_{NN}(\{\mathbf{R}\})$ (eq 2.7) from $\hat{h}_n(\{\mathbf{R}\})$ to $E_n(\{\mathbf{R}\})$, in the sense of

$$E_n(\{\mathbf{R}\}) \leftarrow E_n(\{\mathbf{R}\}) + \sum_{I<J}^{P}\frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} \tag{6.15}$$

6.3 Born-Oppenheimer molecular dynamics

In order to enable the dynamical simulation of real systems with reasonably large size, the propagation of the nuclei is most of the time operated under the *classical nuclei approximation*. By seeing the nuclear wavefunction as a product of Dirac's delta functions $\delta(\mathbf{r} - \mathbf{R}_I)$, and taking the classical limit $\hbar \to 0$, the equations of motion 6.14 can be mapped onto the classical propagation of the nuclei,[9,16] that is

$$M_I \ddot{\mathbf{R}}_I(t) = -\nabla_I E_n(\{\mathbf{R}(t)\}) \tag{6.16}$$

where $\mathbf{R}_I$ are the coordinates of the nuclei I, and $E_n(\{\mathbf{R}(t)\})$ is the nth adiabatic PES provided by the time-independent electronic problem (that includes here the nuclei-nuclei interactions) at time t. The integration of these equations of motion is typically achieved by computing nuclear forces using the Hellmann-Feynman theorem,[336,337] that relates forces to the expectation value of the electronic Hamiltonian operator

$$\mathbf{F}_I := -\nabla_I E_n(\{\mathbf{R}\}) = -\left\langle \Phi_{e,n}(\{\mathbf{R}\}) \right| \frac{\partial \hat{\mathcal{H}}_e(\{\mathbf{R}\})}{\partial \mathbf{R}_I} \left| \Phi_{e,n}(\{\mathbf{R}\}) \right\rangle \tag{6.17}$$

where $\mathbf{F}_I$ is the force acting on nuclei I, and, of course, $n = 0$ in case of ground state dynamics.

The numerical integration of the above Newtonian equations is called *ab initio* (AIMD) or *first-principles* (FPMD) molecular dynamics.[195] The development of AIMD techniques allows for simulations of complex systems without the need for adjustable (empirical) parameters. The key principle behind AIMD is the on-the-fly calculation of energies and forces through accurate electronic structure calculations as seen in Chapter 2. While the concept of AIMD can be applied with any electronic structure method, DFT is currently the most widely utilized theory in combination with MD simulations due to its most optimal cost-accuracy trade-off (cf. Section 2.3).

In Born-Oppenheimer MD (BOMD), the ground state potential energy $E_0(\{\mathbf{R}\})$ is calculated on-the-fly at every propagation time step of the nuclei. This is accomplished via the various computational approaches described in Chapters 2 and 3 that provide the approximate energy $E(\{\mathbf{R}\})$ of the many-electron system. In a classical picture, the Hamiltonian $\mathcal{H}$ of the N-nuclei system is thus given by

$$\mathcal{H} = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2M_i} + E(\{\mathbf{q}\}) \tag{6.18}$$

where the generalized $\mathbf{p}_i$ and $\mathbf{q}_i := \mathbf{R}_i$ represent respectively the nuclear momenta and coordinates. Those define the $6N$-dimensional phase space $\Gamma(\{\gamma\})$ in which any microstate $\gamma = (\{\mathbf{p}\}, \{\mathbf{q}\})$ lies. The Hamiltonian formulation of classical mechanics therefore describes the time-evolution of the particle system in phase space according

to

$$\dot{\mathbf{q}}_i = \frac{\partial \mathcal{H}}{\partial \mathbf{p}_i} \quad ; \quad \dot{\mathbf{p}}_i = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}_i} \tag{6.19}$$

Standard methods for propagating the nuclei with respect to eq 6.19 rely on finite differences: starting from initial positions and velocities at time t, the positions and velocities at a subsequent time $t + \delta t$ are calculated with a satisfactory level of accuracy. The selection of the time step δt is mainly influenced by the propagation scheme employed, and is typically determined by the fastest physical motion (e.g., vibrational modes) expected in the system. Numerous finite-difference algorithms exist, among which the probably most widely used is the velocity Verlet algorithm.[9,34] Finally, the consideration of different thermodynamic ensembles in AIMD can be obtained with the help of thermostats, respectively barostats, which are added to the equations of motion 6.19.[33,338,339]

6.4 Multiple time step algorithms

As mentioned earlier, the ergodic hypothesis necessitates sampling the PES over sufficiently long time scales. In AIMD simulations, the length of trajectories is primarily determined by the computational cost of evaluating the forces using the underlying electronic structure method, such as DFT or wavefunction-based approaches. The more computationally demanding the force computations, the shorter the achievable time scale for the simulation. Therefore, the length of AIMD trajectories is once again constrained by the balance between accuracy and computational efficiency. The number of force evaluations is directly influenced by the time step used to integrate the equations of motion. By using larger time steps, the accessible time scale can be linearly increased. However, it is important to select a time step that is sufficiently small to ensure accurate numerical integration, particularly for the fastest force components. Nuclear forces exhibit a wide range of characteristic time scales in AIMD simulations such that, traditionally, even slowly varying force components need to be integrated at small intervals determined by the fastest components. As a solution to this, multiple time step (MTS) algorithms allow to separate the dynamical propagation of force components depending on their respective time scales, and possibly reduce the number of calculations for the most expensive force components.

The MTS integrator employed in this thesis, which is utilized in AIMD, is derived from the reversible reference system propagator algorithm (r-RESPA) which was initially introduced in the context of classical MD.[98] r-RESPA is a meticulously designed integration scheme that makes the time evolution reversible, ensuring accuracy and energy conservation throughout the simulation. It is derived from the Liouville operator appearing in the Hamiltonian formulation of classical mechanics.[33] The Liouville

operator L is defined by

$$iL = \sum_{j=1}^{3N} \left[\frac{\partial \mathcal{H}}{\partial p_j} \frac{\partial}{\partial q_j} - \frac{\partial \mathcal{H}}{\partial q_j} \frac{\partial}{\partial p_j} \right] = \sum_{j=1}^{3N} \left[\dot{q}_j \frac{\partial}{\partial q_j} + F_j \frac{\partial}{\partial p_j} \right] \tag{6.20}$$

according to the corresponding coordinates q_j and momenta p_j. In Hamiltonian mechanics, the Liouville operator acts as the propagation operator for any point γ in phase space via

$$\gamma(t_0 + t) = e^{iLt} \gamma(t_0) \tag{6.21}$$

where $\gamma(t_0 + t)$ is the phase space element at time $t_0 + t$ obtained from its counterpart at time t_0. From that formalism, if forces were to be decomposed into fast $\mathbf{F}^{\text{fast}}$ and slow $\mathbf{F}^{\text{slow}}$ components eq 6.20 would translate into

$$\begin{aligned} iL &= \sum_{j=1}^{3N} \dot{q}_j \frac{\partial}{\partial q_j} + \sum_{j=1}^{3N} F_j^{\text{fast}} \frac{\partial}{\partial p_j} + \sum_{j=1}^{3N} F_j^{\text{slow}} \frac{\partial}{\partial p_j} \\ &:= iL_p + iL_q^{\text{fast}} + iL_q^{\text{slow}} \end{aligned} \tag{6.22}$$

The split components of the Liouville operator do not commute, but applying a second-order Trotter decomposition to the corresponding propagators allows to transform the propagation eq 6.21 into[33,98]

$$\gamma(t_0 + \Delta t) = e^{iL_q^{\text{slow}}(\Delta t/2)} \left[e^{iL_q^{\text{fast}}(\Delta t/2n)} e^{iL_p(\Delta t/n)} e^{iL_q^{\text{fast}}(\Delta t/2n)} \right]^n e^{iL_q^{\text{slow}}(\Delta t/2)} \gamma(t_0) \tag{6.23}$$

where Δt is a pre-defined *outer* time step. Note that third-order and higher terms have been discarded. In this form, the inner brackets of eq 6.23 correspond to the velocity Verlet propagation of the fast components,[98] over smaller *inner* time steps $\delta t = \Delta t/n$. After a first half-Δt step with slow components, the propagation is executed n times with fast components at the inner time step δt. Finally, the slow components are updated to complete the full time propagation of the system over the entire time step Δt.

As mentioned previously, the r-RESPA algorithm was first developed to separate force components attributed to specific force field terms in classical MD. In this framework, bonded forces are generally faster in nature, while forces resulting from non-bonded interactions exhibit slower variations. In AIMD, the slow and fast forces are rather less well-defined and several attempts have been investigating how to decouple them.[340–344] A recent introduction by Liberatore et al.[345] presents a versatile MTS implementation for AIMD and mixed AIMD-QM/MM simulations employing various electronic structure methods. This AIMD-MTS approach is based on the fundamental concept that accuracy enhancements in quantum chemical methods arise from better treatment of exchange-correlation effects in DFT or correlation contributions in wavefunction-based methods. Since these terms constitute a relatively small fraction of

the total energy, the corresponding force contributions are expected to exhibit smooth variations over time. This provides an opportunity to employ a computationally efficient but approximate description of (exchange-)correlation effects as a lower-level method to compute the fast force components $\mathbf{F}^{\text{fast}} = \mathbf{F}^{\text{L}}$, while a higher-level method is utilized to compute a correction term $\mathbf{F}^{\text{slow}} = \Delta\mathbf{F} = \mathbf{F}^{\text{H}} - \mathbf{F}^{\text{L}}$, which captures the slow force.

By employing the r-RESPA algorithm, AIMD-MTS trajectories can be generated with the same level of accuracy as the high-level method but at a significantly reduced computational cost of approximately $n = \Delta t/\delta t$ times. This approach allows for the combination of different methods, such as semiempirical methods/DFT, DFT/MP2, ..., or even DFT/CCSD(T). Similarly, DFT calculations utilizing higher-rung functionals can be expedited by using lower-rung functionals to compute the fast force components, for example, combining GGA with hybrid DFT functionals. In practical applications, it was indeed observed that the force difference $\Delta\mathbf{F} = \mathbf{F}^{\text{H}} - \mathbf{F}^{\text{L}}$ evolves on a much slower time scale compared to either $\mathbf{F}^{\text{L}}$ or $\mathbf{F}^{\text{H}}$ individually.[345] This characteristics allows for the dynamical decoupling between $\mathbf{F}^{\text{slow}} = \Delta\mathbf{F} = \mathbf{F}^{\text{H}} - \mathbf{F}^{\text{L}}$ from the fast force component $\mathbf{F}^{\text{fast}} = \mathbf{F}^{\text{L}}$.

Since the MTS propagator reproduces the high level of accuracy by construction, it also enables the calculation of high-level structural and dynamical properties at much reduced cost (c.f. Chapter 7). If τ^H and τ^L represent the times required to compute $\mathbf{F}^{\text{H}}$ and $\mathbf{F}^{\text{L}}$ respectively, the ideal speedup s of the algorithm can be estimated as

$$s = \frac{n\tau^H}{n\tau^L + \tau^H} = \frac{n}{n\tau^L/\tau^H + 1} \tag{6.24}$$

which demonstrates that the ideal speedup indeed reaches the time step ratio n in the limiting case of a very expensive (inexpensive) higher (lower) level ($\tau^H \gg \tau^L$). Nevertheless, in practice, the achieved speedups are smaller. This can be attributed to the fact that, as the time step ratio increases, the initial guess for solving the high-level wavefunction deteriorates. This deterioration increases the time required for the wavefunction (re)optimization when using larger ratios between time steps. Typical ratios for production trajectories can vary between 4 and 15, with the largest (speedups) being attainable in the presence of thermostats (as observed in Chapter 7 in the NVT ensemble).

By design, all energy-conserving ΔF-MTS schemes ensure that the generated trajectories maintain the full accuracy of the chosen high-level method, regardless of the specific lower-level method employed. The choice of the lower-level method primarily affects the efficiency gain. Speedups are maximized when the computational cost associated with computing the slow force component $\mathbf{F}^{\text{slow}} = \Delta\mathbf{F} = \mathbf{F}^{\text{H}} - \mathbf{F}^{\text{L}}$ is the dominant factor. This occurs when there is a significant difference in computational

cost between the high- and low-level methods. In turn, efficiency is improved further when the ratio $n = \Delta t/\delta t$ is maximized, in accordance with the possible dynamical decoupling between fast and slow force components.

As a final remark, the value of n in most MTS schemes is constrained by the occurrence of resonances, as discussed in previous studies.[346] These resonances arise when there is a persistent coupling between the fast and slow force components, limiting the efficiency gain of the MTS algorithm. However, the use of specific thermostats can help mitigate these resonances and improve the overall performance of the MTS approach.[347–350]

6.5 Machine learning-aided multiple time step algorithms

The field of machine learning (ML) being particularly extensive and dynamic would require more than this thesis to be presented exhaustively. As references of choice, I invite the reader interested in the details of ML methods to consult refs [91] and [351]. I also would like to recommend to the reader the recent books [20] and [352] that focus on the use of ML in combination with methods from the field of computational chemistry.

Section 6.5 is a postprint version of the section entitled *Machine learning-aided multiple-time-step MD* as published in the article:

Mouvet, F; **Villard, J.**; Bolnykh, V.; Rothlisberger, U. Recent advances in first-principles based molecular dynamics. *Accounts of Chemical Research* 2022, 55, 221–230.

Reproduced under the terms of the CC-BY-NC-ND 4.0 License.

With the advance of artificial intelligence algorithms and their rapid spread over computational chemistry, ML has also become a promising tool to cope with AIMD bottlenecks incurred by the cost of having to solve the electronic structure problem at every time step. Recent applications of ML with MD engines have mostly focused on the replacement of quantum calculations with ML-designed potentials for force field-like dynamics.[353,354] Commonly, kernel methods[355,356] or neural networks[357,358] are employed to learn from training data (originating from presumably accurate DFT or coupled cluster calculations) and predict potential energies and/or forces at a fraction of the quantum reference's cost. In parallel, several attempts have shown how ML can accelerate the solving of DFT equations[359] or improve the cost-accuracy trade-off of DFT with ML-designed functionals.[360,361]

As discussed in the previous section, the ΔF-MTS scheme opens a wealth of possible combinations of high- and low-level methods. The idea of using ML models as the lower-level surrogate or as computationally expedient replacement of the higher-level method emerges therefore quite naturally.

We have recently introduced such a combined ΔF-MTS-ML scheme[99] in which energies and forces are inferred based on the Operator Quantum Machine Learning (OQML) kernel method proposed by Christensen et al.[362,363] In OQML, response properties, such as nuclear gradients, are included in the training process for better data efficiency and high-quality force predictions. OQML owns some similarities with kernel ridge regression (KRR)[364] but rather expands the kernel in a basis of kernel functions that depend on all the atomic environments included in the training set. Illustrating this on energies only, those are obtained for a query system S as local atomic contributions of all atoms a in the system,

$$E_S = \sum_{a \in S} \varepsilon_{\text{local}}(\mathbf{q}_a) = \sum_{a \in S} \sum_{\bar{S} \in \{T\}} \sum_{\bar{b} \in \bar{S}} k(\mathbf{q}_{\bar{b}}, \mathbf{q}_a)\, \alpha_{\bar{b}} := \mathbf{k}_S^T \boldsymbol{\alpha} \tag{6.25}$$

where $\bar{b}$ corresponds to all atoms of all systems $\bar{S}$ included in the training set $\{T\}$, $\mathbf{q}_a$ and $\mathbf{q}_{\bar{b}}$ are the respective atomic environment representations[365] and $k(\mathbf{q}, \mathbf{q}')$ is the user-defined kernel function. $\boldsymbol{\alpha}$ is the vector containing the regression coefficients for each atomic environment in the training set. Considering supplementary query systems, eq 6.25 can be cast into the matrix form

$$\mathbf{E} = \mathbf{K}\boldsymbol{\alpha} \tag{6.26}$$

where the row of $\mathbf{K}$ corresponding to the system S is given by $\mathbf{k}_S^T$. It follows that, in contrast to KRR or Gaussian process regression (GPR),[366] the kernel matrix is neither square nor symmetric. Also, the number of regression coefficients equals the number of atoms in the training set, which for a set of N systems with an average number of atoms A gives a matrix of size $N \times NA$. The OQML method extrapolates this to the simultaneous learning of energies and forces and writes the learning problem as

$$\begin{bmatrix} \mathbf{E} \\ \mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{K} \\ -\dfrac{\partial}{\partial \mathbf{R}_i}\mathbf{K} \end{bmatrix} \boldsymbol{\alpha} := \mathbf{K}^{\text{OQML}} \boldsymbol{\alpha} \tag{6.27}$$

where $-\partial/\partial\mathbf{R}_i$ denotes the force operator acting on atom i of a query system, such that the three rows of $\mathbf{F}$ corresponding to atom i are given by

$$\mathbf{F}_i = -\frac{\partial}{\partial \mathbf{R}_i} E_S = -\sum_{a \in S} \sum_{\bar{S} \in \{T\}} \sum_{\bar{b} \in \bar{S}} \alpha_{\bar{b}} \frac{\partial k(\mathbf{q}_{\bar{b}}, \mathbf{q}_a)}{\partial \mathbf{q}_a} \frac{\partial \mathbf{q}_a}{\partial \mathbf{R}_i} \tag{6.28}$$

In the training phase, $\mathbf{E}$ and $\mathbf{F}$ are known from the training data and $\mathbf{K}^{\text{OQML}}$ is constructed from the atomic representations (i.e. from the atomic coordinates of the training systems). The OQML formalism thus avoids the inclusion of double derivatives in the kernel matrix as well as reduces the number of regression coefficients compared to conventional GPR. Ref [363] discusses the method's scalings in more

details, which become relevant for investigating larger heterogeneous systems. The extended kernel matrix is rectangular of size $(N + 3AN) \times NA$ and the typical matrix inversion encountered in KRR is replaced in practice by a singular-value decomposition (SVD) of $\mathbf{K}^{\mathrm{OQML}}$ to obtain the regression coefficients $\boldsymbol{\alpha}$ from eq 6.27. The choice of SVD is motivated by a better numerical stability than the more common solution of the corresponding normal equations. Once the model is trained and $\boldsymbol{\alpha}$ determined, energies and forces are inferred with the reconstruction of $\mathbf{K}^{\mathrm{OQML}}$ for new atomic configurations and the evaluation of eq 6.27. In addition to enforcing consistency between energies and forces during training, the OQML method has the advantage of reducing the number of regression coefficients to the number of atoms in the training set, consequently not expanding the kernel basis for treating derivatives.

With this machinery at hand, it becomes possible to couple a Δ-ML correction of forces with the MTS algorithm. Training on the difference between a low-level and the target high-level method, such a correction on atom i of a query system S is written as

$$\Delta\mathbf{F}_i^{\mathrm{ML}} = -\frac{\partial}{\partial\mathbf{R}_i}\left(E_S^H - E_S^L\right) = -\sum_{a\in S}\sum_{\bar{S}\in\{T\}}\sum_{\bar{b}\in\bar{S}}\alpha_{\bar{b}}\frac{\partial k(\mathbf{q}_{\bar{b}},\mathbf{q}_a)}{\partial\mathbf{q}_a}\frac{\partial\mathbf{q}_a}{\partial\mathbf{R}_i} \tag{6.29}$$

which requires only one kernel matrix and one set of $\boldsymbol{\alpha}$ coefficients since $\Delta\mathbf{F}$ is now the target quantity. Two possible schemes have been explored for an MTS-decomposition of the forces. The first uses the ML correction to recover the higher-level method at a fraction of the cost and therefore defines (see eqs 6.22-6.23)

$$\mathbf{F}^{\mathrm{fast}} = \mathbf{F}^{\mathrm{L}} \quad , \quad \mathbf{F}^{\mathrm{slow}} = \Delta\mathbf{F}^{\mathrm{ML}} \qquad \text{(scheme \textbf{I})}$$

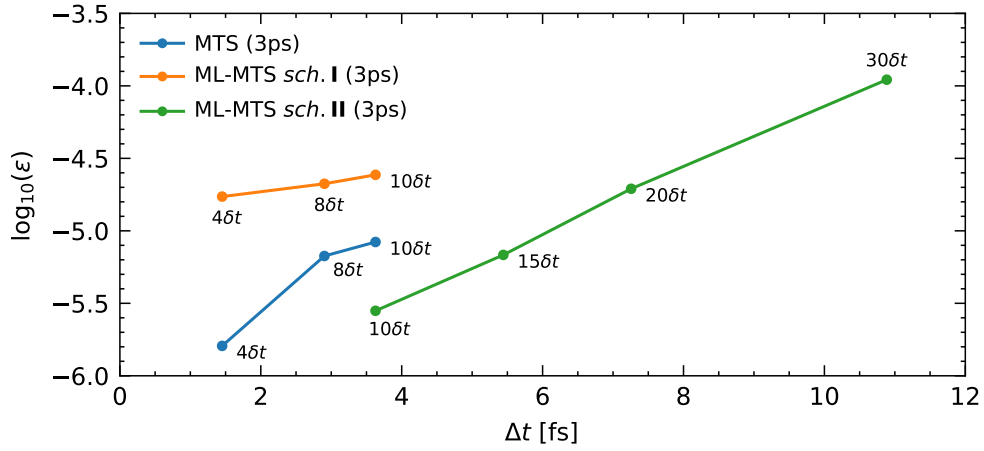
while the second intends to improve the lower-level method to better agree with the higher reference at the outer time step:

$$\mathbf{F}^{\mathrm{fast}} = \mathbf{F}^{\mathrm{L}} + \Delta\mathbf{F}^{\mathrm{ML}} \quad , \quad \mathbf{F}^{\mathrm{slow}} = \mathbf{F}^{\mathrm{H}} - \mathbf{F}^{\mathrm{fast}} \qquad \text{(scheme \textbf{II})}$$

A proof-of-concept was demonstrated on a 32-molecule sample of liquid water with LDA as the MTS low level and the hybrid PBE0 functional as the high-level method (or ML target).[99] The two Δ-learning schemes (schemes **I** and **II**) were implemented in an extension of a development version of the CPMD code[47] with the Gaussian kernel $k(\mathbf{q},\mathbf{q}') = \exp\left(-|\mathbf{q}-\mathbf{q}'|^2/2\sigma^2\right)$ and aSLATM representations.[367] Training data was generated by running MTS-LDA/PBE0 trajectories on liquid water and small water clusters of various sizes, followed by a farthest-point sampling (FPS) in the kernel space to enhance data efficiency. In the end, the OQML model owns a total of 10725 regression coefficients and showed an out-of-sample mean absolute error on $|\Delta\mathbf{F}^{\mathrm{ML}}|$ around 0.3 kcal/(mol Å) as well as a mean absolute error on force directions of about 0.7 degrees on a test set of 50 random frames extracted from an MTS-LDA/PBE0 trajectory.

Figure 6.1a depicts the energy conservation during NVE liquid water MTS trajectories for standard and ML-aided MTS schemes. ML-MTS scheme **I**, for which the high-level forces are predicted directly from the ML model, shows higher energy fluctuations than the standard MTS algorithm, which can be explained by the statistical errors introduced by the ML inference. On the other hand, using ML for low-level predictions while using explicit PBE0 calculations as the high-level method scheme **II** allows to significantly increase the outer time step at which PBE0 has to be evaluated. The ML-MTS scheme **II** still provides trajectories with an acceptable energy conservation of $\log_{10}(\varepsilon) = -4.7$ at $\Delta t = 20\delta t$ while the time step ratio reduces to around 8 for the ML-MTS scheme **I**. Resonance effects affect the energy conservation when the time

(a)



(b)

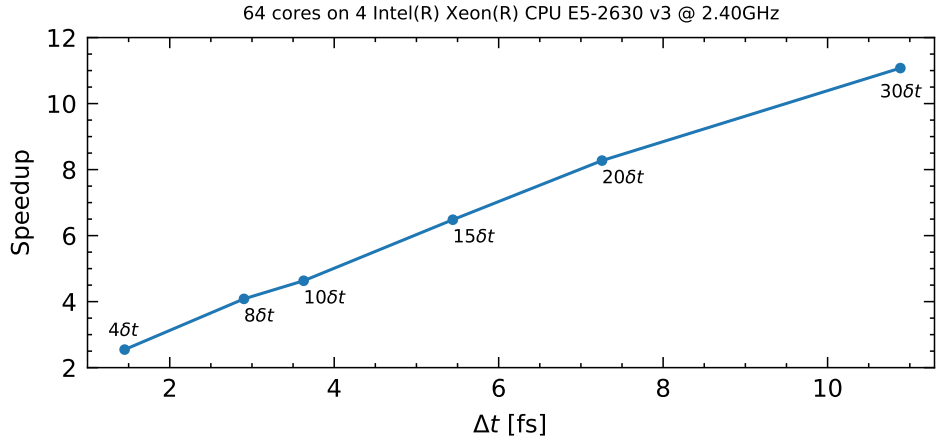


Figure 6.1: (a) Logarithm of the total energy fluctuations for standard MTS, ML-MTS schemes **I** and **II** and different outer time steps $\Delta t = n \cdot \delta t$ with $\delta t = 0.36$ fs. $\varepsilon = \sum_{i=1}^{N} |(E_i - \bar{E})/\bar{E}|/N$ with N the number of outer time steps and E_i the instantaneous, respectively average $\bar{E}$ energies. (b) Standard MTS (4-8δt) and ML-MTS scheme **II** (10-30δt) speedups against conventional Velocity-Verlet Born-Oppenheimer MD.

step ratios increase further.

For a fixed trajectory duration, the effective speedup of the MTS splitting can be measured by eq 6.24 and is reported in Figure 6.1b. While the conventional MTS scheme yields typical speedups over straightforward Velocity-Verlet PBE0 MD of about 2.5 ($4\delta t$) to 4 ($8\delta t$) respectively, the ML-MTS scheme **II** brings additional speedups that reach a factor of 4.6 ($10\delta t$) to 6.5 ($15\delta t$) for equivalent energy fluctuations. Since the ML-MTS scheme **I** completely bypasses the PBE0 calculations, drastic speedups up to 80 can be reached at the price that the generated dynamics fully relies on the accuracy of the ML model in reproducing PBE0 quality results while in scheme **II** this accuracy is fully guaranteed albeit at higher computational cost. Nevertheless, both ML-MTS schemes are able to give radial distribution functions of liquid water in very close agreement with the MTS-LDA/PBE0 reference as illustrated in Figure 6.2, showing that the ML-aided MTS approach is able to reproduce, e.g., structural properties at a high level of accuracy with strongly reduced computational cost.

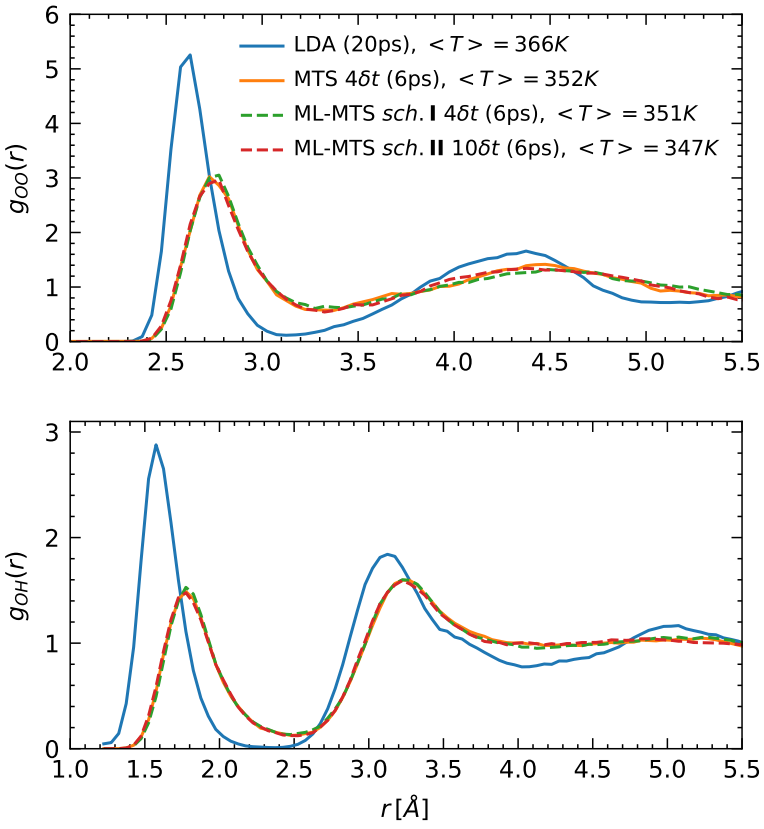


Figure 6.2: Oxygen-oxygen and oxygen-hydrogen radial distribution functions of a 32-molecule water box as calculated from LDA/PBE0-MTS, LDA/ML(PBE0)-MTS scheme **I**, and LDA/ML(PBE0)-MTS scheme **II**, compared with NVE Velocity-Verlet LDA MD. Data from ref [345].

Overall, ML-MTS scheme **I** can achieve speedups of 1 to 2 orders of magnitude (i.e., factors 10-100), while scheme **II** can yield accelerations up to an order of magnitude. Since the speedup of scheme **II** is basically determined by the possible time step ratio (limited most probably due to intrinsic resonances), the performance could be further enhanced when applying stochastic thermostats as discussed and demonstrated in the previous section. On the other hand, for scheme **I**, the accuracy of the Δ-ML corrections is key for keeping faithful dynamics. Scheme **II** finally provides a more reliable reproduction of the high-level dynamics and the quality and computational cost of the ML model are essential to operate at larger Δt. For this scheme, additional efficiency limitations at large n are caused by the increased number of iterations for re-optimizing the PBE0 wavefunction at the outer time steps, which can be improved to some extent by using better extrapolation schemes.[345] ML-aided AIMD certainly holds a lot of promise especially in view of methods for ML error monitoring,[368,369] on-the-fly learning,[370,371] and recent QM/MM[372] and excited-state dynamics[373] implementations.

6.6 Car-Parrinello molecular dynamics

In 1985, Car and Parrinello proposed an AIMD scheme that aimed to reduce computational costs by treating the electronic degrees of freedom as fictitious classical variables, alongside the nuclear motions.[374] This approach, known as Car-Parrinello MD (CPMD), is based on the concept of adiabatic time scale separation between the fast electronic and slow nuclear motions in quantum mechanics. By mapping the original two-component quantum/classical problem onto a two-component classical problem with distinct energy scales, the Car-Parrinello method offers computational efficiency at the expense of losing the explicit time evolution information of the quantum subsystem dynamics.[16,195] A notable distinction between CPMD and BOMD lies in the treatment of orbitals, which no longer undergo optimization at every time step in CPMD, but instead emulate classical entities. This is achieved by attributing a fictitious mass μ and temperature to the orbitals. Notably, it has been demonstrated that this approach upholds the adiabatic separation inherent in the Born-Oppenheimer approximation.[195] Consequently, within the realm of CPMD, the computationally demanding electronic wavefunction optimization step, which poses a bottleneck in BOMD, can be circumvented. Remarkably, this shift in methodology often yields significant speedups, with possible gains of up to an order of magnitude over similar BOMD simulations.

The CPMD time propagation is given by the following extended Lagrangian

$$\mathcal{L}_{\text{CP}} = \frac{1}{2}\sum_{I=1}^{P} M_I \dot{\mathbf{R}}_I^2 + \frac{1}{2}\sum_{i=1}^{N} \mu \langle \dot{\phi}_i | \dot{\phi}_i \rangle - E[\{\phi\}, \{\mathbf{R}\}] + \sum_{i,j}^{N} \Lambda_{ij} \left(\langle \phi_i | \phi_j \rangle - \delta_{ij} \right) \tag{6.30}$$

where N is the number of electronic states, μ is the fictitious electronic mass, ϕ_i is the ith electronic state, and E is the total energy of the many-electron system as provided by DFT (mostly) or wavefunction-based methods. The Λ_{ij} matrix of Lagrange multipliers has been introduced to constrain the orthonormality of electronic orbitals along the dynamics. From the Euler-Lagrange equations, the equations of motion for the nuclear and electronic degrees of freedom become

$$\begin{aligned} M_I\ddot{\mathbf{R}}_I &= -\nabla_I E[\{\phi\},\{\mathbf{R}\}] \\ \mu\ddot{\phi}_i(\mathbf{r},t) &= -\frac{\delta E}{\delta\phi_i^*(\mathbf{r},t)} + \sum_j^N \Lambda_{ij}\phi_j(\mathbf{r},t) \end{aligned} \tag{6.31}$$

with the constant of motion being provided by the Hamiltonian quantity

$$E_{\text{conserved}} = \frac{1}{2}\sum_{I=1}^{P} M_I\dot{\mathbf{R}}_I^2 + \frac{1}{2}\sum_{i=1}^{N}\mu\,\langle\dot{\phi}_i|\dot{\phi}_i\rangle + E[\{\phi\},\{\mathbf{R}\}] \tag{6.32}$$

The equations of motion 6.31 can be integrated using standard techniques such as the (velocity) Verlet algorithm. Similar to the nuclear degrees of freedom, the fictitious electronic mass μ determines the time evolution of the electronic degrees of freedom within a certain energy range, which slightly exceeds the electronic ground state (the Born-Oppenheimer surface). Specifically, the ratio of M_I to μ signifies the relative velocity at which the electronic degrees of freedom propagate compared to the nuclear positions. When μ is much smaller than M_I, the resulting dynamics is adiabatic, as the electronic orbitals promptly adjust to changes in the nuclear positions. However, to maintain the desired adiabatic energy-scale separation between the electronic and nuclear degrees of freedom, the highest frequency of nuclear motion, denoted as ω_I^{max}, must be significantly smaller than the lowest frequency associated with the fictitious motion of the electronic degrees of freedom, denoted as ω_i^{min}. The latter has been demonstrated to exhibit behavior similar to

$$\omega_i^{min} \propto \sqrt{\frac{E_g}{\mu}} \tag{6.33}$$

with E_g the energy (HOMO-LUMO) gap of the system. In case of finite gap, μ can be tuned in order to comply with $\omega_I^{max} \ll \omega_i^{min}$ so that no energy transfer occurs between the electronic and nuclear degrees of freedom. For metallic systems, however, subsequent adaptations of the original CPMD method are required.[48,195]

In what follows, I will present the results obtained from the application of the AIMD methods discussed in this chapter, particularly when combined with Minnesota density functionals for the study of liquid water.

Part IV

Navigating the potential energy surface with artificial intelligence approaches

7 Structure and dynamics of liquid water from ab initio simulations: Adding Minnesota density functionals to Jacob's ladder

Experiment is the only means of knowledge at our disposal.
Everything else is poetry, imagination.
— Max Planck

Chapter 7 is a preprint version of an article in preparation:

Villard, J.; Rothlisberger, U. Structure and dynamics of liquid water from ab initio simulations: Adding Minnesota density functionals to Jacob's ladder. *In preparation* 2023.

7.1 Abstract

The accurate representation of the structural and dynamical properties of water is essential for simulating the unique behavior of this ubiquitous solvent. In this study, we assess the current status of describing liquid water using ab initio molecular dynamics, with a special focus on the performance of all the later generation Minnesota functionals. Findings are contextualized within the current knowledge on DFT for describing bulk water under ambient conditions and compared to experimental data. We find that, contrary to the prevalent idea that local and semilocal functionals overstructure water and underestimate dynamical properties, M06-L, revM06-L, and M11-L understructure water, while MN12-L and MN15-L overdistance water molecules due to weak cohesive effects. This can be attributed to a weakening of the hydrogen bond network, which leads to dynamical fingerprints that are over fast. While most of the hybrid Minnesota functionals (M06, M08-HX, M08-SO, M11, MN12-SX, and MN15) also yield understructured water, their dynamical properties generally improve over their semilocal counterparts. It emerges that exact exchange is a crucial component for accurately describing hydrogen bonds, which ultimately leads to corrections in both the dynamical and structural properties. However, an excessive amount of exact exchange strengthens hydrogen bonds and causes overstructuring and slow dynamics (M06-HF). As a compromise, M06-2X is the best performing Minnesota functional for water, and its D3 corrected variant shows very good structural agreement. From previous studies considering nuclear quantum effects (NQEs), the hybrid revPBE0-D3, and the rung-5 RPA (RPA@PBE) have been identified as the only two approximations that closely agree with experiments. Our results suggest that the M06-2X(-D3) functionals have the potential to further improve the reproduction of experimental properties when incorporating NQEs through path integral approaches. This work provides further proof that accurate modeling of water interactions requires the inclusion of both exact exchange and balanced (non-local) correlation, highlighting the need for higher rungs on Jacob's ladder to achieve predictive simulations of complex biological systems in aqueous environments.

7.2 Introduction

Liquid water is a ubiquitous and essential component of life, playing a critical role in a wide variety of chemical and biological processes.[375–378] A comprehensive understanding of water at the atomic scale is vital for advancing research in diverse domains such as aqueous chemistry,[379–382] biochemistry,[383,384] atmospheric science,[385,386] and environmental engineering.[387,388] Furthermore, unraveling the intricate behavior of water molecules enables deeper insights into solvation dynamics,[389,390] water-materials interactions,[391,392] protein folding,[393–396] enzymatic reactions,[397,398] and the properties of biological membranes,[399,400] ultimately contributing to the development of innova-

tive technologies and therapeutics. Deceivingly simply at first sight, it is well known that liquid water shows anomalous properties that have been extensively observed and documented like the density anomaly,[376,401,402] high heat capacity,[384,399,403] high boiling and melting points,[377,404] high surface tension,[405,406] high dielectric constant,[407,408] and high viscosity.[409,410] Despite substantial advances in the understanding of the behavior of water, the origins of these anomalies are not yet entirely elucidated neither by experiments nor theory, although it has been widely recognized that the structural characteristics of the hydrogen bond network under thermal fluctuations play a pivotal role for these unique features.[378,411–415]

Significant challenges exist in conclusively capturing atomic-scale phenomena in water through experiments like NMR,[416–426] IR,[427] X-ray[428,429] or neutron[429–432] spectroscopy for which measurement interpretations often rely on theoretical models. Although a variety of computationally efficient and relatively accurate empirical force fields have been developed,[33,433–436] those remain intrinsically incapable of describing bond breaking in chemical reactions. Therefore, the quantitative understanding of condensed phase water, in particular its reactivity, and role as a universal solvent can only fully emerge from the development of accurate ab initio molecular dynamics (AIMD) simulations.[437–439] These simulations need to faithfully represent both electronic reorganization and nuclear quantum effects (NQEs) associated with hydrogen bonding but, at present, such a comprehensive predictive quantum picture at ambient conditions remains quite elusive. In addition, the cost of most accurate wavefunction-based approaches such as post-Hartree-Fock[17,18] (e.g., MP2[21] or RPA[251,440–442]), coupled cluster (CCSD(T)),[24,443] or configuration interaction (CI)[3,444] hinders their potential application across the entire water phase diagram.

Balancing accuracy and computational feasibility, Kohn-Sham (KS)[140] density functional theory (DFT)[135] has become the go-to quantum-chemical method for time propagation of molecular systems and computation of statistical averages when combined with molecular dynamics (MD) or Monte-Carlo (MC) engines.[33,195] Although the ground-state energy and electron density are formally exact within DFT, their universal mapping remains unknown, necessitating the use of approximations in the KS formalism. In this approach, many-body interactions are accounted for and incorporated into the approximate exchange-correlation (XC) functional.

Over the past several decades, hundreds of XC functionals have been developed with the aim to capture all relevant physics and achieve chemical accuracy over a broad range of molecules, materials, and organometallic systems. To classify the growing number of functionals, John P. Perdew proposed a hierarchy called *Jacob's ladder*,[142] which organizes functionals based on their complexity. The ladder consists of five rungs: (1) Local Density Approximations (LDA) depend only on the electron density at a given point in space and offer computational efficiency but often lack accuracy.[150,445] (2) Generalized Gradient Approximations (GGA) functionals incorporate the (local)

electron density and its gradient.[446–448] (3) Meta-GGA functionals account for the electron density, its gradient, and the kinetic energy density.[149,449–452] (4) Hybrid functionals mix a portion of Hartree-Fock (HF) exact exchange with XC terms from DFT,[158,161,162,453,454] and (5) Double-hybrid and RPA-based functionals (rung-5), the highest rung on the ladder, combine a hybrid functional with post-HF correlation corrections, e.g., within second-order perturbation theory (MP2)[21,56,165–170] or non-local correlation based on the random phase approximation (RPA).[130,171–174] As one moves up the ladder, the functionals globally tend to provide better descriptions of electronic interactions and improve the overall predictive accuracy.[101,141,142,146,455,456] However, this comes at the price of an increasingly higher computational cost: for instance, the cost of hybrids is roughly two orders of magnitude the one of GGA functionals.[45,101,145,457]

While DFT has demonstrated impressive success in the examination of structures, properties, and reactivities for a wide range of molecules and materials, the prominent challenge persists in identifying the appropriate XC functional for a specific problem, as the performance of a functional can vary significantly depending on the system under study.[458] For liquid water, no local (LDA) or semilocal (GGA, meta-GGA) DFT simulation has yet achieved a conclusive replication of experimental observations, covering both structural and dynamical properties simultaneously. For example, it was established that most of the GGA functionals, like PBE[448] and BLYP,[446,447] provide overstructured oxygen-oxygen pair correlation functions, and dynamical figures that are too slow, therefore not completely remedying the glassy behavior observed with the LDA.[183,459–462] Furthermore, GGA and (even) hybrid levels can underestimate the equilibrium density of liquid water, leading to the incorrect prediction that ice sinks in water.[415,463]

DFT approximations encounter difficulties when describing condensed water due to the intricate nature of concurrent competing interactions that are involved in covalent bonds, hydrogen bonds, and van der Waals (vdW) forces. Hydrogen bonds, though one order of magnitude weaker than intramolecular O-H covalent bonds, remain locally strong and directionally attractive. Another order of magnitude weaker, vdW dispersion forces play a non-negligible role at larger distances, with an attractive and isotropic character. The interplay between varying interaction strengths, length scales, and directionalities makes water a highly sensitive test system for the design and assessment of XC functionals. Indeed, even slight imprecision in the XC description is likely to disrupt the complex balance of interactions, ultimately impacting the H-bond network that is responsible for many of water's properties.[273,462]

While local and semilocal functionals fail to capture intermediate to long-range vdW forces,[462] AIMD simulations have demonstrated that GGAs enhanced with vdW representations typically lead to a softer structure of bulk water, accompanied by increased mobility that aligns more closely with experimental measurements.[183] This improve-

ment is achieved by incorporating dispersion-corrected atom-centered potentials (DCACPs),[181–183] empirical dispersion corrections (e.g., Grimme's D2[184] and D3,[147] or non-local correlation terms (e.g., (r)VV10,[185–187] vdW-DF,[188] TS-vdW[189,190]). However, the performance of such corrections relies on the original GGA to which the combination may not always improve, or may even deteriorate properties.[183,191] Other studies pointed out the necessity of including a fraction of exact exchange, thus resorting to rung-4 hybrids, to effectively describe hydrogen bonding but without reaching a perfect agreement with the experiment.[344,462,464–467]

Altogether, attaining a reliable description of the structural and dynamical properties of liquid water through lower rung (1-3) DFT models remains an issue. The goal of this work is consequently to contribute further understanding to this endeavor by incorporating the popular Minnesota density functionals[61,100,451,468–475] into the array of approximations tested on water at ambient conditions. While having demonstrated success for molecular systems, previous investigations of the performance of Minnesota functionals on condensed water are, to our knowledge, limited to the work of Del Ben et al. who ran MC simulations on water with the M06-L-D3, M06-D3, and M06-2X-D3 functionals,[144] and the work of Pestana et al. that focuses on MD with M06-L-D3.[467] Our work thus fills a gap in the evaluation of the performance of DFT functionals for liquid water. Gaining insights from the performance of various functionals not only helps demystify their promise and limitations for water, but also on a wider range of systems exhibiting a similar delicate balance of interactions such as e.g., in large biomolecules,[476,477] heterogeneous catalysts,[478,479] aqueous solutions[381,480,481] and molecules on surfaces.[482] For this reason, we have made an effort, albeit not exhaustive, to compile in this document previously calculated quantities from DFT-based MC and AIMD. Our aim is to establish a common ground for comparing various studies found in the literature and confront them with experimental measurements.

Information on higher-rung approximations, such as double-hybrids, is limited in this assessment due to their exorbitant computational overhead and infrequent implementation in MD software packages.[144] The substantial cost of hybrid functionals also poses a significant challenge for obtaining extensive results in MD simulations,[465,483] in particular in the context of plane wave based approaches. To tackle this issue, the emergence of machine learning (ML)-based interaction models has shown the potential to attain a similar level of accuracy at a fraction of the cost.[180,484,485] Nevertheless, the effectiveness of such ML potentials primarily depends on their reliability across the entire phase (configurational) space sampled during MD (MC) simulations

Hereafter, we present structural properties (in terms of radial distribution functions, coordination numbers, density, number of H-bonds and angular distributions) and dynamical characteristics (quantified via diffusion coefficients and rotational correlation times) obtained with AIMD and all the later generation Minnesota functionals. Those include some of the most employed meta-GGAs and hybrid meta-GGAs in

computational chemistry.[101,304,486] Meta-GGAs are investigated with Car-Parrinello MD (CPMD), while the much more computationally expensive hybrid meta-GGAs have been run with Born-Oppenheimer MD (BOMD), thanks to the crucial acceleration of a recent ML-aided multiple time step scheme that preserves the target DFT level description by construction.[45,99,487] Both CPMD and BOMD employ classical propagation of nuclei; however, capturing a comprehensive picture of water including nuclear quantum effects (NQEs) requires more sophisticated and considerably costlier (approximately two orders of magnitude[457]) ab initio path integral MD (PIMD) approaches.[33,488] Alternatively, NQEs can be qualitatively evaluated based on very recent studies that employ DFT/ML-based PIMD methods.[414,484,485] This allows an identification of the most promising XC functionals worth further investigation in conjunction with quantum nuclei.

In this regard, this chapter provides benchmarks for the widely-used Minnesota density functionals in simulating liquid water, and places them in the context of existing knowledge of other DFT approximations as well as experimental measurements. This will hopefully assist the scientific community in utilizing, refining, or developing more accurate and transferable XC functionals.

7.3 Theory and methods

7.3.1 Minnesota density functionals

Since 2005, the Minnesota theoretical chemistry group led by Donald Truhlar has focused on the development of post-GGA functionals capable of capturing the chemistry of main group elements as well as transition metals including activation barriers as well as non-covalent interactions. The excellent "across-the-board" performance of these functionals has made them one of the most widely used XC approximations in computational chemistry.[101,304,381,486] The Minnesota functionals are semi-empirical in nature, with functional forms that have been fitted against extensive datasets of reference absolute and relative energies, as well as eventual structures and lattice constants. For brevity's sake, Table 7.1 provides a summary of the XC approximations studied in this work, along with a global overview of their functional components. Interested readers are referred to the corresponding references for more technical and mathematical details.

The generation of the 2006 functionals was ingeniously crafted by merging the characteristics of the earlier M05[492] and VSXC[493] functionals (in turn designed from modifications of the PBE and LSDA functionals for the exchange). These include M06, a versatile hybrid meta-functional that boasts consistent accuracy for main group thermochemistry, barrier heights, medium-range correlation energies, and transition metals. M06-2X, another hybrid meta-GGA, excels in main group chemistry and bar-

Table 7.1: Overview of some characteristic features of Minnesota density functionals, in terms of $E_{xc} = (X/100)E_x^{\mathrm{HF}} + (1 - X/100)E_x^{\mathrm{DFT}} + E_c^{\mathrm{DFT}}$. X is the percentage of exact exchange in the functional. E_x^{DFT} and E_c^{DFT} depicts the origins of the functional form for the exchange (e.g., exchange energy density, correction factors) and the correlation (e.g., correlation energy density, gradient correction)[b]. Also listed are the number of fitted parameters # as well as the satisfaction (✓) or not (×) of the uniform electron gas (UEG) limit.

Functional	Class[a]	X [%]	E_x^{DFT}	E_c^{DFT}	#	UEG	Ref.
			Meta-GGA				
M06-L	L meta-GGA	0	M05+VSXC	M05+VSXC	34	✓	[451]
revM06-L	L meta-GGA	0	M05+VSXC	M05+VSXC	31	x only	[468]
M11-L	RSL meta-GGA	0	SR/LR: LSDA(PBE+RPBE)	LSDA+PBE	44	✓	[469]
MN12-L	L meta-NGA	0	N12	N12+(LSDA+PBE)	58	×	[470]
MN15-L	L meta-NGA	0	N12	N12+(LSDA+PBE)	58	×	[471]
			Hybrid meta-GGA				
M06	GH meta-GGA	27	M05+VSXC	M05+VSXC	33	✓	[61]
M06-HF	GH meta-GGA	100	M05+VSXC	M05+VSXC	32	✓	[100]
M06-2X	GH meta-GGA	54	M05	M05+VSXC	29	✓	[61]
M08-HX	GH meta-GGA	52.23	LSDA(PBE+RPBE)	LSDA+PBE	47	✓	[472]
M08-SO	GH meta-GGA	56.79	LSDA(PBE+RPBE)	LSDA+PBE	44	✓	[472]
M11	RSH meta-GGA	42.8-100	SR: LSDA(PBE+RPBE)	LSDA+PBE	40	✓	[473]
MN12-SX	RSH meta-NGA	25-0	N12	N12+(LSDA+PBE)	58	×	[474]
MN15	GH meta-NGA	44	N12	N12+(LSDA+PBE)	59	×	[475]

[a] L stands for local, RSL for range-separated local, GH for global hybrid, RSH for range-separated hybrid and NGA for non-separable gradient approximation.
[b] SR stands for short-range, LR for long-range. LSDA is the local spin density approximation,[138,489] PBE the Perdew, Burke, Ernzerhof functional,[448] RPBE the secondly revised PBE functional,[490] and N12 Truhlar's non-separable density gradient functional.[491]

rier heights, accurately predicts valence and Rydberg electronic excitation energies, and π-π stacking interactions, while its performance falters in the realm of transition metals. M06-L, a local functional devoid of Hartree-Fock exchange, was skillfully tailored as a cost-effective choice for numerous demanding applications associated with extensive systems. It excels for transition metals, yet its accuracy for barrier heights does not match that of M06 and M06-2X. Finally, M06-HF was designed primarily for spectroscopy, demonstrating good performance for valence, Rydberg, and charge transfer excited states with little compromise on ground-state accuracy. An important point to note is that M06-2X and M06-HF that differ in the amount of exact exchange (54 vs 100 %) share the same training set, which was expanded with transition metals with respect to the one used for the parameterization of the M06 functional. revM06-L, on the other hand, was developed later using an even larger database and additional smoothness restraints to ensure better numerical stability, smoother potential energy curves, and overall improved accuracy compared to M06-L.

The next generation functionals M08-HX and M08-SO resulted from exploring a more flexible functional form, with different formal constraints; while M08-SO respects the exact gradient expansion for slowly varying density up to the second order (SO) and the uniform electron gas (UEG) limit, M08-HX only respects the latter. Both functionals of the M08 generation were found to modestly improve on M06-2X for main-group

thermochemistry, kinetics, and non-covalent interactions. The even more recent M11, on the other hand, is a range-separated version[494] of the M08 functionals, with the same correlation component. The percentage of HF exchange of 100% at large inter-electronic distance reduces to 42.8% at short range. The second-order density gradient expansion is also correct by construction in M11, and good across-the-board accuracy was shown thanks to the use of a further extended training set. A bit later, the M11-L functional was designed as the local analogue of M11, mainly for cost-efficiency and improved accuracy for multi-reference systems. M11-L replaces the exact exchange by a long-range meta-GGA exchange functional, that has different spatial extent and parameters than the exchange at short range.

In 2012, a new functional form called N12 was developed that pushes the limits of local functionals, providing simultaneous accuracy on energetic and structural properties of both solids and molecules.[491] Unlike traditional GGAs, the N12 functional is a non-separable approximation (NGA) between the density and its (reduced) gradient that embodies both exchange and correlation effects, and can be seen as a generalization of the dual-range M11-L. By adding a dependence on the kinetic energy density, and the M08/M11 correlation term, Peverati and Truhlar designed the MN12-L meta-non-separable gradient approximation to obtain even broader accuracy with a local functional. With the inclusion of 25% of short-range exact exchange (that is screened at large distances), the MN12-SX functional yields better results than MN12-L for most chemical properties, and is notably more successful in calculating semiconductor band gaps.[474] Finally, re-optimization of MN12-L using a larger training database and additional smoothness restraints on the functional form resulted in the most recent MN15-L local meta-NGA functional. This latter shows better performance for transition metals and is generally recommended over MN12-L.[471] Its hybrid version, called MN15, was trained using a combination of single-reference chemical data (barrier heights), as well as diverse multi-reference transition-metal bond energies and atomic excitation energies that are challenging to describe with KS-DFT. As a result, it provides broad accuracy for both multi-reference and single-reference systems, and at the same time has demonstrated outstanding performance in describing noncovalent interactions.[475]

7.3.2 Simulations

AIMD simulations were carried out using the CPMD code[47] with PBE Trouiller-Martins norm-conserving pseudopotentials.[495] The wavefunction cutoff energy was set to 80 Ry for all systems. We used a finer integration mesh with a density cutoff energy set to 640 Ry (dual of 8) to ensure proper convergence of the Minnesota functionals with planes waves,[304] therefore affecting the usual computational cost by a factor of 2. The convergence threshold for the DIIS[266] wavefunction optimization was set to 10^{-6} a.u. on the residual gradient on occupied orbitals, except for the M06 functional that is

harder to converge to such a low criterion and for which $5 \cdot 10^{-6}$ a.u. was used instead.

Meta-GGA functionals

For the simulations with meta-GGAs, systems use a cubic 12.445^3 Å^3 periodic box of 64 water molecules corresponding to a density of $\sim$1 g/cm^3, simulated via Car-Parrinello MD. All hydrogens were assigned the mass of deuterium to increase the integration time step. The wavefunction fictitious mass is chosen to be 800 a.u., and $\delta t = 3.5$ a.u. is the default time step that we reduced in case of energy exchange between the ionic and fictitious degrees of freedom to keep the Hamiltonian energy conserved.

A first equilibration phase was performed for each functional. Starting with a pre-equilibrated structure at the classical level, systems were first heated up to 400 K with velocity rescaling for about 1 ps until reaching a stable average temperature. Then, systems were cooled down to 330 K during another picosecond, and the temperature was again decreased more slowly to 300 K during a time interval of about half a picosecond.

After the first initial equilibration, systems were further thermalized with a Nosé-Hoover thermostat on the ions at 300 K for several picoseconds with a coupling frequency of 1500 cm^{-1} before finally switching to the NVE ensemble for the production runs for at least 10 ps. Configurations were saved every 50 steps for analysis. More information about the lengths of the trajectories, time steps and energy conservation are reported in Table C1.

Hybrid meta-GGA functionals

Due to their high computational cost, the AIMD simulations with hybrid functionals were performed with a smaller cubic box of dimensions 9.939^3 Å^3 containing 32 water molecules. A multiple time step (MTS) scheme[343,345,487] was used to further accelerate the simulations, with an inner time step of $\delta t = 15$ a.u. and an outer time step of $\Delta t = n \cdot \delta t$, where the time step ratio n is chosen to maintain sufficient energy conservation. At inner time steps, fast force components are given by a delta-ML model that predicts PBE0 forces based on the LDA ($\mathbf{F}^{\text{inner}} = \mathbf{F}^{\text{LDA}} + \Delta\mathbf{F}_{ML}^{\text{PBE0-LDA}}$), while total forces are corrected at the outer time step with their slow components ($\mathbf{F}^{\text{outer}} = \mathbf{F}^{\text{Minnesota}} - \mathbf{F}^{\text{inner}}$) to fully recover the higher-level Minnesota forces.[45,99] In this approach, ML serves only as a low-level surrogate operating on shorter timescales without impacting the target DFT level. Note that the inner PBE0 level does not need to match the outer Minnesota level entirely, but should be close enough so that their difference slowly varies in time and dynamically decouples from fast force components. Ultimately, the Minnesota level is recovered at larger physical time steps by construction, ensuring that the structural and dynamical properties are not affected,[45,496] unlike in ML-potential MD.

The OQML[497,498] kernel method is used to infer force differences $\Delta\mathbf{F}_{ML}^{\text{PBE0-LDA}}$ from the aSLATM[499] representations of chemical environments. The training set for the OQML model was generated by running PBE0 trajectories on condensed water and small water clusters. Both energies and forces were used in the training. The model demonstrated an out-of-sample mean absolute error of around 0.3 kcal/(mol Å) on $|\Delta\mathbf{F}_{ML}^{\text{PBE0-LDA}}|$, as well as a mean absolute error of 0.7 degrees on force directions, based on a test set of 4800 atomic forces.

Starting from a PBE0 pre-equilibrated configuration, all systems were first thermalized in the NVT ensemble with the ML-MTS acceleration method and a Nosé-Hoover thermostat with a coupling frequency of 1500 cm^{-1} at 300 K for at least 5 ps. After this initial equilibration process, NVE runs were conducted during the production phase, sampling configurations for at least 6 ps. The lengths of the trajectories, time step ratios, and energy conservation are reported in Table C1.

7.3.3 Analysis

Here, we provide information on how the properties were calculated from AIMD trajectories. As the production runs were conducted in the NVE ensemble, the average temperature of each simulation slightly differs. To ensure comparability, care was taken to renormalize the properties either by considering temperature or box volume differences.

We note that the average structural properties are similar in the NVT and NVE ensembles.[496] Additionally, the replacement of hydrogen atoms with deuterium has little effect on structural properties when the ionic motion is treated classically.[459,464] However, the use of deuterated water can affect dynamical properties, such as the diffusion coefficient. Therefore, it is important to rely on heavy water data when validating D_2O simulations against experimental results.

Radial distribution functions and coordination number

Radial distribution functions (RDFs) were computed using the VMD software,[500] accounting for periodic boundary conditions and a bin width of 0.01 Å. The RDFs are then smoothed by interpolation for integration and visualization purposes with negligible differences when compared to the original statistical averages. The coordination numbers n_{OO} of water molecules is obtained as the oxygen-oxygen (O-O) coordination number resulting from the integral of the O-O RDF g_{OO}:[428]

$$n_{\text{OO}} = 4\pi\bar{\rho}\int_0^{r^*_{\text{min}}} r^2 g_{\text{OO}}(r)dr \tag{7.1}$$

where $\bar{\rho}$ is the molecular number density. For consistency with experimental reference data, the value of r^*_{min} is set as the position of the first minimum in the actual integrand $r^2 g_{\text{OO}}(r)$, rather than the first minimum of $g_{\text{OO}}(r)$. For comparison, we also report the coordination number $\bar{n}_{\text{OO}}$ calculated up to the first minimum of $g_{\text{OO}}(r)$ in Table C3.

Density of liquid water

The equilibrium density predicted by the Minnesota functionals is estimated by scanning over volume changes around trajectory snapshots.[460] For each snapshot, total energies are calculated at scaled values of the lattice constant. The intramolecular coordinates are held fixed while the positions of the centers of mass of the water molecules are rescaled to scan over volume reductions and expansions. The equilibrium volume and density are determined by calculating the minimum of the interpolated energy values, at a given snapshot. 30 snapshots were used, each separated by 0.2 ps, to obtain a representative set of configurations. The density is calculated from the average equilibrium volume over all snapshots. Given that the basis set size varies with the volume of the box in plane wave basis sets, a larger wavefunction cutoff energy of 200 Ry was used to ensure reliable energy differences from these calculations.

H-bond number and angular distributions

The number of hydrogen bonds is evaluated from geometrical criteria following refs [501] and [183]. A polynomial function is defined by

$$f(d) = \frac{1 - [(d - d_0)/\Delta]^n}{1 - [(d - d_0)/\Delta]^m} \tag{7.2}$$

where $d_0 = 2.8$, $\Delta = 0.45$, $n = 10$ and $m = 16$. $f(d)$ increases to 1 from $d = 2.3$ up to 2.8 Å, and decreases rapidly near 3.4 Å. These values correspond to the first experimental maximum and minimum of the O-O RDF such that $f(\overline{\text{O}_i\text{O}_j})$ encodes the $\overline{\text{O}_i\text{O}_j}$ distance between two molecules i and j. A second function is used to model the decrease in the probability of hydrogen bonding as the total distance $\overline{\text{O}_i\text{H}} + \overline{\text{HO}_j}$ (between the donor oxygen O_i and its covalently-bound hydrogen H, and its distance $\overline{\text{H}\ldots\text{O}_j}$ to the corresponding acceptor oxygen O_j) increases. This latter metric increases either when the donor-hydrogen direction is tilted or when the donor and acceptor are far away. Thus, the second function used is $f(\overline{\text{O}_i\text{H}} + \overline{\text{HO}_j} - \overline{\text{O}_i\text{O}_j})$ with $d_0 = 0$, $\Delta = 0.4$, $n = 4$ and $m = 8$, that equals 1 at 0 Å, and rapidly decays to 0 when the argument exceeds 0.5 Å. An H-bond is therefore counted if the product of the two functions exceeds 0.5, and not otherwise. The presence or absence of an H-bond is facilitated because the product of these analytical functions is predominantly either close to 0 or to 1. In practice, it is defined whether H is covalently bound to either molecule i or j in order to ensure the correct counting with periodic boundary conditions. We have observed, like others,[501]

that this counting is qualitatively comparable to conventional criteria that involve both the $\overline{O_iO_j}$ distance and the angle between the O_iO_j and O_iH directions.[464]

To compute the H-bond angular distributions, we took into account all the molecules present in the first coordination shell of the reference molecule, i.e. we restricted our analysis to angles for which the donor-acceptor distance is less than 3.4 Å, and the hydrogen-acceptor distance is less than 2.5 Å, based on the experimental RDFs.[428,429]

Diffusion coefficient

The self-diffusion coefficient D_L is calculated from the Einstein relation

$$D_L = \frac{1}{6} \lim_{t\to\infty} \frac{d}{dt} \frac{1}{N} \sum_{i=1}^{N} \langle |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \rangle \tag{7.3}$$

where N is the number of water molecules, $\mathbf{r}_i(t)$ the position of each oxygen atom i at time t, and the brackets indicate an average over the NVE ensemble. Improved statistics were gathered across multiple lag times and time origins according to the default parameters of the Diffusion Coefficient Tool plugin[502] for VMD[500] to finally obtain D_L from the average slope of the mean-squared displacement (MSD).

Since D_L is calculated from the simulation of a L^3 cubic water box, finite size effects are corrected via[503]

$$D_\infty = D_L + \xi \frac{k_B T}{6\pi\eta L} \tag{7.4}$$

where D_∞ is the infinite-size limit, $\xi = 2.837297$, k_B the Boltzmann constant, and η the shear viscosity of the fluid at average temperature T. The viscosity η predicted by each functional approximation is generally not known, and relying on the experimental value[504] was observed not to significantly affect the rescaling of D_L to D_∞.[503] In this regard, theoretical viscosities were computationally derived for SCAN and optB88-vdw.[414] We observed negligible deviations in D_∞ when calculated using either experimental or theoretical viscosities (Table C4). However, if the functionals are too overstructured, they may predict a larger viscosity, leading to an overestimation of D_∞ with respect to the experimental (lower) viscosity. Another reliable approach to compare with experiment is to rescale the experimental coefficients D_∞^{exp} back to D_L^{exp}, which is the hypothetical experimental value for a box of size L.[505]

Orientational correlation times

In addition to the translational motion, the rotational time scale of the water molecules is determined by analyzing the orientational auto-correlation function:

$$C_n(t) = \frac{1}{N}\sum_{i=1}^{N}\langle P_n[\hat{\mathbf{u}}_i(0)\cdot\hat{\mathbf{u}}_i(t)]\rangle \tag{7.5}$$

where P_n is the Legendre polynomial of order $n = 1, 2$ and $\hat{\mathbf{u}}_i$ is the molecular unit vector along either the OH covalent bonds, the HH intramolecular direction, or the direction of the dipole moment μ. The rotational correlation times $\tau_{1,2}$ were determined by fitting the curves $C_{n=1,2}(t)$ with the function $Ae^{-t/\tau_{1,2}}$ in the exponential regime following the initial subpicosecond decay, which is due to the librational motion of the water molecules.[506,507] These relaxation times have been found to be less affected by finite-size effects compared to the self-diffusion coefficient,[508,509] and are of interest because they can be measured experimentally using techniques such as NMR[376,416,417,419,420,424,426,510,511] or IR[512,513] spectroscopy.

