EPFL

Thèse n° 10 799

# Engineering novel protein interactions with therapeutic potential using deep learning-guided surface design

Présentée le 2 mai 2024

Faculté des sciences et techniques de l'ingénieur
Laboratoire de conception de protéines et d'immuno-ingénierie
Programme doctoral en biotechnologie et génie biologique

pour l'obtention du grade de Docteur ès Sciences

par

**Anthony MARCHAND**

Acceptée sur proposition du jury

Prof. E. Oricchio, présidente du jury
Prof. B. E. Ferreira De Sousa Correia, directeur de thèse
Prof. E. Levy, rapporteur
Prof. S. Fleishman, rapporteur
Prof. M. Dal Peraro, rapporteur

École polytechnique fédérale de Lausanne

2024

# Acknowledgement

The present work would not have been possible without the exceptional support from several people within and outside the lab.

First, I am very grateful to my thesis director and supervisor Prof. Dr. Bruno Correia. Thank you for affording me the opportunity to be a part of your research group and for serving as an encouraging and supportive mentor in all tasks and initiatives. I appreciated your way to think “out-of-box” and to challenge me to explore scientific realms beyond my comfort zone. You have always been trying to reassure me and to give the confidence and positive energy to achieve my projects. I am also appreciative of the autonomy and freedom you granted for research, enabling the acquisition of new knowledge and broadening our horizons beyond the laboratory.

Next, I would like to extend my gratitude to all my colleagues who provided support and shared a part of this incredible journey. A special acknowledgment goes to Dr. Pablo Gainza, who has been a wonderful scientist and incredibly supportive in boosting my confidence with computational protein modeling and design. I am sincerely thankful for the valuable scientific insights and the computational groundwork, without which this thesis would not have been possible. I also want to express my appreciation to my colleagues and co-authors who collaborated on my projects, namely Dr. Alexandra Van Hall-Beauvais, Dr. Sarah Wehrle, Dr. Andreas Scheck, Lucia Bonati, Stephen Buckley, and Arne Schneuing. Also a big thank to Dr. Leo Scheller and Dr. Martin Pacesa for their help and scientific advices on multiple projects. To all of them, thank you for the amazing collaborations and teamwork over the last four years.

Furthermore, I would like to thank our outstanding laboratory technicians: Sandrine Georgeon, Stéphane Rosset, and Joseph Schmidt. Your invaluable support, technical guidance, and efforts in organizing the lab have played a crucial role in facilitating research within our group. None of this would be possible without your exceptional work. I would also like to express gratitude to all our collaborators and EPFL facilities for their invaluable assistance, with special acknowledgment to Dr. Kelvin Lau for his advice and expertise in the field of proteins.

Last but certainly not least, my heartfelt thanks go to my partner and close friends who have been very supportive over the last four years and always present to talk about the successful moments or the failed experiments. Thank you for bringing the necessary resilience and motivation. A special appreciation is extended to my parents, who have constantly encouraged me throughout my studies and any life project I've undertaken. I dedicate this work to those who are no longer with us to witness it, and I hope that the scientific advances presented here will enable others to cherish longer moments with their loved ones.

# Abstract

Proteins are foundational biomolecules of life playing a crucial role in a myriad of biological processes. Their function often requires interplay with other biomolecules, including proteins themselves. Protein-protein interactions (PPIs) are essential for maintaining cell homeostasis, but are also involved in the progression of several diseases, being pathogenic, neuro-degenerative or cancer related. Therefore, PPI engineering has always been at the basis of several protein-based therapeutics and other biotechnology tools. However, most PPI engineering strategies so far relied on extensive experimental optimization or computational tools that depend on prior knowledge. Indeed, challenges remain for protein targets where no structural or experimental data are available, or for interfaces that involve non-protein components such as small molecules.

To explore and address these limitations, this work aims to leverage machine-learning and physics-based methods for the design of *de novo* protein interactions with therapeutic potential, that will ultimately be characterized and validated with established laboratory techniques.

The first part of this thesis showcases the translational capabilities of PPI designs. For this purpose, we rationally designed switchable protein-based therapeutics by integrating a previously established chemically-disruptable heterodimer (CDH). To optimize this OFF-switch system, we employed *in silico* methods based on physics-driven predictions, followed by rigorous *in vitro* validations to enhance its switchability in solution. This resulted in the development of a protein therapeutic exhibiting significantly improved drug-based controllability in mice models.

Nevertheless, most antibody and protein therapeutics discovered using experimental methods are agnostic to where and how these proteins engage their respective target. Despite recent advances, predicting an amino acid sequence that binds to a specific interface remains a major challenge for the field. To address this, a geometric deep learning framework, called MaSIF, was developed in our group to predict PPI interfaces and their corresponding binding partners based solely on the vectorized geometric and chemical features of the protein surface, also known as “fingerprints”. In this work, we improved MaSIF by leveraging a database of small binding motifs to design novel protein binders for four therapeutically relevant targets. All protein binders were validated experimentally and reached native-like affinities after pure *in silico* generation.

Finally, we generalized our framework to design drug-bound protein complexes via the formation of neosurfaces that arise upon small molecule binding. The versatility of our approach allowed us to computationally design and experimentally validate binders against three small molecule-protein complexes. All designs exhibited drug-dependent binding with native-like affinities and were functionalized as ON-switch systems for different cell-based applications.

Altogether, this dissertation provides new insights for the design of site-specific *de novo* protein interactions and their potential implementation in therapies by using innovative computational tools. On top of improving our understanding of PPI design, this work represents a new avenue for the development of biotechnology tools with concrete applications that can benefit patients.

**Keywords**: protein design, protein-protein interactions, protein engineering, machine learning, protein switch, protein therapeutics, computational modeling

# Résumé

Les protéines sont des biomolécules de la Vie cruciales dans une myriade de procédés biologiques. Les interactions protéines-protéines (IPP) sont essentielles pour maintenir l'homéostasie cellulaire, mais sont aussi impliquées dans le développement de maladies. C'est pourquoi l'ingénierie d'IPP a toujours été à la base de plusieurs thérapies protéiques et d'autres biotechnologies. Cependant, la plupart des stratégies d'ingénierie dépendent de profondes optimisations expérimentales ou des outils computationnels qui se basent sur des connaissances précédentes. Des défis persistent pour des cibles protéiques pour lesquels aucune donnée expérimentale n'est disponible ou pour l'implication de composants non-protéiques.

Pour répondre à cela, ce travail a pour but de tirer profit de méthodes basées sur la physique et l'apprentissage machine pour la conception d'interactions protéiques *de novo* avec un potentiel thérapeutique qui seront validées avec des techniques de laboratoire reconnues.

La première partie de cette thèse présente les capacités transversales des conceptions d'IPP. Pour ce faire, nous avons rationnellement conçu une thérapie protéique interruptible en intégrant un hétéro-dimère chimiquement perturbé. Pour optimiser ce système d'interrupteur désactivable (OFF), nous avons employé une méthode *in silico* basées sur la physique, suivie d'une validation *in vitro* pour augmenter la capacité d'interrupteur en solution. Cela résulta dans le développement d'une thérapie protéique démontrant une contrôlabilité augmentée dans des modèles de souris.

Cependant, la plupart des protéines thérapeutiques découvertes en utilisant des méthodes expérimentales sont agnostiques de l'endroit et la façon dont elles engagent leur cible. Malgré les avancées, prédire une séquence d'acides aminés se liant à une interface spécifique reste un défi. Pour répondre à cela, un outil basé sur l'apprentissage profond, appelé MaSIF, a été développé dans notre groupe afin de prédire les interfaces d'IPP et leur partenaire de liaison en se basant sur les propriétés géométriques et chimiques de la surface des protéines, aussi appelées « empreintes ». Dans ce travail, nous avons amélioré MaSIF grâce à une base de données de petits motifs de liaisons pour ensuite concevoir de nouveaux ligands protéiques pour quatre cibles d'importance thérapeutique. Tous ces ligands ont été validés expérimentalement et ont atteint des affinités proches des interactions natives après seulement leur génération *in silico*.

Enfin, nous avons généralisé notre outil pour concevoir des complexes protéine-médicament via la formation de néo-surfaces qui apparaissent lors de la liaison à de petites molécules. Cette polyvalence a permis de concevoir de manière computationnelle et de valider des ligands protéiques ciblant trois complexes protéines-médicaments. Toutes les conceptions ont présenté une liaison dépendante du médicament avec des affinités similaires à celles naturelles et ont été fonctionnalisées en tant que systèmes d'interrupteur activable (ON) pour des applications cellulaires.

En somme, cette dissertation procure de nouvelles idées pour la conception *de novo* d'interactions protéines pour des sites spécifiques, ainsi que leur implémentation dans des thérapies en utilisant des outils computationnels innovants. En plus d'amener une meilleure compréhension, ce travail

représente une nouvelle piste pour le développement d'outils biotechnologiques avec des applications concrètes.

**Mots-clés** : Conception de protéines, interactions protéine-protéine, ingénierie de protéines, apprentissage machine, interrupteur protéique, protéine thérapeutique, modélisation computationnelle

# Table of contents

# List of figures

# List of supplementary figures

# List of tables

# List of supplementary tables

# List of abbreviations

| | |
|---|---|
| Å | Angstrom |
| ACE2 | Angiotensin-converting enzyme 2 |
| AF2 | AlphaFold 2 |
| Bcl-2 | B-cell lymphoma 2 |
| Bcl-$X_L$ | B-cell lymphoma-extra large |
| BLI | Bio-layer interferometry |
| $C_\alpha$ | Carbon alpha |
| CAR | Chimeric antigen receptor |
| CD | Circular dichroism |
| CDH | Chemically-disruptable heterodimer |
| CID | Chemically-induced dimerization |
| Cryo-EM | Cryo-electro microscopy |
| CTLA-4 | Cytotoxic T-Lymphocyte Antigen 4 |
| $\Delta\Delta G$ | Change of binding free energy |
| DNA | Deoxyribonucleic acid |
| $EC_{50}$ | Half maximal effective concentration |
| ELISA | Enzyme-linked immunosorbent assay |
| Fab | Fragment antigen-binding region |
| FACS | Fluorescence-activated cell sorting |
| Fc | Fragment crystallizable region |
| GAN | Generative adversial network |
| GEMS | Generalized extracellular molecule sensor |
| GNN | Graph neural network |
| HER2 | Human Epidermal Growth Factor Receptor-2 |
| IL | Interleukine |
| $K_D$ | Constant of dissociation |
| MaSIF | Molecular surface interaction fingerprinting |
| MFI | Mean fluorescence intensity |
| MPNN | Message passing neural networks |
| PDB | Protein Data Bank |
| PD-1 | Programmed cell death protein 1 |
| PDF1 | Peptide deformylase 1 |
| PD-L1 | Programmed death ligand 1 |
| PPI | Protein-protein interaction |
| RBD | Receptor binding domain |
| RNA | Ribonucleic acid |
| R.E.U. | Rosetta energy unit |
| RMSD | Root mean square deviation |
| RU | Response unit |
| SA | Super agonist |
| ScFv | Single-chain variable fragment |
| SPR | Surface plasmon resonance |
| SEC-MALS | Size exclusion chromatography multi-angle light scattering |
| SSM | Site-saturation mutagenesis |
| VAE | Variational autoencoder |
| WT | Wild type |

# Chapter 1

## Introduction

Every house, every structure – being man-made or from Nature itself – is initiated from different kinds of building blocks. Each living cell representing the unit of Life is composed of four different basic elements: fatty acids, carbohydrates, nucleotides, and amino acids. These construction elements will respectively form lipids, sugar, nucleic acids, and proteins, which in turn can form a cell, a tissue and ultimately a living organism. Proteins play a central role in this process as they can fulfill a myriad of functions on their own or when they interact with other proteins, a process called protein-protein interactions (PPIs). Nowadays, thanks to the expansion of computational capabilities, scientists can predict, model and design proteins on a computer screen, and even interactions between proteins, which reflects one of the numerous ways to control Life. Among the tools available, artificial intelligence, and especially machine learning, became an indispensable asset for protein engineers to achieve this aim.

In this journey through the protein universe and beyond, we will first introduce the biochemistry of proteins and their role in Life. Then we will look more deeply into how their mission is fulfilled by interacting with other protein partners and simultaneously present the recent advances in terms of computational protein design and engineering of novel protein-protein interactions using more classical approaches as well as recent deep learning tools.

### 1.1 Biochemistry of proteins

Proteins belong to the most diverse and versatile group of molecules of Life in terms of function, biophysical properties, and complexity. Proteins are complex structures that require four elements to exist: i) a construction plan, ii) some building blocks, iii) a shape, and iv) a function.

Firstly, every construction starts with a project blueprint, which in the case of proteins is encoded in our genetic material composed of deoxyribonucleic acids (DNA). The human genome encodes slightly less than 20'000 genes and as many proteins [1]. The DNA is then transcribed into a messenger ribonucleic acid (mRNA) whose role is to deliver the construction plan to the protein factory within our cells, also called ribosomes. The mRNA is ultimately translated into a polymer called "protein".

Scientists often refer to this sequential process as the “central dogma of molecular biology” [2], which was first described by Francis Crick, one of the first scientists to unravel the helical structure of DNA.

Proteins are built by the ribosomes from a set of 20 amino acids as building blocks. Each amino acid consists of a common architecture, also referred to as backbone, and a more variable region that is specific to each amino acid, also called side chain (Figure 1.1A-B). The backbone of amino acids starts with an amine group (N-terminus) and ends with a carboxyl group (C-terminus). In the middle, the central carbon ($C_\alpha$) can host different side chains that will bring particular features to each of the 20 amino acids (Figure 1.1B). Indeed, these side chains can differ in size, polarity, and charges. For this reason, amino acids have been grouped in different categories, namely hydrophobic, polar uncharged, positively charged, negatively charged and a couple special cases (proline and cysteines) (Figure 1.1A).

Amino acids are comparable to a pearl necklace: they can be attached one after each other, thanks to a peptide bond formation between a carboxyl group and an amine group of another amino acid, and so on (Figure 1.1B). This linear chain of amino acid is also called the “primary structure” of proteins. Each bond between the atoms along this chain can rotate, either on the C-terminal side ($\psi$), the N-terminal side ($\phi$) or at the peptide bond itself ($\omega$). Of note, only certain torsion angles will be energetically favorable, which constrains the infinite amount of folding possibilities to certain regions only, as exemplified by the Ramachandran plot [3] (Figure 1.1C). The polypeptide chain will exploit these torsion angles to fold into a three dimensional structure called a protein.

Polar components of the amino acid backbone, namely the amine hydrogen and the carboxyl oxygen that are proton donors and acceptors respectively, can form local or non-local hydrogen bonds with their respective partner from another amino acid. This process will give rise to “secondary structures” which are local geometric elements such as alpha helices and beta sheets (Figure 1.1C). These repeated elements can be connected by flexible loops and will play a crucial role in the protein packing and folding.

Further energy minimization will aim to reduce the contacts between hydrophobic residues and the water molecules, which are polar. Consequently, during the folding process, the protein will adopt a three-dimensional structure that shields the hydrophobic residues from the solvent. This resulting 3D arrangement is commonly referred to as the protein's “tertiary structure” (Figure 1.1C). While most proteins may be already functional as a single monomer, others need to form a “quaternary structure” and undergo homo- or hetero-oligomerization with other protein subunits to fulfill their function, like antibodies or hemoglobin for example.

Protein folding is one of the most important processes of protein biochemistry, as folding will attribute a specific function to a protein. The Anfinsen principle, or thermodynamic hypothesis, stipulates that the three dimensional structure of a native protein is found where the Gibbs free energy of the whole system is at its lowest [4]. From that postulate, it became evident that the linear amino acid sequence will define the protein structure, and in turn its structure will define its function. To reach this energy minimum, various driving forces are involved. The most dominant contribution comes from the hydrophobic interaction [5] which aims to shield the hydrophobic residues away from the surrounding aqueous solvent

by enclosing them within an internal hydrophobic core[1]. Consequently, hydrophilic amino acids will face the solvent to form hydrogen bonds with water molecules. To properly bury all hydrophobic residues within the protein core, specific backbone torsions will be provided by secondary structure elements. Hydrogen bonds formed within secondary structures also greatly enhance protein stability, therefore secondary structure can be seen as both a cause and a consequence of the protein tertiary structure [6,7].

Once these driving forces sampled possible paths within the funnel-shaped energy landscape and lead to a global minimum, the protein will have a structure that will subsequently define its cellular function [8]. One major category of proteins are enzymes which can fulfill a catalytic function and facilitate chemical reactions. Proteins can also have a structural (e.g. tubulin to form the cell cytoskeleton) or contractile role (e.g. myosin in our muscles). The storage of ions (e.g. iron with ferritin) or their transport through the cell membrane (e.g. potassium channel) is also accomplished by proteins. Immune defense is also a crucial role, which is ensured by antibodies. And finally, proteins are also found in signaling pathways (e.g. insulin, cell receptor) and cell regulation (e.g. cyclin kinase). Overall, the wide majority of these functions involve a binding process with another protein, a lipid, an ion, a sugar or a nucleic acid. Nevertheless, the most exciting route to develop innovative therapies and biotechnology tools remain the protein-protein interactions.

Modulating PPIs represent a major goal for drug development as these interactions are not only involved in healthy cell homeostasis but also in disease progression, either pathogenic, degenerative or cancer-related [9]. Between 130'000-650'000 PPIs are estimated in the human interactome [10,11], but only a fraction of them have been targeted by drugs [12]. A wide majority remain "undruggable" mainly because of flat interfaces that lack a defined binding pocket for small molecules [12,13]. This makes however these sites promising candidate interfaces for protein-based therapeutics such as monoclonal antibodies [14] or *de novo* protein binders [15].

As for protein folding, different driving forces are involved in protein association. It has been shown that Van der Waals interactions and hydrophobic patches are the major contributor for protein-protein binding and are less tolerant to mutations, reason why they are often referred to as hotspots [16,17]. These hotspots can benefit from additional surrounding hydrogen bonds and salt bridges found at the rim to stabilize the interaction and improve binding specificity [18,19]. However, the molecular surface geometry [20] and shape complementarity of both interacting partners are also critical for protein association [21]. Some proteins may undergo obligatory homo-oligomerization with identical subunits (e.g. keratin) or hetero-oligomerization with non-identical subunits (e.g. antibody heavy and light chains) to achieve their full functionality. Alternatively, proteins may interact in a non-obligate way with other protein partners to accomplish their role (e.g. antigen-antibody complex).

These non-obligate interactions can be only transient with affinities ranging from low micromolar to mid-nanomolar, or permanent with affinities ranging from low nanomolar to femtomolar (e.g. E9 endonuclease-Im2 complex) [22,23].

[1] Of note, this assumption is valid for globular soluble proteins. Membrane proteins are exposing their hydrophobic residues to the cell membrane, which is composed of lipids and is therefore hydrophobic as well.

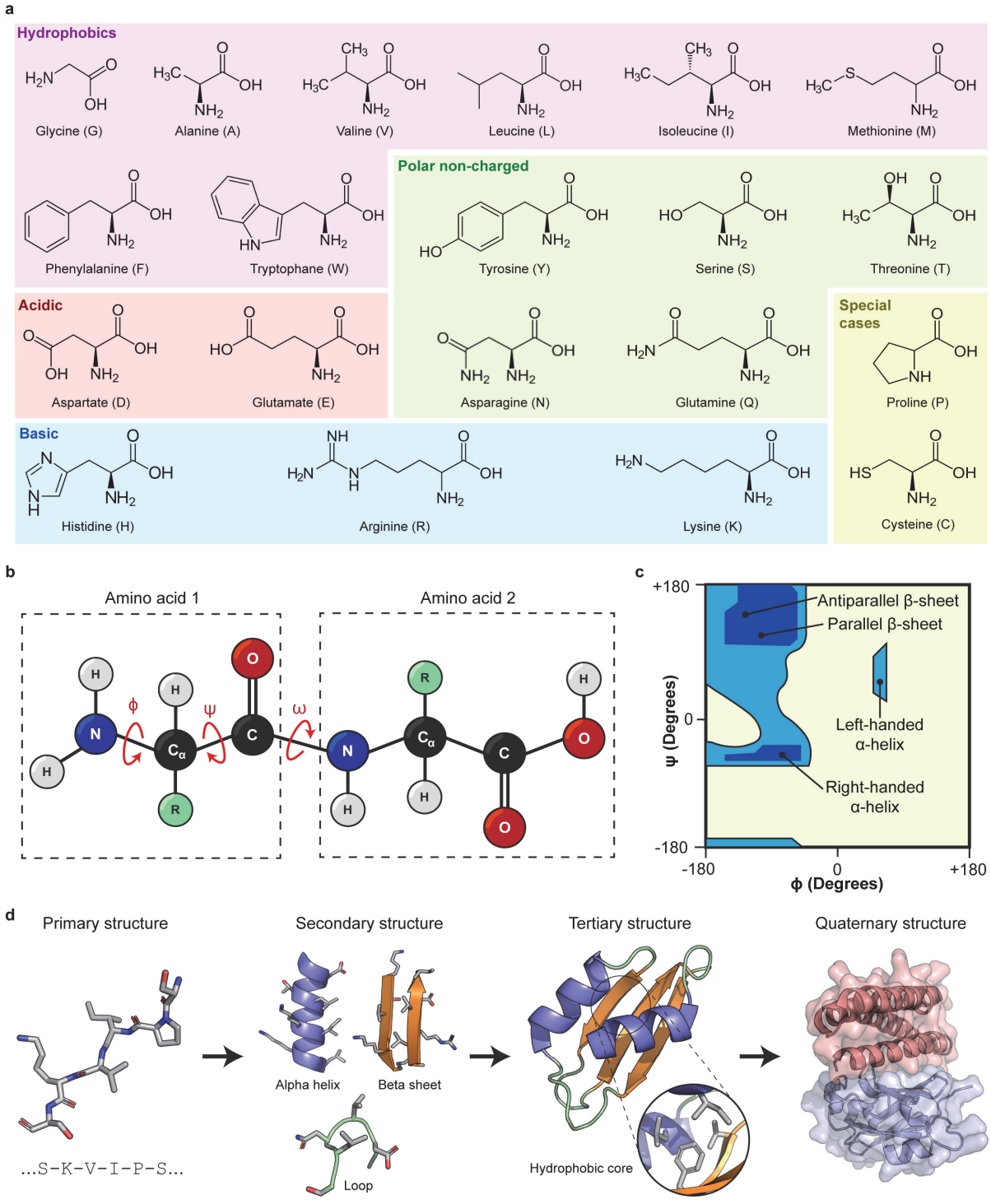



**Figure 1.1 : The biochemistry of amino acids and proteins. A.** Schematic representation of the 20 canonical amino acids involved in protein biosynthesis grouped into four categories (hydrophobic, polar uncharged, polar acidic and polar basic) and one special case category. **B.** Representation of two amino acids linked by a peptide bond. The phi (ϕ) angle is found on the N-terminal side, the psi (ψ) angle on the C-terminal side, and the omega (ω) angle at the peptide linkage. All central carbon ($C_\alpha$) are linked to a side chain (R) that gives a chemical property to the amino acid. **C.** Ramachandran plot summarizing the energetically allowed conformation of the ϕ and ψ angle of an amino acid. **D.** Hierarchy of the protein structure starting from the linear sequence of amino acid (primary), the local secondary structure elements (alpha helix, beta sheet or loop), the three-dimensional folded structure (tertiary), and eventually the complexation with another protein subunit (quaternary).

The PPI universe, also called interactome, is a complex biological collection whose prediction has remained a challenge over the last years [10,24]. Interactome mapping has been done primarily relying on experimental techniques including yeast-two-hybrid (Y2H) system, luminescence-based assays or co-immunoprecipitation coupled with mass spectrometry [10]. In parallel, several computational tools using evolution-derived structural and sequence similarities were proposed [25,26]. Nevertheless, purely independent PPI predictions based solely on the intrinsic surface properties without comparison with a homolog protein remained an ambitious aim until recently. As explained hereinafter, numerous computational methods have been developed for protein interaction prediction and, more generally, for protein modelling and design.

## 1.2 Computational protein modelling and design

With the advent of computers in the second half of the 20$^{th}$ century, numerous calculations and automations once impossible for humans, have become accessible to scientists. Since the 70's computers gained computational capabilities exponentially as observed by the Moore's law who posited that the number of transistors found in an integrated circuit doubles every two years [27]. This significant expansion enables the emergence of computational protein modeling and design since the mid 90's. Numerous tools have been proposed, notably for protein dynamic simulation [28–30], protein structure prediction [31–34], protein docking [35,36], or protein design [37–40].

### 1.2.1 Computational protein structure prediction

Protein modeling has been driven by the desire of folding any protein computationally given a sequence (Figure 1.2A-D). Three approaches have been proposed: i) Template-based methods, ii) template-free methods (including physics-based approaches) or, most recently, iii) neural network predictions [41] (Figure 1.2D). Firstly, as the three dimensional structure is intrinsically contained within the linear sequence of amino acid [42], it became quickly evident that searching for similarities between a query sequence and a database of known tertiary structure could be a rapid way to model the desired protein structure [43–45]. Nevertheless, this approach is hindered by the experimental data availability and the prior characterization of protein homologs that are evolutionary similar.

Following the Anfinsen principle [4] that postulated that a protein folds to its lowest energy state, template-free protein structure prediction tools relied on the laws of physics to find this global minimum and the subsequent three dimensional structure where a defined amino acid sequence will fold (Figure 1.2B). Rosetta, one of the most popular tools exploiting this approach, takes advantage of small existing backbone fragments sharing sequence similarities with the query. A Monte Carlo simulation [46] will introduce small conformational changes and perturbations on a randomly selected region. The acceptance of the move will be evaluated by a Metropolis criterion: sampled moves are accepted if the energy is decreased, while it might be rejected or still accepted depending on a certain probability criteria if the energy state is increased [39,47] (Figure 1.2C) :

$$\text{if } E_{new} < E_{original} : \text{Accepted}$$

$$\text{Otherwise accepted with probability } P = e^{-(E_{new}-E_{original})/T} \text{ ; where } T = \text{Temperature}$$

The conformational-energy landscape of protein folding is a complex funnel with several local minima and one global minimum where the energy state is at its lowest [48,49] (Figure 1.2B). By randomly accepting non-favorable moves, the algorithm mitigates the risk of becoming trapped in a local minimum. For the evaluation of the energy state, Rosetta exploits an all-atom physics-based scoring function, which has been continuously under improvement over the years [39,50]. The scoring function is decomposed in multiple weighted terms that are describing forces and potentials such as: i) Van der Waals energies (split in attractive and repulsive), ii) hydrogen bonds, iii) electrostatics, iv) disulfide bonds, v) residue solvation energy, vi) backbone torsion angles (based on Ramachandran statistics), vii) sidechain rotamer energy, and viii) a reference energy (average unfolded state). However, several limitations remain and notably the lack of consideration of the entropic contribution [51,52] or the absence of an accurate explicit water molecule representation [53] due to the computational cost of such simulations.

Finally, thanks to recent advances in machine learning and the increase of computational power, new tools have been proposed to the protein science community. AlphaFold2 (AF2), released by DeepMind, has pioneered a new era for deep learning tool in protein structure prediction by winning the 14th Critical Assessment of protein Structure Prediction (CASP14) [54] with a median backbone accuracy of 0.96 Å root mean square deviation (RMSD) and an all-atom accuracy of 1.5 Å, a performance that has never been reached formerly [31]. Since the success of AF2, other groups proposed similar structure prediction tools such as RoseTTAFold [32] and ESMFold (Evolutionary Scale Modeling Fold) [55] and many others that includes ligand and other biomolecules are about to be released [56,57].

### 1.2.2 Computational protein design

Protein design is often known as the inverse folding problem [41,58] (Figure 1.2A). While protein structure prediction and folding aims to search for the lowest energy state of a given protein sequence, protein design aims to define a sequence that will fold into a specific structure. Contrary to *in vitro* evolution or experimental optimizations, where proteins evolve or are selected from extensive mutational libraries, computational protein design employs methodical and rational approaches to forecast a limited number of sequences aligned with specific desired objectives. Computational protein design can be split into i) fixed-backbone design (or template-based design) and ii) *de novo* design [41,59].

The early stages of computational protein design mainly focused on template-based design with the aim to repurpose existing protein backbones and sample different side chains to improve stability [60,61], to redesign specificity [62,63] or incorporate a particular function by motif grafting (e.g. binding to another protein, see section 1.3) [64,65].

A longstanding challenge for protein engineering has been to design *de novo* proteins without any known homologs and to discover novel shapes and folds that Nature didn't explore. The protein universe is wide, but it is commonly stated that only a fraction of it has been explored by Nature, leaving a blank space that protein scientists need to uncover [59]. Though, to design a small protein of 100 amino acids without any knowledge, there are $20^{100}$ side chain combinations and a quasi-infinite amount of possible torsion angles. However, narrowing the problem to specific folds and using computational tools that combine sampling and scoring, such as Rosetta, made *de novo* protein folds possible since the early 2000's.

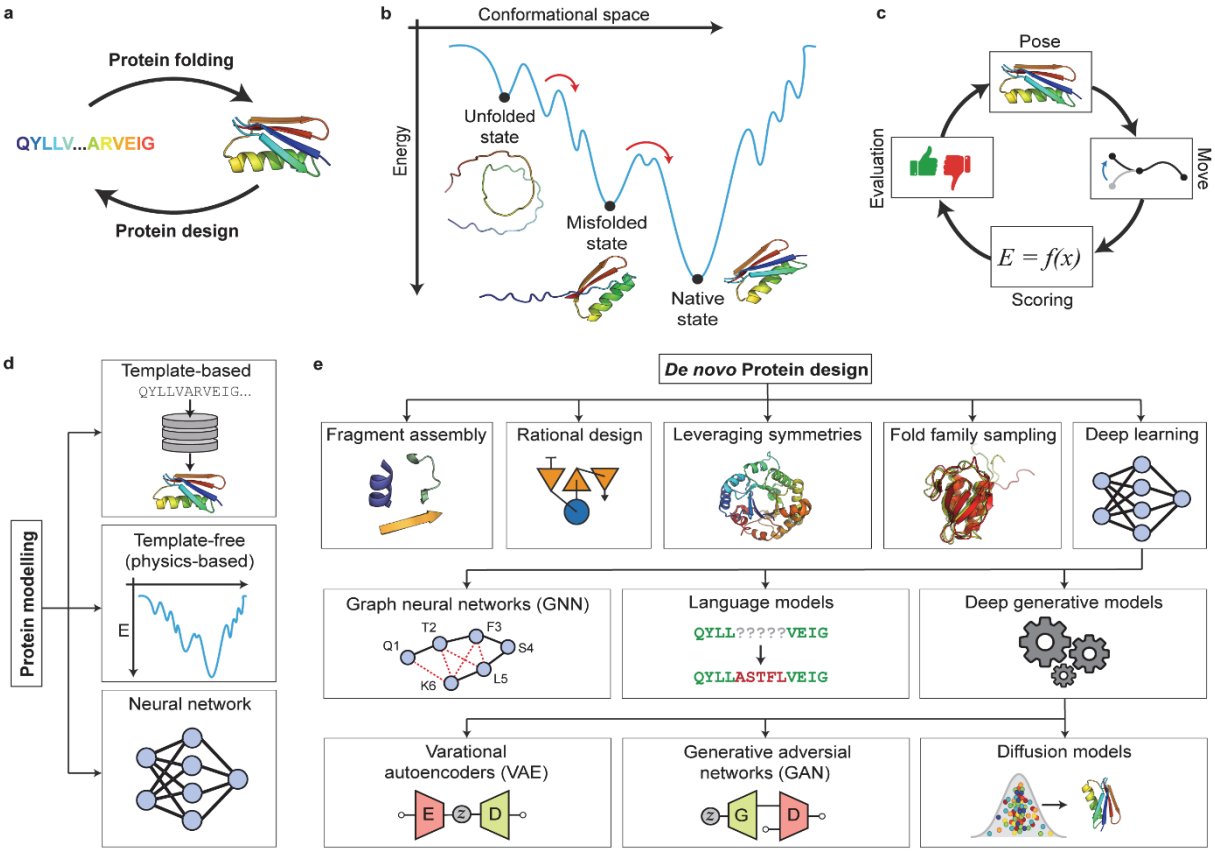



**Figure 1.2 : Computational protein modeling and design. A.** Protein modeling has been primarily focusing on protein folding with the aim to get a structure for a given sequence. Protein design represent the inverse folding problem with the aim to obtain the sequence that will fold into a desire three dimensional structure. **B.** The conformational-energy landscape of protein folding is a complex funnel with several local minima (intermediate and misfolded state) and one global minimum where the energy state is at its lowest (native state). **C.** Rosetta modeling suite follows an iterative principle where conformation moves are sampled, introduced into a protein pose, scored with a physics-based energy function and accepted or rejected following a Metropolis criterion. **D.** Three approaches are used for protein modeling and especially protein structure prediction: Template-based, template-free (e.g. Rosetta) or neural networks (e.g. AlphaFold2). **E.** Various methods were proposed for protein design. More recently, deep learning tools gained popularity and can be decomposed in graph neural networks (e.g. ProteinMPNN) or language models (e.g. Inpainting) of deep generative models, which can themselves be decomposed in variational autoencoders (VAE; E: Encoder; D: Decoder), generative adversarial networks (GAN; G: Generator; D: Discriminator), or diffusion models (e.g. RFDiffusion).

One of the first *de novo* protein designs reported to have explored the "dark space" of the protein universe is exemplified by the Top7 protein, proposed by Kuhlman and colleagues in 2003 [66]. By defining a sketch of the desired α/β-fold and leveraging the sequence design and backbone optimization capabilities of Rosetta, they provided a novel and stable protein fold validated experimentally. Since then, a myriad of *de novo* protein attempts were performed by using various strategies (Figure 1.2F), such as [67]: i) local structure assembly [66,68,69], ii) rational design (bottom-up approach) [70,71], iii) leveraging symmetries [72], or iv) fold family sampling [73]. As of today, multiple examples of small *de novo* protein designs displaying significant stability [74–76] and successful functionalization [77–79] have been reported by using high-throughput screening techniques.

Protein design is, however, entering a new era with the arrival of emerging machine learning algorithms [80]. Machine learning represents a considerable toolbox for scientists with applications in biology and beyond. For the field of protein design, most emerging tools are based on graph neural networks (GNN) [81], language models [82] or deep generative models [80] (Figure 1.2F). Proteins are the perfect examples of graph representation with the amino acids being the nodes, and their spatial relationship (bonded or not) being the edges. Thus, a subset of GNN called message passing neural networks (MPNN) [83], such as ProteinMPNN [37], were developed to (re)design a protein sequence given a defined backbone and proved to be successful in bringing highly stable molecules [84,85].

As opposed to being represented as a graph, proteins can be considered as "words" built from an "alphabet" of 20 amino acids. As for conversational language models generating full sentences or filling missing words, protein language model were trained to generate full-length protein sequences [86] or filling the gap of a partial sequence (also known as "inpainting") [87]. Finally, deep generative models take advantage of probability distribution to generate novel data and, notably, variational autoencoders (VAE) [88,89], generative adversarial neural networks (GAN) [90,91] and diffusion models [38,92–94], which are the most popular frameworks to generate novel molecules. Diffusion models, such as RFdiffusion [38], are gaining more and more popularity due to their ability to generate protein backbones out of a random noise distribution by following a denoising process. However, this approach is still in its early stages, and many diffusion tools do not currently integrate side chain prediction or lack all necessary experimental validations [80]. To sum up, significant advancements have been made in the field of computational protein design over the last two decades, but a new era driven by deep learning tools is set to further evolve in the upcoming years.

## 1.3 Computational design of novel protein-protein interactions

As discussed previously, multiple biological processes involve an interaction between a protein and another protein, a process called protein-protein interaction (PPI). Due to their wide implication in our cell but also in multiple disease progression, PPIs became a target of choice for designing novel protein-based therapeutics such as monoclonal antibodies for example. However, with the advances in terms of computational protein design, innovative methods have been proposed for designing novel protein-protein interactions that can serve as protein-based therapeutics or biotechnology tools for synthetic biology.

This section is adapted from a review published in Current Opinion in Structural Biology in 2022 (doi: 10.1016/j.sbi.2022.102370), as allowed by the publisher.

**Authors**

Anthony Marchand[1,2‡], Alexandra K. Van Hall-Beauvais[1,2‡] & Bruno E. Correia[1,2]

**Affiliations**

[1] Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[2] Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

[‡] Equal contribution

### Author contributions

A.M., A.K.V and B.E.C wrote the review with an equal contribution. Figures were generated by A.M. with inputs from all authors.

### Funding

We thank the support of the European Research Council (starting grant no. 716058), the Swiss National Science Foundation (grant no. 310030_188744), the NCCR Molecular Systems Engineering (www.nccr-mse.ch) and the NCCR Chemical Biology (www.nccr-chembio.ch).

### 1.3.1 Abstract

Protein-protein interactions (PPIs) govern numerous cellular functions in terms of signaling, transport, defense and many others. Designing novel PPIs poses a fundamental challenge to our understanding of molecular interactions. The capability to robustly engineer PPIs has immense potential for the development of novel synthetic biology tools and protein-based therapeutics. Over the last decades, many efforts in this area have relied purely on experimental approaches, but more recently, computational protein design has made important contributions. Template-based approaches utilize known PPIs and transplant the critical residues onto heterologous scaffolds. *De novo* design instead uses computational methods to generate novel binding motifs, allowing for a broader scope of the sites engaged in protein targets. Here, we review successful design cases, giving an overview of the methodological approaches used for templated and *de novo* PPI design.

### 1.3.2 Introduction

Proteins are among the most ubiquitous molecules of life and are likely the most versatile in terms of function, biophysical properties, and diversity. They perform primordial functions for cell signaling, structure, transport, catalysis, regulation, and defense, among others. Many fundamental protein functions involve association with other proteins, referred to as protein-protein interactions (PPIs) [16]. Native PPIs are involved in most cellular functions and their binding affinities span several orders of magnitude, with dissociation constants commonly ranging from picomolar to micromolar [95].

PPIs are involved in cell homeostasis processes that, if disrupted, can lead to numerous disease progressions, either pathogenic, degenerative or cancer-related [9]. Of the more than 645,000 disease-relevant PPIs, few have been successfully targeted by drugs [12]. A wide majority remain “undruggable” mainly due to featureless interfaces that lack defined binding pockets for small molecules [12]. In addition to studying PPIs as a source of potential druggable targets, PPIs are at the core of novel biotechnology tools such as protein-based therapeutics [13,96], cell therapies [97–99], bio-sensors [100–102], vaccine candidates [71,103,104] and other synthetic biology applications [105–108] (Figure 1.3A).

Similar to protein folding processes, protein association is driven by energy minimization. This process has several driving forces, including Van der Waals interactions, hydrophobicity, and electrostatic steering (also called long-range electrostatics) [109]. Hydrogen bonds and salt bridges stabilize the

interaction and improve specificity [18,19,110]. The geometry of the molecular surface [20], with both shape and chemical complementarity of the interacting partners, plays a critical role for protein association [21] (Figure 1.3B).

In order to engineer novel PPIs, approaches such as *in vitro* evolution have been extensively used in the past decades [111–113]. However, one of the most important limitations of *in vitro* evolution is that it is "site agnostic," meaning that it is impossible to predict with certainty where the evolved binder will target the protein of interest. For the biological function of the binder, this is an important challenge that computational approaches attempt to solve. With the rise of computational methodologies numerous bioinformatics tools to predict, design, and engineer protein structures have been developed to address the limitations of the *in vitro* maturation techniques [31,39,114].

In this review, we will highlight successful design cases and discuss challenges in the computational design of PPIs. We group computational PPI design strategies in two categories: I) template-based design and II) *de novo* design. The first approach consists of transplanting a motif that mediates an existing PPI interface onto a new protein scaffold [115]. Despite its robustness and relatively high success rate, this strategy constrains PPI design to existing interfaces and precludes the possibility of targeting new sites. To explore a broader landscape of solutions, *de novo* design strategies aim to create completely new interactions starting from only the structure of the target protein[115]. However, engineering PPIs from scratch remains a non-trivial task requiring a detailed understanding of biomolecular interactions and stands as a stringent test of our understanding of the driving forces of PPIs.

### 1.3.3 Template-based design of protein-protein interactions

The template-based approach consists of transplanting the binding motif of an existing PPI into a new structural context (Figure 1.4). The motif is grafted onto a protein scaffold by sidechain grafting (i.e., backbone mimicry and then sidechain replacement) or backbone grafting (i.e., full motif transplantation including sidechains and backbone). Alternatively, a *de novo* protein scaffold can be built around the binding motif.

One of the first cases of successful computational sidechain grafting design dates from the early 2000s by Liu and colleagues [116]. The Protein Data Bank (PDB) [117] was searched for scaffolds that contained three residues satisfying the geometric relationships of the $C_\alpha$–$C_\beta$ vectors of the three key residues of EPO required for binding to the receptor EPOR. Grafting only these three residues onto an appropriate scaffold resulted in a binder with 24 nM affinity to EPOR, highlighting the crucial contribution of hotspot residues in PPIs [118]. Several years later, a similar strategy [104,119] used backbone similarity searches to find host protein scaffolds onto which continuous viral epitopes were transplanted. To address higher structural epitopes, this approach was extended to transplant discontinuous backbone segments of a viral epitope [120]. In both cases, the epitope transplantation gave binding affinities to the antibody in the nanomolar range and high structural agreement to the original epitope.