7.4 Results and discussion

7.4.1 Structural properties

Radial distribution functions

The radial pair distribution functions (RDFs in terms of g_{OO}, g_{OH}, g_{HH}) provide structural information as modelled by the different Minnesota functionals. In Figure 7.1, we compare respectively the O-O, O-H and H-H RDFs to experimental references. The left panel reports the results of meta-GGAs. Clearly, the O-O RDFs indicate that M06-L and M11-L are understructured, with first g_{OO} minima that are too high and too far. These two functionals also behave alike when it comes to the O-H and H-H distributions that are slightly understructured compared to experiment. Although yielding similar RDFs, it is interesting to recall that M06-L and M11-L do not share the same exchange and correlation functional forms as well as training data, and that M11-L is range-separated (Table 7.1). However, both fulfill the UEG limit. For the remaining functionals where this constraint is lifted (revM06-L, MN12-L, MN15-L), the O-H and H-H RDFs move even further away from the experiment and no longer capture the hydrogen-bond network as shown by the smearing out of the second peak in the g_{OH} distributions. The revM06-L functional has the same XC form as M06-L, but differs in the imposed constraints and training data. In contrast to the others, the MN12-L and MN15-L non-separable functionals result in an overstructured g_{OO} but again lack exactness in the intermolecular distances[514] with typical shifts in the location of the first minimum up to 1 Å. MN15-L was designed from a re-optimization of MN12-L using a larger

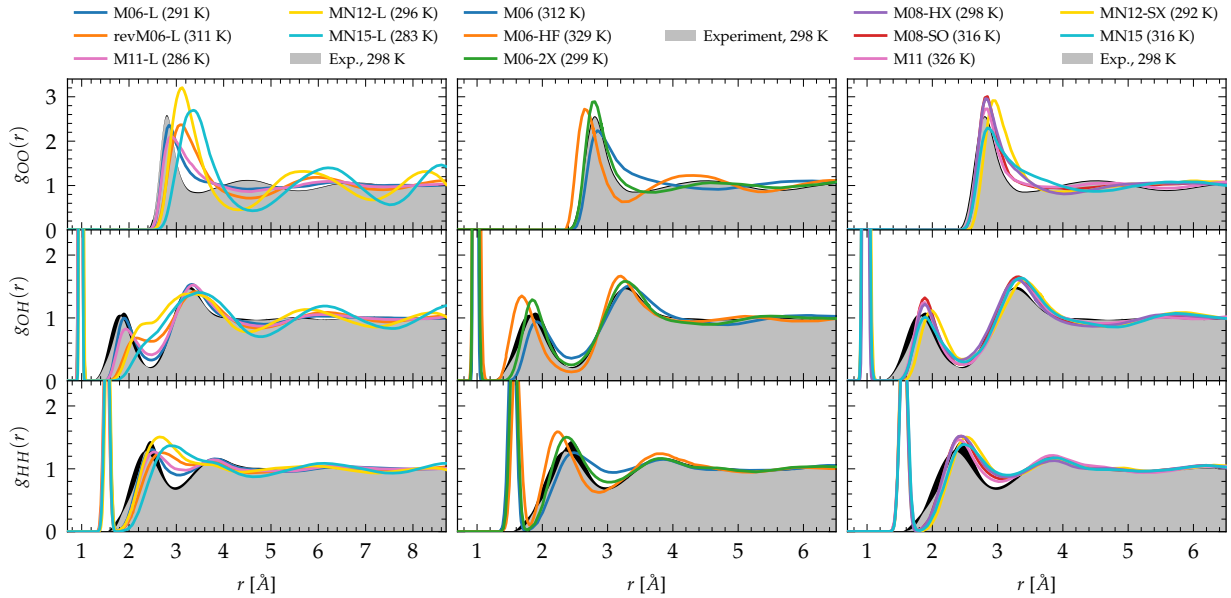


Figure 7.1: Oxygen-oxygen (g_{OO}), oxygen-hydrogen (g_{OH}) and hydrogen-hydrogen (g_{HH}) radial distribution functions of liquid water predicted by Minnesota density functionals. The experimental reference for g_{OO} comes from X-ray diffraction[428,429] interpolated at 298 K[509] and joint X-ray/neutron diffraction experiments were used for g_{OH} and g_{HH}.[432] Black areas represent experimental uncertainties.

database and additional smoothness restraints on the functional form. Therefore, the RDF similarities between M06-L and M11-L (different forms, different training data) and differences between M06-L/revM06-L and MN12-L/MN15-L (same form, different constraints, different training data) would advocate for a larger sensitivity of semi-empirical meta-GGAs to exact constraints rather than training data. Consistent with this, the additional smoothness restraints in revM06-L (versus M06-L), and MN15-L (versus MN12-L) seem to reduce the packing of water molecules and shift the first peak of the O-O RDF to larger distance. Overall, no local meta-GGA Minnesota functional is providing an accurate reproduction of the structure of liquid water, mainly due to failures in the description at intermediate and long-range intermolecular distances.

The hybrid functionals of the M06 family are shown in the center panel of Figure 7.1. Interestingly, M06 predicts RDFs that are very similar to its M06-L sister. M06-L therefore appears as a good local functional fit for the 27%-hybrid M06, but both fail at reproducing the intermolecular structure of water at long range[476] In contrast, the increase of exact exchange to 100% in M06-HF noticeably over-accentuates the structure and shifts the first and second g_{OO} peaks to too short intermolecular distances. This increased cohesive effect that was missing for the local functionals is also observed in the g_{OH} and g_{HH} RDFs.

As a compromise between M06 and M06-HF, M06-2X, with 54% of exact exchange, re-

markably improves the agreement of the RDFs with experiments. Despite the first minimum of g_{OO} being a bit right-shifted by ∼0.3 Å, M06-2X shows better peak positions and an improved second coordination shell according to the second peak in g_{OO}. As observed, the agreement with experimental data is not a trivial matter, as the structure of water is the result of the complex interplay between covalent bonds, hydrogen bonds, and vdW interactions. Many-body effects among hydrogen-bonded water molecules can be observed in the first peak of g_{OO} and the second peak of g_{OH}. The region between the first and second peaks of g_{OO} mainly consists of non-hydrogen-bonded water molecules that occupy the intershell space between the hydrogen-bonded neighbors. The increased number of water molecules in these intershell regions can partly be attributed to the attractive, non-directional vdW interactions.[415,467] Therefore, achieving a balance between exact exchange and vdW dispersion at an intermediate length scale is essential for accurately reproducing the densely packed and disordered structure in the intershell regions. As demonstrated by the RDFs, M06-2X captures these correlations with the highest accuracy and is thus capable of describing both hydrogen bonding and dispersion effects. M06-2X was specifically designed with the absence of transition metals in its training set, the M06-2X focuses on the description of the electron correlation of the main group elements which could be one of the reasons why it performs so well on water compared to M06 for which transition metals were included. M06-HF lacks an adequate amount of correlation to counterbalance the full HF exchange: The second coordination shell has a higher population of water molecules that are not sufficiently drawn out to the intershell region by vdW forces.

The newer generation Minnesota hybrid functionals do not improve the structural description any further (Figure 7.1, right panel). While possessing nearly the same amount of exact exchange as M06-2X, the new functional form introduced in M08-HX (52%) and M08-SO (57%) does not outperform M06-2X. MN12-SX is both range-separated and non-separable, with 25% of exact exchange at short range that decreases to 0% at long range. This functional has the lowest proportion of exact exchange. Notably, it is also the one where the first g_{OO} peak and the second g_{OH} peak are shifted to the right, i.e. to longer intermolecular distances, presumably due to an elongation (weakening) of the intermolecular hydrogen bonds, or a lack of vdW cohesive forces[514] (the analysis of the dynamical properties in Section 7.4.2 confirms the second hypothesis). In general, it is observed that the inclusion of a fraction of exact exchange leads to clearly visible improvements in the g_{OH} and g_{HH} RDFs over local functionals, and addition of the right amount of exact exchange can also better the agreement for g_{OO}. This is particularly the case for M11, MN12-SX and MN15 that improve the second peak of g_{OH} significantly over M11-L, MN12-L and MN15-L, respectively. Moreover, although not perfect, these functionals clearly ameliorate the position and shape of the first g_{OO} peak compared to their local counterparts. Consequently, this emphasizes the crucial importance of exact exchange in accurately describing the hydrogen bond network in general, supporting the notion that hybrid functionals and higher rungs of Jacob's

ladder are indeed the most accurate approaches for depicting complex interactions with KS-DFT.

To evaluate the performance of Minnesota functionals in the broader context of DFT approximations, we compiled a comprehensive dataset from the literature (Table C2). As various functionals were employed at different temperatures, the position and height of the first g_{OO} peak, as well as the first g_{OO} minimum, were rescaled to a common reference point at 298 K based on empirical interpolations fitted to experimental data (Figure C1). The differences between simulated and experimental values are depicted in Figure 7.2. As can be seen, KS-DFT coupled to a classical propagation of the nuclei have the tendency to generally overestimate the height of the first peak and underestimate the first minimum, resulting in an overstructured prediction of liquid water. This is a well-known result for approximations lacking vdW interactions, such as purely local GGAs.[144,183,467] Although dispersion corrections generally represent a step in the right direction, i.e. a less overstructured RDF, their effect depends on the specific functional and correction employed. For instance, BLYP is improved when supplemented with either D3 and DCACP corrections, while PBE is only improved with the D3 correction and deteriorates with DCACP (which was attributed to the presence of artificial dispersion effects in PBE[183]). Notably, the rVV10 non-local functional is also overstructured. In summary, BLYP-DCACP and revPBE-D3

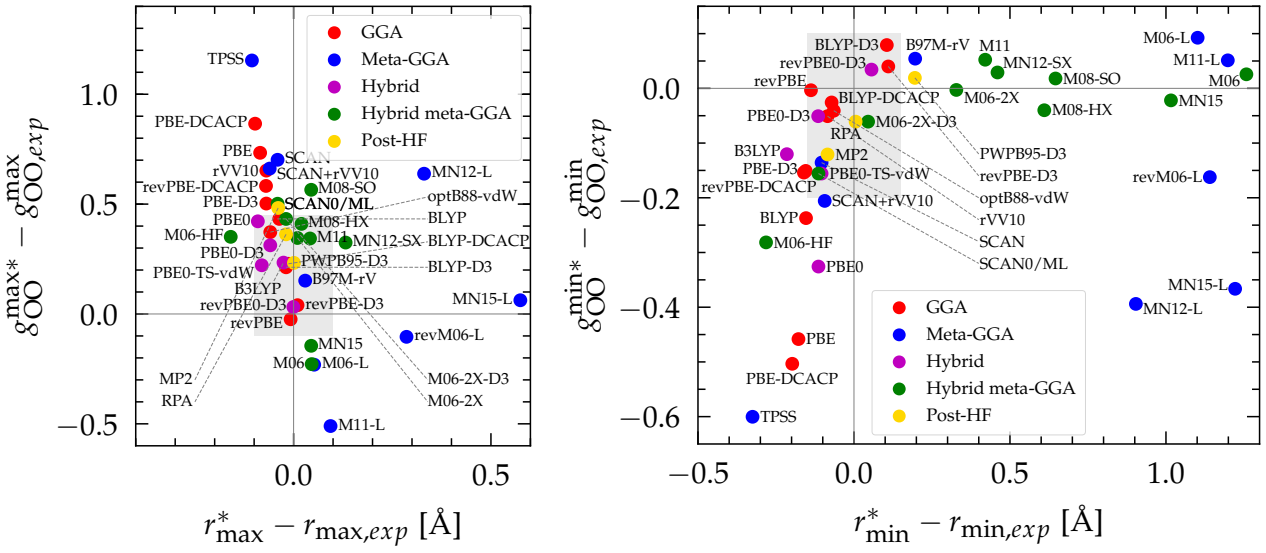


Figure 7.2: Left: Difference between the rescaled position r_{max}^{*} and height g_{OO}^{max*} of the first g_{OO} maximum and the experimental values at 298 K. Right: Difference between the rescaled position r_{min}^{*} and height g_{OO}^{min*} of the first g_{OO} minimum and the experimental values at 298 K. Values for non-Minnesota functionals were extracted from refs [144, 180, 183, 191, 414, 457, 465, 467, 483, 485] and are reported in Table C2. Rescaled values were obtained through empirical interpolation of experimental data.[429] The grey areas represent a visual estimate of the potential deviations resulting from the neglect of nuclear quantum effects as well as statistical and experimental uncertainties (c.f. Section 7.4.1).

are the best GGA functionals reported so far for the structure of liquid water.

The importance of a sensitive tweaking of non-local dispersion effects is likely the primary reason why meta-GGA functionals do not exhibit improvement over the best dispersion-corrected GGAs. Compared to all functionals, the local Minnesota ones are the worst, as they cause substantial right-shifting and broadening of the first g_{OO} peak. In contrast, the SCAN functional appears to capture the intermediate-ranged vdW interactions which seem to help locating the g_{OO} maximum and minimum at good distance,[415] but SCAN remains overstructured. The difference in results between SCAN and its augmentation with the rVV10 non-local correlation functional (SCAN+rVV10) is negligible.[191] However, this add-on does help the B97M-rV functional to become the best meta-GGA reported.

Based on the available data, hybrids provide a good approximation of the first maximum of g_{OO}, which is consistent with our previous observation that the inclusion of exact exchange in Minnesota functionals improves the accuracy of both the position and height of the first peak. This can be attributed to the fraction of exact exchange that mitigates the self-interaction error in local and semilocal XC functionals, which has been correlated with an artificial strengthening of the H_2O tetrahedral structure and the delocalization of protons.[485] Although PBE0 still yields overly structured water, its D3 and TS-vdW variants provide better agreement with experimental data. The most accurate hybrid functional appears to be revPBE0-D3, which is also the best approximation over all functionals for which data on water has been reported (vide infra).

Moving on to hybrid meta-GGAs, indications of the performance of SCAN0, the hybridized version of SCAN, has been obtained from simulations based on a deep neural network potential (ML) which indicate that SCAN0 is still overstructured. With the exception of M06-HF with 100% exact exchange, the hybrid Minnesota functionals are generally accurate in predicting the height of the first minimum, but they fail to accurately predict its position (Figure 7.2, right). However, Del Ben et al. discovered that M06-2X, which appears to be the best performing Minnesota functional for water overall, further improves when coupled with the D3 correction.[144] In general, it appears that the first minimum r^*_{min} is shifted to a smaller intermolecular distance when there is either an excessive amount of exchange (M06-HF) or when the correlation effects overestimate vdW interactions. This highlights the remarkable sensitivity between (exact) exchange and correlation, both of which tend to compress or augment the first coordination peak instead of having compensatory effects. Achieving an accurate description of liquid water with DFT therefore requires finding the correct balance between these two quantum effects. This quest has motivated the refinement of exchange and correlation functionals, where the occupied and virtual KS orbitals contribute to non-local correlation just as the occupied orbitals contribute to the non-local exact exchange. From Figure 7.2, the RPA, which consists of exact exchange plus the RPA

correlation, appears as the most promising post-HF DFT approach in this direction, e.g., yielding very good structural properties outperforming MP2.[144,180]

According to this comprehensive comparison, the most accurate functionals for describing the structure of water with classical nuclei are: revPBE-D3, BLYP-DCACP (GGAs), B97M-rV (meta-GGA), revPBE0-D3 (hybrid), M06-2X-D3 (hybrid meta-GGA) and the RPA (rung 5).

Nuclear quantum effects

The low mass of the hydrogen atom makes nuclear quantum effects (NQEs) significant when simulating water properties.[144,496,515] For example, tunneling effects can affect the formation and breaking of hydrogen bonds and influence the dynamics.[464] The results presented in this work (Figure 7.2) should therefore be interpreted in light of the fact that NQEs are absent in CPMD or BOMD dynamics with a classical propagation of nuclei. As an illustration, taking into account NQEs with revPBE-D3 revealed that its good agreement with water properties using classical nuclei is due to a fortuitous cancellation of errors, where the neglect of exact exchange compensates for the neglect of quantum nuclei.[457,467,516] Advanced path integral molecular dynamics (PIMD) methods are necessary for quantum-mechanical treatment of nuclei, particularly when comparing high-level electronic structure calculations with experimental results.[33,515,517] However, this comes at the cost of approximately two orders of magnitude more computational expense than simulations where the nuclei are treated classically. As a result, it has been common practice to mimic NQEs by performing classical (nuclei) MD at elevated temperatures increased by around 30 K.[415,483] While this ad hoc technique was found to provide reasonable accuracy for RDFs, it often fails to correctly reproduce the dynamical properties that become too fast compared to proper NQEs.[415,457,484,496] Alternatively, recent advances have enabled the acceleration of PIMD dynamics, especially with the help of ML potentials that infer DFT energies and forces at a much reduced cost.[180,414,484,485]

As expected, the general trend observed in PIMD simulations is that NQEs tend to soften the structure of liquid water: for BLYP,[515] SCAN/ML,[414,484] PBE0-D3,[144] SCAN0/ML,[485] RPA/ML[180] and MP2/ML[518] less structured RDFs were found when including NQEs. For other functionals like SCAN,[496] B97M-rV[457] and revPBE0-D3,[457,516] O-O RDFs remain almost unchanged, while the O-H and H-H RDFs become less structured. O-H and H-H RDFs are also less structured for BLYP-D3[496] and revPBE-D3,[516] that however have a slight decrease in the O-O first minimum (by $\sim$0.1) when adding NQEs, with no impact on the first maximum. However, overall NQEs seem to have a marginal influence on the positions of the maxima and minima of the distribution functions. Hence, classical RDFs tend to be either too structured or very similar to their quantum analogues. This is in agreement with experimental isotope studies be-

tween heavy and light water that also showed that NQEs soften the structure of liquid water.[414,483,519] Hence, NQEs can only partially explain why most DFT functionals tend to overstructure water compared to the experimental results in Figure 7.2. The gray areas plotted on this figure represent possible deviations due to the neglect of NQEs. These are based on PIMD references cited previously, potential discrepancies between experimental measurements,[428,429,432] and the variance of the rescaling procedure to 298 K. These areas therefore enclose the most promising XC functionals to be predictive with NQEs.

According to previous studies, the best functionals tested so far for describing the atomic structure of water with the consideration of NQEs are: revPBE-D3[516] (GGA), B97M-rV[457] (meta-GGA), revPBE0-D3[457,516] (hybrid), SCAN0/ML[485] (hybrid meta-GGA) and RPA/ML[180] (rung-5). Note however that good agreement with experiment was only obtained for revPBE0-D3 and the RPA (from insights with ML potentials). The other levels of theory still overstructure water when considering NQEs, except for B97M-rV that remains understructured. From Figure 7.2, other XC approximations that would be worth investigating with PIMD simulations would be: optB88-vdW, BLYP-DCACP (GGA), PBE0-D3(TS-vdW) (hybrid) and M06-2X(-D3) (hybrid meta-GGA). Running PIMD calculations with rung-5 XC descriptions like the RPA, without the aid of ML, would be of interest but their cost currently prevents this.

Finally, we note that NQEs also influence the balance between covalent and hydrogen bonds. In fact, PIMD simulations showed that NQEs broaden the covalent peak of the O-H RDF, meaning that more fluctuations occur for the hydrogen atom positions, accompanied by a weakening of the covalent bonds. In turn, such a delocalization of the protons seems to strengthen the hydrogen bond network by forming statistically more interactions, which slows down dynamical properties.[144,457,484,496,516] Counterintuitively, the disordering due to NQEs smoothes out the structure of water by destabilizing molecules in the intershell region of the O-O RDF, while simultaneously reducing diffusion and rotational times due to stronger hydrogen bonds. It will be therefore important to analyze dynamical properties in light of these findings in Section 7.4.2.

Coordination number

The coordination number n_{OO} predicted by each functional is plotted in Figure 7.3a. Experimentally, Skinner et al. showed that the O-O coordination number of the liquid state has a value of 4.3 and is independent of temperature,[428,429] while previous works reported values between 4 and 5.[183,413,428,429,431] In addition, negligible changes were observed from AIMD and force field simulations at different temperatures,[183] supporting that deviations of n_{OO} directly relate to the quality of the intermolecular interactions as described by the functionals. As seen, a majority of them is in agreement with the tetrahedral configuration of nearest-neighbor water molecules.[415,428]

However, the fact that the O-O RDF does not reach zero after the first peak makes it challenging to determine the first coordination shell unambiguously. This difficulty makes n_{OO} strongly dependent on the distance cutoff selected for the integration of the RDF: In most cases, n_{OO} is slightly underestimated because the O-O RDF tends to be overstructured in the absence of NQEs. On the other hand, the smoothening due to the addition of dispersion corrections makes the theoretical predictions agree more closely with experiments (e.g., BLYP-DCACP, revPBE-DCACP, M06-2X-D3). The lack of accuracy of the Minnesota meta-GGAs is further exemplified by their extended first coordination shell that includes an unphysical number of water molecules. Although still understructured, this is partly corrected for some hybrid functionals such as M06, M06-2X(-D3), M08-SO, M11, and MN15.

Density of liquid water

As illustrated in Figure 7.3b, GGA functionals tend to underestimate the equilibrium density, which is rectified by adding dispersion corrections. The incorrect prediction that ice sinks in water with local DFT is mainly due to the absence of dispersion in plain GGA functionals.[415,462] However, meta-GGA functionals such as SCAN have been shown to correct this issue.[415] The PBE0 hybrid functional faces challenges in achieving the right balance between covalent, hydrogen bonds and vdW forces. It significantly underestimates the density, but this can be improved with the D3 correction. For all other meta-GGAs, hybrids, hybrid meta-GGAs and post-HF/double-hybrids, the density is higher than the experimental value. Overall, vdW interactions increase the density because of their attractive and isotropic nature at intermediate and long range. This increases the population of molecules in the intershell regions of the O-O distribution function, i.e. between the coordination shells, and acts as additional cohesive force in the condensed phase. Consistent with their structural differences (Figure 7.1), the increase in the amount of exact exchange in the M06, M06-2X and M06-HF also correlates with a rise of the density. On the other hand, in a counteracting manner, the delocalization and disordering effects due to NQEs can be expected to reduce the density, explaining why DFT densities with classical nuclei are usually overestimated.

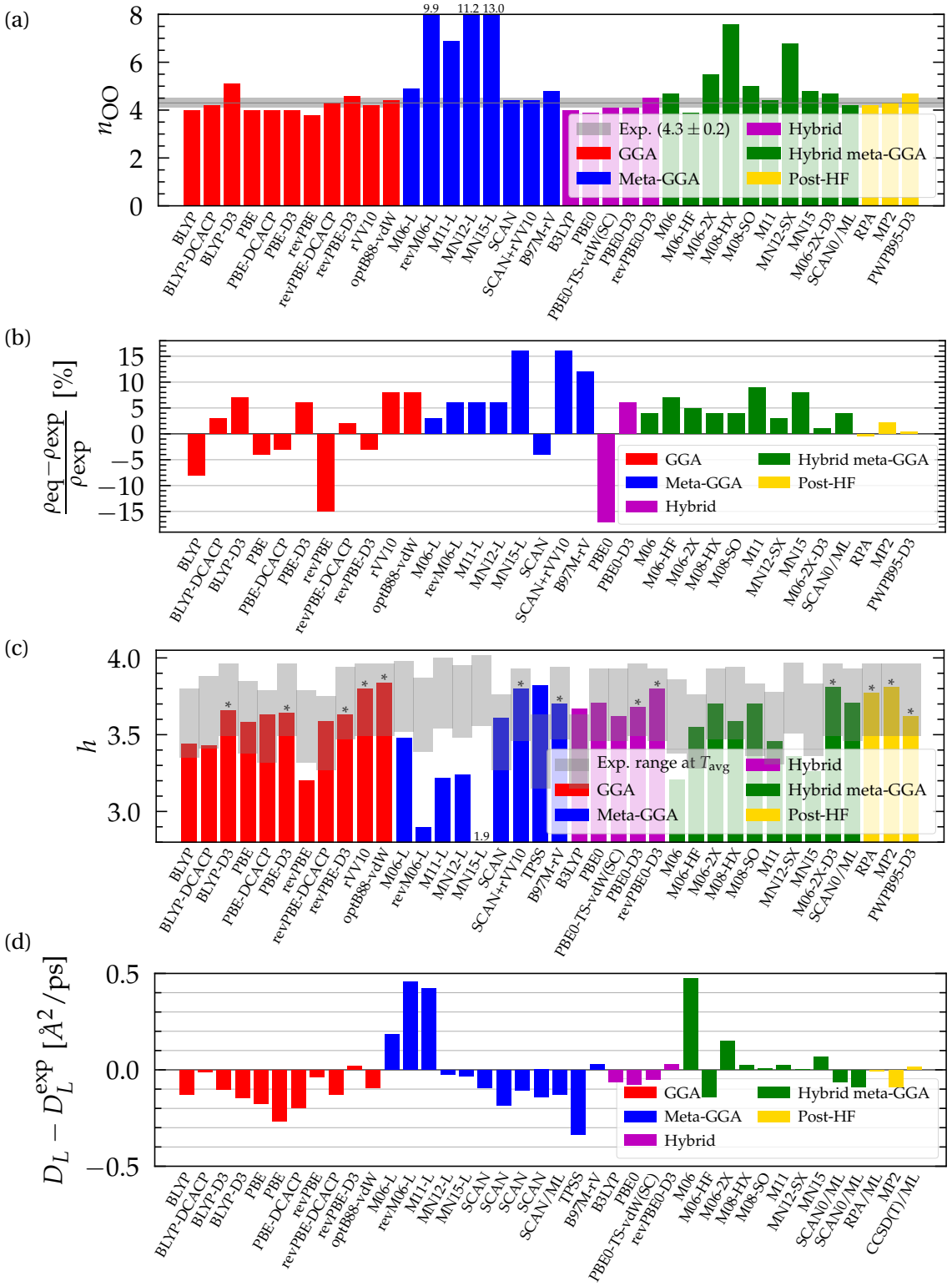


Figure 7.3: Structural and dynamical properties of liquid water from DFT-based ab initio simulations, compared to experimental values.[425,428,429,504,520,521] (a) Coordination number, (b) equilibrium density, (c) average number of H-bonds per water molecule (*upper bound from the integration of g_{OH} instead of geometric criteria), (d) finite-size diffusion coefficient. Results for non-Minnesota functionals were extracted from refs [144, 180, 183, 191, 414, 415, 457, 465, 467, 483–485, 496, 509] and reported in Tables C3 and C4.

H-bond number and angular distributions

From their atomic composition, water molecules in ice ideally arrange in a tetrahedral coordination made of four hydrogen bonds per molecule. In liquid water, entropic effects bend, stretch, break and reform hydrogen bonds such that the average number of H-bonds per molecule is slightly less than 4 (∼3.8) at near ambient conditions.[415,467] This average number h is plotted in Figure 7.3c, where the gray boxes indicate the estimated discrepancy among various experimental methods at the simulated temperature.[520,521] Our observations, and those of others,[464] suggest that the computation of h is relatively insensitive to changes in temperature, with a small deviation of approximately 0.1 for every 10 K increase.

Linked to the fact that Minnesota meta-GGAs are not providing accurate descriptions of the structure of water (Figure 7.1), being either understructured (M06-L, revM06-L, M11-L) or biasing the orientation between neighboring molecules (MN12-L, MN15-L), they are also unable to properly account for hydrogen bonds. Their angular distribution in Figure 7.4a further shows that semilocal Minnesota functionals are incapable of

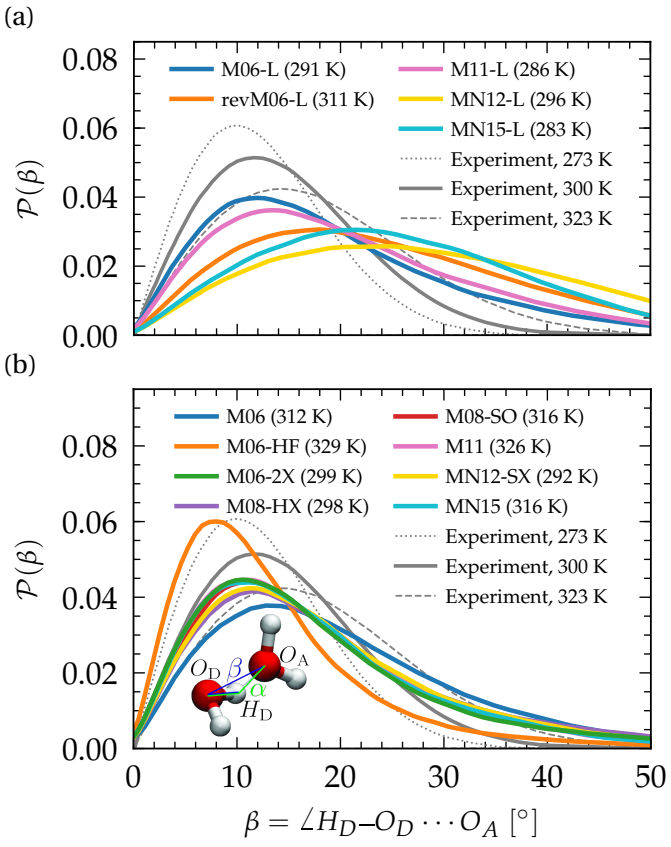


Figure 7.4: Distribution $\mathcal{P}(\beta)$ of the H-bonding angle β, compared to experimental values.[522] (a) Meta-GGA Minnesota functionals, (b) hybrid meta-GGA Minnesota functionals. Distributions of the complementary angle α are provided in Figure C2.

capturing the full details of the hydrogen bond network of water, that is too fluid. The hydrogen bond network of water is composed of a combination of short, straight, and robust bonds as well as longer, weak, and bent interactions. The strength of a hydrogen bond is consequently highly correlated with its length and angular orientation. At finite temperature, the elongation of the H-bonds competes with the cohesive effects of vdW interactions, which explains why the h number is in general higher and in better agreement with experimental data with dispersion corrections without altering significantly the angular distribution.[183] In contrast, both h and the angular distribution vary when considering different fractions of exact exchange; Figure 7.3c shows that h increases for M06-HF (100%), M06-2X (54%), M08-HX (52%), M08-SO (57%), while it is too low for M06 (27%), M11 (43-100%), MN12-SX (25-0%) and MN15 (44%). At the same time, H-bonds become shorter (Figure 7.1) and straighter (Figure 7.4b) when augmenting the fraction of exact exchange from M06 (27%) to M06-2X (54%) to M06-HF (100%). Hydrogen bonds are therefore particularly more sensitive to exchange effects than to correlation ones. Incorporating more exact exchange strengthens the hydrogen bonds and results in a more rigid structure of water.

Of all the structural properties analyzed, and taking also potential variations due to NQEs into account, we conclude that the functionals that provide results closest to experiments are: revPBE-D3, optB88-vdW, BLYP-DCACP (GGA), B97M-rV (meta-GGA), revPBE0-D3, PBE0-D3 (hybrid), M06-2X-D3, SCAN0 (hybrid meta-GGA) and the RPA (rung-5). Satisfactory agreement with experimental results, while directly accounting for NQEs, has only been demonstrated for revPBE0-D3[457,516] and the RPA.[180] The revPBE-D3,[516] PBE0-D3[144] and SCAN0[485] functionals overstructure water with NQEs, while B97M-rV[457] understructures. From a structural perspective, the remaining optB88-vdW and BLYP-DCACP GGAs emerge as intriguing candidates to investigate also in the presence of NQEs. The rung-4 M06-2X-D3 functional is even more promising, as it is slightly overstructured without NQEs and offers accurate density and hydrogen bond characteristics.

7.4.2 Dynamical properties

Diffusion coefficient and orientational correlation times

In Figure 7.3d, we plot the difference between the diffusion coefficient D_L and its experimental counterpart rescaled to a fictitious simulation box. The equivalent comparison with simulated coefficients D_∞, rescaled to infinite size, is presented in Figure 7.5a. While the diffusion coefficient provides information about the translational movement, rotational features are characterized by the orientational relaxation times plotted in Figure 7.5b. These correlation times are highly sensitive to statistical sampling and require trajectories that are sufficiently long (approximately three times their value) to be accurately converged. Additionally, the fitting, respectively integration, methods used

for their calculation vary between studies, and experimental measurements exhibit non-negligible deviations. Nevertheless, these values are presented as a qualitative

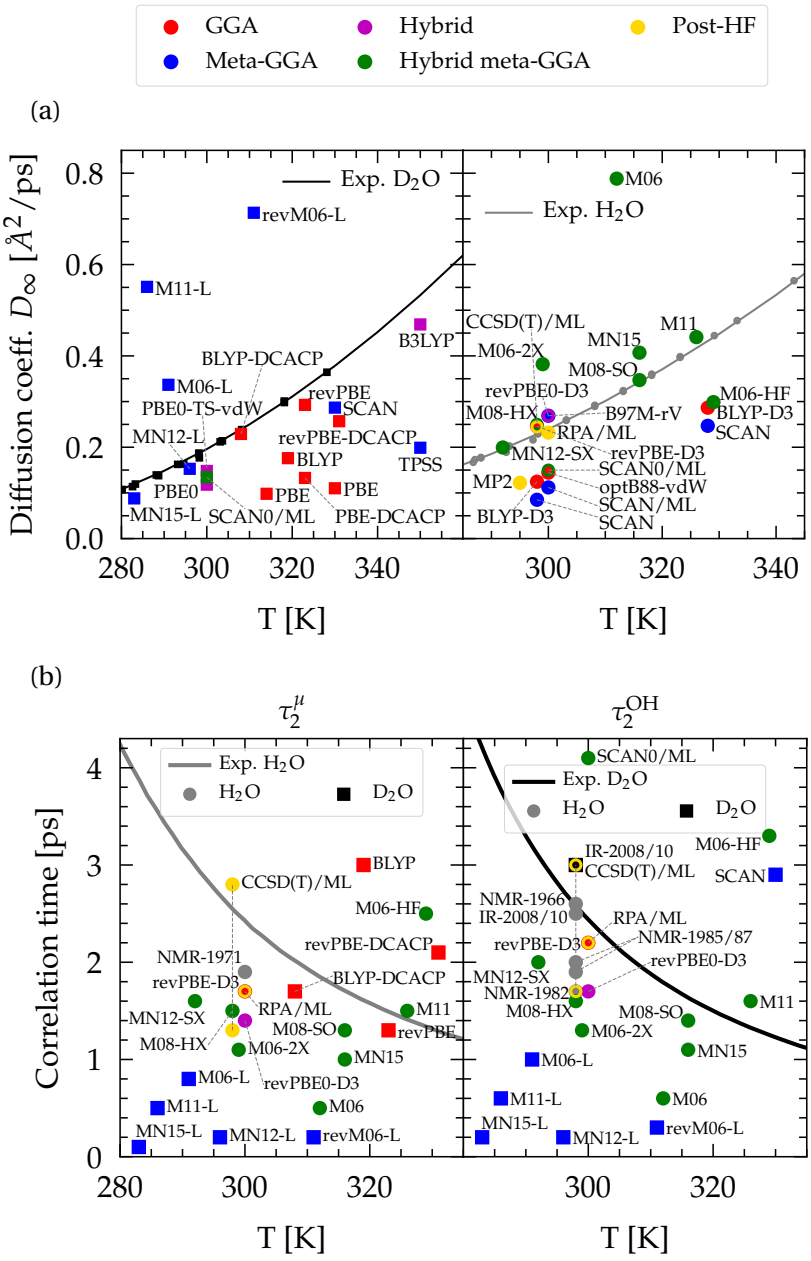


Figure 7.5: (a) Diffusion coefficient rescaled to infinite size for heavy (left) and light (right) water. Experimental data points were compiled from refs [421–423, 425, 523, 524] and fitted according to ref [425]. (b) Qualitative comparison of the orientational relaxation times τ_2^μ and τ_2^{OH} with experimental results.[416,420,424,426,510–513,525] CCSD(T)/ML values are from PIMD simulations including NQEs,[509] and extend through the range of experiments. Non-Minnesota results were extracted from refs [144, 180, 183, 414, 415, 457, 465, 467, 483–485, 496, 516] and reported in Tables C4 and C5.

comparison.

The dynamics predicted by DFT functionals depends on their ability to account for hydrogen bond strength as well as directionality. Diffusion and rotational movements are determined by the dynamic breaking and formation of H-bonds under thermal fluctuations. Therefore, if the description of H-bonds is too strong, it significantly slows down the dynamical properties. Local and semilocal functionals suffer from the self-interaction error that promotes a delocalization of the protons.[467,485] This declocalization facilitates the formation of H-bonds when the proton moves toward the acceptor and thus contributes to the H-bond strengthening, in an analogous manner to the NQEs (Section 7.4.1). As an illustration, the diffusion coefficient is too low for most GGA and meta-GGA functionals, in agreement with their tendency to overstructure. For example, optB88-vdW yields slightly overstructured water, and diffuses too slowly. BLYP-DCACP and revPBE have higher coefficients, more in line with experiment, but this originates from their underestimation of the number of hydrogen bonds (Figure 7.3c). For GGAs, both BLYP-DCACP and revPBE-D3 functionals have a diffusion coefficient and relaxation times very close to the experiment, which is also true for the diffusion modelled by the B97M-rV meta-GGA. In contrast to the statement that the self-interaction error slows down the dynamics of liquid water, we have found that the M06-L, revM06-L and M11-L semilocal functionals exhibit a complete opposite trend, generally leading to faster dynamics. This is obviously due to their distortion of the hydrogen bonding network (Figure 7.4) and incorrect structuring (Figure 7.1). The diffusion coefficients predicted by MN12-L and MN15-L functionals appear to be in good agreement with experimental values, but this is fortuitously caused by an error compensation between the lack of hydrogen bonds (Figure 7.3c) and their overly strong (incorrect) structure (Figure 7.1). Their rotational dynamics is indeed significantly faster than observed experimentally.

As explained earlier, NQEs tend to strengthen H-bonds and slow down the dynamical properties. This was seen in all PIMD calculations with BLYP-D3,[496] revPBE-D3,[516] SCAN,[414,484,496] B97M-rV,[457] revPBE0-D3,[457,516] and RPA/ML.[180] The diffusion coefficients, in the absence of NQEs, should therefore be seen as overestimated, and relaxation times as underestimated. The diffusion with GGAs would therefore become even slower with NQEs. From the available data, hybrid and hybrid meta-GGA functionals generally give faster diffusion than GGAs, which indicates that the exact exchange is also a key ingredient towards achieving accuracy for the dynamics, in the same way as for the structural properties. The revPBE0-D3 functional can be considered as the most effective hybrid in this regard. Other functionals like PBE0 or SCAN0 are likely to remain too slow even upon inclusion of NQEs.

Except for M06-HF, hybrid Minnesota functionals lead either to too fast diffusion or are in good agreement with the reference values. Thus, incorporating NQEs could potentially bring them closer to experimental results. Consistent with the understructuring

tendency of M06 (with 27% of exact exchange) and the overstructuring of M06-HF (with 100%), the M06 family shows once again that the amount of exact exchange tightly regulates the precision of the functional: Dynamical properties are too slow for M06-HF due to the shortening of stronger H-bonds, while M06 is too fast. The balanced M06-2X (54%) is giving results that are inbetween and therefore closer to experimental values. From a first estimation based on ML potentials, the dynamics of the rung-5 RPA decription resembles closely the one revPBE-D3 and is thus highly consistent with the experimental data.

Overall, considering the possible influence of NQEs on the analyzed structural and dynamical properties, the functionals that most closely align with experiments are: revPBE-D3 and BLYP-DCACP (GGAs), B97M-rV (meta-GGA), revPBE0-D3 (hybrid), M06-2X(-D3), SCAN0 (hybrid meta-GGAs) and the RPA (rung-5). Satisfactory agreement for both structural and dynamical properties while accounting for NQEs has only been demonstrated with revPBE0-D3[457,516] and RPA/ML.[180] revPBE-D3[516] and SCAN0/ML[485] descriptions tend to overstructure water yielding too slow dynamics, even when accounting for NQEs, while B97M-rV[457] understructures and slightly accelerates diffusion. Based on our extensive analysis, BLYP-DCACP (GGA) and M06-2X(-D3) (hybrid meta-GGA) functionals therefore emerge as promising competitors to revPBE0-D3 and the RPA, and warrant further investigation with PIMD approaches.

7.5 Conclusions

Water is the most abundant substance on Earth, and its liquid properties are distinct from those of other fluids, posing a challenge for in silico simulations not only of condensed water but also of aqueous chemistry. In this work, we explore the performance of Minnesota meta-GGAs and hybrid meta-GGAs in describing the structure and dynamics of liquid water via ab initio molecular dynamics simulations. Contrary to the prevailing belief that local and semilocal functionals overstructure water, leading to underestimation of dynamical properties, the Minnesota meta-GGAs exhibit the opposite trend. M06-L, revM06-L, and M11-L lead to understructuring of water, while MN12-L and MN15-L lack cohesive effects, resulting in increased intermolecular distances. This behavior can be attributed to the weakening of the hydrogen bond network causing dynamical fingerprints that are far too fast. On the other hand, while most of the hybrid Minnesota functionals remain understructured (M06, M08-HX, M08-SO, M11, MN12-SX, MN15), their dynamical properties generally improve over those obtained with local and semilocal functionals. The inclusion of exact exchange was identified as a key ingredient for the correct description of hydrogen bonds leading to improved structural and dynamical properties. In contrast, we found that an excessive amount of exact exchange (M06-HF) shortens and strengthens the hydrogen bonds between molecules, thus giving water properties that are too glassy.

M06-2X turns out to be the best Minnesota functional for liquid water. Slightly understrucured, its D3 dispersion corrected version shows very good agreement for structural properties. Describing the complete picture of water from small to larger clusters, to the condensed phase, is highly non-trivial with DFT, because functionals showing good performance in the gas phase do not necessarily perform well in the liquid phase and vice versa.[183,462] Very encouragingly, M06-2X has also been identified as one of the most accurate functionals for relative energies of water hexamers[526] and binding energies of 16-mers and 17-mers.[527] Furthermore, from the thorough benchmark by Goerigk and Grimme, M06-2X-D3 was found to be the best among 23 hybrid functionals for general main group thermochemistry, kinetics, and noncovalent interactions.[146]

Previous studies considering explicit nuclear quantum effects (NQEs) in water, have identified the hybrid revPBE0-D3, and the rung-5 RPA (EXX+RPA, RPA@PBE) with the help of machine learning potentials, as the only two approximations that agree closely with experiments so far. This therefore encourages the investigation of the performance of M06-2X(-D3) functionals with NQEs via path integral approaches. Although it is unfortunate that this involves drastic computational overheads, our work provides further evidence that both exact exchange and appropriate (non-local) correlation are essential for accurately describing water interactions. This, in turn, suggests that well-balanced XC functionals from higher rungs of the Jacob's ladder are required for simulating complex biological systems in water with predictive accuracy. In this regard, determining whether M06-2X(-D3) are indeed one of the best functionals would avoid the resort to the significantly more expensive fifth rung of the Jacob's ladder.

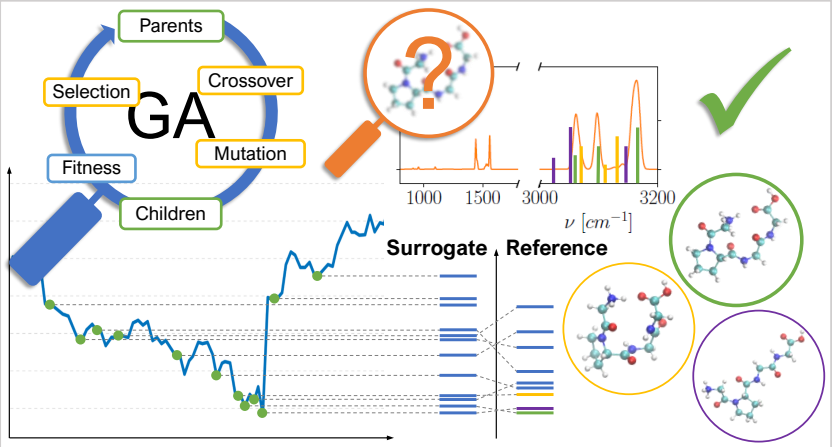
Data availability

Data and analysis scripts will be provided on Zenodo at https://doi.org/10.5281/zenodo.7933087.

Appendix

In Appendix C, readers will find additional simulation details, and interpolations utilized for rescaling the g_{OO} peaks to 298 K. Supplementary angular distributions of hydrogen bonds are also provided. Moreover, the appendix lists all the structural and dynamical properties presented in the study, along with corresponding references to relevant literature. Hopefully this will serve as a shared foundation for future assessments of DFT on water, encouraging further advancements.

8 Surrogate based genetic algorithm method for efficient identification of low-energy peptide structures



Chapter 8 is a postprint version of an article published as:

Villard, J.; Kılıç, M.; Rothlisberger, U. Surrogate based genetic algorithm method for efficient identification of low-energy peptide structures. *Journal of Chemical Theory and Computation* 2023, 19, 1080-1097.

Reproduced under the terms of the CC-BY-NC-ND 4.0 License.

8.1 Abstract

Identification of the most stable structure(s) of a system is a prerequisite for the calculation of any of its properties from first-principles. However, even for relatively small molecules, exhaustive explorations of the potential energy surface (PES) are severely hampered by the dimensionality bottleneck. In this chapter, we address the challenging task of efficiently sampling realistic low-lying peptide coordinates by resorting to a surrogate based genetic algorithm (GA)/density functional theory (DFT) approach (sGADFT) in which promising candidates provided by the GA are ultimately optimized with DFT. We provide a benchmark of several computational methods (GAFF, AMOEBApro13, PM6, PM7, DFTB3-D3(BJ)) as possible prescanning surrogates and apply sGADFT to two test case systems that are (i) two isomer families of the protonated Gly-Pro-Gly-Gly tetrapeptide (Masson, A.; et al. *J. Am. Soc. Mass Spectrom.* 2015, 26, 1444-1454),[107] and (ii) the doubly-protonated cyclic decapeptide gramicidin S (Nagornova, N. S.; et al. *J. Am. Chem. Soc.* 2010, 132, 4040-4041).[528] We show that our GA procedure can correctly identify low-energy minima in as little as a few hours. Subsequent refinement of surrogate low-energy structures within a given energy threshold ($\leq$10 kcal/mol (i), $\leq$5 kcal/mol (ii)) via DFT relaxation invariably led to the identification of the most stable structures as determined from high-resolution infrared (IR) spectroscopy at low temperature. The sGADFT method therefore constitutes a highly efficient route for the screening of realistic low-lying peptide structures in the gas phase as needed for instance for the interpretation and assignment of experimental IR spectra.

8.2 Introduction

Understanding the correlation between composition, structure, properties and functional roles of biomolecules is at the very heart of biochemistry and biophysics. The first step in this hierarchy, i.e., the connection between composition and structure, has thus attracted enormous interest both in the case of, e.g., entire proteins[529–532] and for smaller peptides.[533–536] The latter are especially interesting in view of reducing the complexity of natural systems and studying smaller-size models under controlled conditions. Furthermore, peptides made of few amino acids have attracted much attention in recent years thanks to their promising and wide scope of applications, be it in the fabrication of biomaterials,[537] in the engineering of biomimetic compounds for catalysis,[538] or in drug design.[539] Indeed, in addition to contributing to physiological health,[540] peptides have brought about conclusive benefits as anti-infective drugs due to their antimicrobial activity,[541–543] leading to intense efforts in therapeutics development with peptidomimetic systems.[544,545]

The study of gas-phase peptides alone or with a defined number of solvent molecules constitutes a first step toward the understanding of the *in vivo* properties and al-

lows for a differential picture of well-controlled intramolecular interactions separated from their combination with condensed-phase and intermolecular effects. Moreover, in some cases, the experimentally produced gas-phase systems are able to retain solution-phase features so that scrutinizing native forms in the gas phase and at near zero temperature can provide valuable insight into remanent condensed-phase interactions.[107,546,547]

Experimentally, advances over the past decade have coupled laser desorption and supersonic molecular beam cooling to capture IR spectra of neutral biomolecules in the gas phase.[110,111] Alternatively, combinations of electrospraying, ion-mobility selection, mass spectrometry, and cryogenic ion traps were reported to separate between conformational families of charged molecules prior to, e.g., IR measurements.[107,110,548] In particular, such experiments performed at cryogenic temperatures have been able to produce vibrationally resolved and conformer-selective measurements, but due to the high intrinsic complexity, the identification of the underlying structures and the full assignment of the observed IR spectra can only be achieved with the support of computational methods. In turn, the experimental low-temperature data provide highly sensitive benchmarks for the assessment of the performance of computational methods for biorelevant systems where the availability of accurate quantitative data is often sparse and hard to obtain. Therefore, the present work also illustrates a sensitive test case of the complementarity between simulations and experiments.

At low temperature, conformers are expected to occupy the thermodynamically most stable configuration on the PES or at least some kinetically trapped low-lying metastable states. Therefore, from a computational perspective relevant local minima (LM) are usually searched on the rugged, high-dimensional PES and theoretical IR spectra, commonly computed with DFT including exact exchange at the hybrid level for sufficient accuracy, are compared to the experimentally observed spectra.[105–107,549–552]

The exploration of the PES is typically performed with molecular dynamics (MD), relying on classical force fields and semiempirical or first-principles potentials in combination with replica-exchange and/or simulated annealing (SA) to enhance sampling.[105,108–111,548,553–555] Though highly successful in many cases,[106,107,552] this approach based on traditional quantum chemical tools suffers from severe drawbacks and limitations: On one hand, the quantitative identification of the lowest energy structures at low temperature poses stringent accuracy demands to provide a correct energetic ordering in the 0-2.5 kcal/mol observation range for biomolecules that are characterized by complex interatomic and noncovalent interactions.[553] This imposes the use of higher level computational methods for the determination of realistic relative energetics,[548,554,556] while force fields or semiempirical approaches often fail at providing the necessary accuracy.[548,553,557,558] On the other hand, even small peptides contain of the order of tens to hundreds of atoms, making higher level first-principles calculations time-consuming for all but the smallest molecules. In particular, using

first-principles MD requires long simulation times with ten thousands of energy and force evaluations for typical simulated annealing runs and, in spite of multiple runs with different starting geometries and varying simulated annealing protocols (in terms of highest temperature, simulation length at $T_{\max}$ and subsequent cooling rate), can potentially fail to recover the most stable structures due to the presence of high-energy barriers on the PES.[111,555] Even when successful, DFT-MD based identifications of the lowest-energy structures observed in experiments can take several months and might only be practicable for larger systems when introducing experimental information to guide the search.[548,552]

Here, we tackle the task of rapidly finding the global minimum (GM) as well as low-lying LM, with the help of surrogate based genetic algorithms (GA). By leveraging evolutionary mechanisms, GAs have shown efficiency in solving highly nonlinear and complex global optimization problems[94,103,104] where deterministic or analytical methods fail at finding correct solutions or efficiently search enormous solution spaces. In particular, the capabilities of GAs were for instance found to surpass SA in the search for ground state fullerene clusters[559] or perform better at protein structure predictions compared to Monte Carlo approaches on simplified energy models.[560–562] GAs are also among the most CPU-/search-efficient methods to computationally identify low-energy conformations when applied to small organic molecules[563–565] or peptides in the gas phase.[557,566,567] For example, they outperformed systematic and random search methods for the mycophenolic acid drug-like ligand and were more efficient than replica-exchange MD for dipeptides in terms of low-energy conformational coverage (respectively within 5 and 10 kcal/mol from the GM).[568]

Despite this algorithmic gain, the predictive power of such evolutionary methods evidently depends on whether the energy function is able to faithfully describe the relevant physical interactions. For example, up to now, the lack of fast and sufficiently accurate (free) energy models[569–573] explains why ab initio protein folding predictions have met little success in recovering secondary and tertiary structures in close agreement with native-like conformations.[562,572,574,575] In that case, the enormous space defined by the number of structural degrees of freedom severely challenges search engines, so that protein folding approaches often privilege sequence homology[531,576] or machine learning[532,577] algorithms. However, in the mid-size range of peptidic systems, GA applications have a lot of potential if tractable energy models with sufficient accuracy exist.

Due to the large number of energy evaluations required, GAs for peptide folding are commonly used in conjunction with classical force fields[557,563–566] or expedient semiempirical methods,[567] at the price of loosing accuracy so that identified stable structures might correspond to false LM introduced by the energy function and relative energies between different conformers are far off experimental observations.[557,563,565] As a potential remedy, GA optimizations were recently combined with DFT local

relaxations.[568] However, this approach was rather limited to short GA instances of dipeptides and molecules up to ∼40 atoms so that applications of this fully DFT-based approach to larger systems are currently compromised even when resorting to massively parallel computational resources.

We rather explore here the possibility of using less accurate surrogate models for a faster (pre)evaluation of the PES and demonstrate that a judicious choice of surrogate level can provide satisfactory knowledge for establishing a pool of low-energy candidates, to be ultimately refined at a first-principles level. This seems also reasonable in view of the fact that relative energies and vibrational frequencies can differ markedly upon changing the level of theory, DFT functional, or basis set[107,556,558,578] and that there exists a priori no exact, tractable and universal baseline for the PES to drive the optimization with.

To anticipate our results, it turns out that while the surrogate LM geometries are in general very close to their first-principles analogues for all lower-level methods considered here, the energy hierarchy varies significantly between PES approximations and can considerably deteriorate the search. Nevertheless, our results show that, in combination with a state-of-the-art polarizable force field, the approach is highly successful in generating surrogate low-lying minima that match experimental structures for the two test case systems including two isomers of the protonated Gly-Pro-Gly-Gly tetrapeptide (referred to as GPGG herein) and the doubly protonated gramicidin S cyclodecapeptide. This encourages the use of sGADFT as a straightforward, fast, and automatized way to identify the lowest energy structures of peptides in the gas phase.

In what follows, we first describe the reference test systems in Section 8.3, along with our GA implementation. After presenting the computational details and the investigated surrogate models in Section 8.4, we provide a quantitative assessment of their cost-accuracy performance on a test set of GPGG structures in Section 8.5.1. Respective GA results are then presented in Sections 8.5.2 and 8.5.3, and their computational footprint finally is reported in Section 8.5.4, before drawing conclusions in Section 8.6.

8.3 Methods

8.3.1 Reference data

To test the performance of the sGADFT approach, we have chosen two reference systems of different size for which the lowest energy structures have previously been determined via a combination of high-resolution conformer-selective IR spectroscopy paired with electrospray ionization and cryo-cooled ion traps, supported by a traditional computational approach (as described above) to determine the most stable structures. The first test case system comes from the work of Masson et al., who lever-

aged ion-mobility techniques to identify and separate two conformational families of the protonated GPGG peptide (Figure 8.1) with different collisional cross-sections, and acquired respective spectroscopic data.[107] Major conformers of each family were determined as involving either the cis or trans isomers of the proline residue.

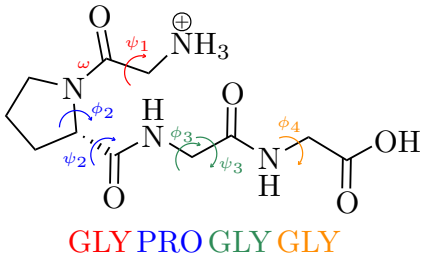


Figure 8.1: Schematic structure of the 39-atom protonated GPGG peptide in its cis ($\omega = 0$) isomer, shown with the respective backbone dihedrals employed for GA optimization.

The 3D structure determination was previously established by running SA ab initio MD starting from random cis/trans structures extracted from the Protein Data Bank (PDB).[579] The search was conducted at the DFT level with the B3LYP functional[580] and a 6-31G basis set, with extensive trials of heating temperatures and annealing rates for total simulation times of several tens to hundreds of picoseconds. After this first exploration, isomers were structurally and energetically selected, and locally relaxed at the B3LYP/6-31G(d,p) level of theory to provide a final set of 13 cis and 29 trans energetically low-lying candidate structures, which serve as the reference pool in this work. Comparison of theoretical harmonic vibrational frequencies (at B3LYP/6-311++G(d,p) level) including isotopic substitutions with the measured spectra clearly confirmed that the lowest-energy configuration of each family of these two sets corresponded indeed to the most abundant of the observed conformers.

Similarly, in 2010, Nagornova et al. published highly resolved IR spectra of the doubly protonated gramicidin S peptide (Figure 8.2) featuring a D rather than an L enantiomer of a phenylalanine.[528] Since the experimental data indicated some symmetry (C_2) for the major conformer, an SA exploration of the high-dimensional PES could be performed by imposing structural constraints over multiple FF99SB[581] and FF02polEP[582] force field trajectories.[552] The 3D structure was finally determined by calculating B3LYP/6-31G(d,p) spectra of few candidates.

8.3.2 Genetic algorithms

Genetic algorithms (GAs) are global optimizers that belong to the larger class of evolutionary algorithms rooted in the mechanisms of biological evolution. As metaheuristic search engines, GAs operate over populations of individuals that each represent a candidate solution of the optimization problem and are progressively modified toward (near-) optimal solutions. GAs are powerful tools when it comes to hard optimization

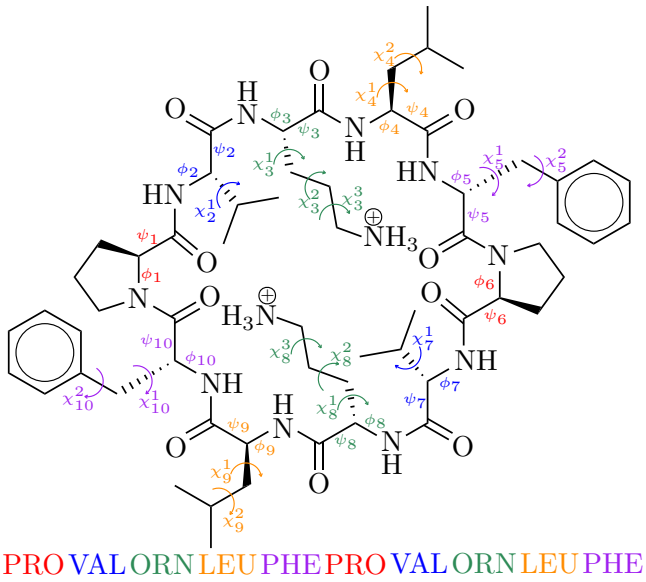


Figure 8.2: Schematic structure of the 176-atom doubly protonated gramicidin S cyclic decapeptide, shown with respective backbone and side-chain dihedrals employed for optimization.

problems for which the solution space is supposedly noisy, unsteady, and involves constraints or many LM as well as many degrees of freedom that do not allow simpler local optimizers or enumeration searches to perform efficiently.

Figure 8.3 depicts a schematic representation of the GA employed in this work built from conventional genetic operations. Generally, first individuals are randomly generated, if no other information or constraints are known, to ensure diversity and prevent any other bias in the solution space originating from the initialization. At each generation (iteration), GA evolves solution individuals with biologically inspired operators. Each individual is assigned a fitness that serves as a metric to drive the genetic evolution of the algorithm. Most of the time, this fitness function is nothing else than the objective function of the optimization problem.

Following the Darwinian principles of mate selection and survival of the fittest, individuals are stochastically selected based on relative fitnesses in the population and give birth to children individuals through crossover of genes. Genes are encoding fragments of a tentative solution that depend on the problem at hand and that must be carefully designed by the practitioner. Examples of such encodings or representations are bit strings, symbols, or vectors that contain relevant information to be transferred from one generation to the next. Children solutions are then randomly mutated to maintain diversity and possibly extend the search over yet uncovered regions of the solution space. Finally, elitism consists of replacing some of the current less-fit individuals with the best individuals of the previous generation, in order to maintain the best traits discovered so far over the generations. Such a selection-crossover-mutation-elitism

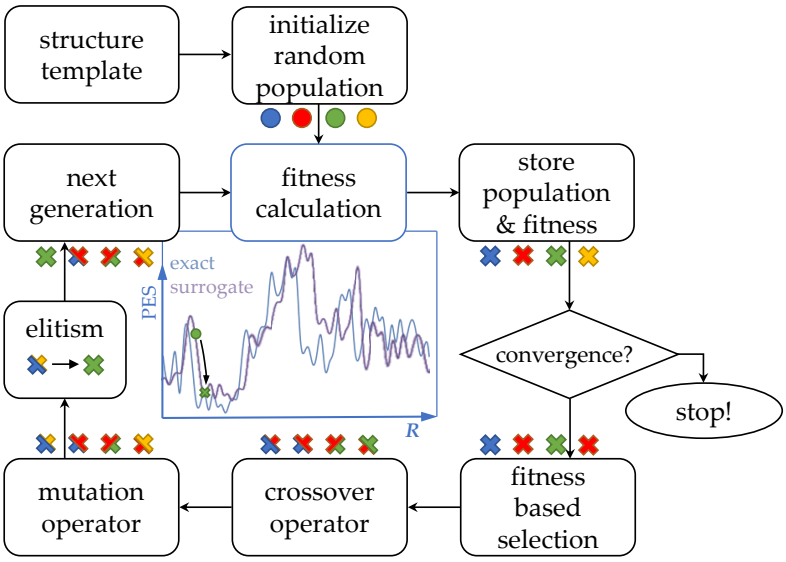


Figure 8.3: Schematic representation of the GA cycle as implemented in EVOLVE for this study.

cycle hence simulates an artificial evolution and propagates relevant and optimal features of the representations across GA iterations. The algorithm terminates after a fixed number of generations or when improvement of the fitness function stagnates over several iterations.

Due to the very nature of the initialization, selection, and mutation, GAs are intrinsically stochastic and provide statistical results that hopefully contain the GM of the optimization problem. In the following, our GA implementation toward the optimization and generation of low-lying peptide geometries is described.

8.3.3 Optimization of peptide conformations with EVOLVE

All of the work presented in this chapter was performed with the in-house implementation of a single-objective and multiobjective GA engine called EVOLVE.[99,583] As a versatile and modular Python code for peptide and protein sequence optimization, EVOLVE was successful in the optimization of a biomimetic peptidic scaffold for the fixation of CO_2[584] and in the engineering of a highly thermostable metalloprotein.[585] It also served in the elaboration of training sets for enhanced machine learning models of molecular properties.[586]

For the compositional optimizations mentioned above, side chain rotamer libraries were used in order to restrict the search space to discrete sets of residue conformations. In contrast here, EVOLVE is extended into a complete in silico optimizer of a peptide structure (including both the backbone conformation and side chain dihedrals) with a fixed amino acid sequence whose degrees of freedom therefore cover a huge space. In

the gas phase at near zero temperature, the objective function is nothing more than the potential energy as a function of atomic coordinates, meaning that the lower the energy (fitness), the better the structure. Such an "ab initio" peptide folder is capable of exploring the low-lying LM or reaching the GM of the PES, which is particularly relevant for assigning 3D structures to measured IR spectra.[105–107,110,552] In addition, it also provides an exhaustive search that enables a quality test of the method used to describe the PES.

In practice, genetic operators have parameters that are fixed before execution, which strongly influence the efficiency and reliability of the algorithm. The optimal choice of these parameters is a multivariate problem in itself and depends on the forms of the operator, the problem to be solved, and the characteristics of the fitness function.[587] We studied the effect of several parameters such as population size, crossover probability, or mutation rate and selected a set for which independent runs progress steadily toward the lowest energies in a small number of iterations. In what follows, we describe the specificities of the algorithm (Figure 8.3) and list the corresponding parameters in Section 8.4.1.

Representation

Each individual or tentative solution of the optimization problem is a peptide conformation. Translated in a GA framework, each geometry is represented by the backbone ϕ and ψ torsional angles as well as possible side chain torsional angles χ. The genes of one individual composed of N amino acids with respective k numbers of side chain dihedrals are therefore

$$\mathbf{\Theta} = (\phi_1, \psi_1, \chi_1^1, ..., \chi_1^{k_1}, ..., \phi_N, \psi_N, \chi_N^1, ..., \chi_N^{k_N}) \tag{8.1}$$

encoded into a single numerical vector $\mathbf{\Theta}$ that defines the internal coordinates of the optimizer. The specific torsional angles used for the optimization of the two test systems studied herein are indicated in Figures 8.1 and 8.2. This choice of representation, inspired by the underlying characteristics of the Ramachandran plot,[588] was already exploited in previous evolutionary methods[557,568] and has the advantage of easily defining genetic operators that preserve the peptide atomic connectivity.

Initialization

The information about the amino acid sequence, the atom types, and atomic connectivity are provided to EVOLVE in the form of a PDB file which serves as an initial template. From this, a $\mathbf{\Theta}^i$ representation of size K is randomly generated in which each individual i of the first population has a uniform distribution of its torsional angles such that $\Theta_k^i \in [-180°, 180°)$ for $k = 1, ..., K$. Technically, such modifications of the

peptide structures are performed with the help of the Open Babel toolbox.[589]

Fitness function

At each generation, the ability of an individual to be among the lowest energy configurations is assessed by calculating its potential energy. In order to avoid the exploration of highly improbable nonphysical structures, e.g., with too close distances or steric overlaps, initial local relaxations are performed before assigning the energy. Indeed, the individuals modified by the genetic operators can be very distorted and far from LM of the fitness function, which prevents the algorithm from progressing rapidly to low energies by stagnation or by bouncing off the PES,[566] in a manner quite similar to a gradient descent with a high learning rate. To improve this, the search space is consequently focused on the physically more meaningful regions that involve LM.