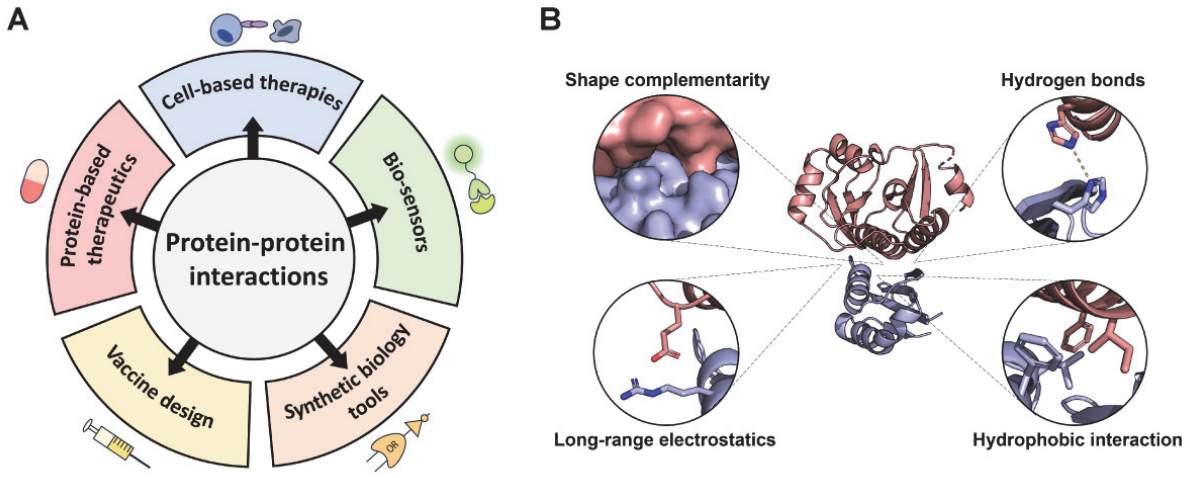



**Figure 1.3 : Overview of potential applications for novel PPIs and molecular features that drive protein association. A.** Protein-protein interactions (PPI) have numerous applications for vaccine design, protein-based therapeutics (e.g. antibodies, inhibitors, etc.), cell-based therapies (e.g. CAR-T), bio-sensors (e.g. diagnostics), or as synthetic biology tools (e.g. ON/OFF-switch) **B.** Different structural features that can be designed by computational methods are necessary to engineer a strong PPI. These include good shape complementarity, hydrophobic patches, hydrogen bonds, and long-range electrostatic interactions (electrostatic steering) that stabilize the interaction and improve specificity.

**Table 1.1 : Key terms in the field of *de novo* PPI design.**

| **Term** | **Definition** |
|---|---|
| Binding motif | Continuous or discontinuous structural segments of amino acids that encompass the interface in a protein-protein interaction. |
| Hotspot | Key residues that have a large energetic contribution for the affinity of the protein-protein interaction. |
| One-sided design | Approach where the binder is designed and the target remains constant. |
| Two-sided design | Approach where both interfaces involved in the protein complex are designed. |
| Scaffold protein | Heterologous protein used as a recipient for the grafting of hotspot residues and/or binding motif(s). |
| *De novo* scaffold protein | Protein scaffolds that have been designed using computational approaches that model protein backbones and find the best sequences to stabilize the fold. |
| *De novo* PPI design | Design of novel protein-protein interactions without using explicit information of binding motifs used in native protein complexes. |

Sidechain grafting has been successfully used to transplant helical motifs onto *de novo* designed scaffolds. Successful examples include the design of candidate protein-based inhibitors against influenza haemagglutinin (HA) and botulinum neurotoxin B (BoNT/B), using known HA binders or natural BoNT/B target respectively as a helical template motif for subsequent grafting on *de novo* miniprotein scaffolds [77]. Future research efforts in protein-based therapeutics will benefit from the generation of highly stable *de novo* scaffolds presenting functional motifs.

Recently, the sidechain grafting approach for PPI engineering demonstrated useful applications for synthetic biology and the design of small molecule-controlled switches. The underlying principle consists of repurposing an existing PPI that can be targeted by a known small molecule to control its dissociation. Giordano-Attianese and colleagues [97] repurposed the binding of BH3-motif to Bcl-XL by grafting the sidechains onto a globular scaffold protein. This led to a 3.9 pM affinity for Bcl-XL and created a protein switch controlled by a Bcl-XL inhibitor. The novel switch was incorporated into the chimeric antigen receptor (CAR) of T cells and was shown to turn off killing activity upon the addition of the small molecule. Work by Shui and colleagues has extended the logical behavior of this system creating a multidomain architecture that, upon the addition of a small molecule, triggers the association of the two protein subunits [105]. These applications demonstrate promising applications for translational research in the domain of cell engineering.

Motif grafting by sidechain replacement faces limitations when the motif is too complex to find a structurally compatible protein scaffold. Grafting approaches have been described where both sidechains and backbone are grafted onto protein scaffolds. Azoitei and colleagues designed epitope-scaffolds by selecting scaffolds based on N- and C- termini alignments to identify sites in proteins where the motif was grafted and the connection regions were further refined and designed [121]. Such strategy was also successfully utilized to transplant a complex binding site from an HIV epitope, composed of two discontinuous segments that were required to present a precise three-dimensional structure to mediate productive binding to the antibody B12 [122]. The two segments of the epitope were grafted in a stepwise fashion and multiple rounds of *in vitro* evolution were performed to optimize the binding affinity of the designed scaffold, highlighting the difficulty of grafting complex sites onto protein scaffolds.

To address more complex epitopes, the Fold From Loops (FFL) protocol was proposed as an alternative by folding *de novo* scaffold proteins to stabilize the binding motif of interest [103]. The FFL approach was first used to embed a viral epitope from RSV onto a *de novo* folded and designed three-helix bundle protein. Several of the designs bound with picomolar affinities to a site-specific monoclonal antibody and the designs showed, for the first time, the ability to elicit neutralizing antibodies in non-human primates. Further, the FFL protocol was utilized by Procko *et al* to design a protein inhibitor against an Epstein Barr-Viral (EBV) Bcl2-homolog called BHRF1 [123]. Extensive *in vitro* maturation was necessary to stabilize and improve the affinity to BHRF1 and the success rate of functional designs was rather low. The FFL methodology was also used by Bryan *et al* [124] to develop small, ultra-stable mini-protein scaffolds designed around a five amino acid stretch of PDL-2, one of the native binding partners of PD-1, resulting in a 100 nM binder for PD-1.

Two main shortcomings of the FFL protocol came to light: I) the lack of compatibility for multiple discontinuous motifs; II) the incorporation of the binding partner during the folding-design simulations for the optimization of additional contacts and as a constraint for the sampling the conformational/sequence space. Bonet *et al* improved FFL, by developing a Rosetta framework called FunFolDes, which addressed these drawbacks [125]. This novel approach successfully functionalized "functionless" folds by incorporating the Respiratory Syncytial Virus protein F (RSVF) site IV on a *de novo* protein called TOP7. Another intrinsic limitation of the FFL approach was its reliance on existing structures, either native or *de novo* designed. To circumvent this drawback, Sesterhenn and colleagues proposed the TopoBuilder, a protocol for the assembly of *de novo* topologies conditioned to the structure of the motif of interest [71]. Upon the assembly of the topologies with the embedded functional/binding motif, the FunFolDes folding and design protocol is used for sequence generation. This work contributed to the development of different candidate vaccine immunogens that elicited neutralizing antibodies against specific viral epitopes and created a series of functional molecules that were used for different synthetic biology applications [70,71]. Other methods of grafting hotspots to *de novo* scaffolds led to rapid design of a nanomolar SARS-CoV-2 binder that neutralized SARS-CoV-2 [126] and the use of *de novo* peptides as a scaffold for PPI disruptors [127]. These methods highlight the potential uses for *de novo* scaffolds, albeit dependent on known interactions.

Overall, these methods allow for PPI design with various levels of complexity, however, they are limited to known binding interactions. To broaden the landscape of targetable protein interfaces, *de novo* approaches to generate motifs that can mediate novel PPIs is needed.

### 1.3.4 *De novo* design of protein-protein interactions

In the context of this review, *de novo* strategies for the design of protein interactions rely only on the structural information of the target, which we generally refer to as one-sided design. *De novo* design strategies are subdivided in: I) dock-&-optimize; II) hotspot-centric approaches [115] (Figure 1.5). The dock-&-optimize approach consists of two stages. First, hundreds of protein scaffolds are computationally docked on the target protein to find configurations with favorable shape complementarities. Second, interface residues of the best candidates are computationally designed to improve the binding propensity. Alternatively, the hotspot-centric approach first requires the placement of a few clustered hotspot residues before grafting onto a suitable scaffold protein that will be further refined [128].

One of the early publications [129] in this field introduced a Rosetta-based protocol, called DDMI (Docking, Design, Minimization and Interface), following the dock-&-optimize approach. The DDMI protocol is a two-step approach which uses rigid-body docking to find a suitable orientation for the partner scaffold and then iterates between sequence design and energy minimization to settle the interface to the lowest energy state. Their best candidate, named "Spider Roll," used a single pre-selected scaffold to target the kinase domain of p21-activated kinase 1 (PAK1), and showed only weak

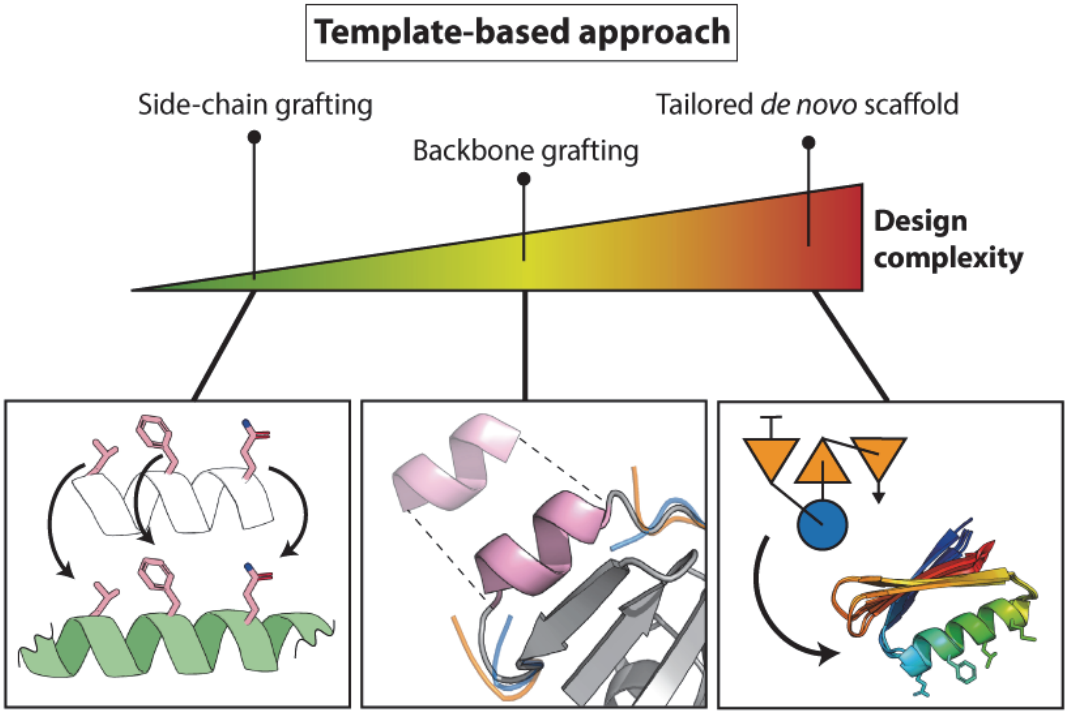



**Figure 1.4 : PPI design methods using the template-based approach.** The template-based approach can be subdivided (from lowest to highest complexity) in I) side-chain grafting, II) backbone grafting, or III) use of a tailored *de novo* scaffold. Side-chain grafting transplants binding motifs from an existing PPI onto a heterologous scaffold that stabilizes the interaction between these side-chains and the binding target. In backbone grafting the transplantation involves the full backbone and side chains of the binding motif involved in a PPI onto a heterologous scaffold. Backbone grafting often poses the difficulty of modeling realistic backbones and finding suitable stabilizing sequences in the connecting segments between the grafted motif and the scaffold. Finally, in a more tailored approach, a *de novo* scaffold could be built around the motif of interest by specifying the arrangement of secondary structure elements to generate a three-dimensional topology.

affinity ($K_D \approx 100$ µM). This study and others using a dock-&-optimize approach [130,131] were strong demonstrations that more accurate energy force fields are needed, as well as larger pools of scaffold candidates and, due to all these limitations, *in vitro* evolution techniques may be required to further optimize the putative binders.

In an alternative route, hotspot-centric approaches were proposed. Fleishmann and colleagues were the first to implement a hotspot-centric method to target a conserved surface site on the stem of the influenza hemagglutinin (HA) [132]. The design approach consisted of docking disembodied residues, selecting suitable scaffolds, and refining the interface with RosettaDesign [66]. Out of 73 designs screened by yeast display, 2 showed binding to HA including one with an apparent affinity of 200 nM. Two rounds of affinity maturation were performed, providing insight into the sub-optimal features of the designed protein: I) void volumes at the interface should be minimized and backbone minimization can facilitate the choice of suitable residues; II) complementary electrostatic charges which remain outside of the hydrogen-bond range (~3 Å) should not be underestimated; III) the energetic cost for the desolvation of charged residues in close contact with non-polar amino acids should not be neglected. In conclusion, the hotspot-centric strategy yielded a higher affinity binder than the dock-&-optimize approach, noting the fact that these were optimized by *in vitro* evolution.

Later on, Procko *et al* [130] targeted the hen egg lysozyme (HEL) using the same approach with two polar hotspot residues. Scaffold candidates were docked, refined, and selected to satisfy both the disembodied hotspot residues and the complementarity for the target. Out of 21 designs, one showed a modest affinity of 7 µM and required two rounds of directed evolution and four mutations to obtain a final affinity of 8 nM. This experiment, as the previous one, had to rely partially on known interacting residues, as well as *in vitro* maturation techniques to improve binders to an acceptable affinity, although requiring only a few mutations. Despite these promising results, both studies showed that hotspot residue placement was a promising approach, however the need for *in vitro* optimization and the low success rates support that improved energy functions and methods are still necessary.

Recently, computational tools such as the rotamer interaction field (RIF) docking have been proposed to search for *de novo* hotspots for PPI and protein-ligand design without prior knowledge. Briefly, billions of disembodied residue conformations are docked on the target interface with the aim of introducing hydrogen bonds and hydrophobic packing interaction to create an energetically favorable interface. All RIF rotamers are stored and can be rapidly sampled for scaffold matching using a docking grid-based search algorithm [133]. RIF docking and a miniprotein scaffold library were used for the rapid generation of protein-based therapeutics against SARS-CoV-2 spike protein, with *de novo* designs having affinity lower than 1 nM after *in vitro* evolution optimization [78].

Ultimately, with the same strategy, the same group was also able to generalize the hotspot-centric approach proposed by Fleishman and colleagues [132] by generating at least one binder for 12 different target proteins [134]. These publications were among the first to demonstrate complete *de novo* hotspot generation for PPI engineering. Intriguingly, most binders designed so far rely on helical structures, limiting the landscape of binding motifs available for PPI designs, especially when working with disembodied residues. This approach still seems dependent on a large library screening (15'000-100'000 designs per target), although it undoubtedly pushed the frontier in de novo PPI design.

### 1.3.5 Challenges and perspectives

The methodology for designing novel PPIs has evolved rapidly in the past years. Templated design approaches take advantage of known binding partners and transplant either the sidechains of hotspot residues or the backbone and sidechains of the region of interest. Although this method is reliable and has led to the design of many successful binders, it is limited by relying on known binding partners. *De novo* design does not rely on such interactions and is a more difficult problem that poses a rigorous test to our understanding of the principles that drive protein-protein interactions. Recently, RIFDock has allowed for *de novo* hotspots to be predicted without prior knowledge of binding partners and from these hotspots, high affinity binders have been developed.

Despite these successes, there are still challenges that need to be addressed. There is a low success rate for *de novo* designs, and the designs that are successful often need rounds of *in vitro* evolution to improve the affinity. One plausible explanation could be the lack of proper energy functions to capture

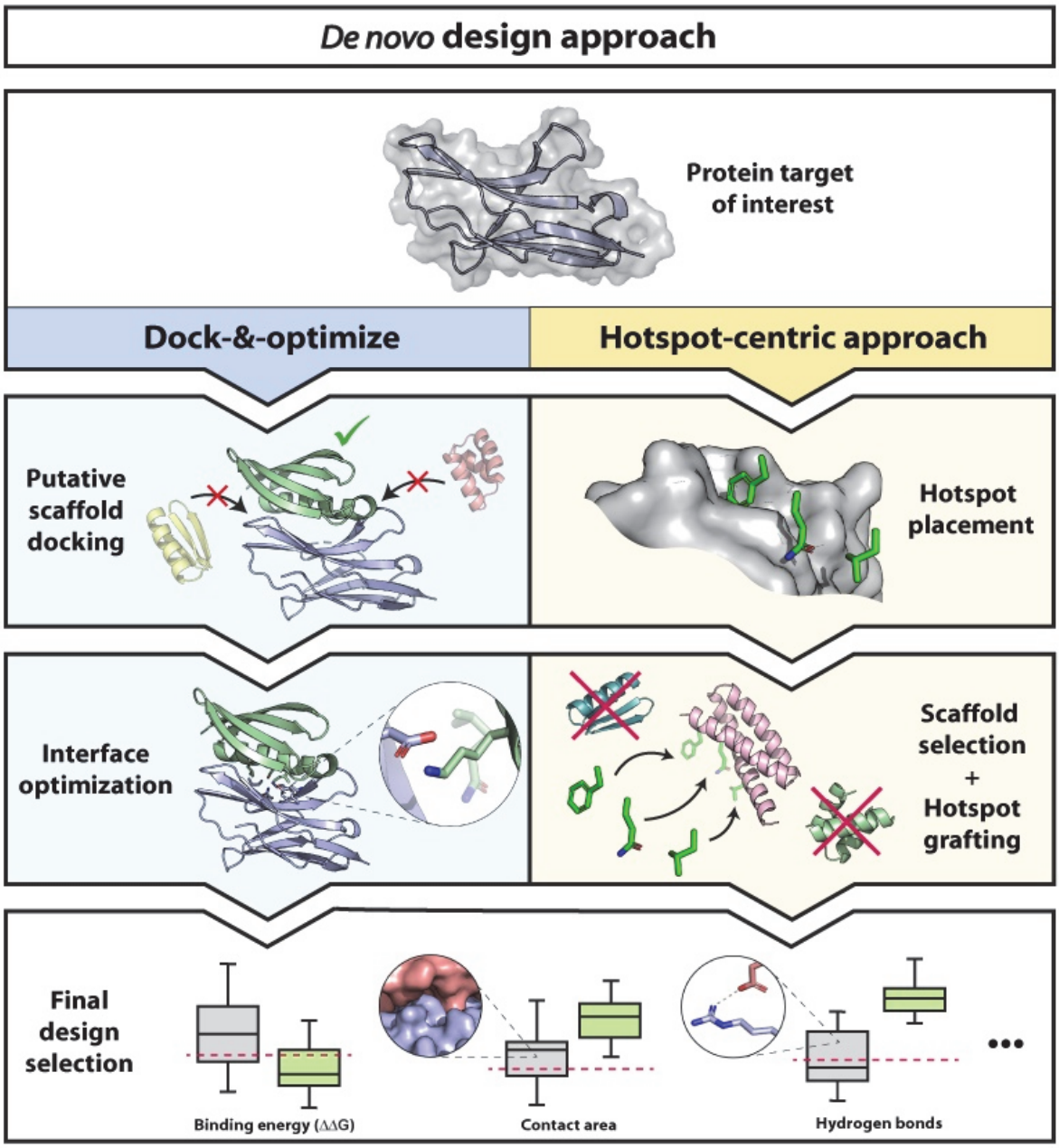



**Figure 1.5 : PPI design methods using the *de novo* approach**. The *de novo* design approach consists of two alternative strategies: I) Dock-&-optimize or II) Hotspot-centric approach. The first is a two-step method that combines the docking of putative scaffolds and then an interface optimization aiming to minimize the binding energy between the target and the most appropriate scaffold. In the second method, hotspot residues are searched, placed on the interface of interest, and then grafted on a scaffold which is suitable for both side-chain orientations and target interface. In both methodologies, a final selection based on different metrics (binding energy, contact area, hydrogen bonds, etc.) is needed to reduce the pool of designs to be tested.

long-range interactions and the effect of water molecules. A study also found that poorly designed buried hydrogen bonds account for most of the failure in *de novo* PPI attempts [135]. Computational tools aiming to design broad hydrogen bond networks, such as HBnet [136], or the introduction of score penalty for buried unsatisfied polar atoms [137] may help future *de novo* design pipelines to tackle challenging polar interfaces. Additionally, more work must be done to extend these methods beyond helical motifs. Although helical binders can be successful, opening this strategy to more than one secondary structure would further increase the breadth of structural space that could be covered. Finally, it seems an emergent theme that most of the *de novo* PPI designs target known PPI interfaces, leaving unsolved challenges in targeting arbitrary target sites that may have low interface forming propensity. Despite these challenges, new tools for protein engineers are being developed that can address these difficulties. Newly introduced machine learning based software such as MaSIF [138] allows protein engineers to predict novel binding sites and possible binding partners. The introduction of Alphafold [31] and RoseTTAFold [32] allows for the prediction of three-dimensional protein structures with just the amino acid sequence. These tools and others will assist protein engineers in further studies. Despite the difficulty of understanding and accurately designing novel PPIs, the number of computational methods available is expanding steadily and will undoubtedly lead to a higher success rates and benefit to translational research with biomedical applications.

## 1.4 Geometric deep learning for the study of protein surfaces

As explained in chapter 1.1, the mapping of the protein interactome predominantly relied on experimental methods or computational prediction using prior knowledge, mostly by structural or sequence similarities. Nevertheless, complicated complexes or protein lacking known homologs remain challenging. But three parameters made new computational approaches possible: i) the computational power capabilities, ii) the amount of data available, and iii) the algorithmic innovations [139]. As said previously, computational power has been doubling every two years, reaching computational potential that was previously unattainable [27]. Thanks to the experimental work done over the last decades, a myriad of data has been generated and can be used to train new algorithm. For instance, the number of entries in the Protein Data Bank (PDB) exceeded 210'000 at the end of 2023, which represent a 3-fold progression compared to ten years ago [117]. Finally, the promotion of open source tools and the contribution of the research field made new training algorithms available. Among them, geometric deep learning emerged as a resource to learn and predict the features of various surfaces, including protein. Geometric deep learning takes advantage of neural networks in high-dimensional non-Euclidean spaces, such as protein surfaces which involve 3D-surface and chemical characteristics [140].

With this novel approach in hands, Gainza *et al* proposed a geometric deep learning framework called MaSIF (Molecular Surface Interaction Fingerprinting) with the aim to capture the crucial determinants of biomolecular interaction. Unlike previous tools that rely on an atomistic representation, MaSIF mainly emphasizes a higher-level depiction of proteins, namely the molecular surface (Figure 1.6A). This representation, also known as the solvent-excluded surface, is derived by 'rolling' a water molecule probe

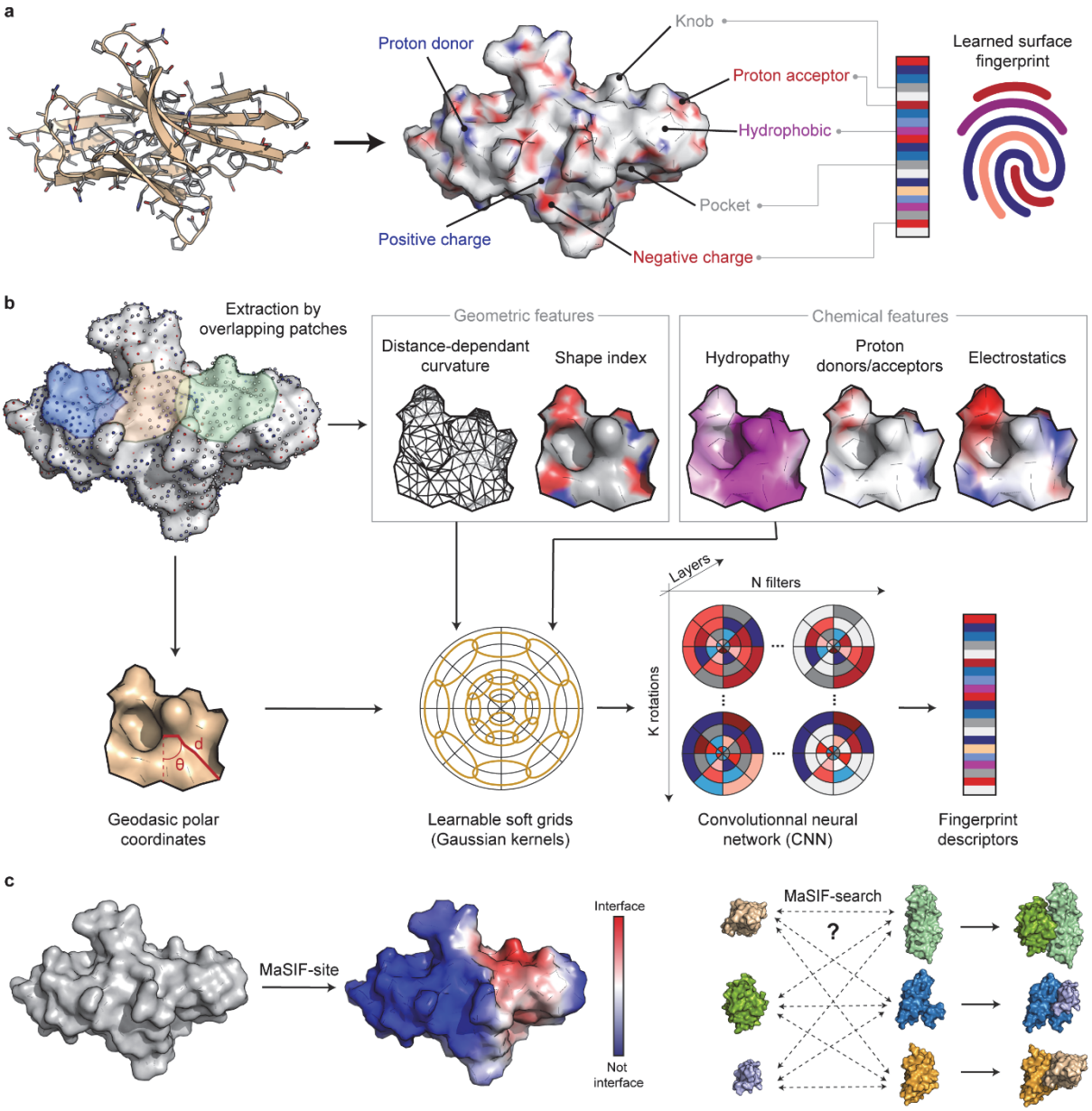



**Figure 1.6 : Overview of the MaSIF framework and its applications. A.** A protein target of interest is represented as a molecular surface representation where different geometrical and chemical features are computed. The vectorization of these surface features in a so-called "descriptor" is comparable to a "fingerprint" that can be learned for different type of predictions. **B.** The molecular surface of the protein of interest is divided in overlapping patches with a geodesic radius of 12Å. The patches are then described with several geometrical and chemical features mapped in space with geodesic polar coordinates. A learnable Gaussian kernels locally average the vertex-wise patch features and a convolutional neural network (CNN) is applied to produce the surface descriptors (fingerprint) that can be used for different applications. **C.** MaSIF-site is an application than predict the interface propensity over the protein surface. MaSIF-search is an ultrafast search algorithm to search and dock protein partners together.

over the protein atoms [20,141]. It notably constitutes the area of significant relevance for most biomolecular interactions with other molecules. MaSIF precomputes various geometrical (shape and curvature) and chemical features (hydrophobicity, electrostatics and proton donors/acceptors) found on the protein molecular surface represented as a mesh. Rather than being considered as a single object, the protein molecular surface is divided in overlapping patches with a radius of 12Å, which corresponds to the average size of a PPI interface[2]. Each patch is then discretized in different vertices mapped with a geodesic polar coordinates and assigned with the corresponding geometric and chemical feature values. These vectorized features are often referred to as fingerprints or descriptors (Figure 1.6B).

MaSIF's descriptor were processed with convolutional neural networks (CNNs) [142] for specific tasks and notably for PPI site prediction (MaSIF-site) and for a fast search of protein partners (MaSIF-search) (Figure 1.6C). Overall, MaSIF has the advantage of not being based on neither evolutionary background nor an explicit protein sequence, but solely on the vectorized intrinsic features of the buried interface. Hence, we are in principle able to predict PPIs – and by extrapolation to design novel PPIs (see chapter 3) – even on interfaces that lack prior knowledge and documentation.

## 1.5 Objectives

With the advances made in the field of computational protein design – but considering the challenges that remain in terms of validation and optimization – my thesis work is found at the interface between the computational and experimental domains. This project aims to take advantage of state-of-the art computational methods, but also to develop novel machine-learning based tools, to seek out novel biomolecular interactions with translational capabilities. Nevertheless, experimental validation and optimization will be performed to endorse the effectiveness of the computational approaches being developed and to further improve them.

### 1.5.1 Aim I: Rationally designing chemically controlled protein therapeutics

Numerous protein-based therapeutics, such as monoclonal antibodies or cytokines, have been developed over the last decades to successfully fight numerous diseases, notably cancer. However, most of these therapies are limited by their toxicities that triggers unwanted side effects and deleterious complications to the patients. Therefore, my thesis work will first aim to rationally design a switchable protein therapeutic by incorporating and optimizing a previously developed chemically-disruptable heterodimer using a clinically approved drug into a potent protein therapeutic system. While this approach has been effectively applied to the field of cell-based therapies, namely chimeric antigen receptor T (CAR-T) cells, I sought here to translate this switchable system into a soluble protein therapy with the help of physics-based computational methods to rationally improve the switchability in solution. The chapter 2 presented in this dissertation illustrates how this OFF-switch system led to the development of a more controllable and safer therapy.

[2] See supplementary Fig S3.2 in chapter 3

### 1.5.2 Aim II: Designing de novo protein interactions using learned surface fingerprinting

Despite the major breakthrough made in the field of protein-protein interaction (PPI) prediction and design, numerous challenges remained. A majority of the methodologies suggested thus far repurposed pre-existing PPIs, relied on prior interface knowledge or used physics-based methods with low energetic resolution as exemplified previously in chapter 1. Designing novel binders for defined protein targets – and especially those with no prior knowledge – requires two pieces of information: i) the site with the highest propensity to form an interface, and ii) the optimal motif to bind to this interface. With this rational, my colleagues and I sought to adapt MaSIF, a geometric deep-learning framework for PPI prediction (see chapter 1.4), for the design of site-specific *de novo* protein interactions. As shown in chapter 3, we hypothesized that protein fragments that show geometrical and chemical complementarity to a defined patch on a protein target of interest constitute the best candidates to design novel protein binders. By using MaSIF to predict sites with the highest interface propensity, and by leveraging the vectorized geometrical and chemical features of this patch (also known as "fingerprint") to search for complementary binding seeds, we successfully designed and validated several binders for proteins of major therapeutic interest.

### 1.5.3 Aim III: Targeting protein-ligand neosurfaces using a generalizable deep learning approach

While a wide choice of chemically-disruptable heterodimer (CDH) systems serving as OFF-switches have been proposed, the number of examples of chemically-induced dimerization (CID) systems serving as ON-switches remains limited. Most CID systems proposed thus far have been obtained by experimental methods or through an extensive *in vitro* evolution. On the other hand, only a limited number of algorithms are accounting for both proteins and small molecules for the purpose of protein design, creating a gap in the design process of novel chemically-induced PPIs. In this last aim, presented in chapter 4, we hypothesized that a small molecule-bound interface constitutes a hybrid neosurface with a unique signature that does not exist in the unbound state. We therefore sought to expand and generalize the surface fingerprinting approach, developed in chapter 3, to account for the presence of small molecules in order to design drug-bound specific protein interactions. Thus, we can explore new application capabilities including biosensors, logic gates, and various other biotechnology tools.

Altogether, this work presents a combination of classical and cutting-edge computational tools for the design of protein interactions with a potential development of innovative therapies and biotechnology tools. Together with this aim, this project will lead to some conclusions for a better understanding of computational PPI design.

# Chapter 2

## Rational design of chemically controlled protein therapeutics

As explained previously in this work, recent advances in terms of computational protein design allowed the emergence of tools in order to repurpose existing protein interactions (template-based approach) or to design completely novel ones (*de novo* approach). Template-based approaches such as side chain grafting have been performed over the last decade with numerous successful examples. The main advantage relies on the absence of extensive hotspot search and scaffold design, as hotspots come from a known protein interface and are grafted on an existing scaffold originating from a structural database. In this chapter, we will present an example of template-based approach that has been previously proposed for a switchable CAR T cell therapy and has been optimized and adapted for different switchable protein therapeutics in solution, including antibodies.

This section is adapted from an article published in ACS Chemical Biology in 2023 (doi: 10.1021/acschembio.3c00012), as allowed by the publisher (License CC-BY 4.0).

**Authors**

Anthony Marchand[1‡], Lucia Bonati[1,2‡], Sailan Shui[1], Leo Scheller[1], Pablo Gainza[1], Stéphane Rosset[1], Sandrine Georgeon[1], Li Tang[2], and Bruno E. Correia[1]

**Affiliations**

[1] Laboratory of Protein Design and Immunoengineering, Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

[2] Laboratory of Biomaterials for Immunoengineering, Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

[‡] Equal contribution

**Author contributions**

A.M. and L.B. contributed equally. A.M., L.B., L.T., and B.E.C. led the project. A.M. and L.B. performed the experimental work. S.S., L.S., and P.G. contributed to the design of the experimental setup. A.M. and P.G. contributed to the computational optimization. S.R. performed the biolayer interferometry. S.G. expressed and purified the proteins. A.M., L.B., L.T., and B.E.C. wrote the manuscript with input from all authors.

**Funding**

The work done by the team of Bruno Correia was supported by the European Research Council (Starting grant — 716058), the Swiss National Science Foundation (310030_197724), the National Center of Competence in Research in Molecular Systems Engineering (182895). LS was supported by the grant #2021-446 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain and by the Anniversary Foundation of Swiss Life for Public Health and Medical Research. Li Tang acknowledges the grant support from Swiss National Science Foundation (315230_204202, IZLCZ0_206035, CRS25_205930), European Research Council under the ERC grant agreement MechanoIMM (805337), and Swiss Cancer Research Foundation (KFS-4600-08-2018).

## 2.1 Abstract

Protein-based therapeutics such as monoclonal antibodies and cytokines are important therapies in various pathophysiological conditions such as oncology, auto-immune disorders, and viral infections. However, the wide application of such protein therapeutics is often hindered by dose-limiting toxicities and adverse effects, namely cytokine storm syndrome, organ failure and others. Therefore, spatiotemporal control of the activities of these proteins is crucial to further expand their application. Here, we report the design and application of small molecule-controlled switchable protein therapeutics by taking advantage of a previously engineered OFF-switch system. We used Rosetta modeling suite to computationally optimize the affinity between B-cell lymphoma 2 (Bcl-2) protein and a previously developed computationally designed protein partner (LD3) to obtain a fast and efficient heterodimer disruption upon addition of a competing drug (Venetoclax). The incorporation of the engineered OFF-switch system into anti-CTLA4, anti-HER2 antibodies or an Fc-fused IL-15 cytokine demonstrated an efficient disruption in vitro, as well as fast clearance in vivo upon addition of the competing drug Venetoclax. These results provide a proof-of-concept for the rational design of controllable biologics by introducing a drug-induced OFF-switch into existing protein-based therapeutics.

## 2.2 Main text

Protein-based therapeutics, such as monoclonal antibodies (mAbs) and cytokines, have shown to mediate potent antitumor effects and are the fastest growing group of therapeutics [143,144]. Nevertheless, their therapeutic use is limited by systemic toxicities arising from excessive immune and inflammatory responses, and by off-target effects [145,146]. Innovative engineering strategies have been applied to increase safety through localized activity of the therapeutic [147–149] or drug-induced ON-switch system [150]. However, none of these approaches allows the direct OFF-switch control of the therapeutics' activity with an external trigger that can be applied as desired. A system that allows the spatiotemporal control of biological activities upon administration of clinically-approved small molecules represents a promising strategy to increase protein therapeutics' safety profile. Several prior studies focused on modulating protein-protein interactions (PPIs) using small molecules to trigger either disruption or dimerization [151–155]. We previously reported a novel chemically-disruptable heterodimer composed (CDH) of a BH3-motif grafted and computationally improved protein (LD3) binding to B-cell lymphoma-extra large (Bcl-$X_L$) or B-cell lymphoma 2 protein (Bcl-2) with high affinity

[97,105]. The heterodimers can be disrupted by A-1155463 and Venetoclax, respectively. However, this approach has never been used to control the activity of a soluble protein therapeutic. Here, we computationally optimized the interface of the CDH for enhanced drug sensitivity and faster disruption. We used the optimized CDH to disrupt the Fc region from a therapeutic protein to control its half-life. Our results demonstrate the potential of designed OFF-switches for generating biologics with enhanced safety and broader applications.

To generate switchable antibodies (SwAbs), we placed the LD3:Bcl-2 complex, that can be disrupted by Venetoclax, between the epitope-binding region and the fragment crystallizable (Fc) region of the antibody (Figure 2.1A). Fc regions are crucial for antibodies as they provide important features such as: i) longer half-life *in vivo* [156], ii) increased avidity effect due to the dimerization [157] and iii) an ability to trigger effector functions [158]. We hypothesized that the addition of Venetoclax would compete for the LD3-binding site on Bcl-2 and trigger disruption between the two components. As a result, the epitope-binding domain would lose the advantages provided by the Fc-region, leading to an indirect OFF-switch of the biological activity.

We first generated a switchable version of a published $\alpha$ CTLA4 fragment antigen-binding region (Fab, Ipilimumab) [159], and tested the disruption efficiency by detecting the complex and monomeric components by size-exclusion chromatography combined with multi-angle light scattering (SEC-MALS). However, Venetoclax did not trigger detectable SwAb disruption as monomeric components were not observed (Figure 2.1B, Supplementary table S2.1). Similar observations were obtained when replacing the therapeutic moiety fused to LD3 by an $\alpha$ HER2 single-chain variable fragment (scFv) or a mouse interleukin-15 superagonist (IL15SA) (Supplementary Fig. S2.1). We therefore hypothesized that the low-nanomolar affinity of the LD3:Bcl-2 complex (Table 2.1) does not allow an efficient competition by the drug, most probably due to the slow dissociation rate ($k_{off}$) that restricts the opportunity of the drug to displace the LD3 binder. With these considerations, we aimed to further engineer LD3 for reduced affinity for Bcl-2. We used the protein modeling framework Rosetta, to conduct a computational alanine-scan on all LD3 interface residues to highlight alanine mutants with increased computed binding energy ($\Delta\Delta G$) (Figure 2.1C). All mutations to alanine increasing the $\Delta\Delta G$ by 2 Rosetta Energy Unit (R.E.U) were considered as potential LD3 variant candidates (v1 to v5), except for G137A which introduces a steric clash likely to be considerably deleterious for binding.

The remaining five LD3 variants were expressed, purified and tested by surface plasmon resonance (SPR) for binding Bcl-2 compared to the original LD3 protein (Figure 2.1C-D, Table 2.1 & Supplementary Fig. S2.2). We sought to find variants with slightly decreased dissociation rate ($k_{off}$) compared to the original LD3, but with unperturbed association rate ($k_{on}$). Variants 2 (I136A) and 5 (K144A) showed only minor differences to the original LD3, and were not further considered. Variant 3 (D138A) had the highest destabilization effect, which is consistent with the high $\Delta\Delta G$ difference predicted by the alanine scan. Both variants 1 (L133A) and 4 (F140A) showed similar mild decreases in dissociation rates, however variant 4 had a less affected association rate and was therefore chosen as a lead candidate for the switchable antibody system.