The algorithm thus operates at two levels: a coarser (and wider) exploration of the configurational space driven by genetic operators acting on $\boldsymbol{\Theta}$, refined by local optimizations of the Cartesian coordinates $\mathbf{R}$, as illustrated in the central graph of Figure 8.3. The relaxed structures $\tilde{\mathbf{R}}$ and energies are stored to construct the pool of putative LM and corresponding fitnesses and are further translated back to their torsional representations $\tilde{\boldsymbol{\Theta}}$ that are updated before selection:

$$\boldsymbol{\Theta} \mapsto \mathbf{R}, \mathbf{R} \xmapsto{\text{“}-\nabla_{\mathbf{R}} E\text{”}} \tilde{\mathbf{R}}, \tilde{\mathbf{R}} \mapsto \tilde{\boldsymbol{\Theta}} \tag{8.2}$$

The computational cost is determined by the number of fitness function calls (equal to the number of generations times the population size) times the cost for a single fitness evaluation. The latter depends crucially on the level of the surrogate method, while the choice of the PES model (and thus the fitness function) is critical in order to reliably reflect experimental results. Thus, compromises have to be made between cost and accuracy. EVOLVE is currently interfaced with several external software programs (Gaussian,[332] Amber,[590] OpenMM[52]) that can be used for local gradient based optimizations at different levels of theory. Note that the modular structure of EVOLVE and the use of the Atomic Simulation Environment (ASE) library[591] to interface with external codes greatly facilitate the integration of new fitness evaluators. Finding an appropriate surrogate model for the PES in terms of speed and accuracy for the relative energetics is investigated in Section 8.5.1. For the GA applications envisioned here, it is not absolutely necessary to reproduce the PES in every detail but for a given surrogate model to be satisfactory, it has to be able to drive the GA optimizations toward regions with a promising set of candidates also likely to belong to low-energy regions at the higher-level reference method.

Sanity checks and constraints

The resulting geometries and energies are checked after each fitness evaluation to ensure that the local optimization was successful, as it may happen that the initial structures $\mathbf{R}$ generated by the GA operators have clashes or are so deformed that it becomes difficult for the local optimizer to converge to a stable (local) minimum, especially within the first few generations. In this uncommon case, individuals from nonconverged optimizations are simply ignored and replaced in the next generation by assigning them a very high (unfavorable) fitness value.

A similar procedure is applied to constrain the GA search. In particular, when running separate optimizations for the cis- and trans-GPGG manifolds, the geometries are checked on-the-fly to ensure that they belong to the chosen isomer class since the local optimizer can, although very rarely, alter the isomerization state of the proline (Figure D6).

For gramicidin, the cyclic structure is enforced by requiring the bond between the PRO1 nitrogen and PHE10 carbon atom not to exceed 2 Å, which again rarely occurs due to the definition of a cyclic topology in the force fields which imposes a bonding potential between these two atoms. If we were to use a surrogate at the electronic structure level, the ring structure would be constrained similarly by an additional penalty potential.

Selection

Individuals are selected with *tournament selection*: a subset of a given size s is randomly created from the population and a competition operates between individuals in this set. The solution with minimal (i.e., most optimal) fitness in the set is added to a pool of mates for crossing over. The process is repeated until the number of mates in the mating pool reaches the population size.

Crossover

The recombination of genetic material to be inherited by the offspring is achieved with the *simulated binary crossover* (SBX)[592] operator, which is a real-valued analogue of the single-point crossover of binary strings that was used in early GAs with discrete degrees of freedom. This simple operator cuts and swaps at one random site in the bit representations. More specifically, SBX is designed to enhance the probability for two parents to give birth to an arbitrary child solution and better explore the fitness landscape. More details about the SBX implementation are provided in Section 8.7.1. This operator demonstrated better performance in finding global optima of multivariable objective functions with numerous LM. To illustrate its enhanced search

power, we report the number of LM visited along a GA run with SBX in Figure D1 compared to simple swaps of $\mathbf{\Theta}$ components (*genewise crossover*) that cover less space on average.

Mutation

Mutations are random disturbances to ensure that all regions of the solution space are accessible during the search. A point in the solution space should in principle be reachable from any other point thanks to mutations (and their combination with crossover). However, in conventional GAs, mutations should not be too strong in order not to scatter promising features out of their optimal regions as long as the search improves. Mutations are therefore usually considered as rather local changes aimed at exploiting the vicinities of current solutions, whereas larger moves (explorations) are driven by crossovers.[104,565]

Dealing with a real-valued search space, an instinctive choice for mutations is the addition of Gaussian noise[593] that mutates an individual $\mathbf{\Theta}^i$ like

$$\tilde{\mathbf{\Theta}}^i = \mathbf{\Theta}^i + \mathbf{P}(p_m, K) \circ \sigma(\mathcal{N}_1(0,1), ..., \mathcal{N}_K(0,1)) \tag{8.3}$$

where $\circ$ denotes the element-wise multiplication between vectors. $\mathbf{P}(p_m, K)$ is a vector of size K filled with 0 or 1 that selects genes to be mutated with probability p_m. For p_m = 1, all genes are mutated, while p_m = 0 turns off the mutation. Selected genes are consequently modified with independent samples from the standard normal distribution $\mathcal{N}(0,1)$ scaled with the parameter σ that controls the mutation strength, along with p_m.

Elitism

The crossover and mutation operators mix and alter the tentative solutions that were among the best individuals in the previous generation. While the solutions are expected to improve along a GA run on average, there is no guarantee that the best fitness at a certain generation is lower than its previous counterpart and genes can drastically change in the case of genetic drift, escaping from a region where the GM actually sits. A way to counteract this is the application of an elitism operator which consists of replacing a fraction f of the worst solutions by the best individuals of the previous generation. This makes sure that the best fragments of information found so far are automatically transferred to the offspring generation, which thus always contains the overall best solution. Such a selective pressure can improve the convergence speed,[587,594] though the efficiency of any GA is dictated by its ability to balance between exploration and exploitation and elitism introduces the risk of losing diversity and converging prematurely to less-fit LM.[595]

8.4 Computational details

8.4.1 GA parameters

We report in Table 8.1 the parameters optimized through a series of test runs and finally used in this study. The algorithm terminates after a fixed number of generations, for which we verified that no significant improvement in fitness was observed anymore.

Table 8.1: Input parameters for EVOLVE

System	GPGG	Gramicidin
Population size	40	48
Number of generations	60	80
Tournament selection, set size s	2	
Mating probability	1.0	
Genewise crossover probability p_c	0.5	
SBX crossover order n	5	
Mutation probability	0.75	
Genewise mutation probability p_m	1/3	1/10
Mutation strength σ	60°	
Elitism fraction f (if applicable)	4/40	5/48

The mating and mutation probabilities fix the fractions of the population that are respectively crossed or mutated. For a solution $\mathbf{\Theta}$, p_c =50% of its components are crossed with the SBX operator while the others remain unchanged. 75% of the population is mutated and the probability p_m of mutating each gene is chosen so that one ϕ and one ψ backbone dihedral are modified on average. For gramicidin, the same probability applies to all 16 side chain dihedrals resulting in an average rate of 1.6 side chain mutants per individual. We choose a reasonable replacement of about 10% of the population by elites, unless otherwise specified, and also study the effect of no or stronger elitism in what follows.

8.4.2 Surrogate fitness function

Among the plethora of available methods, we focus our assessment of surrogate PES on some widely used force fields and semiempirical approaches that are expected to give fairly accurate results over a broad chemical and conformational space, as well as for charged or nonstandard residues.

The first chosen surrogate candidate is the General Amber Force Field (GAFF)[596] as provided in the Amber 2018 suite.[590] Fixed partial charges, atom types, and force field parameters have been assigned with the Antechamber and Leap tools. Atomic charges are derived from the default restrained electrostatic potential (RESP) fit[597] at the HF/6-31(d) level of theory. For the purpose of comparison and to test the sensitivity

with respect to the choice of fixed point charges, we also used charges derived with the faster Austin Model 1 with bond charge correction (AM1-BCC) scheme.[598,599] For both cis- and trans-GPGG, charges are calculated from structures constructed with the amino acid sequence editor of Molden,[600] while the X-ray-resolved crystal structure is used for gramicidin.[601] van der Waals and electrostatic interactions are not truncated in the absence of periodic boundary conditions. Local geometry optimizations are performed using Sander single-core jobs consisting first of 4000 steepest descent steps followed by conjugate gradient optimization until convergence to the default 10^{-4} kcal/(mol Å) root-mean-square deviation of the Cartesian elements of the gradient.

Second, we examine the AMOEBA polarizable force field for proteins[602] in its OpenMM implementation[52] with the L-BFGS minimizer tolerance set to 10^{-4} kcal/mol. In our experience, the GPU-accelerated version significantly speeds up geometry optimizations by up to a factor of 80 compared to the CPU version.

Calculations with the self-consistent-charge (SCC) density functional tight binding method[603] with full third order terms[604] (DFTB3) are performed with the DFTB+[605] code with the SCC tolerance set to 10^{-7} a.u. using the parameter set 3OB.[606] Hydrogen interactions are corrected with a damping exponent of 4.2 in the SCC short-range contribution.[604] DFTB3 is extended with the London dispersion correction D3[147] as parametrized for DFTB3[607] with the Becke-Johnson damping variant.[608] The geometry optimizations are carried out with the L-BFGS algorithm and default convergence criteria.

We also evaluate the ability to rely on hybrid DFT with a small basis set (6-31G) as a possible surrogate. For this, we use the GPU-supported TeraChem software[50,51] with L-BFGS optimizations[609] at the B3LYP level of theory performed on 2 parallel GPU cards with default settings.

Finally, Gaussian16[332] is used for the semiempirical PM6[610] and PM7[611] methods with the Berny optimizer[612] and default convergence criteria. The same is true for the B3LYP/6-31G(d,p) reference calculations with the difference being that very tight (tight) convergence criteria with ultrafine grid were chosen for GPGG (gramicidin). The performances of all these alternative surrogates compared to the B3LYP/6-31G(d,p) reference are discussed in the next section.

8.5 Results and discussion

8.5.1 Performance of different surrogate fitness functions

The success of the search for good candidate structures relies on the matching of the surrogate PES with the one of a reference method capable of reproducing the

experimental results. Ideally, running the GA with the surrogate should lead to a similar coverage of the configurational space as well as a good match of the relative energetics between structures within an affordable computational cost. We therefore seek to establish here which approximation provides the best compromise between accuracy and computational expense.

However, a quantitative evaluation of the performance of a given fitness function is nontrivial due to the stochastic nature of GAs, in addition to the intractable cost of running multiple benchmark instances with, for example, hybrid DFT. Furthermore, assessing accuracy differences between various methods has been one of the major challenges in computational chemistry for decades. For this reason, we rather test the quality of the different PES approximations on a finite test set of GPGG geometries.

In order to maximize the coverage of different regions of the PES, the set was generated from 10 high mutation rate GA instances with the GAFF force field and with the structures that resulted from GA crossovers and mutations before local relaxations (and fitness evaluations). Therefore, these latter are not LM of the GAFF force field which is only used to drive the sampling. From all visited configurations (20000 in total), 200 diverse geometries were initially selected using a farthest-point sampling (FPS) algorithm[613] in the space defined by the radius of gyration R_G and the number of hydrogen bonds N_H (see Section 8.7.2) that turned out to be useful for differentiating polypeptide configurations.[614] Among these, 146 geometries were successfully relaxed to distinct LM at the B3LYP/6-31G(d,p) reference level, which we augmented with 42 structures derived from ab initio SA (cf. Section 8.3.1) that we know correspond to low energy minima. Hence, the test set finally contains 69 cis and 119 trans nonrelaxed individuals that are representative of points potentially visited during GA runs.

As it would happen for a GA process, the different surrogates are employed to locally optimize the set and produce pools of respective LM. Therefore, the evaluation of a surrogate's performance must be based on its ability to not only approximate the energy but also the coordinates of the reference LM; a satisfactory model should provide target structures with relative energies following the B3LYP/6-31G(d,p) ranking at best. Illustratively, the wells of the surrogate in Figure 8.4 must be as "close" as possible to the reference wells, in terms of both energy and structure. However, as also depicted in Figure 8.4, we note that a direct (one-to-one) comparison between surrogate and reference LM is not possible because similar initial points may relax into very different geometries depending on the method and optimizer used. Consequently, in the absence of side chains for GPGG, the backbone RMSD (bb-RMSD) was chosen as a metric to identify "closest" LM structures and rely on a more faithful measure of proximity than a direct comparison from the shared initial point.

We opted for a statistical analysis to mimic the various fitness evaluations during a GA run, which also informs about the performance of the surrogates for different

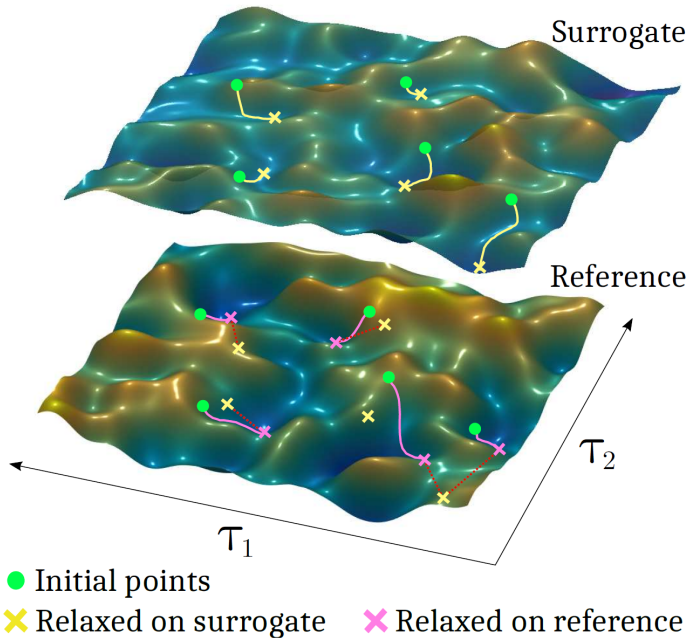


Figure 8.4: Illustration of the local relaxation of test structures on the surrogate and reference PES projected along two arbitrary reaction coordinates. The LM resulting from the same initial points are not necessarily close in energy and/or geometry.

population sizes: For a random subset of S initial structures taken from the test set, each reference B3LYP/6-31G(d,p) LM is associated with its closest (in terms of bb-RMSD) surrogate LM. Then, the relative energies within the subset are used to calculate the mean absolute error (MAE) of the surrogate energy:

$$\begin{aligned} \mathrm{MAE}(\Delta E) &= \frac{1}{S}\sum_{i=1}^{S}\left|\Delta E_i^{\mathrm{ref}} - \Delta E_i^{\mathrm{surr}}\right| \\ &= \frac{1}{S}\sum_{i=1}^{S}\left|E_i^{\mathrm{ref}} - E_{0,sub}^{\mathrm{ref}} - E_i^{\mathrm{surr}} + E_{0,sub}^{\mathrm{surr}}\right| \end{aligned} \tag{8.4}$$

with $E_{0,sub}$ being the minimum energy in the respective subset. This represents the ranking on which the GA selection would operate and avoids giving too much importance to whether the surrogate was able to correctly find the GM of the entire set or not.

Figure 8.5a shows the MAE(ΔE) for different subset sizes and surrogates, and Figure 8.5b gives the average bb-RMSD between the closest reference and surrogate LM structures from which the ΔE were calculated. As could be expected, the B3LYP/6-31G(d,p) LM are best reproduced using the same method (B3LYP) but with the smaller (nonpolarized) 6-31G basis set, with structures showing on average 0.2-0.4 Å bb-RMSD and $\sim$4 kcal/mol energy differences. While all other surrogates show similar differences in geometries that saturate at best around 0.3 Å bb-RMSD for DFTB3, the relative energies

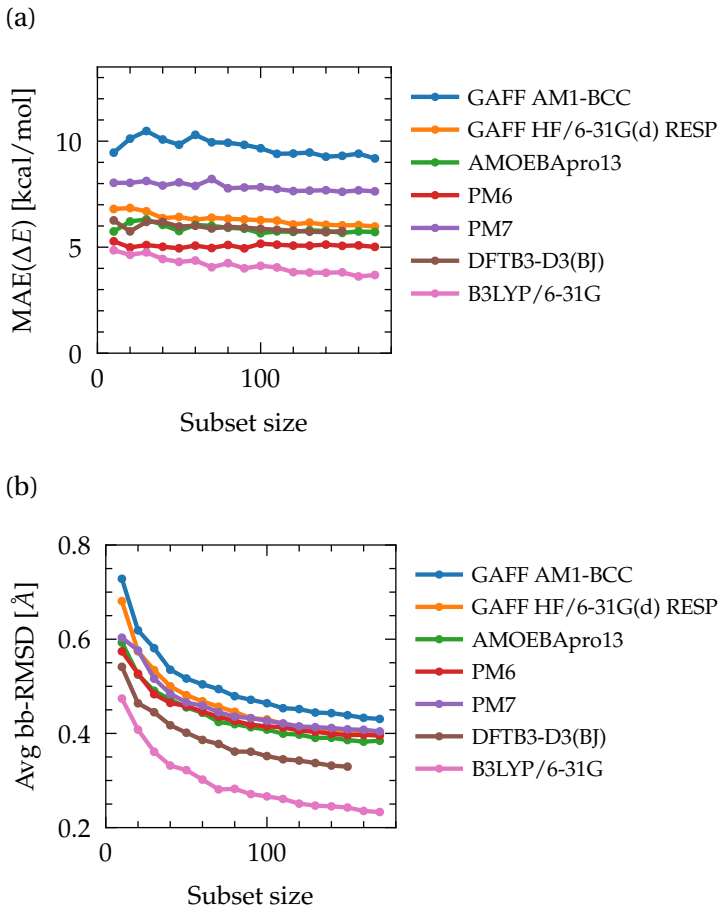


Figure 8.5: (a) MAE of relative energies between surrogate LM and their bb-RMSD closest B3LYP/6-31G(d,p) counterparts for the GPGG test set. (b) Average bb-RMSD between the surrogate LM and their closest B3LYP/6-31G(d,p) counterparts. Average values over $\max(S, 70)$ random subsets for each size S are plotted; standard deviations are of the order of 1 kcal/mol, respectively, 0.03 Å, and are provided in Appendix D (Figure D5a,b). The energies of the reference LM span 30 kcal/mol with two outliers around 40 and 60 kcal/mol.

between methods are more variable. The PM6 semiempirical method appears to perform best with average energy deviations of 5 kcal/mol, followed by the GAFF(RESP) and AMOEBA force fields, as well as DFTB3, which all have energy differences of about 6 kcal/mol while these exceed 7 kcal/mol for the remaining surrogates. The worst approach is the GAFF force field with AM1-BCC charges, which were only used here to explicitly test the influence of the effective charge set but are indeed not recommended for common practice.[590]

Regarding the quality of the structural prediction, Figure D2 gives an illustration of some bb-RMSD between reference and surrogate LM. In general, structures with a bb-RMSD of less than 0.5 Å are very similar and thus more likely to relax to their B3LYP/6-31G(d,p) counterpart upon re-optimization. As for the energetic performance, the

closest reproduction of the reference geometries is found for the smaller basis set B3LYP variant but also all remaining surrogate methods perform relatively well in terms of geometric predictions, yielding LM geometries with a bb-RMSD around 0.5 Å from a subset size of 40, which is the population size of the GA chosen for GPGG. Therefore, we conclude that over a set of representative geometries encountered in a GA optimization, all non-DFT surrogates show a similar performance in terms of reproducing B3LYP/6-31G(d,p) structures. However, correct relative energies are more difficult to approximate and differ between methods, with PM6 slightly outperforming GAFF(RESP), AMOEBA and DFTB3 in terms of overall MAE.

Apart from the fact that the surrogate method should be able to generate a diverse set of structures, we are particularly interested in the performance for the low-energy regime, whose members will drive GAs to the most optimal regions of the PES. The pool of low-energy LM at the surrogate level also represents the candidates that will be selected for a reoptimization with a higher level reference. For all cis, respectively trans isomers, Table 8.2 presents the MAE of the relative energies of LM at less than 10 kcal/mol of the respective GM in the set. Again, the energies are compared between corresponding pairs of surrogate-reference geometries that exhibit the smallest bb-RMSD.

Table 8.2: Assessment of surrogate methods in the low-energy regime $\Delta\tilde{E} \leq 10$ kcal/mol. N_{LM} is the number of local minima within the range. $\mathrm{MAE}(\Delta\tilde{E}) = \frac{1}{N_{\mathrm{LM}}}\sum_{i=1}^{N_{\mathrm{LM}}}\left|E_i^{\mathrm{ref}} - \tilde{E}_0^{\mathrm{ref}} - E_i^{\mathrm{surr}} + \tilde{E}_0^{\mathrm{surr}}\right|$ in kcal/mol where $\tilde{E}_0^{\mathrm{ref}}$ and $\tilde{E}_0^{\mathrm{surr}}$ are the respective cis or trans putative GM found over the entire test set. The energies of the surrogate and the reference are compared according to the smallest bb-RMSD match, whose average value and standard deviation are reported in Å.

	cis			**trans**		
Surrogate	MAE($\Delta\tilde{E}$)	Av bb-RMSD	N_{LM}	MAE($\Delta\tilde{E}$)	Av bb-RMSD	N_{LM}
GAFF AM1-BCC	7.1 ± 3.1	0.48 ± 0.25	10	5.9 ± 4.3	0.41 ± 0.12	15
GAFF HF/6-31G(d) RESP	6.1 ± 7.6	0.29 ± 0.25	10	6.1 ± 4.9	0.35 ± 0.17	29
AMOEBApro13	3.6 ± 4.1	0.36 ± 0.11	23	4.8 ± 4.7	0.27 ± 0.16	49
PM6	3.7 ± 4.6	0.37 ± 0.14	25	4.4 ± 3.4	0.23 ± 0.16	32
PM7	5.5 ± 6.7	0.43 ± 0.18	31	2.6 ± 1.9	0.21 ± 0.06	13
DFTB3-D3(BJ)	6.9 ± 4.7	0.36 ± 0.21	34	6.0 ± 6.7	0.36 ± 0.22	45
B3LYP/6-31G	1.6 ± 2.8	0.16 ± 0.17	30	1.1 ± 0.9	0.13 ± 0.13	35
B3LYP/6-31G(d,p), ref.			27			21

All methods provide on average satisfactory geometries with a difference in bb-RMSD of less than 0.5 Å with the reference LM. In addition to providing the closest structural match, the B3LYP/6-31G PES is the best surrogate with respect to relative energies in the low-energy realm. However, the performance of some of the other tested surrogates can markedly deviate from the overall energetic performance shown in Figure 8.5. Both GAFF(RESP) and DFTB3 are comparatively less accurate with MAEs between 6 and 7 kcal/mol for both cis and trans configurations. Although PM7 performs well for trans low-lying minima, it shows more weaknesses in ranking higher energy configurations

(Figure 8.5a) as well as cis isomers in the low-energy range, which highlights the fact that the performance of surrogates can be system-dependent. Finally, AMOEBA and PM6 exhibit the smallest MAEs of all non-DFT methods with balanced accuracies for the two configurational classes.

While the previous analyses assessed the quality of the structural as well as energetic predictions of the different surrogate methods, their overall computational cost also plays a major role in the choice of the most appropriate fitness function. To give an overview of the different time scales involved we give estimates of the average time needed for a local geometry optimization for each surrogate method in Table 8.3. For the sake of comparison, running a GA search with the B3LYP/6-31G(d,p) reference would take more than 2.5 months on a desktop workstation for the 39-atom GPGG molecule, highlighting the need for more expedient approaches. Although it is found that resorting to a smaller basis set provides the best accuracy, a GPU-accelerated implementation only reduces the elapsed time to the order of a month, while other surrogates bring it down to less than a day for semiempirical methods (PM6, PM7, DFTB3) and only few minutes for force fields (GAFF, AMOEBA).

Table 8.3: Average elapsed time $\bar{t}$ for local GPGG geometry optimization on N_{cores} cores (or GPU) for different surrogate methods based on the GPGG test set. $\bar{t}_{GA}$ is an estimate of the average time spent on fitness evaluations for a 60-generation 40-individual GA run if executed on a 24-core 2-GPU workstation[a]. The calculation details of each method are reported in Section 8.4.2.

Surrogate	$\bar{t}$ [min]	N_{cores}	$\bar{t}_{GA}$[a]
GAFF AM1-BCC	0.024	1[a]	3 min
GAFF HF RESP	0.028	1[a]	3 min
AMOEBApro13	0.005	1 GPU[b]	5 min
PM6	1.072	8[a]	15 h
PM7	1.578	8[a]	22 h
DFTB3-D3(BJ)	1.331	1[a]	2.7 h
B3LYP/6-31G	19.508	2 GPUs[b]	1 mth
B3LYP/6-31G(d,p)	150.235	8[a]	2.7 mths

[a] 24-core Intel Xeon E5-2650 v4 @ 2.20GHz CPU, 2 Nvidia GeForce GTX 1060 GPUs.
[b] 16-core Intel Xeon E5-2630 v3 @ 2.40GHz CPU, 2 Nvidia GeForce GTX 970 GPUs.

The small improvement in accuracy of PM6 does not seem to justify its use over AMOEBA, which is about 180 times faster. From these tests on the GPGG tetrapeptide, AMOEBA is thus emerging as a promising surrogate for GA optimization of peptides in terms of cost and accuracy and it will therefore be our choice in the following sections along with the fast but presumably less accurate GAFF (HF/6-31G(d) RESP) force field for comparison.

8.5.2 GA optimization of GPGG

Global minimum search

The results presented here are all based on a common pool of surrogate geometries generated after 10 GA runs, for which the minimum energy progressions are plotted in Appendix D (Figures D7 and D8). Without prior knowledge about structures and energies, the GM is assumed to be the lowest-energy individual found over all instances.

In terms of GA performance, it is worth mentioning that elitism markedly increases the chance of finding the GM as reported in Figure 8.6 that shows the cumulative success of reaching the GM at a given iteration. A 10% replacement of the current population with the best parent individuals substantially improves the GM search for all schemes but the cis-GPGG on the AMOEBA PES due to its rapid convergence (the energy decrease between the first and last generations is only 0.1/0.6 kcal/mol as shown in Figure D7). For the other cases, the GA might not always succeed in finding the GM but elitism allows enhancement of the convergence rate by 30%. Since the minimum energy will fix the overall ranking of surrogate LM, and consequently the selection of candidates to be reoptimized, it is essential for the GA to reach the surrogate GM or at least low-lying structures within a few kcal/mol from it.

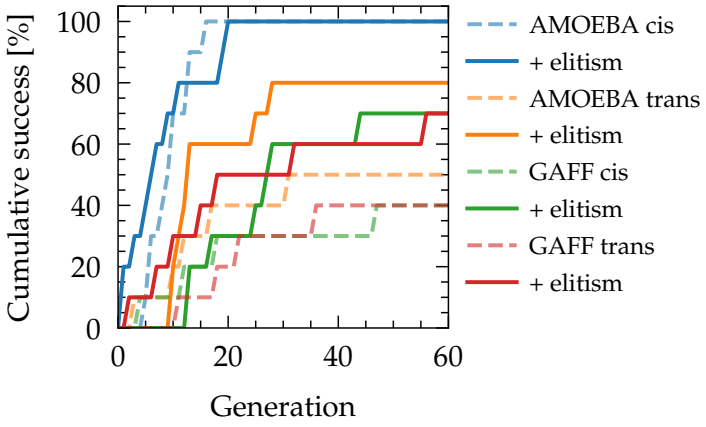
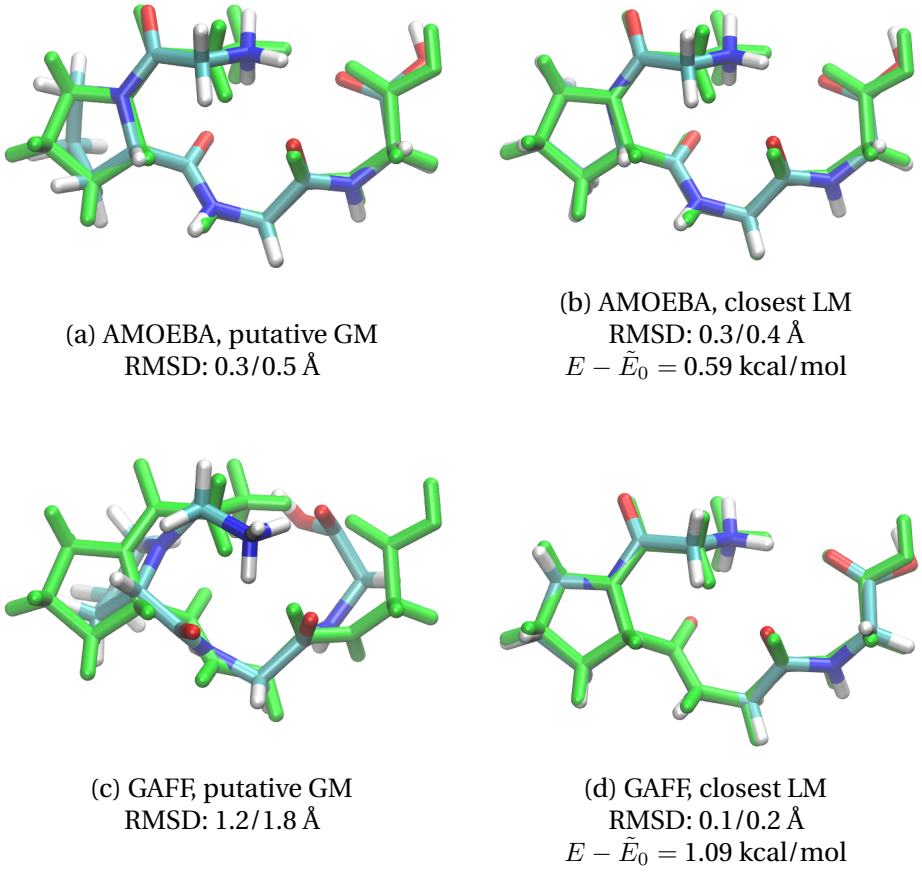


Figure 8.6: GPGG: cumulative success of finding the surrogate GM at each generation, averaged over 10 GA optimizations per surrogate/isomer combination.

Comparing the sampled structures of the cis isomer with the DFT-resolved GM, it is found that the putative GM of AMOEBA has a heavy-atom RMSD of only 0.5 Å (Figure 8.7a). However, this is not the most similar structure found, as the GA was able to provide an even closer structure with an RMSD of 0.4 Å about 0.6 kcal/mol higher in energy that better reproduces the configuration of the proline cycle (Figure 8.7b). At the GAFF level, although the lowest energy structure is more compact (Figure 8.7c)



(a) AMOEBA, putative GM
RMSD: 0.3/0.5 Å

(b) AMOEBA, closest LM
RMSD: 0.3/0.4 Å
$E - \tilde{E}_0 = 0.59$ kcal/mol

(c) GAFF, putative GM
RMSD: 1.2/1.8 Å

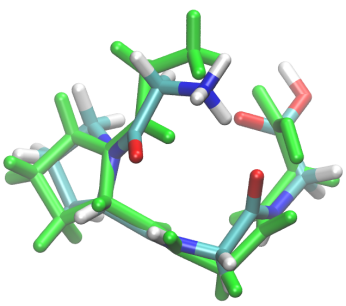
(d) GAFF, closest LM
RMSD: 0.1/0.2 Å
$E - \tilde{E}_0 = 1.09$ kcal/mol

Figure 8.7: cis-GPGG: putative GM and closest LM found on surrogate PES after 10 GA runs for (a, b) AMOEBA and (c, d) GAFF. $E - \tilde{E}_0$ is the relative energy of the LM with respect to the putative GM. The B3LYP/6-31G(d,p) GM is depicted in green with respective backbone/heavy-atom RMSD. Similar structures are obtained with or without elitism.

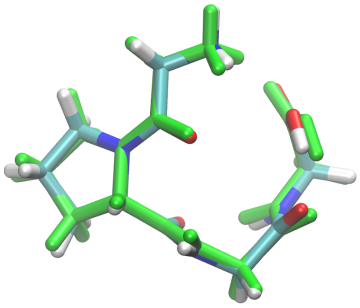
and therefore shows a larger RMSD from the reference, a geometry almost equal to the DFT GM is also discovered about 1 kcal/mol higher in energy (Figure 8.7d).

For the trans isomer, the AMOEBA GM is more distant from the DFT reference (Figure 8.8a) than it is with GAFF (Figure 8.8c) with respective RMSDs of 1.1 Å against 0.6 Å, but both force fields yield almost identical structures to the DFT GM within 2 kcal/mol above their putative GM (Figures 8.8b and 8.8d).

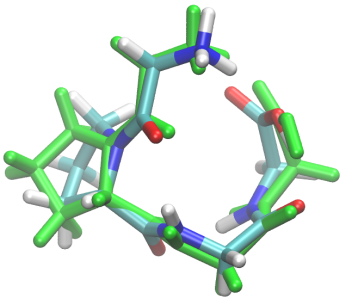
Hence, for both cis- and trans-GPGG, GAFF and AMOEBA are able to identify the DFT GM as a low-lying surrogate structure within a maximum of 2 kcal/mol above their (putative) GM demonstrating that a surrogate approach can indeed be beneficial before resorting to higher-level refinement as it is done in the next section.



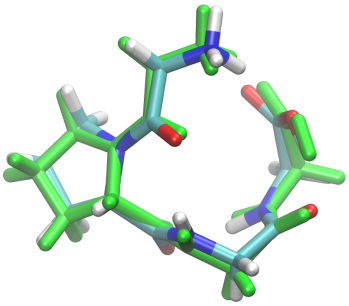
(a) AMOEBA, putative GM
RMSD: 0.7/1.1 Å



(b) AMOEBA, closest LM
RMSD: 0.1/0.2 Å
$E - \tilde{E}_0 = 1.99$ kcal/mol



(c) GAFF, putative GM
RMSD: 0.4/0.6 Å



(d) GAFF, closest LM
RMSD: 0.3/0.3 Å
$E - \tilde{E}_0 = 1.83$ kcal/mol

Figure 8.8: trans-GPGG: putative GM and closest LM found on surrogate PES after 10 GA runs for (a, b) and AMOEBA (c, d) GAFF. $E - \tilde{E}_0$ is the relative energy of the LM with respect to the putative GM. The B3LYP/6-31G(d,p) GM is depicted in green with respective backbone/heavy-atom RMSD. Similar structures are obtained with or without elitism.

Low-lying minima search and refinement

GA optimization offers the additional advantage that one can profit from all of the LM visited during evolution. Maximizing the number of low-energy structures is therefore important in order to capture all surrogate candidates likely to relax to the desired reference minimum. As an example, the progression of the number of new minima explored is shown in Figure 8.9 for a single GA as well as after several executions. The average number of minima found over the GA iterations is very similar with or without elitism and reaches a plateau after a certain number of GA generations (Figure 8.9a). Figure 8.9b shows that it is more efficient to perform independent runs in parallel to improve the search and sample more LM, rather than extending a single execution with more generations. In this case, however, the use of elitism can alter diversity and reduce the exploration of low-lying LM, which is observed for all schemes (Figure D9).

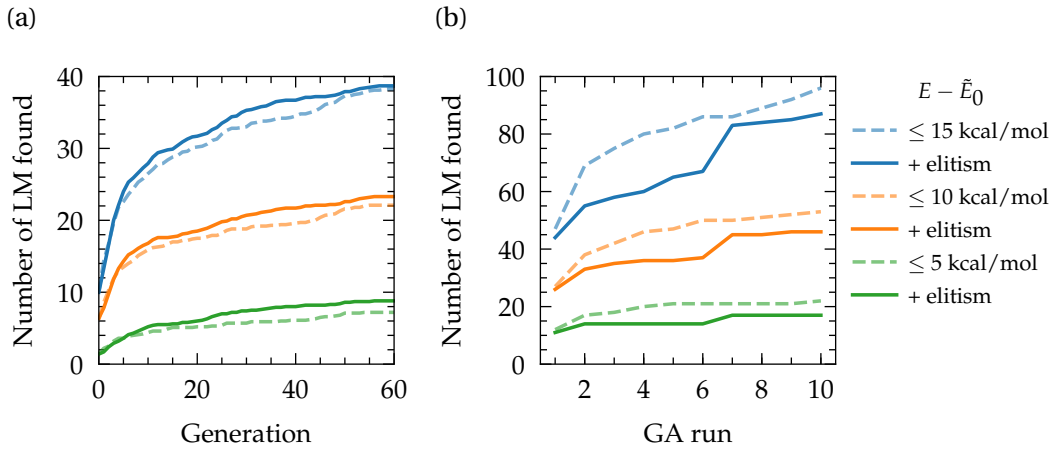


Figure 8.9: trans-GPGG: number of low-lying minima found on the AMOEBA PES within 15, 10, and 5 kcal/mol with respect to the putative GM. (a) Per GA generation, averaged over 10 GA runs. (b) By running independent GAs. Distinct LM are taken to be at least separated by 10^{-4} kcal/mol and 0.2 heavy-atom RMSD.

The collection of thousands of structures provided by the GA is followed by their ultimate reoptimization at the reference level. For this purpose, only surrogate structures within 10 kcal/mol of their putative GM are selected and locally relaxed with DFT (B3LYP/6-31G(d,p)). To establish the actual accuracy of the surrogate, Figure 8.10 compares the energies and coordinates between the AMOEBA LM and their reoptimized counterparts and Table 8.4 reports respective MAE on energy and bb-RMSD for all schemes.

Table 8.4: GPGG: MAE$(\Delta\tilde{E})^a$ in kcal/mol and average backbone RMSD in Å between the surrogate LM and their reoptimized counterparts at B3LYP/6-31G(d,p). N_{LM} is the number of LM reoptimized within 10 kcal/mol from the putative GM.

Surrogate/isomer	MAE$(\Delta\tilde{E})$	Av bb-RMSD	N_{LM}
AMOEBA/cis	1.9 ± 1.5	0.28 ± 0.10	31
GAFF/cis	2.9 ± 1.9	0.29 ± 0.14	94
AMOEBA/trans	4.4 ± 2.9	0.31 ± 0.21	64
GAFF/trans	4.0 ± 2.5	0.38 ± 0.23	87

[a] as defined in Table 8.2.

As expected, relative energies are not perfectly reproduced and the largest errors (outliers) cannot be systematically attributed to larger RMSDs. However, AMOEBA provides a rather good MAE of only 1.9 kcal/mol for the cis isomers, while it increases to 4.4 kcal/mol for the trans structures. Surprisingly, the relative energies obtained for the GAFF (fixed point charge) force field are only slightly worse than the ones of AMOEBA for cis and even slightly better for trans isomers. In spite of the few kcal/mol error in the predictive power of the surrogates, LM are generally very close to their DFT

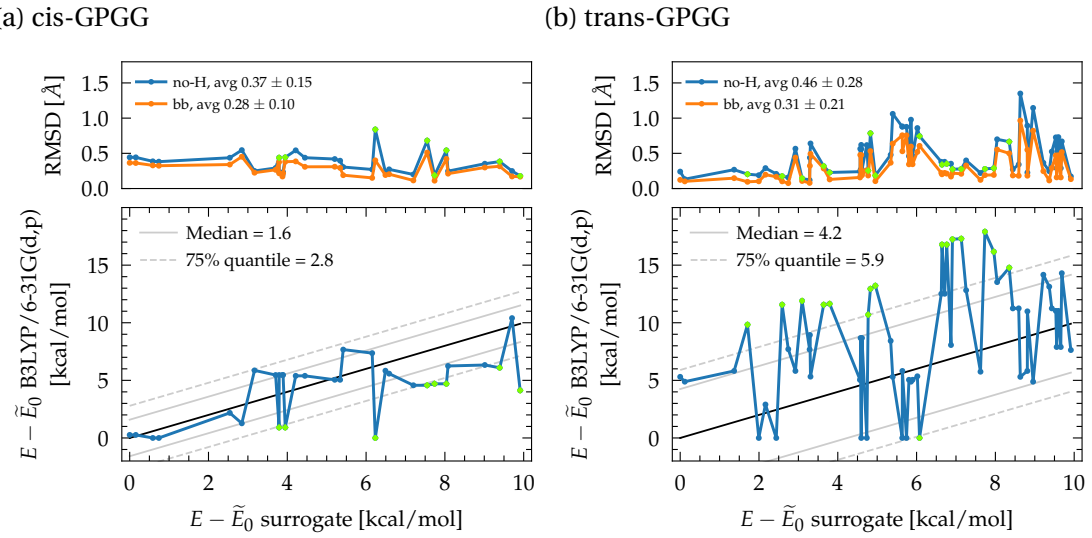


Figure 8.10: GPGG: predictive performance of AMOEBA in reproducing geometries and relative energies of LM at the B3LYP/6-31G(d,p) level for (a) cis isomers and (b) trans isomers. Backbone (bb) and heavy-atom (no-H) RMSDs are reported. Also indicated are the median and 75% quantile of absolute errors on energies. The 75% quantile outliers are marked in green with their respective RMSD. Corresponding plots for the GAFF force field are given in Appendix D (Figure D10).

counterparts with a small backbone (heavy-atom) RMSD around 0.3 (0.5) Å on average. Some of them relax into identical minima on the DFT PES, but the GA candidates still provide an extensive set of realistic low-lying minima: We note that all LM that were identified as closest to the DFT GM for AMOEBA (Figures 8.7b and 8.8b) and GAFF (Figures 8.7d and 8.8d) did indeed relax to the DFT GM. Therefore, the surrogate GA approach was overall successful in retrieving the target DFT GM structures that were assigned to experimental IR spectra.

Compared to a previous SA search,[107] the sGADFT found more (theoretical) LM on the B3LYP PES within the convergence criteria and basis set employed, as it is shown in Figure 8.11. In the cis subspace, the AMOEBA GA gave four similar lowest-energy geometries to SA within 2 kcal/mol and misses four of them within 5 kcal/mol. Nevertheless, it provides additional structures that were not found in the SA search. Ditto for the GAFF force field, except for some very low energies that are not recovered. For the trans-GPGG, the very low region is more sparse and a structure at 0.05 kcal/mol is missed with AMOEBA, as is another one close to 5 kcal/mol that was spotted with GAFF. Overall, this demonstrates that GA-sampled structures are indeed relevant for the low-energy resolution of the ab initio PES. Should the results not be satisfactory, there is always the possibility of running additional GAs and/or performing a higher number of reoptimizations.

The obtained ab initio LM can describe very similar structures that are chemically

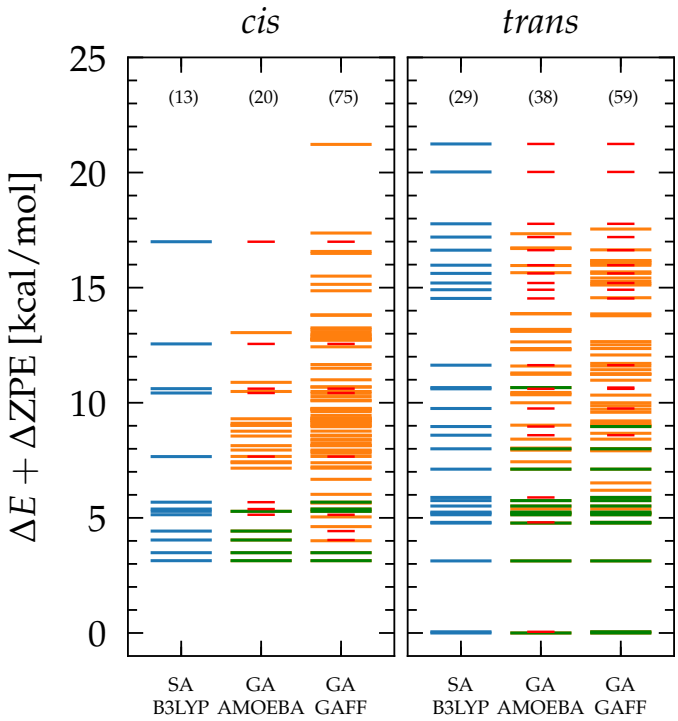


Figure 8.11: GPGG: zero point energy-corrected energies of B3LYP/6-31G(d,p) reoptimized structures obtained with ab initio simulated annealing (SA) at B3LYP/6-31G[107] and GA with AMOEBA or GAFF force fields. The green and shorter red levels are respectively matches and misses compared to SA. The respective number of LM is indicated in parentheses.

indistinguishable. To group essentially identical structures, a minimum RMSD can be imposed and the number of distinct LM becomes of the same order for both AMOEBA and GAFF (Figures D11 and D12). This shows that it is important to sample not only as many low-lying minima as possible at the surrogate level but also those that are farthest away and likely to relax into distinct DFT minima. An effective approach in this sense would be to select distant structures using clustering,[615] FPS,[613] or RMSD analysis prior to reoptimization and avoid irrelevant relaxations due to small numerical differences.

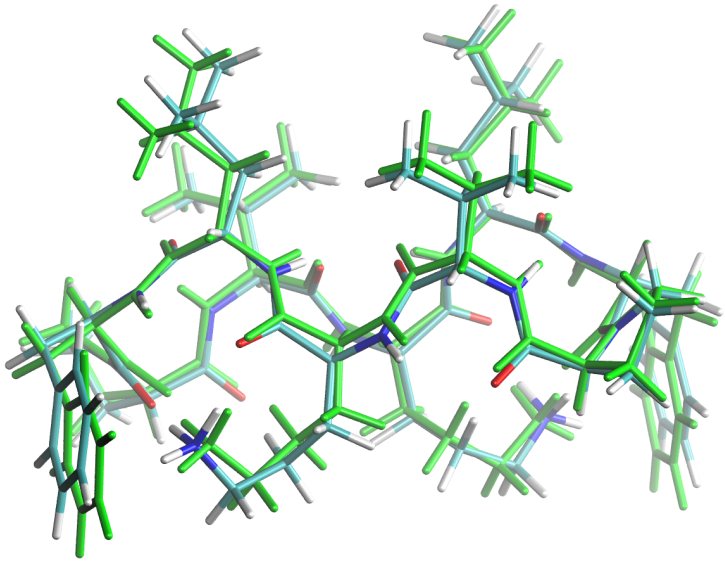
8.5.3 GA optimization of Gramicidin

Global minimum search

An even harder performance test is represented by the larger gramicidin system with explicit side chain optimization. As reported in Figure D13 using the AMOEBA surrogate PES the energy progression is clearly hampered or stagnates after a few tens of

iterations in the absence of elitism. On the other hand, a (too) high fraction of elitism of 20% increases the variance and does not reach the lowest energies, whereas the putative GM is finally found in 3 over 10 GA runs using a medium 10% rate of elites. Astonishingly, the surrogate GM is also the closest geometry to the DFT-resolved structure, which are both reproduced in Figure 8.12. The agreement between the AMOEBA and the B3LYP geometries is remarkable with a backbone (heavy-atom) RMSD of only 0.2 (0.4) Å.

(a)



(b)

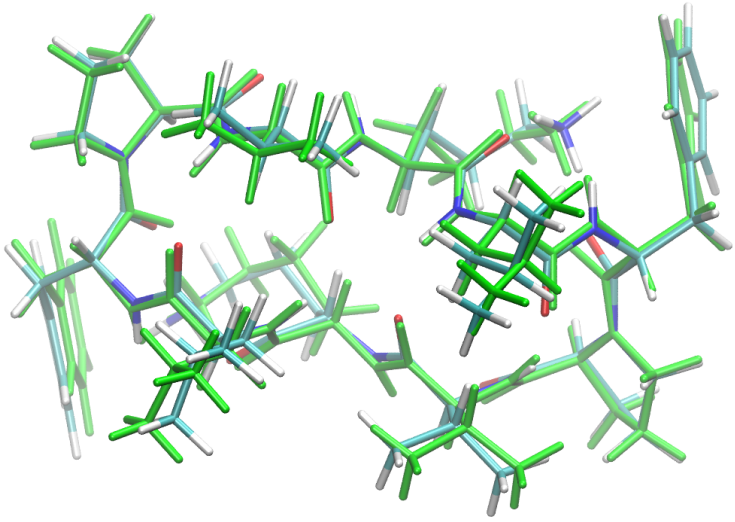


Figure 8.12: Gramicidin: two views (a) and (b) of the AMOEBA putative GM that is also the closest LM found over 10 GA runs (with 10% elitism). The DFT B3LYP/6-31G(d,p) reference GM is depicted in green. Backbone/heavy-atom RMSDs are 0.2/0.4 Å.

In contrast, the GAFF putative GM is found with a 20% elitism rate (Figure D13), which highlights the fact that elitism is an essential factor in the search for GM on complicated PES, but its magnitude may be system- and method-dependent and remains a parameter to be assessed or adjusted in order to find an ideal exploration-exploitation trade-off. As opposed to AMOEBA, the GAFF putative GM is very far from the DFT GM with a large (2.1 Å) backbone RMSD (Figure D14a). Over 30 GA runs, the closest LM found is only located within 26 kcal/mol (!!) from the putative GM, has a 0.5 (1.5) Å backbone (heavy-atom) RMSD and does not relax to the DFT GM (Figure D14b). In order to assess if this poor performance originates from the limitation of the GA search or the quality of the surrogate PES, we relaxed the DFT GM with the GAFF force field and obtained a very similar structure (0.1 (0.3) RMSD) located 18 kcal/mol above the GAFF putative GM. Therefore, the DFT GM is indeed a LM on the GAFF PES but does not lie in the low-energy regime which definitely renders the GAFF force field unsuitable for the GM search, in particular because several hundreds of structures were found within 18 kcal/mol (Figure D15) and, in the absence of prior knowledge, reoptimizing all would be far from tractable.

Low-lying minima search and refinement

As seen previously, the number of explored minima depends on the ability of the GA to reach different low-energy regions and varies with the fraction of elitism and the choice of fitness function. Running multiple GA instances starting from different initial structures is again more efficient than extending a single run whose variance decreases with the number of iterations (Figure D16). Elitism reduces in principle the overall diversity of the LM (cf. Section 8.5.2) but becomes essential to explore the very low energy regions of more complex systems. Indeed, for gramicidin, the greater number of low-energy minima was obtained by the elitism fraction capable of identifying the putative GM (Figures D15 and D16). Therefore, mitigating elitism with other mutation-like operators could potentially improve the search power in the low-energy regime by providing a certain diversity of structures visited while maintaining low energies.

B3LYP/6-31G(d,p) reoptimizations of gramicidin candidates are much more CPU intense than those of the smaller GPGG peptides, so that only structures sampled with 10% elitism and located within 5 kcal/mol could be retained for subsequent optimizations. It is therefore crucial that the surrogate, although not optimal, provides relevant candidates located in a range of only a few kcal/mol, as the number of structures and their refinement cost increase considerably with the size of the system. Again, we plot relative energies and RMSDs against DFT reoptimized geometries in Figure 8.13, and report averages in Table 8.5.

Table 8.5: Gramicidin: MAE($\Delta\tilde{E}$)[a] in kcal/mol and average backbone RMSD in Å between the surrogate LM and their reoptimized counterparts at B3LYP/6-31G(d,p). N_{LM} is the number of LM reoptimized within 5 kcal/mol from the putative GM, separated at least by 10^{-4} kcal/mol and 0.75 heavy-atom RMSD.

Surrogate	MAE($\Delta\tilde{E}$)	Av bb-RMSD	N_{LM}
AMOEBA	4.3 ± 3.3	0.29 ± 0.11	48
GAFF	17.9 ± 2.0	0.25 ± 0.05	28

[a] as defined in Table 8.2.

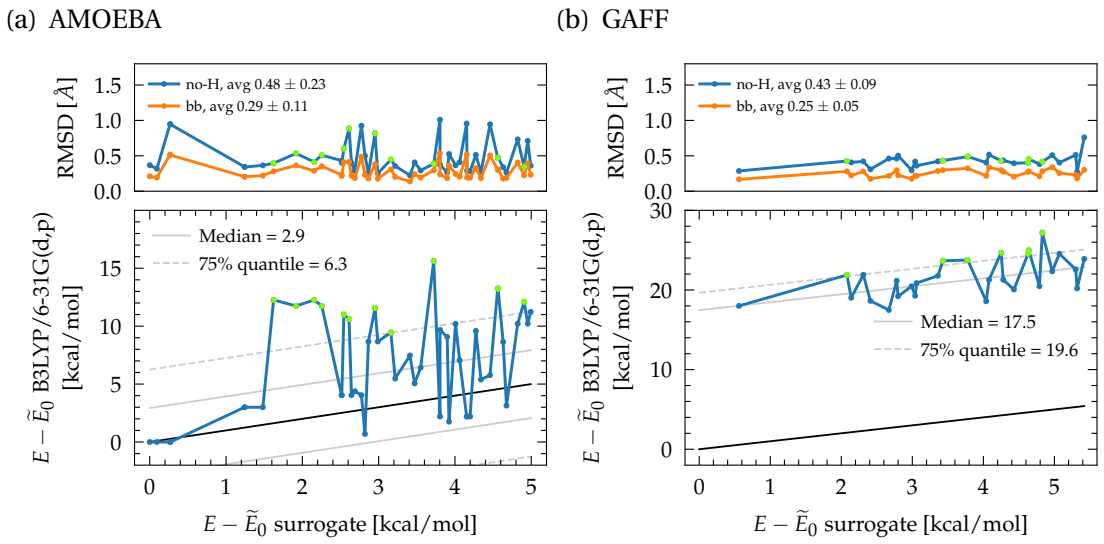
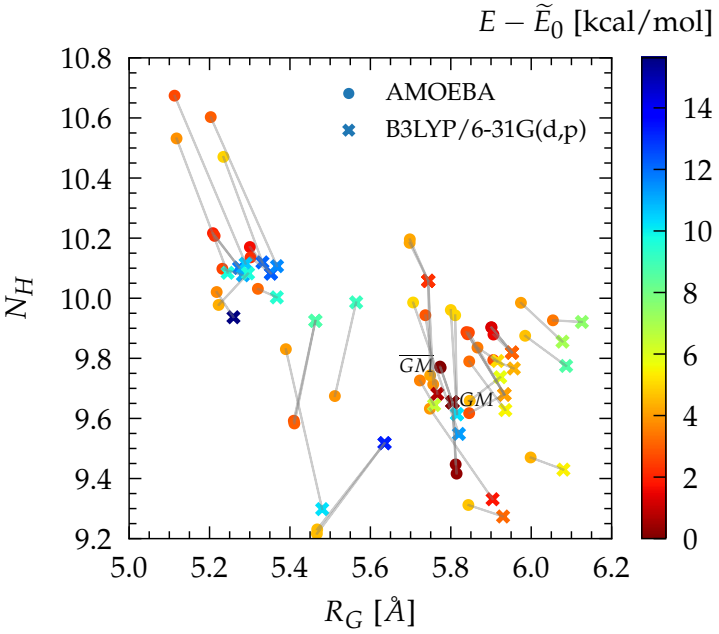


Figure 8.13: Gramicidin: predictive performance of AMOEBA and GAFF in reproducing geometries and relative energies of LM at the B3LYP/6-31G(d,p) level. Backbone (bb) and heavy-atom (no-H) RMSDs are reported. Also indicated are the median and 75% quantile of absolute errors on energies. The 75% quantile outliers are marked in green with their respective RMSD.

For gramicidin, AMOEBA performs as well as for GPGG with a MAE around 4 kcal/mol and small 0.3 Å average backbone RMSD. Successfully, the three lowest candidate structures relax to the DFT GM (Figure 8.13a). For GAFF, the geometries of the candidate structures are also very similar to their closest DFT minima, but relative energies are significantly off due to GAFF's inability to correctly reproduce the lower regions of the DFT PES of this system. By visualizing the LM in the R_G-N_H space in Figure 8.14, we notice that GAFF biases the search toward higher-energy DFT regions. In these, GAFF does rather well on relative energies despite a large energy offset (∼18 kcal/mol, Figure 8.13b). Hence, we conclude that GAFF cannot reliably approximate the energetics for screening low-energy gramicidin structures and sampling realistic regions of the PES, which the polarizable AMOEBA force field, on the other hand, seems to achieve surprisingly well.

(a) AMOEBA



(b) GAFF

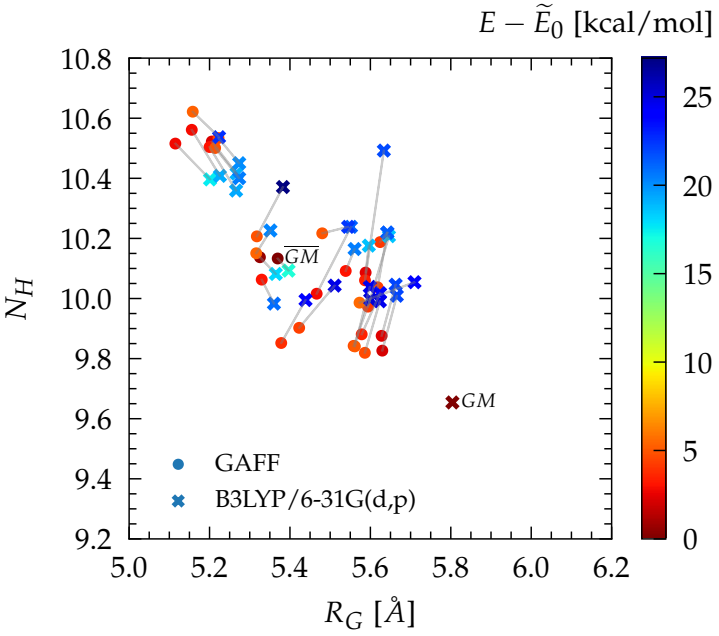


Figure 8.14: Gramicidin: surrogate AMOEBA and GAFF LM candidates within 5 kcal/mol in the R_G (radius of gyration) and N_H (number of hydrogen bonds) space, connected by lines to their reoptimized structures at the B3LYP/6-31G(d,p) level of theory. $\tilde{E}_0$ are the respective energies of the putative GM for each PES, indicated by $\overline{GM}$ for the surrogates and GM for B3LYP.

Figure 8.15 finally demonstrates that the straightforward GA approach with AMOEBA produces an extensive set of low-lying B3LYP/6-31G(d,p) structures with little effort, as opposed to the more technically involved restrained SA simulations that were used in the initial search for the experimentally observed structure[552] (cf. Section 8.3.1). Although the overall sGADFT method did not find similar LM, its explored energy space is denser in the low-energy range and the experimental GM is retrieved, advocating the use of surrogate GAs for low-energy sampling with little setup management and cost, as discussed in the next section.

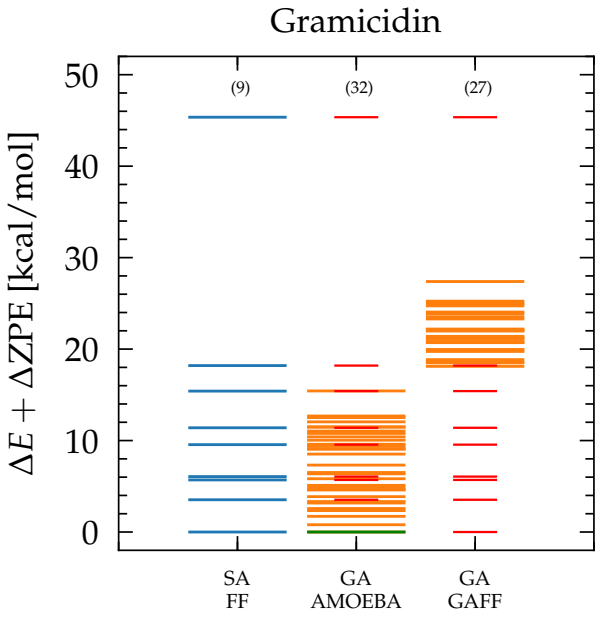


Figure 8.15: Gramicidin: zero point energy-corrected energies of B3LYP/6-31G(d,p) reoptimized structures obtained with SA based on classical force fields (SA/FF),[552] and GAs with AMOEBA and GAFF force fields. The green and shorter red levels are respectively matches and misses compared to SA/FF. The respective number of LM is indicated in parentheses. A similar plot restricted to clearly distinct structures (only LM differing by at least two side chain dihedrals) is provided in Figure D18.

8.5.4 Computational performance

Thanks to the use of surrogates that allows one to bypass a direct exploration of the PES at the first-principles level, searching for GPGG conformers takes less than 15 min on a conventional workstation as indicated in Table 8.6, albeit requiring more than 2400 local relaxations per GA execution.

Table 8.6: Wall time $\bar{t}$ per GA execution for the AMOEBApro13(OpenMM[52]) and GAFF(Amber[590]) surrogates. Computational settings correspond to Section 8.4. Averages and standard deviations are given for 20 GA runs on a workstation with 24-core Intel Xeon E5-2650 v4 @ 2.20GHz CPU and 2 Nvidia GeForce GTX 1060 GPUs. Time differences with or without elitism are insignificant.

System	$\bar{t}$ AMOEBA [min]	$\bar{t}$ GAFF [min]
cis GPGG	13.9 ± 0.6	14.3 ± 1.6
trans GPGG	12.5 ± 0.1	7.5 ± 0.8
Gramicidin	43.7 ± 3.4	122.8 ± 11.2

For GAFF, benefits come from a parallel split of fitness evaluations over multiple cores. The timing difference between cis- and trans-GPGG originates from the longer initialization of a complete cis population that has the tendency to relax to trans structures. Apart from that, for the smaller GPGG system, GAFF is generally faster than the more sophisticated (polarizable) AMOEBA force field but the recent GPU-accelerated implementation[52] of AMOEBA makes the optimizations significantly faster for the larger gramicidin peptide; thanks to a load split of EVOLVE over two parallel GPUs, the evaluation of more than 3800 fitness evaluations can be achieved in less than 45 min for this 176 atom molecule.

All in all, in the case of AMOEBA, the pools of low-energy candidate structures for GPGG and gramicidin were sampled in respectively 2.5 and 7 h on a single workstation for 10 serial GA runs, without monitoring or restart procedures, in contrast to the previously employed B3LYP/6-31G SA search for GPGG that took several days with multiple runs, with different heating temperatures and annealing rates and a postprocessing analysis of trajectories to extract promising candidates.[107] Such an ab initio exploration is simply out of reach for gramicidin and only SA based on classical force fields employing additional experimentally observed constraints could provide the GM.[552]

Regardless of the search approach employed, a final ab initio refinement with a large basis set is necessary for calculating properties, e.g., reliably assigning IR frequencies to experimental spectra. DFT reoptimizations and (harmonic) vibrational analyses are far more demanding than the GA searches themselves; in fact, they required 6 days on two workstations for all AMOEBA/GAFF GPGG LM (276 structures) while 4 days on 8 16-core compute nodes were needed for the AMOEBA gramicidin (48 structures). However, similar to the previous SA searches, experimental information like ion-mobility cross-sections[107] or symmetry constraints derived from typical vibrational fingerprints[552] can be used as additional prefilter for GA applications, to further narrow down the pool of candidate geometries instead of retaining all structures within a given energy range. This would drastically reduce the computational demand when treating larger systems.

8.6 Conclusions and outlook

In this chapter, we have presented a GA based search method to efficiently sample low-energy structures of peptides and its implementation in our in-house code EVOLVE.[583] Rather than aiming for a full first-principles exploration, we argue that resorting to more expedient surrogates allows significant reduction of the computational expense in the screening of candidate structures to be later reoptimized at the ab initio level. This is motivated by the fact that coordinates of local minimum candidates are in general well-approximated by surrogates, while getting reliable energies is the main difficulty.

Among several approximate methods investigated, the AMOEBApro13 polarizable force field showed the best compromise between cost and accuracy. Tested on three systems that are the cis-, trans-proline protonated GPGG and the doubly protonated gramicidin S decapeptide, the approach was successful in identifying B3LYP DFT GM within a maximum 2 kcal/mol from the putative surrogate GM. The GAFF force field also succeeded for GPGG isomers but failed for gramicidin due to a large offset in the energy predictions. As opposed to the more cumbersome and expensive ab initio simulated annealing employed in earlier studies, GPGG local minima were generated over 10 serial GA runs in less than 3 h on a single workstation, and only 7 h were necessary for the larger gramicidin system. Obviously, these timings can be further improved by parallelizing between multiple GA instances.

Overall, this demonstrates that the AMOEBA based surrogate GA alternative can provide substantial advantages in the three-dimensional determination of trapped metastable or global minimum peptide structures, as observed in ultracold spectroscopy, because all resulting GM coordinates were indeed correctly identified.