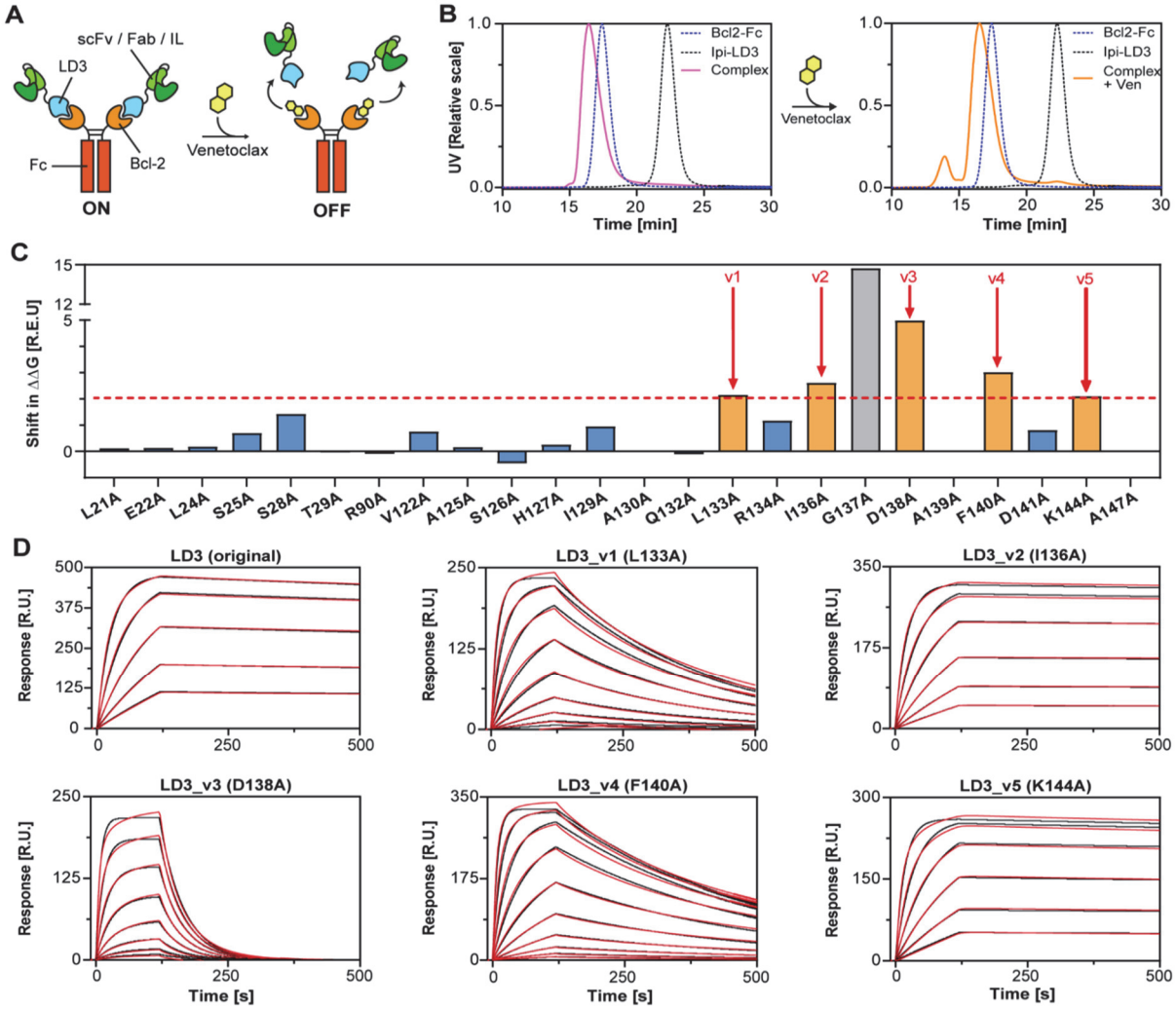



**Figure 2.1 : Computational design and improvement of a switchable antibody system. A.** Schematic representation of the switchable antibody system. A single-chain variable fragment (scFv) or fragment antigen-binding region (Fab) or interleukin (IL) is fused to a computational design (LD3) with high affinity to the Fc-fused Bcl-2. The addition of Venetoclax binds to the LD3-binding site on Bcl-2 and triggers disruption of the switchable antibody. **B.** SEC-MALS of an $\alpha$ CTLA4 Fab fused to LD3 (Ipi-LD3, gray dashed line), a Fc-fused Bcl-2 (blue dashed line), the switchable antibody complex (pink, left part) and the switchable antibody complex incubated with Venetoclax (orange, right part). **C.** Computational alanine scan obtained with Rosetta. Mutations to alanine giving an increase of the computed binding energy (ΔΔG) of at least two Rosetta Energy Unit (R.E.U.) were considered as variant candidates (Orange bars). G137A mutation was not considered (Gray bar). **D.** Surface plasmon resonance (SPR) with Bcl-2 binding to different immobilized LD3-variants (v1 to v5). Measurements are indicated in red and fit curves in black. Highest concentration of Bcl-2 starts at 2000 nM for LD3 variant 1, 3 and 4, and starts at 500 nM for original LD3, variant 2 and 5. A 2-fold dilution factor was then applied between each concentration.

**Table 2.1 : Summary table of the affinities surface plasmon resonance data of the different LD3 variants**. Data was collected using SPR showing the association rate ($k_{on}$), dissociation rate ($k_{off}$) and dissociation constant ($K_D$) of the original LD3 and the different variants obtained by computational alanine scanning. Data represent mean ± standard deviation from three independent experiments.

| | LD3 (Original) | LD3_v1 | LD3_v2 | LD3_v3 | LD3_v4 | LD3_v5 |
|---|---|---|---|---|---|---|
| $k_{on}$ [$10^4 M^{-1}s^{-1}$] | 29.1 ± 2.7 | 4.08 ± 0.34 | 36.0 ± 3.15 | 15.4 ± 0.26 | 24.0 ± 2.12 | 45.5 ± 3.70 |
| $k_{off}$ [$10^{-4}s^{-1}$] | 1.23 ± 0.65 | 26.3 ± 7.75 | 0.89 ± 0.44 | 184 ± 30.4 | 19.7 ± 5.45 | 0.66 ± 0.67 |
| $K_D$ [nM] | 1.40 ± 0.86 | 65.8 ± 25.1 | 0.74 ± 0.28 | 358 ± 55.8 | 27.3 ± 15.9 | 0.46 ± 0.41 |

We used LD3 variant 4 (LD3_v4) to generate an improved version of the switchable Ipilimumab-based $\alpha$ CTLA4 antibody by fusing the Ipilimumab Fab to LD3_v4. After complex formation with Bcl2-Fc, we assessed the switchability using SEC-MALS as described above. While only 3% ($m_{uncomplex}/m_{total}$) of the switchable antibodies were disrupted on SEC-MALS upon Venetoclax treatment with the original LD3 protein, more than 90% of the complex efficiently disrupted with LD3_v4 (Figure 2.2A, Supplementary Table S2.1). Similarly, we noticed comparable results with the $\alpha$ HER2 and IL-15SA switchable therapeutics, demonstrating the modularity of the system (Supplementary Fig. S2.3). We evaluated disruption kinetics by biolayer interferometry (BLI) and detected 30% disruption at the highest tested concentration of Venetoclax (10 $\mu$ M) after 200 seconds (Supplementary Fig. 2.2B). During that time, the switchable antibody complex remained stable in solution without addition of Venetoclax.

To confirm these results in a cell-based assay, we substituted the antigen-targeting domain of the SwAb with an $\alpha$ HER2 scFv that allowed the labeling of HER2-expressing cells. We stained MC38-HER2 cells, a murine colon adenocarcinoma cell line stably expressing HER2, with the switchable $\alpha$ HER2 antibody and treated the cells with or without Venetoclax (Figure 2.2C and Supplementary Fig. S2.4). One hour after adding Venetoclax, the Fc fragment detected on MC38 cell surface decreased by 2-fold. Among other possibilities, the reduced Venetoclax-induced antibody disruption might be explained by the avidity provided by the two Fabs binding simultaneously, which may reduce drug sensitivity. The switchable $\alpha$ HER2 antibody showed similar binding to MC38-HER2 cells compared to a conventional $\alpha$ HER2 antibody, which did not respond to Venetoclax, and no disruption was observed when using the original LD3 protein (Supplementary Fig. S2.4). Altogether, these results confirm the improved switchability of the engineered antibody.

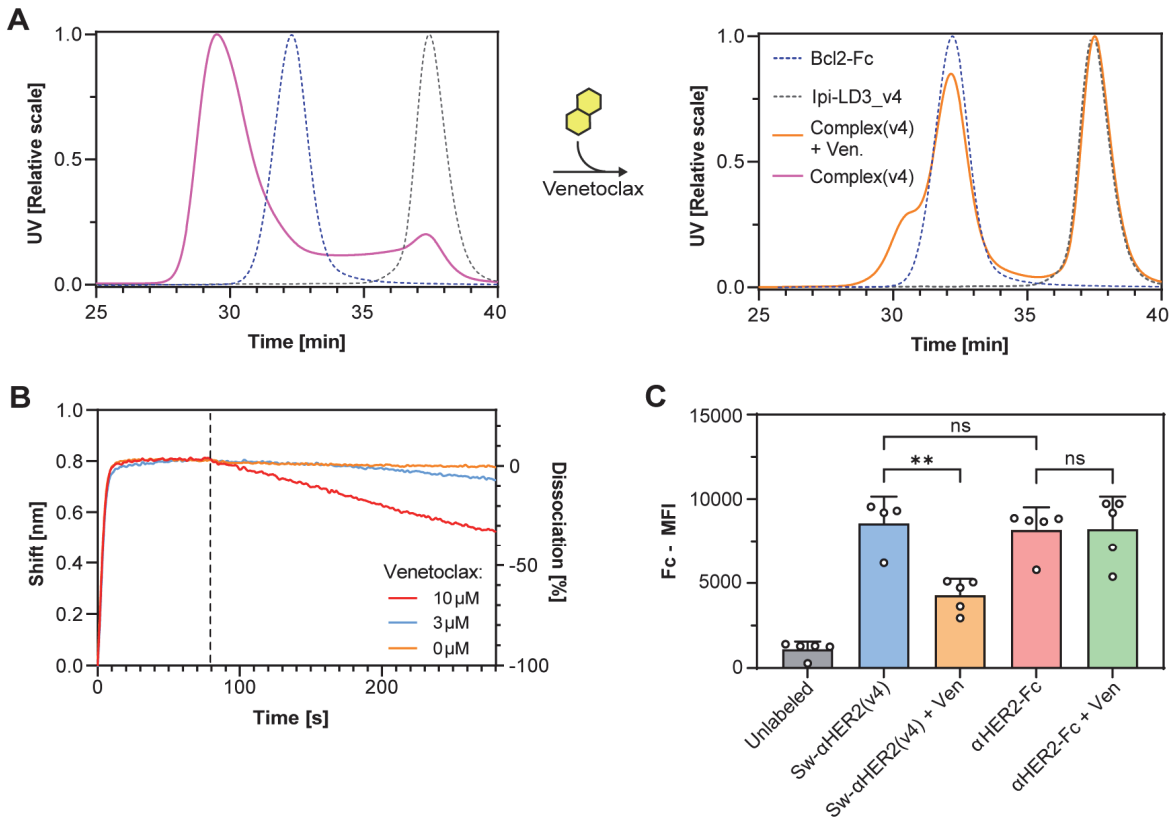



**Figure 2.2 : Disruption efficiency of a switchable antibody with LD3_v4. A.** SEC-MALS of a Bcl2-Fc alone (blue dashed line), an $\alpha$ CTLA4 Fab (Ipilimumab) fused to LD3 variant 4 (gray dashed line) and the switchable antibody complex in absence (pink) and presence (orange) of 100 $\mu$ M Venetoclax. **B.** Biolayer interferometry (BLI) measurements of the switchable anti-CTLA4 antibody with increasing concentration of Venetoclax. **C.** Quantification of the mean fluorescence intensity (MFI) measured on the surface of MC38 cells unlabeled or labeled with with a switchable or conventional $\alpha$ Her2 antibody (Sw-$\alpha$ HER2 and $\alpha$ HER2-Fc respectively) treated without or with 10 $\mu$ M Venetoclax. Tukey's multiple comparisons test, $p<0.01$ (**), non-significant (ns). Data points represent technical replicates with mean and standard deviation.

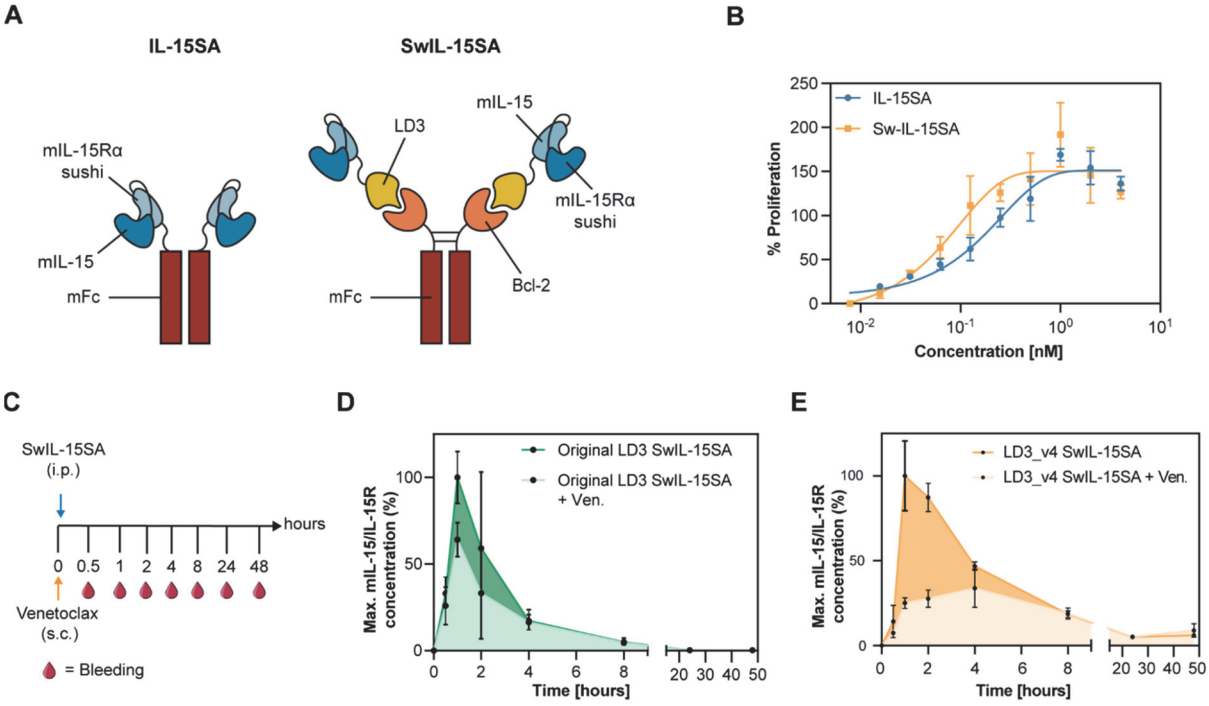



**Figure 2.3 : Functional assessment and *in vivo studies* using an Fc-fused switchable cytokine.** **A.** Schematic representation of the switchable interleukin system. In IL-15SA, the sushi domain of mouse IL-15R$\alpha$ is fused to mIL-15 binding a mousse Fc (left). In SwIL-15SA, the sushi domain of mouse IL-15R$\alpha$ is fused to the optimized LD3 binding to mouse Fc-fused Bcl-2 (right). **B.** Activated mouse T cell proliferation in response to IL-15SA or SwIL-15SA. **C.** C57BL/6 mice were first injected subcutaneously (s.c.) with Venetoclax (25.0 mg/kg) and subsequently injected intraperitoneally (i.p.) with 100 pmol SwIL-15SA. Mice were bled overtime after 0.5, 1, 2, 4, 8, 24, and 48 hours after treatment. **D.** Pharmacokinetic properties of SwIL-15SA composed of the IL-15/IL-15R complex fused to the original LD3 with (light green) or without (dark green) the administration of Venetoclax. **E.** Pharmacokinetic properties of SwIL-15SA composed of the IL-15/IL-15R complex fused to LD3_v4 with (light orange) or without (dark orange) the administration of Venetoclax.

We next tested the function of the engineered switchable proteins in vitro and in vivo by measuring cell proliferation and the half-life in mice blood. To do so, we extended the strategy to the generation of switchable cytokines. We chose mouse IL-15 superagonist (IL-15SA) and generated switchable IL-15SA (SwIL-15SA) by fusing IL-15 and the IL-15 receptor $\alpha$ domain (IL-15R$\alpha$ ) to LD3 assembled with Bcl2-Fc (Figure 2.3A).To assess the functionality of SwIL-15SA, we stimulated primary murine T cells *ex vivo* with either IL-15SA or SwIL-15SA and measured cell proliferation. Proliferation of murine primary T cells induced by SwIL-15SA was comparable to conventional IL-15SA, indicating that fusing LD3 to the sushi domain of IL-15R$\alpha$ did not hinder its functionality (Figure 2.3B).

In a second step, we assessed the switchability of SwIL-15SA *in vivo*. C57BL/6 mice were first injected subcutaneously (s.c.) with or without Venetoclax and then intraperitoneally (i.p.) with SwIL15-SA containing the original LD3. Mice were bled overtime after treatment and IL-15/IL15R complex concentration was measured by enzyme-linked immunosorbent assay (ELISA) (Figure 2.3C). IL-15/IL-15R complex concentration in blood of mice treated with Venetoclax peaked at 64% of the maximum IL-15/IL15R concentration of the control group, confirming that Venetoclax administration does not lead to the efficient disruption of the original LD3:Bcl-2 complex, as demonstrated in *in vitro* experiments (Figure 2.3D and Supplementary Fig. S2.5). To investigate whether the affinity of the Bcl-

2:LD3 complex could provide a parameter to tune the switchability efficiency of the system, we further tested a variant of SwIL-15SA being composed of the IL-15 sushi domain fused to LD3_v4. Here, blood concentrations in control mice peaked at 1 hour after injection and then decreased overtime (Figure 2.3E and Supplementary Fig. S2.5). Unlike the control group, in mice treated with Venetoclax the IL-15/IL15R complex concentration reached only about 25% of the maximum IL-15/IL15R concentration of the control group. This observation suggests that the disruption efficiency and the half-life of the system can be tuned with the affinity of the Bcl-2:LD3 complex. Overall, these results show that Venetoclax disrupts the interaction between Bcl-2 and LD3, leading to the fast clearance of monomeric IL-15/IL-15R-LD3 *in vivo*.

Altogether, we show a modular and generalizable OFF-switch approach for the design of safe antibody and cytokine therapeutics by introducing a chemically-disruptable heterodimer between the therapeutic domain and the Fc moiety. Loss of the Fc-fragment leads to a decrease of the avidity effect and a drastic reduction of the protein half-life. We took advantage of a previously designed CDH that can be competed by a clinically-approved drug, Venetoclax, which makes it a good candidate for translational applications. Of note, one strength of our system is its modularity with the ease to adapt it to several therapeutic proteins by exchanging the therapeutic domain fused to LD3. But the large size of the protein complex (of about ~250 kDa for a switchable antibody, compared to ~150 kDa for a normal antibody) may limit tissue penetration [160]. However, for highly toxic therapies, such as immunostimulatory therapies, these limitations would be outweighed by the improved safety profile. Our presented workflow to reduce heterodimer affinity to increase drug sensitivity can likely be readily extended to other examples of CDHs. These types of switchable biologics could serve as a basis for safer biologics for therapeutic use.

## 2.3 Methods

*Computational design*

Previously solved crystal structure of Bcl-2 in complex with LD3 was used for computational modeling (PDB ID: 6IWB). Using the Rosetta modeling suite, the pose was relaxed with the "FastRelax" mover, before the computational alanine scan was performed using an "Alascan" filter. Residues where a mutation to alanine lead to an increase of the computed binding energy of >2 Rosetta energy units (R.E.U.) were considered as potential candidates to lower the affinity of LD3 for Bcl2. Mutations exceeding 5 R.E.U. were not considered due to the introduction of clashes that may abrogate binding.

*Protein expression and purification*

The engineered IL-15SA construct (gWIZ-mIL-15SA) was a gift from D. J. Irvine (MIT). IL-15SA contains a mouse IL-15 fused at the C terminus of Sushi domain of a mouse IL-15R$\alpha$ , which is next fused at the C terminus with a mouse IgG2c Fc. A previously optimized version of Bcl-2 [108] was fused to either human IgG or mouse IgG2 Fc-fragment (see Supplementary Table S2.2). Switchable antibodies were composed of either a previously published $\alpha$ CTLA4 antibody Ipilimumab [159] as a Fab or an $\alpha$ HER2

4D5 clone [161] as an scFv fused to LD3 protein N-terminal with a $(GGGS)_3$-linker. As a switchable cytokine, we used a fusion protein composed of mouse IL-15 C-terminally fused to the IL-15 receptor $\alpha$ domain (IL-15R$\alpha$ ), itself fused to the LD3 variant C-terminal with a $(GGGS)_3$-linker. DNA sequences were ordered from Twist Bioscience and Gibson cloning used to clone into bacterial (pET11) or mammalian (pHLSec) expression vectors. Mammalian expressions were performed using the Expi293TM expression system from Thermo Fisher Scientific. Supernatant was collected 6 days post transfection, filtered, and purified. E. coli expressions were performed using BL21 (DE3) cells and IPTG induction (1 mM at OD 0.6-0.8) and growth overnight at 16-18° C. Pellets were lysed in lysis buffer (50 mM Tris, pH 7.5, 500 mM NaCl, 5% Glycerol, 1 mg/ml lysozyme, 1 mM PMSF, and 1 µg/ml DNase) with sonication, the lysate clarified, and purified. Proteins were then purified using an ÄKTA pure system (GE healthcare) with Ni-NTA affinity columns followed by size exclusion chromatography with PBS.

*Surface plasmon resonance*

SPR measurements were performed on a Biacore 8K (GE Healthcare) with HBS-EP+ as running buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.005% v/v Surfactant P20, GE Healthcare). Original LD3 and mutants were immobilized on a CM5 chip (GE Healthcare # 29104988) via amine coupling. 500-1000 response units (RU) were immobilized and Bcl-2 was injected as an analyte in serial dilutions. The flow rate was 30 µl/min for a contact time of 120s followed by 400s dissociation time. After each injection, the surface was regenerated using 50 mM NaOH. SPR Data were fit with 1:1 Langmuir binding model within the Biacore 8K analysis software (GE Healthcare #29310604).

*Bio-layer interferometry (BLI)*

Measurements were performed on a Gator BLI system. The running buffer was PBS. Fc-tagged Bcl-2 were diluted to 5 $\mu$ g/mL and immobilized on the anti-human IgG tips for 80 seconds (1-2 nm immobilized). The loaded tips were then dipped into 500 nM LD3-fused Ipilimumab Fab (or PBS for the reference) for 80 seconds and then in different concentrations of Venetoclax (10, 3 and 0 $\mu$ M) diluted in PBS for 210 seconds. Each measurement was subtracted with the reference (channel with Fc-fused Bcl2 immobilized, no associated LD3 and a corresponding concentration of Venetoclax diluted in PBS).

*Size exclusion chromatography multi-angle light scattering (SEC-MALS)*

Size exclusion chromatography with an online multi-angle light scattering device (miniDAWN TREOS, Wyatt) was used to determine the oligomeric state and molecular weight for the switchable antibodies in solution. Purified LD3-Fab and Bcl2-Fc proteins were mixed with a 2:1 molar ratio and incubated at room temperature for 5 min to form a complex. Assembled complexes received 100 $\mu$ M Venetoclax or PBS and incubated 1h at 37°C. Final concentration was approximately 1 mg/ml in PBS (pH 7.4), and 100 µl of the sample was injected into a Superdex 75 300/10 GL column (GE Healthcare) with a flow rate of 0.5 ml/min, and UV280 and light scattering signals were recorded. Molecular weight was determined using the ASTRA software (version 6.1, Wyatt).

*In vitro cell binding assay*

100'000 HER2-transduced MC38 mouse colon cancer cells were collected in a tube. Purified HER2-specific LD3-Fab and Bcl-2-Fc proteins were mixed at a 2:1 ratio and incubated at room temperature for 5 minutes to form a complex. MC38-HER2+ cells were then stained with $\alpha$ HER2 SwAb at concentrations of 100 nM and incubated at 4 °C for 30 min. An Fc-fused $\alpha$ HER2 ($\alpha$ HER2_Fc) was used as a positive control. Cells were washed twice with FACS buffer (PBS containing bovine serum albumin, 0.2% (w/v)) and 10 μM Venetoclax was added to the cells and incubated at 37 °C for 1 hour. Following, cells were washed and stained with anti-human Fc antibody at 4 °C fr 30 min. Cells were then washed, stained with 4′,6-diamidino-2-phenylindole (DAPI; Sigma-Aldrich) and analyzed by FACS.

*T-cell proliferation assay*

Activated Pmel T cells were collected by centrifugation, re-suspended in mouse T-cell media and seeded at a density of 10'000 T cells/well in a 96-well flat bottom tissue culture plate. T cell growth was stimulated by addition of serial dilutions of IL-15SA or Sw-IL15SA to a total volume of 100 μL and cultured for 48 hours at 37 °C. On day 2, cells were collected, washed once with FACS buffer and stained with DAPI. Cell counts for each condition were quantified by FACS using the Attune NxT flow cytometer (Invitrogen/Thermo Fisher Scientific).

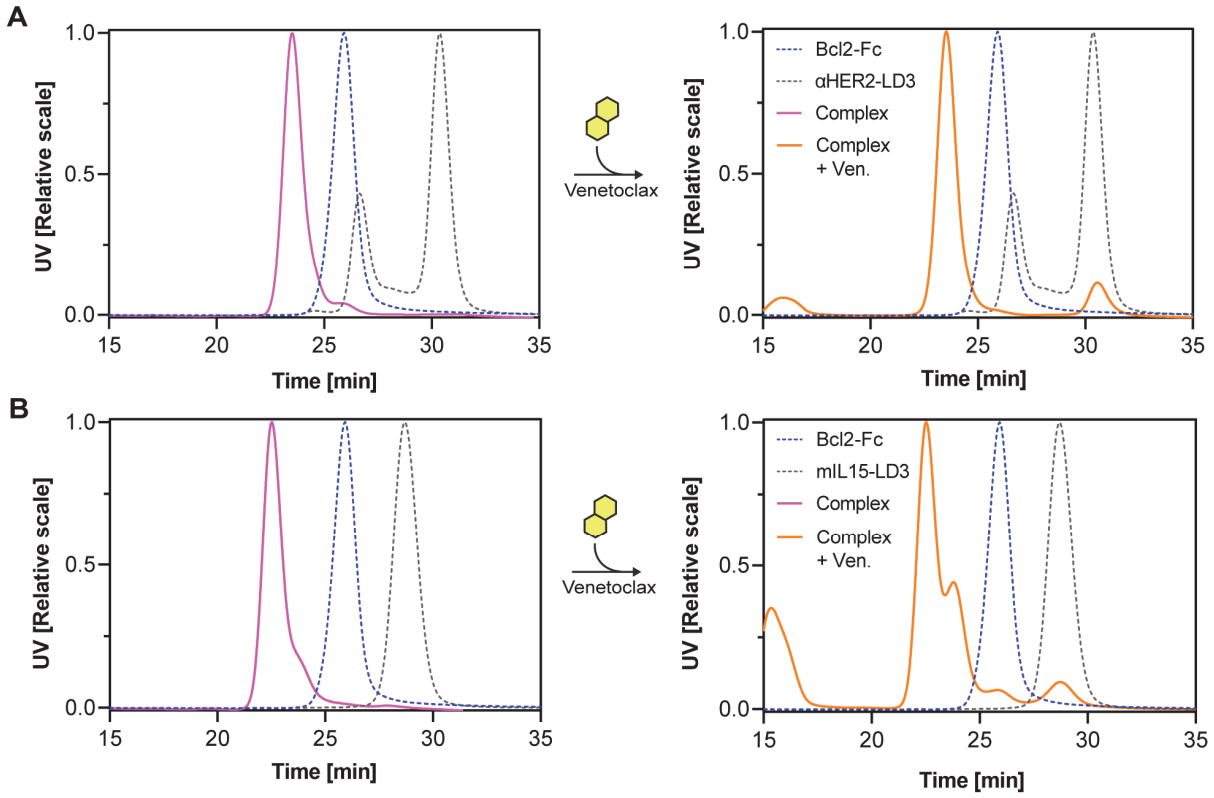
*Animal studies*

6-8 week-old female C57BL/6 mice were purchased from Charles River Laboratories and maintained in the animal core facility [Center of Phenogenomics (CPG)] of École Polytechnique Fédérale de Lausanne (EPFL). All experiments were conducted according to the Swiss Federal Veterinary Office guidelines and were approved by the Cantonal Veterinary Office. In evaluating the switchability potential of SwIL-15SA, C57BL/6 mice were injected subcutaneously with 100 μl Venetoclax dissolved at 25 mg/kg in a solution of saline and 2% dimethyl sulfoxide (DMSO). Following, the animals were injected intraperitoneally with 100 pmol of Sw-IL15SA in 100 μl and bled overtime at 0.5, 1, 2, 4, 8, 24, and 48 hours after treatment. IL-15/IL-15R complex concentration in blood was quantified using a commercial enzyme-linked immunosorbent assay (ELISA) kit following manufacturer's instructions (Thermo Fisher Scientific, 88-7215-88).

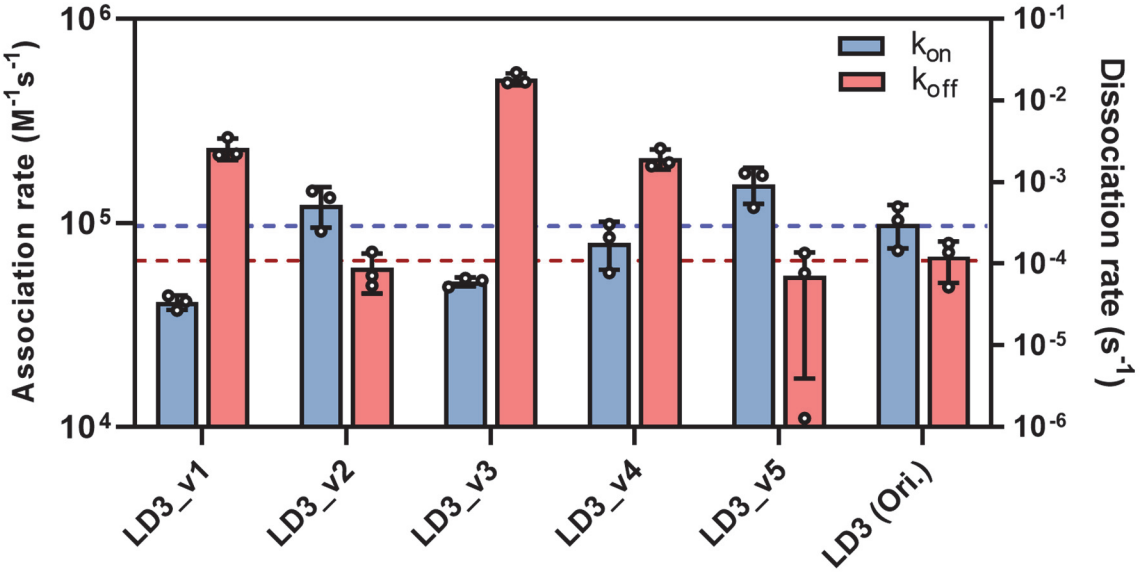
*Acknowledgment*

We thank the EPFL animal facility (CPG), for their support for conducting animal experiments, and the flow cytometry core facility (FCCF) for their assistance. We also thank the high-performance computing facility at EFPL – SCITAS for the computational resources.

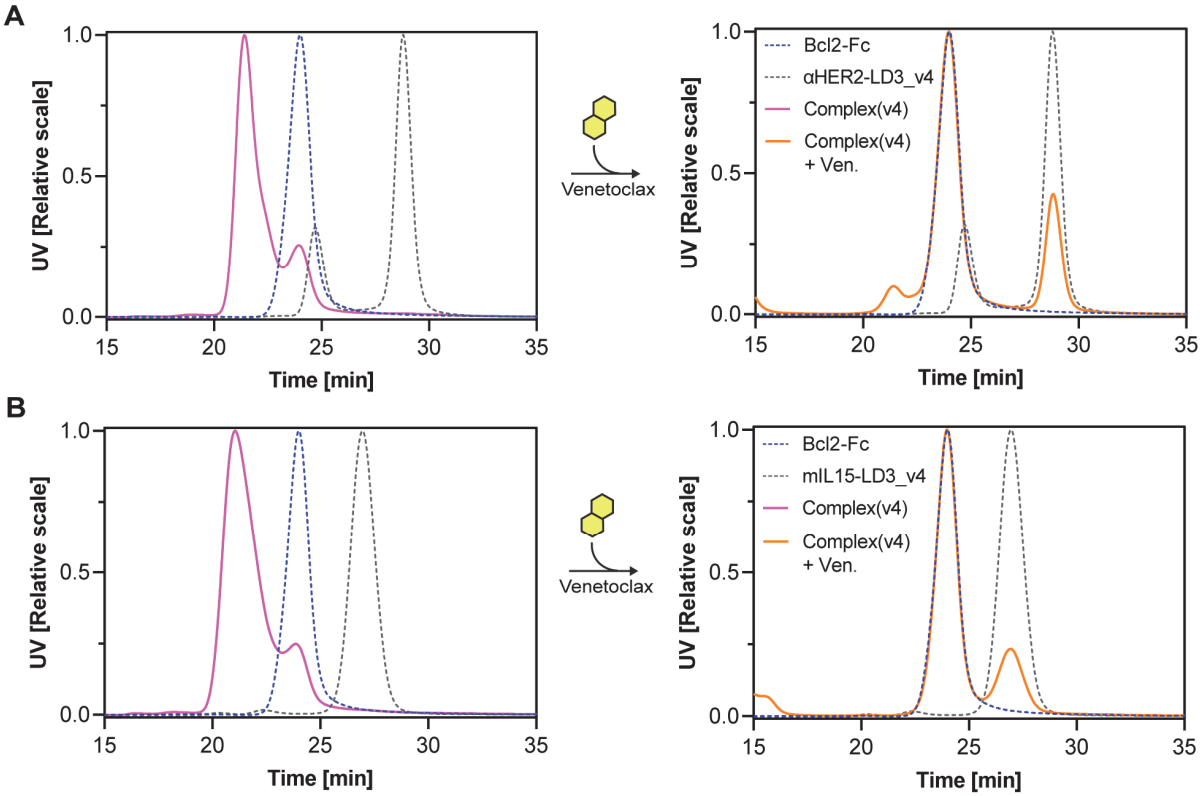
## 2.4 Supplementary materials



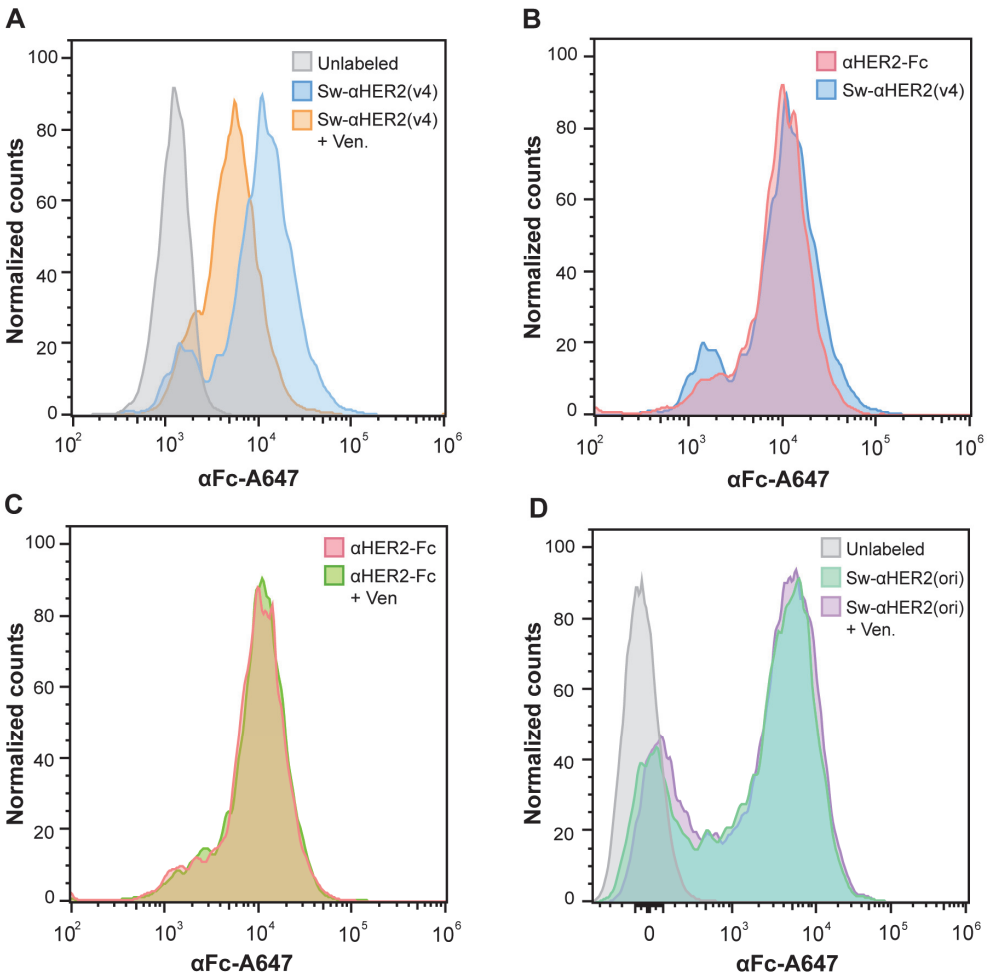

**Supplementary Figure S 2.1 : Size exclusion chromatography of anti-HER2 antibody and mouse interleukin 15 fused to the original LD3 protein.** Size exclusion chromatography of an $\alpha$ HER2 single-chain variable fragment (**A**) or a mouse interleukin 15 (**B**) fused to the original LD3. Plots show the switchable protein therapeutic complex in absence (pink) or in presence (orange) of Venetoclax, compared to the Bcl2-Fc (blue, dashed line) or original LD3-fused moiety alone (gray, dashed line)



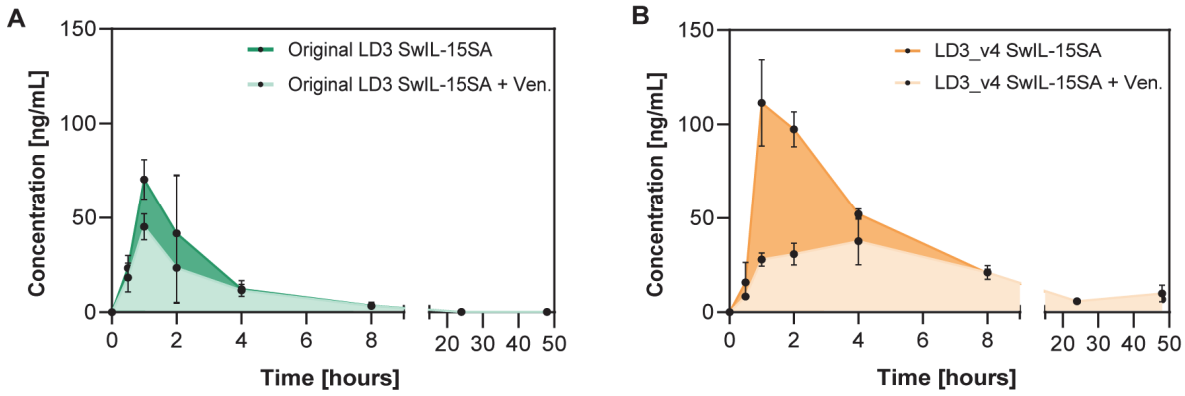

**Supplementary Figure S 2.2 : Kinetic measurements of the different LD3 variants.** Association rates ($k_{on}$) are shown with blue bars and dissociation rate ($k_{off}$) with red bars. Dashed lines show the mean values for the original LD3 (LD3 Ori.) for comparison. Data points represent mean ± standard deviation from three independent experiments.




**Supplementary Figure S 2.3 : Size exclusion chromatography of *α* HER2 antibody and mouse interleukin 15 fused to the LD3_v4 protein.** Size exclusion chromatography of an *α* HER2 single-chain variable fragment (**A**) or a mouse interleukin 15 superagonist (IL-15SA) (**B**) fused to LD3_v4. Plots show the switchable protein therapeutic complex in absence (pink) or in presence (orange) of Venetoclax, compared to the Bcl2-Fc (blue, dashed line) or LD3_v4-fused moiety alone (gray, dashed line)




**Supplementary Figure S 2.4 : HER2-Overexpressing cells MC38 labeling and controls. A.** Histogram of MC38 cells labeled with switchable *α* Her2 antibody (Sw-*α* HER2) composed of LD3_v4 in presence or absence of venetoclax (Ven.), **B.** Histogram of MC38 cells labeled with switchable *α* Her2 antibody (Sw-*α* HER2) or conventional *α* Her2 antibody (*α* HER2-Fc). **C.** Histogram of MC38 cells labeled with conventional *α* Her2 antibody (*α* HER2-Fc) in presence or absence of Venetoclax (Ven.) **D.** Histogram of MC38 cells labeled with switchable *α* Her2 antibody (Sw-*α* HER2) composed of the original LD3 (ori.) in presence or absence of Venetoclax (Ven.)




**Supplementary Figure S 2.5 : *In vivo* studies using an Fc-fused switchable cytokine (Absolute scale) A.** Pharmacokinetic properties of SwIL-15SA composed of the IL-15/IL-15R complex fused to the original LD3 with (light green) or without (dark green) the administration of Venetoclax. **B.** Pharmacokinetic properties of SwIL-15SA composed of the IL-15/IL-15R complex fused to LD3_v4 with (light orange) or without (dark orange) the administration of Venetoclax. Data points represent mean ± standard deviation from three biological replicates.