Thinking ahead, such a comprehensive generation of low-energy minima can also be advantageous for a wider range of research studies: for example as starting points for MD simulations, free-energy sampling, transition state searches, and nudged elastic band methods, or as templates for protein-ligand complexes in the rational design of analogues, or finally as training data for a variety of machine learning approaches.[8,87,616]

8.7 Additional details

8.7.1 Simulated binary crossover

Let Θ_k^i be the kth real-coded gene (component) of the parent individual i with representation of size K (eq 8.1). For a two-to-two crossover between individuals i and j, the spread factor β is defined as the ratio of the distance between children points $\tilde{\Theta}_k^i$ to that of the parent points:

$$\beta = \frac{|\tilde{\Theta}_k^i - \tilde{\Theta}_k^j|}{|\Theta_k^i - \Theta_k^j|} \tag{8.5}$$

such that for $\beta < 1$ ($\beta > 1$), the spread of the children points is smaller (larger) than that of the parents and has a contracting (expanding) effect on the children extent. Deb and Agrawal[592] showed that the probability distribution $\mathcal{P}$ of having a contracting or expanding single-point binary crossover with spread β can be approximated by polynomial functions, such that

$$\mathcal{P}(\beta) = \begin{cases} \frac{1}{2}(n+1)\beta^n & \beta \leq 1 \\ \frac{1}{2}(n+1)\beta^{-(n+2)} & \beta > 1 \end{cases} \tag{8.6}$$

is used to design a real-value crossover, where n between 2 and 5 appeared to match closely with single-point crossover results. It is easy to show that contracting or expanding the distance between children genes is equiprobable (with 0.5 probability) by integrating $\mathcal{P}$ in the respective ranges. Figure 8.16 shows the probability distribution of eq 8.6 for different n. Generally, the probability of creating children close to their parents ($\beta = 1$) is higher than creating very different children. Larger values of n accentuates this effect. In practice, a fixed n is chosen although one could broaden the initial search with small n and progressively narrow the exploration over generations with larger n.

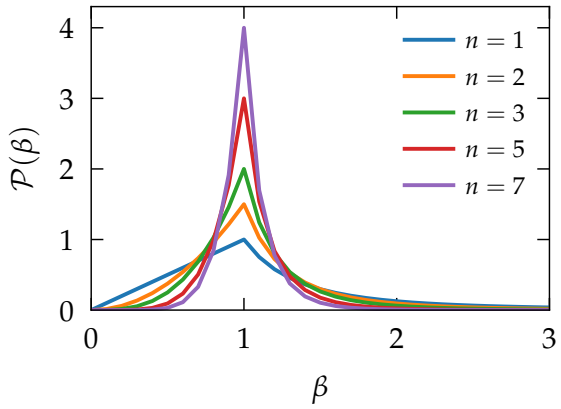


Figure 8.16: Probability distributions $\mathcal{P}(\beta)$ (eq 8.6) of contracting and expanding SBX crossover to mimic binary single-point crossover distributions.

A sample from this probability distribution is generated by choosing the point $\bar{\beta}$ for which the cumulative probability $\int_0^{\bar{\beta}} \mathcal{P}(\beta)d\beta = u$, where u is a uniformly generated random number in $[0,1)$ and the change in contracting or expanding $\mathcal{P}(\beta)$ occurs at $u = 1/2$. For such a $\bar{\beta}$, the children's genes are crossed according to

$$\begin{aligned} \tilde{\Theta}_k^i &= \frac{1}{2}\left[(1-\bar{\beta})\Theta_k^i + (1+\bar{\beta})\Theta_k^j\right] \\ \tilde{\Theta}_k^j &= \frac{1}{2}\left[(1+\bar{\beta})\Theta_k^i + (1-\bar{\beta})\Theta_k^j\right] \end{aligned} \tag{8.7}$$

Up to now, the SBX operator has been presented for unbounded variables, whereas peptide dihedrals are periodic. To restrict the search space to specified lower ($l_b = -180°$) and upper ($u_b = 180°$) bounds used throughout the GA, the probability distributions are modified so that the probability of creating dihedrals outside of the bounds is equal to zero; without loss of generality in what follows, one can assign the largest value to Θ_k^i ($\tilde{\Theta}_k^i$) and the lowest to Θ_k^j ($\tilde{\Theta}_k^j$). It is straightforward to notice from eq 8.5 that a maximum spread allowed for $\tilde{\Theta}_k^i - \tilde{\Theta}_k^j$ can be chosen as

$$\beta_{max} = 1 + \frac{2\min(\Theta_k^j - l_b, u_b - \Theta_k^i)}{\Theta_k^i - \Theta_k^j} \tag{8.8}$$

which provides a scaling factor α for the probability distribution in order to make the overall cumulative probability in the bounds equal to one:

$$\alpha := \int_{\beta_{min}=0}^{\beta_{max}\geq 1} \mathcal{P}(\beta)d\beta = 1 - \frac{1}{2}(\beta_{max})^{-(n+1)} \tag{8.9}$$

Therefore, the bounded crossover operates with eq 8.7 and $\bar{\beta}$ is generated from the normalized cumulative probability $\int_0^{\bar{\beta}} \frac{1}{\alpha}\mathcal{P}(\beta)d\beta = u$, where u is a uniformly sampled random number in $[0,1)$. The normalized probability distribution consequently changes at $u = 1/2\alpha$ such that

$$\bar{\beta} = \begin{cases} (2\alpha u)^{\frac{1}{(n+1)}} & u \leq \frac{1}{2\alpha} \\ \left(\frac{1}{2-2\alpha u}\right)^{\frac{1}{(n+1)}} & u > \frac{1}{2\alpha} \end{cases} \tag{8.10}$$

The extension of the single-variable (Θ_k^i) SBX operator to the multivariate problem is straightforward: setting a probability p_c of crossing over, $p_c K$ respective components of the solutions $\mathbf{\Theta}^i$ and $\mathbf{\Theta}^j$ are selected and crossed with the single-variable operator described above.

8.7.2 Radius of gyration and number of hydrogen bonds

The radius of gyration used in this work is the geometric radius rather than its mass-weighted analogue, defined as

$$R_G = \sqrt{\frac{1}{N_{\text{bb}}} \sum_{i=1}^{N_{\text{bb}}} \left(\mathbf{r}_i - \frac{1}{N_{\text{bb}}} \sum_{i=1}^{N_{\text{bb}}} \mathbf{r}_i \right)^2} \tag{8.11}$$

where N_{bb} is the number of backbone heavy atoms located at positions $\mathbf{r}_i$, so that R_G represents the RMSD of the backbone coordinates with respect to the average center of the backbone chain. It therefore differentiates between linear or more globular structures. For gramicidin, all heavy atoms are rather considered in eq 8.11 to establish a finer resolution of the side chains packing around the cyclic backbone.

The number of hydrogen bonds is evaluated as

$$N_H = \sum_{i \in \text{O}} \sum_{j \in \text{H}} \frac{1 - [\frac{\mathbf{r}_i - \mathbf{r}_j}{d_0}]^6}{1 - [\frac{\mathbf{r}_i - \mathbf{r}_j}{d_0}]^{12}} \tag{8.12}$$

with $d_0 = 1.8$ Å and i,j running over all oxygen and hydrogen atoms of the peptide, excluding their covalent bonds. This second quantity informs about the secondary structure and distinguishes between molten globular geometries or properly folded peptides. The R_G and N_H geometric descriptors are for example used as collective variables in the context of metadynamics.[614]

Code and data availability

A snapshot version of the EVOLVE code[583] as used in this work is provided on Zenodo at https://doi.org/10.5281/zenodo.7251981, along with the data and analysis scripts needed to reproduce the results.

Appendix

Appendix D contains supplementary information on the effect of the SBX versus genewise crossover; examples of surrogate versus reference structures; relative energies and RMSD plots for the different surrogate fitness functions; minimum energy progression against GA iterations; number of local minima found along and for different GA instances; and energy levels recovered by the sGADFT method when imposing a minimum RMSD between the resulting structures.

9 Enhanced screening of low-energy peptide structures with genetic algorithms and clustering-based novelty search

9.1 Introduction

In the previous chapter, we illustrated how the combination of genetic algorithms (GAs) and molecular mechanics can expedite the navigation of the configurational space of peptides, thus enabling the swift identification of low-energy candidates that are likely to also inhabit the lower regions of more accurate potential energy surfaces (PES). In this chapter, we expand upon this study from an algorithmic perspective, demonstrating how integrating an unsupervised machine learning (ML) technique with GA optimization can promote novelty search, leading to the identification of a significantly larger number of low-energy minima. Our preliminary results, focusing on the trans isomer of the Gly-Pro-Gly-Gly tetrapeptide,[107] are derived from extensive sets of GA executions. The findings suggest that GAs enhanced with clustering techniques not only cover the solution space more comprehensively and discover additional local minima (LM), but also increase the chances of reaching the global minimum (GM) of the PES under investigation. This development could significantly enhance the computational efficiency when searching for larger peptide structures or when GAs are applied with more computationally intensive, yet potentially more accurate surrogate PES.[568] Further advancements could be realized by adaptively tuning GA parameters on-the-fly.

9.2 Clustering-enhanced genetic algorithms

Although GAs are generally highly efficient heuristics for optimizing complex (multi-)objective functions, the tuning of their parameters, and thus their performance, is heavily dependent on the specific problem being addressed and their algorithmic design.[103,587] In this regard, novelty search is a unique approach that focuses on promoting diversity and exploration rather than solely aiming for optimization.[104,617] Instead of only seeking optimal solutions for a given problem, novelty search rewards individuals that exhibit distinct characteristics or behaviors not previously encountered. This helps to counteract the issue of premature convergence, where a GA becomes trapped in LM, unable to find the GM because it overly exploits current areas of the PES that are deemed most promising.[587,618] By encouraging exploration and divergence in the search space, novelty search allows the algorithm to probe various regions that might otherwise be overlooked in traditional optimization-focused approaches.[564–566,570] This can lead to the discovery of innovative and potentially more effective solutions, improving the overall performance of the algorithm. As a result, novelty search can be particularly beneficial in complex, high-dimensional, and rugged problem landscapes, where maintaining diversity and exploration capabilities is crucial for success.

Various strategies exist to introduce novelty search in GAs. The first is the dynamic adjustment of crossover, mutation, and elitism rates during the evolutionary process that can help maintain diversity and promote exploration. For instance, increasing the mutation strength (σ, cf. Table 8.1) when the population converges or when the algorithm is stuck in a LM can spur novelty.[104,619] Second, niching is a technique that detects attractive basins in the solution space and separates the search within these niches. By maintaining multiple coexisting subpopulations, the algorithm is more likely to explore diverse regions and discover novel solutions. Third, a metric is defined to quantify the difference between individuals. This metric can be either based on individuals' characteristics or on the distribution of fitness functions among the population. With this, individuals in the search space that are similar can be identified, leading to penalization of more populated regions and continued exploration in underrepresented areas. The comparison of individuals can be performed not just within the current population but also against a periodically updated archive that preserves the most unique individuals encountered throughout the evolutionary process. In this context, ML descriptors can be useful in defining a metric for similarity and pinpointing new structural patterns within the search space during GA optimization.[104,617] This on-the-fly application differs appreciably from the more conventional use of ML as a post-processing tool to differentiate the resulting structures.[557,615,620]

Expanding upon the latter strategy, Jørgensen et al. introduced a clustering-enhanced evolutionary algorithm.[621] Their approach intelligently selects parent structures by scrutinizing the multitude of intermediate LM structures generated as the search

progresses. These structures, in turn, become essential hints for the exploration of the PES: with the aid of clustering, outlier structures are identified and reintroduced into the population to enhance diversity. When tested on the 2D optimization of a 17-atom organic molecule, the integration of clustering proved to significantly reduce the number of iterations needed to locate the GM on a density functional tight binding PES.

Given our interest in not only the GM but also the extensive set of low-lying minima, we developed the modified GA illustrated in Figure 9.1, drawing inspiration from but also differing from Jørgensen's approach. First, our algorithm is designed for three-dimensional optimization of peptide structures, typically involving more variables than the 2D system examined by Jørgensen et al., and employs a dihedral-based internal encoding. Inclusion of mutations, which were omitted in Jørgensen's work, allows for exploration of small local deformations around low-lying minima. The clustering feature space we use is also different, being more suited to peptides and being smaller, lighter, and quicker to compute. We also suggest a distinct mechanism for reintroducing outliers into the population and eventually consider elitism. Ultimately, we will show that these modifications allow our clustering-enhanced GAs to outperform Jørgensen's version for the specific task of optimizing peptide structures.

More specifically, our ML-enhanced GAs build upon the GA developed in Chapter 8, incorporating the unsupervised learning algorithm before the selection and crossover processes. At each iteration (generation), every low-lying minimum relaxed on the surrogate PES is projected onto a two-dimensional feature space defined by the radius

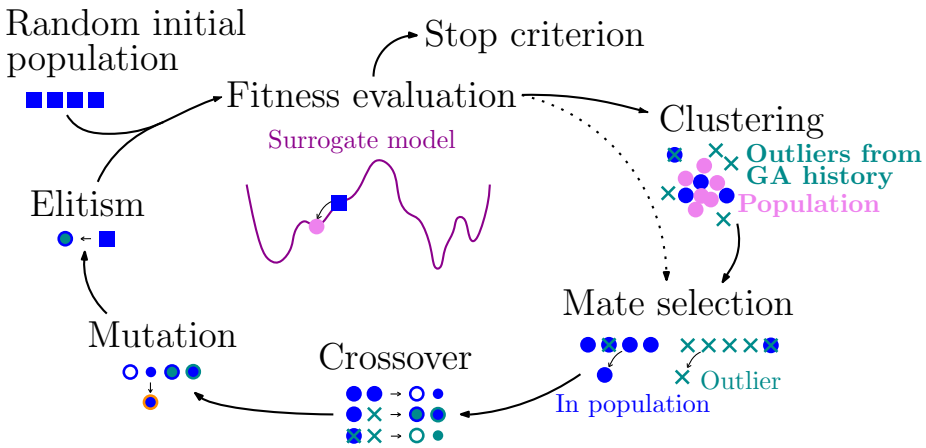
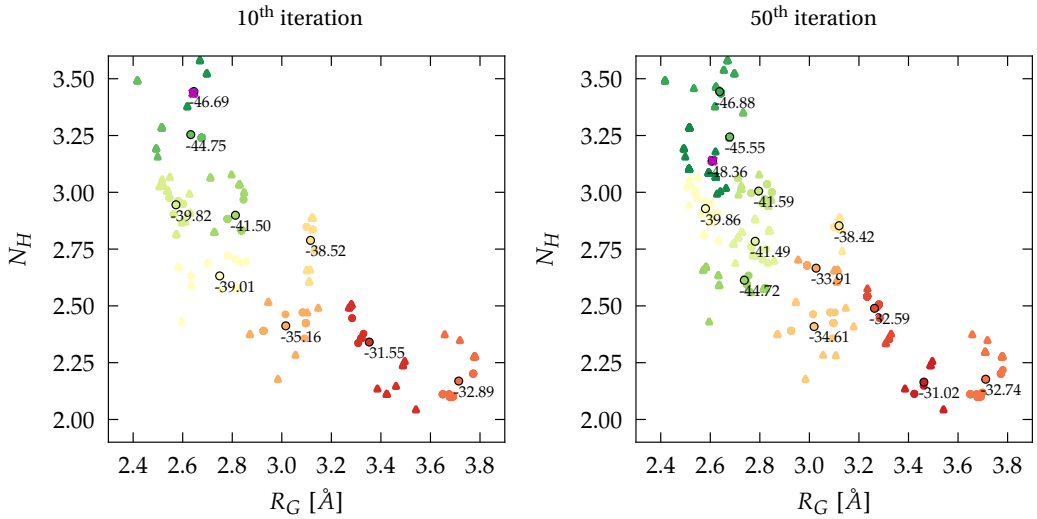


Figure 9.1: A schematic depiction of the clustering-enhanced GA as implemented in EVOLVE. Couples for mating are created by pairing one individual selected from the current population, with one outlier selected from the pool of the clustering algorithm. Other pairing schemes are not considered herein. The deviation in workflow with respect to the previously-developed GA (Chapter 8) is indicated by the dashed arrow. When applicable, elitism is executed after the mutation operator.

of gyration R_G (eq 8.11) and the number of hydrogen bonds N_H (eq 8.12).[614] This space is stored in memory and updated at each generation before being segmented using the agglomerative hierarchical clustering (AHC) technique (Section 9.2.1). For example, Figure 9.2 displays the clusters obtained at various iterations. By doing

(a) Mutations, elitism



(b) Outliers in population (pop outliers), mutations

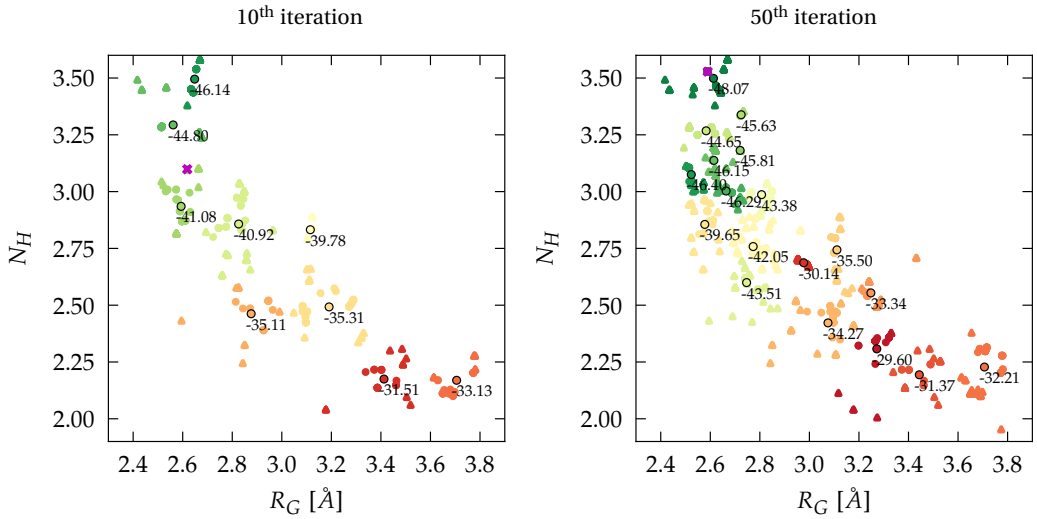


Figure 9.2: Accumulated feature space defined by the radius of gyration R_G and number of hydrogen bonds N_H, as clustered with the agglomerative hierarchical clustering (AHC) method at the 10th (left) and 50th (right) iterations. (a) During a reference GA with mutations and elitism. (b) During our clustering-enhanced GA with mutations (no elitism). Centroids of different clusters are indicated by circles accompanied by cluster average energies. Outliers are shown as triangles. The putative GM during the progression of the GA is indicated by a purple cross. Clustering schemes visibly improve space coverage.

this, the pairing of individuals can leverage the entire knowledge of the GA up to the current iteration. The average of points in each cluster defines the centroid, and the cluster width is calculated as the average distance of all points to the centroid. Based on this information, individuals are classified as outliers if their distance to the centroid exceeds the cluster width. This provides several possible options for reintroducing outliers into the algorithm. In this work, we propose two effective schemes that consistently pair one selected individual from the current population, using tournament selection (Section 8.3.3), with one outlier:

1. The outlier is selected via tournament selection within the entire pool of outliers - Scheme *all outliers*

2. The outlier is selected via tournament selection within the pool of outliers that are also members of the current population - Scheme *pop outliers*

Consequently, both schemes still incorporate a certain level of selective pressure. The first scheme, which encompasses the entire history of the GA, adopts a more relaxed and exploratory approach. In contrast, the second scheme relies on the current population but placed in the broader context of the accumulated knowledge.

In what follows, we investigate the performance of these clustering-enhanced schemes on the search for Gly-Pro-Gly-Gly (GPGG) tetrapeptides in the trans isomer form of the proline residue. The GAFF force field as implemented in Amber acts as the surrogate PES, with fixed-point charges obtained by RESP fitting at the HF/6-31(d) level of theory.[590] The choice of the PES is of minor importance in this chapter since the main focus is on the performance of the search algorithm rather than the comparison with experimental results. Furthermore, it is anticipated that the GA is transferable to other energy models without significant modifications, provided that they own a similar ruggedness as the GAFF PES.[587] The computational details are those described for GPGG in our previous study, which can be found in Section 8.4. The subsequent section provides a brief overview of the AHC algorithm before proceeding to the presentation of results in Section 9.3.

9.2.1 Agglomerative hierarchical clustering

Agglomerative hierarchical clustering (AHC) is a bottom-up clustering technique belonging to the class of unsupervised ML algorithms.[91,112,622] It is used to group data points based on their similarity or distance in a defined feature space. The algorithm starts by considering each data point as a separate cluster and iteratively merges the closest pairs of clusters until all data points belong to a single cluster or a specified stopping criterion is reached. This process of clustering generates a dendrogram like the one reported in Figure 9.3, which is a tree-like structure that represents the hierarchy of clusters and the order in which they are merged.

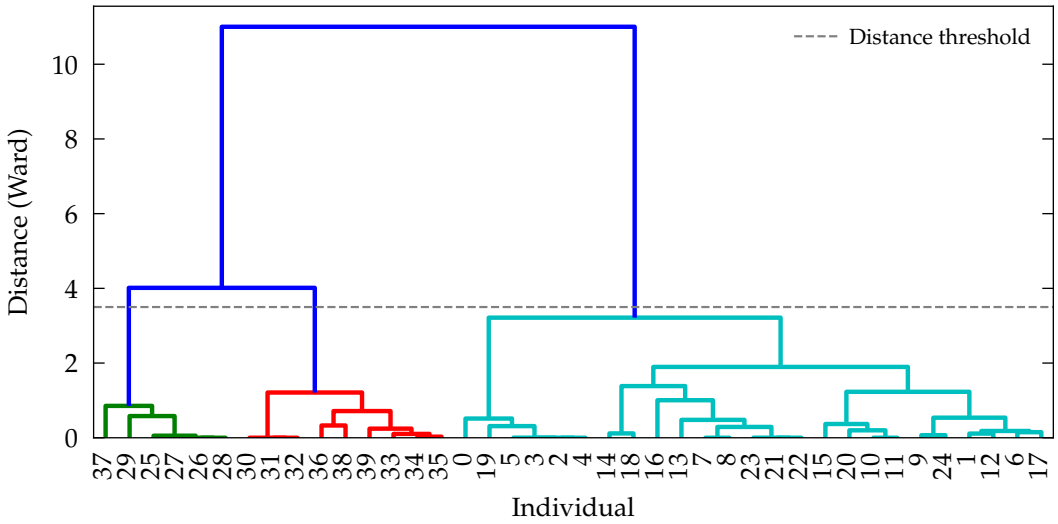


Figure 9.3: Dendogram of the AHC method applied to 40 individuals in the R_G-N_H feature space with Ward's linkage. At the specified distance threshold of 3.5, three clusters are formed.

To build the dendogram, the key aspect of AHC is the choice of 1) a metric, such as the (squared) Euclidean norm, to measure the distance between individual points and 2) a linkage criterion to specify how sets are dissimilar based on the pairwise distances between points within the sets. Several linkage methods exist, such as single-linkage, complete-linkage, average-linkage, and Ward's linkage.[623] In single linkage, the distance between two clusters is defined as the minimum distance between any pair of data points, with one point belonging to each cluster. This means that the proximity between clusters is determined by the closest pair of points, regardless of the overall structure or distribution of the data. In contrast, complete linkage evaluates the distance between two clusters according to the maximum distance between any pair of data points. This means that the proximity between clusters is determined by the pair of points that are farthest apart. As a compromise, in average linkage, the distance between two clusters is calculated as the average distance between all possible pairs of data points, thus determining the proximity between clusters through the average dissimilarity of their members.

Both, the selection of the metric and linkage criterion influence the outcome of clustering. While the metric determines the measure of similarity between objects, the linkage shapes the formation and structure of the resulting clusters. In this work, we use Ward's linkage, also known as the minimum variance method, that aims to minimize the total within-cluster variance, or equivalently maximize the between-cluster variance.[624] Hence, Ward's linkage is particularly effective in generating compact and balanced clusters with similar variances with uniform spread over the feature space. Generally, it is important to preprocess the data by normalizing or standardizing the features since

Ward's linkage is sensitive to the distribution of the data points in the feature space. The Ward's linkage criterion can be defined as follows:

$$d(A,B) = \frac{|A| \cdot |B|}{|A| + |B|} ||\bar{x}_A - \bar{x}_B||^2 \tag{9.1}$$

where $d(A,B)$ is the distance between clusters A and B, $|A|$ and $|B|$ are their cardinalities, $\bar{x}_A$ and $\bar{x}_B$ are the centroids of clusters A and B, respectively, and $||\cdot||$ denotes the Euclidean norm. The Ward's linkage criterion measures the increase in the total within-cluster variance that results from merging clusters A and B. In details, the AHC algorithm using Ward's linkage method can be described as follows:

1. Compute the distance matrix $D_{IJ} = d(I,J)$ using the Ward's linkage criterion for all pairs of data points I and J in feature space.
2. Represent each data point as a separate cluster.
3. Find the pairs of clusters with the smallest distance $d(I,J)$ and merge them.
4. Update the distance matrix by recalculating the distances between the newly formed clusters using the Ward's linkage criterion.
5. Repeat steps 2-4 until all data points belong to a single cluster or a specified stopping criterion is reached.

An advantage of AHC compared to other methods like e.g., K-means, is that the number of clusters does not need to be predefined by the user.[91] This allows for dynamic adaptation of the cluster count during the iterations of the GA. The iterative merging of clusters can be halted once a distance threshold is reached, effectively determining the final number of clusters. In our study, we observed that a consistent distance threshold value of 3.5 yields a reasonable coverage of the feature space by clusters. Importantly, as the number of data points increases over GA iterations, this approach results in a denser clustering and thus a finer resolution of outliers as demonstrated in Figure 9.2.

It should be noted that clustering methods, such as AHC, are susceptible to the curse of dimensionality. As the number of features increases, these methods become less effective in separating data into meaningful information. In their related study, Jørgensen et al. utilized the high-dimensional feature space created by the molecular Bag of Bonds ML descriptor.[621,625] In our investigation, we explored a similar approach using the FCHL representation,[362,626] but did not find any superior features beyond the radius of gyration and the number of hydrogen bonds. Dimensionality reduction techniques like Principal Component Analysis (PCA),[627] applied to the FCHL descriptors before clustering, did not yield further improvements. The explained variance ratios of the first two (three) components accounted for less than 40% (50%) of the variability among the data, indicating limited effectiveness. Hence, we believe that R_G and N_H are suitable features for effectively distinguishing acyclic peptide structures

due to their lightweight nature, quick evaluation, and physical relevance. However, ML descriptors or alternative geometrical quantities may prove more helpful for the classification of more globular conformers.

9.3 Results and discussion

In our pursuit of developing means to efficiently screen the low-energy regions of the PES, the performance evaluation of a GA is based on two criteria. Firstly, we assess the number of low-lying LM retrieved by the algorithm. This assessment is conducted either within a single execution or by considering multiple runs of the GA, as these runs are stochastic and may explore different regions in each execution. Ideally, we aim for the GA to generate a maximum number of LM in a minimal number of iterations or across a few GA runs. Secondly, the GA's effectiveness is determined by its ability to consistently reach the GM. The GM serves as the reference energy baseline for ranking other structures. However, in practice, the GM remains unknown and is assumed based on the lowest-energy structure found by the GA. These two evaluation criteria allow us to gauge the efficiency and reliability of the GA in exploring the low-energy regions of the PES.

To ensure a fair assessment, we sought to determine the best variant of conventional GA as a reference for comparison. A comprehensive grid search was conducted, exploring various combinations of GA parameters (Table 8.1) such as the crossover SBX operator with weak, normal, and strong effects ($n \in [1, 5, 7]$), Gaussian mutations with different standard deviations ($\sigma \in [10°, 60°, 180°]$), and elitism with varying fractions ($f \in [0.00, 0.05, 0.10, 0.40]$). After evaluating the performance in terms of GM retrieval and coverage of the low-energy regions, the algorithm employed in Chapter 8 ($n = 5, \sigma = 60°, f = 0.10$) emerged as the fastest to converge towards the GM and among the top performers overall. Only the addition of stronger mutations resulted in a marginally higher (maximum three) number of LM found within 5 kcal/mol from the GM. Hence, for subsequent benchmarking, the reference algorithm to be surpassed is the one employed in Chapter 8, referred to hereafter as the *mutations, elitism* algorithm.

9.3.1 Number of local minima

The coverage of LM by a single execution of the GA search is illustrated in Figure 9.4, where the results are averaged over 100 independent runs. A comparison reveals that random search exhibits greater efficiency in covering a large number of high-lying LM (within 15 kcal/mol from the GM) as opposed to targeting low-energy structures (within 10 and 5 kcal/mol). Hence, as a validation of our research, GAs emerge as more effective tools for sampling the low-energy regions of the PES than random search.

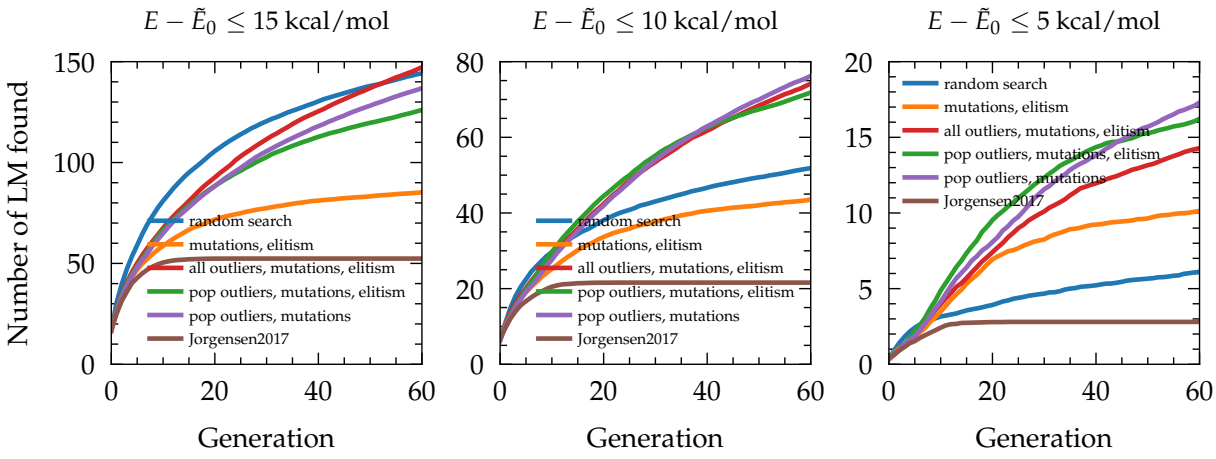


Figure 9.4: Number of LM found as a function of GA generation compared to random search and Jørgensen's clustering scheme.[621] Results are shown for LM located within respectively 15, 10, and 5 kcal/mol from the putative GM, and were averaged over 100 independent runs. Distinct LM are separated by at least 0.0001 kcal/mol.

In all 15, 10, and 5 kcal/mol energy ranges, our clustering-enhanced GAs outperform the reference *mutations, elitism* scheme by yielding higher numbers of LM in the low-energy regions (10, 5 kcal/mol) without sacrifying efficient exploration for higher energies (15 kcal/mol). It is worth noting that the combination of *all outliers, mutations* is not reported, as it only outperforms the *mutations, elitism* GA (and Jørgensen's method) but not the other clustering-based schemes. This observation highlights the delicate balance between exploration, which probes unknown regions of the landscape, and exploitation, which focuses on regions known to contain low-energy structures. In the *all outliers, mutations, elitism* scheme, exploration is encouraged by considering all outliers in the clustered pool, which is efficiently counterbalanced by the exploitation process of elitism. However, the absence of elitism in the *all outliers, mutations* scheme causes the algorithm to spend more time exploring higher-energy landscapes for less sampling efficiency in low-lying regions. In the *pop outliers* schemes, the selection of outliers from the current population promotes exploitation. Consequently, when combined with elitism, the *pop outliers, mutations, elitism* algorithm performs slightly worse than the *pop outliers, mutations* GA. Thus, the *pop outliers, mutations* algorithm stands out as the most effective approach for generating a significant number of low-energy structures within a single execution.

In contrast, Jørgensen's sampling approach encounters limitations after several iterations due to multiple factors. Firstly, in Jørgensen's method, the population exclusively comprises the best individuals found thus far, which inadvertently reduces population diversity and intensifies exploitation. Secondly, mutations are not incorporated, leading to the fact that exploration has to rely solely on the effectiveness of the applied

clustering scheme, i.e. in Jorgensen's approach: "if an outlier is present in the population, it will always be chosen as the first parent. If more than one outlier is present in the population, one of them is chosen randomly to be the first parent. The second parent is chosen randomly from all other structures in the population." Hence, the clustering-based exploration is further compromised by exclusively utilizing outliers from the current population. In our investigation of sampling trans-GPGG peptides, we observed that Jørgensen's scheme exhibits premature convergence due to the lack of sufficient exploration. Without the inclusion of mutations, the number of outliers within the population quickly diminishes, resulting in the algorithm primarily performing crossovers of randomly selected individuals that become closely-related. These crossovers are not targeted or efficient enough to generate diverse offspring from the parents. Such factors contribute to the stagnation observed in Jørgensen's sampling approach.

It is also interesting to observe how LM are sampled when multiple GA search instances are executed. Figure 9.5 illustrates the cumulative number of LM found by our clustering-enhanced GAs, random sampling, and Jørgensen's algorithm. In the case of a large number of runs, random sampling proves to be efficient in retrieving peptide structures, particularly within the wider energy range of 15 kcal/mol from the putative GM. However, as the focus shifts to lower energies and fewer searches are performed, random sampling becomes less efficient compared to our clustering-enhanced GAs. Furthermore, the GAs combined with clustering demonstrate superior performance compared to the reference *mutations, elitism* algorithm in this context. Although the *pop outliers, mutations, elitism* scheme yields a few more minima within 5 kcal/mol

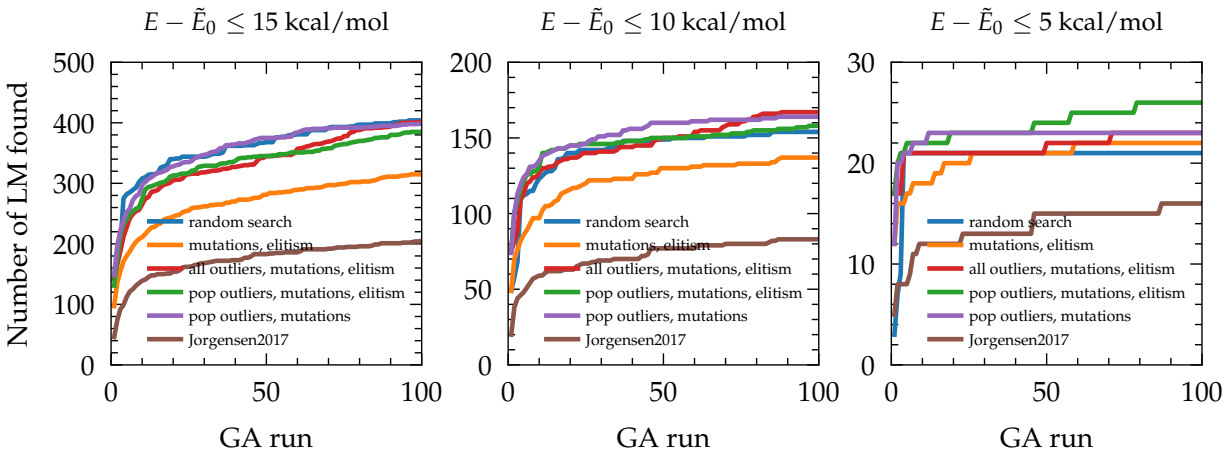


Figure 9.5: Cumulative number of LM found after several GA independent runs, compared to random search and Jørgensen's clustering scheme.[621] Results are shown for LM located within respectively 15, 10, and 5 kcal/mol from the putative GM. Distinct LM are separated by at least 0.0001 kcal/mol.

when elitism is included (after a considerable number of runs), it is the *pop outliers, mutations* scheme that exhibits greater consistency across energy ranges, as previously observed in the single run analysis.

When considering approximately ten algorithms running in parallel, it may appear that clustering-based GAs and random search offer similar performance in terms of sampling low-energy minima. However, important distinctions emerge when the objective shifts towards targeting the GM, as demonstrated in the next section.

9.3.2 Global minimum detection

The ability to identify the GM or structures that are energetically very close is crucial to ensure that the sampled LM belong to the low-energy regime. An ideal GA should reach the GM in most of its runs within a minimal number of generations. This convergence speed can be assessed by the cumulative success in finding the GM across a given number of GA generations, as depicted in Figure 9.6. In this regard, Jørgensen's scheme performs poorly due to the aforementioned reasons. Moreover, random search, while exploring the PES with qualitative efficiency when multiple (too many) trials are conducted, lacks exploitation mechanisms so that the GM and the low-lying space are generally overlooked.

The *mutations, elitism* GA presented in Chapter 8 demonstrates a higher probability of approximately 45% in finding the GM at the end of a run. Despite this, the introduction

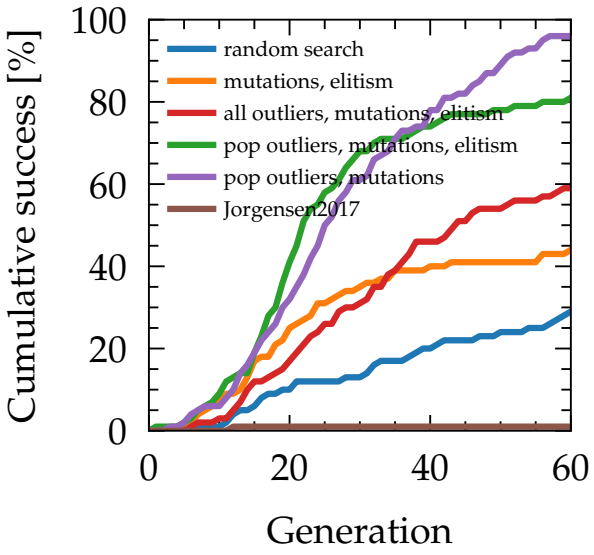


Figure 9.6: Cumulative success, i.e. probability of finding the GM at each generation of the GAs, compared to random search and Jørgensen's clustering scheme.[621] Results were averaged over 100 independent runs.

of clustering-enhanced GAs yields significant improvements in the search for the GM, particularly with the *pop outliers* schemes which exhibit superior performance. These schemes employ more selection pressure, emphasizing exploitation for faster convergence, in comparison to the *all outliers* option. Elitism generally accelerates convergence but limits further exploration as the GA progresses. As a compromise, mixing the exploratory nature of mutations with the exploitation-driven nature of the *pop outliers* scheme results in a highly efficient algorithm for global optimization. Thus, the *pop outliers, mutations* GA successfully and efficiently explores numerous low-energy minima while effectively retrieving the GM and enhancing the screening process of low-energy peptide structures.

9.4 Conclusions and outlook

This chapter has demonstrated the substantial performance enhancements achieved by integrating unsupervised ML techniques into GAs. By leveraging a lightweight feature space defined by the radius of gyration and number of hydrogen bonds, the ML method facilitates the clustering of the complete history of the GA and promotes novelty search in highly dissimilar regions. As a result, not only is the exploration of low-lying regions improved, but there is also a significant increase in the likelihood of identifying the most probable GM of the PES. Importantly, these enhancements are accomplished without compromising execution times, with only a marginal overall increase of a few minutes, making this integrated approach a powerful and efficient tool for PES exploration.

Results were obtained for the optimization of the trans isomer of the GPGG 39-atom peptide, therefore encouraging the use of clustering-enhanced GAs for the sampling of either larger linear systems or more computationally expensive energy models. Also, further improvements could be achieved from the incorporation of adaptive parameters (e.g., mutation strength, elitism fraction, fraction of outliers) that vary during GA execution according to simple considerations such as in the Rechenberg's scheme.[617,619] For instance, as the GA progresses and discovers better structures, the number of outliers could be reduced to focus the search on the most promising low-lying structures, akin to the concept of simulated annealing. Similarly, mutations could be adjusted to be weaker when good individuals are found, enabling a more focused exploitation of interesting regions of the PES without escaping to higher energies too rapidly. Looking ahead, such hybridizations of GAs with clustering techniques hold promise for addressing other complex optimization problems, including those involving multi-objective functions.

Part V
Conclusions and outlook

Summary and perspectives

Computational chemistry endeavors to accurately simulate chemical reactions and molecular properties at the atomic scale using computational tools. Its overarching goal is to contribute to the design of new chemical compounds and materials. However, current numerical methods encounter challenges related to the trade-off between computational cost and accuracy, as well as the impractical execution times required for efficient sampling of the potential energy surface of larger systems. This thesis has thus centered around exploring unconventional approaches such as stochastic sampling and artificial intelligence to overcome these challenges in specific problems.

The main outcomes of this thesis and related perspectives are as follows:

–1– The implementation and validation of the MP2 energy in the CPMD plane wave code, as well as technical details to ensure accurate convergence within the plane wave basis set. — Chapter 4

MP2 is the most expedient wavefunction-based method for considering electron correlation in quantum chemical calculations and, when integrated within DFT, gives rise to the most accurate double-hybrid density functionals. The plane wave basis set, in turn, allows to systematically converge reference energies to the complete basis set limit, devoid of basis set superposition errors, and enables the MP2 (eventually double-hybrid) calculation of periodic systems at the basis set limit. However, only a limited number of plane wave codes have focused on providing access to the MP2 energy due to certain inherent complexities.[64,65] The first resides in the enormous number of basis functions that is required to converge properties accurately (typically of the order of 10^5). Also, the MP2 correlation energy is obtained by summing an astronomically large number of small contributions which substantially hampers convergence. The prohibitive quintic scaling of the method with respect to system size also makes the approach very expensive for all but the smallest systems. Our implementation paves the way for access to double-hybrid accuracy in CPMD, which may one day be used to propagate molecular dynamics (MD) simulations, possibly aided by multiple time step integrators and/or machine learning models. Currently, plain MP2 calculations with CPMD require large memory and remain expensive for

systems with more than ∼100 electrons and are thus still prohibitive for performing straightforward longer time dynamics for larger systems. Also, forces have not been implemented yet. However, we have shown that single-point calculations with plane waves are feasible, taking up to few days, and yield MP2 energies free of basis set superposition errors at difference to atom-centered basis sets. A further advantage of plane waves is completeness and a systematic way to reach the basis set limit, which allows to provide accurate reference energies that help identifying possible biases of other basis set types. The convergence analysis of double-hybrid functionals when used with plane waves against atom-centered bases[125] would be an interesting project for the future.

–2– The caveat that Gaussian-type correlation-consistent (aug-)cc-pVXZ basis sets, due to their intrinsic incompleteness, may bias correlation energies, especially in the limit of increasing system size. — Chapter 4

The comparison of non-covalent interaction energies between plane-wave and correlation-consistent atom-centered basis sets has highlighted the accuracy and importance of the counterpoise correction for the latter. In addition, the recovering of the full MP2 correlation energy by the (aug-)cc-pVXZ bases was found to depend on the number of electrons, therefore questioning their ability to accurately converge correlation energies in the limit of (very) large systems. For the system sizes routinely investigated nowadays with MP2, such deviations are marginal compared to the errors of the method itself. However, thinking further, once very accurate methods will become feasible for large systems,[27] basis set deviations will dominate and potentially bias results by more than chemical accuracy. Time and the investigation of larger and larger systems with different types of bases will confirm or refute this statement.

–3– The establishment of suggestions to accurately extrapolate correlated energies to the basis set limit when resorting to the (aug-)cc-pVXZ basis sets. — Chapter 4

Converging energies to the basis set limit is essential to assess the intrinsic performance of a quantum chemical method. Despite the minor discrepancies mentioned above between plane waves and (aug-)cc-pVXZ atom-centered bases, minimizing their deviations in the complete basis set limit allowed to find the best extrapolation scenarios when resorting to the latter. Those are reported in Section 4.6 and include the newly proposed *Rovibi34* scheme established by our investigations. Hopefully this will serve the community in general applications and be further validated by comparing (aug-)cc-pVXZ bases to other accurate references, obtained either via correlated wavefunction-based methods or experimental values.

–4– A Monte Carlo MP2 method capable of speedups of up to three orders of magnitude and reduced scaling with plane waves, while keeping stochastic errors at marginal levels. — Chapter 5

Plane-wave MP2 calculations are fundamentally hampered by the gigantic basis set required to converge energies. This, in turn, affects the number of virtual orbitals that contribute to the MP2 correlation energy. To cope with this number, a stochastic treatment of the integral contributions from the virtual space was introduced. Above a certain eigenvalue threshold, virtual orbitals are considered as part of a continuum-like space, so that the distribution of their contributions is smooth and well-behaved. This allows a stochastic sampling of the MP2 integrand based on Monte Carlo summation over continuum orbitals. Within the most expensive continuum regime, the algorithmic scaling reduces from quintic to quadratic, and bypasses the calculation and summation of millions to billions of terms. Therefore, the method gives access to the calculation of unbiased correlation energies in the basis set limit for systems containing hundreds of electrons with unprecedented efficiency gains. The stochastic errors and execution times are mitigated by two parameters that are the continuum eigenvalue cutoff and the number of samples per virtual orbital. Although a set of parameters was found to be transferable between the systems studied here, with conservative performance (errors below 0.1 kcal/mol), further work should elaborate on getting reliable estimates of the stochastic deviations. Indeed, noise reduction techniques or methods to estimate the statistical error bars[26,76,78] would ensure that an accuracy close to the exact MP2 value is preserved. With this, new bottlenecks are likely to appear; these are large memory requirements and the diagonalization of the large virtual space prior to the MP2 Monte Carlo treatment.[231] This Monte Carlo approach is simple and can easily be extended to other ab initio methods that involve numerous virtual states, e.g., the RPA as shown in Chapter 5. It should be noted, however, that the speed gains would be less for the RPA since it originally has a quartic scaling. A somewhat similar stochastic sampling has been used recently to speed up the evaluation of the correlation part of the self-energy in GW calculations of two-dimensional materials.[79] Therefore, Monte Carlo integration offers a promising acceleration in the calculation of correlation energies, as long as statistical errors are well controlled.

–5– The benchmark of the Minnesota density functionals for the structural and dynamical description of liquid water under ambient conditions thanks to a machine learning enhanced multiple-time-step scheme. — Chapter 7

Car-Parrinello MD simulations with the Minnesota meta-GGA functionals revealed that, contrary to the prevalent idea that local and semilocal functionals overstructure and slow down dynamical properties of liquid water, M06-L, revM06-L, and M11-L understructure, while MN12-L and MN15-L lead to too large intermolecular distances between water molecules due to too weak cohesive effects. This has been attributed to a weakening due to disruption of the hydrogen bond network, which leads to excessively fast dynamical fingerprints. Hybrid functionals are about two orders of magnitude more expensive than meta-GGAs due to the inclusion of exact exchange.

This has so far challenged their use in the context of ab initio MD. The extensive benchmark of hybrid Minnesota functionals performed in this thesis was made possible thanks to the multiple time step (MTS) scheme implemented recently in the CPMD software.[45] With the help of a fast machine learning (ML)-based low level, that infers forces and drives the dynamics at shorter time steps, the ML-MTS propagation allowed for speedups of about 6 to 15 as compared to standard Born-Oppenheimer MD without affecting the accuracy. While most of the hybrid Minnesota functionals remain understructured (M06, M08-HX, M08-SO, M11, MN12-SX, and MN15), their dynamical properties generally improve over their semilocal counterparts. Water is the most abundant substance on Earth, yet its liquid properties are distinct from those of other fluids, posing a challenge for in silico simulations not only of condensed water but also of aqueous chemistry. Chapter 7 not only provides benchmarks for the widely-used Minnesota density functionals but also places them in the context of other DFT approximations as well as experimental measurements. This will hopefully serve as a shared foundation for future assessments of DFT on water, and assist the scientific community in utilizing, refining, or developing more accurate and transferable exchange-correlation functionals.

–6– The finding that an over-inclusion of exact exchange in DFT functionals shortens and strengthens hydrogen bonds, leading to water properties that are too glassy.
— Chapter 7

Correlation effects, and specifically the consideration of dispersion, are known to stabilize water molecules between coordination shells and consequently improve the description of liquid water. We found that the inclusion of exact exchange is another key ingredient for the correct description of hydrogen bonds, that ultimately corrects both dynamical and structural properties of water. However, an excessive amount of exact exchange like in M06-HF shortens and strengthens the hydrogen bonds between water molecules, thus yielding properties that are too glassy. This highlights the primordial balance between exchange and correlation effects, not only for water but also for more sophisticated phenomena that involve interacting with water molecules or in which hydrogen bonds are important, as for example in the reactions of biological systems in physiological environments. This delicate and subtle balance can only be achieved by using the higher rungs of Jacob's ladder, i.e. hybrids and double-hybrids/RPA.

–7– M06-2X(-D3) are not only the most accurate Minnesota functionals for liquid water but also emerge as top contenders for reproducing experimental results when compared to other functionals that have previously been tested.
— Chapter 7

M06-2X emerges as the best Minnesota functional for liquid water. Slightly understrucured and fast, its D3 dispersion corrected version shows even better agreement

for structural properties. Based on previous studies taking nuclear quantum effects (NQEs) into account, the revPBE0-D3 hybrid, and the RPA (RPA@PBE) have been the only two approximations that agree with experiments. The examination of the impact of NQEs on the results shows that M06-2X(-D3) would perform equally well or better if NQEs were explicitly included, thus competing with revPBE0-D3 and RPA@PBE. In this regard, determining whether M06-2X(-D3) are indeed one of the best functionals in conjunction with NQEs would obviate the need for the significantly more expensive fifth rung of Jacob's ladder (RPA@PBE).

–8– The multiple-time-step propagation has the advantage of training a low-level machine learning model only once, which is data-efficient due to transferability between simulations of different high levels. — Chapter 7

The conventional role of ML in MD typically involves training models on ab initio data to bypass the computationally demanding nature of quantum methods. However, this approach necessitates meticulous training and lacks control over model errors in the extrapolative regime. In contrast, the MTS scheme relaxes the accuracy constraints on the ML model to some extent. The low-level model is only tasked with reproducing the high-level model with sufficient fidelity, allowing for high ratios (speedups) between inner and outer time steps. The advantage of MTS lies in the choice of the low-level model, which is not required to precisely reflect the accuracy of the high-level model as long as their differences vary slowly over time. By construction, the accuracy of the resulting MTS dynamics is automatically the one of the high-level model, independently of the choice of the low level which has only an impact on the computational efficiency. The study on Minnesota functionals showcased that training a single ML model on PBE0 data can expedite simulations involving different levels of theory. This ML-MTS approach enables the rapid assessment of computationally intense high-level methods for a given system without the need to retrain the ML model for each level. As a result, the method is highly data-efficient and transferable. Future applications will therefore accelerate the testing of different quantum chemical methods when combined with MD.

–9– The implementation of a genetic algorithm for the optimization of peptide structures in the EVOLVE code, targeting the exploration of low-energy regions of the potential energy surface. — Chapter 8

Even for relatively small molecules, the exhaustive exploration of the potential energy surface (PES) is severely hampered by the dimensionality bottleneck. The challenging task of efficiently sampling realistic peptide geometries was addressed by resorting to a surrogate based genetic algorithm (GA)/DFT approach in which promising candidates provided by the GA were ultimately optimized with hybrid DFT. Tests and tuning of the algorithm led to good performance in retrieving not only the global minimum

but also many low-lying minima of the surrogate PES. Subsequent developments could extend the structural optimization to mutli-objective problems involving the structure and additional criteria such as thermostability, composition or amino-acid specific constraints. Looking forward, the systematic generation of low-energy minima could prove beneficial in practical applications such as providing starting points for MD simulations, free-energy sampling, transition state searches, and nudged elastic band methods, or serving as templates for protein-ligand complexes in drug discovery. Additionally, the manifold of minima provided by the GA can serve as valuable training data for various ML approaches.

–10– The efficient generation of low-lying peptide structures as observed in ultracold infrared spectroscopy thanks to genetic algorithms and the AMOEBA polarizable force field as surrogate method, ultimately saving weeks of search with conventional approaches.

— Chapter 8

Computational identification of the most stable structures generated in ultracold gas phase experiments poses stringent accuracy demands to provide a correct energetic ordering in the 0-2.5 kcal/mol observation range. This accuracy is usually only attainable by resorting to high-level quantum chemical methods. However, the computational expense of these approaches makes them unsuitable for direct combinations with GAs that involve typically the evaluations of a few tens of thousands of structures. A comparison of several more cost-effective approaches (GAFF, AMOEBApro13, PM6, PM7, DFTB3-D3(BJ)) indicated that the AMOEBApro13 polarizable force field offers the best compromise between cost and accuracy. Moreover, in three test systems, the GA combined with AMOEBA managed to find DFT global minima within a maximum of 2 kcal/mol above the assumed AMOEBA global minimum. As a result, subsequent DFT relaxation of AMOEBA low-energy structures within 2 kcal/mol consistently led to the identification of the most stable structures on the DFT PES. This demonstrates that the AMOEBA GA approach can offer significant benefits in the three-dimensional determination of trapped metastable or global minimum peptide structures, as observed in ultracold spectroscopy. Indeed, the spectra computed for all the resulting DFT global minima accurately correspond with results from high-resolution infrared spectroscopy. Furthermore, the GA was able to generate local minima in just a few hours, compared to the more laborious ab initio simulated annealing used in earlier studies, which could take up to several weeks of computational trials. To further validate this success, it would be interesting to test our GA approach on new ultracold systems with infrared spectra for which the structures have not yet been resolved.

–11– The development of an alternative genetic algorithm coupled with unsupervised learning for an even better coverage of low-lying regions on the PES, without affecting execution times. — Chapter 9

Having demonstrated the benefits of GAs for the sampling of low-energy peptide structures with a surrogate method, it was also found that integrating GAs with a ML clustering algorithm further enhances both the speed and probability of convergence towards the global minimum, while also augmenting the generation of local minima. By projecting at each iteration the GA history onto a feature space, and subsequently segmenting this space into clusters, instantaneous outlier structures can be selected and re-injected into the population. This encourages diversity and exploration in the low-energy regions of the PES with a near doubling of the number of minima generated compared to a traditional algorithm. Furthermore, the clustering algorithm operates with negligible computational overhead. Ongoing work is examining the application of this enhancement to larger systems and in conjunction with adaptive GA parameters.[619] From an algorithmic standpoint, it remains to be tested if such a hybrid GA/ML framework brings improvements for other optimization problems, including those that are multi-objective. A last obvious extension would be to also use ML to predict expensive fitness functions much more rapidly, like energies, vibrational frequencies or any other property of interest.[617]

Advancements in hardware and high-performance computing infrastructures continually drive computational chemistry towards delivering more predictive insights. Simultaneously, computational chemists are making strides in theoretical development, method implementation, and innovative algorithm design. In this context, the increasing use of artificial intelligence in natural sciences and fundamental research presents new opportunities to circumvent the computationally expensive solution of the Schrödinger equation. However, it is important to remember that data serves as the foundation of approaches like ML, and the quality of this data ultimately sets the upper limit on attainable accuracy.

This thesis began by pushing current boundaries in reference accuracy and system size with the development of an MP2 method in plane waves, which served to attain reference MP2 energies at the full basis set limit that in turn were used to assess the errors arising in atom-centered basis sets. At the same time, this development lays the foundations for future MP2 and double-hybrid simulations of condensed phase systems. Subsequently, the performance of Minnesota functionals on water was evaluated through an ML-accelerated MD approach that could identify M06-2X(-D3) as the best performing functionals. Lastly, we achieved comprehensive sampling of the PES with GAs operating on surrogate PESs, further enhanced by ML, and refined with DFT.

My work has traversed the various stages required for the development of increasingly efficient electronic structure and ML models. Moreover, it has demonstrated how artificial intelligence can be strategically combined with quantum chemistry methods

to maintain an ab initio degree of accuracy in results, rather than being solely reliant on ML predictions. With this, I hope to have shown how both stochastic and artificial intelligence approaches can advance computational chemistry towards simultaneously achieving higher accuracy, accommodating larger system sizes, and reducing sampling time.

```
Finishing the last lines, saving my files, closing my terminal,
and turning off my laptop, I think back to that teenager who wanted
to understand how mathematics can describe matter in a predictive way.
Thirteen years later, the more learned, the more humble, but I enjoyed
trying to figure out such a formidable question.
```

A Appendix of chapter 4: Plane-wave vs correlation-consistent MP2

MP2 implementation in CPMD

Algorithm 1 presents the details of the calculation of $E_{c,n}^{\mathrm{MP2}}$ (eq 4.9) that uses the existing mixed distributed/shared (MPI/OpenMP) parallelization strategy[225,226] of CPMD (CP_GROUPS are currently not supported). The work load is divided into blocks. Within a block, a list of summand indices is created. Then, the partial two-electron integrals are saved in an array, and summed across tasks only once the loop has been completed. This allows for the OpenMP parallelization of the outermost loop, leading to a much smaller shared-memory overhead from thread creation. Additionally, inter- and intra-task-communication becomes much cheaper, as one big array is distributed once per block, instead of a single number being distributed for every $ijab$ tuple. This saves a lot of overhead. Then, the summed partial integrals are used to calculate the final MP2c energy. Note that at the Γ-point, the orbital coefficients can be chosen to be real, introducing the symmetry $\tilde{\phi}_{i,a}(\mathbf{G}) = \tilde{\phi}^{*}_{i,a}(-\mathbf{G})$ which can be exploited to speed-up the code. In order to extrapolate the MP2 correlation energy at the basis set limit (eq 4.11), $E_{c,n}^{\mathrm{MP2}}$ values are printed out whenever n is incremented by b_{incr} virtual orbitals until n_{max}. This increment is user-defined; the larger the value, the more efficient the calculation, which however becomes more memory intensive and provides less extrapolation points. In this work, $b_{incr} = 100$ was found as a good compromise. In addition, n_{max} is defined sufficiently large as to reach a reliable extrapolation regime according to eq 4.11 that is valid at high virtual index. Values between $n_{\mathrm{max}} = 10000$-20000 were used herein for relative energies. A RESTART mechanism is available in order to diagonalize supplementary virtual orbitals and continue the computation of $E_{c,n}^{\mathrm{MP2}}$ for larger n_{max}. This implementation corresponds to the one used in our recent study on the acceleration of the MP2c energy calculation by stochastic sampling of the virtual space integrands based on Monte Carlo summation[232] (Chapter 5). In that case, the stochastic sampling is carried out by selecting which $ijab$ tuples will be effectively included in the *list* for later evaluation and proper renormalization.

Input : $n_{\max}$, $E_{cut}^{\rho_{ia}}$, virtual block increment b_{incr}
Output: $E_{c,n}^{\mathrm{MP2}}$ for $n \bmod b_{incr} = 0$

Wavefunction optimization with HF $\rightarrow E^{\mathrm{HF}}$, $\tilde{\phi}_i(\mathbf{G})$, ε_i;
Diagonalization of the $n_{\max}$ lowest virtuals $\rightarrow \tilde{\phi}_a(\mathbf{G})$, ε_a;
/* The $\mathbf{G}$ vectors are distributed among the MPI tasks for all i,a orbitals */
$\phi_{i,a}(\mathbf{r}) \leftarrow \mathrm{FFT}^{-1}_{E_{cut}^{\phi}}[\tilde{\phi}_{i,a}(\mathbf{G})]$;
for $i \leftarrow 1$ **to** N_{occ} **do**
  for $a \leftarrow 1$ **to** n_{max} **do**
    $\rho_{ia}(\mathbf{r}) \leftarrow \phi_i^*(\mathbf{r})\phi_{a+N_{\mathrm{occ}}}(\mathbf{r})$;
    $\rho_{ia}(\mathbf{G}) \leftarrow \mathrm{FFT}_{E_{cut}^{\rho_{ia}}}[\rho_{ia}(\mathbf{r})]$;
  end
end
$E_{c,n}^{\mathrm{MP2}} \leftarrow 0$, $n_{blocks} \leftarrow \mathrm{ceil}(n_{\max}/b_{incr})$;
for $n_{low} \leftarrow 1$ **to** n_{blocks} **by** b_{incr} **do**
  /* Block contribution from adding b_{incr} new virtuals, create *list* of contributing *ijab* tuples and make use of the symmetries of the integrals and MP2c summands */
  $n_{high} \leftarrow \min(n_{low} + b_{incr} - 1, n_{\max})$;
  for $i \leftarrow 1$ **to** N_{occ} **do**
    for $j \leftarrow i$ **to** N_{occ} **do**
      for $a \leftarrow n_{low}$ **to** n_{high} **do**
        for $b \leftarrow a$ **to** n_{high} **do**
          $ijab \leftarrow (i,j,a,b)$;
          $list$.append$(ijab)$;
        end
      end
      for $a \leftarrow 1$ **to** n_{low} **do**
        for $b \leftarrow n_{low}$ **to** n_{high} **do**
          $ijab \leftarrow (i,j,a,b)$;
          $list$.append$(ijab)$;
        end
      end
    end
  end
  $\langle ij|ab\rangle \leftarrow$ Array(size: $list$.size, elements: 0);
  $\langle ij|ba\rangle \leftarrow$ Array(size: $list$.size, elements: 0);
  /OMP parallelized loop/;
  forall $ijab \in list$ **do**
    forall $\mathbf{G}$ *defined by* $E_{cut}^{\rho_{ia}}$ **do**
      /* Within a MPI task */
      $\langle ij|ab\rangle\,[ijab] \leftarrow \langle ij|ab\rangle\,[ijab] + \Phi(\mathbf{G})\rho_{ia}(\mathbf{G})\rho_{jb}(\mathbf{G})$;
      $\langle ij|ba\rangle\,[ijab] \leftarrow \langle ij|ba\rangle\,[ijab] + \Phi(\mathbf{G})\rho_{ib}(\mathbf{G})\rho_{ja}(\mathbf{G})$;
    end
  end
  MPI_SUM $\langle ij|ab\rangle\,[1,\ldots,list.\mathrm{size}]$ across all MPI tasks;
  MPI_SUM $\langle ij|ba\rangle\,[1,\ldots,list.\mathrm{size}]$ across all MPI tasks;
  /OMP parallelized loop, + reduction/;
  forall $ijab \in list$ **do**
    /* Factors for symmetries */
    if $i = j$ and $a = b$ **then**
      $f \leftarrow 1$;
    else
      if $i = j$ **or** $a = b$ **then**
        $f \leftarrow 2$;
      else
        $f \leftarrow 4$;
      end
    end
    $E_{c,n}^{\mathrm{MP2}} \leftarrow E_{c,n}^{\mathrm{MP2}} + \frac{f}{\Omega^2}\frac{(\langle ij|ab\rangle[ijab])^2 - \langle ij|ab\rangle[ijab]\langle ij|ba\rangle[ijab] + (\langle ij|ba\rangle[ijab])^2}{\varepsilon_i+\varepsilon_j-\varepsilon_a-\varepsilon_b}$
  end
  print $E_{c,n}^{\mathrm{MP2}}$ $(n = n_{high})$ **for extrapolation at consecutive** b_{incr}
end

Algorithm 1: Pseudocode for the calculation of the MP2c energy in CPMD.[47]

Table A1: HF and MP2c contributions to the MP2 interaction energy for some S22 systems and wavefunction cutoff energy E^{ϕ}_{cut}. Energies are in [kcal/mol]. $r_{x,y,z}$ are the respective x, y, z ratios of the orthorhombic supercell dimensions with respect to the HF electron density measured at an isosurface of 0.002 a.u., while Ω is the volume of the supercell. σ_c^{MP2} corresponds to the standard deviation of $\Delta E_c^{\mathrm{MP2}}$ values extrapolated on different fitting ranges according to eq 4.38. The density cutoff energy is $E^{\rho}_{cut} = 4E^{\phi}_{cut}$ and its analogue for the MP2c pair densities is $E^{\rho_{ia}}_{cut} = E^{\phi}_{cut}$.