**Supplementary Table S 2.1 : Mass fraction of the different SwAb components measured by the SEC-MALS upon Venetoclax treatment.** Switchable antibody/interleukin complexes were assembled *in vitro* with either the original LD3 or the variant 4 (LD3_v4), treated with venetoclax, and analyzed by size-exclusion chromatography coupled to multi-angle light scattering (SEC-MALS). The mass fraction of the different peaks shown in figures 2.1B, 2.2A, S2.1 and S2.3 were measured.

| | | Mass fraction (%) | | | |
|---|---|---|---|---|---|
| | | Full complex | Partial complex | Bcl2-Fc | LD3-fused moiety |
| Ipilimumab | Original LD3 | 97% | N/D | 3% | N/D |
| | LD3_v4 | 9.6% | N/D | 36% | 54.4% |
| $\alpha$ HER2 | Original LD3 | 87.6% | N/D | 1.4% | 11% |
| | LD3_v4 | 7% | N/D | 65.3% | 27.7% |
| IL-15/IL-15R | Original LD3 | 62.1% | 22.1% | 5.2% | 10.6% |
| | LD3_v4 | N/D | N/D | 62.8% | 37.2% |

**Supplementary Table S 2.2 : Amino acid sequences of the different proteins used. A** stabilized version of Bcl2 fused to either human IgG1 (Bcl2-hmFc) or mouse IgG2 (Bcl2-mFc), a previously defined Bcl2-binding protein (LD3), an anti-CTLA4 Fab (Ipilimumab_H and Ipilimumab_L), a mouse interleukin 15 (mIL15), an anti-HER2 single chain fragment clone 4D5 ($\alpha$ HER2_scFv) and the same fused to an IgG ($\alpha$ HER2_Fc).

| | |
|---|---|
| Bcl2-hmFc | AHAGRTGYDNREIVMKYIHYKLSQRGYEWDAGDDAEENRTEAPEGTESEVVHRALRDAGD<br>DFERRYRRDFAEMSSQLHLTPDTARQRFETVVEELFRDGVNWGRIVAFFEFGGVMCVESVN<br>REMSPLVDNIAEWMTEYLNRHLHTWIQDNGGWDAFVELYGPSMRGGGGSGTDKTHTCP<br>PCPAPELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAK<br>TKPREEQYNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYT<br>LPPSREEMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTV<br>DKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGKHHHHHH |
| Bcl2-mFc | AHAGRTGYDNREIVMKYIHYKLSQRGYEWDAGDDAEENRTEAPEGTESEVVHRALRDAGD<br>DFERRYRRDFAEMSSQLHLTPDTARQRFETVVEELFRDGVNWGRIVAFFEFGGVMCVESVN<br>REMSPLVDNIAEWMTEYLNRHLHTWIQDNGGWDAFVELYGPSMRGGGGSEPRVPITQNP<br>CPPLKECPPCAAPDLLGGPSVFIFPPKIKDVLMISLSPMVTCVVVAVSEDDPDVQISWFVNN<br>VEVHTAQTQTHREDYNSTLRVVSALPIQHQDWMSGKEFKCKVNNRALPSPIEKTISKPRGP<br>VRAPQVYVLPPPAEEMTKKEFSLTCMITGFLPAEIAVDWTSNGRTEQNYKNTATVLDSDGSY<br>FMYSKLRVQKSTWERGSLFACSVVHEGLHNHLTTKTISRSLGKGTKHHHHHH |
| LD3 | GQRWELALGRFLEYLSWVSTLSEQVQEELLSSQVTQELRALMDETMKELKAYKSELEEQLTP<br>VAEETRARLSKELQAAQARLGADMEDVRGRLVQYRGEVQAMLGQSTEELRVRLASHLIALQ<br>LRLIGDAFDLQKRLAVYQAGA |
| Ipilimumab<br>_H | QVQLVESGGGVVQPGRSLRLSCAASGFTFSSYTMHWVRQAPGKGLEWVTFISYDGNNKYY<br>ADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAIYYCARTGWLGPFDYWGQGTLVTVSSAST<br>KGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLS<br>SVVTVPSSSLGTQTYICNVNHKPSNTKVDKRVEPKSC |
| Ipilimumab<br>_L | EIVLTQSPGTLSLSPGERATLSCRASQSVGSSYLAWYQQKPGQAPRLLIYGAFSRATGIPDRFS<br>GSGSGTDFTLTISRLEPEDFAVYYCQQYGSSPWTFGQGTKVEIKRTVAAPSVFIFPPSDEQLKS<br>GTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEK<br>HKVYACEVTHQGLSSPVTKSFNRGEC |
| mIL15 | GTTCPPPVSIEHADIRVKNYSVNSRERYVCNSGFKRKAGTSTLIECVINKNTNVAHWTTPSLK<br>CIRDPSLAGGSGGSGGSGGSGGSGGSGGNWIDVRYDLEKIESLIQSIHIDTTLYTDSDFHPS<br>CKVTAMNCFLLELQVILHEYSNMTLNETVRNVLYLANSTLSSNKNVAESGCKECEELEEKTF<br>TEFLQSFIRIVQMFINTSHHHHHH |
| $\alpha$ HER2_scF<br>v | EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYA<br>DSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDYWGQGTLVTVSSG<br>GGGSGGGGSGGGGSDIQMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKL<br>LIYSASFLYSGVPSRFSGSRSGTDFTLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVEIK |
| $\alpha$ HER2_Fc | DYKDIVMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPS<br>RFSGSRSGTDFTLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVELKRATPSHNSHQVPSAG<br>GPTANSGEVKLVESGGGLVQPGGSLRLSCATSGFNIKDTYIHWVRQAPGKGLEWVARIYPT<br>NGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDYWGQGT<br>TVTVSSTGVHSEPRVPITQNPCPPLKECPPCAAPDLLGGPSVFIFPPKIKDVLMISLSPMVTC<br>VVVAVSEDDPDVQISWFVNNVEVHTAQTQTHREDYNSTLRVVSALPIQHQDWMSGKEFKC<br>KVNNRALPSPIEKTISKPRGPVRAPQVYVLPPPAEEMTKKEFSLTCMITGFLPAEIAVDWTSN<br>GRTEQNYKNTATVLDSDGSYFMYSKLRVQKSTWERGSLFACSVVHEGLHNHLTTKTISRSL<br>GKASGSRSLLANKRSEL |

# Chapter 3

## *De novo* design of protein interactions with learned surface fingerprints

Template-based approaches, such as the one described before, have demonstrated numerous successful examples but remain limited to known protein interaction motifs. Emergent virus, hard-to-drug onco-targets and novel synthetic biology tools often require to target novel sites with no native protein partner reported, which justifies the need of *de novo* design approaches. However, targeting a protein surface without prior information rises two fundamental questions: i) which site should be targeted and ii) what motif can target this specific site. Thanks to the emergence of machine learning, new methods can now be leveraged for the study of protein surfaces and the design of novel protein interactions. In this chapter, we will repurpose a geometric deep learning framework called MaSIF (Molecular Surface Interaction Fingerprinting) which has initially been proposed for the prediction of protein interfaces and partners based solely on the molecular surface features. Here, MaSIF has been adapted for the design of site-specific novel protein interactions to provide *de novo* protein binders straight from a computer.

This section is adapted from an article published in Nature in 2023 (doi: 10.1038/s41586-023-05993-x), as allowed by the publisher (License CC-BY 4.0).

**Authors**

Pablo Gainza*[1,†], Sarah Wehrle[1,†], Alexandra Van Hall-Beauvais[1,†], Anthony Marchand[1,†], Andreas Scheck[1, †], Zander Harteveld[#1], Stephen Buckley[#1], Dongchun Ni[#2], Shuguang Tan[#3], Freyr Sverrisson[1], Casper Goverde[1], Priscilla Turelli[4], Charlène Raclot[4], Alexandra Teslenko[5], Martin Pacesa[1], Stéphane Rosset[1], Sandrine Georgeon[1], Jane Marsden[1], Aaron Petruzzella[6], Kefang Liu[3], Zepeng Xu[3], Yan Chai[3], Pu Han[3], George F. Gao[3], Elisa Oricchio[6], Beat Fierz[5], Didier Trono[4], Henning Stahlberg[2], Michael Bronstein[7,◊], Bruno E. Correia[1,◊]

**Affiliations**

[1] Laboratory of Protein Design and Immunoengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne, and Swiss Institute of Bioinformatics, Lausanne, Switzerland

[2] Laboratory of Biological Electron Microscopy, Institute of Physics, School of Basic Science, École Polytechnique Fédérale de Lausanne, and Dep. of Fund. Microbiology, Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland

[3] CAS Key Laboratory of Pathogen Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

[4] Laboratory of Virology and Genetics, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[5] Laboratory of Biophysical Chemistry of Macromolecules, School of Basic Sciences, Institute of chemical sciences and engineering (ISIC), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[6] Swiss Institute for Experimental Cancer Research, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[7] Department of Computer Science, University of Oxford, Oxford, UK

[‡] Equal contribution

[#] Equal contribution

**Author contributions**

P.G., S.W., A.V., A.M., A.S., contributed equally to this work. Z.H, S.B., D.N. contributed equally to this work. P.G., S.W., A.V., A.M., A.S., M.B. and B.E.C. conceived the work and designed the experiments. P.G., A.M., A.S. and Z.H. performed the computational design and S.W., A.V., S.B. and A.M. performed experimental characterization and optimization. P.G., F.S., A.M., Z.H., and A.S. developed the MaSIF-seed method. D.N. and H.S. solved the cryo-EM structure. S.T., M.P., K.L., Z.X., Y.C., P.H. and G.F.G solved the crystal structures. A.P. and E.O. performed the PD-L1 cell binding assay. A.T. and B.F. synthesized peptides. P.T., C.R., and D.T. performed SARS-CoV-2 binding and neutralization studies. F.S., C.G., S.R., S.G. and J.M. performed experiments and acquired data. P.G., S.W., A.V., A.M., A.S. and B.E.C. wrote the manuscript with input from all authors.

**Funding**

The DCI is an initiative of the EPFL, University of Lausanne and University of Geneva. DN and HS were supported by the Swiss National Science Foundation and the NCCR Transcure. MB was supported by an ERC Consolidator grant No. 724228. BC was supported by the Swiss National Science Foundation, the NCCR in Chemical Biology, the NCCR in Molecular Systems Engineering and the ERC Starting grant no. 716058. P.G. was sponsored by an EPFL-Fellows grant funded by an H2020 Marie Sklodowska-Curie.

## 3.1 Abstract

Physical interactions between proteins are essential for most biological processes governing life [16]. However, the molecular determinants of such interactions have been challenging to understand, even as genomic, proteomic, and structural data grows. This knowledge gap has been a major obstacle for the comprehensive understanding of cellular protein-protein interaction (PPI) networks and for the *de novo* design of protein binders that are crucial for synthetic biology and translational applications

[71,132,162–167]. We exploit a geometric deep learning framework operating on protein surfaces that generates fingerprints to describe geometric and chemical features critical to drive PPIs [138]. We hypothesized these fingerprints capture the key aspects of molecular recognition that represent a new paradigm in the computational design of novel protein interactions. As a proof-of-principle, we computationally designed several *de novo* protein binders to engage four protein targets: SARS-CoV-2 spike, PD-1, PD-L1, and CTLA-4. Several designs were experimentally optimized while others were purely generated *in silico*, reaching nanomolar affinity with structural and mutational characterization showing highly accurate predictions. Overall, our surface-centric approach captures the physical and chemical determinants of molecular recognition, enabling a novel approach for the *de novo* design of protein interactions and, more broadly, of artificial proteins with function.

## 3.2 Introduction

Designing novel protein-protein interactions (PPIs) remains a fundamental challenge in computational protein design, with broad basic and translational applications in biology. The challenge consists of generating amino acid sequences that engage a target site and form a quaternary complex with a given protein. This represents a stringent test of our understanding of the physicochemical determinants that drive biomolecular interactions [168]. Robust computational methods to design *de novo* PPIs could be used to rapidly engineer protein-based therapeutics such as antibodies and protein inhibitors or vaccines, among others, and therefore are of major interest for biomedical and translational applications [71,132,162–167].

Despite recent advances in rational PPI design [132,162,166] and prediction [31], designing novel protein binders against specific targets is very challenging , particularly when no structural elements from preexisting binders are known. Current state-of-the-art methods for *de novo* PPI design [131,132,162,169], such as hotspot-centric approaches [132] and rotamer information fields [162,166], rely on placing disembodied residues on the target interface and then optimizing their presentation on a protein scaffold. Intrinsic limitations of these approaches relate to the very weak energetic signatures provided by scoring functions to single-side chain placements, which is compounded in flat interfaces that lack deep pockets. These methods also face the challenge of finding compatible protein scaffolds to precisely display the generated constellations of residues. To circumvent these limitations, new approaches are needed to design *de novo* binders to various surface types and protein sites.

A long-standing model of molecular recognition postulates that PPIs form between protein molecular surfaces with chemical and geometric complementarity [170,171]. The complementarity features arise as a consequence of the energetic contributions that are critical to stabilize PPIs, including van der Waals interactions (geometric complementarity), hydrophobic effect, and electrostatics interactions (chemical complementarity) [170]. At the structural level, most protein interfaces contain surface regions that become inaccessible to solvent upon complex formation, which we refer to as *buried* or *core interface*, as well as patches that are involved in the interface but remain solvent-exposed, which we refer to as the *interface rim*. Residues within the buried areas tend to be much less tolerant to mutations [16,17] and have a large energetic contribution towards the PPI formation, often referred to

as hotspots. Rim regions are generally more polar and tolerant to mutations, giving also important contributions to affinity and, more notably, specificity [16,172]. Guided by these general principles of molecular recognition, we introduce a novel protein design approach based on the critical importance of the fully buried patches of the interface to drive protein interactions. We implemented these design principles by exploiting surface fingerprints learned from interacting protein surfaces which capture features that are determinant for molecular recognition. Our novel approach allows for ultra-fast and accurate prediction of privileged sites for PPI design, and reduces the complexity for hotspot search and grafting. We leveraged this design workflow to successfully engineer and characterize binders against four therapeutic targets of interest, namely SARS-CoV-2 spike, PD-1, PD-L1, and CTLA-4.

## 3.3 Results

### 3.3.1 Design strategy and in silico validation

In previous work, we introduced a geometric deep learning framework, MaSIF (Molecular Surface Interaction Fingerprinting), to generate surface fingerprints from the geometric and chemical features of molecular surfaces and learn patterns that determine the propensity of protein interactions [138]. Within this framework we developed the MaSIF-site tool to predict areas with propensity to form PPIs on the surface of proteins. MaSIF-site receives as input a protein decomposed into patches and outputs a per-vertex regression score on the propensity of each surface point to become a buried site within a PPI. We also developed MaSIF-search, another tool to evaluate surface complementarity between binding partners. MaSIF-search was designed as a Siamese neural network architecture [173] trained to produce similar fingerprints for the target patch vs. the binder patch, and dissimilar fingerprints for the target patch vs. the random patch. As MaSIF tools had robust performance in PPI-related *prediction* tasks, we hypothesized that we could leverage them to *design* novel PPIs by targeting sites only using structural information from the target protein. To address the *de novo* PPI design problem we devised a three-stage computational approach depicted in Fig. 3.1: I) prediction of target buried interface sites with high binding propensity using MaSIF-site (Fig. 3.1A); II) surface fingerprint-based search for complementary structural motifs (*binding seeds*) that display the required features to engage the target site, a protocol we refer to as MaSIF-seed (Fig. 3.1A-B); III) binding seed transplantation to protein scaffolds to confer stability and additional contacts on the designed interface (Fig. 3.1C) using established transplantation techniques [64].

The new MaSIF-seed protocol tackles the problem of identifying binding seeds that can mediate productive binding interactions (Fig. 3.1, Supplementary Fig. S3.1). This task stands as a remarkable challenge in protein design due to the vast space of structural possibilities to explore, as well as the required precision given that subtle atomic-level changes, such as misplaced methyl groups [64,104], uncoordinated water molecules in the interface, or incompatible charges, are sufficient to disrupt PPIs [135].

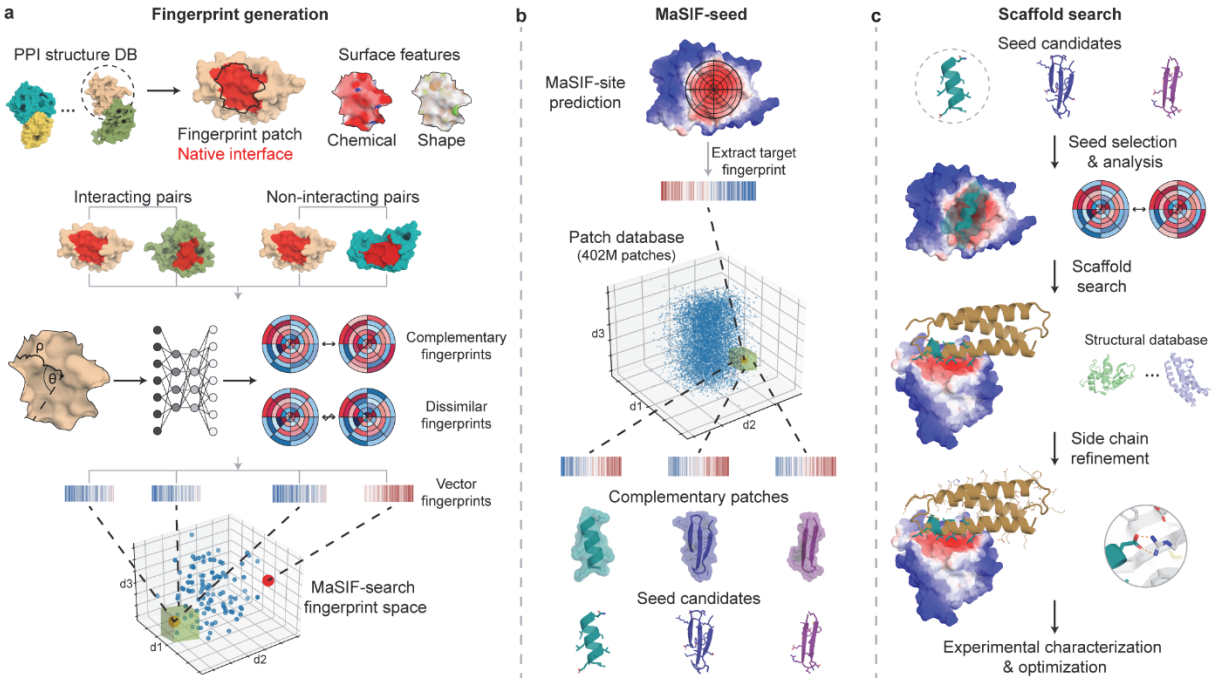



**Figure 3.1 : Surface-centric design of de novo site-specific protein binders. A.** Schematic of fingerprint generation. Protein binding sites are spatially embedded as vector fingerprints. Protein surfaces are decomposed into overlapping radial patches, and a neural network trained on native interacting protein pairs learns to embed the fingerprints such that complementary fingerprints are placed in a similar region of space. We show an illustration for a subsample of the fingerprints projected in a space reduced to three dimensions. The green box highlights a region of complementary fingerprints. **B.** MaSIF-seed—a method to identify new binding seeds. A target patch is identified by MaSIF-site based on the propensity to form buried interfaces. Using MaSIF-seed, fingerprint complementarity is evaluated between the target patch and all fingerprints in a large database (around 402 million patches); the pairs of fingerprints are subsequently ranked. The top patches are aligned and rescored to enable a more precise evaluation of the seed candidates. **C.** Scaffold search, seed grafting and interface redesign. The selected seeds are transferred to protein scaffolds and the rest of the interface is redesigned using Rosetta. The top designs are selected and tested experimentally.

In MaSIF-seed, protein molecular surfaces are decomposed into overlapping radial patches with a 12 Å radius, capturing on average nearly 400 Å$^2$ of surface area, consistent with the buried surface areas observed in native interfaces (Supplementary Fig. S3.2). For each point within the patch, we compute chemical and geometric features, as well as a local geodesic polar coordinate system to locate points within the patch relative to each other. A neural network is then trained to output vector fingerprint descriptors that are complementary between patches of interacting protein pairs and dissimilar between non-interacting pairs [138] (Fig. 3.1A, Supplementary Fig. S3.1). Matched surface patches are aligned to the target site and scored with a second neural network, outputting an interface post-alignment (IPA) score to further improve the discrimination performance of the surface descriptors (see methods).

To benchmark our method, we assembled a test set composed of 114 dimeric complexes, which contained 31 complexes where the binding motif was a single alpha-helical segment and 83 where the binding motif was composed of less than 50% helical segments (Supplementary Fig. S3.3). As decoy sets, we used 1000 motifs (ranging from 600K-700K patches) which in the case of the helical set also had helical secondary structure and in the non-helical set were composed of two- and three-strand beta sheets.

We benchmarked MaSIF-seed relative to other docking methods to identify the true binder from the co-crystal structure in the correct orientation (<3 Å iRMSD) among 1000 decoys (Supplementary Fig. S3.4). MaSIF-seed identified the correct binding motif in the correct orientation as the top scoring result in 18 out of 31 cases, and 41 out of 83 cases for the helical and non-helical sets, respectively. While the best performing method, ZDock+ZRank2 [174–176] identified only 6 out of 31 as top results in the helical set, 21 out of 83 in the non-helical set. In addition to superior performances MaSIF-seed was considerably faster, showing speed increases between 20-200 fold, which mostly depend on the number of patches derived from each motif. In our benchmark we also performed comparisons with faster methods which showed much lower performances than ZDock+ZRank2 (Table 3.1 and Supplementary Table S3.1).

An analysis of the cases where MaSIF-seed performed best showed that its success relied first on PPIs where the interaction site could be correctly identified by the method, and second to those where the majority of contacts lie on a radial patch at the interface core, and with a high shape complementarity in that region (Supplementary Fig. S3.5A). This is consistent with how MaSIF-seed was designed to capture protein interfaces using a radial geodesic patch.

**Table 3.1 : Benchmark of MaSIF-seed against other docking methods.** Recovering the native binder in the correct conformation from co-crystal structures for 31 helix-receptor complexes or 83 non-helix seed-receptor complexes, discriminating between 1000 decoys. [a]Benchmarked method. [b-d]Number of receptors for which the method recovered the native binding motif (<3 Å iRMSD) within the [b]top 1, [c]top 10, and [d]top 100 results. [e]Number of receptors for which the method did not recover the native binding motif in the top 100 results. [f]Average running time in minutes, excluding pre-computation time.

| | **Method[a]** | **# in top 1[b]** | **# in top 10[c]** | **# in top 100[d]** | **>100[e]** | **Avg time (m)[f]** |
|---|---|---|---|---|---|---|
| **Helical seeds** | MaSIF-seed | 18 | 18 | 20 | 11 | 15 |
| | ZDock | 3 | 4 | 8 | 23 | 2715 |
| | ZDock+ ZRank2 | 6 | 12 | 21 | 10 | 2946 |
| **Non-helical seeds** | MaSIF-seed | 41 | 47 | 49 | 34 | 118 |
| | ZDock | 7 | 9 | 22 | 61 | 2206 |
| | ZDock+ZRank2 | 21 | 33 | 45 | 38 | 2400 |

Encouraged by MaSIF-seed's speed and accuracy in discriminating the true binders from decoys based on rich surface features, we sought to design *de novo* protein binders to engage challenging and disease-relevant protein targets. We thus assembled a motif database including approximately 640 K structural fragments (402 M surface patches/fingerprints) with distinct secondary structures (approximately 390 K and 250 K of non-helical and helical motifs, respectively), extracted from the PDB (see methods). We computationally designed and experimentally validated binders against four structurally diverse targets: the receptor binding domain (RBD) of the SARS-CoV-2 spike protein where we identified a neutralization-sensitive site; the two partners of the PD-1/PD-L1 complex, an important protein interaction in immuno-oncology that displays a flat interface considered "hard-to-drug" by small molecules (Supplementary Fig. S3.6); CTLA-4, another important target for immuno-oncology. We show that our method can be applied to a variety of structural motifs as binding seeds (helical and non-helical), generating functional designs directly from the computational simulations.

### 3.3.2 Targeting a predicted SARS-COV2 site

We applied our surface-centric approach to design *de novo* binders to target the SARS-CoV-2 RBD. First, we used MaSIF-site to predict surface sites on the RBD with high propensity to be engaged by protein binders. We selected a site distinct from the ACE2 binding region, but overlapping such that a putative binder could inhibit the ACE2-RBD interaction (Fig. 3.2A). At the time, there were no known binders to this site. We searched a subset of our database containing 140 million surface fingerprints derived from helical fragments to find binding seeds that could target the selected site. The 7713 binding seeds MaSIF-seed provided showed two prominent features: I) a contact surface devoid of residues with strong binding hotspot features (e.g. large hydrophobic residues); II) an equivalent distribution of binding seeds in two distinct orientations of the helical fragment, with the seeds binding at 180° from each other (Fig. 3.2B), hinting that both binding modes are plausible. Remarkably, both orientations of the binding seeds present very similar signatures at the surface fingerprint level (Supplementary Fig. S3.7) and at the sequence level (Fig. 3.2B).

We synthesized one of the top ranked binding seeds as a linear peptide, but no binding interaction was detected by Surface Plasmon Resonance (SPR) (Supplementary Fig. S3.8). Therefore, using the Rosetta MotifGraft protocol we identified several protein scaffolds compatible with both binding modes of the seed (Fig. 3.2C), transplanted the seed hotspot side chains from a top-ranking seed onto the scaffolds, and used Rosetta (v3.13) to optimize the binder interface (Fig. 3.1C). Sixty-three designs based on twenty scaffolds, ranging from 7 to 23 mutations relative to the native proteins, were screened with yeast display (Supplementary Fig. S3.9). From this initial round of designs, DBR3_01 showed weak binding in yeast display experiments. Moreover, binding of DBR3_01 was competitive with soluble ACE2 (Supplementary Fig. S3.9), suggesting that the binder was targeting the correct RBD site. Furthermore, DBR3_01 showed slightly increased binding compared to the native scaffold protein and a double point mutant on the designed interface residues, further supporting that the seed residues were participating in the binding interaction (Supplementary Fig. S3.9, Supplementary Table S3.2).

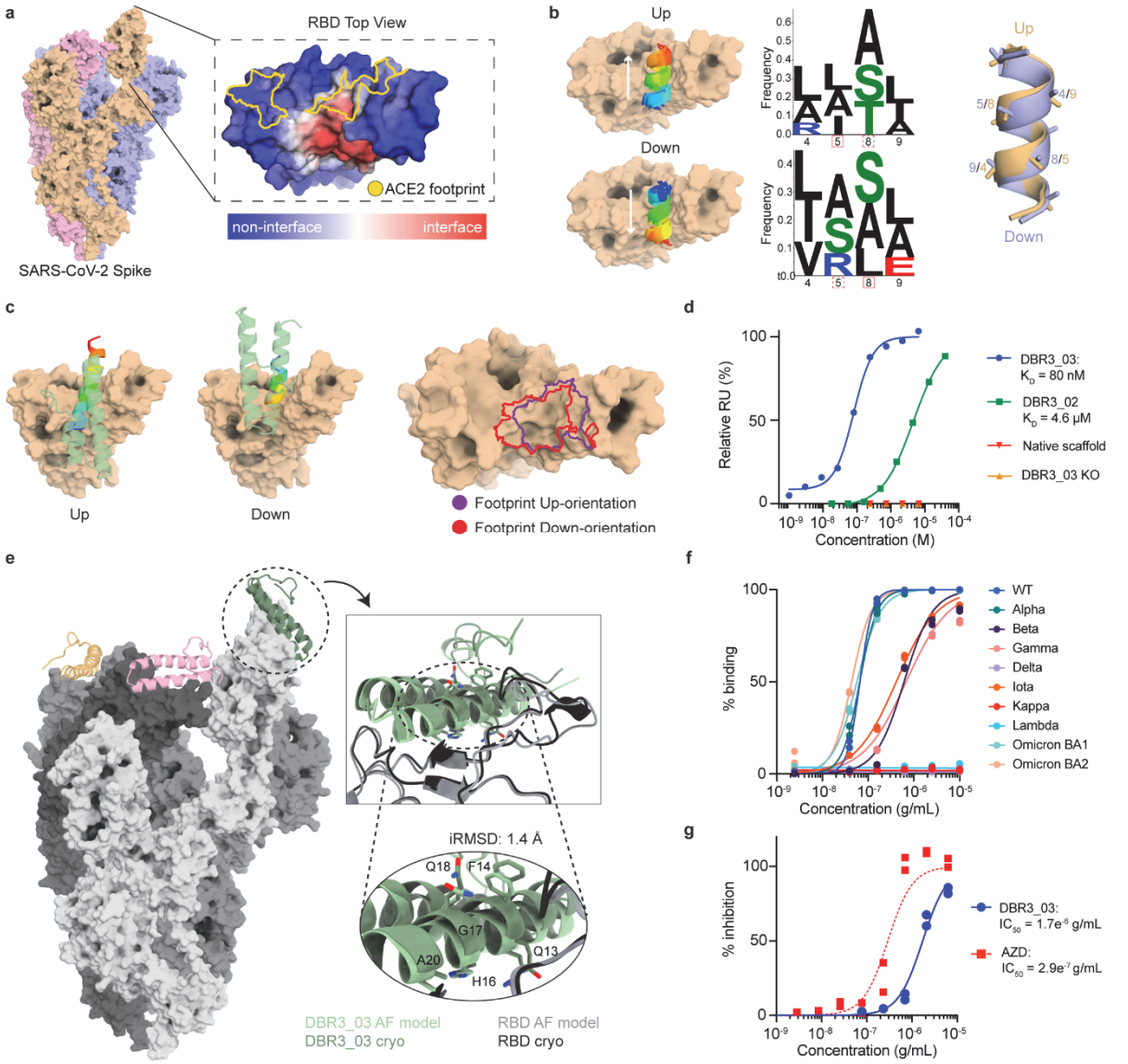



**Figure 3.2 : Design and optimization of a SARS-CoV-2 binder targeting the RBD. A.** MaSIF-site prediction of the interface propensity of the RBD. The ACE2-binding footprint (yellow outline) is distinct from the predicted binding site (red). **B.** MaSIF-seed predicts helical seeds that cluster into anti-parallel orientations, referred to as up or down configurations. Sequence logo plots highlight the similarity between the sequences of the two seed clusters, regardless of orientation. **C.** The scaffold (PDB: 5VNY) used to make DBR3_01 allows for binding in the up or down orientation, sharing similar footprints. **D.** SPR data of improved DBR3 binders with controls. DBR3_03 has an affinity of 80 nM with RBD. **E.** A cryo-EM structure (dark green) aligns to the AlphaFold prediction with an iRMSD of 1.4 Å. The trimeric spike protein (grey) has one DBR3_03 bound per RBD (orange, pink, green). **F.** Fc–DBR3_03 binds to the spike protein of most variants of concern, except for those with the L452R mutation. A list of half-maximal effective concentration (EC50) values of DBR3_03 is provided in Supplementary Table S3.3. The fits were calculated from technical replicates (n = 2) using a nonlinear four-parameter curve fitting analysis. **G.** Fc–DBR3_03 neutralizes live Omicron virus in cell-based inhibition assays with an half-maximal inhibitory concentration ($IC_{50}$) of $1.7 \times 10^{-6}$ g ml−1, compared with the AstraZeneca (AZD8895 and AZD1061) mix, which has an $IC_{50}$ of $2.9 \times 10^{-7}$ g ml−1. The fits were calculated from biological replicates (n = 2) using a nonlinear four-parameter curve fitting analysis.

Next, we sought to improve the binding affinity of the design by performing two mutagenesis libraries: first, a directed library in the designed interface was prepared (Supplementary Fig. S3.10), which yielded DBR3_02 with 4 mutations and a $K_D$ of 4.6 μM determined by SPR (Fig. 3.2D, Supplementary Fig. S3.10); second, we screened a site saturation mutagenesis (SSM) library which resulted in the enrichment of 3 point mutants, one of which overlapped with a mutation from the first library (Supplementary Fig. S3.11). Adding these 3 mutations to DBR3_02 resulted in DBR3_03 that showed a $K_D$ of 80 nM and was folded and stable (Fig. 3.2D, Supplementary Fig. S3.12). Here, we started from a computationally designed binder with very low affinity as observed with yeast display, yet undetectable by SPR, and after introducing 6 mutations we observed an improvement greater than 60 fold in binding affinity. The mutations all occurred in the binding helix of the design. Of these mutations, A17G and S20A, residing in the core of the interface, appear to have relieved steric clashes and reduced buried unsatisfied polar atoms, respectively.

To structurally characterize the binding mode of DBR3_03 we solved a cryo-EM structure of the design in complex with the trimeric spike protein at 2.9 Å local resolution (Fig. 3.2E and Supplementary Fig. S3.13-3.15). The structure confirmed the predicted binding sites on both partners. Importantly, the binder adopted the orientation of the helical binding seed that was marginally less favored by MaSIF's fingerprint descriptors (down-orientation) (Fig. 3.2B). Interestingly, the initial design DBR3_01 showed similar metrics when the interfaces were analyzed in both directions (Supplementary Fig. S3.7), pointing to known limitations of surface fingerprints in unbound docking type of problems[10]. This led us to attempt another state-of-the-art protein docking method, AlphaFold (AF) multimer [177] to predict the complex of DBR3_03 with the spike RBD and obtained a 1.4 Å iRMSD between the AF prediction and the experimental structure (Fig. 3.2E). This result presents a powerful demonstration of the synergies between machine learning techniques purely based on structural features and those that leveraged large sequence-structure datasets for structure prediction tasks. At the structural level DBR3_03 engages the RBD with a 1452 $Å^2$ of buried interface area (surface area buried on both sides of the complex), which is much smaller than the average buried surface area of antibodies (approximately 2071 ± 456 $Å^2$ [178]), yet still results in a high affinity interaction. The designed interface lacks canonical hotspot residues and engages the RBD through small residues and is composed of 21% backbone and 79% side chain contacts. Given the pandemic situation with SARS-CoV-2 and the general need for rational design of protein-based therapeutics to fight viral infections, we next engineered an Fc-fused DBR3_03 (Fc-DBR3_03) construct and tested its neutralization capacity on a panel of SARS-CoV-2 variants in virus-free and pseudovirus surrogate assays (Fig. 3.2F-G, Supplementary Fig. S3.16, Supplementary Table S3.3) [179]. We compared the breadth and potency of our design to those of clinically approved monoclonal antibodies. In virus-free assays we observed that Fc-DBR3_03 had comparable potency to that of Imdevimab (REGN10987), an antibody used clinically, for the WT spike and bound to the omicron strain while RGN87 did not (Supplementary Fig. S3.16). Neutralization activity in pseudovirus assays was tested and Fc-DBR3_03 neutralized omicron, albeit less potently than the AstraZeneca (AZN) clinically approved antibody mix (Fig. 3.2G). A cryo-EM structure showed that the binding mode was nearly identical (1.4 Å backbone RMSD) between DBR3_03-WT-RBD

complex and DBR3_03-omicron-RBD complex (Supplementary Fig. S3.17-3.19). Importantly, Fc-DBR3_03 showed a very broad reactivity to many SARS-CoV-2 variants (Fig. 3.2F) which is attributable to the sequence conservation of the targeted site and the small binding footprint of the design. The design was sensitive to the L452R/Q mutation present in the delta, lambda and kappa variants (Supplementary Fig. S3.16B, Fig. 3.2F), but introducing a single point mutation (L24G) to relieve the clash between L452R and the binder led to the design binding to delta (Supplementary Fig. S3.16). Our results highlight the value of the surface fingerprinting approach to reveal target sites in viral proteins and for the subsequent design of functional antivirals with broad activity.

### 3.3.3 Targeting a flat surface site in PD-L1

Surface sites presenting flat structural features are difficult to target with small molecule drugs, leading to their categorization as undruggable. To test our fingerprint-based approach, we sought to design binders to target the PD-1/PD-L1 interaction, which is central to the regulation of T-cell activity in the immune system [180]. We used MaSIF-site to find high propensity protein binding sites in PD-L1, and unsurprisingly, the identified site overlapped significantly with the native binding site engaged by PD-1 (Fig. 3.3A). This site is extremely flat at the structural level, ranking in the 99$^{th}$ percentile in terms of interface flatness (ranked #7 among 1068 transient interfaces, details in methods) (Supplementary Fig. S3.20), one of the dominant structural features that makes this site hard-to-drug by small molecules. Next, we used MaSIF-seed to find binding motifs to engage the site, among the top results helical motifs clustered in both orientations packing in the beta-sheets of PD-L1 (Supplementary Fig. S3.21). In the most populated cluster (Supplementary Fig. S3.21), we observed sequence convergence for a 12 residue fragment (Fig. 3.3B). We then used Rosetta MotifGraft to search for putative scaffolds to display this fragment and used RosettaDesign to optimize contacts at the interface. We tested 16 designs based on 5 different scaffolds for binding to PD-L1 on the surface of yeast. Two designs based on two different scaffolds showed low binding signals (Supplementary Fig. S3.22), which we refer to as DBL1_01 and DBL2_01 (Fig. 3.3C). The specificity of the interaction was confirmed by testing hotspot knockout controls of each design (Supplementary Fig. S3.22). To improve the binding affinity of DBL1_01 we constructed a combinatorial library with mutations in the predicted binding region, while maintaining the hotspot residues predicted by MaSIF-seed (Supplementary Fig. S3.23). From this library we selected a variant, DBL1_02 with 5 mutations found mostly in the interface rim of the design and improving the formation of polar contacts. The most substantial change occurred at position 53, a mutation of alanine to glutamine that introduces a hydrogen bond with PD-L1 (Supplementary Fig. S3.23). To improve the design's expression and stability we constructed a second library targeting residues in the protein core to optimize core packing (Supplementary Fig. S3.23). Combining mutations from both libraries, we obtained DBL1_03 with 11 mutations from the starting design, which was folded and monomeric in solution, and showed a binding affinity of 2 μM (Fig. 3.3D, Supplementary Fig. S3.12), comparable to that of PD-1 ($K_D$ = 8.2 μM) [181].

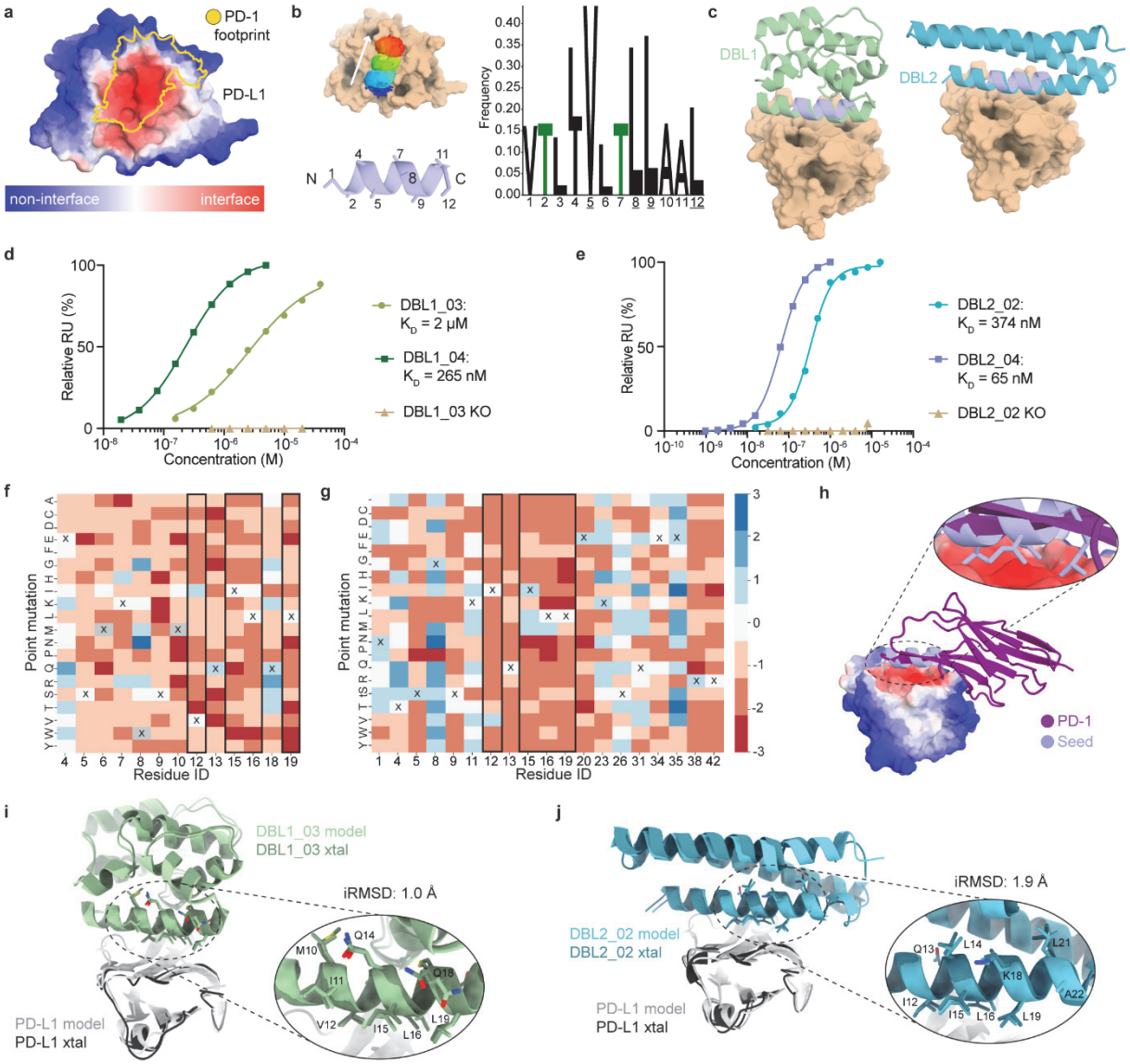



**Figure 3.3 : De novo design and optimization of PD-L1 binders targeting a flat surface. A.** MaSIF-site prediction of the interface propensity of PD-L1. The predicted interface (red) overlaps with the binding site of the native interaction partner PD-1 (yellow). **B.** Helical seeds were predicted by MaSIF-seed and clustered. The dominant cluster showed strong amino acid preferences (Z-score > 2). Hotspot residues are underlined. **C.** Binders based on two different scaffold proteins using the selected seed were identified. **D.** The binding affinities of DBL1 designs after combinatorial (light green) and SSM library optimization (dark green), measured using SPR. Mutation of a hotspot residue (V12R) ablates binding of DBL1_03 (wheat). **E.** The binding affinities of DBL2 designs after combinatorial (light blue) and SSM library optimization (dark blue), measured using SPR. Mutation of a hotspot residue (V12R) knocks out binding of DBL2_02 (wheat). **F.** SSM analysis of regions of interest in the binding interface of DBL1_03. The original residue of DBL1_03 is indicated by a cross and hotspot residue positions are shown in black boxes. Enrichment in the binding population (blue) and in the non-binding population (red) is indicated. **G.** SSM data in the binding interface of DBL2_03. The original residue of DBL2_02 is indicated by a cross. **H.** The binding mode of the selected seed in comparison to the native interaction partner PD-1. **I.** Crystal (xtal) structure of DBL1_03 in a complex with PD-L1. The computational model (light green) is aligned with the crystal structure (dark green). Inset: the alignment of the residues in the binding seed. **J.** Crystal structure of DBL2_02 in a complex with PD-L1, shown by aligning the computational model (light blue) with the crystal structure (dark blue). Inset: the alignment of the residues in the binding seed represented as sticks.