S22 system	$r_x\ r_y\ r_z$	Ω [Å^3]	E^{ϕ}_{cut} [Ry]	ΔE^{HF}	$\Delta E_c^{\mathrm{MP2}}$	ΔE^{MP2}	σ_c^{MP2}
$(NH_3)_2$	2.0 2.0 2.0	987.84	150	-1.428	-1.763	-3.191	0.004
			180	-1.429	-1.762	-3.191	0.002
	2.0 2.9 2.9	1822.02	150	-1.430	-1.752	-3.182	0.006
			180	-1.430	-1.751	-3.181	0.006
	2.0 3.6 3.5	2744.00	150	-1.430	-1.758	-3.188	0.011
	(*cubic*)		180	-1.431	-1.751	-3.182	0.011
$(H_2O)_2$	1.7 1.9 2.2	648.86	150	-3.618	-1.354	-4.972	0.004
			160	-3.622	-1.347	-4.969	0.004
			180	-3.638	-1.345	-4.983	0.004
	2.0 2.0 2.0	686.16	150	-3.599	-1.358	-4.957	0.003
			180	-3.618	-1.344	-4.962	0.001
Formamide	1.4 1.4 1.4	539.82	150	-11.768	-3.658	-15.426	0.004
			180	-11.821	-3.626	-15.447	0.006
	2.0 2.0 2.0	1573.83	150	-12.166	-3.549	-15.715	0.008
			180	-12.221	-3.499	-15.720	0.005
PD benzene	1.6 1.8 1.7	1550.20	150	6.207	-10.751	-4.543	0.012
			180	6.206	-10.750	-4.544	0.012
Benzene · H_2O	2.0 2.0 2.0	2954.17	150	-0.924	-2.444	-3.367	0.010
			180	-0.929	-2.428	-3.357	0.013

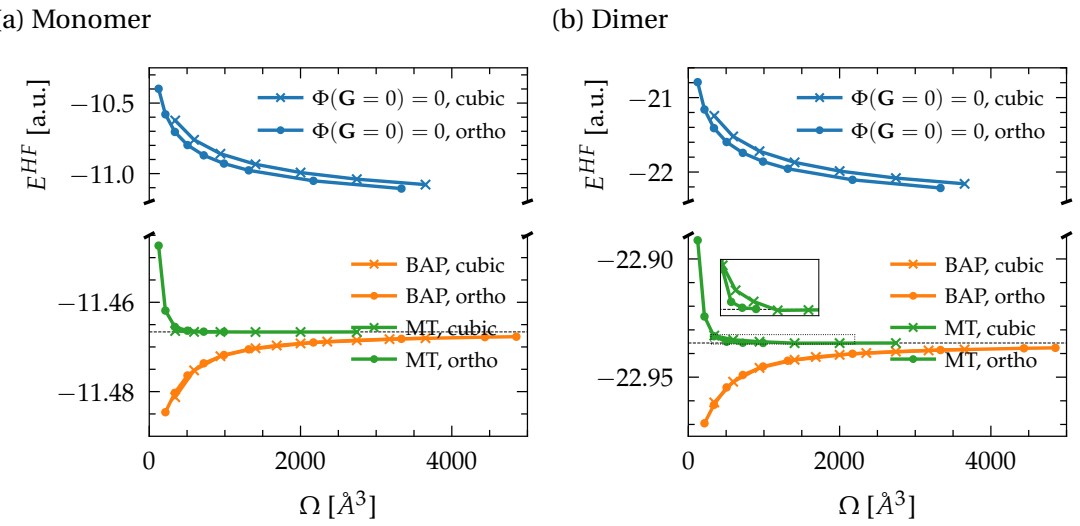


Figure A1: HF energy of the (a) NH_3 monomer and (b) $(NH_3)_2$ dimer for different exchange (Coulomb) potentials. Ω is the volume of expanding cubic or orthorombic supercells around the dimer electron density.

Appendix A. Appendix of chapter 4: Plane-wave vs correlation-consistent MP2

Table A2: MP2 interaction energies in [kcal/mol] of the S22* test systems, uncorrected and with CP correction for the 5 zeta GTO basis sets. The PW values are given with the standard deviation σ resulting from the two consecutive extrapolations: with respect to the virtual orbitals (eq 4.38) and the supercell volume (eq 4.43). Mean signed deviations (MSD) and mean absolute errors (MAE) against PWs are indicated, respectively for each dominant interaction type and over all systems.

S22* test set		cc-pV5Z		aug-cc-pV5Z		PWs
No.	complex	uncorr.	CP-corr.	uncorr.	CP-corr.	CBS $\pm 1\sigma$
Hydrogen-bonded						
1	$(NH_3)_2$	-3.21	-3.08	-3.17	-3.12	-3.19 $\pm$ 0.01
2	$(H_2O)_2$	-5.14	-4.85	-5.04	-4.90	-4.95 $\pm$ 0.01
3	Formic acid dimer	-18.74	-18.22	-18.78	-18.33	-18.37 $\pm$ 0.02
4	Formamide dimer	-15.94	-15.51	-15.96	-15.64	-15.72 $\pm$ 0.01
5	Uracil dimer	-20.57	-20.10	-20.64	-20.21	-20.19 $\pm$ 0.03
6	2-pyridoxine · 2-aminopyridine	-17.54	-17.08	-17.61	-17.20	-17.25 $\pm$ 0.02
		-0.24	*0.14*	*-0.25*	*0.04*	*MSD*
		0.24	*0.14*	*0.26*	*0.05*	*MAE*
Predominant dispersion						
8	$(CH_4)_2$	-0.48	-0.46	-0.51	-0.49	-0.50 $\pm$ 0.01
9	$(C_2H_4)_2$	-1.58	-1.50	-1.63	-1.56	-1.59 $\pm$ 0.01
10	Benzene · CH_4	-1.84	-1.75	-1.90	-1.79	-1.84 $\pm$ 0.01
11	Parallel-displaced benzene dimer	-5.06	-4.78	-5.16	-4.90	-5.06 $\pm$ 0.02
12	Pyrazine dimer	-6.96	-6.67	-7.09	-6.83	-6.92 $\pm$ 0.02
13	Uracil dimer	-11.32	-10.78	-11.46	-11.00	-10.91 $\pm$ 0.04
14	Stacked indole · benzene	-8.27	-7.85	-8.38	-8.01	-8.10 $\pm$ 0.05
		-0.08	*0.16*	*-0.17*	*0.05*	*MSD*
		0.09	*0.16*	*0.17*	*0.08*	*MAE*
Mixed complexes						
16	Ethene · ethine	-1.66	-1.61	-1.71	-1.64	-1.67 $\pm$ 0.01
17	Benzene · H_2O	-3.72	-3.42	-3.64	-3.50	-3.37 $\pm$ 0.02
18	Benzene · NH_3	-2.73	-2.57	-2.74	-2.63	-2.66 $\pm$ 0.01
19	Benzene · HCN	-5.19	-5.05	-5.29	-5.11	-5.13 $\pm$ 0.04
20	T-shaped benzene dimer	-3.71	-3.53	-3.80	-3.59	-3.66 $\pm$ 0.04
21	T-shaped indole · benzene	-7.13	-6.83	-7.22	-6.92	-6.88 $\pm$ 0.02
22	Phenol dimer	-7.91	-7.56	-7.98	-7.66	-7.69 $\pm$ 0.02
		-0.14	*0.07*	*-0.19*	*0.00*	*MSD*
		0.15	*0.09*	*0.19*	*0.05*	*MAE*
		-0.15	0.12	-0.20	0.03	MSD
		0.16	0.13	0.20	0.06	MAE

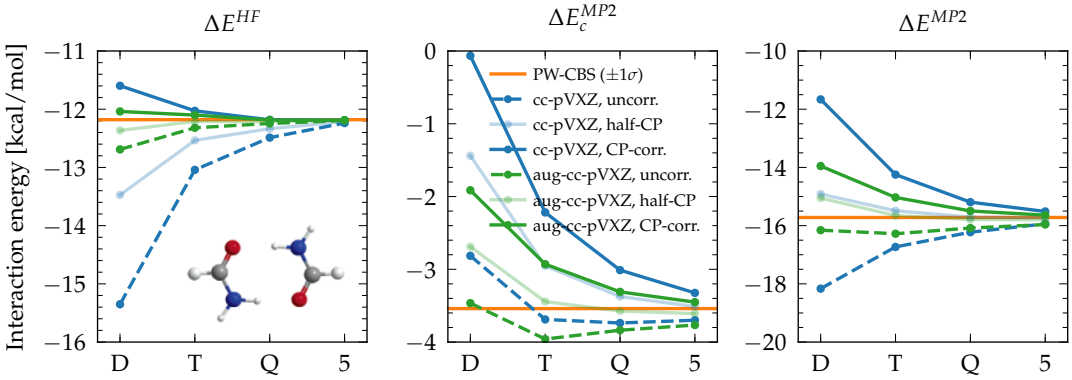


Figure A2: Formamide dimer - Convergence of the HF and MP2c energy contributions to the total MP2 interaction energy for the (aug-)cc-pVXZ basis sets and different treatments of the BSSE. The CBS PW value is also indicated.

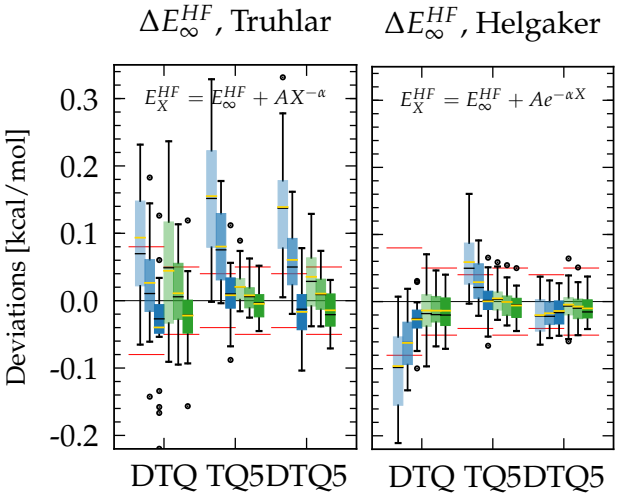


Figure A3: Box plots of the differences $\Delta E^{\mathrm{HF}}_{\mathrm{GTO}} - \Delta E^{\mathrm{HF}}_{\mathrm{PW}}$ between extrapolated GTO and PW HF interaction energies of the S22* test systems. Medians are shown as horizontal black lines and yellow lines stand for the mean signed deviation (MSD). The solid red lines correspond to the smallest maximum deviation obtained with plain Q and 5 zeta basis sets reported in Table 4.2. The legend is given in Figure 4.2.

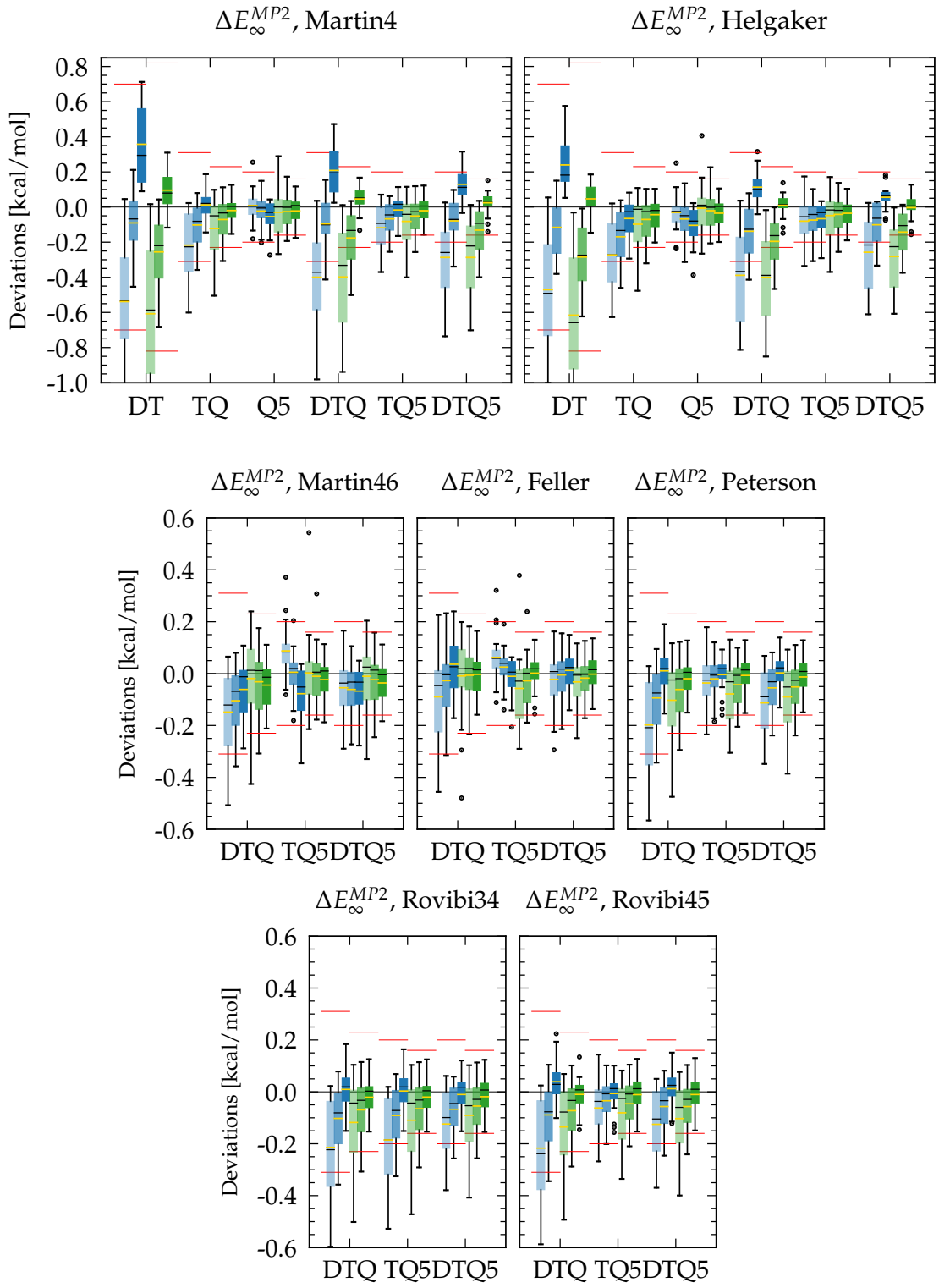


Figure A4: Box plots of the differences $\Delta E_{\mathrm{GTO}}^{\mathrm{MP2}} - \Delta E_{\mathrm{PW}}^{\mathrm{MP2}}$ between extrapolated GTO and PW MP2 interaction energies of the S22* test systems. Medians are shown as horizontal black lines and yellow lines stand for the mean signed deviation (MSD). The solid red lines correspond to the smallest maximum deviation obtained with plain T, Q or 5 zeta basis sets respectively, as reported in Table 4.2. The legend is given in Figure 4.2.

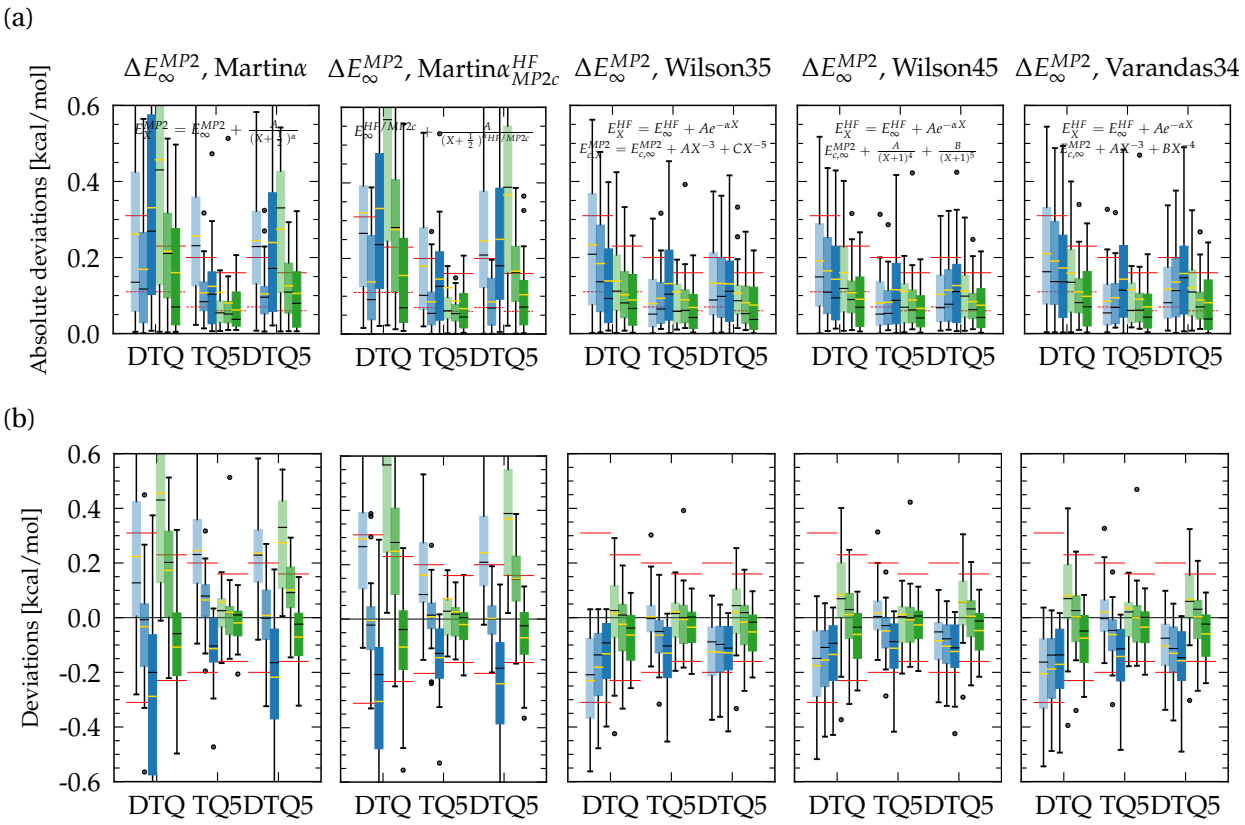
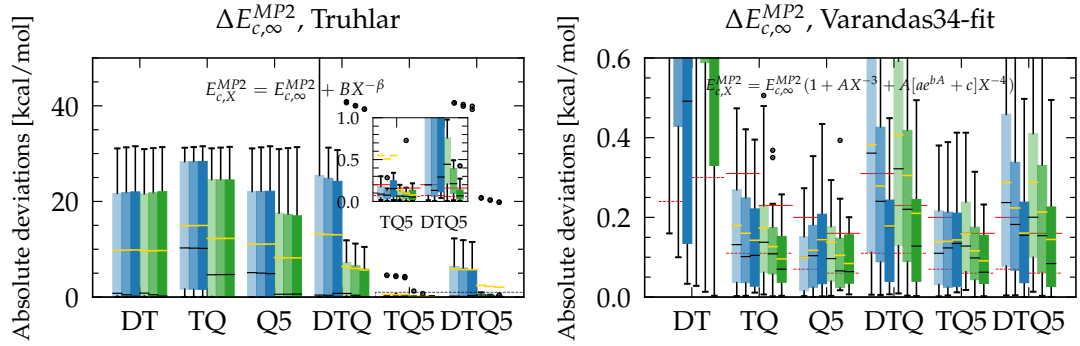


Figure A5: Box plots of the differences $\Delta E_{\mathrm{GTO}}^{\mathrm{MP2}} - \Delta E_{\mathrm{PW}}^{\mathrm{MP2}}$ between extrapolated GTO and PW MP2 interaction energies of the S22* test systems. If applicable, HF and MP2c contributions to the total MP2 energies have been extrapolated separately. Absolute differences are given in (a) while (b) reports signed values. Medians are shown as horizontal black lines and yellow lines stand respectively for the mean absolute error (MAE) in (a) and the mean signed deviation (MSD) in (b). The dashed(solid) red lines correspond to the smallest MAE(maximum deviation) obtained with plain Q or 5 zeta basis sets respectively, as reported in Table 4.2. The legend is given in Figure 4.2.

(a)



(b)

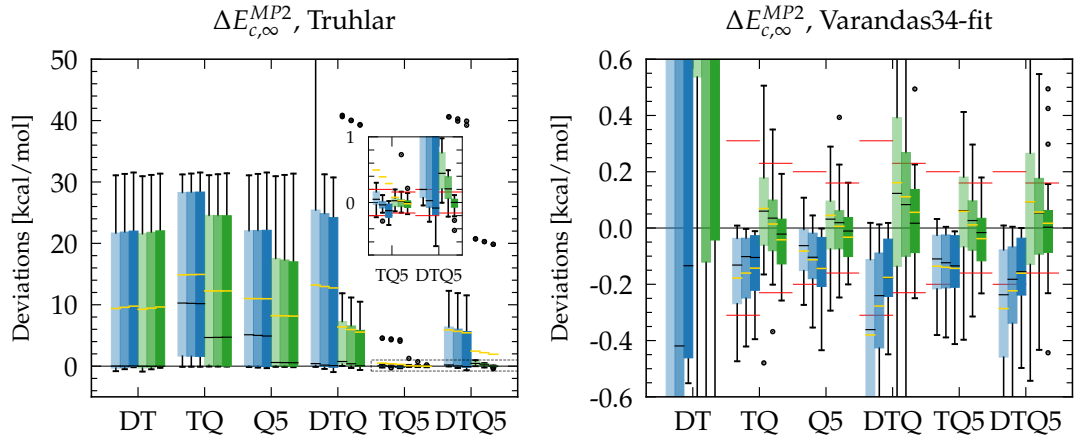
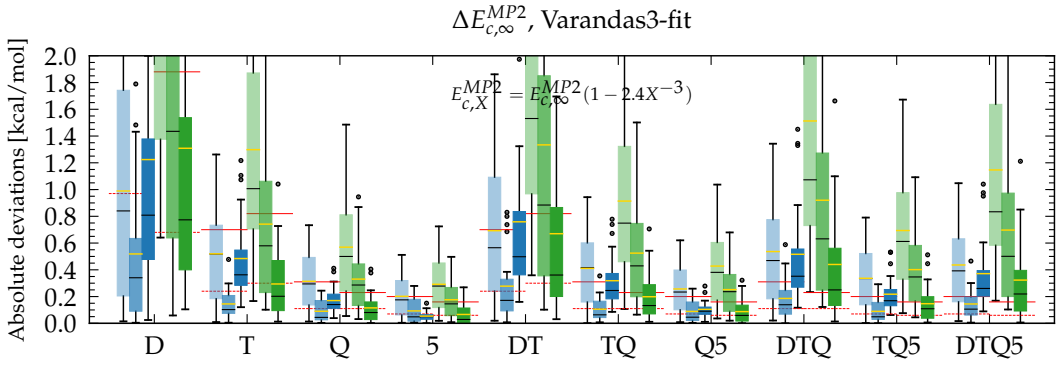


Figure A6: Box plots of the differences $\Delta E_{c,\mathrm{GTO}}^{\mathrm{MP2}} - \Delta E_{c,\mathrm{PW}}^{\mathrm{MP2}}$ between extrapolated GTO and PW MP2c interaction energies of the S22* test systems. Absolute differences are given in (a) while (b) reports signed values. Medians are shown as horizontal black lines and yellow lines stand respectively for the mean absolute error (MAE) in (a) and the mean signed deviation (MSD) in (b). The dashed(solid) red lines correspond to the smallest MAE(maximum deviation) obtained with plain T, Q or 5 zeta basis sets respectively, as reported in Table 4.2. The legend is given in Figure 4.2. Note that the small deviations due to the HF contribution would not counterbalance the results of MP2c.

(a)



(b)

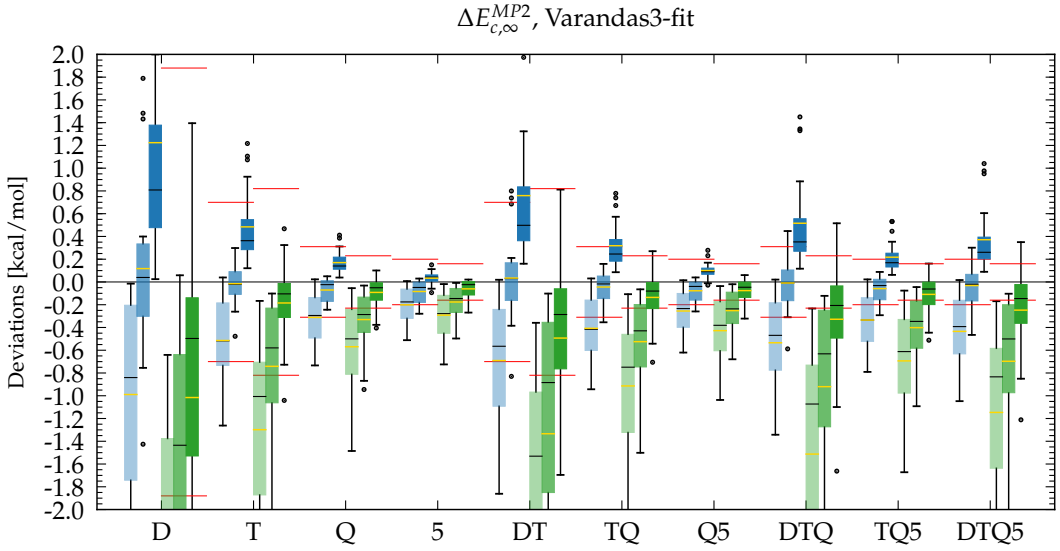
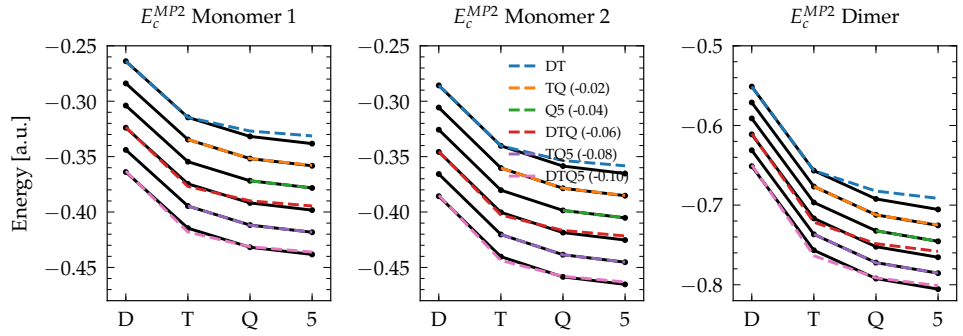
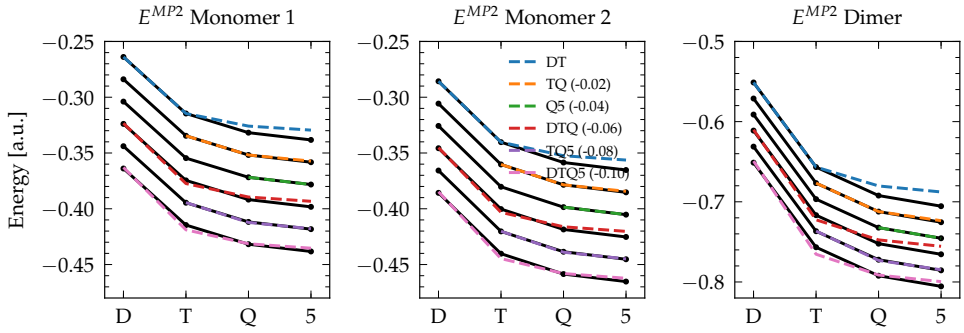


Figure A7: Box plots of the differences $\Delta E_{c,\mathrm{GTO}}^{\mathrm{MP2}} - \Delta E_{c,\mathrm{PW}}^{\mathrm{MP2}}$ between extrapolated GTO and PW MP2c interaction energies of the S22* test systems. Absolute differences are given in (a) while (b) reports signed values. Medians are shown as horizontal black lines and yellow lines stand respectively for the mean absolute error (MAE) in (a) and the mean signed deviation (MSD) in (b). The dashed(solid) red lines correspond to the smallest MAE(maximum deviation) obtained with plain D, T, Q or 5 zeta basis sets respectively, as reported in Table 4.2. The legend is given in Figure 4.2. Note that the small deviations due to the HF contribution would not counterbalance the results of MP2c.

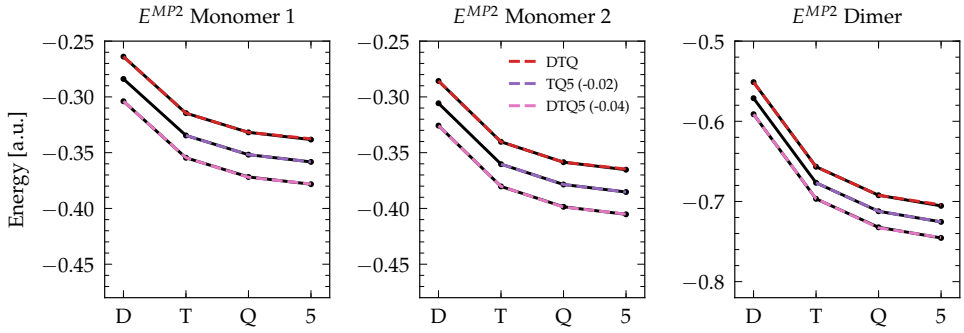
(a) Helgaker (MP2c), $E_{c,X}^{MP2} = E_{c,\infty}^{MP2} + BX^{-3}$



(b) Martin4, $E_X^{MP2} = E_\infty^{MP2} + A(X+\frac{1}{2})^{-4}$



(c) Rovibi34, $E_X^{MP2} = E_\infty^{MP2} + A(X-\frac{1}{2})^{-3} + B(X+\frac{1}{2})^{-4}$



(d) Rovibi45, $E_X^{MP2} = E_\infty^{MP2} + AX^{-4} + B(X+1)^{-5}$

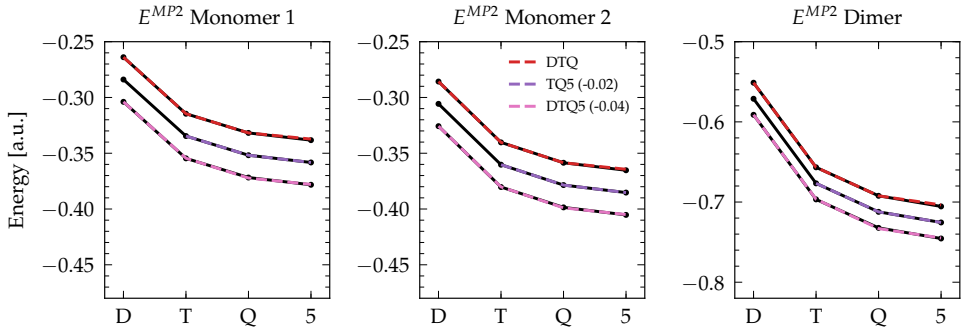
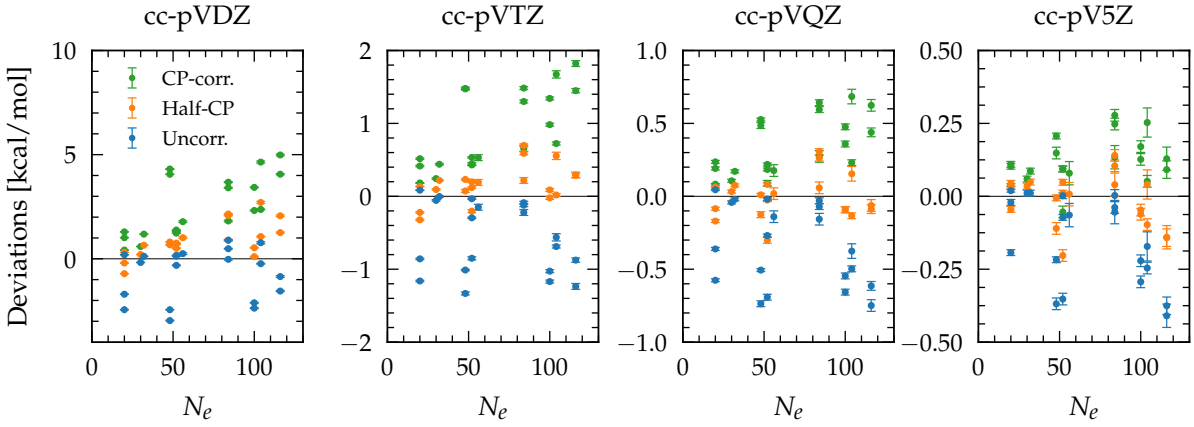


Figure A8: Examples of fitting curves on aug-cc-pVXZ/CP-corrected data points for the ethene-ethine complex. Including the D zeta point in the sequence for (a) *Helgaker* and (b) *Martin4* deteriorates the interpolation while these are better for (c) *Rovibi34* and (d) *Rovibi45*.

Table A3: Quality of interpolation per GTO extrapolation. The mean absolute error (MAE) and root-mean-square deviation (RMSD), as well as the coefficient of determination R^2 are calculated between fitting curves and (aug-)cc-pVXZ data points, always evaluated up to the 5 zeta values (like shown in Figure A8) to reflect the predictive power at larger Xs. Averages on all test systems are reported and include errors on both dimer and monomer energies that are fitted separately. MAE and RMSD are in [kcal/mol]. For simplicity and due to the small *Helgaker*/HF errors, *Helgaker*/MP2 results are based on the MP2c energies only.

Points	Set	Scheme	BSSE corr.	MAE	RMSD	R^2	Scheme	BSSE corr.	MAE	RMSD	R^2
HF - Helgaker vs Truhlar											
DTQ	non-aug	Helgaker	CP	0.058403	0.116806	0.999953	Truhlar	half-CP	0.353033	0.706064	0.999265
	aug	Helgaker	CP	0.099620	0.199236	0.999909	Truhlar	half-CP	0.374430	0.748858	0.998828
TQ5	non-aug	Helgaker	CP	0.000004	0.000005	0.999999	Truhlar	CP	0.000001	0.000001	0.999999
	aug	Helgaker	CP	0.000006	0.000007	0.999999	Truhlar	CP	0.000020	0.000022	0.999999
DTQ5	non-aug	Helgaker	CP	0.051213	0.062020	0.999986	Truhlar	CP	0.292207	0.352576	0.999814
	aug	Helgaker	CP	0.086686	0.104778	0.999974	Truhlar	CP	0.307017	0.370197	0.999712
MP2 - Helgaker vs Martin4											
		Best agreement with PWs according to Table 4.3					Best complementary Helgaker or Martin4				
DT	non-aug	Helgaker	half-CP	8.3797	12.0569	0.972987	Martin4	half-CP	10.5614	15.1795	0.976575
	aug	Helgaker	CP	6.6717	9.5909	0.977574	Martin4	CP	8.6849	12.4628	0.978886
TQ	non-aug	Martin4	CP	0.7062	1.2232	0.998475	Helgaker	CP	0.4706	0.8150	0.998973
	aug	Martin4	CP	0.4649	0.8052	0.999060	Helgaker	CP	0.3253	0.5635	0.999291
Q5	non-aug	Martin4	half-CP	$4 \cdot 10^{-9}$	$4 \cdot 10^{-9}$	1.000000	Helgaker	none	$6 \cdot 10^{-14}$	$7 \cdot 10^{-14}$	1.000000
	aug	Martin4	CP	$5 \cdot 10^{-9}$	$6 \cdot 10^{-9}$	1.000000	Helgaker	CP	$5 \cdot 10^{-14}$	$6 \cdot 10^{-14}$	1.000000
DTQ	non-aug	Helgaker	CP	5.9349	6.8382	0.991316	Martin4	half-CP	7.5922	8.7549	0.992187
	aug	Helgaker	CP	4.7317	5.4354	0.992792	Martin4	CP	6.2159	7.1282	0.993074
TQ5	non-aug	Martin4	CP	0.6558	0.7258	0.999463	Helgaker	CP	0.4303	0.4733	0.999654
	aug	Martin4	CP	0.4317	0.4778	0.999669	Helgaker	CP	0.2975	0.3272	0.999761
DTQ5	non-aug	Helgaker	CP	4.8106	5.8782	0.993582	Martin4	half-CP	6.1927	7.5687	0.994169
	aug	Helgaker	CP	3.8443	4.6838	0.994650	Martin4	CP	5.0918	6.1902	0.994785
MP2 - Rovibi34 vs Rovibi45											
DTQ	non-aug	Rovibi34	CP	0.2955	0.5910	0.999958	Rovibi45	CP	0.5313	1.0627	0.999872
	aug	Rovibi34	CP	0.1537	0.3075	0.999982	Rovibi45	CP	0.3601	0.7202	0.999919
TQ5	non-aug	Rovibi34	CP	$5 \cdot 10^{-10}$	$6 \cdot 10^{-10}$	1.000000	Rovibi45	CP	$1 \cdot 10^{-10}$	$2 \cdot 10^{-10}$	1.000000
	aug	Rovibi34	CP	$6 \cdot 10^{-10}$	$7 \cdot 10^{-10}$	1.000000	Rovibi45	CP	$3 \cdot 10^{-10}$	$4 \cdot 10^{-10}$	1.000000
DTQ5	non-aug	Rovibi34	CP	0.2699	0.3393	0.999986	Rovibi45	CP	0.4874	0.6121	0.999958
	aug	Rovibi34	CP	0.1404	0.1765	0.999994	Rovibi45	CP	0.3303	0.4149	0.999973
MP2 - Peterson vs Feller											
DTQ	non-aug	Peterson	CP	0.0691	0.1381	0.999996	Feller	CP	0.4250	0.8501	0.999887
	aug	Peterson	CP	0.1317	0.2634	0.999989	Feller	CP	0.3234	0.6468	0.999914
TQ5	non-aug	Peterson	CP	$1 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	1.000000	Feller	CP	$9 \cdot 10^{-7}$	$1 \cdot 10^{-6}$	1.000000
	aug	Peterson	CP	$9 \cdot 10^{-9}$	$1 \cdot 10^{-8}$	1.000000	Feller	CP	$4 \cdot 10^{-6}$	$4 \cdot 10^{-6}$	1.000000
DTQ5	non-aug	Peterson	CP	0.0618	0.0767	0.999999	Feller	CP	0.3460	0.4083	0.999973
	aug	Peterson	CP	0.1178	0.1463	0.999997	Feller	CP	0.2648	0.3130	0.999979

(a) Non-extrapolated



(b) Extrapolated, closest to PWs in the CBS limit

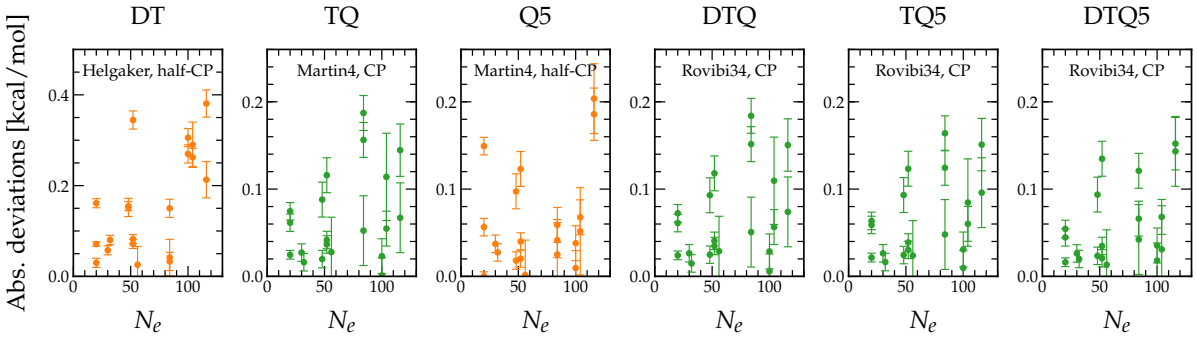
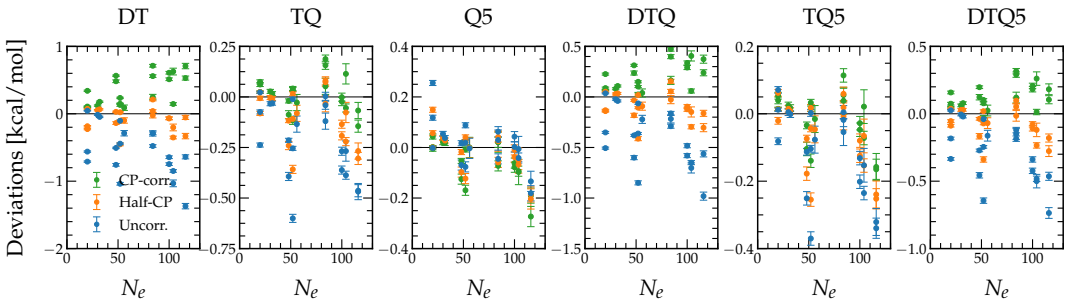
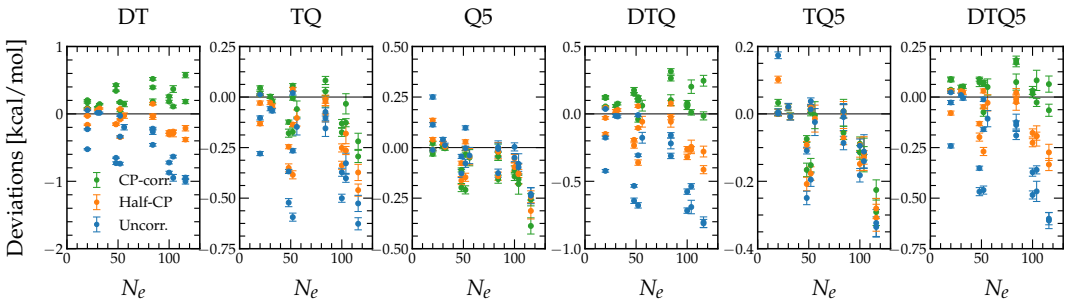


Figure A9: Deviations between the cc-pVXZ and PW MP2 interaction energies as a function of the number of electrons N_e in the dimer system. (a) for plain basis sets, (b) for energies extrapolated to the CBS limit with best schemes of Tables 4.3 and 4.4.

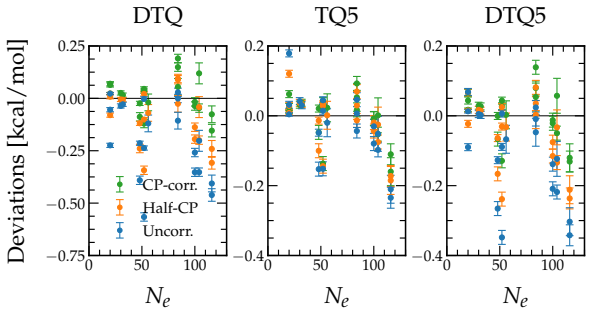
(a) cc-pVXZ, Martin4



(b) cc-pVXZ, Helgaker



(c) cc-pVXZ, Peterson



(d) cc-pVXZ, Rovibi34

(e) cc-pVXZ, Rovibi45

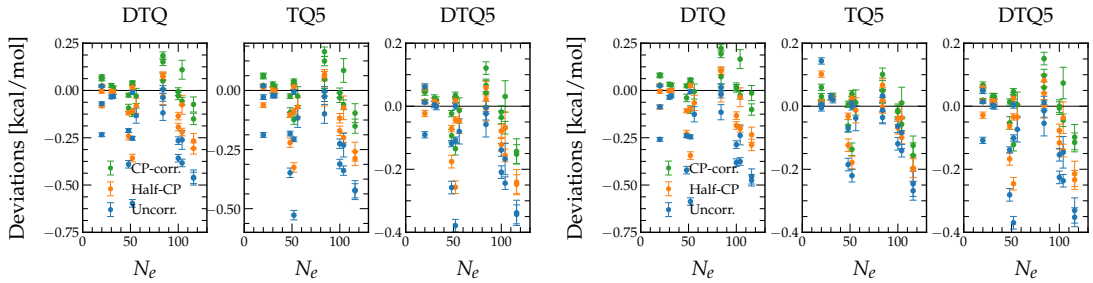
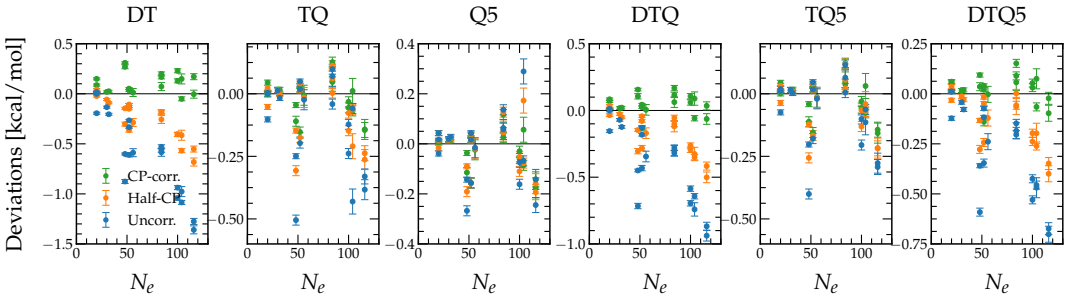
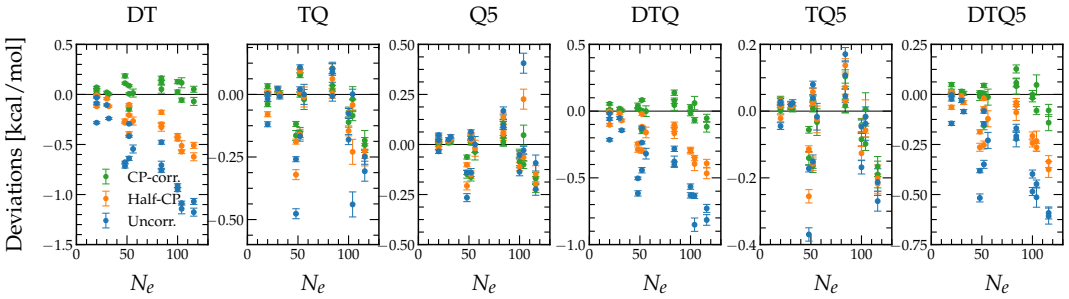


Figure A10: Deviations between the extrapolated cc-pVXZ and PW MP2 interaction energies as a function of the number of electrons N_e in the dimer system. Extrapolated to the CBS limit with (a) Martin4, (b) Helgaker, (c) Peterson, (d) Rovibi34, (e) Rovibi45. For all schemes, the deviations (in absolute value) tend to increase with N_e.

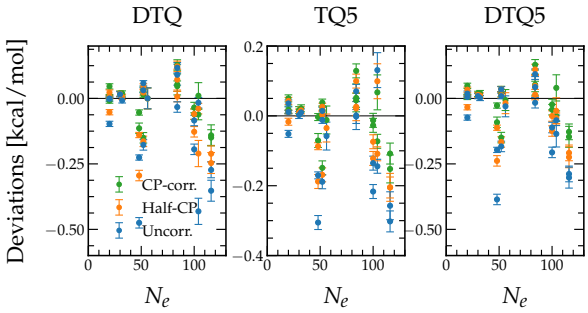
(a) aug-cc-pVXZ, Martin4



(b) aug-cc-pVXZ, Helgaker



(c) aug-cc-pVXZ, Peterson



(d) aug-cc-pVXZ, Rovibi34

(e) aug-cc-pVXZ, Rovibi45

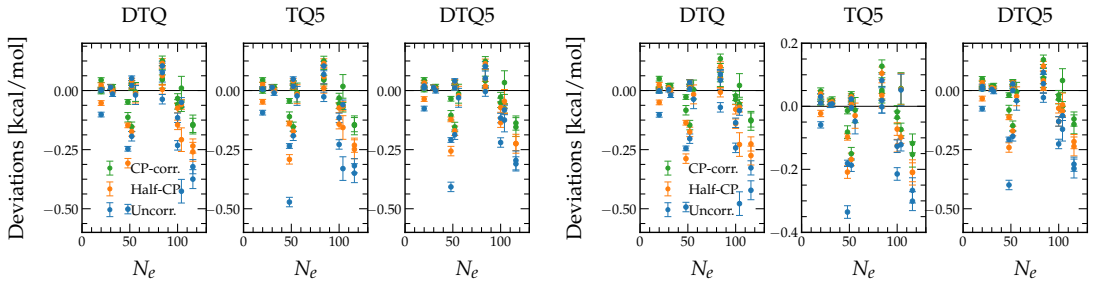


Figure A11: Deviations between the extrapolated aug-cc-pVXZ and PW MP2 interaction energies as a function of the number of electrons N_e in the dimer system. Extrapolated to the CBS limit with (a) Martin4, (b) Helgaker, (c) Peterson, (d) Rovibi34, (e) Rovibi45. For all schemes, the deviations (in absolute value) tend to increase with N_e.

B Appendix of chapter 5: Plane-wave Monte Carlo MP2

Summary of results and fitting errors

For fitting curves, cf. the respective system subsections below.

Table B1: MP(s)2 energies and differences in a.u. obtained from extrapolation. $\varepsilon_c^{\text{gap}}$ denotes the threshold for stochastic sampling and N_{MC} is the number of terms sampled per virtual contribution, as explained in Chapter 5. e^- is the number of electrons in the system. If not explicitly stated, all values are given in atomic units.

System	Ethylene	Benzene			
	Crystal	Crystal	Monomer	Dimer	Binding
E_c^{MP2}	-0.78054±0.00010	-4.69164±0.00073	-1.05681±0.00094	-2.12780±0.00215	-0.01417
E_c^{MPs2}	-0.78056±0.00011	-4.69149±0.00093	-1.05695±0.00088	-2.12777±0.00216	-0.01387
$\Delta E_{\text{MPs2}}^{\text{MP2}}$	$2 \cdot 10^{-5}$	$-1.5 \cdot 10^{-4}$	$1.4 \cdot 10^{-4}$	$-3 \cdot 10^{-5}$	$-3.0 \cdot 10^{-4}$
$\Delta E_{\text{MPs2}}^{\text{MP2}}/e^-$	$8 \cdot 10^{-7}$	$-1 \cdot 10^{-6}$	$5 \cdot 10^{-6}$	$-5 \cdot 10^{-7}$	$-5 \cdot 10^{-6}$
$\varepsilon_c^{\text{gap}}$ [eV]	120	120	120	120	120
N_{MC}	12000	12000	12000	12000	12000

Table B2: MPs2 energies in a.u. obtained from extrapolation. $\varepsilon_c^{\text{gap}}$ denotes the threshold for stochastic sampling and N_{MC} is the number of terms sampled per virtual contribution, as explained in Chapter 5.

System	Hydronium ion solvated in 32 water molecules		
	Crystal	Crystal	Crystal
E_c^{MPs2}	-10.22771±0.01009	-10.23190±0.01047	-10.22952±0.01087
$\varepsilon_c^{\text{gap}}$ [eV]	60	90	120
N_{MC}	12000	12000	12000

Ethylene crystal

Monoclinic cell:

$a = 6.620$ Å, $b = 4.626$ Å, $c = 4.067$ Å, $\alpha = 94.39$ °, $\beta = 90.00$ °, $\gamma = 90.00$ °.

Total number of electronic states: 11400.

Plane-wave cutoff energy for orbitals: $E^{\phi}_{\text{cut}} = 140$ Ry.

Plane-wave cutoff energy for densities: $E^{\rho}_{\text{cut}} = 560$ Ry.

Element	X	Y	Z	Element	X	Y	Z
C	0.10029	0.61226	0.21569	H	-0.89666	-1.21379	0.17293
C	-0.10029	-0.61226	-0.21569	H	0.52701	-1.04440	-0.96901
C	4.82653	-1.37057	-0.52784	H	5.39184	-1.48263	0.37489
C	3.76335	-0.60043	-0.57359	H	5.15893	-1.90883	-1.39251
H	0.89666	1.21379	-0.17293	H	3.19804	-0.48837	-1.47633
H	-0.52701	1.04440	0.96901	H	3.43095	-0.06218	0.29107

Table B3: Atomic coordinates of the ethylene crystal in Å.

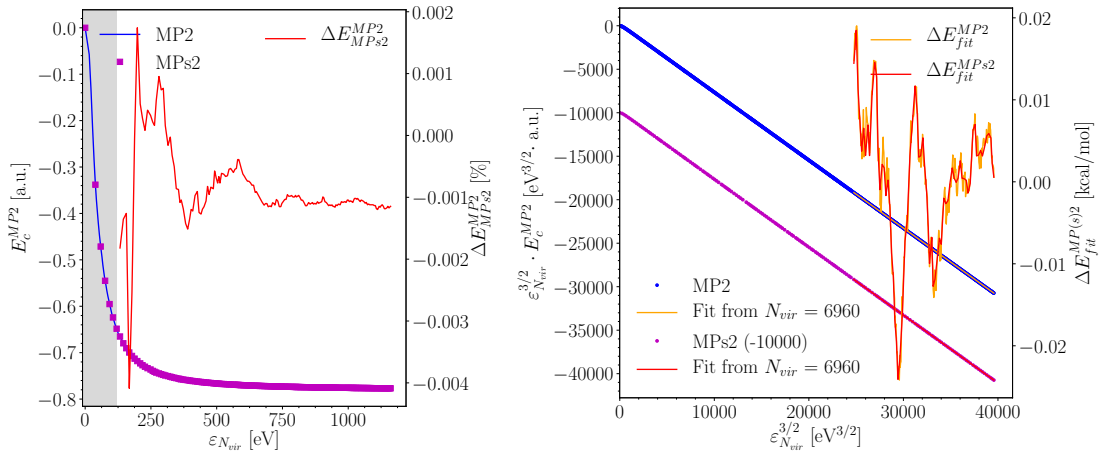


Figure B1: Left: MP(s)2 correlation energy of the ethylene crystal as a function of the highest eigenvalue truncating the sum. The difference between MP2 and MPs2 is given along the secondary axis relatively to the MP2 reference value (red). Right: Fitting curves over a large range of MP(s)2 energies and difference between fit and calculated values on the secondary axis. $N_{\text{vir}}^{\text{max}} = 11158$.

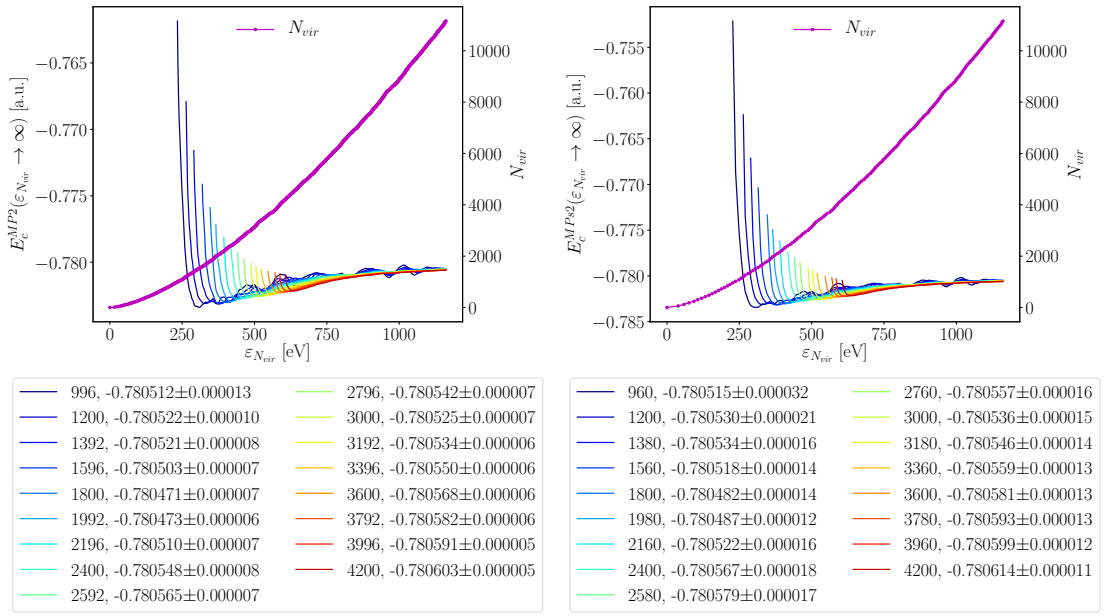


Figure B2: Left: Extrapolated MP2 energy of the ethylene crystal for a window range ending at N_{vir}, as a function of the corresponding eigenvalue $\varepsilon_{N_{\text{vir}}}$. Right: Same for MPs2 energy. Legends show numbers of orbitals in fitting windows and extrapolated values in a.u. at $N_{\text{vir}}^{\text{max}}$ with asymptotic standard deviation. The number of virtual orbitals corresponding to the choice of $\varepsilon_{N_{\text{vir}}}$ is shown on the secondary axis. Windows containing at least 1800 orbitals are taken for calculating the MP(s)2 values reported in Table B1.

Benzene crystal

Orthorhombic cell:

$a = 9.550$ Å, $b = 6.640$ Å, $c = 6.920$ Å, $\alpha = 90.00$ °, $\beta = 90.00$ °, $\gamma = 90.00$ °.

Total number of electronic states: 12000.

Plane-wave cutoff energy for orbitals: $E^{\phi}_{\text{cut}} = 150$ Ry.

Plane-wave cutoff energy for densities: $E^{\rho}_{\text{cut}} = 600$ Ry.

Element	X	Y	Z	Element	X	Y	Z
C	1.30336	-0.36606	0.32000	H	2.30301	-0.66474	0.50752
C	0.35092	-0.23805	1.32843	H	0.60089	-0.45635	2.36092
C	0.95009	-0.11336	-0.99976	H	1.67633	-0.20865	-1.78117
C	-0.95009	0.11336	0.99976	H	-1.67633	0.20865	1.78117
C	-0.35092	0.23805	-1.32843	H	-0.60089	0.45635	-2.36092
C	-1.30336	0.36606	-0.32000	H	-2.30301	0.66474	-0.50752
C	6.28437	3.17055	-2.19721	H	7.31273	3.34626	-2.38504
C	5.38596	4.19140	-1.89523	H	5.71169	5.22500	-1.85332
C	5.83638	1.85787	-2.27835	H	6.52051	1.06682	-2.50925
C	4.04539	3.88904	-1.70661	H	3.36125	4.68009	-1.47571
C	4.49581	1.55551	-2.08973	H	4.17007	0.52191	-2.13165
C	3.59739	2.57636	-1.78775	H	2.56903	2.40065	-1.59992
C	3.26493	1.76649	3.16518	H	2.24442	1.48500	3.21991
C	4.07921	1.87949	4.28975	H	3.69258	1.66660	5.28039
C	3.79213	2.01171	1.90330	H	3.17292	1.92785	1.03345
C	5.41781	2.20845	4.13460	H	6.03702	2.29231	5.00445
C	5.13073	2.34068	1.74815	H	5.51736	2.55357	0.75750
C	5.94501	2.45367	2.87271	H	6.96552	2.73517	2.81798
C	-1.28315	5.38412	0.02158	H	-2.27496	5.57700	-0.29904
C	-0.41484	6.38995	0.43970	H	-0.72571	7.42896	0.43977
C	-0.85068	4.06395	-0.00168	H	-1.51199	3.28434	-0.32102
C	0.88419	6.06515	0.80183	H	1.54551	6.84475	1.12118
C	0.44836	3.73915	0.36046	H	0.75923	2.70014	0.36038
C	1.31667	4.74498	0.77857	H	2.30847	4.55210	1.09919

Table B4: Atomic coordinates of the benzene crystal in Å.

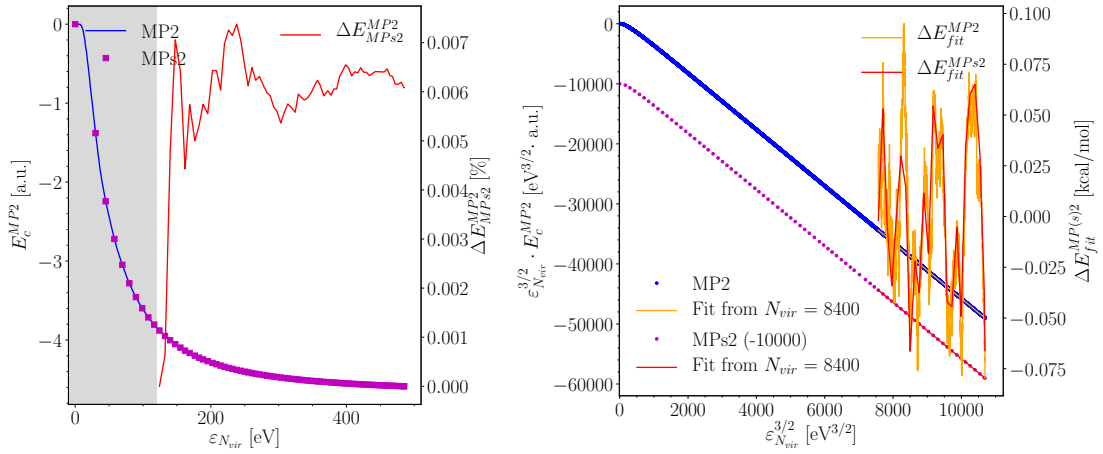


Figure B3: Left: MP(s)2 correlation energy of the benzene crystal as a function of the highest eigenvalue truncating the sum. The difference between MP2 and MPs2 is given along the secondary axis relatively to the MP2 reference value (red). Right: Fitting curves on large range of MP(s)2 results and difference between fit and calculated values on the secondary axis. $N_{\text{vir}}^{\text{max}} = 11880$.

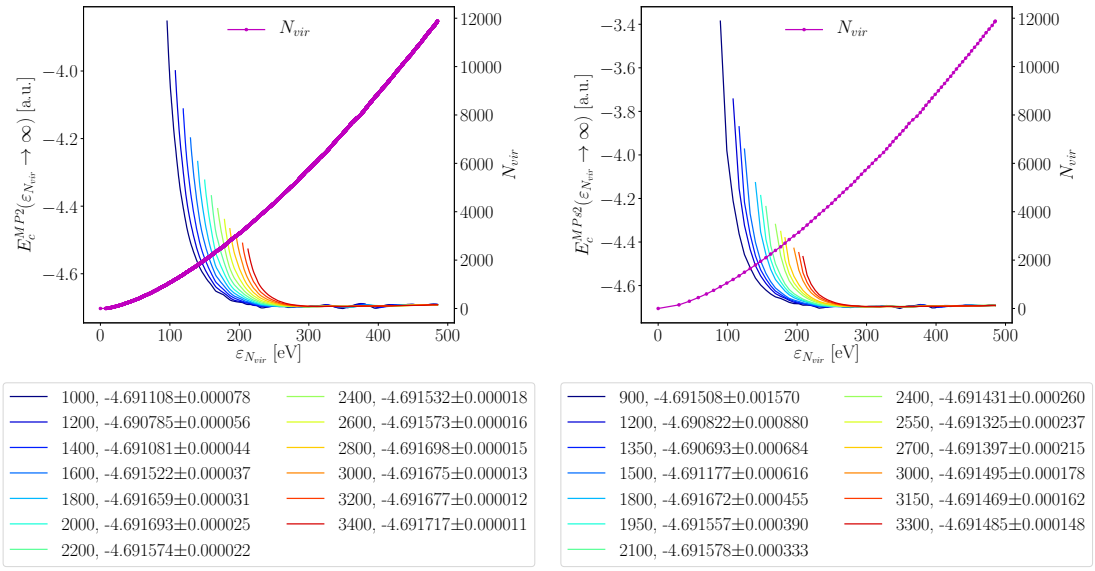


Figure B4: Left: Extrapolated MP2 energy of the benzene crystal for a window range ending at N_{vir}, as a function of the corresponding eigenvalue $\varepsilon_{N_{\text{vir}}}$. Right: Same for MPs2 energy. Legends show numbers of orbitals in fitting windows and extrapolated values in a.u. at $N_{\text{vir}}^{\text{max}}$ with asymptotic standard deviation. Respective number of virtual orbitals is shown on the secondary axis. Windows containing at least 1800 orbitals are taken for calculating MP(s)2 values reported in Table B1.

Dependency of MPs2 on the continuum cutoff $\varepsilon_c^{\text{gap}}$

We performed 5 additional independent runs of MPs2 calculations of the benzene crystal in order to assess the stochastic variance at different cutoffs $\varepsilon_c^{\text{gap}}$. Results are depicted in Figure B5.

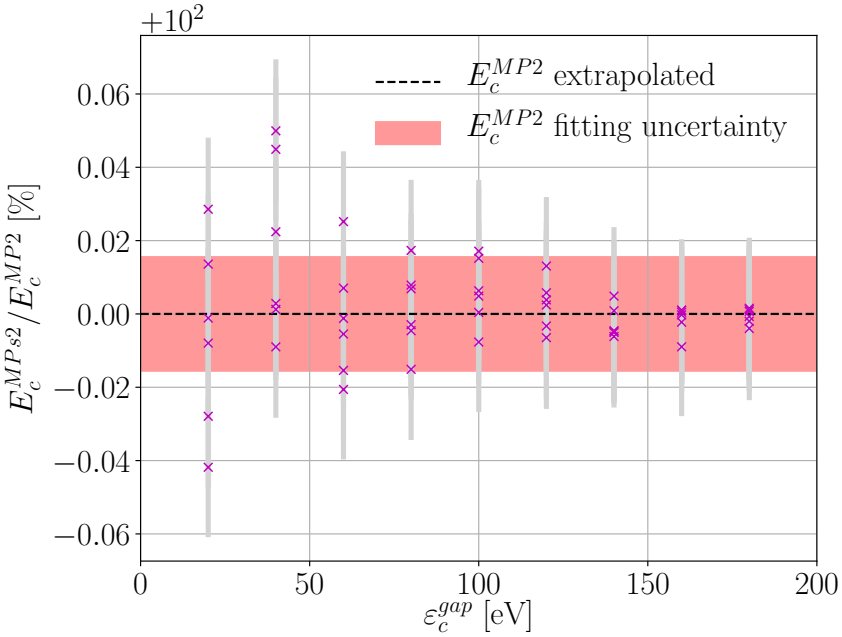


Figure B5: Deviation of the MPs2 energy from the MP2 reference energy of the benzene crystal. 6 independent values are plotted for different $\varepsilon_c^{\text{gap}}$ between 20 eV and 180 eV. Each point and errorbar correspond to the mean extrapolated value and its standard deviation like obtained in Figure B4-right.