To further assess the optimality of each residue at the interface of the designed binder we screened a SSM library sampling 19 positions, based on DBL1_03. The most relevant positions are shown in Figure 3.3F (all positions in Supplementary Fig. S3.24). The SSM results revealed that the four hotspot residues placed by MaSIF-seed were crucial, as any other residue was deleterious for binding (Fig. 3.3F). However, in the interface rim many mutations could provide affinity improvements strongly suggesting that this region of the interface was suboptimal (Fig. 3.3F). Based on these data, we generated the DBL1_04 variant which resulted in a 10-fold increase of the binding affinity showing a $K_D$ of 256 nM to PD-L1 (Fig. 3.3D). Both DBL1_03 and DBL1_04 showed cell-surface binding, comparable to PD-1, on cells expressing PD-L1. The specificity of the designed interaction was confirmed by the binding inability of single-residue mutants at the interface (Supplementary Fig. S3.24).

The second lead design, which utilizes the same seed but is based on a different scaffold, DBL2_01, could not be solubly expressed and therefore we designed a combinatorial library to improve expression and binding affinity (Supplementary Fig. S3.23). From this library we isolated the variant DBL2_02 which had six mutations and expressed in *E. coli*. From the six mutations, three were predicted to be in the interface (Y23K, Q35E, Q42R) and improved binding affinity by forming additional salt bridges with PD-L1 (Supplementary Fig. S3.23). The $K_D$ to PD-L1 determined by SPR was 374 nM, more than 10-fold higher than the native ligand PD-1. Since both designs shared the same binding seed we transplanted the SSM mutations of the DBL1_04 design and generated the DBL2_03, which showed a 3-fold improvement in binding affinity ($K_D$ = 120 nM) (Supplementary Fig. S3.25), indicating that the binding seed was engaging PD-L1 in a similar fashion to that of DBL1_03. To further assess the influence of each residue in the designed binding interface we performed an SSM analysis on 19 interface residues of DBL2_03 (Fig. 3.3F, Supplementary Fig. S3.25). The SSM profile reiterated that the hotspot residues placed by MaSIF-seed were very restricted in variability, showing that these residues were accurately predicted. In contrast, several positions on the interface rim were suboptimal and mutations to polar amino acids resulted in affinity enhancements. Based on the SSM data, we generated the DBL2_04 design with additional polar mutations (Fig. 3.3G, Supplementary Fig. S3.25) which showed an improved $K_D$ of 65 nM (Fig. 3.3E). To experimentally validate the binding mode, we co-crystallized the designs with PD-L1 (Supplementary Fig. S3.26). Overall for both designs, the structures (Fig. 3.3I-J) showed excellent agreement with our computational models with 0.8 Å and 2.0 Å for the overall backbone and 1.0 Å and 1.9 Å for the full atom interface RMSDs of DBL1_03 and DBL2_02, respectively, showing an exquisite accuracy of the predictions in the interface region. The buried interface area of the designs with PD-L1 was between 1424 Å$^2$ and 1438 Å$^2$, compared to 1648 Å$^2$ for the buried interface area of PD-1 (PDB ID: 4ZQK). The chemical composition of the designed interface is similar in both designs, ~59% of the surface area is hydrophobic and the remaining area is hydrophilic for DBL1_03 and correspondingly for DBL2_02. These values are comparable to those of the PD-1/PD-L1 interaction (52% hydrophobic surface), showing that we have designed interfaces with similar chemical compositions of the native interaction using a distinct backbone conformation (Fig. 3.3H). The discovery of novel binding motifs by MaSIF-seed is striking when comparing the backbone motif used by the native PD-L1 binding partner, PD-1, and the designed binders. While the native PD-

1 uses a beta-hairpin to engage the site, the designed binders do so through an alpha-helix motif, illustrating the capability of our approach to explore outside of the structural repertoire of native binding motifs. The general trend arising from the designed PD-L1 binders is that despite the accurate predictions of core residues in the interface, through mutagenesis studies, the designed polar interactions are suboptimal. To address these and other limitations of our computational approach, we performed additional computational design steps to improve the pipeline and tested it on the design of binders to target PD-1.

### 3.3.4 One-shot designs with native affinities

Despite the successes in designing site-specific binders to engage two different targets, the computational designs still required *in vitro* evolution to enable expression and detectable binding affinities that could be biochemically characterized. To address these issues, we used a structurally diverse library of binding seeds (helical and beta-sheet motifs) and assembled a more comprehensive design pipeline (Fig. 3.4A) performing: I) sequence optimization of selected seeds; and II) biased design for polar contacts in the scaffold interface [137]. To test this approach, we designed *de novo* binders to target three proteins (PD-L1, PD-1 and CTLA-4). For each of the design targets we selected the top 2000 designed sequences according to several structural metrics (see methods) and tested them using yeast display coupled with deep sequencing readout. According to our deep sequencing readout we obtained binders for all three targets using diverse structural motifs to mediate the binding interaction (Supplementary Table S3.4). Several binders were biochemically characterized to varying degrees. For PD-1 we found three designs based on *de novo* miniprotein [74,75] scaffolds with interfaces mediated by helical motifs (DBP13_01, DBP40_01 and DBP52_01) (Fig. 3.4B, Supplementary Fig. S3.27) that showed a moderate to strong binding signal on the surface of yeast. The most promising candidate binding to PD-1, DBP13_01, was investigated in more detail (Fig. 3.4B-E). To confirm whether the binding interaction was mediated through the designed interface, we tested several control constructs, which included the native miniprotein scaffold and DBP13_01 variants with predicted knockout mutations (Fig. 3.4B), all of which abolished binding (Fig. 3.4C). The interaction site on PD-1 was further probed via a competition assay with Nivolumab [182], which blocked the DBP13_01/PD-1 interaction as expected due to the overlapping binding footprints (Supplementary Fig. S3.28). DBP13_01 did not bind to a close sequence homologue (porcine PD-1) supporting the specificity of the designed interactions (Supplementary Fig. S3.28). The DBP13_01/PD-1 interaction showed a $K_D$ of 4.2 ± 2 μM (n = 3, Fig. 3.4D) as determined by SPR, similar to the affinity of the native PD-L1/PD-1 interaction ($K_D$ = 8.2 μM) [181]. This was a promising result given that the design was not subjected to experimental optimization by *in vitro* evolution. Next, we performed an SSM experiment and observed that mutations at the predicted core interface positions (L23, L27, I30, M31) were generally deleterious for binding, supporting the structural and sequence accuracy of the design (Fig. 3.4E, Supplementary Fig. S3.29). Moreover, we readily improved the affinity to sub-micromolar by introducing two mutations identified in the SSM data (M31F+H33S, DBP13_02) (Fig. 3.4D). The predicted complex structure by AlphaFold Multimer (AF) was in agreement with that of MaSIF, with an interface footprint that is largely overlapping with the designed residues, and 3.3 Å of backbone RMSD and 2.9 Å of interface full atom

RMSD (Supplementary Fig. S3.30). Although these results are supported by the SSM data, they are a predictive exercise and cannot be interpreted as absolute evidence that the designed binding mode is occurring, which ultimately will require an experimental structure.

Similarly, we experimentally confirmed the specificity of a beta-sheet based-binder to PD-L1 (DBL3_01) (Fig. 3.4F) with a predicted knock-out mutant and a competition assay with high-affinity PD-1 (Fig. 3.4G and Supplementary Fig. S3.28). These data were supported by an AF prediction matching our design model with a 0.97 Å backbone RMSD (Supplementary Fig. S3.30). Binding to PD-L1 was further improved on yeast by mutating two exposed cysteines to serines in the scaffold, which may stabilize the protein and avoid unwanted disulfide bonds (DBL3_02, Fig. 3.4G and Supplementary Fig. S3.28). This design adopts a different backbone conformation than the native PD-1:PD-L1 interaction which further demonstrates MaSIF-seed's ability to generalize beyond interactions found in nature (Supplementary Fig. S3.28). We also estimated the affinity on a yeast display-based assay determining an apparent $K_D$ of 21.8 nM, 42.7-fold higher than the known high-affinity PD-1, which has been reported to have a true $K_D$ of 110 pM [183] (Fig. 3.4H).

We also performed experimental characterization for two other binders targeting PD-L1 (DBL4_01) and CTLA-4 (DBC2_01) and observed that the binding interactions are specific to targeted sites by competition and mutagenesis experiments performed using yeast display (Supplementary Fig. S3.28 and S3.31). It is important to note that for several of these binders the AF predictions were not in agreement with our models but that nevertheless the experimental results provide solid evidence that the correct interfaces are involved in the designed interactions (Supplementary Table S3.4).

Overall, the results show that by starting the interface design process driven by surface fingerprints and introducing additional features of native interfaces (e.g. hotspot optimization, polar contacts) we can design site-specific binders, using a variety of structural motifs with native-like affinities purely by computational design.

## 3.4 Discussion

Physical interactions between proteins in living cells are one of the hallmarks of function [184]. Our incomplete understanding of the complex interplay of molecular forces that drive PPIs has greatly hindered the comprehension of fundamental biological processes as well as the capability to engineer such interactions from first principles. It has been particularly challenging for protein modeling methodologies that use discrete atomic representations to perform *de novo* design of PPIs [131,132,162,169]. In large part, this is due to the small number of molecular interactions involved in most protein interfaces and to the very small energetic contributions that determine binding affinities, making physics-based energy functions less reliable [185].

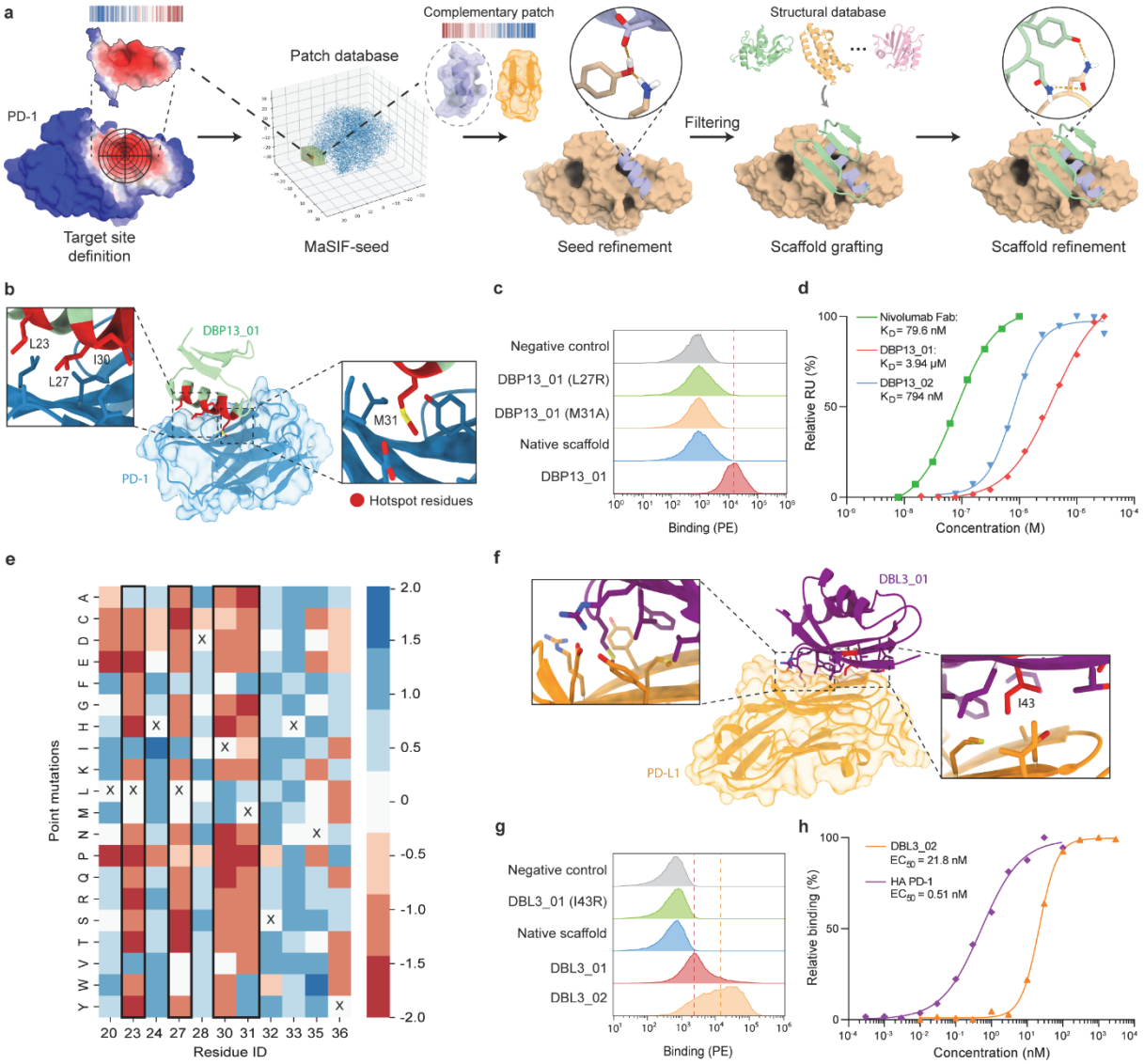



**Figure 3.4 : Optimized workflow and de novo binders for PD-1. A.** Improved design computational workflow in which two steps of design are used, at the seed and at the scaffold level, with an emphasis on building new hydrogen bond networks. **B.** PD-1 (blue) targeted by DBP13_01 (green); hotspot residues from the binding seed (red) are highlighted. Insets: crucial residues for binding. **D.** Histogram of the binding signal (PE, phycoerythrin) measured by flow cytometry for DBP13_01, the native miniprotein scaffold, two variants of DBP13_01 with crucial residues mutated and a negative control with unlabelled yeast. The dashed line indicates the geometric mean of the DBP13_01 binding signal. **D.** Binding affinities determined by SPR of the nivolumab Fab (green squares), DBP13_01 (red diamonds) and DBP13_02 (blue triangles). The dissociation constant of DBP13_01 was obtained with three independent measurements. **E.** SSM heat map showing interface residues and the enrichment of each point mutation. The original amino acids in DBP13_01 are indicated by a cross. Enrichment in the binding population (blue) and in the non-binding population (red) is indicated. Hotspot residues are highlighted with a black box. **F,** PD-L1 (orange) targeted by DBL3_01 (purple). Insets: magnification of interface residues, including one crucial residue tested for knockout mutants (Ile43, red). **G.** The binding signal measured using flow cytometry for DBL3_01, DBL3_02, the native protein scaffold, one knockout mutant and a negative control with unlabelled yeast. **H.** PD-L1 ligand titration on yeast displaying DBL3_02 (orange triangle) or high-affinity PD-1 (HA–PD-1, purple diamonds).

To address this gap, we developed an enhanced data-driven framework to represent proteins as surfaces and learn the geometric and chemical patterns that ultimately determine the propensity of two molecules to interact. We proposed a new geometric deep learning tool, MaSIF-seed, to overcome the PPI design challenge by both identifying patches with a high propensity to form buried surfaces and binding seeds with complementary surfaces to those patches. By computing fingerprints from protein molecular surfaces, we rapidly and reliably identify complementary surface fragments that can engage a specific target within 402 million candidate surfaces. This, in practice, solves an important challenge in protein design by efficiently handling search spaces of daunting scales.

The identified binding seeds were then used as the interface driving core to design novel binding proteins against challenging targets: a novel predicted interface in the SARS-CoV-2 spike protein, which ultimately yielded a SARS-CoV-2 inhibitor, PD-1/PD-L1 protein complex and CTLA-4, exemplifying sites that are difficult to target with small molecules due to its flat surface. Several designed binders showed close mimicry to computationally predicted models and achieved high binding affinities, often, after experimental optimization. In the case of purely computationally designed binders, the PD-1 binder showed low micromolar affinity without experimental optimization, which is the range of many native PPIs [95], and several other binders targeting PD-L1 and CTLA-4 were shown to be specific to the targeted sites. By using surface fingerprints, we identified novel structural motifs that can mediate *de novo* PPIs presenting a route to expand the landscape of motifs that can be used to functionalize proteins and be critical for the *de novo* design of function.

For all targets, the original binding seed arguably provided the principal driver of molecular recognition representing the design's binding interface core (Supplementary Fig. S3.32), maintaining a high surface similarity in this region between the original seed and the final design (Supplementary Fig. S3.33). However, contacts at the buried interface region are necessary though in most cases, likely not sufficient for high affinity binding, and in the three designed binders for PD-L1 and RBD, optimization of the polar interface rim through libraries was necessary to improve binding to a biochemically detectable range ($K_D$ at the micromolar level). Our *de novo* designs agree with previous findings [130,132] that small changes in the polar interface rim (for example in the hydrogen bond network surrounding the interface) can result in substantial differences in binding affinities. Encouragingly, by using a larger and more structurally diverse library of binding seeds together with an optimized design pipeline we obtained several *in silico* only designed binders to a variety of targets, which represents a major step forward for the robust design of *de novo* PPIs.

In our study several limitations of the approach became evident, namely the absence of conformational flexibility and adaptation of the protein backbone to mutations and the difficulty of designing polar interactions that balance the hydrophobic patches of the interface contributing for affinity and specificity, which has also been observed by other authors [129,130,135]. In future methodological developments, neural network architectures could be optimized to capture such features of native interfaces. The emergence of generative algorithms that can construct backbones conditioned to the target binding sites or the seed motifs, as recently described by other groups [87,186], present another

exciting route where our conceptual framework based on surfaces is likely to become more useful to overcome important challenges on the design of molecular recognition.

Here we presented a surface-centric design approach that leveraged molecular representations of protein structures based on learned geometrical and chemical features. We showed that these structural representations can be efficiently used for the design of *de novo* protein binders, one of the most challenging problems in computational protein design. We anticipate that this conceptual framework for generation of rich descriptors of molecular surfaces can open possibilities in other important biotechnological fields like drug design, biosensing or biomaterials in addition to providing a means to study interaction networks in biological processes at the systems levels.

## 3.5 Methods

*Computing buried surface areas*

A dataset of protein-protein interactions was downloaded from the PDBBind database [187] containing all interactions with a reported affinity stronger than 10 μM; since these PPIs have a reported affinity, all were assumed to be transient. The PDBBind database does not report the chains involved in the interaction with the reported affinity; thus, for simplicity, only those complexes containing exactly two chains in the PDB crystal structure were considered for the analysis.

The MSMS program [141] was used to compute all molecular surfaces in this work (density = 3.0, water radius = 1.5 Å). Since MSMS produces molecular surfaces with highly irregular meshes, PyMESH (v.0.2.1) [188] was used to further regularize the meshes at a resolution of 1.0 Å. For a given protein subunit that appears in a complex, we define the subunit's buried surface as the patch that becomes inaccessible to water molecules upon complex formation. Since in our implementation a surface is defined by a discretized mesh, we compute the buried surface region as follows. The buried surface of both the subunit and the complex are first independently computed. Then, the minimum distance between every subunit surface vertex and any complex surface vertex are computed. Subunit vertices that are farther than 2.0 Å from a vertex in the surface of the complex are labeled as part of the buried surface, as these vertices no longer exist in the surface of the complex. The size of buried areas was determined by computing the area of each vertex labeled as a buried surface vertex.

We note that computing buried surface areas using this method can result in measurements that are different from those widely used in the field, which use the solvent accessible surface area and count the buried interface of all subunits into a single value (the buried SASA area). Here we use the molecular surface (also known as solvent excluded surface) and count a single subunit. Therefore, while in Supplementary Fig. 3.1 we show areas computed using this method to compare to patch sizes, throughout the rest of this work we refer to the more widely used buried SASA areas.

*Patch generation in the MaSIF framework - Decomposing surfaces into radial patches*

In order to process protein surface information, all molecular surfaces were decomposed into overlapping radial patches. This means that each vertex on the surface becomes the center of a radial

patch of a given radius. To compute the geodesic radius of patches, throughout this work we used the Dijkstra algorithm [189], a fast and simple approximation to the true geodesic distance in the patch. We used a radius size of 12 Å for patches, limited to at most 200 points, which we found corresponds roughly to 400 Å$^2$ (Supplementary Fig. S3.2), a value close to the median size of the buried interface of transient interactions (Supplementary Fig. S3.2). Exceptionally, for the MaSIF-site application (described below) we limited the patch to 9 Å or 100 points to reduce the required GPU RAM for this application [138].

*Patch generation in the MaSIF framework - Computing angular and radial coordinates*

An essential geometric deep learning component in our pipeline is to compute angular and radial coordinates in the patch that enable MaSIF to map features in a 2D plane. The radial coordinate is computed using the Dijkstra algorithm, where the geodesic distance (meaning the distance taken to 'walk' along the surface) from the center of the patch to every vertex is computed. To compute the angular coordinate, all pairwise geodesic distances between vertices in the patch are computed, and then the multidimensional scaling algorithm [190] in scikit-learn [191] is used to map all vertices to the 2D plane. Then, a random direction in the 2D plane is computed as the 0$^o$ frame of reference, and the angle of every vertex in the plane with respect to this frame of reference is computed. Computing the angular and radial coordinates is the slowest step in the MaSIF precomputation. However, we have provided experimental code to compute these coordinates much faster in our github repository under a branch called "fast-masif-seed".

*Patch generation in the MaSIF framework - Geometric and chemical features*

Each point in a patch of the computed molecular surface was assigned an array of two geometric features (shape index [192], distance-dependent curvature [193]), and three chemical features (hydrophobicity [194], Poisson-Boltzmann electrostatics [195], and a hydrogen bond potential [196]). These features are identical to those described in Gainza *et al* [138].

*Patch generation in the MaSIF framework - Largest circumscribed patch computation*

From each labeled interface point, we used the Dijkstra algorithm to compute the shortest distance to a non-interface point. The interface point with the largest distance to a non-interface point was labeled as the center of the interface, and the distance to the nearest non-interface point as the radius of the largest circumscribed patch.

*Calculation of surface planarity*

The surface planarity of all target interfaces, with respect to a database of PPIs (Supplementary Fig. S3.20) was calculated as follows. 690 PPIs crystallized as dimers from the PDBBind database were used as the dataset, resulting in 1380 interfaces as each chain was analyzed separately. Interfaces with an approximate area lower than 150 Å$^2$ or more than 1000 Å$^2$ were discarded, resulting in 1068 interfaces. The vertices in the buried interface area of each chain were computed, as explained in section Computing buried surface areas in protein complexes above**,** and the 3D coordinates of those vertices

in the interface were extracted from each chain. Then, the multidimensional scaling method[190] from scikit-learn [191] was used to position interface vertices in a 2D plane, with the optimization goal of maintaining the distances between all pairs of vertices as close as possible in the 2D embedding as they were in 3D space. The root mean square difference between the distances in the original 3D space vs. the 2D space was used as the measure of planarity. Interfaces that are very planar in 3D have small values under this metric as an embedding in 2D preserves the distance between vertices, while non-planar interfaces have larger values as an embedding in 2D must significantly alter their 3D distances.

*Geometric deep learning layer in MaSIF*

Geometric deep learning enables the application of traditional techniques from deep learning to data that does not lie in Euclidean spaces, such as a protein molecular surface. At the core of MaSIF lies a mapping from a molecular surface patch to a 2D Euclidean Tensor. The mapping is performed through a learned soft polar grid around each patch center vertex, using the angular and radial coordinates. Once the mapping is performed, a traditional convolutional neural network layer is performed, with an angular max pooling layer, which deals with the rotation ambiguity of geodesic patches. Further details on these techniques are detailed in Gainza *et al* [138] and Monti *et al* [197].

*Prediction of protein interaction sites*

The MaSIF-site tool [138] was trained to predict areas with propensity to form protein-protein interactions on the surface of proteins. Here, MaSIF-site was used to predict surface areas with propensity to form a PPI in 114 targets of our benchmark (Supplementary Fig. S3.4) and all the design targets (SARS-CoV-2 RBD, PD-L1, PD-1 and CTLA4). MaSIF-site receives as input a protein decomposed into patches and outputs a per-vertex regression score on the propensity of each point to become a buried surface area within a PPI. MaSIF-site computes a regression score on each point of the surface, yet it becomes necessary to identify the precise patch that we will use to define each interface. Thus, to select interface patches in target proteins, the output of MaSIF-site was decomposed into 12 Å overlapping patches, and the per-vertex prediction for all points in the patch are averaged to obtain a score for each patch.

*Training of MaSIF-site*

MaSIF-site was trained on a database of protein-protein interactions sourced from PRISM[198], PDBBind [187], the ZDock benchmark [199], and SabDab [200]. Proteins from these databases that failed to run through the MaSIF pipeline due to, for example, too many incomplete residues in the deposited structure, were discarded. Each instance of these databases, which we refer to as 'subunits' could consist of one or multiple chains (e.g., an antibody), and was crystallized in complex with a partner subunit. In total 12002 subunits from deposited structures passed the threshold. These subunits were then clustered by sequence identity at 30% identity and up to one representative from each cluster was selected, resulting in 3362 subunits. Then, a matrix of all pairwise TM-scores for this set was computed, and a hierarchical clustering algorithm was used on this matrix to split the dataset into 3004 subunits for the training set and 358 for the testing set.

The molecular surface for each subunit was computed using MSMS [141] and the buried interface area was labeled as described above. The architecture of MaSIF-site (Supplementary Fig. S3.1B and further described in Gainza *et al* [138]) consisted of three layers of geodesic convolution. The network received as input the full surface of a protein (with batch size of 1) decomposed into overlapping patches of size 9 Å. During training, each vertex of the input was labeled with the ground truth, with a value of 1 if the vertex belonged to the buried area and a label of zero otherwise. The output of the network is a per-vertex assignment between 0 and 1 for the prediction of that vertex on whether it belongs to the buried surface area or not. A sigmoid activation function was used as the output layer, and a binary cross function as the loss function. Adam [201] was used as the optimization function. MaSIF-site was implemented in Tensorflow (v1.12) [202], and trained for 40 hours on a single GPU machine, which allowed for 43 epochs. The MaSIF-site neural network implementation in Tensorflow contains a total of 9267 parameters.

*Complementary surface identification*

MaSIF-search [138] was used to compute fingerprints for every overlapping patch in proteins of interest. MaSIF-search was trained on a dataset of 6001 protein-protein interactions (described in Gainza *et al* [138]) to receive as input the features of the target, a binder, and a random patch from a different protein. MaSIF-search was designed as a Siamese neural network architecture [173] trained to produce similar fingerprints for the target patch vs. the binder patch, and dissimilar fingerprints for the target patch vs. the random patch. In order to decrease training time and improve performance, the features of the target were multiplied by -1 (with the exception of hydropathy), turning the problem from one of complementarity to one of similarity.

*Training of MaSIF-search*

MaSIF-search was trained on a database of 6001 protein-protein interactions in co-crystal structures sourced from PRISM [198], PDBBind [187], the ZDock benchmark [199], and SabDab [200]. A split between the training and testing set was performed by extracting the atoms at the interface for all 6001 PPIs and computing a TM-score between all pairs using TM-align. A hierarchical clustering algorithm was used to cluster the pairwise matrix, which was used to split the data into a training set of 4944 PPIs and a testing set of 957 PPIs. As in MaSIF-site, each side of the interaction could consist of one or multiple chains (e.g. an antibody), and we refer to each side as a subunit. In each PPI, pairs of surface vertices within 1.0 Å of each other were selected as interacting pairs.

MaSIF-search produces fingerprints for patches with a radius of up to 12.0 Å in geodesic distance from a central vertex, and is trained to make these patches similar for interacting patches and dissimilar for non-interacting patches (Fig. 3.1A, Supplementary Fig. S3.1). We find that MaSIF-search performs best when trained on interacting pairs that lie in the center of highly complementary interfaces and these pairs were filtered to remove points outside of the interfaces or in interfaces with poor complementarity (further described in Gainza *et al* [138]).

The MaSIF-search network receives as input the features of a patch from one of these pairs (the 'binder'), the inverted input features of its interacting patch (the 'target'), and a patch randomly chosen from a different interface in the training set (the 'random' patch) (Supplementary Fig. S3.1). The neural network was trained on a Siamese neural network architecture to produce fingerprints that are similar for the 'binder' and 'target' patches while at the same time being dissimilar between 'target' and 'random'. Similarity and dissimilarity was measured as the Euclidean distance between the fingerprints. A total of 85,652 true interacting pair patches and 85,652 noninteracting pair patches were used for training/validation, and 12,678 true interacting and 12,678 noninteracting pairs were used for the testing set.

Each of the 5 input features was computed in a separate channel consisting of a MaSIF geometric deep learning convolutional layer. Then the output from all channels was concatenated, and a Fully Connected Layer was used to output a fingerprint of size 80. In each batch, 32 pairs of interacting patches and 32 pairs of non-interacting patches were used. Adam was used as the optimizer, and a learning rate of $10^{-3}$ was used. The d-prime cost function [203] was used as the loss function. MaSIF-search was trained for 40 hours in a GPU, after which it was automatically killed, resulting in 260,000 iterations of the data. The MaSIF-search neural network implementation contained a total of 66080 trainable parameters and was implemented in Tensorflow.

*Patch alignment and IPA scoring*

In the MaSIF-search pipeline, surfaces are computed for each protein of interest, and both a MaSIF-search fingerprint and a MaSIF-site prediction are computed for each surface vertex. All fingerprints within a user defined threshold for similarity to a target patch (defined at 1.7 by default) are then selected for a second-stage alignment and rescoring. In this step, the patch is extracted from the source protein, along with all the fingerprints for all vertices in the patch (since they were all precomputed). The random sample consensus (RANSAC) algorithm implemented in Open3D [204] then uses the fingerprints of all the vertices in the target and matched patch to find an alignment between the patches. The RANSAC algorithm chooses three random points in the binder patch and computes the Euclidean distance of the surface MaSIF-search fingerprints between these points and all those points in the target patch; the most similar fingerprints provide the RANSAC algorithm with 3 correspondences to compute a transformation between the patches.

Once a candidate patch is aligned, the interface post-alignment (IPA) neural network (NN) is used to score the alignment with a score between 0 and 1 on the prediction of whether the alignment corresponds to a real interaction or not. Upon patch alignment, each vertex in the candidate patch is matched to the closest vertex in the target patch, and three features are computed per pair of vertices: (i) 1/(distance), the euclidean distance in 3D between the vertices; (ii) the product of the normal between the vertices, and (iii) 1/(fingerprint distance), the euclidean distance between the MaSIF-search fingerprints between the two vertices. A fourth feature, which we call 'penetration' is computed by computing the distance between each of the vertices in the candidate patch and all the atoms in the target. Thus, the IPA NN receives as input a vector of size Nx4, where N is the number of vertices in the

candidate patch (up to 200 vertices). The IPA NN consists of 5 layers of 1D convolution, followed by a Global Averaging Pool layer and 7 fully connected layers. The 5 layers of 1D convolution contain 16, 32, 64, 128, and 256 filters, respectively, with a kernel size of 1 and a stride of 1, and each layer was followed by a batch normalization layer and a Rectified Linear Unit layer. The fully connected layers contained 128, 64, 32, 16, 8, 4, and 2 dimensions. Each fully connected layer was also followed by a Rectified Linear Unit layer, with the exception of the last layer which was followed by a softmax layer. The network was optimized with Adam [201], with a learning rate of $10^{-4}$ and a categorical cross entropy loss function.

The IPA NN was trained as follows. The same dataset used for MaSIF-search, containing 4944 PPIs and a testing set of 957 PPIs was used. For each protein pair, one protein was chosen as the target, and the patch at the center of the interface was selected as the target patch. Then the partner protein along with 10 randomly chosen other proteins were aligned to it. Any alignment of the true partner within 3 Å RMSD of the co-crystal structure was considered as a positive. Any alignment from the true partner at greater than that RMSD or of any other protein was considered as a negative. Features were computed for all alignments and used for the IPA NN training. The IPA NN was trained with batches of 32 for 50 epochs.

*Binding seed database - Alpha-helix seed library generation*

A snapshot of the non-redundant set of the PDB was downloaded and decomposed into alpha helices, removing all non-helical elements. The DSSP program [205] was used to label each residue according to their secondary structure. Fragments with 10 or more consecutive residues with a helical ('H') label assigned by DSSP were extracted. Each extracted helical fragment was treated as a monomeric protein, and surface features were computed for each one. MaSIF-search fingerprints and MaSIF-site labels were then computed for all extracted helices. MaSIF-seed uses both fingerprint similarity and interface propensity to identify suitable seeds. Ultimately, our binding seed database was composed of approximately 250 K helical motifs from which 140 M fingerprints were extracted.

*Binding seed database - Beta-strand seed library generation*

To collect beta-strand motifs, a snapshot of the non-redundant set of the PDB was preprocessed with the MASTER software [206] to allow for fast structural matches. Two template motifs, one consisting of two beta strands and one consisting of three beta strands, were deprived of loops and served as input to MASTER to find sets of structurally similar motifs that would ultimately become the motif dataset for MaSIF. The search allowed for a variable backbone length of 1-10 amino acids connecting the beta strands of the template. RMSD cutoffs were set at 2.1 Å and 3 Å for two-stranded and three-stranded beta sheets, respectively. Similar to the preparation of helical motifs, each beta fragment was treated as a monomeric protein and surface features were generated, followed by the generation of MaSIF-search fingerprints and MaSIF-site labels. Ultimately, our beta-strand binding seed database was composed of approximately 390 K motifs from which 260 M fingerprints were extracted.

*Binding seed identification*

Based on the different modules within the MaSIF framework [138], we developed a novel pipeline to identify potential binding seeds to targets. For each target, first MaSIF-seed was used to label each point in the surface for the propensity to form a buried surface region. Then, a fingerprint was computed for the target site. Finally, after scanning the entire protein, the best patch was selected. In one case, the SARS-CoV-2 RBD, the fourth best site was selected as it was the site with the highest potential to disrupt binding to the natural receptor. Then, a MaSIF-search fingerprint was computed for the target patch, inverting the target features before inputting them to the MaSIF-search network. The Euclidean distances between the target fingerprint and the millions of fingerprints in the binding seed database were then computed, and all patches with a fingerprint distance below defined thresholds were accepted. In this paper the thresholds utilized were <2.0 for PD-L1, PD-1 and CTLA-4, and <1.7 for the RBD.

Once fingerprints are matched, a second-stage alignment and scoring method uses the RANSAC algorithm as described above. After RANSAC produces an alignment, the IPA neural network classifies true binders vs. non-binders [138] and outputs an IPA score (described above). Those candidate binders with an IPA score of more than 0.90-0.97 in the neural network score were accepted.

*Computational benchmark - Helix:receptor motifs*

A set of transient interactions from PDBBind was scanned to identify proteins that bind to helical motifs. A binding motif was determined to be a helix if 80% of residues are helical and the total number of residues does not exceed 60. The selected complexes were filtered to remove pairs of PPIs with high homology and a set of 31 unique PPIs was used, subsequently MaSIF-search fingerprints and MaSIF-site fingerprints were computed. MaSIF-seed was benchmarked against a hybrid pipeline of existing, fast, well-established docking tools on the dataset of helix:receptor proteins: PatchDock [207], ZDock [174,208], and ZRank2 [175]. For each helix:receptor pair, the helix from the co-crystal structure was placed along 1000 randomly selected helices from the motif database. Then the methods were benchmarked to evaluate their capacity to rank the correct helix from the co-crystal structure, with an alignment RMSD <3.0 Å from the conformation of the co-crystal structure, versus the remaining 1000 helices. We note that each helix can potentially bind in many possible orientations, and in the case of methods that were not preceded by a MaSIF-site identification of the target site, the helix can bind on many sites on the receptor. The measured time for all methods included only the scoring time, except for MaSIF-seed where the alignment time was also included in the calculation. MaSIF-seed**:** All of MaSIF-seed's neural networks (MaSIF-search, MaSIF-site and the IPA score) were retrained for this benchmark to remove helix:receptor pairs from the training set. In each case, MaSIF-site was used to identify the patch in the target protein with the highest interface propensity, and the fingerprint for the selected patch was compared to the fingerprints of all patches in the database. The rigid orientation of each helix in the benchmark was randomly rotated and translated prior to any alignment. Patches were discarded if their MaSIF-search fingerprint's euclidean distance to that of the target site was greater than 1.7. After alignment, patches were further filtered if the IPA score was less than 0.96.

PatchDock+MaSIF-site**:** On each receptor protein, MaSIF-site was used to identify and label the target site, while PatchDock[64] was used to dock all 1001 helices, setting the target site based on a specific residue using the ReceptorActiveSite flag in PatchDock. The PatchDock score was used to produce the ranking of all conformations for all 1001 helices. ZDock: Was run on standard parameters and its standard scoring was used similar to PatchDock. In the ZDock+MaSIF-site case, all residues outside of the MaSIF-site selected patch were blocked using the compute_blocked_res_list.sh provided in ZDock. ZDock+ZRank2: In this variant, the top 2000 results from ZDock with each of the 1001 peptides for each of the 31 receptors were re-scored using ZRank2. The ZRank2 score was then used to score all of the docking poses.

*Computational benchmark - Non-helix:receptor motifs*

The same set of transient interactions from PDBBind was filtered for proteins interacting through non-helical motifs. The secondary structure types of the proteins were annotated with DSSP [205], followed by computing the contribution of helical segments (DSSP annotation of H, G, or I) to the interface. Only interfaces with less than 50% helical segments were selected. Additional filtering was performed by requiring a mean shape complementary at the interface of >0.55 and a maximum inscribed patch area of >150 Å$^2$. From these native complexes, seeds were extracted by selecting residues within 4 Å distance to the receptor and extending the backbone of these residues on their N- and C-terminus until the DSSP annotation changed to capture complete secondary structure elements. In total 83 complexes were collected for the benchmark.

The decoy set was constructed from 1000 randomly selected beta-strand seeds from the MaSIF-seed pipeline, containing 500 two-stranded and 500 three-stranded beta motifs. The benchmark was performed similarly to the helix:receptor benchmark described above with adapting the fingerprint's euclidean distance cutoff to a value of 2.5 and allowing MaSIF-seed to evaluate the top two sites in each receptor. These modifications were performed for this benchmark as it increased the accuracy while still performing at least 20 times faster than comparable competing tools. Only ZDock and ZDock/ZRank2 were benchmarked in the non-helical benchmark as ZDock/ZRank2 was shown to be the best in the helical benchmark.

*Clustering of seed solutions*

In each design case all of the top matched seeds were clustered by first computing the root mean square deviation between all pairwise helices, computed on the C-alpha atoms of each pair of helices, in the segment overlapping over the buried surface area. The pairwise distances were then clustered using metric multidimensional scaling [209] implemented in scikit-learn [191].

*Seed and interface refinement*

For the "one-shot" protocol, seed candidates proposed by MaSIF were refined using Rosetta and a FastDesign protocol with a penalty for buried unsatisfied polar atoms in the scoring function [137]. Beta sheet-based seeds containing >33% contact residues found in loop regions were discarded. 33, 200, and

109 refined seeds were selected based on the computed binding energy, shape complementarity, number of hydrogen bonds and counts of buried unsatisfied polar atoms for PD-1, CTLA-4, and PD-L1, respectively.

*Seed grafting and computational design*

A representative seed was selected from each solution space, and then matched using Rosetta MotifGraft to a database of 1300 monomeric scaffolds in the case of the RBD and PD-L1 designs. For the optimized protocol selected seeds were grafted to a database of 4,347 small globular proteins (<100 amino acids), originating from the PDB [117], two computationally designed miniprotein databases [74,75] and one AF2 proteome prediction database [31,210]. Seeds were cropped to the minimum number of side chains making contact before grafting. Moreover, loop regions from beta sheet-based seeds were completely removed. After side-chain grafting by Rosetta (v3.13), a computational design protocol was used to design the remaining interface. Final designs were selected for experimental characterization based on the computed Rosetta binding energy, the shape complementarity, number of hydrogen bonds and counts of buried unsatisfied polar atoms.

*Yeast surface display of single designs*

DNA sequences of designs were purchased from Twist Bioscience containing homology overhangs for cloning. DNA was transformed with linearized pCTcon2 (Addgene #41843) or a modified pNTA vector with V5 tag into EBY-100 yeast using the Frozen-EZ Yeast Transformation II Kit (Zymo Research). Transformed yeast were passaged once in minimal glucose medium (SDCAA) before induction of surface display in minimal galactose medium (SGCAA) overnight at 30°C. Transformed cells were washed in cold PBS with 0.05-0.1% BSA and incubated with the binding target for 2 hours at 4°C. Cells were washed once and incubated for an additional 30 minutes with appropriate antibodies (Supplementary Table S3.5). Cells were washed and analyzed using a Gallios flow cytometer (Beckman Coulter). For quantitative binding measurements, binding was quantified by measuring the fluorescence of a PE-conjugated anti-human Fc antibody (Invitrogen) detecting the Fc-fused protein target. Yeast cells were gated for the displaying population only (V5, Myc or HA positive) (Supplementary Fig. 3.9A).

*Yeast libraries*

Combinatorial sequence libraries were constructed by assembling multiple overlapping primers (Supplementary Table S3.6) containing degenerate codons at selected positions for combinatorial sampling of the binding interface, core residues or hydrophobic surface residues. Primers were mixed (10 μM each) and assembled in a PCR reaction (55 °C annealing for 30 sec, 72 °C extension time for 1 min, 25 cycles). To amplify full-length assembled products, a second PCR reaction was performed, with forward and reverse primers specific for the full-length product. For SSM libraries and oligo pools, DNA was ordered from Twist Biosciences and amplified with primers to give homology to the pCTcon2/pNTA backbone. In all cases, the PCR product was desalted and used for transformation.

*Yeast surface display of libraries*

Combinatorial libraries, SSM libraries, and oligo pools were transformed as linear DNA fragments in a 5:1 ratio with linearized pCTcon2 or pNTA_V5 vector as described previously into EBY-100 yeast [111]. Transformation efficiency generally yielded around $10^7$ transformants per cuvette. Transformed yeast were passaged at least once in minimal glucose medium (SDCAA) before induction of surface display in minimal galactose medium (SGCAA) overnight at 30°C. Induced cells were labeled in the same manner as the single designs. Labeled cells were washed and sorted on a Sony SH800 cell sorter (acquired with LE-SH800SZFCPL Cell Sorter software, v2.1.5). For combinatorial libraries and oligo pool libraries, sorted cells were grown in SDCAA and prepared similarly for two additional rounds of sorting. After the third sort cells were plated on SDCAA agar and single colonies were sequenced. SSM libraries were sorted once, collecting both binding and nonbinding populations, and grown in liquid culture for plasmid preparation. For flowcytometry analysis of single clones, data were collected with a Galios (Beckman Coulter) cytometer using Kaluza software (Beckman Coulter, v1.1.20388.18228). Flowcytometry data were analyzed with FlowJo software (BD Bioscience, v10.8.1).

*MiSeq Sequencing*

After sorting, yeast cells were grown in SDCAA medium, pelleted and plasmid DNA was extracted using Zymoprep Yeast Plasmid Miniprep II (Zymo Research) following the manufacturer's instructions. The coding sequence of the designed variants was amplified using vector-specific primer pairs, Illumina sequencing adapters and Nextera barcodes were attached using an additional overhang PCR, and PCR products were desalted with Qiaquick PCR purification kit (Qiagen) or AMPure XP selection beads (Beckman Coulter). Next generation sequencing was performed using Illumina MiSeq with appropriate read length, yielding between 0.5-1 million reads/sample. For bioinformatic analysis, sequences were translated in the correct reading frame, and enrichment values were computed for each sequence.

*Protein expression and purification*

DNA sequences were ordered from Twist Bioscience and Gibson cloning or T7 ligation used to clone into bacterial (pET21b) or mammalian (pHLSec) expression vectors. Protein binder and target constructs are listed in Supplementary Table S3.2 and S3.7 respectively. Mammalian expressions were performed using the Expi293™ expression system from Thermo Fisher Scientific (A14635). Cells were authenticated and tested negative for mycoplasma contamination (qPCR) by the provider and no additional authentication and tests have been done. Supernatant was collected 6 days post transfection, filtered, and purified. *E. coli* expressions were performed using BL21 (DE3) cells and IPTG induction (1 mM at OD 0.6-0.8) and growth overnight at 16-18° C. Pellets were lysed in lysis buffer (50 mM Tris, pH 7.5, 500 mM NaCl, 5% Glycerol, 1 mg/ml lysozyme, 1 mM PMSF, and 1 µg/ml DNase) with sonication, the lysate clarified, and purified. All proteins were purified using an ÄKTA pure system (GE healthcare) with either Ni-NTA affinity or protein A affinity columns followed by size exclusion chromatography. If

TEV cleavage was necessary, fused proteins were dialyzed overnight at 4°C (dialysis buffer 20 mM Tris pH 7.5, 150 mM NaCl, 10% glycerol) with excess TEV enzymes.

*Surface plasmon resonance*

SPR measurements were performed on a Biacore 8K (Cytiva, software v4.0.8.19879) with HBS-EP+ as running buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.005% v/v Surfactant P20, GE Healthcare). Ligands were immobilized on a CM5 chip (GE Healthcare # 29104988) via amine coupling. 500-1000 response units (RU) were immobilized and designed proteins were injected as an analyte in serial dilutions. The flow rate was 30 µl/min for a contact time of 120 s followed by 800 s dissociation time. After each injection, the surface was regenerated using 3 M magnesium chloride (for PD-L1) or 10 mM glycine, pH 3.0 (for RBD). Data were fit with 1:1 Langmuir binding model within the Biacore 8K analysis software (Cytiva, v4.0.8.19879).

*Biolayer Interferometry*

BLI measurements were performed on a Gator BLI system using the GatorOne software (Gator Bio, v2.7.3.0728). The running buffer was 150 mM NaCl, 10 mM HEPES pH 7.5. Fc-tagged designs were diluted to 5 ug/mL and immobilized on the tips (1-2 nm immobilized). The loaded tips were then dipped into serial dilutions of either spike protein or RBD. Curves were fit using a 1:1 model on the Gator software after subtracting the background.

*SEC-MALS*

Size exclusion chromatography (controlled by Chromeleon software; ThermoFischer Sci, v7.2.10) with an online multi-angle light scattering device (miniDAWN TREOS, Wyatt) (SEC-MALS) was used to determine the oligomeric state and molecular weight for the protein in solution. Purified proteins were concentrated to 1 mg/ml in PBS (pH 7.4), and 100 µl of the sample was injected into a Superdex 75 300/10 GL column (GE Healthcare) with a flow rate of 0.5 ml/min, and $UV_{280}$ and light scattering signals were recorded. Molecular weight was determined using the ASTRA software (Wyatt, v8.0.2.5).

*Circular Dichroism*

Far-UV circular dichroism spectra were measured using a Chirascan™ spectrometer (AppliedPhotophysics) in a 1-mm path-length cuvette. The protein samples were prepared in a 10 mM sodium phosphate buffer at a protein concentration between 20 and 50 µM. Wavelengths between 200 nm and 250 nm were recorded with a scanning speed of 20 nm $min^{-1}$ and a response time of 0.125 secs. All spectra were averaged two times and corrected for buffer absorption. Temperature ramping melts were performed from 20 to 90°C with an increment of 2°C/min. Thermal denaturation curves were plotted by the change of ellipticity at the global curve minimum to calculate the melting temperature ($T_m$).

*Cell binding analysis*

Karpas-299 cells were purchased from Sigma (06072604-1VL) with the approval of the European Collection of Authenticated Cell Cultures (ECACC). Cells were authenticated (PCR) and tested negative for mycoplasma contamination (PCR & Vero indicator) by the provider. For flow cytometry analysis of DBL1 designs binding to PD-L1 on Karpas-299 cells, $2x10^5$ cells were incubated with 50 µL Fc Block (BD Biosciences, cat #553142) that was pre-diluted 1:50 in FACS buffer (PBS (Gibco/Thermofisher scientific, cat #10010-015) and 2% BSA (Sigma Aldrich, cat #A7906)) for 15 minutes on ice. Samples were subsequently supplemented with 50 µL of PD-L1 binders prepared as follows: high-affinity PD-1_Fc serially diluted 1:2 for 20 dilutions in FACS buffer, starting at 62.5 µg/ml; DBL1_03_Fc and DBL1_04_Fc serially diluted 1:2 for 16 dilutions in FACS buffer, starting at 125 µg/ml; DBL1_03_KO_Fc and PD-1_Fc serially diluted 1:2 for 14 dilutions in FACS buffer, starting at 125 µg/ml. The cell solutions were incubated for 30 minutes. Samples were then washed three times, resuspended in 100 µL of FACS buffer containing secondary R-PE Goat Anti-Human IgG antibody diluted 1:100 (Jackson ImmunoResearch, cat #109-117-008), and incubated for 30 minutes. Samples were then washed three times to remove unbound antibody, resuspended in 100 µL of FACS buffer, and analyzed using LSR Fortessa flow cytometer (BD Biosciences).

*Protein purification for crystallography*

PD-L1 extracellular domain fragment (Uniprot #Q9NZQ7; from F19 to R238) was over-expressed as inclusion bodies in the BL21 (DE3) strain of *E. coli.* Renaturation and purification of PD-L1 was performed as previously described[211]. Briefly, inclusion bodies of PD-L1 was diluted against a refolding buffer (100 mM Tris, pH 8.0; 400 mM L-Arginine; 5 mM EDTA-Na; 5 mM Glutathione (GSH); 0.5 mM Glutathione disulfide (GSSG)) at 4°C for 24 h. Then the PD-L1 was concentrated and exchanged into a buffer of 20 mM Tris-HCl (pH 8.0) and 15 mM NaCl and further analyzed by HiLoad 16/60 Superdex 75 pg (Cytiva) chromatography. PD-L1 binder designs, DBL1_03 and DBL2_02, were over expressed in *E. coli* as inclusion bodies. Renaturation and purification of the PD-L1 binder designs was performed as the PD-L1 protein. PD-L1 and binder designs were then mixed together at a molar ratio of 1:2 and incubated for 1h on ice. The binder/PD-L1 complex was further purified by HiLoad 16/60 Superdex 75 pg (Cytiva) chromatography.

*Data collection and structure determination*

For crystal screening, 1 µl of binder/PD-L1 complex protein solution (10 mg/mL) was mixed with 1 µl of crystal growing reservoir solution. The resulting mixture was sealed and equilibrated against 100 µl of reservoir solution at 4° or 18°C. Crystals of the DBL1_03/PD-L1 complex were grown in 0.2 M potassium formate and 20% w/v PEG 3350. Crystals of the DBL2_02/PD-L1 complex were grown in 0.2 M potassium/sodium tartrate, 0.1 M Bis Tris propane, pH 6.5 and 20 % w/v PEG 3350. Crystals were flash-cooled in liquid nitrogen after incubating in anti-freezing buffer (reservoir solution containing 20% (v/v) glycerol). Diffraction data of crystals were collected at Shanghai Synchrotron Radiation Facility (SSRF) BL19U. The collected intensities were subsequently processed and scaled using the XDS package [212] (vJan 10 2022, BUILT=20220220). The structures were determined using molecular

replacement with the program Phaser MR in PHENIX (v1.20.1-4487), with the reported PD-L1 structure (PDB: 3RRQ) as the search model [213]. COOT (v0.9.5) and PHENIX (v1.20.1-4487) were used for subsequent model building and refinement [214,215]. The stereochemical qualities of the final model were assessed with MolProbity [216] (v4.5.1). Data collection details and refinement statistics are in Supplementary Table S3.8.

*Luminex binding assays*

Luminex beads were prepared as previously published [179]. Briefly, MagPlex beads were covalently coupled to SARS-CoV-2 spike proteins of different variants. The serial dilutions of the antibodies or design were performed and binding curves were fit using Prism (GraphPad, v9) nonlinear four parameter curve fitting analysis of the log(agonist) versus response.

*Live virus neutralization assays*

The virus neutralization assays were performed as previously published [179]. Briefly, VeroE6 cells were seeded in 96 well plates the day before the infection. The DBR3_03-Fc compound in serial dilutions was mixed with omicron-spike virus and incubated at 37°C for one hour before addition to the cells. The cells with virus were kept a further 48 hours at 37°C, then washed and fixed for crystal violet staining and analysis. Neutralization $EC_{50}$ calculations were performed using Prism (GraphPad, v9) nonlinear four parameter curve fitting analysis.

*Cryo-EM preparation and data acquisition*

For cryo-electron microscopy investigations, 3.0 µl aliquots at a concentration of 0.87 or 1.0 mg/ml of the $spike_{D614G}$-binder sample or the $spike_{Omicron}$-binder sample were applied onto glow-discharged carbon-coated copper grids (Quantifoil R2/1, 400 mesh), blotted for 4.0-8.0 s, and flash-frozen in a liquid ethane/propane mixture cooled to liquid nitrogen temperature, using Vitrobot Mark IV (Thermo Fisher Scientific) with 100% humidity and the sample chamber operated at 4 °C. Grids were screened in a Thermo Fisher Scientific (TFS) 200kV Glacios cryo-EM instrument. Suitable grids were transferred to TFS Titan Krios instruments for data collection. Cryo-EM data-collection statistics of this study are summarized in Supplementary Table S3.9. The $spike_{D614G}$-binder data composed of 20,794 movies was collected on a Titan Krios G4 microscope, equipped with a cold-FEG electron source and operated at 300 kV acceleration voltage. Movies were recorded with the automation program EPU (ThermoFisher Sci., v2.12.1) on a Falcon4 direct electron detector in counting mode at a physical pixel size of 0.40 Å per pixel and a defocus ranging from -0.8 to -2.0 µm. Exposures were collected as electron event recordings (EER) with a total dose of 80 $e^-/Å^2$ over approximately 3 seconds, corresponding to a dose rate of 4.53 $e^-$/px/s. For $spike_{Omicron}$-binder data, 22,266 movies were recorded on a Titan Krios G4 microscope, equipped with TFS SelectrisX imaging filter and Falcon4 camera. Exposures were collected at 60 $e^-/Å^2$ total dose with a physical pixel size of 0.726 Å per pixel over approximately 6 seconds, corresponding to a dose rate of 5.4 $e^-$/px/s, at a defocus range of -0.8 to -2.5 µm. Data was analyzed by cryoSPARC (v3.3.1) [217].

*Cryo-EM image processing*

Details of the image processing are shown in Supplementary Figures S3.13-3.15, S3.17-3.19 and Supplementary Table S3.9. Recorded movies in EER format were imported into cryoSPARC (v3.3.1) [217] and gain-normalized, motion-corrected and dose-weighted using the cryoSPARC implementation of patch-based motion correction. CTF estimation was performed using the patch-based option in cryoSPARC. A small set of particles were manually selected and followed by 2D classification to create a 2D template for the subsequent automatic particle picking. For the sample of spike$_{D614G}$ in complex with the de-novo designed binder, 832,816 particles were automatically selected by template-based picker and subjected to three rounds of 2D classifications, resulting in a particle set of 184,763 particles. The particles were grouped into three classes, using the ab-initio and hetero-refine implementations in cryoSPARC. The best 3D class composed of 97,804 particles was further subjected to another round of ab-initio reconstruction and hetero-refinement. The well-resolved class consisting of 67 432 particles resulted in a 2.6 Å overall resolution global map in C1 symmetry. The binder-RBD region was refined with a soft mask, resulting in a local map at 3.1 Å resolution. For the data processing of the spike$_{Omicron}$-binder complex sample, 1,820,333 particles were picked with the cryoSPARC template-based picker. After two rounds of 2D classifications, 981,561 particles were selected and subjected to ab-initio reconstruction and hetero-refinement, resulting in a set of 595,599 particles. Subsequently, the selected particle set was classified by multiple rounds of 3D classifications in cryoSPARC. The best-resolved 3D class containing 50,758 particles resulted in a 2.8 Å overall resolution map and the binder-RBD region was further improved by performing focused refinement with a soft mask, resulting in a map at 3.3 Å resolution. Resolution for all 3D maps was estimated based on the Fourier shell correlation (FSC) with a cutoff value of 0.143.

For model building of the spike$_{D614G}$-binder, the previous model (PDB: 7BNO, spike$_{D614G}$) was used for the region of spike$_{D614G}$ as a starting model. The model was rigid-body fit into the cryo-EM density in UCSF Chimera [218] and adjusted manually in Coot 0.9.4 [219]. *De novo* building for the binder parts was performed manually in Coot 0.9.4. For building the spike$_{Omicron}$-binder structure, the model (PDB: 7QO7, spike$_{Omicron}$) was fitted into the density and rebuilt and adjusted manually, using UCSF Chimera and Coot 0.9.4. After the structural rebuilding, all the atomic models were refined using the Phenix (v1.19.2-4158) implementation of real.space.refine with general structural restraints [220,221]. Comprehensive validation (cryo-EM), model quality assessment and statistics are in Supplementary Table S3.9. EM densities and atomic models were visualized in ChimeraX (UCSF, v1.3) [222] and Pymol (Schrödinger, v2.0).

*Data availability*

Cryo-EM maps were deposited in the Electron Microscopy Data Bank under the access codes of EMD-14947 (spike$_{D614G}$-binder full and spike$_{D614G}$-binder local maps), EMD-14922 (spike$_{Omicron}$-binder full), and EMD-14930 (spike$_{Omicron}$-binder local). Atomic models were deposited in Protein Data Bank under the access codes of PDB-7ZSS (spike$_{D614G}$-binder), PDB-7ZRV (spike$_{Omicron}$-binder full) and PDB-7ZSD (spike$_{Omicron}$-binder local). Crystal structures have been deposited in the Protein Data Bank under

accession codes 7XYQ (DBL1_03/PD-L1 complex) and 7XAD (DBL2_02/PD-L1 complex). The PDBbind database (2018 release), PRISM database, ZDock benchmark and SabDab database are available with the following links respectively: http://pdbbind.org.cn/index.php, http://cosbi.ku.edu.tr/prism, https://zlab.umassmed.edu/benchmark/ and http://opig.stats.ox.ac.uk/webapps/sabdab. The scaffold database generated for grafting the seed provided by MaSIF-seed is available at https://zenodo.org/record/7643697#.Y-z533ZKhaQ

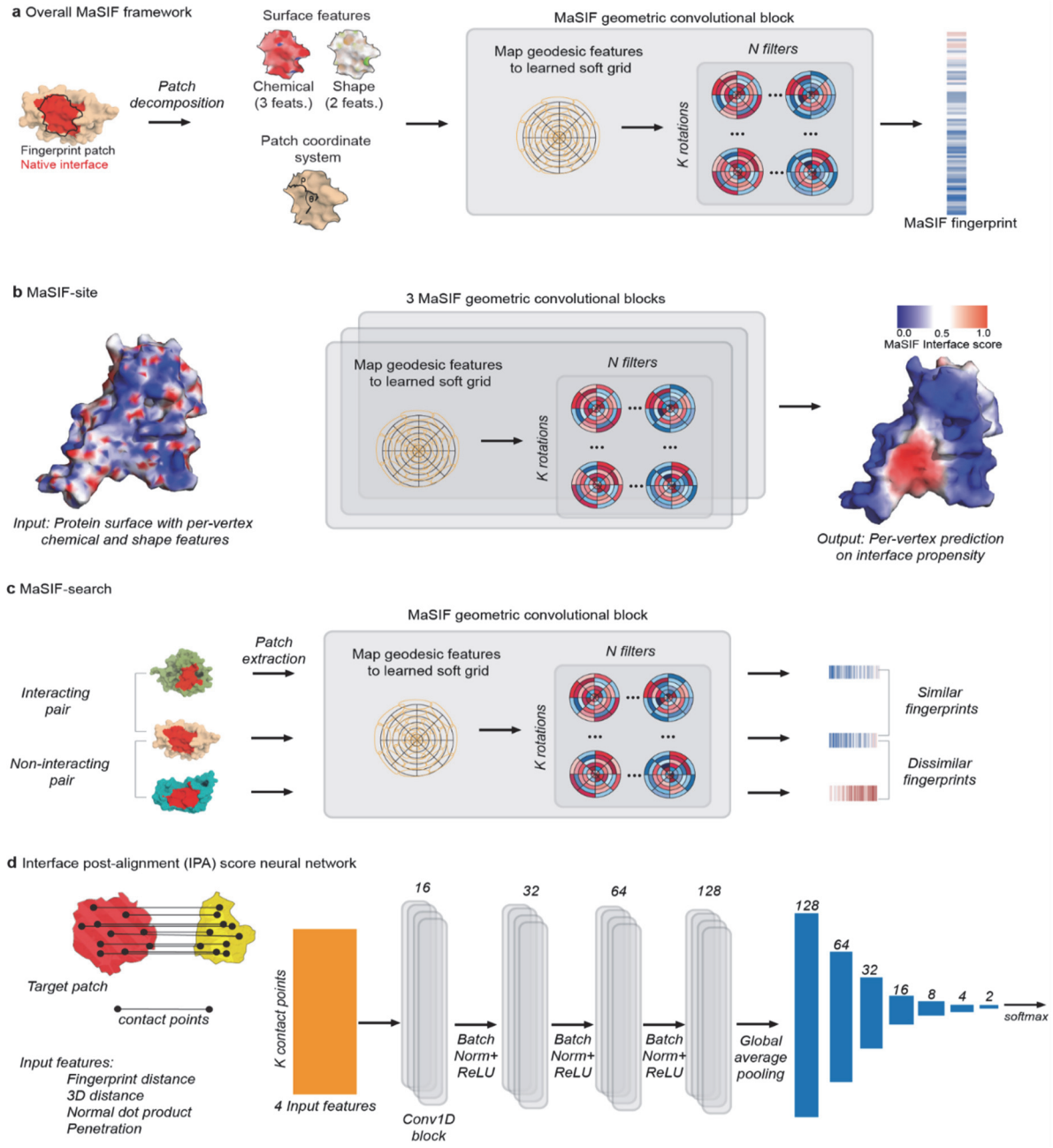
*Code availability*

MaSIF-seed and the Rosetta design scripts are available at https://github.com/LPDI-EPFL/masif_seed

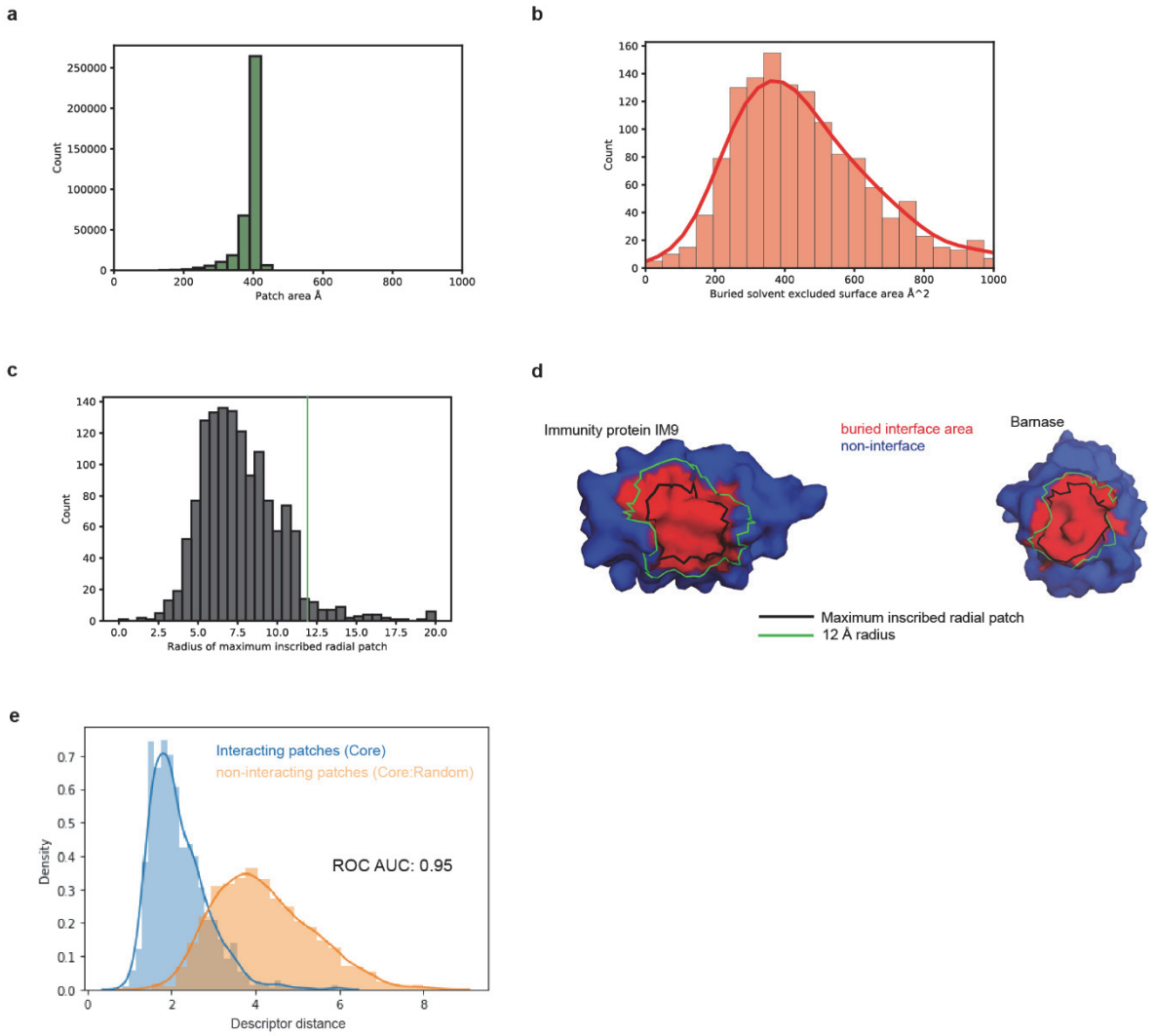
*Acknowledgments*

We thank the Dubochet Center for Imaging (DCI) in Lausanne for cryo-EM data collection. The DCI is an initiative of the EPFL, University of Lausanne and University of Geneva. We thank K. Lau and F. Pojer from the PTPSP facility at EPFL for providing SARS-CoV-2 spike proteins and assistance with cryo-EM. We thank SCITAS at EPFL for support in the computational simulations. We thank the GECF for assistance in deep sequencing and FCCS for assistance in FACS. We thank Emmanuel Levy, Sarel Fleishman, and Mihai Azoitei for their feedback on the manuscript.

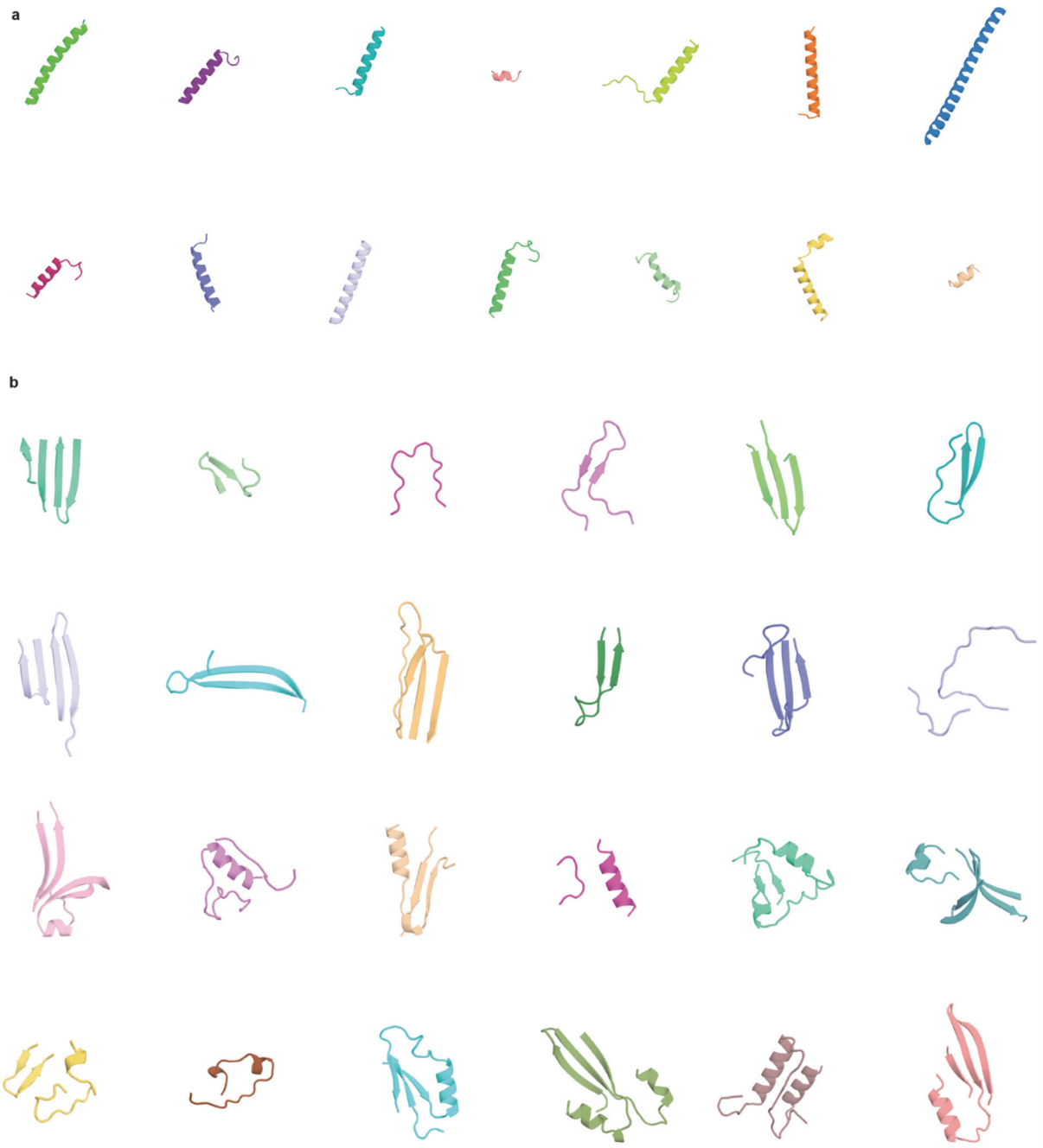
## 3.6 Supplementary materials



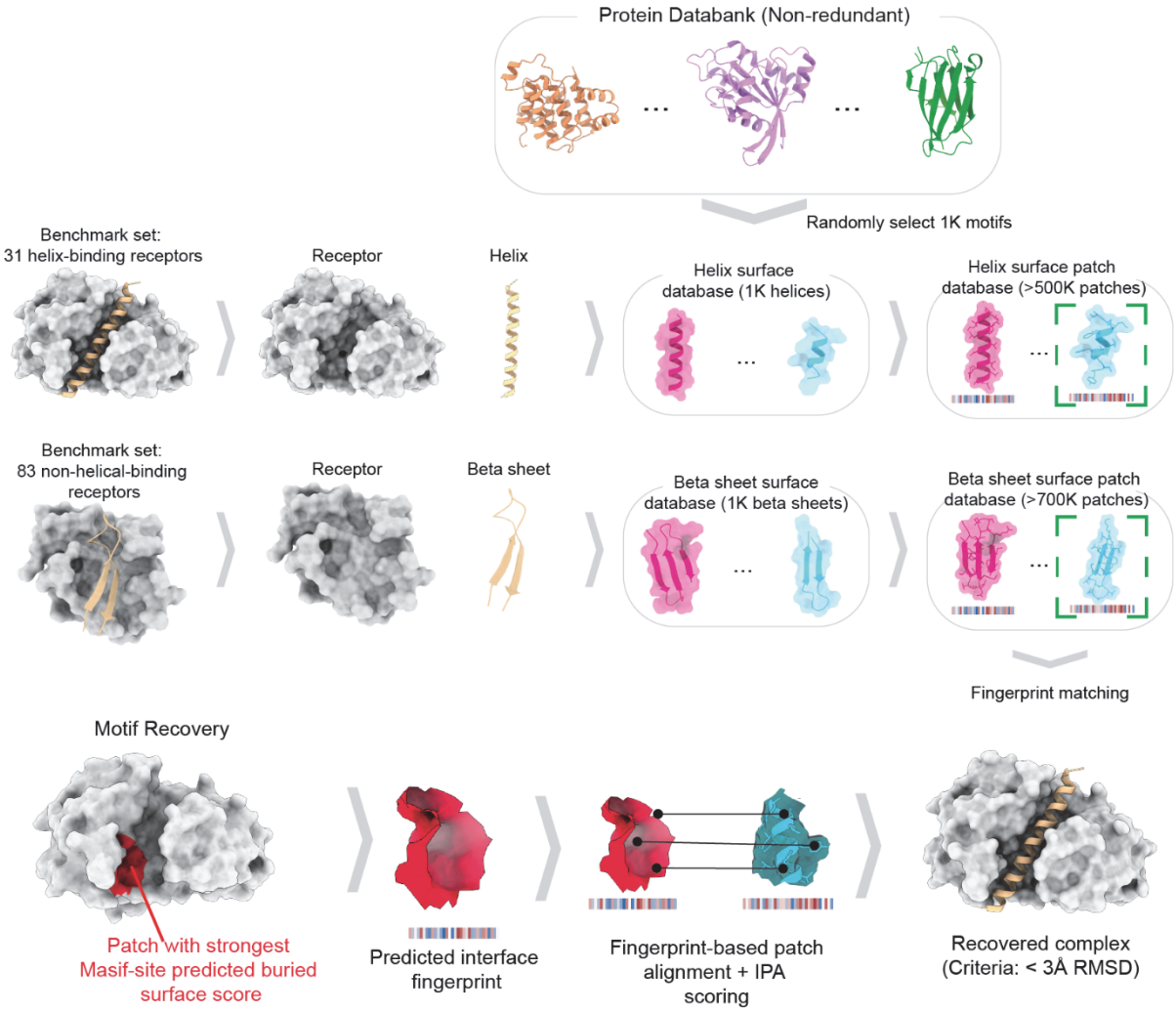

**Supplementary Figure S 3.1 : Overview of the neural network architectures used in the MaSIF protocols**. **A.** General MaSIF framework. Molecular surfaces are decomposed into patches which are annotated with chemical and shape features. The MaSIF network translates these input features into fingerprints that describe the original surface patch. **B.** MaSIF-site neural network. MaSIF-site predicts partner-independent protein interface propensities based on per-vertex chemical and shape features of the protein surface. **C.** MaSIF-search neural network. MaSIF-search embeds protein patches into a space where complementary patches are close to each other. The network was trained on discriminating interacting patches from non-interacting protein surface patches. The network uses MaSIF fingerprints to identify which are compatible and therefore to predict likely interacting proteins. **D.** Interface post-alignment (IPA) scoring neural network. The IPA scoring neural network enables the scoring of protein interfaces based on several input features: fingerprint distance between contacting points, 3D distance of corresponding points, normal dot product, and the distance between surface points in the seed and the closest atom in the target, which we call 'penetration'.



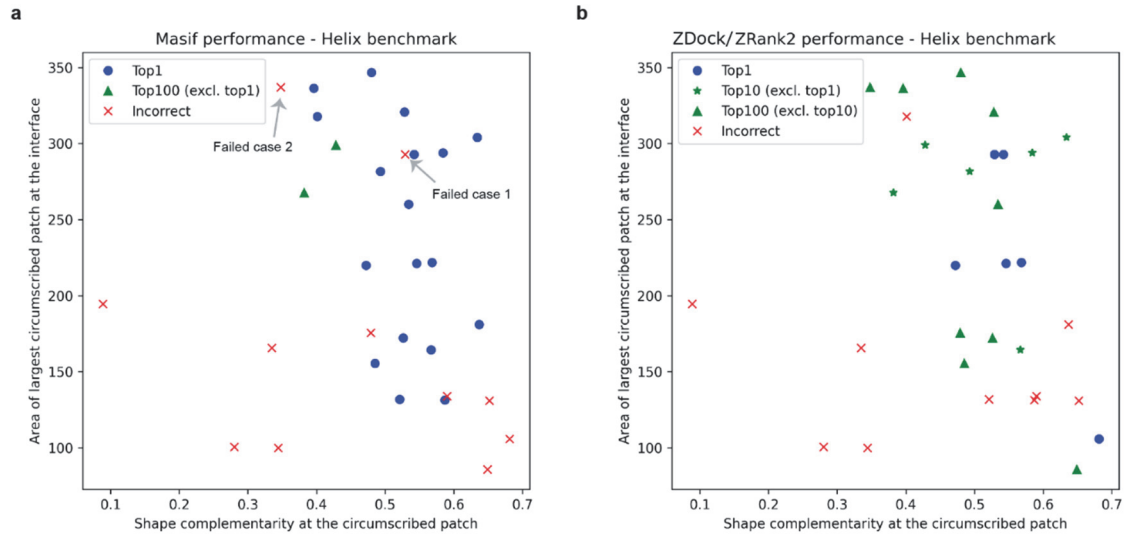

**Supplementary Figure S 3.2 : Modeling buried surfaces as radial patches. A.** Histogram of the patch areas of thousands of randomly selected protein patches with a fixed radius of 12 Å. **B.** Histogram of the area of the buried surface area on 1380 dimeric PPIs. We note that areas are computed for only one of the proteins (i.e. each subunit in a PPI is computed separately), and that we used the solvent excluded surface area, while other authors report buried areas on the solvent accessible area that include the buried surface area of both proteins (see methods). **C.** Size of the maximum inscribed radial patch for the 1380 proteins (see methods). Patch area for the radius used here (12 Å), using a set of 30,000 randomly selected patches. **D.** Example of the buried interface area for two well known, high affinity binders, Immunity Protein IM9 (PDB ID: 1EMV) and the protein Barnase (PDB ID: 1BRS). The buried interface of each protein when bound to its partner is shown in red. The maximum inscribed radial patch's circumference is shown in black, and the circumference of a patch with radius 12 Å is shown in green. **E.** Histogram of similarities between MaSIF-search fingerprint similarity between: (blue) pairs of patches that are co-crystallized from transient PPIs, with the fingerprint computed for the patch centered on the largest inscribed radial patch, and (orange) pairs of patches where one was taken from the center of the interface of a random PPI and the other was taken from a random patch surface.




**Supplementary Figure S 3.3 : Overview of helical and non-helical seeds used in the recovery benchmark.** Examples of **A.** helical seed, **B.** non-helical seeds that were extracted for the recovery benchmark.

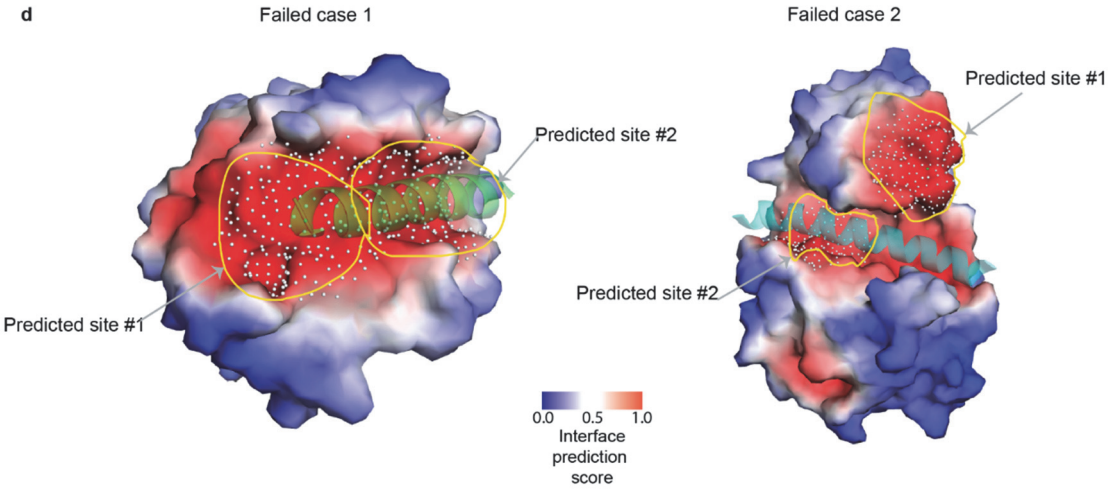



**Supplementary Figure S 3.4 : MaSIF-seed benchmarking for the discrimination of helical or non-helical binding motifs.** A non-redundant set of 31 helical and 83 non-helical fragments that bind to known protein receptors was selected as a benchmark set to evaluate MaSIF-seed's capacity to recover true binding motifs from decoys, and to correctly rank them among the top results. To generate the decoy set, a non-redundant set of all protein chains in the Protein Data Bank was decomposed into continuous helical segments (left) and two/three-stranded beta sheets (right), resulting in over 250K helical and over 380K beta motifs, respectively. One thousand of these motifs each were randomly selected to act as decoys in the respective benchmarks. The surfaces for the two sets of 1000 motifs were computed and decomposed into radial patches and for each patch a fingerprint was computed. Recovered complexes were considered correct if an iRMSD < 3 Å was obtained. A comparable procedure was applied to the benchmark tools.