Benzene monomer and parallel sandwich dimer

Cubic box:

$a = 15.00$ Å, $b = 11.25$ Å, $c = 11.25$ Å, $\alpha = 90.00$ °, $\beta = 90.00$ °, $\gamma = 90.00$ °.

Total number of electronic states: 15000.

Plane-wave cutoff energy for orbitals: $E^{\phi}_{\text{cut}} = 150$ Ry.

Plane-wave cutoff energy for densities: $E^{\rho}_{\text{cut}} = 600$ Ry.

Elem.	X	Y	Z	Elem.	X	Y	Z
C	0.000000	0.000000	1.394259	C	3.700000	0.000000	1.394259
C	0.000000	1.207465	0.697130	C	3.700000	1.207465	0.697130
C	0.000000	1.207464	-0.697130	C	3.700000	1.207464	-0.697130
C	0.000000	0.000000	-1.394260	C	3.700000	0.000000	-1.394260
C	0.000000	-1.207465	-0.697130	C	3.700000	-1.207465	-0.697130
C	0.000000	-1.207465	0.697130	C	3.700000	-1.207465	0.697130
H	0.000000	0.000000	2.476431	H	3.700000	0.000000	2.476431
H	0.000000	2.144653	1.238216	H	3.700000	2.144653	1.238216
H	0.000000	2.144653	-1.238216	H	3.700000	2.144653	-1.238216
H	0.000000	0.000000	-2.476432	H	3.700000	0.000000	-2.476432
H	0.000000	-2.144653	-1.238216	H	3.700000	-2.144653	-1.238216
H	0.000000	-2.144653	1.238216	H	3.700000	-2.144653	1.238216

Table B5: Atomic coordinates of the benzene sandwich in Å. Coordinates for the benzene monomer correspond to only one of the two columns.

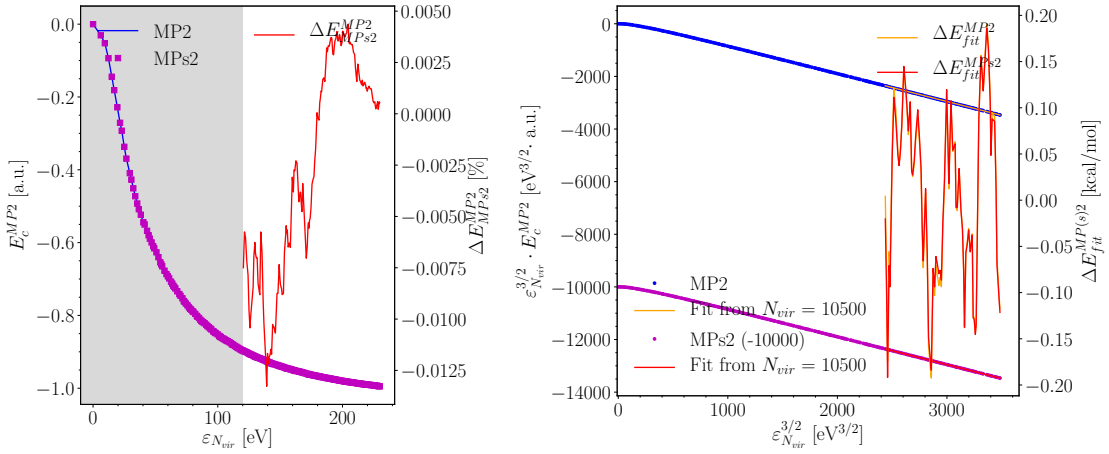


Figure B6: Left: MP(s)2 correlation energy of the benzene monomer as a function of the highest eigenvalue truncating the sum. The difference between MP2 and MPs2 is given along the secondary axis relatively to the MP2 reference value (red). Right: Fitting curves on large range of MP(s)2 results and difference between fit and calculated value on the secondary axis. $N_{\text{vir}}^{\text{max}} = 14985$.

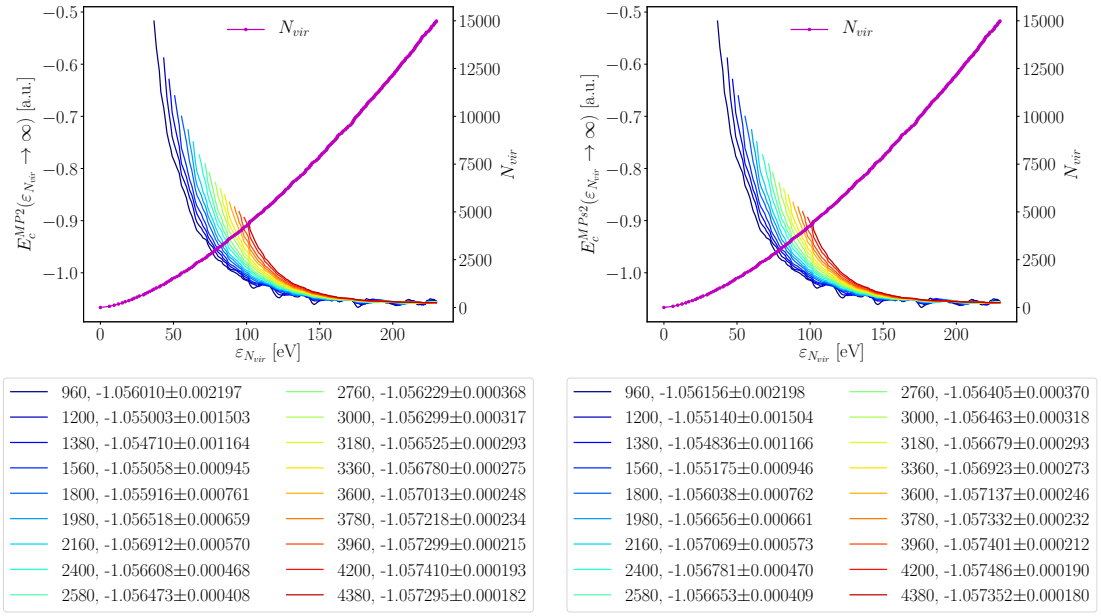


Figure B7: Left: Extrapolated MP2 energy of the benzene monomer for a window range ending at N_{vir}, as a function of the corresponding eigenvalue $\varepsilon_{N_{\text{vir}}}$. Right: Same for MPs2 energy. Legends show numbers of orbitals in fitting windows and extrapolated values in a.u. at $N_{\text{vir}}^{\text{max}}$ with asymptotic standard deviation. Respective number of virtual orbitals is shown on the secondary axis. Windows containing at least 1980 orbitals are taken for calculating MP(s)2 values reported in Table B1.

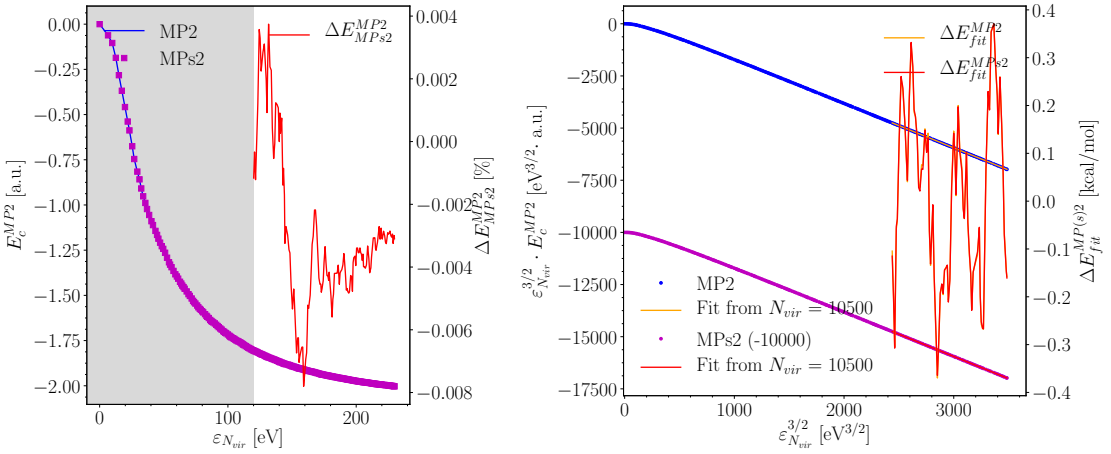


Figure B8: Left: MP(s)2 correlation energy of the benzene sandwich as a function of the highest eigenvalue truncating the sum. The difference between MP2 and MPs2 is given along the secondary axis relatively to the MP2 reference value (red). Right: Fitting curves on large range of MP(s)2 results and difference between fit and calculated values on the secondary axis. $N_{\text{vir}}^{\text{max}} = 14970$.

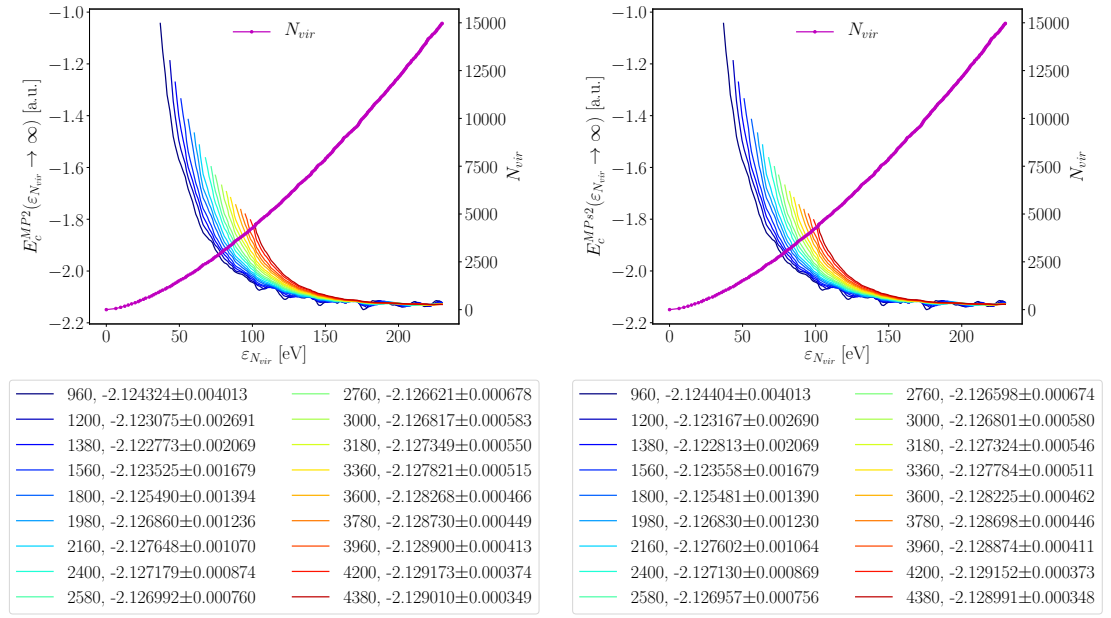


Figure B9: Left: Extrapolated MP2 energy of the benzene sandwich for a window range ending at N_{vir}, as a function of the corresponding eigenvalue $\varepsilon_{N_{\text{vir}}}$. Right: Same for MPs2 energy. Legends show numbers of orbitals in fitting windows and extrapolated values in a.u. at $N_{\text{vir}}^{\text{max}}$ with asymptotic standard deviation. Respective number of virtual orbitals is shown on the secondary axis. Windows containing at least 1980 orbitals are taken for calculating MP(s)2 values reported in Table B1.

Comparison between PW, GPW and Gaussian basis sets

For the sake of comparison to other basis sets, MP2 energies with atom-centered all-electron and Gaussian/plane-wave (GPW) bases are given for the benzene monomer and a dimer in sandwich configuration.

Table B6: Comparison of our pseudopotential/PW MP2 correlation energies with atom-centered all-electron[332] and GPW calculations[296] for the benzene monomer and sandwich dimer. Values are given in atomic units. GPW calculations have been performed with the same GTH-HF pseudopotentials used for PW calculations and a density cutoff of $E^{\rho}_{\text{cut}} = 600$ Ry.

Basis function	Basis set	Monomer	Dimer	Binding
Gaussian, MP2	aug-cc-pVDZ	-0.81038	-1.63681	-0.01606
Gaussian, MP2	aug-cc-pVTZ	-0.96343	-1.94210	-0.01523
Gaussian, MP2	aug-cc-pVQZ	-1.01553	-2.04544	-0.01438
GPW, MP2	GPW, cc-DZ	-0.75888	-1.53066	-0.01290
GPW, MP2	GPW, cc-TZ	-0.93486	-1.88456	-0.01484
GPW, MP2	GPW, cc-QZ	-1.00305	-2.02027	-0.01416
PW, MP2	E^{ϕ}_{cut} = 150 Ry	-1.05681	-2.12780	-0.01417
PW, MPs2	E^{ϕ}_{cut} = 150 Ry	-1.05695	-2.12777	-0.01387

Dependency of MPs2 on the continuum cutoff $\varepsilon_c^{\text{gap}}$ (monomer)

We performed 5 additional independent runs of MPs2 calculations of the benzene monomer in order to assess the stochastic variance at different cutoffs $\varepsilon_c^{\text{gap}}$. Results are depicted in Figure B10.

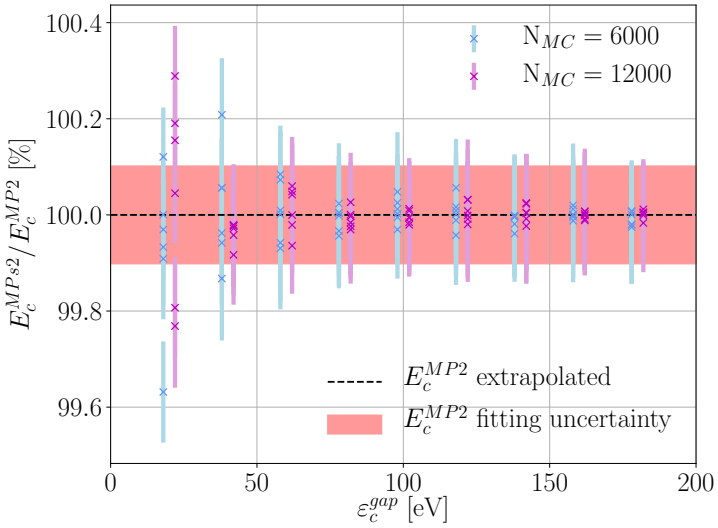


Figure B10: Deviation of the MPs2 energy from the MP2 reference energy of the benzene monomer. 6 independent values are plotted for different $\varepsilon_c^{\text{gap}}$ between 20 eV and 180 eV at two sampling parameters $N_{\text{MC}} = 6000$ and $N_{\text{MC}} = 12000$. Each point and errorbar correspond to the mean extrapolated value and its standard deviation like obtained in Figure B7-right.

Hydronium ion solvated in 32 water molecules

Cubic cell:

$a = 9.95991$ Å, $b = 9.95991$ Å, $c = 9.95991$ Å, $\alpha = 90.00$ °, $\beta = 90.00$ °, $\gamma = 90.00$ °.

Total number of electronic states: 12122.

Plane-wave cutoff energy for orbitals: $E^{\phi}_{\text{cut}} = 150$ Ry.

Plane-wave cutoff energy for densities: $E^{\rho}_{\text{cut}} = 600$ Ry.

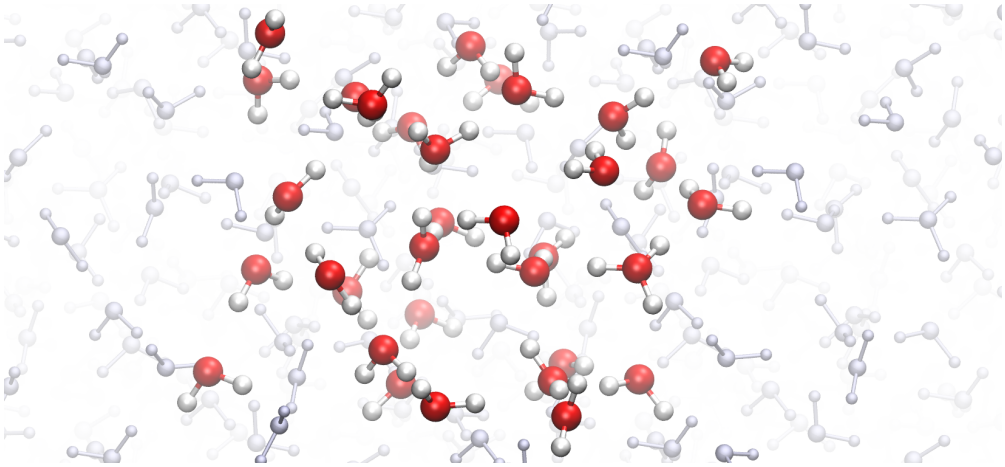


Figure B11: Hydronium ion solvated in 32 water molecules (264 electrons explicitly accounted for). Calculation of E_c^{MPs2} takes between 5 to 15 h on 25 16-core compute nodes. Molecules within the simulation supercell are highlighted. Using $\varepsilon_c^{\text{gap}} = 120$ eV and $N_{\text{MC}} = 12000$ yields an extrapolated $E_c^{\text{MPs2}} = -10.22952$ a.u.

Element	X	Y	Z	Element	X	Y	Z
O	0.7327	-4.2440	2.1886	H	-2.7608	-4.2804	3.0891
O	-0.8347	-4.7227	4.3691	H	-3.5930	-0.0389	-3.2716
O	2.6518	-1.5714	-0.1309	H	-3.0803	1.1482	-2.2743
O	3.2623	-4.4641	2.6462	H	3.3247	1.5747	2.2926
O	-4.7361	4.8510	-2.0459	H	3.0265	2.7969	3.3338
O	-2.3920	0.3963	3.9231	H	4.7065	1.8430	-2.7522
O	-0.5493	-4.5740	-2.0499	H	3.1319	2.0416	-3.1370
O	2.2782	-3.3931	-4.3667	H	-1.9556	-1.5794	-0.0367
O	-3.5142	-4.3400	2.4342	H	-1.7355	0.0251	0.1726
O	-3.8442	0.5428	-2.4979	H	3.1616	-4.3000	0.7191
O	2.6625	2.2735	2.5634	H	2.7360	-5.0244	-0.6811
O	4.0301	2.4810	-3.1204	H	1.5443	-1.0463	-1.4591
O	-1.3385	-0.8110	-0.2060	H	-0.0408	-0.7072	-1.2621
O	3.3646	-4.3723	-0.2574	H	5.5462	3.4712	4.3092
O	0.7415	-0.6233	-1.8793	H	4.1734	3.2213	5.1573
O	4.5585	3.6260	4.3279	H	1.3289	3.2796	-0.5124
O	1.4453	3.8230	-1.3437	H	1.3588	3.2301	-2.1443
O	0.7650	1.2950	4.2396	H	1.2158	1.6650	3.4272
O	-2.5713	3.1043	-4.8839	H	-0.2192	1.4615	4.1796
O	-4.1207	3.2564	1.0086	H	-2.3393	2.1786	-5.1828
O	0.4496	3.1128	1.4121	H	-2.3958	3.1917	-3.9033
O	4.0998	0.0228	4.2341	H	-4.0469	3.9980	1.6755
O	-0.5407	-2.4217	-3.9645	H	-4.8820	3.4416	0.3872
O	-4.5374	-1.1430	-0.2440	H	1.3081	2.8588	1.8577
O	-1.6032	1.7664	0.5732	H	0.2815	4.0896	1.5449
O	-3.1993	-1.5826	-4.4357	H	4.4722	0.2911	3.3456
O	0.5592	-1.4779	1.7951	H	4.8505	-0.1433	4.8736
O	-1.8413	2.8694	-2.0738	H	-0.5906	-3.0520	-3.1897
O	0.9314	1.6635	-3.0237	H	-0.0056	-1.6175	-3.7060
O	4.8851	0.6455	1.8791	H	-4.2505	-0.6992	-1.0929
O	-4.7775	-3.9950	-4.5931	H	-3.9724	-1.9523	-0.0834
O	-2.9360	-3.2201	-0.0728	H	-2.3879	2.3426	0.8019
O	1.8651	-1.3161	4.1881	H	-0.7686	2.1804	0.9365
H	0.1804	-4.3091	3.0196	H	-3.8725	-2.3207	-4.3898
H	0.4583	-3.4352	1.6684	H	-2.2782	-1.9717	-4.4453
H	-0.8907	-3.8505	4.8551	H	1.3775	-1.3132	1.2443
H	-1.4470	-5.3849	4.8012	H	-0.2523	-1.2143	1.2736
H	2.6813	-2.5699	-0.0836	H	-1.1466	2.2371	-2.4168
H	3.5705	-1.2189	-0.3089	H	-2.0815	2.6232	-1.1348
H	2.2641	-4.5114	2.6070	H	1.0048	0.8099	-2.5081
H	3.6143	-5.2567	3.1440	H	0.8075	1.4570	-3.9943
H	-5.2043	3.9908	-2.2480	H	5.3467	1.4906	1.6095
H	-5.2257	5.3344	-1.3202	H	4.9527	-0.0223	1.1379
H	-2.8338	0.3816	3.0261	H	-4.4389	-4.2859	-3.6983
H	-2.6987	-0.3905	4.4587	H	-5.0180	-4.7996	-5.1361
H	0.2035	-5.0868	-1.6374	H	-2.9873	-3.6787	0.8143
H	-1.2460	-5.2114	-2.3791	H	-2.9264	-3.9045	-0.8018
H	3.1464	-3.8096	-4.0969	H	1.8677	-2.1810	4.7284
H	1.5234	-3.8864	-3.9344	H	2.7693	-0.8629	4.0578
H	-4.2719	-3.7659	2.7445	H	1.2248	-1.2816	3.3950

Table B7: Atomic coordinates of the hydronium ion solvated in 32 water molecules in Å.

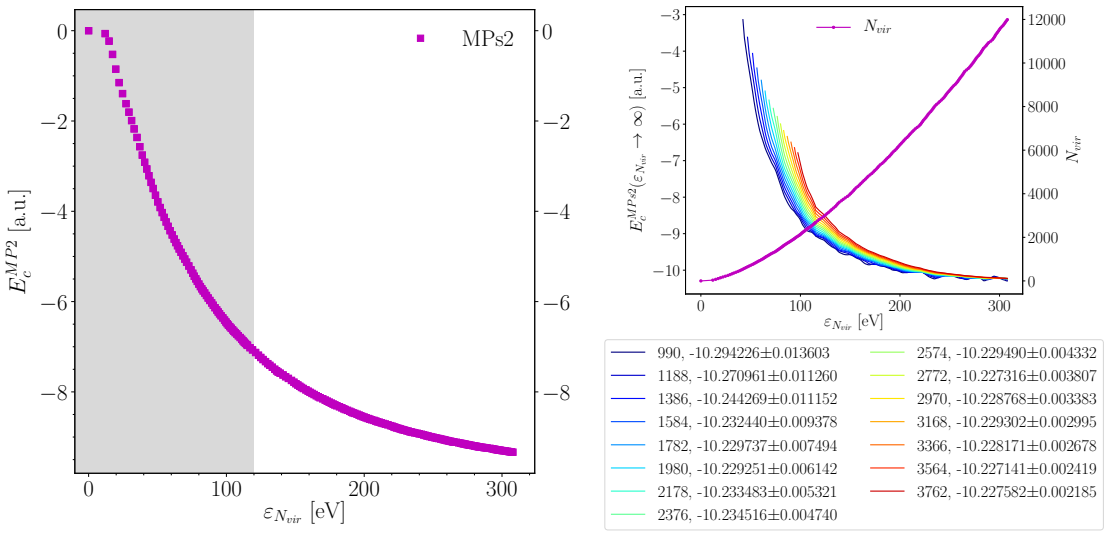


Figure B12: Left: Stochastic MPs2 correlation energy of the solvated hydronium ion as a function of the highest eigenvalue truncating the sum for $\varepsilon_c^{\text{gap}} = 120$ eV. Right: Extrapolated MPs2 energy at $\varepsilon_c^{\text{gap}} = 120$ eV of the solvated hydronium ion for a window range ending at N_{vir}, as a function of the corresponding eigenvalue $\varepsilon_{N_{\text{vir}}}$. Legend shows numbers of orbitals in fitting windows and extrapolated values in a.u. at $N_{\text{vir}}^{\text{max}} = 11990$ with asymptotic standard deviation. Respective number of virtual orbitals is shown on the secondary axis. Windows containing at least 1782 orbitals are taken for calculating MPs2 values reported in Table B2.

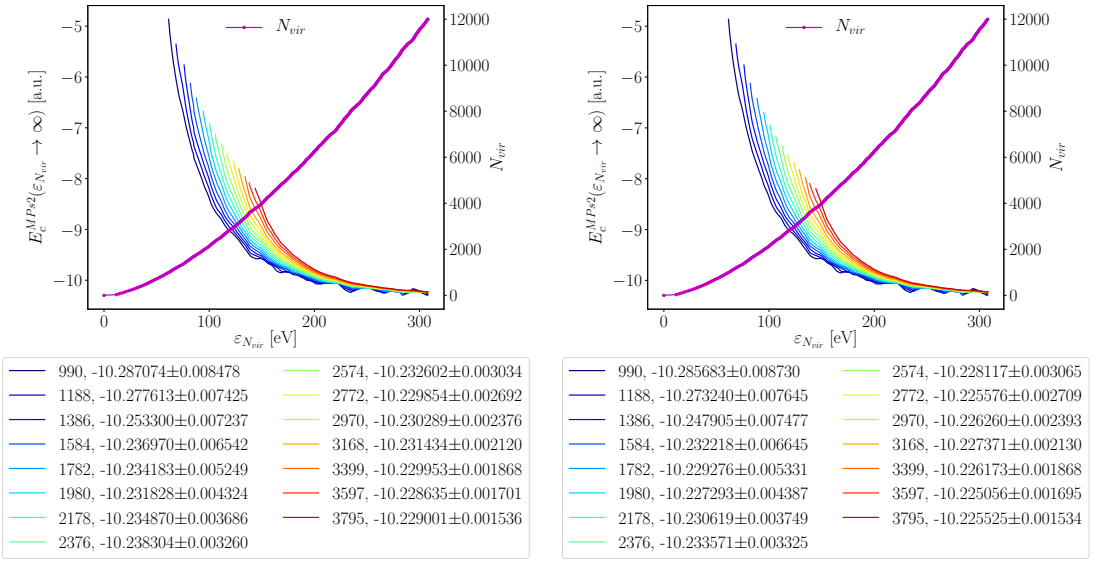


Figure B13: Left: Extrapolated MPs2 energy at $\varepsilon_c^{\text{gap}} = 90$ eV of the solvated hydronium ion for a window range ending at N_{vir}, as a function of the corresponding eigenvalue $\varepsilon_{N_{\text{vir}}}$. Right: Same for $\varepsilon_c^{\text{gap}} = 60$ eV. Legends show numbers of orbitals in fitting windows and extrapolated values in a.u. at $N_{\text{vir}}^{\text{max}} = 11990$ with asymptotic standard deviation. Respective number of virtual orbitals is shown on the secondary axis. Windows containing at least 1782 orbitals are taken for calculating MPs2 values reported in Table B2.

C Appendix of chapter 7: Minnesota functionals on liquid water

Simulation details

Table C1: Simulation details for the different density functionals studied. The BLYP GGA functional is shown for comparison.

N_{mol} is the number of heavy water molecules (D_2O), respectively light water (H_2O), simulated with meta-GGAs and hybrid meta-GGAs. CP stands for Car-Parrinello dynamics while ML-MTS means Born-Oppenheimer (BO) dynamics accelerated with a machine-learning enhanced multiple time step scheme.

t_{traj} [ps] is the simulation length of the production phase. n is the ratio between inner and outer time steps when the MTS scheme is used. $\Delta t = \delta t$ corresponds to the time step for CP dynamics, while for the ML-MTS scheme $\Delta t = n \cdot \delta t$ corresponds to the outer (physical) time step.

$\bar{t}^{\text{outer/inner}}$ [min] are the averaged elapsed times taken per outer/inner time step. We also report the ML-MTS speedup against standard BO dynamics from $\bar{t}^{\text{outer}}$ and $\bar{t}^{\text{inner}}$. $\bar{t}_{\text{sim}}$ [days/ps] is the running time in order to get 1 ps of trajectory. Timings are reported for a full distribution of MPI tasks over 16 (13) Intel Xeon E5-2690 v3 @ 2.60GHz nodes with 12 cores each for respectively the meta-GGAs (hybrid meta-GGAs).

Finally, $(dE/dt)_{\text{max}}$ [a.u./ps | %] represents the maximum energy variation per time observed along each trajectory, in absolute and relative value compared to the average energy of the system.

Functional	N_{mol}	Dynam.	t_{traj}	n	Δt [a.u. | fs]	$\bar{t}^{\text{outer}}$	$\bar{t}^{\text{inner}}$	Speed.	$\bar{t}_{\text{sim}}$	$(dE/dt)_{\text{max}}$
Meta-GGA										
M06-L	64	CP	10.0	1	2.0 | 0.048	-	0.030	-	0.43	0.000015 | $1 \cdot 10^{-6}$
revM06-L	64	CP	10.0	1	3.5 | 0.085	-	0.035	-	0.28	0.000004 | $4 \cdot 10^{-7}$
M11-L	64	CP	10.2	1	3.5 | 0.085	-	0.037	-	0.30	0.000919 | $8 \cdot 10^{-5}$
MN12-L	64	CP	10.2	1	3.5 | 0.085	-	0.031	-	0.26	0.000003 | $3 \cdot 10^{-7}$
MN15-L	64	CP	10.2	1	3.0 | 0.073	-	0.027	-	0.26	0.000009 | $8 \cdot 10^{-7}$
Hybrid meta-GGA										
M06	32	ML-MTS	6.0	6	90.0 | 2.177	94.74	0.08	5.97	30	0.004805 | $9 \cdot 10^{-4}$
M06-HF	32	ML-MTS	6.0	6	90.0 | 2.177	42.91	0.11	5.91	14	0.003075 | $6 \cdot 10^{-4}$
M06-2X	32	ML-MTS	7.3	10	150.0 | 3.628	21.60	0.10	9.57	4	0.002696 | $5 \cdot 10^{-4}$
M08-HX	32	ML-MTS	8.7	6	90.0 | 2.177	16.94	0.09	5.81	6	0.000621 | $1 \cdot 10^{-4}$
M08-SO	32	ML-MTS	6.8	10	150.0 | 3.628	21.43	0.10	9.56	4	0.005239 | $9 \cdot 10^{-4}$
M11	32	ML-MTS	6.0	10	150.0 | 3.628	83.02	0.14	9.83	16	0.003982 | $7 \cdot 10^{-4}$
MN12-SX	32	ML-MTS	6.0	6	90.0 | 2.177	18.59	0.13	5.77	6	0.000833 | $2 \cdot 10^{-4}$
MN15	32	ML-MTS	6.0	10	150.0 | 3.628	36.94	0.10	9.73	7	0.007592 | $1 \cdot 10^{-3}$
GGA										
BLYP	32	BO	20.0	1	15.0 | 0.363	-	0.029	-	0.06	0.003504 | $6 \cdot 10^{-4}$

Structural properties

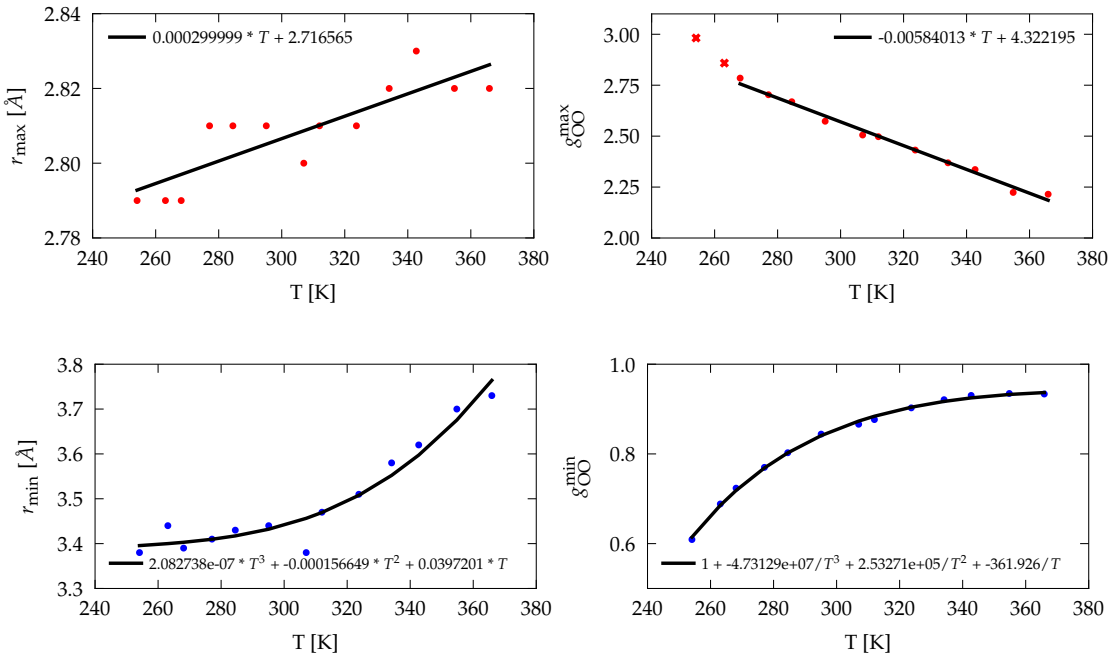


Figure C1: Position r_{max} and height g_{OO}^{max} of the first peak of the g_{OO} distribution at different temperatures extracted from X-ray measurements,[429] as well as their first minimum analogues (r_{min}, g_{OO}^{min}). Shown are the fitting curves used to rescale the simulated data to a common 298 K temperature, assuming a temperature dependency of DFT functionals similar to the experiment.

Table C2: Structure of liquid water. Position [Å] and height of the first maximum (r_{max}, $g_{\mathrm{OO}}^{\mathrm{max}}$) and first minimum ($r_{\mathrm{min}}$, $g_{\mathrm{OO}}^{\mathrm{min}}$) of the oxygen-oxygen radial distribution function as obtained from MD or MC simulations with various DFT functionals at temperature T_{avg} [K]. Their normalized analogues (r_{max}^{*}, $\mathrm{g}_{\mathrm{OO}}^{\mathrm{max}*}$) and ($r_{\mathrm{min}}^{*}$, $\mathrm{g}_{\mathrm{OO}}^{\mathrm{min}*}$) rescaled to 298 K were calculated from the experimental fits of Figure C1. Corresponding references are provided alongside the functional names.

		$T = T_{\mathrm{avg}}$				$T = 298$ K			
Functional	T_{avg}	r_{max}	$\mathrm{g}_{\mathrm{OO}}^{\mathrm{max}}$	r_{min}	$\mathrm{g}_{\mathrm{OO}}^{\mathrm{min}}$	r_{max}^{*}	$\mathrm{g}_{\mathrm{OO}}^{\mathrm{max}*}$	r_{min}^{*}	$\mathrm{g}_{\mathrm{OO}}^{\mathrm{min}*}$
GGA									
BLYP[183]	319	2.77	2.86	3.31	0.66	2.76	2.98	3.26	0.61
BLYP-DCACP[183]	308	2.79	2.72	3.36	0.85	2.79	2.78	3.34	0.82
BLYP-D3[144]	295	2.78	2.78	3.51	0.92	2.78	2.76	3.52	0.93
PBE[183]	314	2.72	3.19	3.27	0.43	2.72	3.28	3.23	0.39
PBE-DCACP[183]	323	2.71	3.27	3.28	0.40	2.70	3.42	3.21	0.35
PBE-D3[144]	295	2.73	3.07	3.25	0.69	2.73	3.05	3.26	0.70
revPBE[183]	323	2.80	2.38	3.34	0.90	2.79	2.53	3.27	0.85
revPBE-DCACP[183]	331	2.74	2.94	3.35	0.76	2.73	3.13	3.25	0.70
revPBE-D3[467]	298	2.81	2.59	3.52	0.89	2.81	2.59	3.52	0.89
rVV10[144]	295	2.73	3.22	3.32	0.79	2.73	3.20	3.33	0.80
optB88-vdW[144]	295	2.74	2.94	3.34	0.80	2.74	2.92	3.35	0.81
Meta-GGA									
M06-L	291	2.85	2.36	4.50	0.92	2.85	2.32	4.51	0.94
revM06-L	311	3.09	2.37	4.58	0.72	3.09	2.45	4.55	0.69
M11-L	286	2.89	2.11	4.59	0.86	2.89	2.04	4.61	0.90
MN12-L	296	3.13	3.20	4.31	0.45	3.13	3.19	4.31	0.46
MN15-L	283	3.37	2.70	4.61	0.43	3.37	2.61	4.63	0.48
SCAN[414]	300	2.76	3.24	3.31	0.72	2.76	3.25	3.31	0.71
SCAN+rVV10[191]	300	2.74	3.20	3.32	0.65	2.74	3.21	3.32	0.64
TPSS[465]	350	2.71	3.40	3.29	0.33	2.69	3.70	3.08	0.25
B97M-rV[457]	300	2.83	2.69	3.61	0.91	2.83	2.70	3.61	0.90
Hybrid									
B3LYP[465]	350	2.79	2.48	3.40	0.81	2.77	2.78	3.19	0.73
PBE0[483]	300	2.71	2.96	3.30	0.53	2.71	2.97	3.30	0.52
PBE0-TS-vdW(SC)[483]	300	2.72	2.76	3.31	0.70	2.72	2.77	3.31	0.69
PBE0-D3[144]	295	2.74	2.88	3.29	0.79	2.74	2.86	3.30	0.80
revPBE0-D3[457]	300	2.80	2.57	3.47	0.89	2.80	2.58	3.47	0.88
Hybrid meta-GGA									
M06	312	2.85	2.24	4.70	0.91	2.85	2.32	4.67	0.88
M06-HF	329	2.65	2.72	3.22	0.63	2.64	2.90	3.13	0.57
M06-2X	299	2.81	2.89	3.74	0.85	2.81	2.90	3.74	0.85
M08-HX	298	2.82	2.96	4.02	0.81	2.82	2.96	4.02	0.81
M08-SO	316	2.85	3.01	4.10	0.91	2.84	3.12	4.06	0.87
M11	326	2.85	2.73	3.91	0.96	2.84	2.89	3.83	0.90
MN12-SX	292	2.93	2.91	3.86	0.86	2.93	2.87	3.87	0.88
MN15	316	2.85	2.30	4.47	0.87	2.84	2.41	4.43	0.83
M06-2X-D3[144]	295	2.78	3.00	3.45	0.78	2.78	2.98	3.46	0.79
SCAN0/ML[485]	300	2.76	3.04	3.30	0.70	2.76	3.05	3.30	0.69
Post-HF, double-hybrid									
RPA[144]	295	2.78	2.93	3.41	0.78	2.78	2.91	3.42	0.79
RPA/ML[180]	300	2.79	2.89	3.41	0.83	2.79	2.90	3.41	0.82
MP2[144]	295	2.76	3.05	3.32	0.72	2.76	3.03	3.33	0.73
PWPB95-D3[144]	295	2.80	2.80	3.60	0.86	2.80	2.78	3.61	0.87
Experimental									
X-ray 2014[429,509]	298	2.80	2.55	3.41	0.85	2.80	2.55	3.41	0.85

Appendix C. Appendix of chapter 7: Minnesota functionals on liquid water

Table C3: Structure of liquid water. Coordination number $\bar{n}_{OO}$ calculated by integrating $g_{OO}(r)$ up to its first minimum. n_{OO} is the coordination number calculated by the same integration up to the first minimum of the radial distribution $4\pi r^2 g_{OO}(r)$ (eq 7.1). Average number h of hydrogen bonds per water molecule and estimated equilibrium density ρ_{eq} [g/cm^3] relative to the experimental one ρ_{exp}[504] at same temperature. Results were obtained from MD or MC simulations with various DFT functionals at temperature T_{avg} [K]. Corresponding references are provided alongside the functional names.

Functional	T_{avg}	$\bar{n}_{OO}$	n_{OO}	h	ρ_{eq}	ρ_{eq}/ρ_{exp}	System
			GGA				
BLYP[183]	319	4.2	4.0	3.44	1.010	0.92	D_2O
BLYP-DCACP[183]	308	4.5	4.2	3.43	1.135	1.03	D_2O
BLYP-D3[144]	295	5.6	5.1	*3.66	1.066	1.07	H_2O
PBE[183]	314	4.0	4.0	3.58	1.056	0.96	D_2O
PBE-DCACP[183]	323	4.1	4.0	3.63	1.063	0.97	D_2O
PBE-D3[144]	295	4.3	4.0	*3.64	1.053	1.06	H_2O
revPBE[183]	323	4.2	3.8	3.20	0.931	0.85	D_2O
revPBE-DCACP[183]	331	4.7	4.3	3.59	1.114	1.02	D_2O
revPBE-D3[467]	298	5.6	4.6	*3.63	0.97	0.97	H_2O
rVV10[144]	295	4.6	4.2	*3.80	1.078	1.08	H_2O
optB88-vdW[144]	295	4.7	4.4	*3.84	1.081	1.08	H_2O
			Meta-GGA				
M06-L	291	12.2	4.9	3.48	1.136	1.03	D_2O
revM06-L	311	12.8	9.9	2.90	1.171	1.06	D_2O
M11-L	286	12.9	6.9	3.22	1.171	1.06	D_2O
MN12-L	296	11.6	11.2	3.24	1.174	1.06	D_2O
MN15-L	283	13.4	13.0	1.93	1.280	1.16	D_2O
SCAN[415]	330	-	-	3.61	1.050	0.96	D_2O
SCAN[414]	300	4.7	4.4	-	-	-	H_2O
SCAN+rVV10[191]	300	4.6	4.4	*3.80	1.16	1.16	H_2O
TPSS[465]	350	-	-	3.82	-	-	D_2O
B97M-rV[457,467]	298	[457]5.8	[457]4.8	[457]*3.70	[467]1.12	1.12	[457,467]H_2O
			Hybrid				
B3LYP[465]	350	4.4	4.0	3.67	-	-	D_2O
PBE0[144,483]	300	[483]4.1	[483]3.9	[483]3.71	[144]0.832	0.83	[483]D/[144]H_2O
PBE0-TS-vdW(SC)[483]	300	4.2	4.1	3.62	-	-	D_2O
PBE0-D3[144]	295	4.4	4.1	*3.68	1.053	1.06	H_2O
revPBE0-D3[457]	300	5.3	4.5	*3.80	-	-	H_2O
			Hybrid meta-GGA				
M06	312	13.7	4.7	3.21	1.031	1.04	H_2O
M06-HF	329	3.9	3.9	3.55	1.051	1.07	H_2O
M06-2X	299	6.6	5.5	3.70	1.043	1.05	H_2O
M08-HX	298	8.7	7.6	3.59	1.035	1.04	H_2O
M08-SO	316	9.0	5.0	3.70	1.033	1.04	H_2O
M11	326	7.6	4.4	3.46	1.074	1.09	H_2O
MN12-SX	292	7.8	6.8	3.36	1.025	1.03	H_2O
MN15	316	11.9	4.8	3.26	1.073	1.08	H_2O
M06-2X-D3[144]	295	5.1	4.7	*3.81	1.004	1.01	H_2O
SCAN0/ML[485]	300	4.5	4.2	3.71	1.032	1.04	H_2O
			Post-HF, double-hybrid				
RPA[144]	295	4.7	4.2	*3.77	0.994	0.996	H_2O
MP2[144]	295	4.7	4.3	*3.81	1.020	1.022	H_2O
PWPB95-D3[144]	295	5.8	4.7	*3.62	1.002	1.004	H_2O
			Experimental				
X-ray 2014[429]	285	4.8	4.5		0.99952	1.00	H_2O
X-ray 2014[429,509]/Neutron 2013[432]	298	[429,509]4.6	[429,509]4.3	[432]*3.80	0.99709	1.00	H_2O
X-ray 2014[429]	307	4.6	4.3		0.99442	1.00	H_2O
X-ray 2014[429]	324	5.2	4.5		0.98765	1.00	H_2O

*estimated from the integration of the second peak of the oxygen-hydrogen radial distribution function g_{OH}, in the same way as n_{OO}.

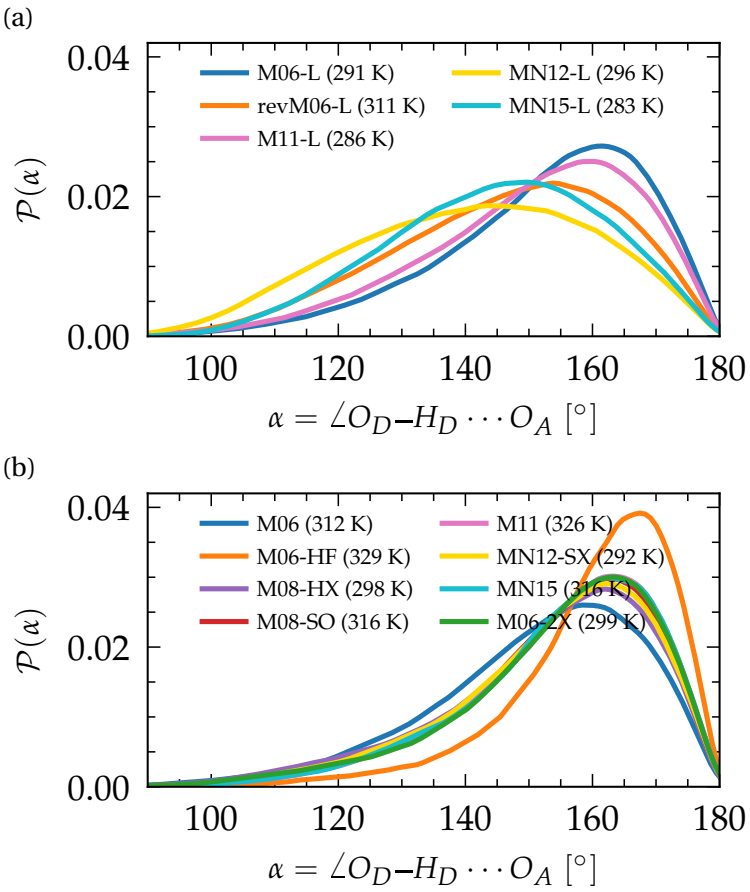


Figure C2: Distribution $\mathcal{P}(\alpha)$ of the H-bond donor angle α for donors in the first coordination shell. (a) Meta-GGA Minnesota functionals, (b) hybrid meta-GGA Minnesota functionals.

Dynamical properties

Table C4: Dynamics of liquid water. L [Å] is the side of the cubic simulation cell, D_L the finite-size diffusion coefficient from simulation and D_∞ its analogue rescaled to infinite size (eq 7.4). Results were obtained from MD simulations with various DFT functionals at temperature T_{avg} [K]. Corresponding references are provided alongside the functional names. D_∞^{exp} is the experimental diffusion coefficient at T_{avg}, as provided by a fractional-power law[425] fitted to experimental results.[421–423,425,523,524] D_L^{exp} is the experimental coefficient rescaled back to finite size (eq 7.4). All diffusion coefficients are in [Å^2/ps]. η [mPa·s] is the experimental shear viscosity[504] of light/heavy water used for rescaling.

Functional	T_{avg}	L	D_L	D_L^{exp}	D_∞	D_∞^{exp}	$\eta(T_{\text{avg}})$	System
GGA								
BLYP[183]	319	12.420	0.10	0.23	0.18	0.30	0.70167	D_2O
BLYP-DCACP[183]	308	12.420	0.17	0.18	0.23	0.24	0.87277	D_2O
BLYP-D3[496]	298	15.640	0.08	0.18	0.12	0.23	0.88982	H_2O
BLYP-D3[496]	328	15.640	0.20	0.35	0.29	0.43	0.50354	H_2O
PBE[183]	314	12.420	0.03	0.21	0.10	0.27	0.77176	D_2O
PBE[415]	330	13.108	0.02	0.29	0.11	0.38	0.58027	D_2O
PBE-DCACP[183]	323	12.420	0.05	0.25	0.13	0.33	0.65294	D_2O
revPBE[183]	323	12.420	0.21	0.25	0.29	0.33	0.65294	D_2O
revPBE-DCACP[183]	331	12.420	0.16	0.29	0.26	0.39	0.57100	D_2O
revPBE-D3[467]	298	12.420	0.19	0.17	0.25	0.23	0.88982	H_2O
optB88-vdW[414]	300	9.850	0.07	0.17	0.14	0.24	0.85072	H_2O
Meta-GGA								
M06-L	291	12.445	0.30	0.12	0.34	0.15	1.32310	D_2O
revM06-L	311	12.445	0.65	0.19	0.71	0.26	0.81967	D_2O
M11-L	286	12.445	0.52	0.10	0.55	0.13	1.53360	D_2O
MN12-L	296	12.445	0.11	0.13	0.15	0.18	1.15640	D_2O
MN15-L	283	12.445	0.06	0.09	0.09	0.12	1.68780	D_2O
SCAN[415]	330	12.217	0.19	0.28	0.29	0.38	0.58027	D_2O
SCAN[496]	328	12.660	0.14	0.32	0.25	0.43	0.50354	H_2O
SCAN[414]	300	9.850	0.06	0.17	*0.09	0.24	0.85072	H_2O
SCAN[496]	298	12.660	0.03	0.17	0.08	0.23	0.88982	H_2O
SCAN/ML[484]	300	11.817	0.05	0.18	0.11	0.24	0.85072	H_2O
TPSS[465]	350	9.939	0.03	0.36	0.20	0.53	0.43331	D_2O
B97M-rV[457]	300	12.420	0.21	0.18	0.27	0.24	0.85072	H_2O
Hybrid								
B3LYP[465]	350	9.939	0.30	0.36	0.47	0.53	0.43331	D_2O
PBE0[483]	300	12.400	0.07	0.15	0.12	0.20	1.04660	D_2O
PBE0-TS-vdW(SC)[483]	300	12.400	0.10	0.15	0.15	0.20	1.04660	D_2O
revPBE0-D3[457]	300	12.420	0.21	0.18	0.27	0.24	0.85072	H_2O
Hybrid meta-GGA								
M06	312	9.939	0.69	0.22	0.79	0.31	0.66506	H_2O
M06-HF	329	9.939	0.16	0.30	0.30	0.44	0.49563	H_2O
M06-2X	299	9.939	0.31	0.16	0.38	0.23	0.86991	H_2O
M08-HX	298	9.939	0.18	0.16	0.25	0.23	0.88982	H_2O
M08-SO	316	9.939	0.24	0.23	0.35	0.34	0.61743	H_2O
M11	326	9.939	0.31	0.29	0.44	0.42	0.52001	H_2O
MN12-SX	292	9.939	0.14	0.14	0.20	0.20	1.02640	H_2O
MN15	316	9.939	0.30	0.23	0.41	0.34	0.61743	H_2O
SCAN0/ML[485]	300	24.575	0.11	0.17	0.13	0.20	1.04660	D_2O
SCAN0/ML[485]	300	24.575	0.12	0.21	0.15	0.24	0.85072	H_2O
Post-HF								
RPA/ML[180]	300	11.817	0.17	0.18	0.23	0.24	0.85072	H_2O
MP2[144]	295	12.335	0.07	0.16	0.12	0.21	0.95417	H_2O
CCSD(T)/ML PIMD[509]	298	15.660	0.20	0.18	0.24	0.23	0.88982	H_2O

*rescaled to infinite size with the actual viscosity obtained with the SCAN functional.[414]

Table C5: Dynamics of liquid water. First-order τ_1 and second-order τ_2 orientational relaxation times [ps] calculated from the orientational auto-correlation function (eq 7.5), between respectively the geometric dipoles μ, OH, and HH vectors. Results were obtained from MD simulations with various DFT functionals at temperature T_{avg} [K]. Corresponding references are provided alongside the functional names. Note that $\tau_{1,2}$ are highly sensitive to statistical sampling and require trajectories that are sufficiently long (approximately three times higher than their value) to be accurately converged, in addition to a sufficient equilibration phase at the beginning of the NVE sampling. Additionally, the fitting or integration methods used for their calculation vary between studies, and experimental results exhibit non-negligible deviations. Nevertheless, we provide these values as a qualitative comparison.

Functional	T_{avg}	τ_1^{μ}	τ_2^{μ}	τ_1^{OH}	τ_2^{OH}	τ_1^{HH}	τ_2^{HH}	System
GGA								
BLYP[183]	319	[a]7.5	[a]3.0	-	-	-	-	D_2O
BLYP-DCACP[183]	308	[a]3.6	[a]1.7	-	-	-	-	D_2O
PBE[183]	314	[a]36.9	[a]15.6	-	-	-	-	D_2O
PBE[415]	330	-	-	-	[b]7.1	-	-	D_2O
PBE-DCACP[183]	323	[a]32.7	[a]10.0	-	-	-	-	D_2O
revPBE[183]	323	[a]2.7	[a]1.3	-	-	-	-	D_2O
revPBE-DCACP[183]	331	[a]5.4	[a]2.1	-	-	-	-	D_2O
revPBE-D3[516]	300	[b]4.6	[b]1.7	[b]5.4	[b]2.2	[b]5.9	[b]2.6	H_2O
Meta-GGA								
M06-L	291	1.8	0.8	2.3	1.0	2.7	1.3	D_2O
revM06-L	311	0.4	0.2	0.5	0.3	0.6	0.3	D_2O
M11-L	286	1.0	0.5	1.3	0.6	1.4	0.7	D_2O
MN12-L	296	0.4	0.2	0.5	0.2	0.5	0.3	D_2O
MN15-L	283	0.4	0.1	0.4	0.2	0.4	0.2	D_2O
SCAN[415]	330	-	-	-	[b]2.9	-	-	D_2O
SCAN/ML[484]	300	-	[b]12.9	-	[b]15.7	-	[b]21.5	H_2O
Hybrid								
revPBE0-D3[516]	300	[b]3.4	[b]1.4	[b]4.3	[b]1.7	[b]4.8	[b]2.0	H_2O
Hybrid meta-GGA								
M06	312	1.2	0.5	1.2	0.6	1.2	0.6	H_2O
M06-HF	329	4.7	2.5	6.2	3.3	7.3	3.6	H_2O
M06-2X	299	2.7	1.1	3.0	1.3	3.2	1.4	H_2O
M08-HX	298	2.8	1.5	3.0	1.6	3.2	1.8	H_2O
M08-SO	316	2.8	1.3	3.0	1.4	3.2	1.6	H_2O
M11	326	3.0	1.5	3.6	1.6	3.9	1.9	H_2O
MN12-SX	292	3.7	1.6	4.4	2.0	4.9	2.9	H_2O
MN15	316	2.0	1.0	2.0	1.1	2.0	1.3	H_2O
SCAN0/ML[485]	300	-	-	-	[b]4.6	-	-	D_2O
SCAN0/ML[485]	300	-	-	-	[b]4.1	-	-	H_2O
Post-HF								
RPA/ML[180]	300	-	[b]1.7	-	[b]2.2	-	[b]2.6	H_2O
CCSD(T)/ML PIMD[509]	298	-	[b]1.3	-	[b]1.7	-	[b]2.1	H_2O
CCSD(T)/ML PIMD[509]	298	-	2.8	-	3.0	-	3.3	H_2O
Experimental								
NMR 1970[376,511]	300	4.8	-	-	-	-	-	H_2O
NMR 1971[511,525]	300	-	1.9	-	-	-	-	H_2O
NMR 1967 (for various T)[416]	300	-	2.4	-	-	-	-	H_2O
Infrared 2008[512]/2010[513]	298	-	-	-	2.5	-	2.5	H_2O
Infrared 2008[512]/2010[513]	298	-	-	-	3.0	-	-	D_2O
NMR 2001 (for various T)[426]	300	-	-	-	2.4	-	-	D_2O
NMR 1985[419]/1987[420]	298	-	-	-	1.9-2.0	-	-	H_2O
NMR 1982[424]	298	-	-	-	1.7	-	-	H_2O
NMR 1966[510]	298	-	-	-	2.6	-	-	H_2O
NMR 1976[417]	303	-	-	-	-	-	2.1	H_2O
NMR 1976[417]	303	-	-	-	-	-	2.5	D_2O

[a] $\tau_{1,2}$ were calculated from fitting the auto-correlation functions $C_{1,2}(t)$ with the exponential form $A\exp[-(t/\tau_{1,2})^{\alpha}]$.
[b] The tail of the auto-correlation function was fitted with $\exp(-t/\tau_{1,2})$ and integrated from zero to ∞ to give $\tau_{1,2}$. Others were obtained with the fit $A\exp(-t/\tau_{1,2})$ in the exponential regime after the initial subpicosecond librational decay.

D Appendix of chapter 8: Genetic algorithms for peptide structures

Crossover operator

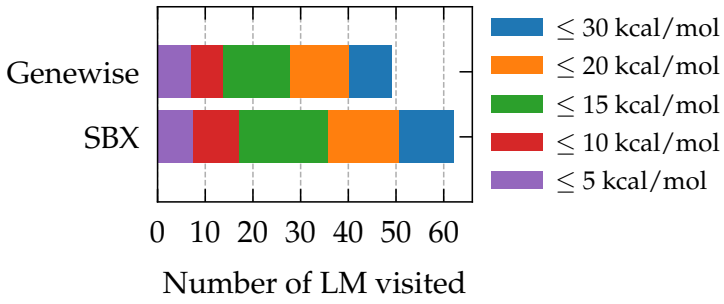
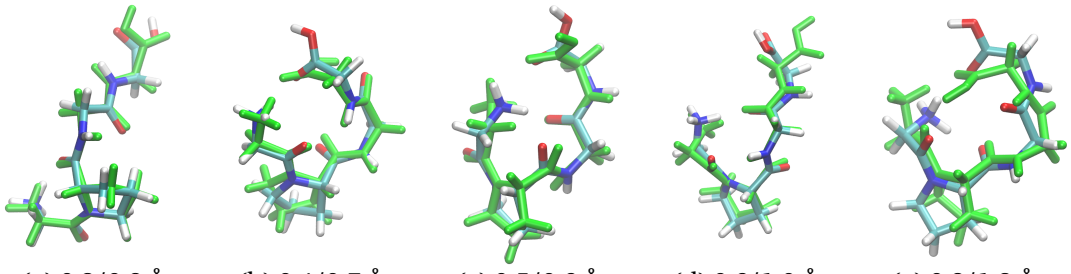


Figure D1: Number of LM visited during a GA run with genewise and SBX crossovers for trans GPGG on the GAFF PES averaged over 30 instances. Energy ranges relative to the putative GM are indicated. No mutations nor elitism were applied, other parameters correspond to Table 8.1 of the main text.

Surrogate fitness function



(a) 0.3/0.8 Å (b) 0.4/0.7 Å (c) 0.5/0.6 Å (d) 0.6/1.0 Å (e) 0.8/1.3 Å

Figure D2: Examples of GPGG LM structures at a surrogate and the B3LYP/6-31G(d,p) (in green) level with respective backbone/heavy-atom RMSD.

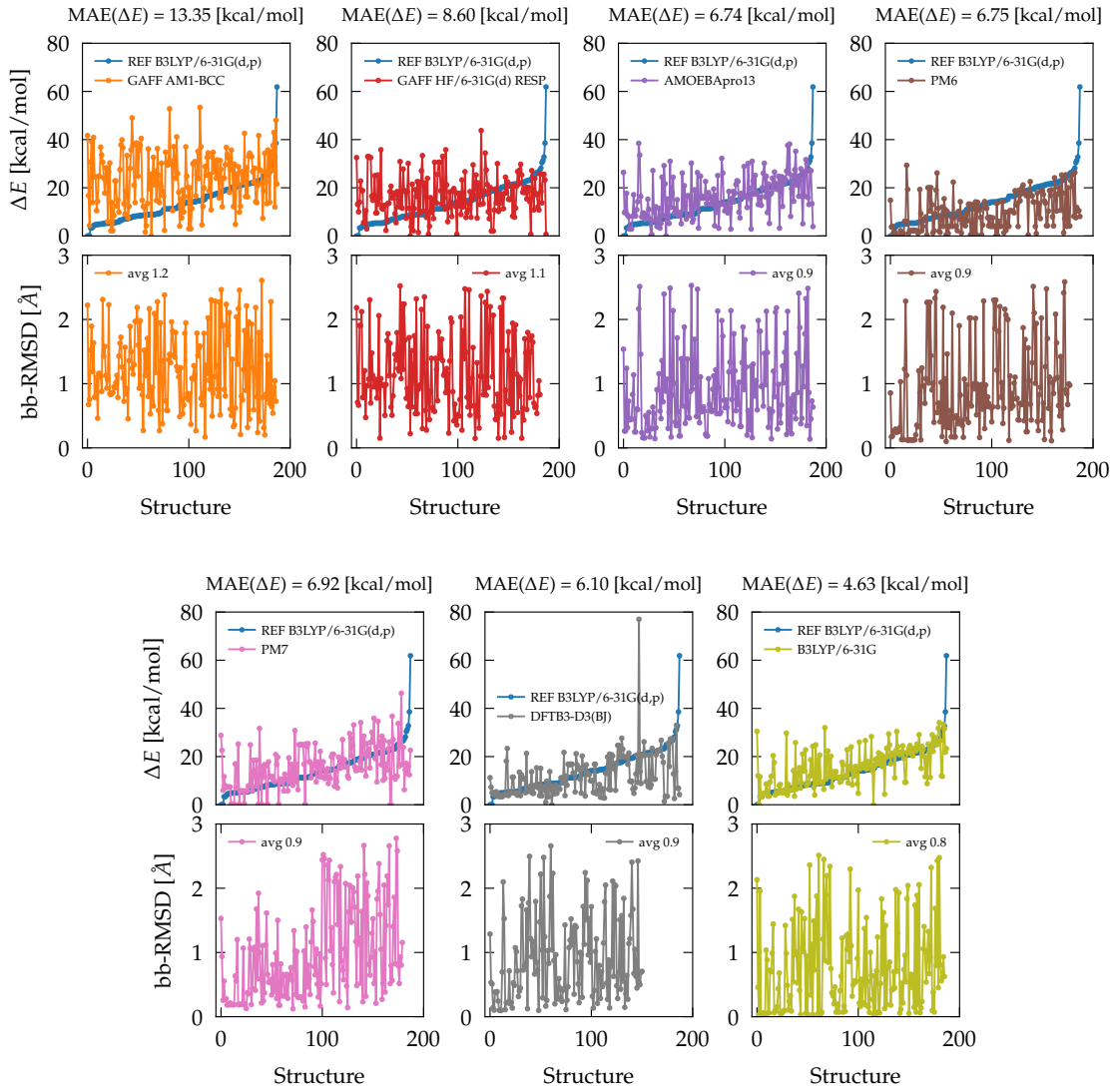


Figure D3: Relative energies with respect to the putative GM in the GPGG test set (188 representative geometries) for surrogate methods, compared to the B3LYP/6-31G(d,p) reference. Energies and corresponding backbone RMSDs stand for structures relaxed from the same initial geometry in the set (one-to-one match).