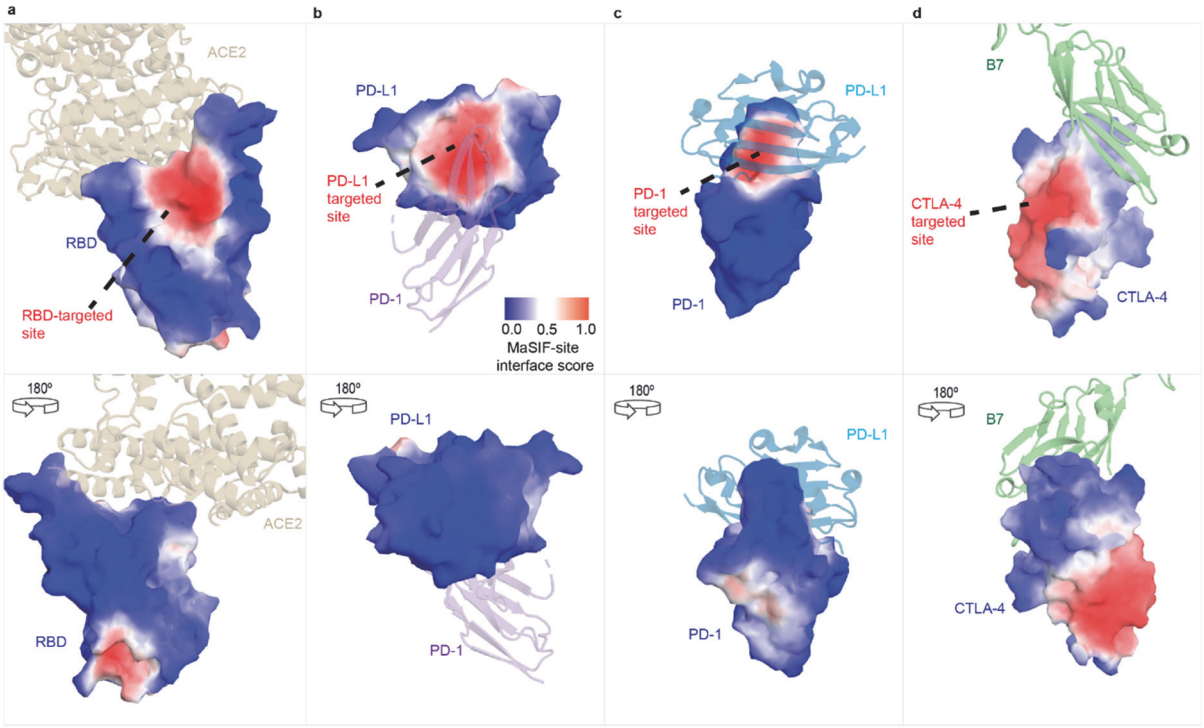



c

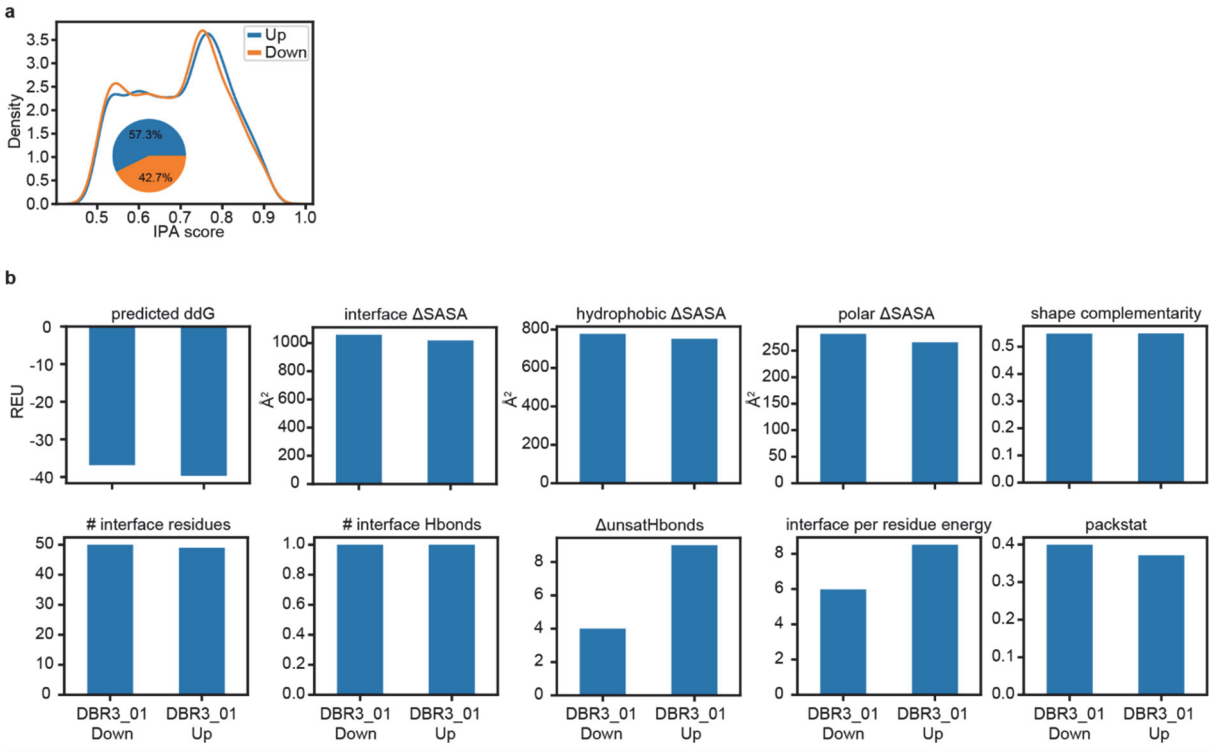
| | Benchmark cases solved by MaSIF-seed but not by ZDock/ZRank2 | Benchmark cases solved by ZDock/ZRank2 but not by MaSIF-seed | Benchmark cases solved by both MaSIF-seed and ZDock/ZRank2 |
|---|---|---|---|
| In the Top 1 | 14 | 2 | 4 |
| In the Top 10 | 10 | 4 | 8 |
| In the Top 100 | 4 | 5 | 16 |



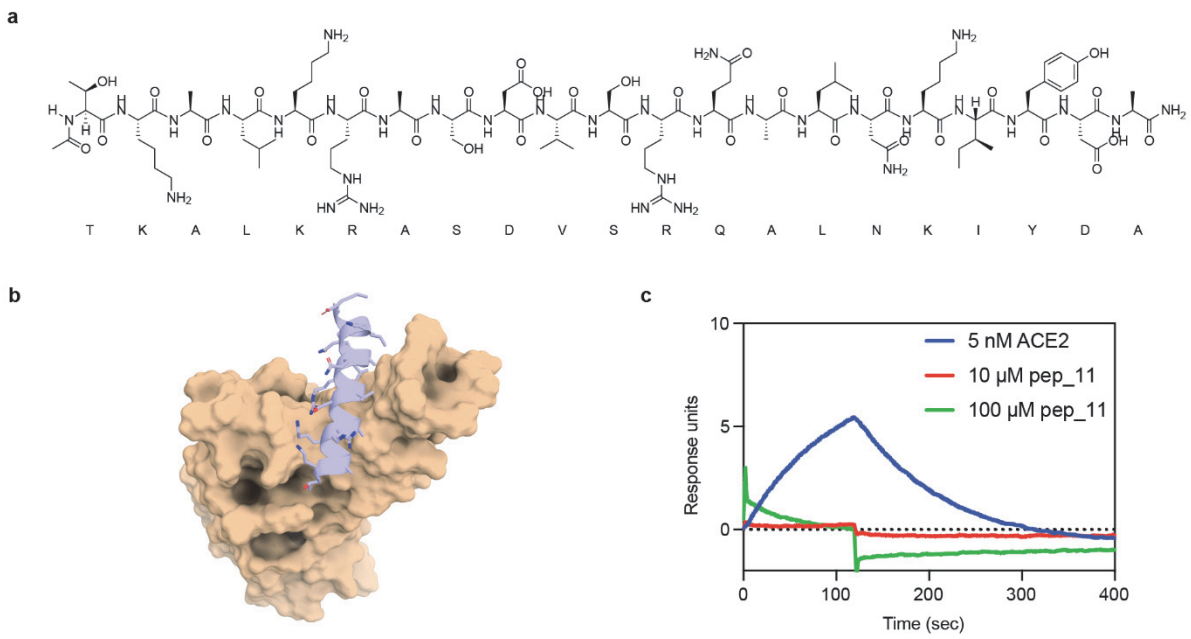

**Supplementary Figure S 3.5 : Analysis of successful/failed helical benchmark cases and comparison between MaSIF-seed and ZDock/ZRank2 performance. A-B**. Plotting of Top 1, Top 10, Top 100 and failed cases for MaSIF-seed and ZDock/ZRank2, showing the maximum circumscribed patch area in the buried interface (y-axis) and the median shape complementarity for vertices of that patch (x-axis) for a, MaSIF-seed, and b, ZDock/ZRank2. **C.** Comparison of cases solved by only MaSIF-seed, only ZDock/ZRank2, or both MaSIF-seed and ZDock/ZRank2 in the Top 1, top 10 or Top 100 rank. **D.** Analysis of two cases that showed both a large circumscribed patch and high complementarity at that patch where MaSIF-seed failed. Left (Failed case 1) shows the BHRF1:Bak BH3 complex (PDB ID: 2XPX); right (Failed case 2) shows proteinase A complexed with a IA3 mutant (PDB ID: 1G0V). In both cases, MaSIF-seed failed because it identified a different site as the top site, but increasing the number of sites explored to the top two resulted in successful predictions. The white dots on the surface denote the predicted site patches.



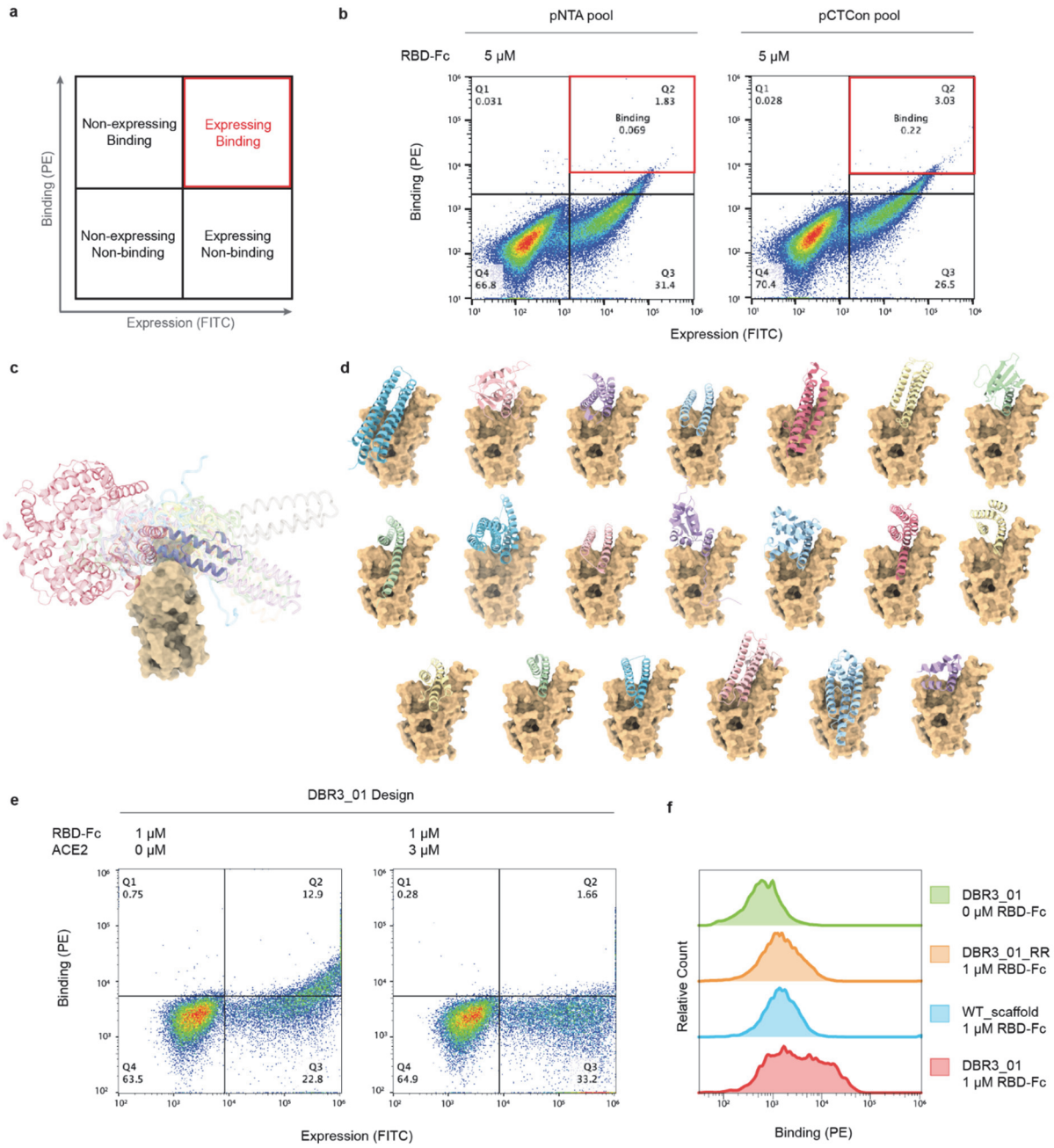

**Supplementary Figure S 3.6 : MaSIF-site target site prediction on SARS-CoV-2 RBD, PD-L1, PD-1, and CTLA-4.** Surface mode shows a MaSIF-site per-surface-vertex regression score on the propensity of each point on the surface to form an interface ranging from 0 (blue) to 1 (red) a-c, Predictions on each target, with the natural ligand of the target shown in cartoon representation as a reference. The structures highlight the predicted site and the bottom row shows a 180 degree rotation. **A**. MaSIF-site prediction on SARS-CoV-2 RBD (PDB ID: 6M17), with the RBD shown in surface and the ACE2 in beige. **B.** Prediction on PD-L1 (PDB ID: 5JDS), with PD-1 shown in purple. **C.** Prediction on PD-1 (PDB ID: 4ZQK) with the natural binder PD-L1 shown in cyan. **D.** Prediction on CTLA-4 (PDB ID: 5GGV) with the natural binding partner B7 (PDB ID: 1I8L) shown in light green.



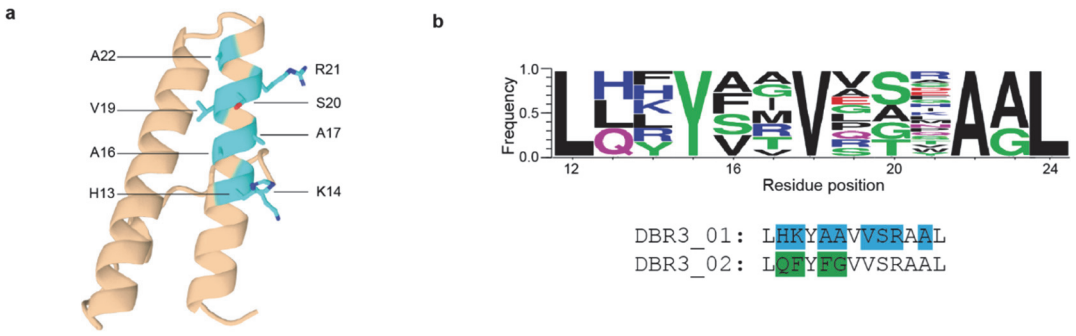

**Supplementary Figure S 3.7 : RBD-binder metrics for up- and down-orientations**. **A.** Distribution of the IPA scores for the seeds of the up- and down-orientations and respective cluster sizes. **B.** Interface metrics (n=1) of the DBR3_01 model in complex with ACE2 for the up- and down-orientations were computed using Rosetta's interface analyzer. The following Rosetta metrics are shown: predicted ddG = change in Rosetta energy of separated versus complexed binding partners, interface ΔSASA = solvent accessible surface area buried at the interface, hydrophobic ΔSASA = solvent accessible surface area buried at the interface that is hydrophobic, polar ΔSASA = solvent accessible surface area buried at the interface that is polar, shape complementarity = Lawrence and Coleman shape complementarity of the interface surfaces, # interface residues = number of residues at the interface, # interface Hbonds = number of hydrogen bonds across the interface, ΔunsatHbonds = number of buried, unsatisfied hydrogen bonds at the interface, interface per residue energy = average Rosetta energy of each interface residue, packstat = Rosetta's packing statistic score for the interface ranging from 0 (low packing) to 1 (high packing).



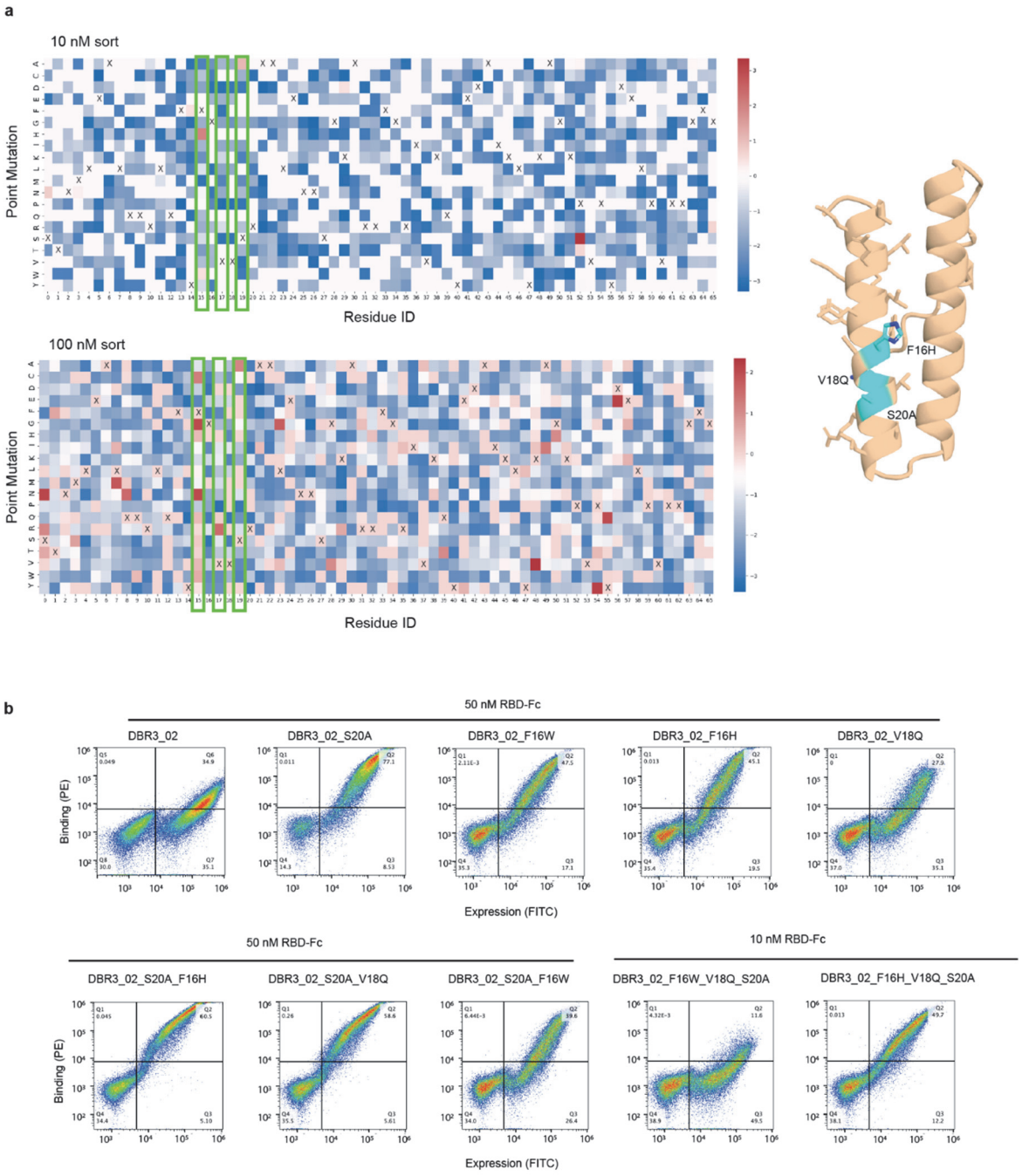

**Supplementary Figure S 3.8 : Binding seed identified by MaSIF tested as a synthetic peptide**. **A.** Structure of the synthesized binding seed. **B.** MaSIF prediction of seed (lavender) binding to RBD (wheat). **C.** SPR data of high concentration of the peptide flowing over RBD. No binding signal is observed for the peptide.



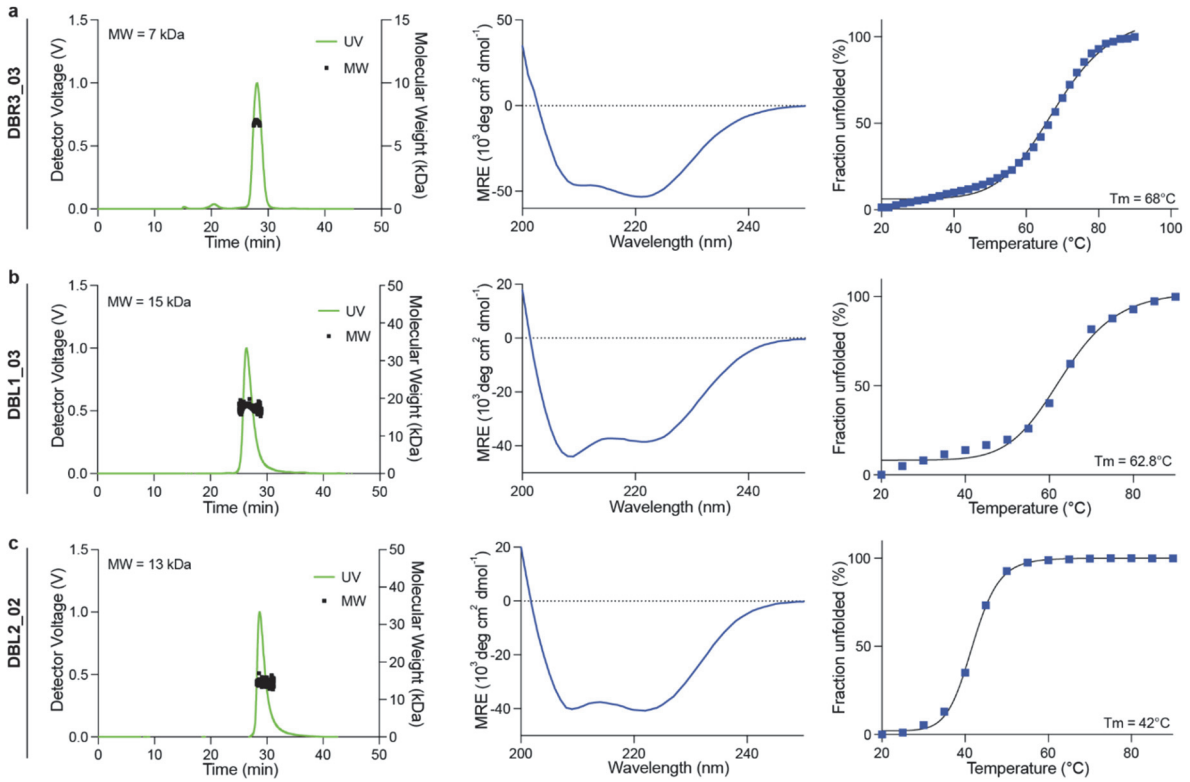

**Supplementary Figure S 3.9 : RBD-binder designs displayed on yeast. A.** The yeast display protocol utilizes PE to label binding and FITC to label expression. Yeast appearing in the double positive quadrant are considered potential binders and sorted for enrichment. **B.** Pools of approximately 30 designs were displayed on the surface of yeast and the highest binding populations (red box) sorted for further analysis**. C.** Schematic of RBD (wheat) bound to the various members of the library (transparent silhouettes and purple for DBR3_01) and ACE2 (red) overlapping with the designed binders. **D.** Individual designs DBR1-DBR20. **E.** DBR3_01 design displayed on yeast binds to RBD-Fc (left panel) but the binding is blocked when the RBD-Fc is preincubated with an excess of ACE2, indicating a competitive binding mode. **F.** A point mutant in the binding interface (DBR3_01_RR) and the original scaffold protein (WT_scaffold) show lower binding signal than DBR3_01 with 1 µM RBD-Fc, indicating that the design is engaging the RBD with the predicted interface.



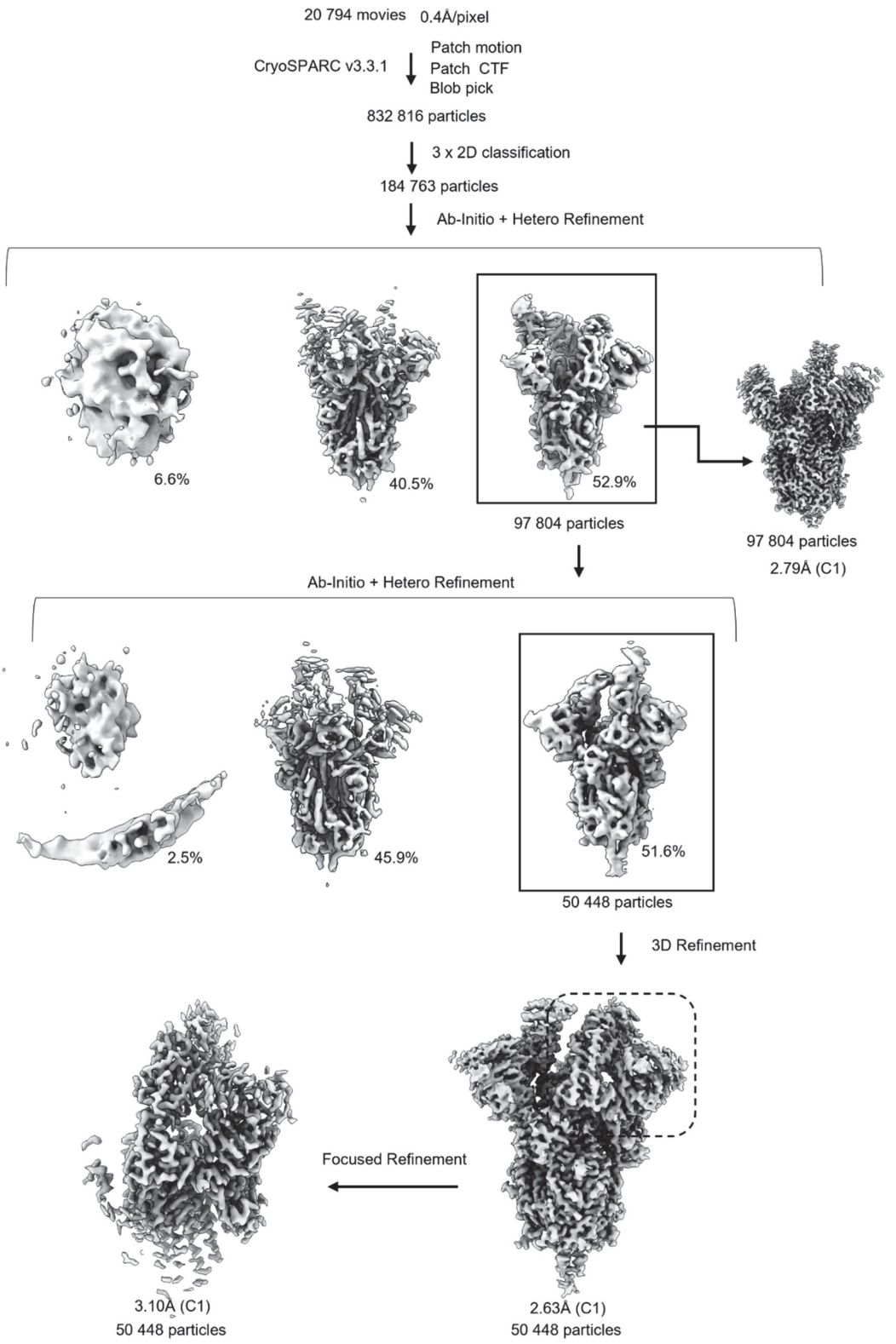

**Supplementary Figure S 3.10 : Directed Library for DBR3_01. A.** Position of residues included in a combinatorial library to improve binding affinity. **B.** Sequence logo plot of specific mutations allowed within the library. The sequences list the residues mutated in DBR3_01 (highlighted in blue) and the mutations gained through the library in DBR3_02 (green).



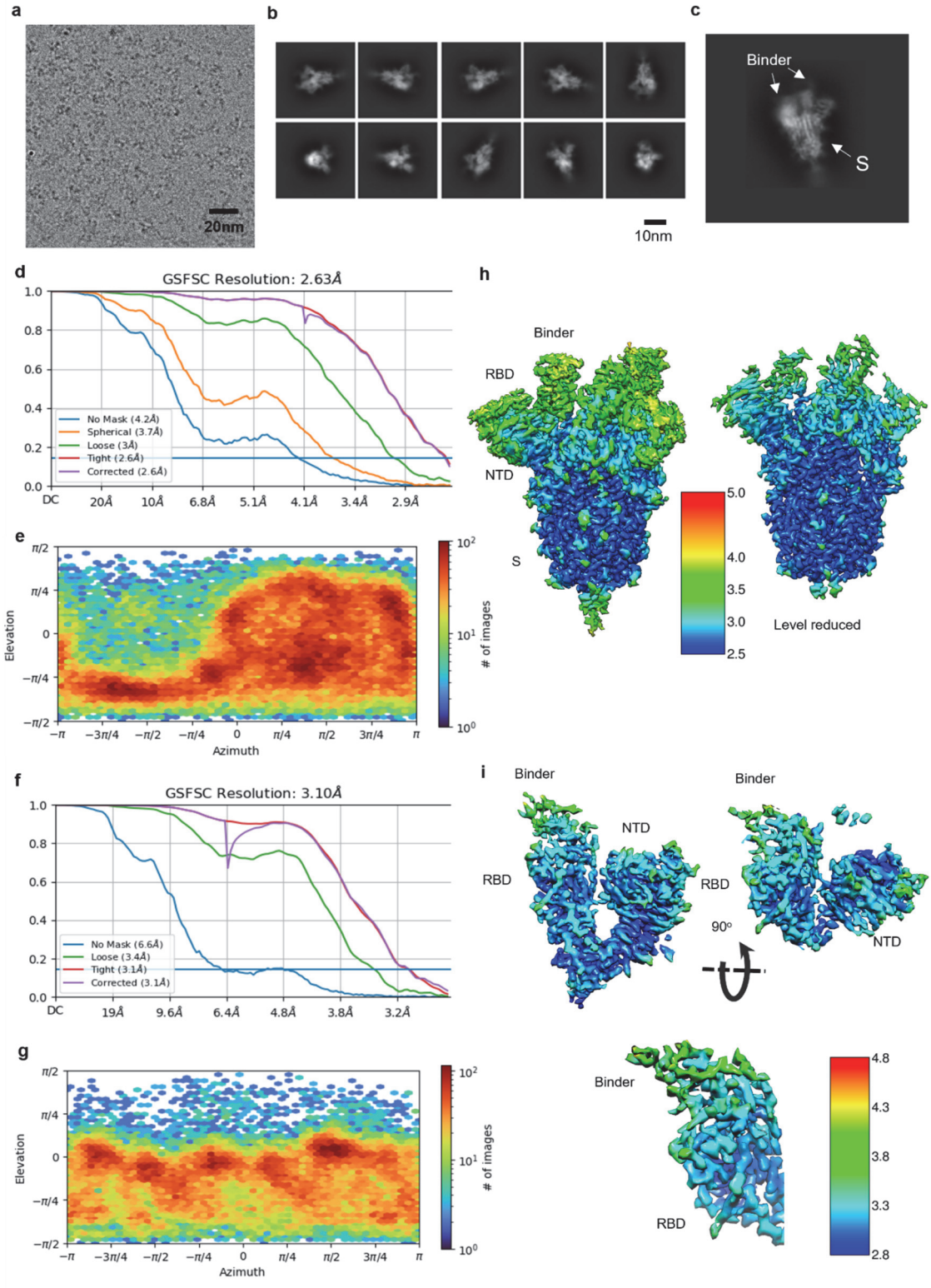

**Supplementary Figure S 3.11 : SSM of DBR3_02. A.** Heat maps of DBR3_02 SSM at two concentrations of RBD-Fc. X indicates the original amino acid of DBR3_02. Red indicates an enrichment of the mutation in the binding population, blue indicates an enrichment in the non-binding population. Three positions, green box, were enriched in both concentrations. The positions of these mutations are highlighted on the DBR3_03 structure. **B.** Yeast display of DBR3_02 with mutations from the SSM introduced shows increase in affinity to RBD.



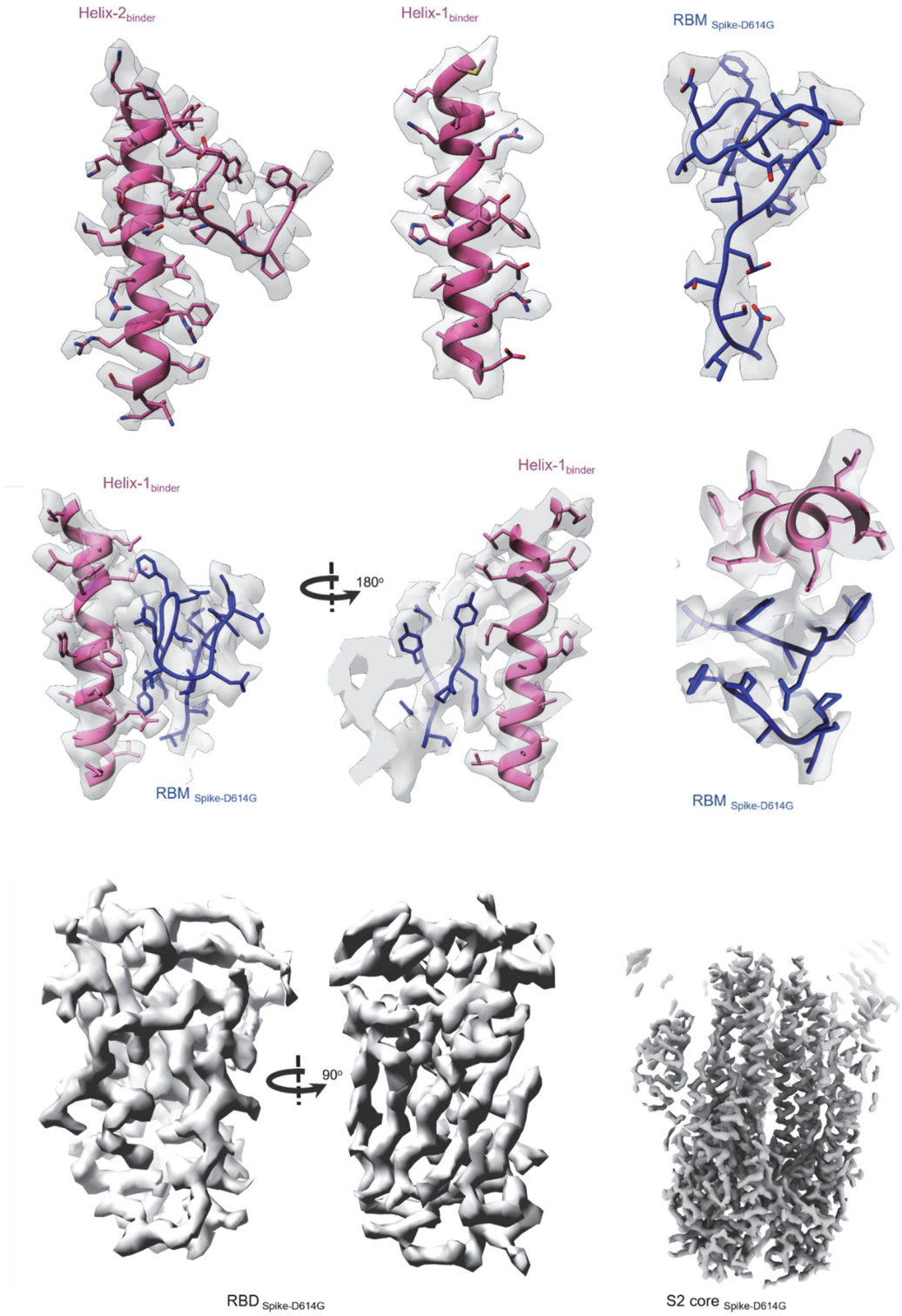

**Supplementary Figure S 3.12 : Biophysical characterization of the designed binders.** From left to right: The oligomeric status was determined via multi-angled light scattering (MALS). Folding was measured using circular dichroism. Thermal stability was determined by plotting the ellipticity at 218 nm at increasing temperatures. a, DBR3_03, b, DBL1_03, c, DBL2_02.



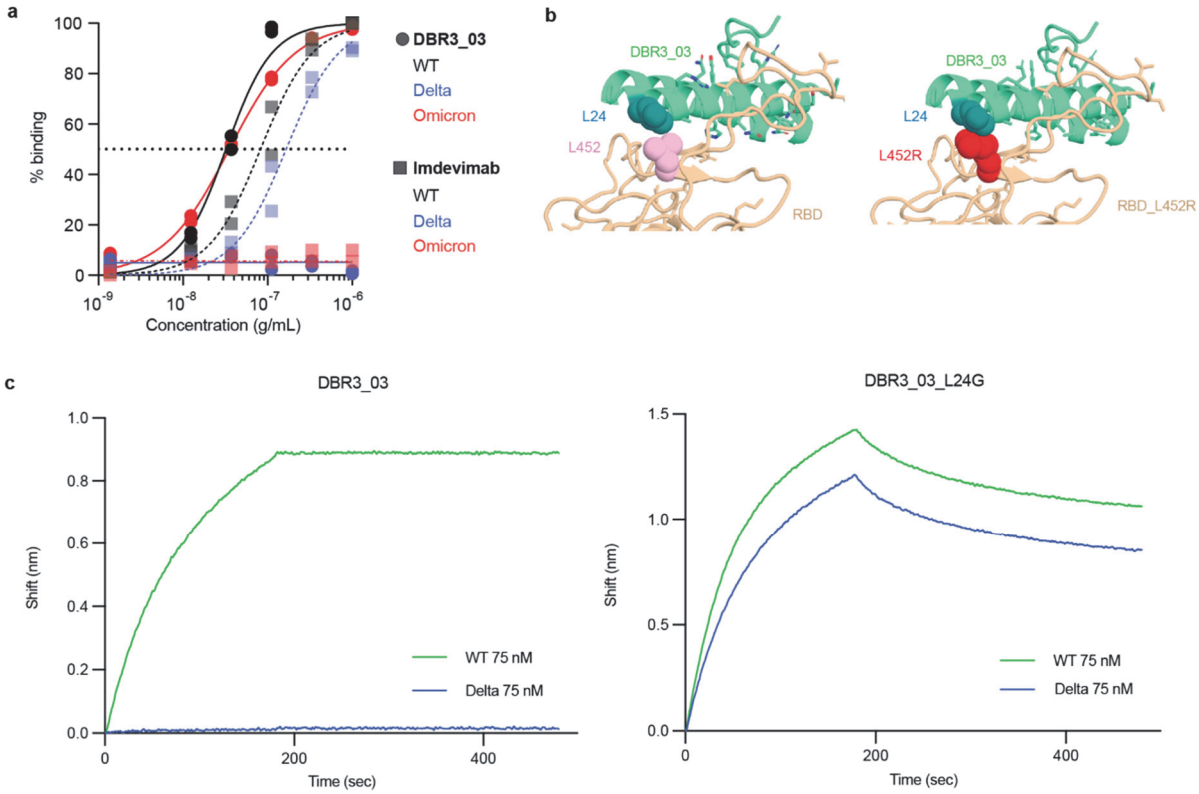

**Supplementary Figure S 3.13 : Cryo-EM data processing of the D614G Spike-DBR3_03 complex.** Image processing workflows performed in CryoSPARC v.3.3.1.