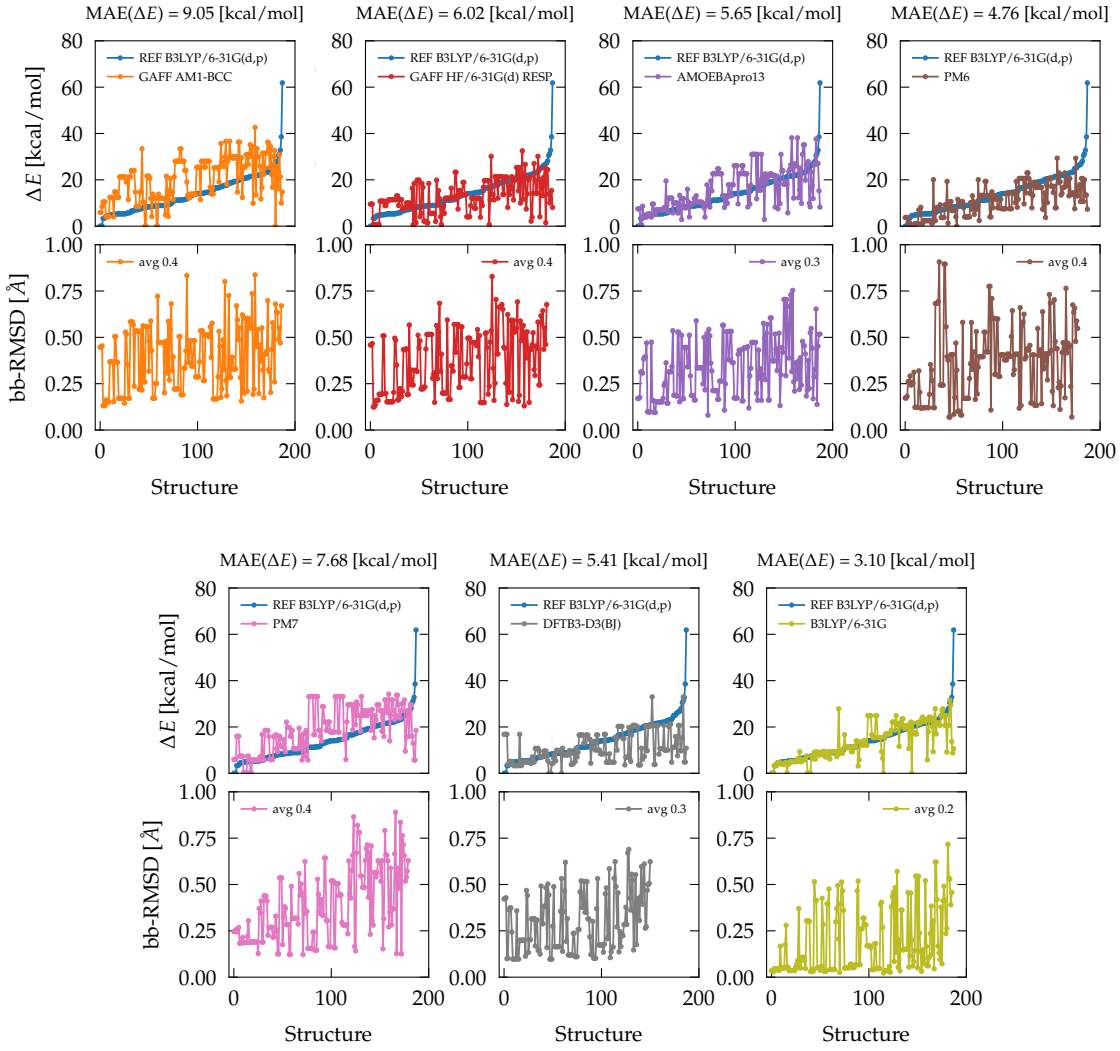


Figure D4: Relative energies with respect to the putative GM in the GPGG test set for surrogate methods, compared to the B3LYP/6-31G(d,p) reference. Energies and corresponding backbone RMSDs stand for relaxed structures that have the smallest backbone RMSD over the entire test set (match according to “closest” geometries). The better agreement of B3LYP/6-31G justifies the bb-RMSD match rather than the less faithful one-to-one match in Figure D3.

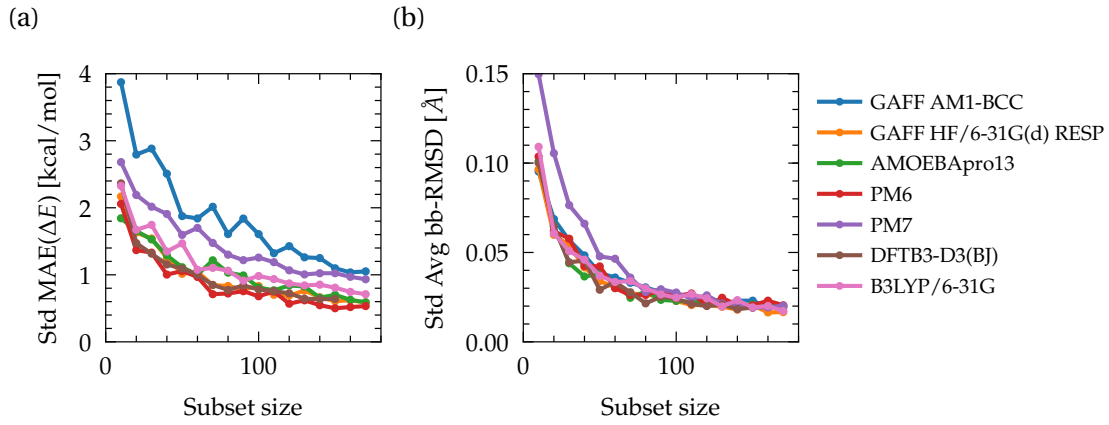


Figure D5: (a) Standard deviation of the MAE of relative energies between surrogate local minima and their bb-RMSD closest B3LYP/6-31G(d,p) counterparts. (b) Standard deviation of the average bb-RMSD between the surrogate LM and their closest B3LYP/6-31G(d,p) counterparts. Statistics were produced over $\max(S, 70)$ random subsets for each size S, extracted from the GPGG test set.

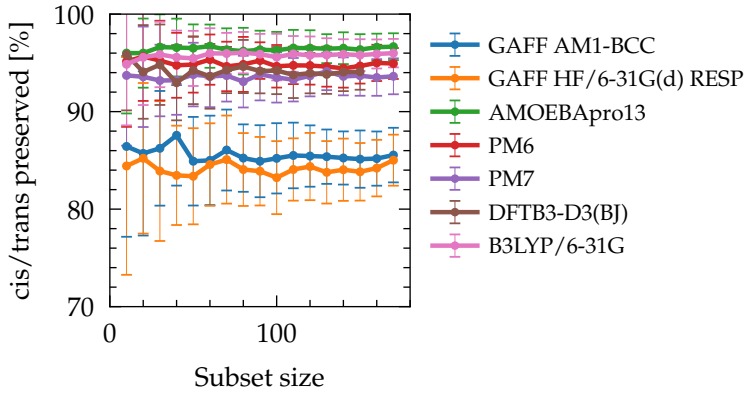


Figure D6: Average percentage of the cis/trans proline isomerizations that were preserved after local relaxation on surrogate models. Statistics were produced over $\max(S, 70)$ random subsets for each size S, extracted from the GPGG test set. Hence, it can happen that local relaxations alter the isomerization and wrongly drive the GA. This is prevented with constraints (cf. Section 8.3.3 of Chapter 8).

GPGG

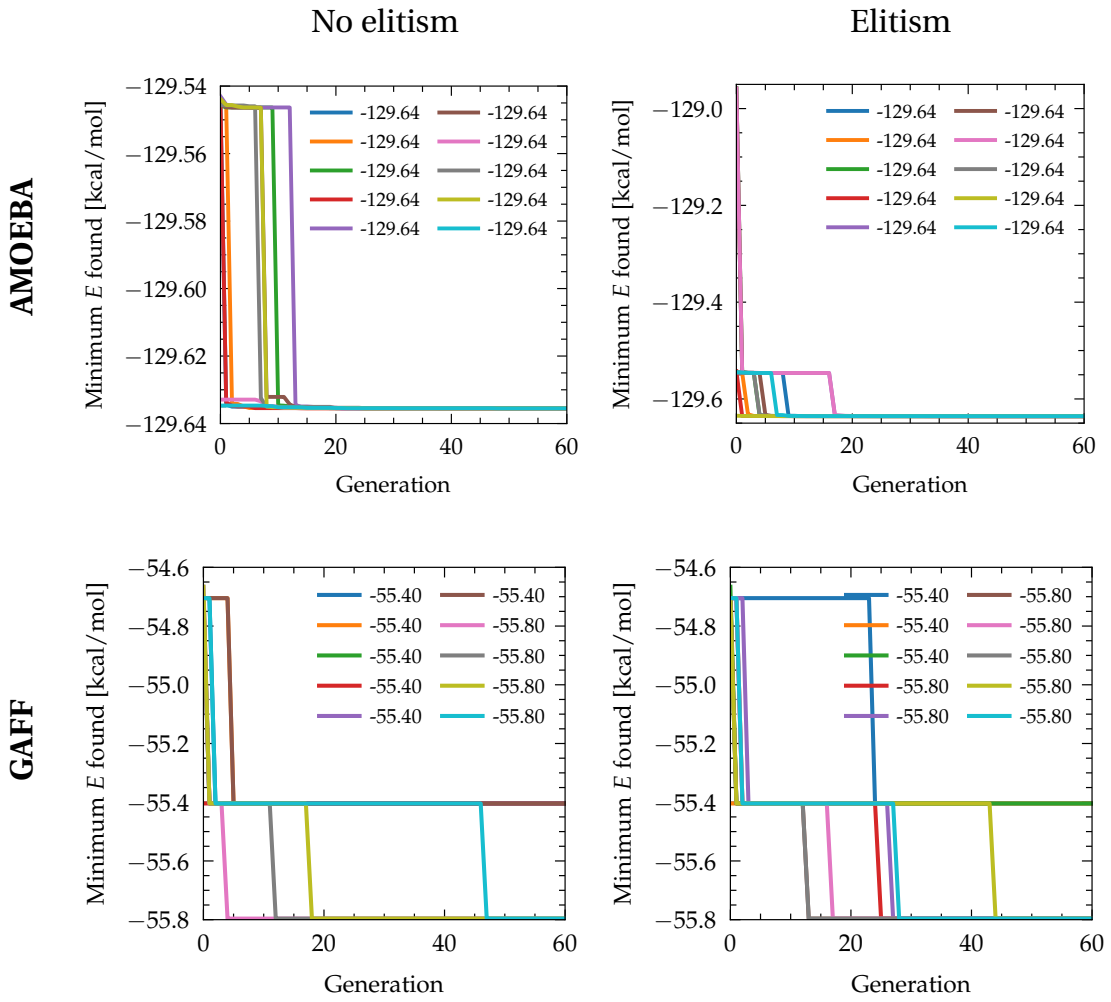


Figure D7: cis-GPGG: minimum energy found along GA runs on surrogate PES for AMOEBA (upper row) and GAFF (lower row) with and without 10% elitism. The respective final energies are reported in the legends.

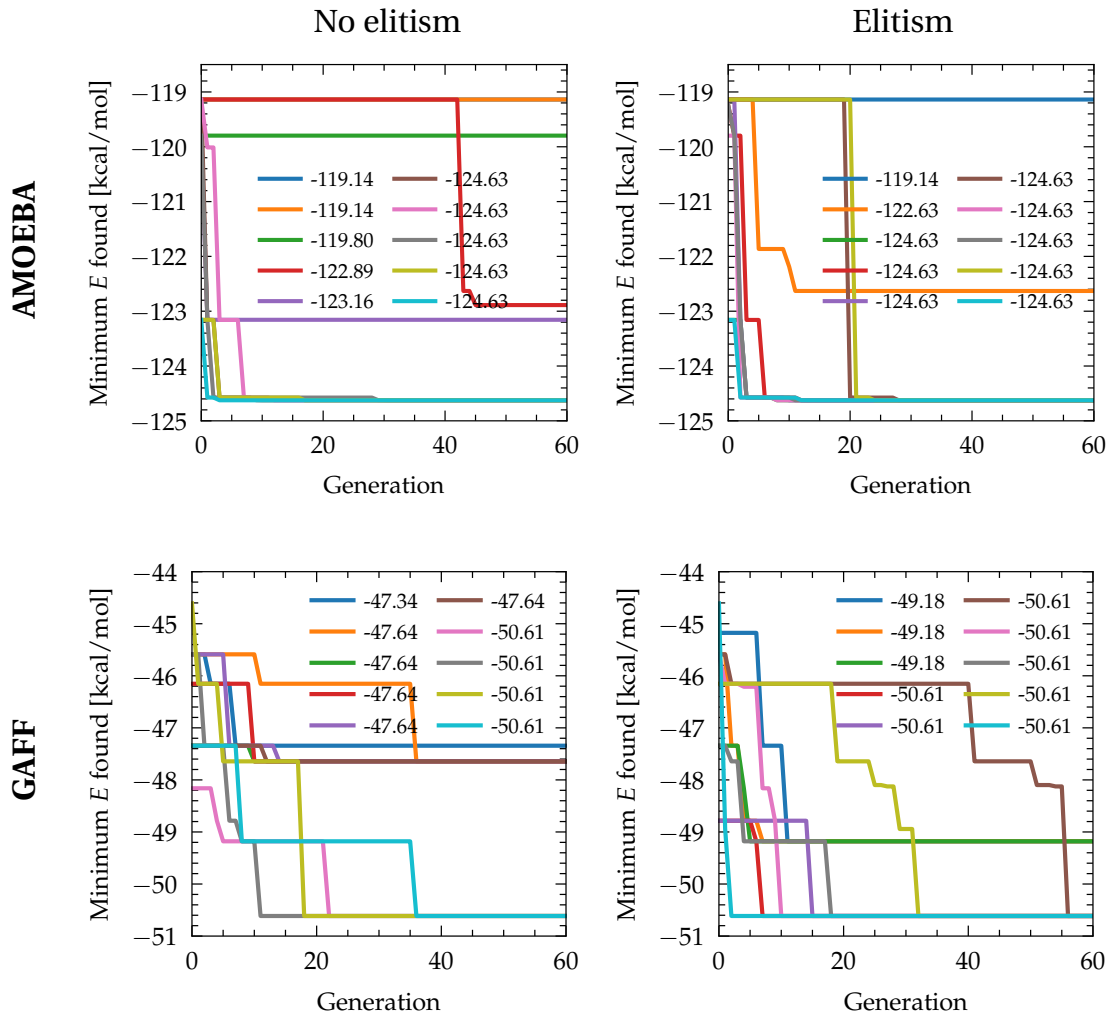


Figure D8: trans-GPGG: minimum energy found along GA runs on surrogate PES for AMOEBA (upper row) and GAFF (lower row) with and without 10% elitism. The respective final energies are reported in the legends.

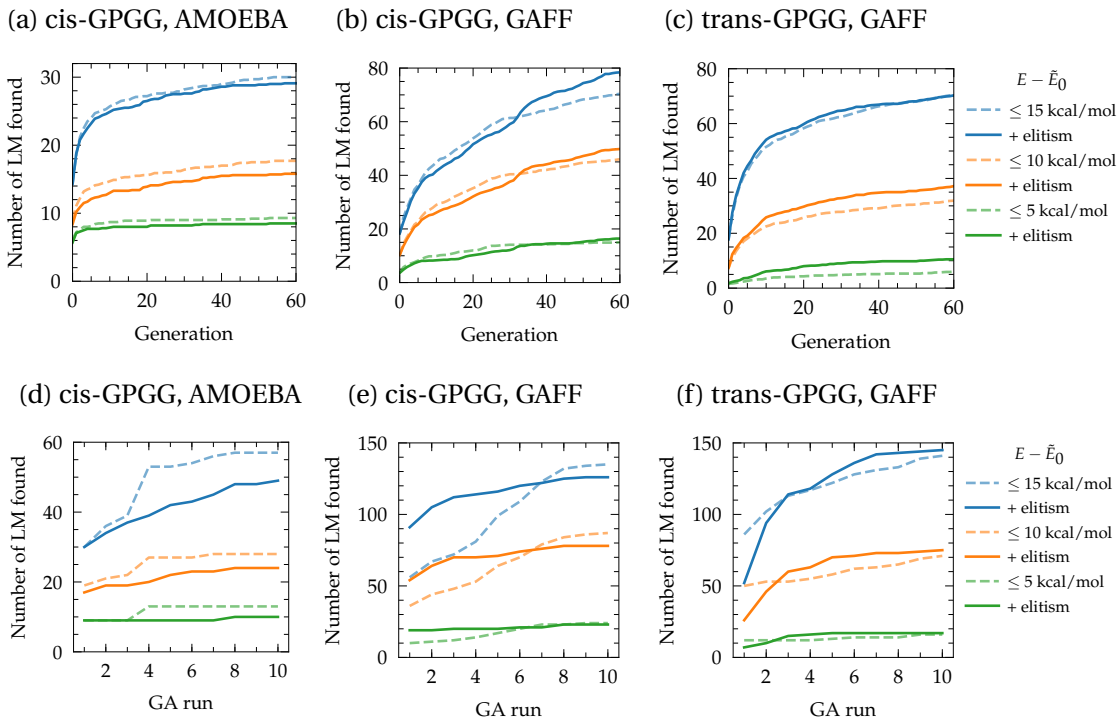


Figure D9: GPGG: number of low-lying minima found on the AMOEBA/GAFF surrogate PES within 15, 10, and 5 kcal/mol with respect to the putative GM, per GA generation averaged over 10 GA runs (upper row) and by running independent GAs (lower row). Distinct LM are taken to be at least separated by 10^{-4} kcal/mol and 0.2 heavy-atom RMSD.

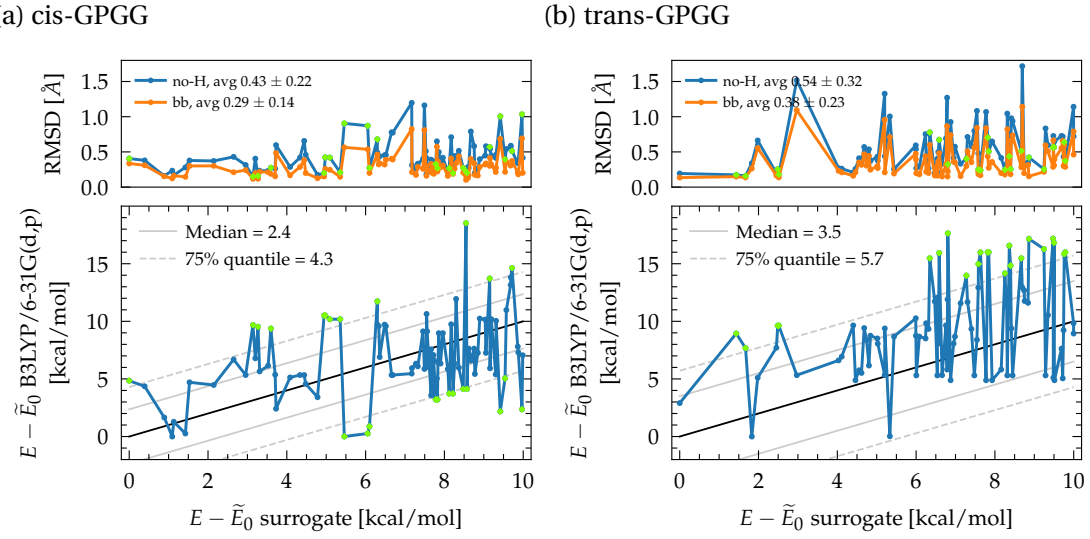


Figure D10: GPGG: predictive performance of GAFF in giving B3LYP/6-31G(d,p) LM coordinates and relative energies. The median and 75% quantile of absolute errors are indicated. The 75% quantile outliers are marked in green with their respective RMSD.

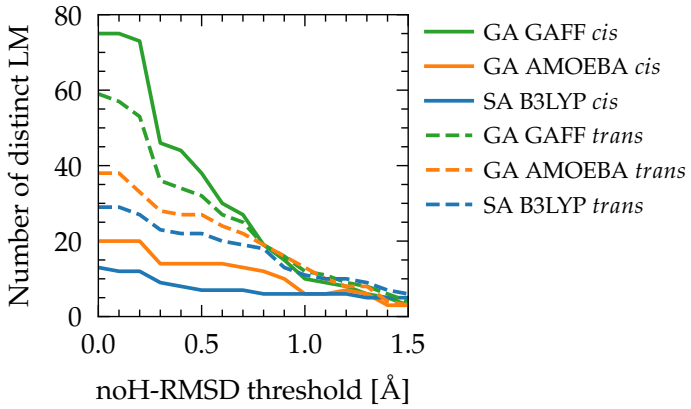


Figure D11: GPGG: number of distinct reoptimized LM at B3LYP/6-31G(d,p) in function of the minimum heavy-atom (no-H) RMSD imposed between all structures.

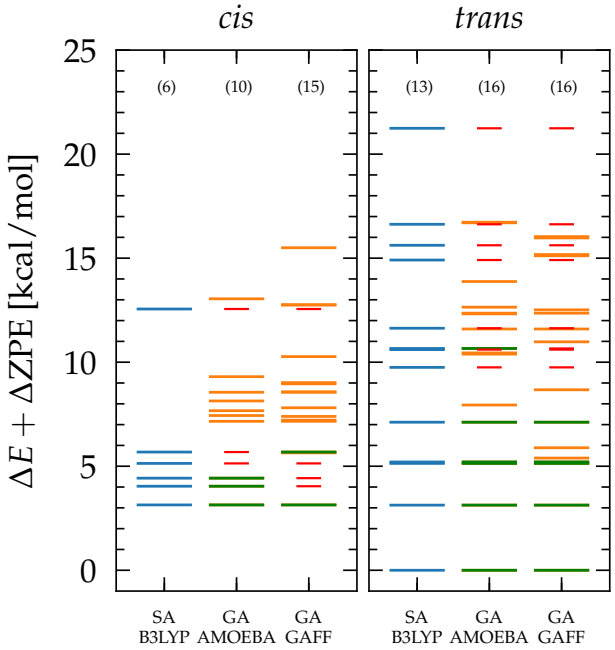


Figure D12: GPGG: zero point energy-corrected energies of distinct B3LYP/6-31G(d,p) LM from Figure 8.11 of Chapter 8, separated at least by 0.9 Å heavy-atom RMSD, which corresponds to at least one backbone dihedral noticeably different from visual inspection. The number of those is indicated in parentheses.

Gramicidin

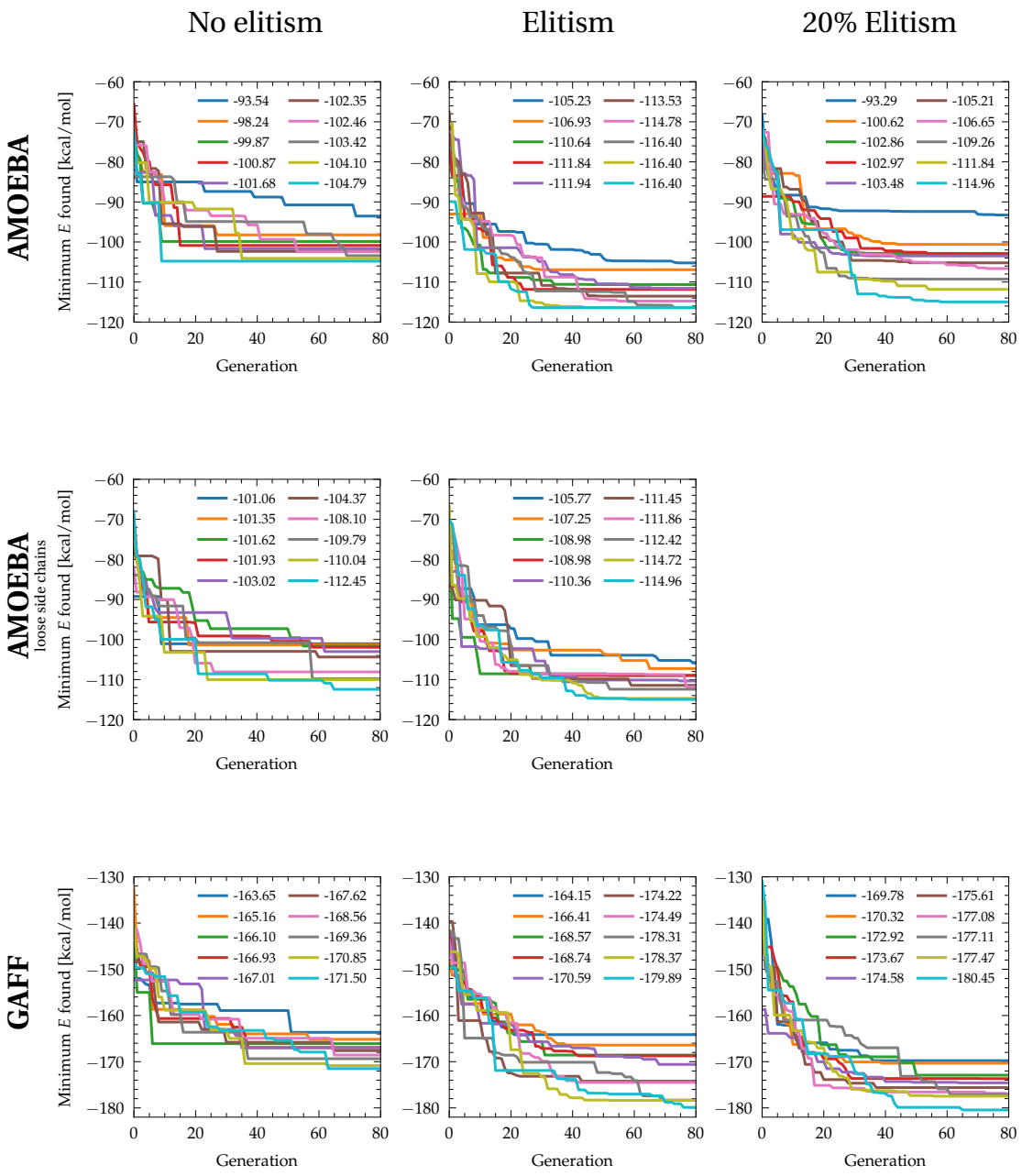
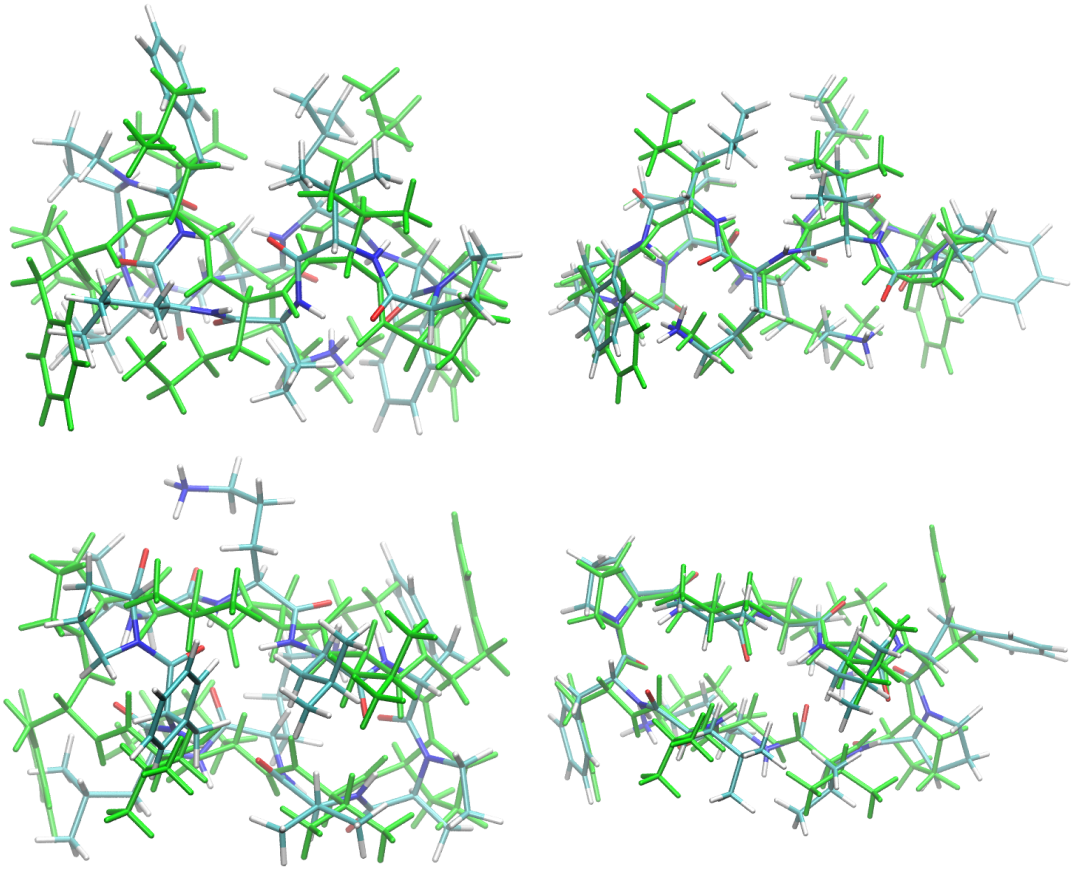


Figure D13: Gramicidin: minimum energy found along GA runs on surrogate PES for AMOEBA (upper and center row) and GAFF (lower row) without elitism, with 10% elitism and with a stronger 20% elitism. The respective final energies are reported in the legends. The AMOEBA with loose side chains corresponds to GA optimizations of backbone dihedrals only. Similarly to the prolines cycles, the side chain dihedrals are only updated after energy evaluations (with locally relaxed values). This scheme does not surpass the complete GA optimizer with elitism.



(a) GAFF, putative GM (from 20% elitism)
RMSD: 2.1/4.3 Å

(b) GAFF, closest LM (from no elitism)
RMSD: 0.5/1.5 Å
$E - \tilde{E}_0 = 26.09$ kcal/mol

Figure D14: Gramicidin: putative GM and closest LM found on the GAFF surrogate PES over 30 GA runs (10 without elitism, 10 with 10% elitism and 10 with 20% elitism). $E - \tilde{E}_0$ is the relative energy of the LM with respect to the putative GM. The DFT B3LYP/6-31G(d,p) GM is depicted in green with respective backbone/heavy-atom RMSD. For runs with 10% elitism, the closest LM has $\Delta\tilde{E} = 47.01$ kcal/mol with 0.7/2.0 Å RMSD. For runs with 20% elitism, the closest LM has $\Delta\tilde{E} = 30.39$ kcal/mol with 0.7/1.9 Å RMSD. We note that these closest LM do not finally relax to the B3LYP/6-31G(d,p) GM when reoptimized.

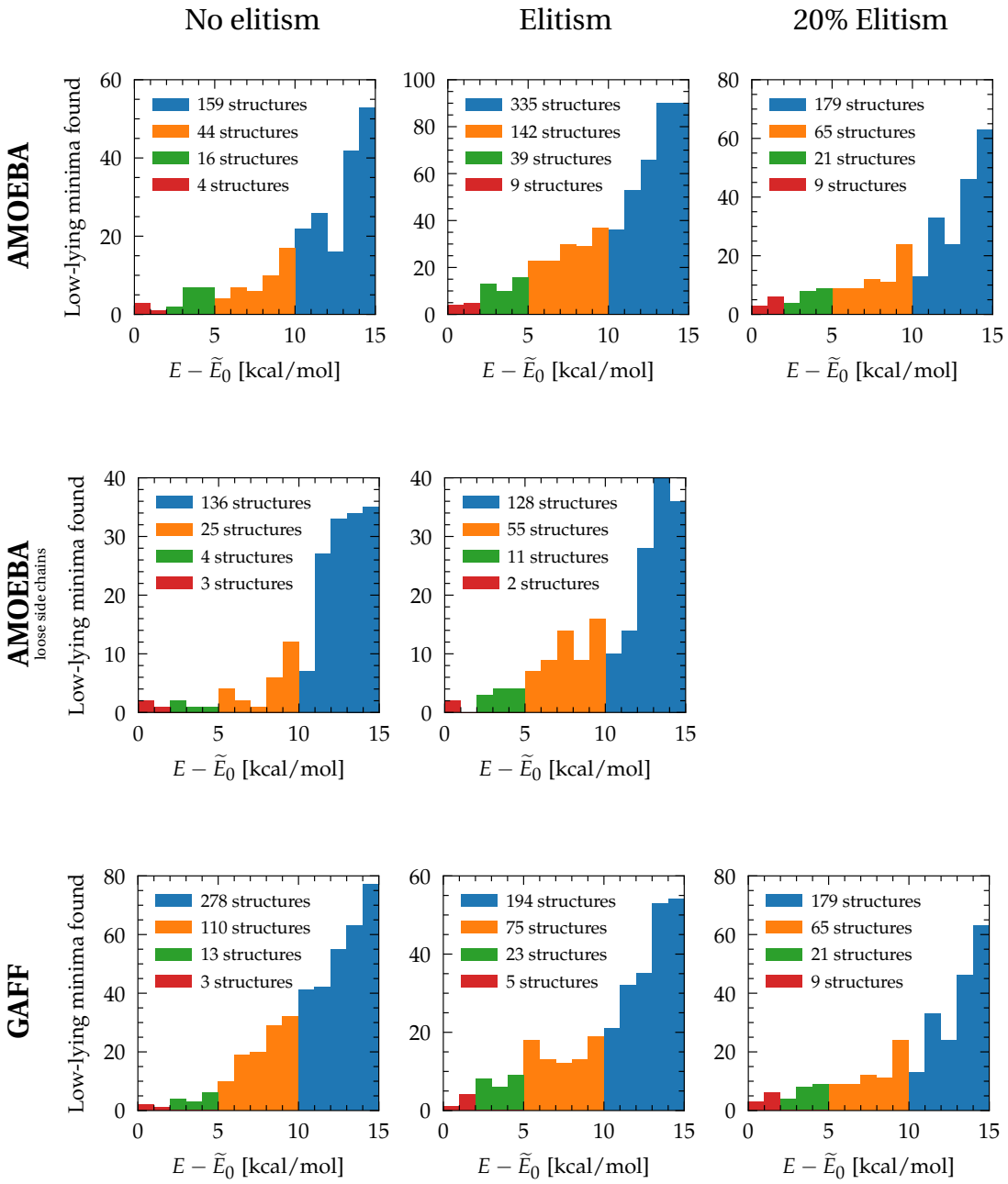


Figure D15: Gramicidin: energy distribution of low-lying LM found on surrogate PES after 10 GA runs for AMOEBA (upper row), AMOEBA with loose side chains (see Figure D13 for the description) and GAFF (lower row). $\widetilde{E}_0$ is the energy of the surrogate putative GM found within 10 runs without elitism, with 10% elitism and with a stronger 20% elitism, respectively. Distinct LM are taken to be at least separated by 10^{-4} kcal/mol and 0.75 heavy-atom RMSD.

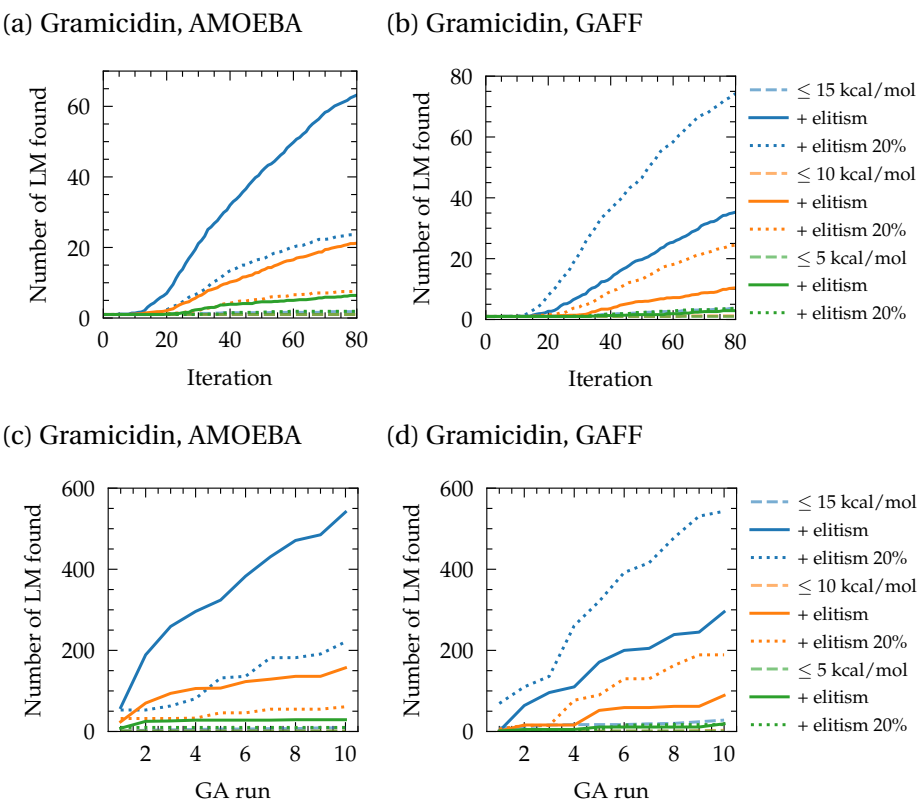


Figure D16: Gramicidin: number of low-lying minima found on the AMOEBA/GAFF surrogate PES within 15, 10, and 5 kcal/mol with respect to the putative GM, per GA generation averaged over 10 GA runs (upper row) and by running independent GAs (lower row). Distinct LM are taken to be at least separated by 10^{-4} kcal/mol and 0.75 heavy-atom RMSD.

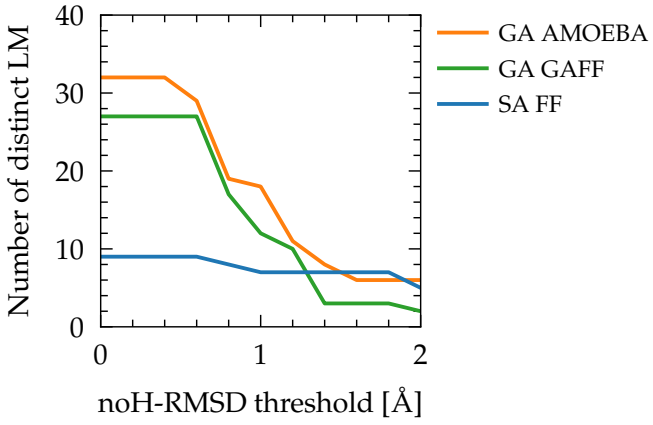


Figure D17: Gramicidin: number of distinct reoptimized LM at B3LYP/6-31G(d,p) in function of the minimum heavy-atom RMSD imposed between all structures.

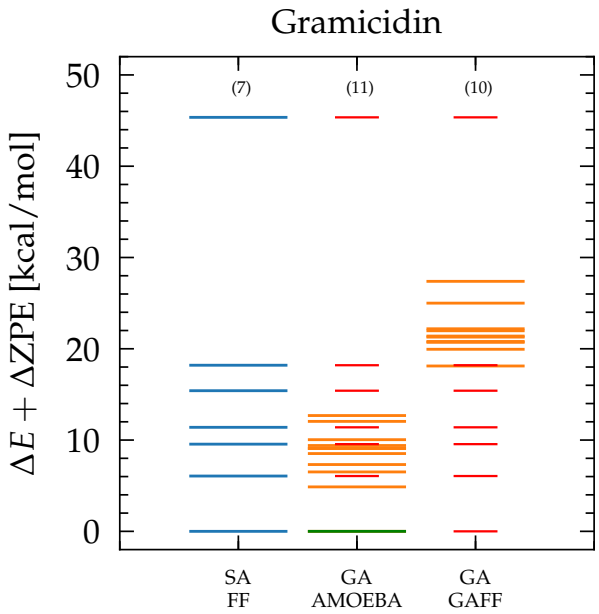


Figure D18: Gramicidin: zero point energy-corrected energies of distinct B3LYP/6-31G(d,p) LM from Figure 8.15 of Chapter 8, separated at least by 1.2 Å heavy-atom RMSD, which corresponds to at least two side chain dihedrals noticeably different from visual inspection. The number of those is indicated in parentheses.

Bibliography

(1) https://en.wikipedia.org/wiki/ENIAC, *Electronic Numerical Integrator and Computer*; Wikipedia: 2023, May.

(2) https://en.wikipedia.org/wiki/Human_Brain_Project, *Human Brain Project*; Wikipedia: 2023, May.

(3) Cramer, C. J., *Essentials of Computational Chemistry: Theories and Models*, 2nd ed.; John Wiley & Sons: Chichester, 2004.

(4) Jensen, F., *Introduction to Computational Chemistry*, 3rd ed.; John Wiley & Sons: Chichester, 2017.

(5) Leszczynski, J.; Kaczmarek-Kedziera, A.; Puzyn, T.; Papadopoulos, M. G.; Reis, H.; Shukla, M. K., *Handbook of computational chemistry*, 2nd ed.; Springer: Cham, 2017.

(6) Ohno, K.; Satoh, H., *Exploration on Quantum Chemical Potential Energy Surfaces: Towards the Discovery of New Chemistry*; Theoretical and Computational Chemistry Series; The Royal Society of Chemistry: Cambridge, 2022.

(7) Schneider, G. *Nature Reviews Drug Discovery* **2010**, *9*, 273–276.

(8) Von Lilienfeld, O. A.; Müller, K. R.; Tkatchenko, A. *Nature Reviews Chemistry* **2020**, *4*, 347–358.

(9) Frenkel, D.; Smit, B., *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: San Diego, 2001.

(10) Monticelli, L.; Tieleman, D. P. In *Biomolecular Simulations*, Monticelli, L., Salonen, E., Eds., 1st ed.; Springer: New York, 2013.

(11) Atkins, P. W.; Friedman, R. S., *Molecular Quantum Mechanics*, 5th ed.; Oxford University Press: Oxford, 2010.

(12) David Sherrill, C.; Schaefer, H. F. *Advances in Quantum Chemistry* **1999**, *34*, 143–269.

(13) Szabo, A.; Ostlund, N. S., *Modern Quantum Chemistry*, 1st ed.; Dover Publications: New York, 1996.

(14) Helgaker, T.; Jørgensen, P.; Olsen, J., *Molecular Electronic-Structure Theory*, 1st ed.; John Wiley & Sons: Chichester, 2000.

(15) Martin, R. M., *Electronic structure : basic theory and practical methods*, 1st ed.; Cambridge University Press: Cambridge, 2004.

(16) Kohanoff, J. J., *Electronic structure calculations for solids and molecules : theory and computational methods*, 1st ed.; Cambridge University Press: Cambridge, 2006.

(17) Hartree, D. R. *Mathematical Proceedings of the Cambridge Philosophical Society* **1928**, *24*, 89–110.

(18) Fock, V. *Zeitschrift für Physik* **1930**, *61*, 126–148.

(19) Slater, J. C. *Physical Review* **1928**, *32*, 339–348.

(20) Dral, P. O., *Quantum Chemistry in the Age of Machine Learning*, 1st ed.; Elsevier Science: Amsterdam, Oxford, Cambridge, 2022.

(21) Møller, C.; Plesset, M. S. *Physical Review* **1934**, *46*, 618–622.

(22) Martin, R. M.; Reining, L.; Ceperley, D. M., *Interacting Electrons: Theory and Computational Approaches*, 1st ed.; Cambridge University Press: Cambridge, 2016.

(23) Giustino, F., *Materials Modelling Using Density Functional Theory: Properties and Predictions*, 1st ed.; Oxford University Press: Oxford, 2014.

(24) Čížek, J. *Journal of Chemical Physics* **1966**, *45*, 4256–4266.

(25) Paldus, J.; Čížek, J.; Shavitt, I. *Physical Review A* **1972**, *5*, 50–67.

(26) Becca, F.; Sorella, S., *Quantum Monte Carlo approaches for correlated systems*, 1st ed.; Cambridge University Press: Cambridge, 2017.

(27) Al-Hamdani, Y. S.; Nagy, P. R.; Zen, A.; Barton, D.; Kállay, M.; Brandenburg, J. G.; Tkatchenko, A. *Nature Communications* **2021**, *12*, 3927.

(28) Parr, R. G.; Yang, W., *Density-functional theory of atoms and molecules*, 1st ed.; Oxford University Press: Oxford, 1989.

(29) Koch, W.; Holthausen, M. C., *A Chemist 's Guide to Density Functional theory*, 2nd ed.; Wiley-VCH Verlag GmbH: Weinheim, 2001.

(30) Engel, E.; Dreizler, R. M., *Density Functional Theory: An Advanced Course*; Theoretical and Mathematical Physics; Springer-Verlag: Berlin Heidelberg, 2011.

(31) Klopper, W.; Bak, K. L.; Jørgensen, P.; Olsen, J.; Helgaker, T. *Journal of Physics B: Atomic, Molecular and Optical Physics* **1999**, *32*, R103–R130.

(32) Marx, D.; Hutter, J., *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*, Reprint; Cambridge University Press: Cambridge, 2012.

(33) Tuckerman, M. E., *Statistical Mechanics: Theory and Molecular Simulation*, 1st ed.; Oxford University Press: Oxford, 2010.

(34) Allen, M. P.; Tildesley, D. J., *Computer Simulation of Liquids*, 2nd ed.; Oxford University Press: Oxford, 2017.

(35) Scheraga, H. A.; Khalili, M.; Liwo, A. *Annual Review of Physical Chemistry* **2007**, *58*, 57–83.

(36) Gomes, C. M.; Faísca, P. F. N., *Protein Folding: An Introduction*, 1st ed.; Springer: Cham, 2019.

(37) Torrie, G. M.; Valleau, J. P. *Journal of Computational Physics* **1977**, *23*, 187–199.

(38) Dellago, C.; Bolhuis, P. G.; Chandler, D. *Journal of Chemical Physics* **1999**, *110*, 6617–6625.

(39) Laio, A.; Parrinello, M. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

(40) Barducci, A.; Bonomi, M.; Parrinello, M. *WIREs Computational Molecular Science* **2011**, *1*, 826–843.

(41) Hao, G.-F.; Xu, W.-F.; Yang, S.-G.; Yang, G.-F. *Scientific Reports* **2015**, *5*, 15568.

(42) Bernardi, R. C.; Melo, M. C. R.; Schulten, K. *Biochimica et Biophysica Acta (BBA) - General Subjects* **2015**, *1850*, 872–877.

(43) Borrero, E. E.; Dellago, C. *The European Physical Journal Special Topics* **2016**, *225*, 1609–1620.

(44) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. *Science* **1983**, *220*, 671–680.

(45) Mouvet, F.; Villard, J.; Bolnykh, V.; Rothlisberger, U. *Accounts of Chemical Research* **2022**, *55*, 221–300.

(46) Brunk, E.; Rothlisberger, U. *Chemical Reviews* **2015**, *115*, 6217–6263.

(47) IBM-MPI-CPMD Car-Parrinello Molecular Dynamics code, http://www.cpmd.org, 2019.

(48) Kühne, T. D. et al. *Journal of Chemical Physics* **2020**, *152*, 194103.

(49) Trobec, R.; Slivnik, B.; Bulić, P.; Robič, B., *Introduction to Parallel Computing*, 1st ed.; Springer: Cham, 2018.

(50) Ufimtsev, I. S.; Martinez, T. J. *Journal of Chemical Theory and Computation* **2009**, *5*, 2619–2628.

(51) Titov, A. V.; Ufimtsev, I. S.; Luehr, N.; Martinez, T. J. *Journal of Chemical Theory and Computation* **2013**, *9*, 213–221.

(52) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. *PLOS Computational Biology* **2017**, *13*, 1–17.

(53) Manathunga, M.; Götz, A. W.; Jr, K. M. M. *Current Opinion in Structural Biology* **2022**, *75*, 102417.

(54) https://en.wikipedia.org/wiki/Moore%27s_law, *Moore's law*; Wikipedia: 2023, May.

(55) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *Journal of Chemical Physics* **2005**, *123*, 062201.

(56) Grimme, S. *Journal of Chemical Physics* **2006**, *124*, 034108.

(57) Mardirossian, N.; Head-Gordon, M. *Journal of Chemical Physics* **2018**, *148*, 241736.

(58) Görling, A.; Levy, M. *Physical Review A* **1994**, *50*, 196–204.

(59) Seidl, A.; Görling, A.; Vogl, P.; Majewski, J.; Levy, M. *Physical Review B - Condensed Matter and Materials Physics* **1996**, *53*, 3764–3774.

(60) Riley, K. E.; Hobza, P. *Journal of Physical Chemistry A* **2007**, *111*, 8257–8263.

(61) Zhao, Y.; Truhlar, D. G. *Theoretical Chemistry Accounts* **2008**, *120*, 215–241.

(62) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Physical Chemistry Chemical Physics* **2006**, *8*, 1985–1993.

(63) Marzari, N., *Intro to DFT - Day 1*; https://www.materialscloud.org/learn/videos/LwBs6523NhY6/2020-marzari-day-1-density-functional-theory: 2020, April.

(64) Marsman, M.; Grüneis, A.; Paier, J.; Kresse, G. *Journal of Chemical Physics* **2009**, *130*, 184103.

(65) Pisani, C.; Maschio, L.; Casassa, S.; Halo, M.; Schütz, M.; Usvyat, D. *Journal of Computational Chemistry* **2008**, *29*, 2113–2124.

(66) Shepherd, J. J.; Grüneis, A.; Booth, G. H.; Kresse, G.; Alavi, A. *Physical Review B* **2012**, *86*, 035111.

(67) Thom, A. J.; Alavi, A. *Physical Review Letters* **2007**, *99*, 5–8.

(68) Booth, G. H.; Thom, A. J. W.; Alavi, A. *Journal of Chemical Physics* **2009**, *131*, 54106.

(69) Willow, S. Y.; Kim, K. S.; Hirata, S. *Journal of Chemical Physics* **2012**, *137*, 204122.

(70) Booth, G. H.; Grüneis, A.; Kresse, G.; Alavi, A. *Nature* **2013**, *493*, 365–370.

(71) Baer, R.; Neuhauser, D.; Rabani, E. *Physical Review Letters* **2013**, *111*, 106402.

(72) Blunt, N. S.; Smart, S. D.; Kersten, J. A. F.; Spencer, J. S.; Booth, G. H.; Alavi, A. *Journal of Chemical Physics* **2015**, *142*, 184107.

(73) Johnson, C. M.; Doran, A. E.; Zhang, J.; Valeev, E. F.; Hirata, S. *Journal of Chemical Physics* **2016**, *145*, 154115.

(74) Holmes, A. A.; Changlani, H. J.; Umrigar, C. J. *Journal of Chemical Theory and Computation* **2016**, *12*, 1561–1571.

(75) Takeshita, T. Y.; De Jong, W. A.; Neuhauser, D.; Baer, R.; Rabani, E. *Journal of Chemical Theory and Computation* **2017**, *13*, 4605–4610.

(76) Schäfer, T.; Ramberger, B.; Kresse, G. *Journal of Chemical Physics* **2018**, *148*, 064103.

(77) Luo, H.; Alavi, A. *Journal of Chemical Theory and Computation* **2018**, *14*, 1403–1411.

(78) Chen, M.; Baer, R.; Neuhauser, D.; Rabani, E. *Journal of Chemical Physics* **2021**, *154*, 204108.

(79) Guandalini, A.; D'Amico, P.; Ferretti, A.; Varsano, D. *npj Computational Materials* **2023**, *9*, 44.

(80) Kulik, H. J. *Israel Journal of Chemistry* **2021**, *02139*, 1–14.

(81) Kalita, B.; Li, L.; McCarty, R. J.; Burke, K. *Accounts of Chemical Research* **2021**, *54*, 818–826.

(82) Kulik, H. J. et al. *Electronic Structure* **2022**, *4*, 023004.

(83) Hermann, J.; Spencer, J.; Choo, K.; Mezzacapo, A.; Foulkes, W. M. C.; Pfau, D.; Carleo, G.; Noé, F. *arXiv* **2022**, 2208.12590.

(84) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. *Nature* **2018**, *559*, 547–555.

(85) Von Lilienfeld, O. A.; Burke, K. *Nature Communications* **2020**, *11*, 10–13.

(86) Morawietz, T.; Artrith, N. *Journal of Computer-Aided Molecular Design* **2021**, *35*, 557–586.

(87) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. R. *Chemical Reviews* **2021**, *121*, 10142–10186.

(88) Westermayr, J.; Gastegger, M.; Schütt, K. T.; Maurer, R. J. *Journal of Chemical Physics* **2021**, *154*, 230903.

(89) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K. R.; Tkatchenko, A. *Chemical Reviews* **2021**, *121*, 9816–9872.

(90) Fedik, N.; Zubatyuk, R.; Kulichenko, M.; Lubbers, N.; Smith, J. S.; Nebgen, B.; Messerly, R.; Li, Y. W.; Boldyrev, A. I.; Barros, K.; Isayev, O.; Tretiak, S. *Nature Reviews Chemistry* **2022**, *6*, 653–672.

(91) Hastie, T.; Tibshirani, R.; Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, 2009.

(92) Bailer-Jones, C. A. L., *Practical Bayesian Inference: A Primer for Physical Scientists*, 1st ed.; Cambridge University Press: Cambridge, 2017.

(93) Heard, N., *An Introduction to Bayesian Inference, Methods and Computation*, 1st ed.; Springer: Cham, 2021.

(94) Holland, J. H., *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, 1st ed.; MIT Press: Cambridge, 1992.

(95) Back, T., *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*, 1st ed.; Oxford University Press: Oxford, 1996.

(96) Ashlock, D., *Evolutionary Computation for Modeling and Optimization*, 1st ed.; Springer New York: New York, 2006.

(97) Vikhar, P. A. *International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)* **2016**, 261–265.

(98) Tuckerman, M. E.; Berne, B. J.; Martyna, G. J. *Journal of Chemical Physics* **1992**, *97*, 1990–2001.

(99) Browning, N. J. Applications of Artificial Intelligence to Computational Chemistry, Ph.D. Thesis, Lausanne: EPFL, 2019.

(100) Zhao, Y.; Truhlar, D. G. *Journal of Physical Chemistry A* **2006**, *110*, 13126–13130.

(101) Verma, P.; Truhlar, D. G. *Trends in Chemistry* **2020**, *2*, 302–318.

(102) Goldberg, D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning*; Artificial Intelligence; Addison-Wesley Publishing Company: Boston, 1989.

(103) Mitchell, M., *An Introduction to Genetic Algorithms*, 1st ed.; MIT Press: Cambridge, 1998.

(104) Kramer, O., *Genetic Algorithm Essentials*, 1st ed.; Springer: Cham, 2017.

(105) Chin, W.; Dognon, J. P.; Piuzzi, F.; Tardivel, B.; Dimicoli, I.; Mons, M. *Journal of the American Chemical Society* **2005**, *127*, 707–712.

(106) Jaeqx, S.; Du, W.; Meijer, E. J.; Oomens, J.; Rijs, A. M. *Journal of Physical Chemistry A* **2013**, *117*, 1216–1227.

(107) Masson, A.; Kamrath, M. Z.; Perez, M. A.; Glover, M. S.; Rothlisberger, U.; Clemmer, D. E.; Rizzo, T. R. *Journal of the American Society for Mass Spectrometry* **2015**, *26*, 1444–1454.

(108) Aseev, O.; Perez, M. A.; Rothlisberger, U.; Rizzo, T. R. *Journal of Physical Chemistry Letters* **2015**, *6*, 2524–2529.

(109) Loquais, Y.; Gloaguen, E.; Habka, S.; Vaquero-Vara, V.; Brenner, V.; Tardivel, B.; Mons, M. *Journal of Physical Chemistry A* **2015**, *119*, 5932–5941.

(110) Anouk M. Rijs; Oomens, J., *Gas-Phase IR Spectroscopy and Structure of Biological Molecules*, 1st ed.; Springer: Cham, 2015.

(111) Bakels, S.; Gaigeot, M. P.; Rijs, A. M. *Chemical Reviews* **2020**, *120*, 3233–3260.

(112) Nielsen, F. In *Introduction to HPC with MPI for Data Science*, Nielsen, F., Ed.; Springer International Publishing: Cham, 2016, pp 195–211.

(113) Heitler, W.; London, F. *Zeitschrift für Physik* **1927**, *44*, 455–472.

(114) Bloch, F. *Zeitschrift für Physik* **1929**, *52*, 555–600.

(115) Schrödinger, E. *Annalen der Physik* **1926**, *384*, 361–376.

(116) Dreizler, R. M.; Gross, E. K. U., *Density Functional Theory*; Springer Berlin Heidelberg: Berlin, Heidelberg, 1990.

(117) Marsman, M.; Grüneis, A.; Paier, J.; Kresse, G. *Journal of Chemical Physics* **2009**, *130*, 184103.

(118) Grüneis, A.; Marsman, M.; Kresse, G. *Journal of Chemical Physics* **2010**, *133*, 074107.

(119) Del Ben, M.; Hutter, J.; Vandevondele, J. *Journal of Chemical Theory and Computation* **2012**, *8*, 4177–4188.

(120) Del Ben, M.; Hutter, J.; Vandevondele, J. *Journal of Chemical Physics* **2015**, *143*, 102803.

(121) Schäfer, T.; Ramberger, B.; Kresse, G. *Journal of Chemical Physics* **2017**, *146*, 104101.

(122) Gruber, T.; Liao, K.; Tsatsoulis, T.; Hummel, F.; Grüneis, A. *Physical Review X* **2018**, *8*, 21043.

(123) Bartlett, R. J. *Journal of Chemical Physics* **2019**, *151*, 160901.

(124) Santra, G.; Martin, J. M. *AIP Conference Proceedings* **2022**, *2611*, 2–5.

(125) Stein, F.; Hutter, J.; Rybkin, V. V. *Molecules* **2020**, *25*, 5174.

(126) Koopmans, T. *Physica* **1934**, *1*, 104–113.

(127) Cohen-Tannoudji, C.; Laloe, F.; Diu, B., *Quantum mechanics, Volume 2*, 1st ed.; Wiley: New York, 1977.

(128) Zhechkov, L.; Heine, T.; Patchkovskii, S.; Seifert, G.; Duarte, H. A. *Journal of Chemical Theory and Computation* **2005**, *1*, 841–847.

(129) Shang, H.; Yang, J. *Frontiers in Chemistry* **2020**, *8*, 956.

(130) Ren, X.; Rinke, P.; Joas, C.; Scheffler, M. *Journal of Materials Science* **2012**, *47*, 7447–7471.

(131) Del Ben, M.; Hutter, J.; Vandevondele, J. *Journal of Chemical Theory and Computation* **2013**, *9*, 2654–2671.

(132) Ihrig, A. C.; Wieferink, J.; Zhang, I. Y.; Ropo, M.; Ren, X.; Rinke, P.; Scheffler, M.; Blum, V. *New Journal of Physics* **2015**, *17*, 093020.

(133) Almlöf, J. *Chemical Physics Letters* **1991**, *181*, 319–320.

(134) Neuhauser, D.; Rabani, E.; Baer, R. *Journal of Chemical Theory and Computation* **2013**, *9*, 24–27.

(135) Hohenberg, P.; Kohn, W. *Physical Review* **1964**, *136*, B864–B871.

(136) Thomas, L. H. *Mathematical Proceedings of the Cambridge Philosophical Society* **1927**, *23*, 542–548.

(137) Fermi, E. *Rendiconti dell'Accademia Nazionale dei Lincei* **1928**, *6*, 602–607.

(138) Dirac, P. A. M. *Mathematical Proceedings of the Cambridge Philosophical Society* **1930**, *26*, 376–385.

(139) Witt, W. C.; del Rio, B. G.; Dieterich, J. M.; Carter, E. A. *Journal of Materials Research* **2018**, *33*, 777–795.

(140) Kohn, W.; Sham, L. J. *Physical Review* **1965**, *140*, A1133–A1138.

(141) Mardirossian, N.; Head-Gordon, M. *Molecular Physics* **2017**, *115*, 2315–2372.

(142) Perdew, J. P.; Schmidt, K. *AIP Conference Proceedings* **2001**, *577*, 1–20.

(143) Broqvist, P.; Alkauskas, A.; Pasquarello, A. *Physical Review B - Condensed Matter and Materials Physics* **2009**, *80*, 1–13.

(144) Del Ben, M.; Hutter, J.; VandeVondele, J. *Journal of Chemical Physics* **2015**, *143*, 054506.

(145) Bircher, M. P.; Rothlisberger, U. *Journal of Physical Chemistry Letters* **2018**, *9*, 3886–3890.

(146) Goerigk, L.; Grimme, S. *Physical Chemistry Chemical Physics* **2011**, *13*, 6670–6688.

(147) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *Journal of Chemical Physics* **2010**, *132*, 154104.

(148) Perdew, J. P.; Kurth, S. In *A Primer in Density Functional Theory*, Fiolhais, C., Nogueira, F., Marques, M. A. L., Eds., 1st ed.; Springer: Berlin, Heidelberg, 2003, pp 1–55.

(149) Sun, J.; Ruzsinszky, A.; Perdew, J. P. *Physical Review Letters* **2015**, *115*, 36402.

(150) Ceperley, D.; Alder, B. J. *Physical Review Letters* **1980**, *45*, 566–569.

(151) Vosko, S. H.; Wilk, L.; Nusair, M. *Canadian Journal of Physics* **1980**, *58*, 1200–1211.

(152) Perdew, J. P.; Wang, Y. *Physical Review B* **1992**, *45*, 244–249.

(153) Harris, J. *Physical Review B* **1985**, *31*, 1770–1779.

(154) Becke, A. D. *Journal of Chemical Physics* **1986**, *84*, 4524–4529.

(155) Becke, A. D. *Journal of Chemical Physics* **1988**, *88*, 2547–2553.

(156) Colle, R.; Salvetti, O. *Theoretica chimica acta* **1975**, *37*, 329–334.

(157) Schmider, H. L.; Becke, A. D. *Journal of Molecular Structure: THEOCHEM* **2000**, *527*, 51–61.

(158) Becke, A. D. *Journal of Chemical Physics* **1993**, *98*, 1372–1377.

(159) Langreth, D. C.; Perdew, J. P. *Solid State Communications* **1975**, *17*, 1425–1429.

(160) Gunnarsson, O.; Lundqvist, B. I. *Physical Review B* **1976**, *13*, 4274–4298.

(161) Perdew, J. P.; Ernzerhof, M.; Burke, K. *Journal of Chemical Physics* **1996**, *105*, 9982–9985.

(162) Becke, A. D. *Journal of Chemical Physics* **1993**, *98*, 5648–5652.

(163) Zhang, M.-Y.; Cui, Z.-H.; Wang, Y.-C.; Jiang, H. *WIREs Computational Molecular Science* **2020**, *10*, e1476.

(164) Baer, R.; Livshits, E.; Salzner, U. *Annual Review of Physical Chemistry* **2010**, *61*, 85–109.

(165) Görling, A.; Levy, M. *Physical Review B* **1993**, *47*, 13105–13113.

(166) Görling, A.; Levy, M. *Physical Review A* **1995**, *52*, 4493–4499.

(167) Engel, E.; Dreizler, R. M. *Journal of Computational Chemistry* **1999**, *20*, 31–50.

(168) Ernzerhof, M. *Chemical Physics Letters* **1996**, *263*, 499–506.

(169) Schwabe, T.; Grimme, S. *Physical Chemistry Chemical Physics* **2007**, *9*, 3397–3406.

(170) Goerigk, L.; Grimme, S. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, *4*, 576–600.

(171) Langreth, D. C.; Perdew, J. P. *Physical Review B* **1980**, *21*, 5469–5493.

(172) Furche, F. *Physical Review B* **2001**, *64*, 195120.

(173) Fuchs, M.; Gonze, X. *Physical Review B* **2002**, *65*, 235109.

(174) Furche, F. *Journal of Chemical Physics* **2008**, *129*, 114105.

(175) Goerigk, L.; Grimme, S. *Journal of Chemical Theory and Computation* **2011**, *7*, 291–309.

(176) Kozuch, S.; Gruzman, D.; Martin, J. M. L. *Journal of Physical Chemistry C* **2010**, *114*, 20801–20808.

(177) Langreth, D. C.; Perdew, J. P. *Physical Review B* **1977**, *15*, 2884–2901.

(178) Olsen, T.; Thygesen, K. S. *Physical Review B* **2013**, *87*, 075111.

(179) Nguyen, N. L.; Colonna, N.; de Gironcoli, S. *Physical Review B* **2014**, *90*, 45138.

(180) Yao, Y.; Kanai, Y. *Journal of Physical Chemistry Letters* **2021**, *12*, 6354–6362.

(181) Von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Physical Review Letters* **2004**, *93*, 1–4.

(182) Lin, I. C.; Seitsonen, A. P.; Coutinho-Neto, M. D.; Tavernelli, I.; Rothlisberger, U. *Journal of Physical Chemistry B* **2009**, *113*, 1127–1131.

(183) Lin, I. C.; Seitsonen, A. P.; Tavernelli, I.; Rothlisberger, U. *Journal of Chemical Theory and Computation* **2012**, *8*, 3902–3910.

(184) Grimme, S. *Journal of Computational Chemistry* **2006**, *27*, 1787–1799.

(185) Vydrov, O. A.; Van Voorhis, T. *Journal of Chemical Physics* **2010**, *132*, 164113.

(186) Sabatini, R.; Gorni, T.; De Gironcoli, S. *Physical Review B* **2013**, *87*, 4–7.

(187) Miceli, G.; De Gironcoli, S.; Pasquarello, A. *Journal of Chemical Physics* **2015**, *142*, 034501.

(188) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Physical Review Letters* **2004**, *92*, 22–25.

(189) Tkatchenko, A.; Distasio, R. A.; Car, R.; Scheffler, M. *Physical Review Letters* **2012**, *108*, 1–5.

(190) Ferri, N.; Distasio, R. A.; Ambrosetti, A.; Car, R.; Tkatchenko, A. *Physical Review Letters* **2015**, *114*, 1–5.

(191) Wiktor, J.; Ambrosio, F.; Pasquarello, A. *Journal of Chemical Physics* **2017**, *147*, 2016–2018.

(192) Jensen, F. *Journal of Physical Chemistry A* **2017**, *121*, 6104–6107.

(193) Roothaan, C. C. J. *Reviews of Modern Physics* **1951**, *23*, 69–89.

(194) Hall, G. G. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **1951**, *205*, 541–552.

(195) Marx, D.; Hutter, J., *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*, 1st ed.; Cambridge University Press: Cambridge, 2009.

(196) Ashcroft, N. W.; Mermin, N. D., *Solid State Physics*, 1st ed.; Thomson Learning: Boston, 1976.

(197) Martyna, G. J.; Tuckerman, M. E. *Journal of Chemical Physics* **1999**, *110*, 2810–2821.

(198) Genovese, L.; Deutsch, T.; Neelov, A.; Goedecker, S.; Beylkin, G. *Journal of Chemical Physics* **2006**, *125*, 074105.

(199) Genovese, L.; Deutsch, T.; Goedecker, S. *Journal of Chemical Physics* **2007**, *127*, 054704.

(200) Hockney, R. W. *Methods Computational Physics* **1970**, *9*, 135–211.

(201) Colony R.; Reynolds R. R. *Spring Joint Computer Conference* **1970**, *70*, 409–416.

(202) Boys, S. F.; Bernardi, F. *Molecular Physics* **1970**, *19*, 553–566.

(203) Van Duijneveldt, F. B.; van de Rijdt, J. G. D.; van Lenthe, J. H. *Chemical Reviews* **1994**, *94*, 1873–1885.

(204) Daudey, J. P.; Claverie, P.; Malrieu, J. P. *International Journal of Quantum Chemistry* **1974**, *8*, 1–15.

(205) Schwenke, D. W.; Truhlar, D. G. *Journal of Chemical Physics* **1985**, *82*, 2418–2426.

(206) Loushin, S. K.; Liu, S. Y.; Dykstra, C. E. *Journal of Chemical Physics* **1985**, *84*, 2720–2725.

(207) Gutowski, M.; Van Lenthe, J. H.; Verbeek, J.; Van Duijneveldt, F. B.; Chałasinski, G. *Chemical Physics Letters* **1986**, *124*, 370–375.

(208) Feller, D. *Journal of Chemical Physics* **1992**, *96*, 6104–6114.

(209) Gutowski, M.; Van Duijneveldt-Van De Rijdt, J. G.; Van Lenthe, J. H.; Van Duijneveldt, F. B. *Journal of Chemical Physics* **1993**, *98*, 4728–4737.

(210) Van Mourik, T.; Wilson, A. K.; Peterson, K. A.; Woon, D. E.; Dunning, T. H. *Advances in Quantum Chemistry* **1998**, *31*, 105–135.

(211) Mentel, M.; Baerends, E. J. *Journal of Chemical Theory and Computation* **2014**, *10*, 252–267.

(212) Halkier, A.; Klopper, W.; Helgaker, T.; Jørgensen, P.; Taylor, P. R. *Journal of Chemical Physics* **1999**, *111*, 9157–9167.

(213) Dunning, T. H. *Journal of Chemical Physics* **1989**, *90*, 1007–1023.

(214) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *Journal of Chemical Physics* **1992**, *96*, 6796–6806.

(215) Woon, D. E.; Dunning, T. H. *Journal of Chemical Physics* **1993**, *98*, 1358–1371.

(216) Feller, D.; Peterson, K. A. *Journal of Chemical Physics* **1998**, *108*, 154–176.

(217) De Lara-Castells, M. P.; Krems, R. V.; Buchachenko, A. A.; Delgado-Barrio, G.; Villarreal, P. *Journal of Chemical Physics* **2001**, *115*, 10438–10449.

(218) Dunning, T. H. *Journal of Physical Chemistry A* **2000**, *104*, 9062–9080.

(219) Martin Jan, M. *Chemical Physics Letters* **1996**, *259*, 669–678.

(220) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *Journal of Chemical Physics* **1997**, *106*, 9639–9646.

(221) Kutzelnigg, W. *Theoretica chimica acta* **1985**, *68*, 445–469.

(222) Marchetti, O.; Werner, H. J. *Physical Chemistry Chemical Physics* **2008**, *10*, 3400–3409.

(223) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chemical Physics Letters* **1998**, *286*, 243–252.

(224) Wilson, A. K.; Dunning, T. H. *Journal of Chemical Physics* **1997**, *106*, 8718–8726.

(225) Hutter, J.; Curioni, A. *Parallel Computing* **2005**, *31*, 1–17.

(226) Hutter, J.; Curioni, A. *ChemPhysChem* **2005**, *6*, 1788–1793.

(227) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. *Journal of Chemical Physics* **2010**, *132*, 144104.

(228) Harl, J.; Kresse, G. *Physical Review B - Condensed Matter and Materials Physics* **2008**, *77*, 1–8.

(229) Lanczos, C. *Journal of Research of the National Bureau of Standards* **1950**, *45*, 255–282.

(230) Paige, C. C. *IMA Journal of Applied Mathematics* **1972**, *10*, 373–381.

(231) Davidson, E. R. *Journal of Computational Physics* **1975**, *17*, 87–94.

(232) Bircher, M. P.; Villard, J.; Rothlisberger, U. *Journal of Chemical Theory and Computation* **2020**, *16*, 6550–6559.

(233) Gygi, F.; Baldereschi, A. *Physical Review B* **1986**, *34*, 4405–4408.

(234) Massidda, S.; Posternak, M.; Baldereschi, A. *Physical Review B* **1993**, *48*, 5058–5068.

(235) Chawla, S.; Voth, G. A. *Journal of Chemical Physics* **1998**, *108*, 4697–4700.

(236) Sorouri, A.; Foulkes, W. M.; Hine, N. D. *Journal of Chemical Physics* **2006**, *124*, 1–7.

(237) Carrier, P.; Rohra, S.; Görling, A. *Physical Review B - Condensed Matter and Materials Physics* **2007**, *75*, 1–10.

(238) Blöchl, P. E. *Journal of Chemical Physics* **1995**, *103*, 7422–7428.

(239) Paier, J.; Hirschl, R.; Marsman, M.; Kresse, G. *Journal of Chemical Physics* **2005**, *122*, 234102.

(240) Ewald, P. P. *Annalen der Physik* **1921**, *369*, 253–287.

(241) Del Bene, J. E. *Journal of Physical Chemistry* **1993**, *97*, 107–110.

(242) Woon, D. E.; Dunning, T. H. *Journal of Chemical Physics* **1994**, *100*, 2975–2988.

(243) Peterson, K. A.; Dunning, T. H. *Journal of Chemical Physics* **1995**, *102*, 2032–2041.

(244) Peterson, K. A.; Dunning, T. H. *Journal of Physical Chemistry A* **1997**, *101*, 6280–6292.

(245) Woon, D. E. *Chemical Physics Letters* **1993**, *204*, 29–35.

(246) Woon, D. E. *Journal of Chemical Physics* **1994**, *100*, 2838–2850.

(247) Woon, D. E.; Dunning, T. H.; Peterson, K. A. *Journal of Chemical Physics* **1996**, *104*, 5883–5891.

(248) Raghavachari, K.; Anderson, J. B. *Journal of Physical Chemistry* **1996**, *100*, 12960–12973.

(249) Truhlar, D. G. *Chemical Physics Letters* **1998**, *294*, 45–48.

(250) Varandas, A. J. *Journal of Chemical Physics* **2000**, *113*, 8880–8887.

(251) Eshuis, H.; Furche, F. *Journal of Chemical Physics* **2012**, *136*, 084105.

(252) Woon, D. E.; Dunning, T. H. *Journal of Chemical Physics* **1993**, *99*, 1914–1929.

(253) Peterson, K. A.; Kendall, R. A.; Dunning, T. H. *Journal of Chemical Physics* **1993**, *99*, 1930–1944.

(254) Feyereisen, M. W.; Feller, D.; Dixon, D. A. *Journal of Physical Chemistry* **1996**, *100*, 2993–2997.

(255) Peterson, K. A.; Woon, D. E.; Dunning, T. H. *Journal of Chemical Physics* **1994**, *100*, 7410–7415.

(256) Woon, D. E.; Dunning, T. H. *Journal of Chemical Physics* **1994**, *101*, 8877–8893.

(257) Feller, D.; Sordo, J. A. *Journal of Chemical Physics* **2000**, *112*, 5604–5610.

(258) Schwartz, C. *Physical Review* **1962**, *126*, 1015–1019.

(259) Kutzelnigg, W.; Morgan, J. D. *Journal of Chemical Physics* **1992**, *96*, 4484–4508.

(260) Carroll, D. P.; Silverstone, H. J.; Metzger, R. M. *Journal of Chemical Physics* **1979**, *71*, 4142–4163.

(261) Hill, R. N. *Journal of Chemical Physics* **1985**, *83*, 1173–1196.

(262) Martin, J. M. *Chemical Physics Letters* **1996**, *259*, 679–682.

(263) Martin, J. M. L.; Taylor, P. R. *Journal of Chemical Physics* **1997**, *106*, 8620–8623.

(264) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Olsen, J. *Chemical Physics Letters* **1999**, *302*, 437–446.

(265) Goedecker, S.; Teter, M. *Physical Review B* **1996**, *54*, 1703–1710.

(266) Hutter, J.; Lüthi, H. P.; Parrinello, M. *Computational Materials Science* **1994**, *2*, 244–248.

(267) Neese, F. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 73–78.

(268) Neese, F. *WIREs Computational Molecular Science* **2022**, *16*, e1606.

(269) TURBOMOLE V7.1, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, 2016.

(270) Burns, L. A.; Marshall, M. S.; Sherrill, C. D. *Journal of Chemical Theory and Computation* **2014**, *10*, 49–57.

(271) Miliordos, E.; Aprà, E.; Xantheas, S. S. *Journal of Physical Chemistry A* **2014**, *118*, 7568–7578.

(272) Del Ben, M.; Hutter, J.; Vandevondele, J. *Journal of Chemical Theory and Computation* **2013**, *9*, 2654–2671.

(273) Morawietz, T.; Singraber, A.; Dellago, C.; Behler, J. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, *113*, 8368–8373.

(274) Mazzola, G.; Yunoki, S.; Sorella, S. *Nature Communications* **2014**, *5*, 3487.

(275) Gao, W.; Tkatchenko, A. *Physical Review Letters* **2013**, *111*, 45501.

(276) Jensen, F., *Introduction to Computational Chemistry*, 3rd ed.; John Wiley & Sons: Chichester, 2017.

(277) Peverati, R.; Truhlar, D. G. *Philosophical Transactions of the Royal Society A* **2014**, *372*, 20120476.

(278) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. *Physical Chemistry Chemical Physics* **2017**, *19*, 32184–32215.

(279) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *Journal of Physical Chemistry A* **2004**, *108*, 4786–4791.

(280) Bartlett, R. J.; Grabowski, I.; Hirata, S.; Ivanov, S. *Journal of Chemical Physics* **2005**, *122*, 034104.

(281) Peverati, R.; Head-Gordon, M. *Journal of Chemical Physics* **2013**, *139*, 24110.

(282) Su, N. Q.; Xu, X. *Journal of Chemical Physics* **2014**, *140*, 18A512.

(283) Zhao, Y.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2006**, *2*, 1009–1018.

(284) Schwabe, T.; Grimme, S. *Physical Chemistry Chemical Physics* **2006**, *8*, 4398–4401.

(285) Zheng, J.; Zhao, Y.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2009**, *5*, 808–821.

(286) Medvedev, M. G.; Bushmarinov, I. S.; Sun, J.; Perdew, J. P.; Lyssenko, K. A. *Science* **2017**, *355*, 49–52.

(287) Su, N. Q.; Zhu, Z.; Xu, X. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, 2287–2292.

(288) Ayala, P. Y.; Scuseria, G. E. *Journal of Computational Chemistry* **2000**, *21*, 1524–1531.

(289) Ayala, P. Y.; Konstantin, K. N.; Kudin, N.; Scuseria, G. E. *Journal of Chemical Physics* **2001**, *115*, 9698–9707.

(290) Pisani, C.; Busso, M.; Capecchi, G.; Casassa, S.; Dovesi, R.; Maschio, L.; Zicovich-Wilson, C.; Schütz, M. *Journal of Chemical Physics* **2005**, *122*, 1–12.

(291) Casassa, S.; Halo, M.; Maschio, L.; Roetti, C.; Pisani, C. *Theoretical Chemistry Accounts* **2007**, *117*, 781–791.

(292) Pisani, C.; Schütz, M.; Casassa, S.; Usvyat, D.; Maschio, L.; Lorenz, M.; Erba, A. *Physical Chemistry Chemical Physics* **2012**, *14*, 7615–7628.

(293) Paulus, B. *Physics Reports* **2006**, *428*, 1–52.

(294) Manby, F. R.; Alfè, D.; Gillan, M. J. *Physical Chemistry Chemical Physics* **2006**, *8*, 5178–5180.

(295) Gillan, M. J.; Alfè, D.; de Gironcoli, S.; Manby, F. R. *Journal of computational chemistry* **2008**, *29*, 2098–2106.

(296) Del Ben, M.; Hutter, J.; Vandevondele, J. *Journal of Chemical Theory and Computation* **2012**, *8*, 4177–4188.

(297) Hutter, J.; Wilhelm, J.; Rybkin, V. V.; Ben, M. D.; VandeVondele, J. *Handbook of Materials Modeling* **2018**, 1–21.

(298) Grüneis, A.; Shepherd, J. J.; Alavi, A.; Tew, D. P.; Booth, G. H. *Journal of Chemical Physics* **2013**, *139*, 084112.

(299) Grüneis, A. *Physical Review Letters* **2015**, *115*, 1–6.

(300) Grüneis, A. *Handbook of Materials Modeling* **2018**, 1–16.

(301) Sharkas, K.; Toulouse, J.; Maschio, L.; Civalleri, B. *Journal of Chemical Physics* **2014**, *141*, 44105.

(302) Loos, P.-F.; Pradines, B.; Scemama, A.; Toulouse, J.; Giner, E. *Journal of Physical Chemistry Letters* **2019**, *10*, 2931–2937.

(303) Mardirossian, N.; Head-Gordon, M. *Journal of Chemical Theory and Computation* **2013**, *9*, 4453–4461.

(304) Bircher, M. P.; López-Tarifa, P.; Rothlisberger, U. *Journal of Chemical Theory and Computation* **2019**, *15*, 557–571.

(305) Halo, M.; Casassa, S.; Maschio, L.; Pisani, C. *Physical Chemistry Chemical Physics* **2009**, *11*, 586–592.

(306) Kelly, H. P. *Physical Review* **1963**, *131*, 684–699.

(307) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Computer Physics Communications* **2005**, *167*, 103–128.

(308) Adler, S. L. *Physical Review* **1962**, *126*, 413–420.

(309) Wiser, N. *Physical Review* **1963**, *129*, 62–69.

(310) Cremer, D. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 509–530.

(311) Shepherd, J. J.; Grüneis, A. *Physical Review Letters* **2013**, *110*, 226401.

(312) Gygi, F.; Baldereschi, A. *Physical Review B* **1986**, *34*, 4405–4408.

(313) Broqvist, P.; Alkauskas, A.; Pasquarello, A. *Physical Review B* **2009**, *80*, 85114.

(314) Huzinaga, S.; Arnau, C. *Physical Review A* **1969**, *1*, 1285–1288.

(315) Huzinaga, S.; Arnau, C. *Journal of Chemical Physics* **1971**, *54*, 1948–1951.

(316) Csepes, Z.; Kozmutza, C. *Croatica Chemica Acta* **1984**, *57*, 855–864.

(317) Palmieri, P.; Tarroni, R.; Rettrup, S. *Journal of Chemical Physics* **1994**, *100*, 5849–5856.

(318) Neogrády, P.; Pitoňák, M.; Urban, M. *Molecular Physics* **2005**, *103*, 2141–2157.

(319) Ge, Q.; Gao, Y.; Baer, R.; Rabani, E.; Neuhauser, D. *Journal of Physical Chemistry Letters* **2014**, *5*, 185–189.

(320) Neuhauser, D.; Baer, R.; Zgid, D. *Journal of Chemical Theory and Computation* **2017**, *13*, 5396–5403.

(321) Willow, S. Y.; Kim, K. S.; Hirata, S. *Physical Review B - Condensed Matter and Materials Physics* **2014**, *90*, 1–5.

(322) Häser, M.; Almlöf, J. *Journal of Chemical Physics* **1992**, *96*, 489–494.

(323) Wilson, A. K.; Almlöf, J. *Theoretica Chimica Acta* **1997**, *95*, 49–62.

(324) Kats, D.; Usvyat, D.; Schütz, M. *Physical Chemistry Chemical Physics* **2008**, *10*, 3430–3439.

(325) Kutzelnigg, W.; Klopper, W. *Journal of Chemical Physics* **1991**, *94*, 1985–2001.

(326) Irmler, A.; Gallo, A.; Hummel, F.; Grüneis, A. *Physical Review Letters* **2019**, *123*, 156401.

(327) Hedin, L. *Physical Review* **1965**, *139*, A796–A823.

(328) Vlček, V. *Journal of Chemical Theory and Computation* **2019**, *15*, 6254–6266.

(329) Gell-Mann, M.; Brueckner, K. A. *Physical Review* **1957**, *106*, 364–368.

(330) Goedecker, S.; Teter, M.; Hutter, J. *Physical Review B* **1996**, *54*, 1703–1710.

(331) IBM-MPI-CPMD Car-Parrinello Molecular Dynamics code, http://www.cpmd.org.

(332) Frisch, M. J. et al. Gaussian 16 Revision A.03, Wallingford CT, 2016.

(333) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Computer Physics Communications* **2005**, *167*, 103–128.

(334) Liberatore, E.; Meli, R.; Rothlisberger, U. *Journal of Chemical Theory and Computation* **2018**, *14*, 2834–2842.

(335) Tuckerman, M. E.; Martyna, G. J.; Berne, B. J. *Journal of Chemical Physics* **1990**, *93*, 1287–1291.

(336) Hellmann, H., *Einführung in die Quantenchemie*, 1st ed.; Franz Deuticke: Leipzig, 1937.

(337) Feynman, R. P. *Physical Review* **1939**, *56*, 340–343.

(338) Nosé, S. *Journal of Chemical Physics* **1984**, *81*, 511–519.

(339) Hoover, W. G. *Physical Review A* **1985**, *31*, 1695–1697.

(340) Tuckerman, M. E.; Parrinello, M. *Journal of Chemical Physics* **1994**, *101*, 1316–1329.

(341) Hartke, B.; Gibson, D. A.; Carter, E. A. *International Journal of Quantum Chemistry* **1993**, *45*, 59–70.

(342) Luehr, N.; Markland, T. E.; Martínez, T. J. *Journal of Chemical Physics* **2014**, *140*, 84116.

(343) Steele, R. P. *Journal of Chemical Physics* **2013**, *139*, 011102.

(344) Guidon, M.; Schiffmann, F.; Hutter, J.; Vandevondele, J. *Journal of Chemical Physics* **2008**, *128*, 214104.

(345) Liberatore, E.; Meli, R.; Rothlisberger, U. *Journal of Chemical Theory and Computation* **2018**, *14*, 2834–2842.

(346) Ma, Q.; Izaguirre, J. A.; Skeel, R. D. *SIAM Journal on Scientific Computing* **2003**, *24*, 1951–1973.

(347) Morrone, J. A.; Markland, T. E.; Ceriotti, M.; Berne, B. J. *Journal of Chemical Physics* **2011**, *134*, 14103.

(348) Ceriotti, M.; Bussi, G.; Parrinello, M. *Journal of Chemical Theory and Computation* **2010**, *6*, 1170–1180.

(349) Minary, P.; Tuckerman, M. E.; Martyna, G. J. *Physical Review Letters* **2004**, *93*, 1–4.

(350) Abreu, C. R. A.; Tuckerman, M. E. *Journal of Chemical Theory and Computation* **2020**, *16*, 7314–7327.

(351) Bishop, C. M., *Pattern Recognition and Machine Learning*, 1st ed.; Springer: New York, 2006.

(352) Schütt, K. T.; Chmiela, S.; von Lilienfeld, O. A.; Tkatchenko, A.; Tsuda, K.; Müller, K. R., *Machine Learning Meets Quantum Physics*, 1st ed.; Lecture Notes in Physics; Springer: Cham, 2020.

(353) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. *Chemical Reviews* **2021**, *121*, 10142–10186.

(354) Gkeka, P. et al. *Journal of Chemical Theory and Computation* **2020**, *16*, 4757–4775.

(355) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. *Physical Review Letters* **2010**, *104*, 1–4.

(356) Chmiela, S.; Sauceda, H. E.; Müller, K. R.; Tkatchenko, A. *Nature Communications* **2018**, *9*, 3887.

(357) Behler, J.; Parrinello, M. *Physical Review Letters* **2007**, *98*, 146401.

(358) Gastegger, M.; Behler, J.; Marquetand, P. *Chemical Science* **2017**, *8*, 6924–6935.

(359) Kalita, B.; Li, L.; McCarty, R. J.; Burke, K. *Accounts of Chemical Research* **2021**, *54*, 818–826.

(360) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K. R.; Burke, K. *Nature Communications* **2020**, *11*, 5223.

(361) Dick, S.; Fernandez-Serra, M. *Nature Communications* **2020**, *11*, 3509.

(362) Christensen, A. S.; Faber, F. A.; von Lilienfeld, O. A. *Journal of Chemical Physics* **2019**, *150*, 064105.

(363) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole Von Lilienfeld, O. *Journal of Chemical Physics* **2020**, *152*, 44107.

(364) Pronobis, W.; Müller, K.-R. In *Machine Learning Meets Quantum Physics*, Schütt, K. T., Chmiela, S., von Lilienfeld, O. A., Tkatchenko, A., Tsuda, K., Müller, K.-R., Eds.; Springer International Publishing: Cham, 2020, pp 25–36.

(365) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. *Chemical Reviews* **2021**, *121*, 9759–9815.

(366) Rasmussen, C. E.; Williams, C. K. I., *Gaussian Processes for Machine Learning*, 1st ed.; Adaptive Computation and Machine Learning; MIT Press: Cambridge, 2006.

(367) Huang, B.; von Lilienfeld, O. A. *Nature Chemistry* **2020**, *12*, 945–951.

(368) Shapeev, A.; Gubaev, K.; Tsymbalov, E.; Podryabinkin, E. In *Machine Learning Meets Quantum Physics*, Schütt, K. T., Chmiela, S., von Lilienfeld, O. A., Tkatchenko, A., Tsuda, K., Müller, K.-R., Eds.; Springer International Publishing: Cham, 2020, pp 309–329.

(369) Imbalzano, G.; Zhuang, Y.; Kapil, V.; Rossi, K.; Engel, E. A.; Grasselli, F.; Ceriotti, M. *Journal of Chemical Physics* **2021**, *154*, 074102.

(370) Li, Z.; Kermode, J. R.; De Vita, A. *Physical Review Letters* **2015**, *114*, 096405.

(371) Jinnouchi, R.; Miwa, K.; Karsai, F.; Kresse, G.; Asahi, R. *Journal of Physical Chemistry Letters* **2020**, *11*, 6946–6955.

(372) Böselt, L.; Thürlemann, M.; Riniker, S. *Journal of Chemical Theory and Computation* **2021**, *17*, 2641–2658.

(373) Westermayr, J.; Marquetand, P. *Chemical Reviews* **2020**, *121*, 9873–9926.

(374) Car, R.; Parrinello, M. *Physical Review Letters* **1985**, *55*, 2471–2474.

(375) Franks, F., *Water: A Matrix of Life*, 2nd ed.; RSC Paperbacks; Royal Society of Chemistry: Cambridge, 2000.

(376) Eisenberg, D.; Kauzmann, W., *The Structure and Properties of Water*, 1st ed.; Oxford Classic Texts in the Physical Sciences; Oxford University Press: Oxford, 1969.

(377) Atkins, P.; de Paula, J., *Physical Chemistry: Thermodynamics, Structure, and Change*, 10th ed.; W. H. Freeman: New York, 2014.

(378) Chaplin, M. F., *Structure and Properties of Water in its Various States*; John Wiley & Sons: New York, 2019.

(379) Jencks, W. P. *Chemical Reviews* **1972**, *72*, 705–718.

(380) Geissler, P. L.; Dellago, C.; Chandler, D.; Hutter, J.; Parrinello, M. *Science* **2001**, *291*, 2121–2124.

(381) Zhao, Y.; Truhlar, D. G. *Reviews in Mineralogy & Geochemistry* **2010**, *71*, 19–37.

(382) Chen, M.; Zheng, L.; Santra, B.; Ko, H.-Y.; DiStasio Jr, R. A.; Klein, M. L.; Car, R.; Wu, X. *Nature Chemistry* **2018**, *10*, 413–419.

(383) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. *Proceedings of the National Academy of Sciences* **1999**, *96*, 9997–10002.

(384) Ball, P. *Chemical Reviews* **2008**, *108*, 74–108.

(385) Manabe, S.; Wetherald, R. T. *Journal of Atmospheric Sciences* **1967**, *24*, 241–259.

(386) Ramanathan, V.; Cess, R. D.; Harrison, E. F.; Minnis, P.; Barkstrom, B. R.; Ahmad, E.; Hartmann, D. *Science* **1989**, *243*, 57–63.

(387) Chapra, S. C., *Surface Water-Quality Modeling*; Waveland Press: Long Grove, 2008.

(388) Simon, M.-O.; Li, C.-J. *Chemical Society Reviews* **2012**, *41*, 1415–1427.

(389) Maroncelli, M.; Fleming, G. R. *Journal of Chemical Physics* **1987**, *86*, 6221–6239.

(390) Pal, S. K.; Peon, J.; Bagchi, B.; Zewail, A. H. *Journal of Physical Chemistry B* **2002**, *106*, 12376–12395.

(391) Cohen-Tanugi, D.; Grossman, J. C. *Nano Letters* **2012**, *12*, 3602–3608.

(392) Werber, J. R.; Osuji, C. O.; Elimelech, M. *Nature Reviews Materials* **2016**, *1*, 16018.

(393) Tanford, C. *Science* **1978**, *200*, 1012–1018.

(394) Zhou, R.; Huang, X.; Margulis, C. J.; Berne, B. J. *Science* **2004**, *305*, 1605–1609.

(395) Camilloni, C.; Bonetti, D.; Morrone, A.; Giri, R.; Dobson, C. M.; Brunori, M.; Gianni, S.; Vendruscolo, M. *Scientific Reports* **2016**, *6*, 1–9.

(396) Baldwin, R. L.; Rose, G. D. *Proceedings of the National Academy of Sciences* **2016**, *113*, 12462–12466.

(397) Warshel, A.; Levitt, M. *Journal of Molecular Biology* **1976**, *103*, 227–249.

(398) Warshel, A.; Bora, R. P. *Journal of Chemical Physics* **2016**, *144*, 180901.

(399) Chaplin, M. *Nature Reviews Molecular Cell Biology* **2006**, *7*, 861–866.

(400) Marrink, S. J.; Corradi, V.; Souza, P. C. T.; Ingólfsson, H. I.; Tieleman, D. P.; Sansom, M. S. P. *Chemical Reviews* **2019**, *119*, 6184–6226.

(401) Pablo G Debenedetti *Journal of Physics: Condensed Matter* **2003**, *15*, R1669.

(402) Kell, G. S. *Journal of Chemical & Engineering Data* **1975**, *20*, 97–105.

(403) Nilsson, A.; Pettersson, L. G. M. *Chemical Physics* **2011**, *389*, 1–34.

(404) Stillinger, F. H. *Science* **1980**, *209*, 451–457.

(405) Fennell Evans, D.; Wennerström, H., *The Colloidal Domain: Where Physics, Chemistry, Biology, and Technology Meet*, 2nd ed.; Wiley-VCH: Weinheim, 1999.

(406) Adamson, A. W.; Gast, A. P., *Physical Chemistry of Surfaces*, 6th ed.; John Wiley & Sons: New York, 1997.

(407) Hasted, J. B., *Aqueous Dielectrics*; Chapman and Hall: London, 1973.

(408) Kaatze, U. *Journal of Chemical & Engineering Data* **1989**, *34*, 371–374.

(409) Gallo, P. et al. *Chemical Reviews* **2016**, *116*, 7463–7500.

(410) Ni, K.; Fang, H.; Yu, Z.; Fan, Z. *Journal of Molecular Liquids* **2019**, *278*, 234–238.

(411) Laage, D.; Hynes, J. T. *Science* **2006**, *311*, 832–835.

(412) Clark, G. N. I.; Cappa, C. D.; Smith, J. D.; Saykally, R. J.; Head-Gordon, T. *Molecular Physics* **2010**, *108*, 1415–1433.

(413) Teixeira, J. *Molecular Physics* **2012**, *110*, 249–258.

(414) Herrero, C.; Pauletti, M.; Tocci, G.; Iannuzzi, M.; Joly, L. *Proceedings of the National Academy of Sciences of the United States of America* **2022**, *119*, 1–8.

(415) Chen, M.; Ko, H. Y.; Remsing, R. C.; Calegari Andrade, M. F.; Santra, B.; Sun, Z.; Selloni, A.; Car, R.; Klein, M. L.; Perdew, J. P.; Wu, X. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, *114*, 10846–10851.

(416) Glasel, J. A. *Proceedings of the National Academy of Sciences of the United States of America* **1967**, *58*, 27–33.

(417) Jonas, J.; DeFries, T.; Wilbur, D. J. *Journal of Chemical Physics* **1976**, *65*, 582–588.

(418) Krynicki, K.; Green, C. D.; Sawyer, D. W. *Faraday Discussions of the Chemical Society* **1978**, *66*, 199–208.

(419) Van der Maarel, J. R. C.; Lankhorst, D.; de Bleijser, J.; Leyte, J. C. *Chemical Physics Letters* **1985**, *122*, 541–544.

(420) Struis, R. P.; De Bleijser, J.; Leyte, J. C. *Journal of Physical Chemistry* **1987**, *91*, 1639–1645.

(421) Price, W. S.; Ide, H.; Arata, Y. *Journal of Physical Chemistry A* **1999**, *103*, 448–450.

(422) Holz, M.; Heil, S. R.; Sacco, A. *Physical Chemistry Chemical Physics* **2000**, *2*, 4740–4742.

(423) Price, W. S.; Ide, H.; Arata, Y.; Söderman, O. *Journal of Physical Chemistry B* **2000**, *104*, 5874–5876.

(424) Lankhorst, D.; Schriever, J.; Leyte, J. C. *Berichte der Bunsengesellschaft für physikalische Chemie* **1982**, *86*, 215–221.

(425) Hardy, E. H.; Zygar, A.; Zeidler, M. D.; Holz, M.; Sacher, F. D. *Journal of Chemical Physics* **2001**, *114*, 3174–3181.

(426) Ropp, J.; Lawrence, C.; Farrar, T. C.; Skinner, J. L. *Journal of the American Chemical Society* **2001**, *123*, 8047–8052.

(427) Lawrence, C. P.; Skinner, J. L. *Journal of Chemical Physics* **2003**, *118*, 264–272.

(428) Skinner, L. B.; Huang, C.; Schlesinger, D.; Pettersson, L. G.; Nilsson, A.; Benmore, C. J. *Journal of Chemical Physics* **2013**, *138*, 074506.

(429) Skinner, L. B.; Benmore, C. J.; Neuefeind, J. C.; Parise, J. B. *Journal of Chemical Physics* **2014**, *141*, 214507.

(430) Soper, A. K. *Journal of Chemical Physics* **1994**, *101*, 6888–6901.

(431) Soper, A. K. *Chemical Physics* **2000**, *258*, 121–137.

(432) Soper, A. K. *ISRN Physical Chemistry* **2013**, *2013*, 1–67.

(433) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *Journal of Chemical Physics* **1983**, *79*, 926–935.

(434) Errington, J. R.; Debenedetti, P. G. *Nature* **2001**, *409*, 318–321.

(435) Matsumoto, M.; Saito, S.; Ohmine, I. *Nature* **2002**, *416*, 409–413.

(436) Teng, X.; Liu, B.; Ichiye, T. *Journal of Chemical Physics* **2020**, *153*, 104510.

(437) Car, R.; Parrinello, M. *Physical Review Letters* **1985**, *55*, 2471–2474.

(438) Payne, M. C.; Teter, M. P.; Allan, D. C.; Arias, T. A.; Joannopoulos, J. D. *Reviews of Modern Physics* **1992**, *64*, 1045–1097.

(439) Tuckerman, M. E. *Journal of Physics: Condensed Matter* **2002**, *14*, R1297.

(440) Bohm, D.; Pines, D. *Physical Review* **1951**, *82*, 625–634.

(441) Pines, D.; Bohm, D. *Physical Review* **1952**, *85*, 338–353.

(442) Bohm, D.; Pines, D. *Physical Review* **1953**, *92*, 609–625.

(443) Purvis, G. D.; Bartlett, R. J. *Journal of Chemical Physics* **1982**, *76*, 1910–1918.

(444) Löwdin, P.-O. *Physical Review* **1955**, *97*, 1474–1489.

(445) Perdew, J. P.; Zunger, A. *Physical Review B* **1981**, *23*, 5048–5079.

(446) Becke, A. D. *Physical Review A* **1988**, *38*, 3098–3100.

(447) Lee, C.; Yang, W.; Parr, R. G. *Physical Review B* **1988**, *37*, 785–789.

(448) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Physical Review Letters* **1996**, *77*, 3865–3868.

(449) Perdew, J. P.; Kurth, S.; Zupan, A.; Blaha, P. *Physical Review Letters* **1999**, *82*, 2544–2547.

(450) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Physical Review Letters* **2003**, *91*, 3–6.

(451) Zhao, Y.; Truhlar, D. G. *Journal of Chemical Physics* **2006**, *125*, 194101.

(452) Sun, J.; Remsing, R. C.; Zhang, Y.; Sun, Z.; Ruzsinszky, A.; Peng, H.; Yang, Z.; Paul, A.; Waghmare, U.; Wu, X.; Klein, M. L.; Perdew, J. P. *Nature Chemistry* **2016**, *8*, 831–836.

(453) Burke, K.; Ernzerhof, M.; Perdew, J. P. *Chemical Physics Letters* **1997**, *265*, 115–120.

(454) Adamo, C.; Barone, V. *Journal of Chemical Physics* **1999**, *110*, 6158–6170.

(455) Peverati, R.; Truhlar, D. G. *Philosophical Transactions of the Royal Society A* **2014**, *372*, 20120476.

(456) Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S. *Angewandte Chemie* **2022**, *134*, e202205735.

(457) Pestana, L. R.; Marsalek, O.; Markland, T. E.; Head-Gordon, T. *Journal of Physical Chemistry Letters* **2018**, *9*, 5009–5016.

(458) Rappoport, D.; Crawford, N. R. M.; Furche, F.; Burke, K., *Approximate Density Functionals: Which Should I Choose?*; John Wiley & Sons: New York, 2009.

(459) Grossman, J. C.; Schwegler, E.; Draeger, E. W.; Gygi, F.; Galli, G. *Journal of Chemical Physics* **2004**, *120*, 300–311.

(460) McGrath, M. J.; Siepmann, J. I.; Kuo, I. F. W.; Mundy, C. J.; Vandevondele, J.; Hutter, J.; Mohamed, F.; Krack, M. *ChemPhysChem* **2005**, *6*, 1894–1901.

(461) Schmidt, J.; VandeVondele, J.; Kuo, I.-F. W.; Sebastiani, D.; Siepmann, J. I.; Hutter, J.; Mundy, C. J. *Journal of Physical Chemistry B* **2009**, *113*, 11959–11964.

(462) Gillan, M. J.; Alfè, D.; Michaelides, A. *Journal of Chemical Physics* **2016**, *144*, 130901.

(463) Gaiduk, A. P.; Gygi, F.; Galli, G. *Journal of Physical Chemistry Letters* **2015**, *6*, 2902–2908.

(464) Sit, P. H.; Marzari, N. *Journal of Chemical Physics* **2005**, *122*, 204510.

(465) Todorova, T.; Seitsonen, A. P.; Hutter, J.; Kuo, I. F. W.; Mundy, C. J. *Journal of Physical Chemistry B* **2006**, *110*, 3685–3691.

(466) Lee, H. S.; Tuckerman, M. E. *Journal of Chemical Physics* **2007**, *126*, 164501.

(467) Ruiz Pestana, L.; Mardirossian, N.; Head-Gordon, M.; Head-Gordon, T. *Chemical Science* **2017**, *8*, 3554–3565.

(468) Wang, Y.; Jin, X.; Yu, H. S.; Truhlar, D. G.; He, X. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, *114*, 8487–8492.

(469) Peverati, R.; Truhlar, D. G. *Journal of Physical Chemistry Letters* **2012**, *3*, 117–124.

(470) Peverati, R.; Truhlar, D. G. *Physical Chemistry Chemical Physics* **2012**, *14*, 13171–13174.

(471) Yu, H. S.; He, X.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2016**, *12*, 1280–1293.

(472) Zhao, Y.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2008**, *4*, 1849–1868.

(473) Peverati, R.; Truhlar, D. G. *Journal of Physical Chemistry Letters* **2011**, *2*, 2810–2817.

(474) Peverati, R.; Truhlar, D. G. *Physical Chemistry Chemical Physics* **2012**, *14*, 16187–16191.

(475) Yu, H. S.; He, X.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2016**, *12*, 1280–1293.

(476) Marom, N.; Tkatchenko, A.; Rossi, M.; Gobre, V. V.; Hod, O.; Scheffler, M.; Kronik, L. *Journal of Chemical Theory and Computation* **2011**, *7*, 3944–3951.

(477) Klimeš, J.; Michaelides, A. *Journal of Chemical Physics* **2012**, *137*, 120901.

(478) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. *Nature Chemistry* **2009**, *1*, 37–46.

(479) Dorcier, A.; Dyson, P. J.; Gossens, C.; Rothlisberger, U.; Scopelliti, R.; Tavernelli, I. *Organometallics* **2005**, *24*, 2114–2123.

(480) Thapa, B.; Schlegel, H. B. *Journal of Physical Chemistry A* **2016**, *120*, 5726–5735.

(481) Röhrig, U. F.; Frank, I.; Hutter, J.; Laio, A.; VandeVondele, J.; Rothlisberger, U. *ChemPhysChem* **2003**, *4*, 1177–1182.

(482) Tkatchenko, A.; Romaner, L.; Hofmann, O. T.; Zojer, E.; Ambrosch-Draxl, C.; Scheffler, M. *MRS Bulletin* **2010**, *35*, 435–442.

(483) Distasio, R. A.; Santra, B.; Li, Z.; Wu, X.; Car, R. *Journal of Chemical Physics* **2014**, *141*, 084502.

(484) Yao, Y.; Kanai, Y. *Journal of Chemical Physics* **2020**, *153*, 044114.

(485) Zhang, C.; Tang, F.; Chen, M.; Xu, J.; Zhang, L.; Qiu, D. Y.; Perdew, J. P.; Klein, M. L.; Wu, X. *Journal of Physical Chemistry B* **2021**, *125*, 11444–11456.

(486) Mardirossian, N.; Head-Gordon, M. *Journal of Chemical Theory and Computation* **2016**, *12*, 4303–4325.

(487) Tuckerman, M. E.; Berne, B. J.; Martyna, G. J. *Journal of Chemical Physics* **1992**, *97*, 1990–2001.

(488) Tuckerman, M. E. In *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, Grotendorst, J., Marx, D., Muramatsu, A., Eds.; John von Neumann Institute for Computing: Jülich, 2002; Vol. 10, pp 269–298.

(489) Oliver, G. L.; Perdew, J. P. *Physical Review A* **1979**, *20*, 397–403.

(490) Hammer, B.; Hansen, L. B.; Nørskov, J. K. *Physical Review B* **1999**, *59*, 7413–7421.

(491) Peverati, R.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2012**, *8*, 2310–2319.

(492) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *Journal of Chemical Physics* **2005**, *123*, 161103.

(493) Van Voorhis, T.; Scuseria, G. E. *Journal of Chemical Physics* **1998**, *109*, 400–410.

(494) Toulouse, J.; Colonna, F.; Savin, A. *Physical Review A* **2004**, *70*, 62505.

(495) Troullier, N.; Martins, J. L. *Physical Review B* **1991**, *43*, 1993–2006.

(496) Li, C.; Paesani, F.; Voth, G. A. *Journal of Chemical Theory and Computation* **2022**, *18*, 2124–2131.

(497) Christensen, A. S.; Faber, F. A.; von Lilienfeld, O. A. *Journal of Chemical Physics* **2019**, *150*, 064105.

(498) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole Von Lilienfeld, O. *Journal of Chemical Physics* **2020**, *152*, 044107.

(499) Huang, B.; von Lilienfeld, O. A. *Nature Chemistry* **2020**, *12*, 945–951.

(500) Humphrey, W.; Dalke, A.; Schulten, K. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(501) Raiteri, P.; Laio, A.; Parrinello, M. *Physical Review Letters* **2004**, *93*, 6–9.

(502) Giorgino, T. *Journal of Open Source Software* **2019**, *4*, 1698.

(503) Yeh, I. C.; Hummer, G. *Journal of Physical Chemistry B* **2004**, *108*, 15873–15879.

(504) Lemmon, E. W.; Bell, I. H.; Huber, M. L.; McLinden, M. O. In *NIST Chemistry webbook, NIST Standard reference database*, Linstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg MD, 20899, USA, 1998; Vol. 69.

(505) Lee, K.; Yu, J.; Morikawa, Y. *Physical Review B - Condensed Matter and Materials Physics* **2007**, *75*, 1–5.

(506) Laage, D.; Hynes, J. T. *Journal of Physical Chemistry B* **2008**, *112*, 14230–14242.

(507) Wilkins, D. M.; Manolopoulos, D. E.; Pipolo, S.; Laage, D.; Hynes, J. T. *Journal of Physical Chemistry Letters* **2017**, *8*, 2602–2607.

(508) Miller, T. F.; Manolopoulos, D. E. *Journal of Chemical Physics* **2005**, *123*, 154504.

(509) Daru, J.; Forbert, H.; Behler, J.; Marx, D. *Physical Review Letters* **2022**, *129*, 226001.

(510) Smith, D. W. G.; Powles, J. G. *Molecular Physics* **1966**, *10*, 451–463.

(511) Sansom, M. S.; Kerr, I. D.; Breed, J.; Sankararamakrishnan, R. *Biophysical Journal* **1996**, *70*, 693–702.

(512) Bakker, H. J.; Rezus, Y. L.; Timmer, R. L. *Journal of Physical Chemistry A* **2008**, *112*, 11523–11534.

(513) Bakker, H. J.; Skinner, J. L. *Chemical Reviews* **2010**, *110*, 1498–1517.

(514) Goerigk, L. *Journal of Physical Chemistry Letters* **2015**, *6*, 3891–3896.

(515) Morrone, J. A.; Car, R. *Physical Review Letters* **2008**, *101*, 017801.

(516) Marsalek, O.; Markland, T. E. *Journal of Physical Chemistry Letters* **2017**, *8*, 1545–1551.

(517) Feynman, R. P.; Hibbs, A. R.; Styer, D. F., *Quantum Mechanics and Path Integrals*, Emended; Dover Books on Physics; Dover Publications: New York, 2010.

(518) Lan, J.; Wilkins, D. M.; Rybkin, V. V.; Iannuzzi, M.; Hutter, J. J. *ChemRxiv* **2021**, 10.26434/chemrxiv-2021-n32q8–v2.

(519) Soper, A. K.; Benmore, C. J. *Physical Review Letters* **2008**, *101*, 65502.

(520) https://water.lsbu.ac.uk/water/water_hydrogen_bonding.html, *Hydrogen Bonding in Water*; Chaplin, M.: 2023, April.

(521) Rastogi, A.; Ghosh, A. K.; Suresh, S. J. In *Thermodynamics – Physical Chemistry of Aqueous Systems*, Moreno-Pirajan, J. C., Ed., 1st ed.; IntechOpen: London, 2011; Chapter 13, pp 351–364.

(522) Modig, K.; Pfrommer, B. G.; Halle, B. *Physical Review Letters* **2003**, *90*, 075502.

(523) Mills, R. *Journal of Physical Chemistry* **1973**, *77*, 685–688.

(524) Easteal, A. J.; Price, W. E.; Woolf, L. A. *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases* **1989**, *85*, 1091–1097.

(525) Rahman, A.; Stillinger, F. H. *Journal of Chemical Physics* **1971**, *55*, 3336–3359.

(526) Dahlke, E. E.; Olson, R. M.; Leverentz, H. R.; Truhlar, D. G. *Journal of Physical Chemistry A* **2008**, *112*, 3976–3984.

(527) Leverentz, H. R.; Qi, H. W.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2013**, *9*, 995–1006.

(528) Nagornova, N. S.; Rlzzo, T. R.; Boyarkln, O. V. *Journal of the American Chemical Society* **2010**, *132*, 4040–4041.

(529) Whitford, D., *Proteins: Structure and Function*, 1st ed.; John Wiley & Sons Inc: New York, 2005.

(530) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128–132.

(531) Lee, D.; Redfern, O.; Orengo, C. *Nature Reviews Molecular Cell Biology* **2007**, *8*, 995–1005.

(532) Jumper, J. et al. *Nature* **2021**, *596*, 583–589.

(533) Dyson, H.; Wright, P. E. *Current Opinion in Structural Biology* **1993**, *3*, 60–65.

(534) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angewandte Chemie International Edition* **1999**, *38*, 236–240.

(535) Venkatraman, J.; Shankaramma, S. C.; Balaram, P. *Chemical Reviews* **2001**, *101*, 3131–3152.

(536) Thomas, A.; Deshayes, S.; Decaffmeyer, M.; Van Eyck, M. H.; Charloteaux, B.; Brasseur, R.; Sanchez, G. *Proteins: Structure, Function, and Bioinformatics* **2006**, *65*, 889–897.

(537) Zhang, S. *Nature Biotechnology* **2003**, *21*, 1171–1178.

(538) Maeda, Y.; Makhlynets, O. V.; Matsui, H.; Korendovych, I. V. *Annual Review of Biomedical Engineering* **2016**, *18*, 311–328.

(539) Fosgerau, K.; Hoffmann, T. *Drug Discovery Today* **2015**, *20*, 122–128.

(540) Bhandari, D.; Rafiq, S.; Gat, Y.; Gat, P.; Waghmare, R.; Kumar, V. *International Journal of Peptide Research and Therapeutics* **2020**, *26*, 139–150.

(541) Zasloff, M. *Nature* **2002**, *415*, 389–395.

(542) Brogden, K. A. *Nature Reviews Microbiology* **2005**, *3*, 238–250.

(543) Hancock, R. E.; Sahl, H. G. *Nature Biotechnology* **2006**, *24*, 1551–1557.

(544) Qvit, N.; Rubin, S. J.; Urban, T. J.; Mochly-Rosen, D.; Gros, E. R. *Drug Discovery Today* **2017**, *22*, 454–462.

(545) Lee, A. C. L.; Harris, J. L.; Khanna, K. K.; Hong, J. H. *International Journal of Molecular Sciences* **2019**, *20*, 1–21.

(546) Hoaglund-Hyzer, C. S.; Counterman, A. E.; Clemmer, D. E. *Chemical Reviews* **1999**, *99*, 3037–3079.

(547) Wyttenbach, T.; Bowers, M. T. *Journal of Physical Chemistry B* **2011**, *115*, 12266–12275.

(548) Scutelnic, V.; Perez, M. A.; Marianski, M.; Warnke, S.; Gregor, A.; Rothlisberger, U.; Bowers, M. T.; Baldauf, C.; Von Helden, G.; Rizzo, T. R.; Seo, J. *Journal of the American Chemical Society* **2018**, *140*, 7554–7560.

(549) Hünig, I.; Kleinermanns, K. *Physical Chemistry Chemical Physics* **2004**, *6*, 2650–2658.

(550) Bakker, J. M.; Plutzer, C.; Hünig, I.; Häber, T.; Compagnon, I.; Von Helden, G.; Meijer, G.; Kleinermanns, K. *ChemPhysChem* **2005**, *6*, 120–128.

(551) Häber, T.; Seefeld, K.; Kleinermanns, K. *Journal of Physical Chemistry A* **2007**, *111*, 3038–3046.

(552) Nagornova, N. S.; Guglielmi, M.; Doemer, M.; Tavernelli, I.; Rothlisberger, U.; Rizzo, T. R.; Boyarkin, O. V. *Angewandte Chemie - International Edition* **2011**, *50*, 5383–5386.

(553) Gloaguen, E.; Mons, M.; Schwing, K.; Gerhards, M. *Chemical Reviews* **2020**, *120*, 12490–12562.

(554) Schubert, F.; Rossi, M.; Baldauf, C.; Pagel, K.; Warnke, S.; von Helden, G.; Filsinger, F.; Kupser, P.; Meijer, G.; Salwiczek, M.; Koksch, B.; Scheffler, M.; Blum, V. *Physical Chemistry Chemical Physics* **2015**, *17*, 7373–7385.

(555) Hao, G. F.; Xu, W. F.; Yang, S. G.; Yang, G. F. *Scientific Reports* **2015**, *5*, 1–10.

(556) Rossi, M.; Chutia, S.; Scheffler, M.; Blum, V. *Journal of Physical Chemistry A* **2014**, *118*, 7349–7359.

(557) Damsbo, M.; Kinnear, B. S.; Hartings, M. R.; Ruhoff, P. T.; Jarrold, M. F.; Ratner, M. A. *Proceedings of the National Academy of Sciences* **2004**, *101*, 7215–7222.

(558) Doemer, M.; Guglielmi, M.; Athri, P.; Nagornova, N. S.; Rizzo, T. R.; Boyarkin, O. V.; Tavernelli, I.; Rothlisberger, U. *International Journal of Quantum Chemistry* **2013**, *113*, 808–814.

(559) Deaven, D. M.; Ho, K. M. *Physical Review Letters* **1995**, *75*, 288–291.

(560) Unger, R.; Moult, J. *Journal of Molecular Biology* **1993**, *231*, 75–81.

(561) Pedersen, J. T.; Moult, J. *Current Opinion in Structural Biology* **1996**, *6*, 227–231.

(562) Unger, R. In *Applications of Evolutionary Computation in Chemistry*, 1st ed.; Springer: Berlin, 2004; Vol. 110, pp 153–175.

(563) Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M.; Peterson, M. L. *Journal of Computational Chemistry* **1993**, *14*, 1407–1414.

(564) Nair, N.; Goodman, J. M. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 317–320.

(565) Vainio, M. J.; Johnson, M. S. *Journal of Chemical Information and Modeling* **2007**, *47*, 2462–2474.

(566) Mcgarrah, D. B.; Judson, R. S. *Journal of Computational Chemistry* **1993**, *14*, 1385–1395.

(567) Herrmann, F.; Suhai, S. *Journal of Computational Chemistry* **1995**, *16*, 1434–1444.

(568) Supady, A.; Blum, V.; Baldauf, C. *Journal of Chemical Information and Modeling* **2015**, *55*, 2338–2348.

(569) Pedersen, J. T.; Moult, J. *Proteins: Structure, Function, and Bioinformatics* **1995**, *23*, 454–460.

(570) Pedersen, J. T.; Moult, J. *Proteins: Structure, Function, and Bioinformatics* **1997**, *29*, 179–184.

(571) Pedersen, J. T.; Moult, J. *Journal of Molecular Biology* **1997**, *269*, 240–259.

(572) Le Grand, S. M.; Merz, K. M. *Molecular Simulation* **1994**, *13*, 299–320.

(573) Le Grand, S. M.; Merz, K. M. *Journal of Global Optimization* **1993**, *3*, 49–66.

(574) Schulze-Kremer, S. In *Protein Structure Prediction: Methods and Protocols*, Webster, D., Ed.; Methods in Molecular Biology; Humana Press: Totowa, 2000; Chapter 9, pp 175–222.

(575) Mijajlovic, M.; Biggs, M. J.; Djurdjevic, D. P. *Evolutionary Computation* **2010**, *18*, 255–275.

(576) Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S. J.; Söding, J. *BMC bioinformatics* **2019**, *20*, 473.

(577) Dor, O.; Zhou, Y. *Proteins: Structure, Function, and Bioinformatics* **2007**, *66*, 838–845.

(578) Sitkiewicz, S. P.; Zaleśny, R.; Ramos-Cordoba, E.; Luis, J. M.; Matito, E. *Journal of Physical Chemistry Letters* **2022**, *13*, 5963–5968.

(579) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Research* **2000**, *28*, 235–242.

(580) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *Journal of Physical Chemistry* **1994**, *98*, 11623–11627.

(581) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Structure, Function, and Bioinformatics* **2006**, *65*, 712–725.

(582) Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. *Journal of Computational Chemistry* **2006**, *27*, 781–790.

(583) LCBC-EPFL Releases of EVOLVE will be available on GitHub, 2023.

(584) Brunk, E.; Perez, M. A.; Athri, P.; Rothlisberger, U. *ChemPhysChem* **2016**, *17*, 3831–3835.

(585) Bozkurt, E.; Perez, M. A.; Hovius, R.; Browning, N. J.; Rothlisberger, U. *Journal of the American Chemical Society* **2018**, *140*, 4517–4521.

(586) Browning, N. J.; Ramakrishnan, R.; von Lilienfeld, O. A.; Roethlisberger, U. *Journal of Physical Chemistry Letters* **2017**, *8*, 1351–1359.

(587) Brain, Z. E.; Addicoat, M. A. *Journal of Chemical Physics* **2011**, *135*, 174106.

(588) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. *Journal of Molecular Biology* **1963**, *7*, 95–99.

(589) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *Journal of Cheminformatics* **2011**, *3*, 33.

(590) Case, D. A. et al. AMBER 2018, University of California, San Francisco, 2018.

(591) Hjorth Larsen, A. et al. *Journal of Physics: Condensed Matter* **2017**, *29*, 273002.

(592) Deb, K.; Agrawal, R. B. *Complex Systems* **1995**, *9*, 115–148.

(593) Beyer, H.-G.; Schwefel, H.-P. *Natural Computing* **2002**, *1*, 3–52.

(594) Ahn, C. W.; Ramakrishna, R. S. *IEEE Transactions on Evolutionary Computation* **2003**, *7*, 367–385.

(595) Du, H.; Wang, Z.; Zhan, W.; Guo, J. *IEEE Access* **2018**, *6*, 44531–44541.

(596) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

(597) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *Journal of Physical Chemistry* **1993**, *97*, 10269–10280.

(598) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *Journal of Computational Chemistry* **2000**, *21*, 132–146.

(599) Jakalian, A.; Jack, D. B.; Bayly, C. I. *Journal of Computational Chemistry* **2002**, *23*, 1623–1641.

(600) Schaftenaar, G.; Vlieg, E.; Vriend, G. *Journal of Computer-Aided Molecular Design* **2017**, *31*, 789–800.

(601) Llamas-Saiz, A. L.; Grotenbreg, G. M.; Overhand, M.; Van Raaij, M. J. *Acta Crystallographica Section D: Biological Crystallography* **2007**, *63*, 401–407.

(602) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. *Journal of Chemical Theory and Computation* **2013**, *9*, 4046–4063.

(603) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Physical Review B* **1998**, *58*, 7260–7268.

(604) Gaus, M.; Cui, Q.; Elstner, M. *Journal of Chemical Theory and Computation* **2011**, *7*, 931–948.

(605) Hourahine, B. et al. *Journal of Chemical Physics* **2020**, *152*, 124101.

(606) Gaus, M.; Goez, A.; Elstner, M. *Journal of Chemical Theory and Computation* **2013**, *9*, 338–354.

(607) Brandenburg, J. G.; Grimme, S. *Journal of Physical Chemistry Letters* **2014**, *5*, 1785–1789.

(608) Grimme, S.; Ehrlich, S.; Goerigk, L. *Journal of Computational Chemistry* **2011**, *32*, 1456–1465.

(609) Kästner, J.; Carr, J. M.; Keal, T. W.; Thiel, W.; Wander, A.; Sherwood, P. *Journal of Physical Chemistry A* **2009**, *113*, 11856–11865.

(610) Stewart, J. J. *Journal of Molecular Modeling* **2007**, *13*, 1173–1213.

(611) Stewart, J. J. *Journal of Molecular Modeling* **2013**, *19*, 1–32.

(612) Li, X.; Frisch, M. J. *Journal of Chemical Theory and Computation* **2006**, *2*, 835–839.

(613) Cersonsky, R. K.; Helfrecht, B. A.; Engel, E. A.; Kliavinek, S.; Ceriotti, M. *Machine Learning: Science and Technology* **2021**, *2*, 35038.

(614) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. *Journal of the American Chemical Society* **2006**, *128*, 13435–13441.

(615) Mancini, G.; Fusè, M.; Lazzari, F.; Barone, V. *Digital Discovery* **2022**, *1*, 790–805.

(616) Chmiela, S.; Sauceda, H. E.; Poltavsky, I.; Müller, K. R.; Tkatchenko, A. *Computer Physics Communications* **2019**, *240*, 38–45.

(617) Kramer, O., *Machine Learning for Evolution Strategies*, 1st ed.; Springer: Cham, 2016.

(618) Jørgensen, M. S.; Larsen, U. F.; Jacobsen, K. W.; Hammer, B. *Journal of Physical Chemistry A* **2018**, *122*, 1504–1509.

(619) Rechenberg, I., *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*; Frommann-Holzboog: Stuttgart, 1973.

(620) Mancini, G.; Fusè, M.; Lazzari, F.; Chandramouli, B.; Barone, V. *Journal of Chemical Physics* **2020**, *153*, 124110.

(621) Jørgensen, M. S.; Groves, M. N.; Hammer, B. *Journal of Chemical Theory and Computation* **2017**, *13*, 1486–1493.

(622) Raschka, S.; Mirjalili, V., *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 3rd ed.; Packt Publishing: Birmingham, 2019.

(623) Pedregosa, F.; Varoquaux, G.; Buitinck, L.; Louppe, G.; Grisel, O.; Mueller, A. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(624) Ward, J. H. *Journal of the American Statistical Association* **1963**, *58*, 236–244.

(625) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. *Journal of Physical Chemistry Letters* **2015**, *6*, 2326–2331.

(626) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; von Lilienfeld, O. A. *arXiv* **2019**, 1909.01946v2.

(627) Jolliffe, I. T.; Cadima, J. *Philosophical Transactions of the Royal Society A* **2016**, *374*, 20150202.

Justin Villard

Contact

justinvillard@hotmail.com

ORCID 0000-0003-4606-319X

justin-villard

31 years old
Swiss, single

Education

2023 **Doctor of Sciences PhD in Physics**, *Ecole Polytechnique Fédérale de Lausanne (EPFL)*, *expected in August*.

Thesis in the Laboratory of Computational Chemistry and Biochemistry, Prof. Ursula Röthlisberger:

Advancing computational chemistry with stochastic and artificial intelligence approaches

Work in a group specialized in computational tools for the investigation of chemical phenomena at the atomic scale, focusing on solar cells and drug development.

- Programming of innovative methods in a high-performance parallel computer code to accelerate and improve the quantum description of electrons in matter;
- Design and implementation of algorithms belonging to artificial intelligence (AI), coupled to unsupervised machine learning (ML), for the search of biological molecules;
- Application of a kernel-based ML model to accelerate the dynamical simulation of chemical processes;

↪ Organized a conference to bring together researchers from computational chemistry with those from AI/ML;

↪ Attended several conferences and workshops related to applications of ML to chemistry and materials science;

↪ Teaching assistant for courses on electronic structure methods, molecular dynamics and Monte Carlo simulations, and solid-state physics; supervisor of a student project aiming to enhance quantum calculations with ML; popularized presentations of computational chemistry for high school students.

2017 **Master of Science MSc in Applied Physics**, *EPFL*, Lausanne.

Thesis in the Chair of Condensed Matter Theory, Prof. Frédéric Mila:

Theoretical study of a partially decorated kagome antiferromagnet inspired by $Cu_2O(SO_4)$

Theoretical and numerical investigation of magnetic phases in a crystal with complex spin structure.

↪ Results argue in favour of different phases made of decoupled spin chains or a potential spin liquid.

2014 **Bachelor of Science BSc in Physics**, *EPFL*, Lausanne.

2010 **Maturity certificate, Baccalaureat** (high-school degree), *Gymnase de Beaulieu*, Lausanne.

Specialization in physics and applied mathematics, complementary specialization in chemistry.

↪ Maturity project on particle physics at CERN passed with honours;

↪ *Merit Award.*

Professional experience

Vocational

2017–2018 **R&D support specialist, remote working student**,
ABB Switzerland, Corporate Research Center, Baden-Dättwil, (part-time, 10%).
Computer-aided design studies and optimization of semiconductor modules for the automotive and solar industry. Electromagnetic and thermal simulations.

↪ Characterization of a power converter in current product development stage.

2016–2017 **Multiphysics simulation of advanced power semiconductor modules**,
ABB Switzerland, Corporate Research Center, Baden-Dättwil, (internship, 6 months).
Design study of power inverters for next hybrid/electric car generations using cutting-edge semiconductors.

- Computer design and assessment of novel power semiconductor layouts according to design guidelines, defined figures of merit and cost aspects;
- Electromagnetic simulations to investigate and reduce parasitic effects during module switching;
- Thermal simulations of power modules to investigate power handling capability;
- Data analysis for the benchmark and recommendation of most promising modules;

↪ Some of the layouts designed have shown better performance than the current products;
↪ Results helped an automotive partner to make decisions on future developments;
↪ Brought in an idea that has been part of an invention disclosure and a patent application.

2014 **Test of radio frequency devices**,
Swiss Plasma Center, Lausanne, (internship, 2 months).
Electromagnetic waves in the millimiter wavelength domain are used for diagnostics of plasmas for industrial applications or fusion energy production.

↪ Set up a test bench and checked the reliability of state-of-the-art components (diodes, attenuators, amplifiers, converters, detectors) in this very specific wavelength domain;
↪ Substantial savings were made by recycling tested components for a value of about hundreds of thousand dollars on the test bench.

Miscellaneous

2012–2017 **Tutor for apprentices**,
Groupement pour l'Apprentissage, Echallens, Morges, Nyon, (2h/week).
Responsible for a group of 15-20 year-old apprentices with difficulties in mathematics and physics.

2013–2016 **Participation in the development of Massive Online Open Courses (MOOC)**,
EPFL, Lausanne, (4h/week).
Online courses of mechanics and thermodynamics of Prof. Jean-Philippe Ansermet.

- Conception and management of self-correcting exercises, supervision of teaching assistants, answer and help on the forum.

Specific skills

knowledge Computational chemistry, electronic structure theory, molecular dynamics, solid state physics, genetic algorithms, machine learning, statistics, data analysis, magnetic systems, quantum mechanics, quantum information, statistical mechanics

programming Python, C++, Fortran, git, bash, High-Performance Computing: MPI, OpenMP, slurm

molecular simulation CPMD, CP2K, TeraChem, Quantum Espresso, OpenMM, Amber, Tinker, Gromacs, GULP, Molden, Avogadro, OpenBabel, VMD

electronic structure	Gaussian, PySCF, Turbomole, Orca, Molpro
machine learning	Python: scikit-learn, PyTorch, pandas, NumPy, SciPy, matplotlib, seaborn, Jupyter Notebook
engineering	Matlab, Mathematica, COMSOL Multiphysics, Ansys Electronics Desktop, Q3D
editing	Microsoft Office, Google Docs, LibreOffice, Latex, vi, Emacs, Sublime Text, Visual Studio Code
os	Linux, Windows, MacOS X

Languages

French	**Mother tongue**	*C2*
English	**Fluent**, Cambridge Certificate in Advanced English (CAE), grade A (2015)	*C1*
German	**High-school knowledge**	*B2*

Interests

politics	Former town councillor of Daillens (2010-2020) and member of the finance commission (2016-2020)
community life	Former member (2007-2014), treasurer (2009-2010) and president (2010-2011) of the village youth society of Daillens. Former committee member of the local brass band of Penthalaz-Daillens (2017-2022)
music	Trumpet, holder of upper level certificates of studies, military recruit school accomplished within the Swiss Military Music. Player in the local brass band of Penthalaz-Daillens (2002-)
sports	Football, tennis, biking, skiing, hiking

Publications

2023 **Justin Villard**, Ursula Rothlisberger. Enhanced screening of low-energy peptide structures with genetic algorithms and clustering-based novelty search. *In preparation.*

Justin Villard, Ursula Rothlisberger. Structure and dynamics of liquid water from ab initio simulations: Adding Minnesota density functionals to Jacob's ladder. *In preparation.*

Justin Villard, Martin P. Bircher, Ursula Rothlisberger. Plane waves versus correlation-consistent basis sets: A comparison of MP2 non-covalent interaction energies in the complete basis set limit. *ChemRxiv*, 10.26434/chemrxiv-2023-203z9. *To be submitted.*

Justin Villard, Murat Kılıç, Ursula Rothlisberger. Surrogate Based Genetic Algorithm Method for Efficient Identification of Low-Energy Peptide Structures. *Journal of Chemical Theory and Computation*, 19, 3, 1080–1097.

2022 François Mouvet, **Justin Villard**, Viacheslav Bolnykh, Ursula Rothlisberger. Recent Advances in First-Principles Based Molecular Dynamics. *Accounts of Chemical Research*, 55, 3, 221-230.

2020 Martin P. Bircher, **Justin Villard** (co-first author), Ursula Rothlisberger. Efficient Treatment of Correlation Energies at the Basis-Set Limit by Monte Carlo Summation of Continuum States. *Journal of Chemical Theory and Computation*, 16, 10, 6550-6559.