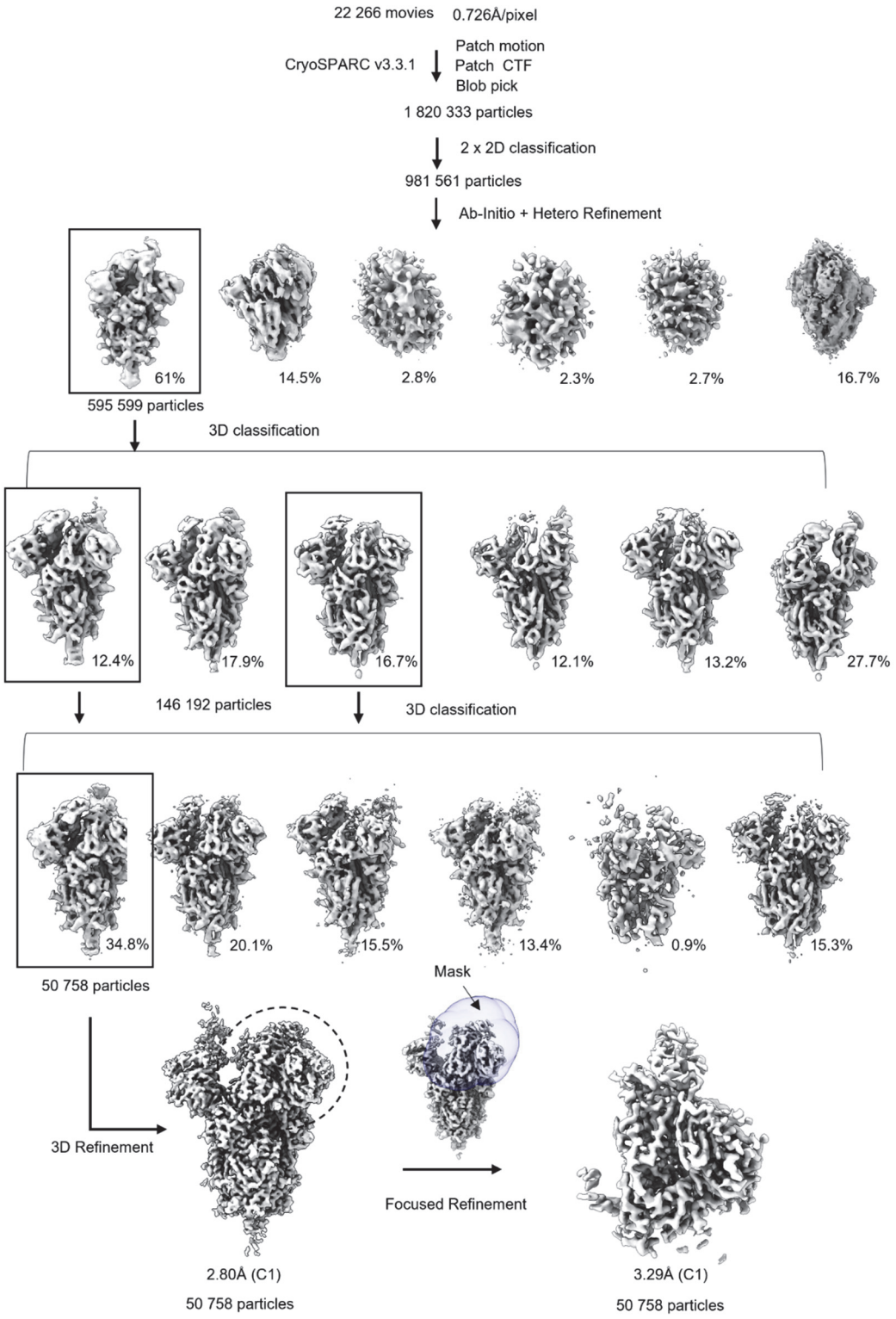

**Supplementary Figure S 3.14 : Details of Cryo-EM data processing for D614G Spike-DBR3_03 complex**. **A.** A representative raw micrograph of the Cryo-EM sample for D614G Spike-binder complex. 20,794 micrographs of such similar quality were acquired for this complex. **B.** The 2D classes of the D614G Spike-binder complex. **C.** A representative 2D class. **D.** Direction distribution of the particle alignment and **E.** FSC curves of the final overall map. **F.** Direction distribution and **G.** FSC curves of the locally refined map. **H-I**. Local resolution distribution of the overall and focused refined maps.



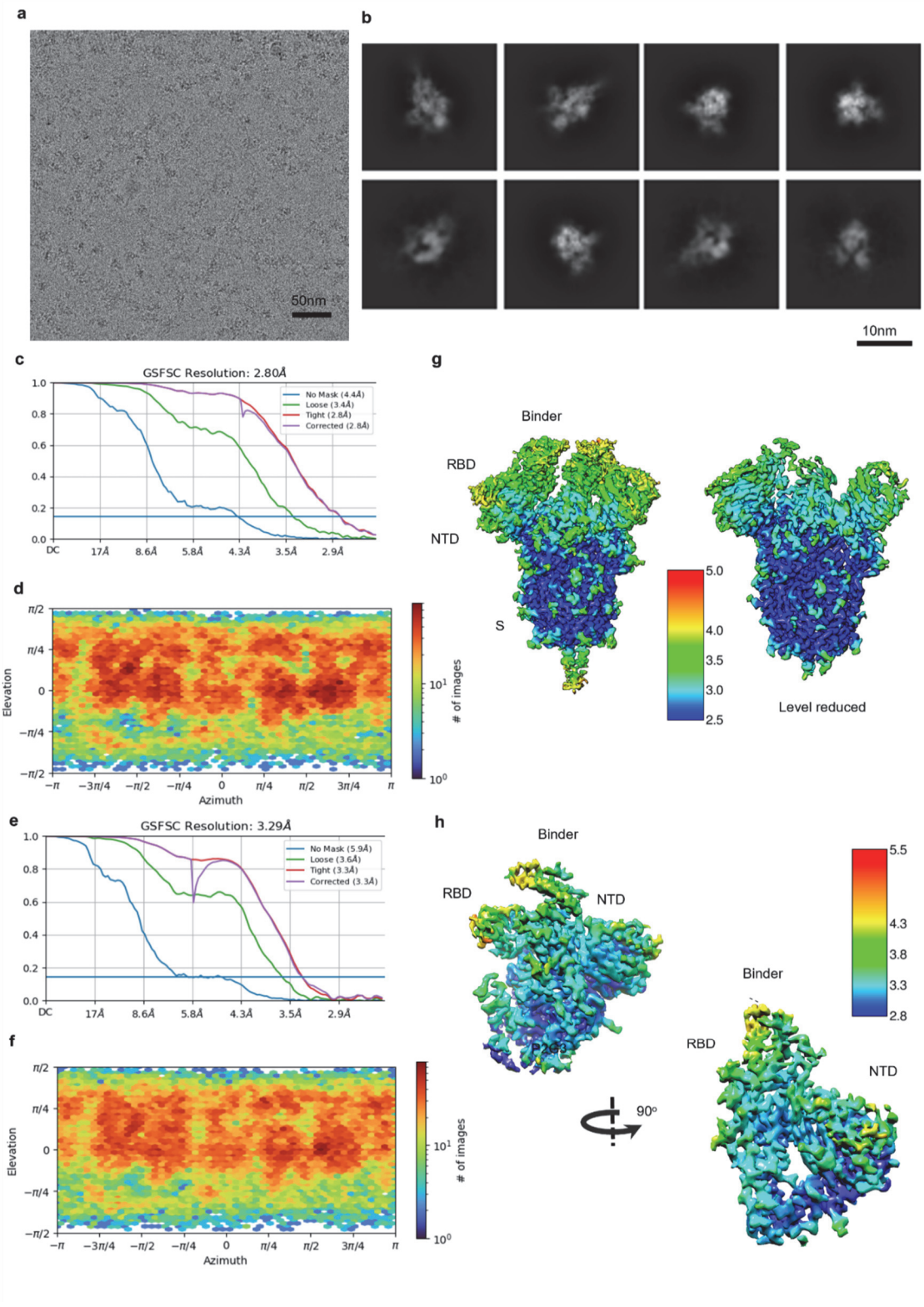

**Supplementary Figure S 3.15 : Highlights of the Cryo-EM densities of DBR3_03 with D614G spike.** Cryo-EM densities are shown as surfaces. RBM (receptor binding motif) in blue with DBR3_03 in pink. The atomic model is shown as stick or ribbon representation.



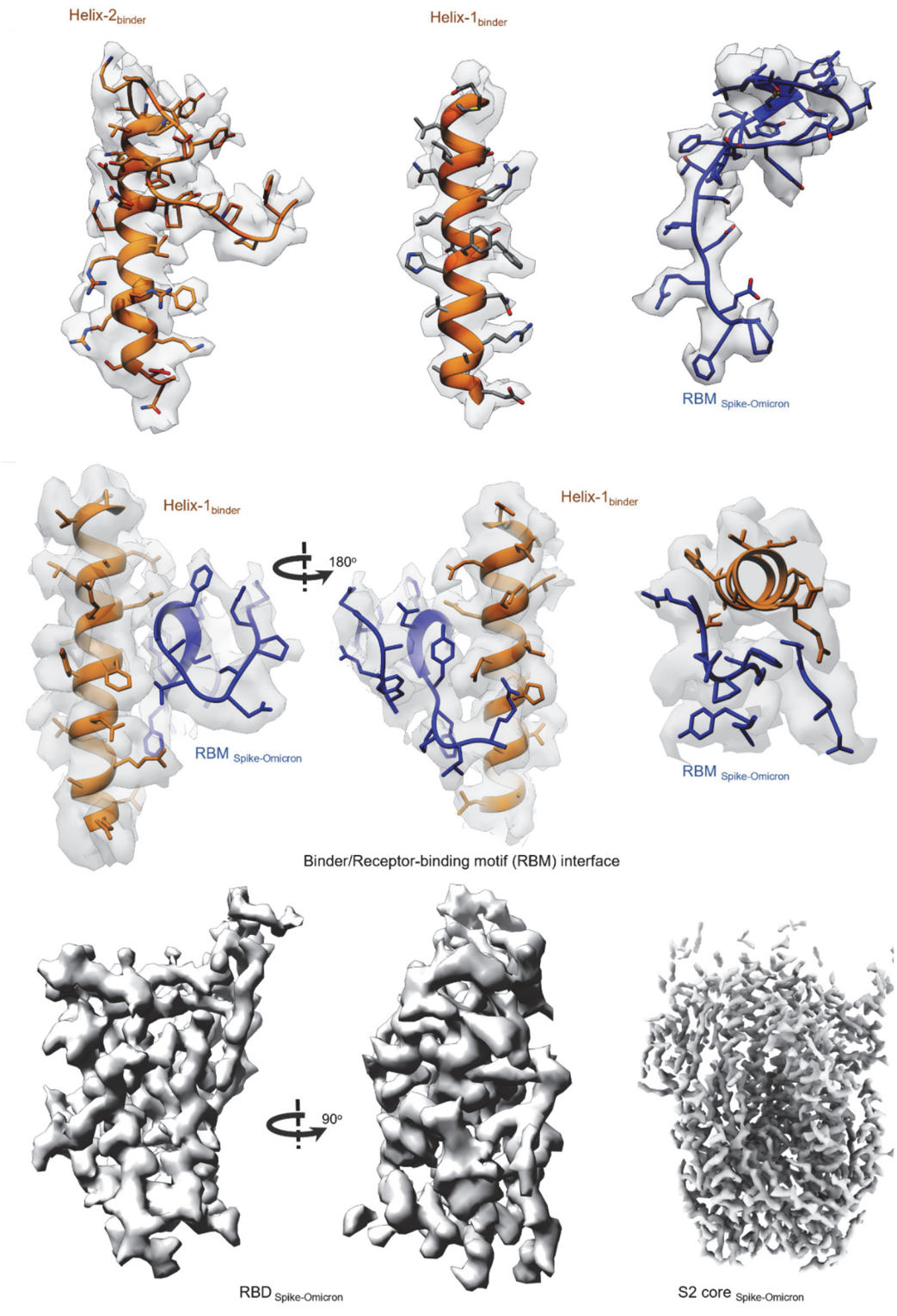

**Supplementary Figure S 3.16 : DBR3_03 binding is sensitive to the L452R mutation in the spike protein**. **A.** Luminex binding assay of DBR3_03 or Imdevimab (REGN10987) with beads functionalized with SARS-CoV-2 spike protein of indicated variants. DBR3_03 has an $EC_{50}$ of $3.2e^{-8}$ g/mL with WT and $3.5e^{-8}$ g/mL with omicron. Imdevimab has an $EC_{50}$ of $8.2e^{-8}$ g/mL with WT and $1.7e^{-7}$ g/mL with delta. The fits were calculated from technical replicates (n=2) using a nonlinear four parameter curve fitting analysis. **B.** The L452R mutation on the spike protein leads to a clash with the DBR3_03 binding. A L24G mutation is proposed to avoid the clash. **C.** BLI data with DBR3_03 (WT $K_D$<0.1 nM, delta $K_D$ not detected) or DBR3_03_L24G (delta $K_D$=6 nM, WT $K_D$=6 nM) immobilized on the tips, dipped into spike protein of different variants.



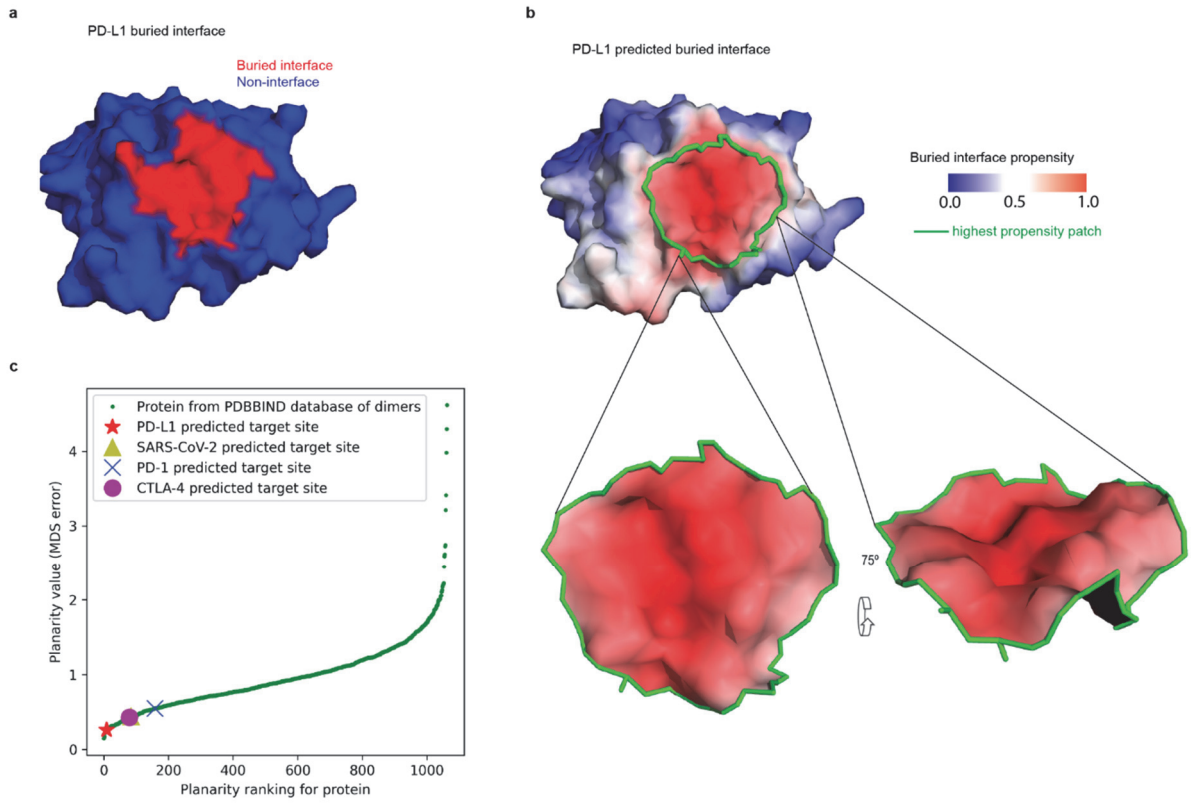

**Supplementary Figure S 3.17 : Cryo-EM data processing of the Omicron Spike-DBR3_03 complex**. Image processing workflows performed in CryoSPARC.



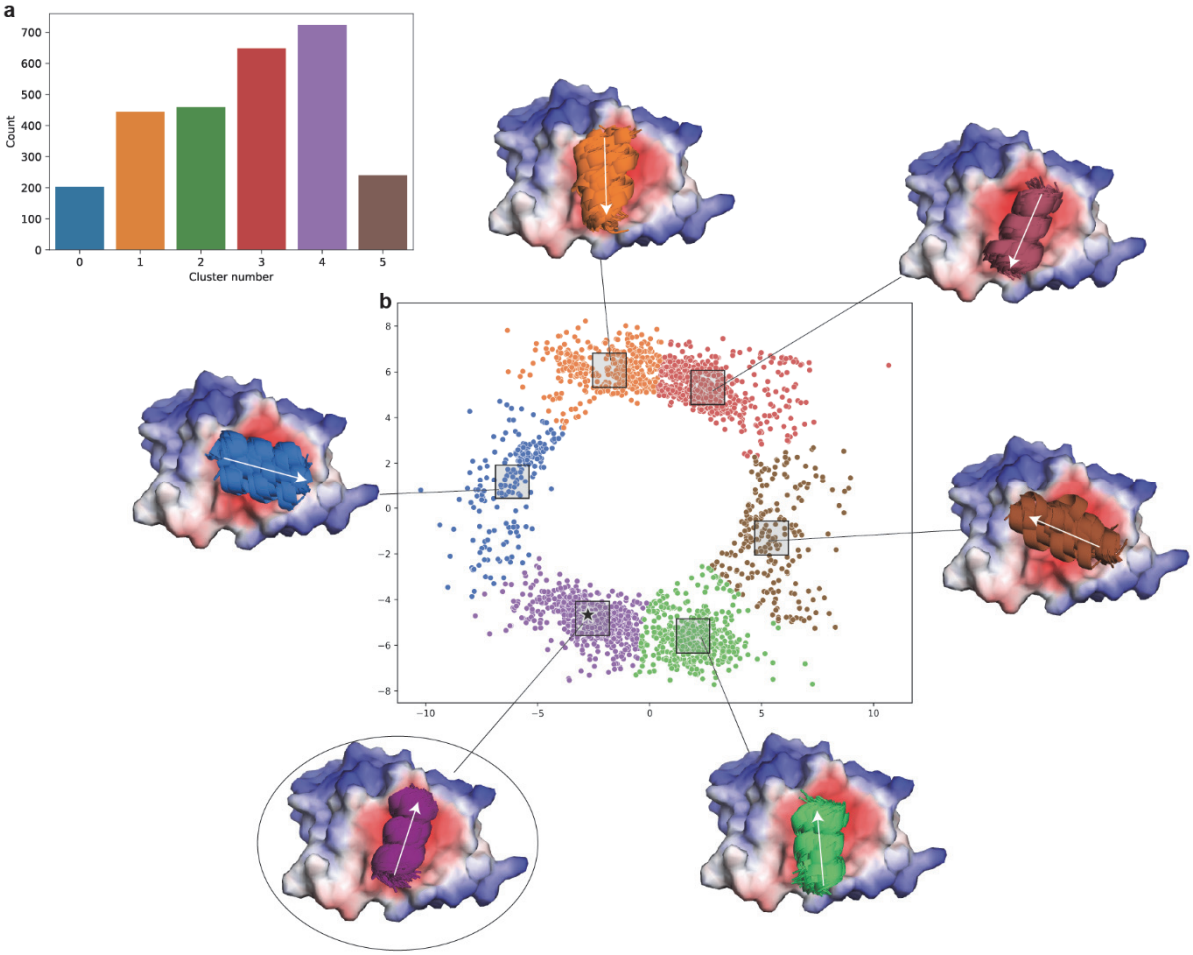

**Supplementary Figure S 3.18 : Details of Cryo-EM data processing for Omicron Spike-DBR3_03 complex.** **A.** A representative Cryo-EM micrograph for the D614G Spike-binder complex. 22,266 micrographs of such similar quality were acquired for this complex. **B.** The representative 2D classes of the omicron Spike-binder complex. **C.** Direction distribution of the particle alignment and **D.** FSC curves of the final overall map. **E.** Direction distribution and **F.** FSC curves of the locally refined map. **G-H.** Local resolution distribution of the overall and focused refined maps.



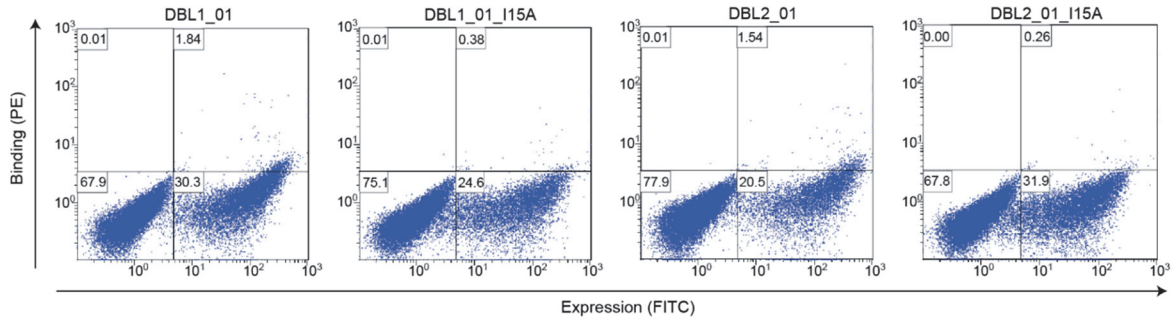

**Supplementary Figure S 3.19 : Highlights of the cryo-EM densities of DBR3_03 with Omicron spike.** Cryo-EM densities are shown as surfaces. RBM (receptor binding motif) in blue with DBR3_03 in orange. The atomic model is rendered as stick or ribbon representation.



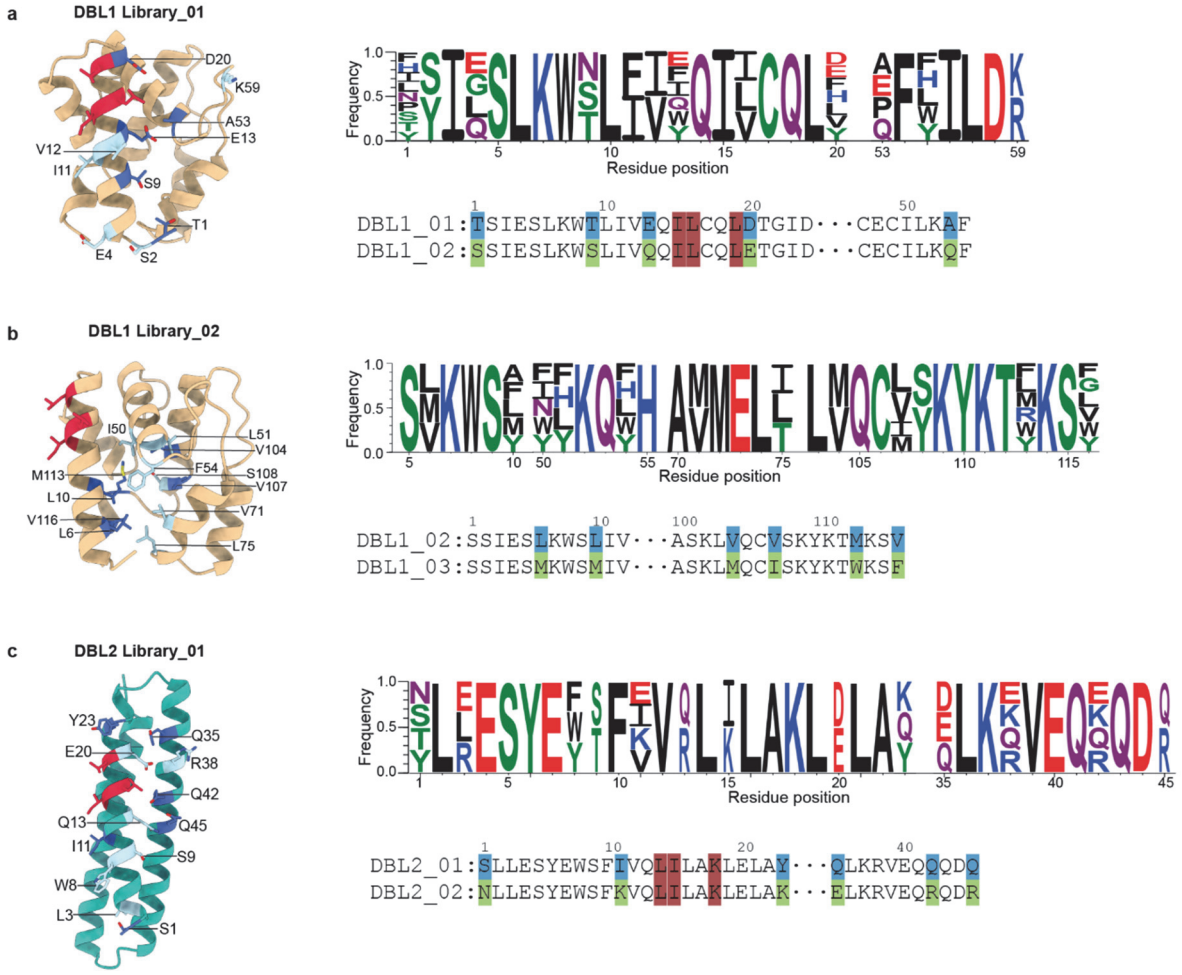

**Supplementary Figure S 3.20 : Planarity of the targeted interface sites. A.** Buried interface on PD-L1 upon complex formation with PD-1. **B.** (Top) PD-L1 predicted buried interface, with selected target patch marked with a green contour. (bottom) View of the selected target patch to show its planarity. **C.** Plotting of the planarity of each of 1068 dimeric protein interfaces. Y-axis: error in multidimensional scaling when flattening the patch from 3D to 2D. X-axis: ranking of each protein according to the planarity value with respect to the dataset of 1068 dimeric protein interfaces. The PD-L1 interface targeted in this work is marked with a red star, SARS-CoV-2 with a gold triangle, PD-1 with a blue X, and CTLA-4 with a magenta circle.



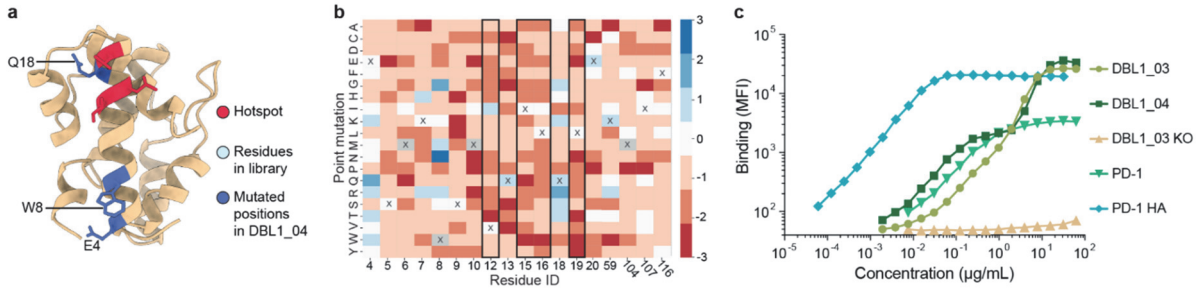

**Supplementary Figure S 3.21 : Clusters of putative binding seeds identified by MaSIF-seed docked on the PD-L1 surface (PDB ID: 5JDS).** 140 million patches from ~250,000 helices extracted from the PDB were compared and docked to the predicted interface in PD-L1 using MaSIF-seed. The top scoring seeds were selected for further processing. Twelve-amino acid fragments of these seeds that occupied the largest buried surface were then clustered using metric multidimensional scaling of all pairwise RMSDs between all seeds. **A.** Histogram of clusters, showing the prevalence of each orientation. **B.** Binding seed clusters in the multidimensional scaling plot. A box is drawn around the center of each cluster and the picture shows the selected helix orientation for all points inside the box. The circled binding seed cluster shows the helix orientation of the seed used for the PD-L1 designs. A star symbol shows the PD-L1 seed used for the designs.



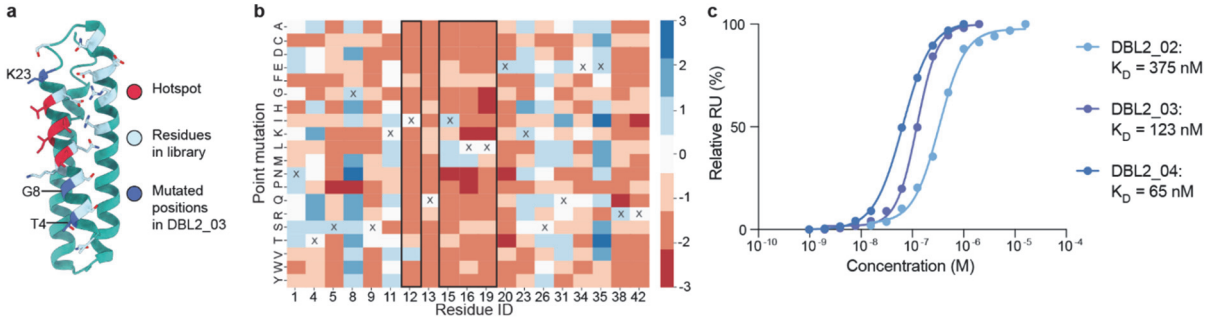

**Supplementary Figure S 3.22 : Binding signals of initial PD-L1 binder designs.** Binding measured on the surface of yeast with 15 µM PD-L1-Fc. Comparison of DBL1_01 and DBL2_01 with corresponding interface mutants.



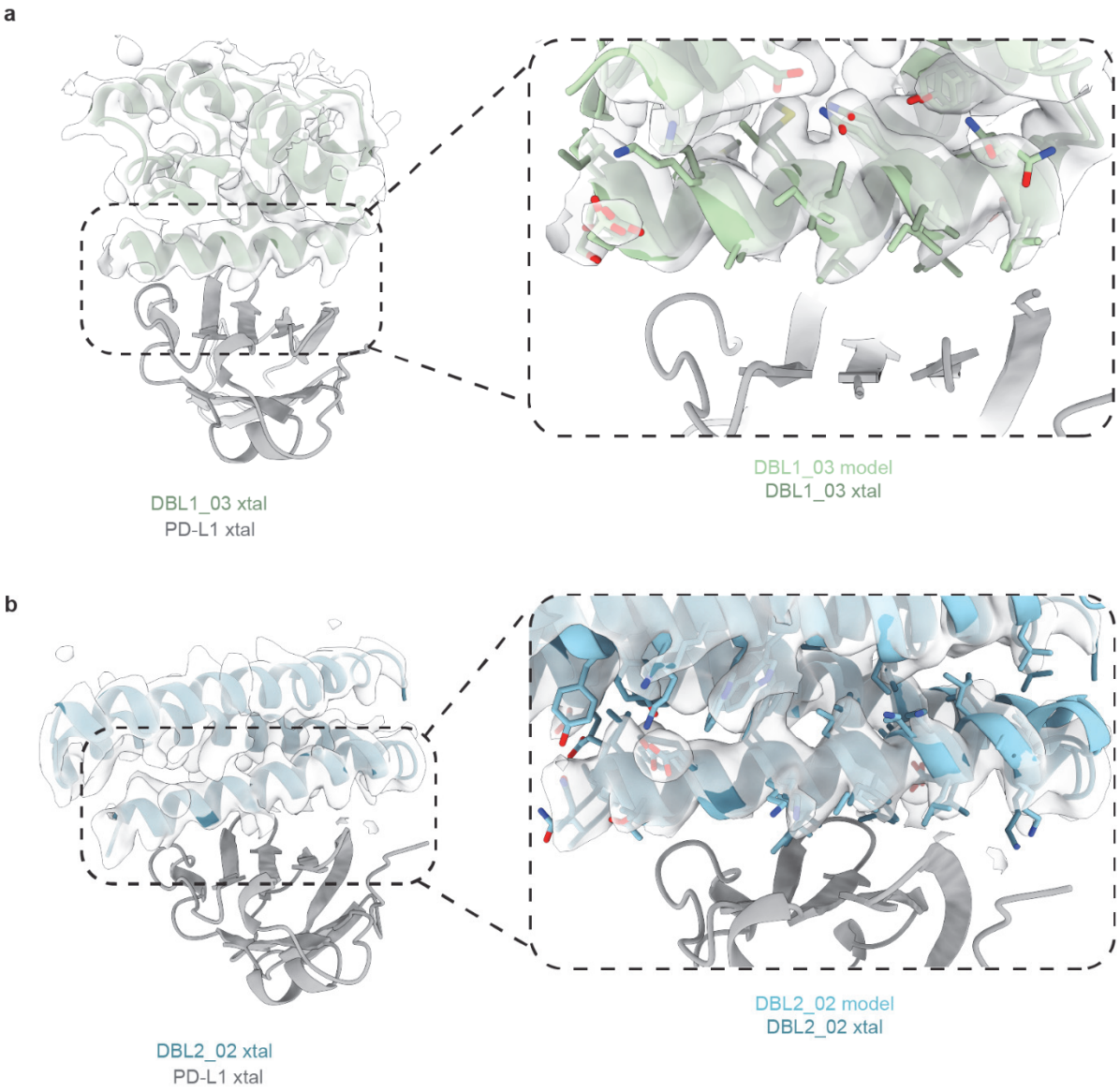

**Supplementary Figure S 3.23 : Composition and outcome of yeast display libraries. A.** Position of targeted residues in the structure of DBL1_01 to improve binding affinity. Logo plot of the allowed mutations in the library and alignment of initial design with library enriched design. **B.** Position of targeted residues in the structure of DBL1_02 to improve core packing. Logo plot of the allowed mutations in the library and alignment of DBL1_02 with library enriched design. **C.** Position of targeted residues in the structure of DBL2_01 to improve binding affinity and solubility. Logo plot of the allowed mutations in the library and alignment of initial design with library enriched design. Hotspot residues red, targeted residues light blue, mutated residues dark blue.



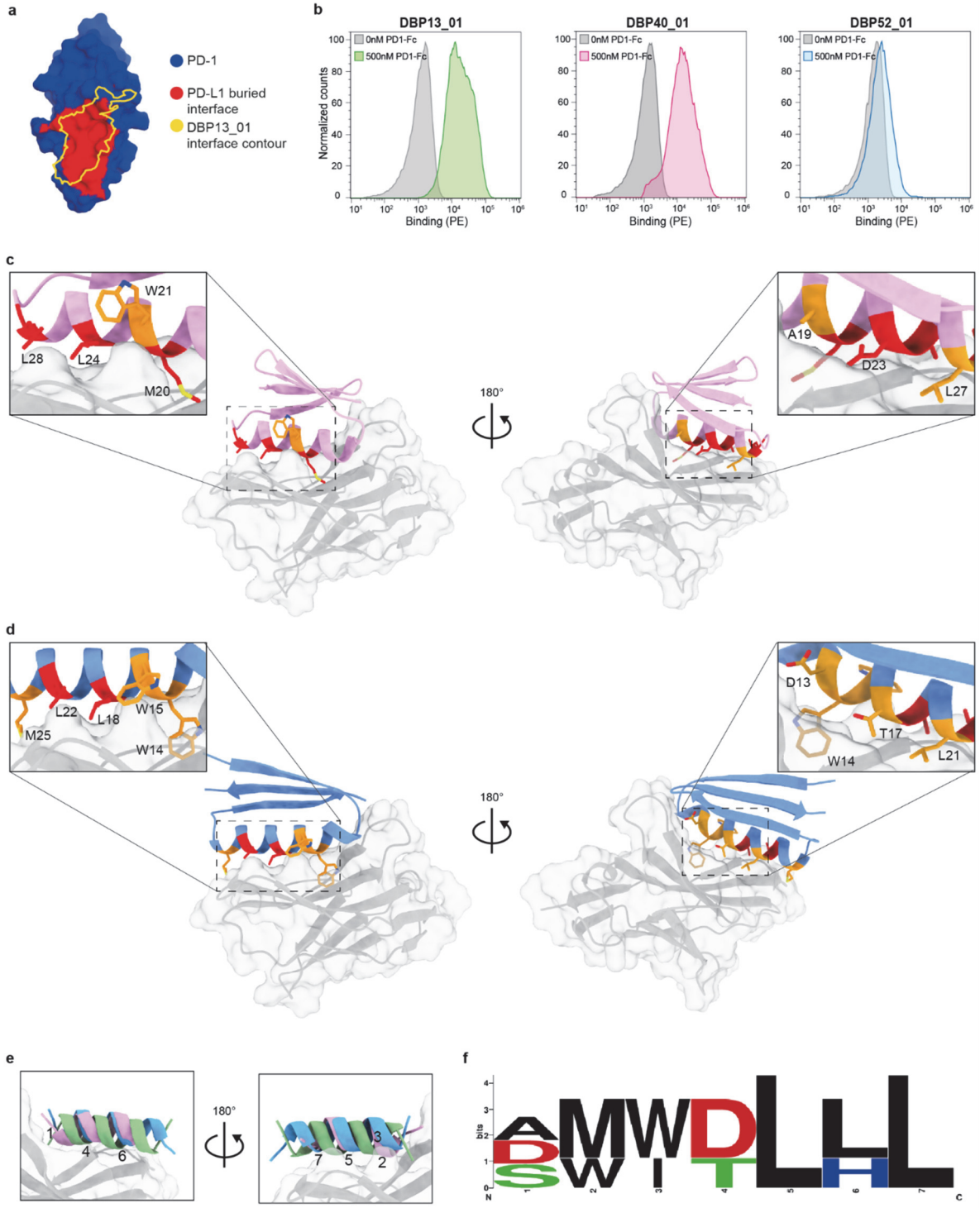

**Supplementary Figure S 3.24 : Complete SSM library of DBL1_03 and cell binding data. A.** Structural representation of all positions sampled in the SSM library (light blue). The four hotspot residues (red) were also sampled. Three positions were mutated in DBL1_04 (dark blue). **B.** Outcome of the entire SSM library. Blue indicates enrichment in the binding population, while red shows enrichment in the non-binding population. **C.** Binding of DBL1_03 and DBL1_04 to KARPAS299 cells expressing PD-L1 compared to binding of WT PD-1, a high affinity version of PD-1 (PD-1_HA)[35] and a V12R mutation of DBL1_03 (KO). All proteins contained a Fc domain.



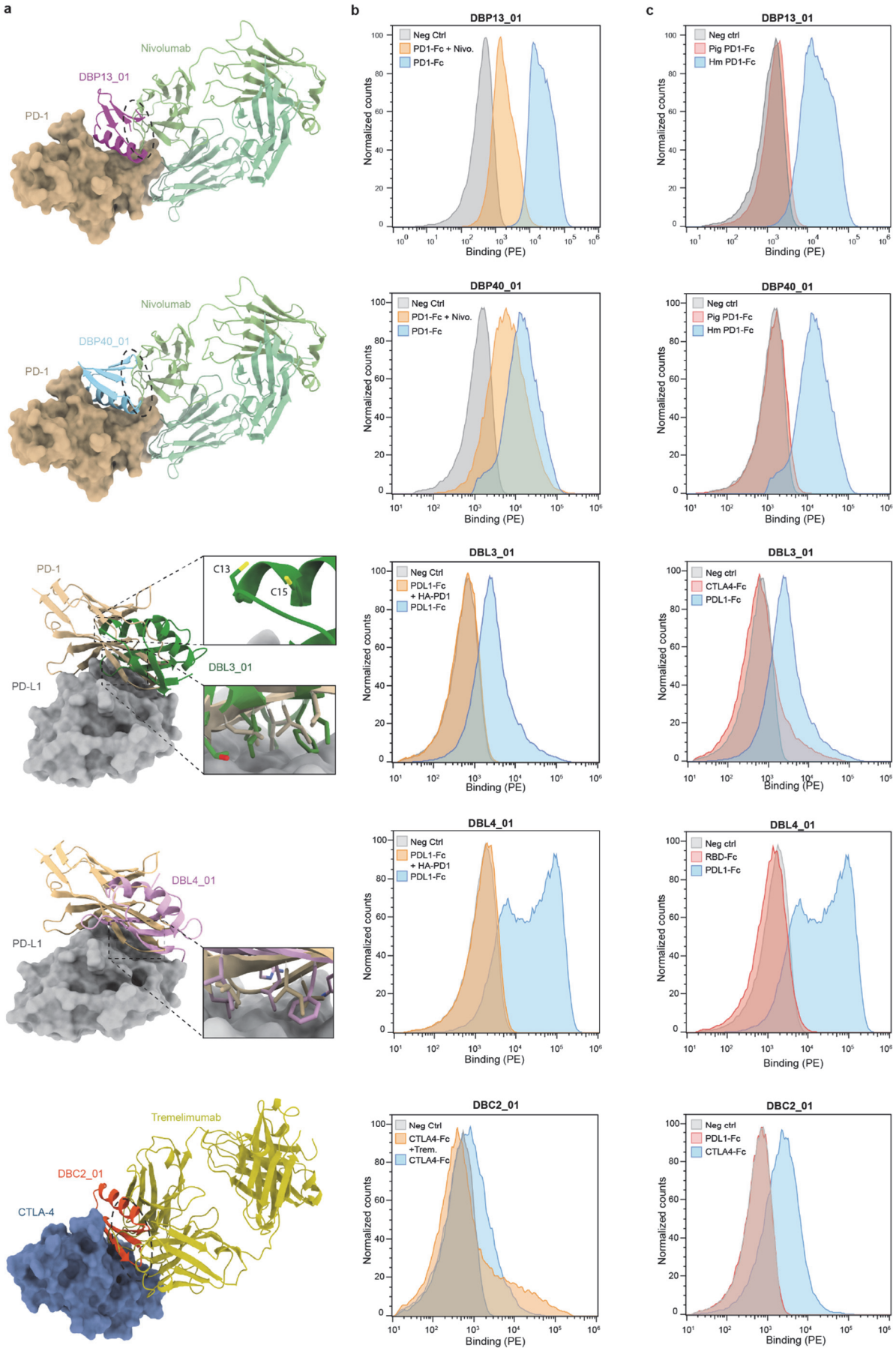

**Supplementary Figure S 3.25 : Overview of DBL2_03 SSM library**. A. Structural representation of all positions sampled in the SSM library (light blue). The four hotspot residues (red) were also sampled. Three positions were mutated in DBL2_04 (dark blue). Position 35 was not mutated in DBL_04, because all mutations in this position led to the inability of the soluble expression of the protein. **B.** Outcome of the entire SSM library. Blue indicates enrichment in the binding population, while red shows enrichment in the non-binding population. **C.** Binding affinities measured by SPR for the different versions of DBL2.



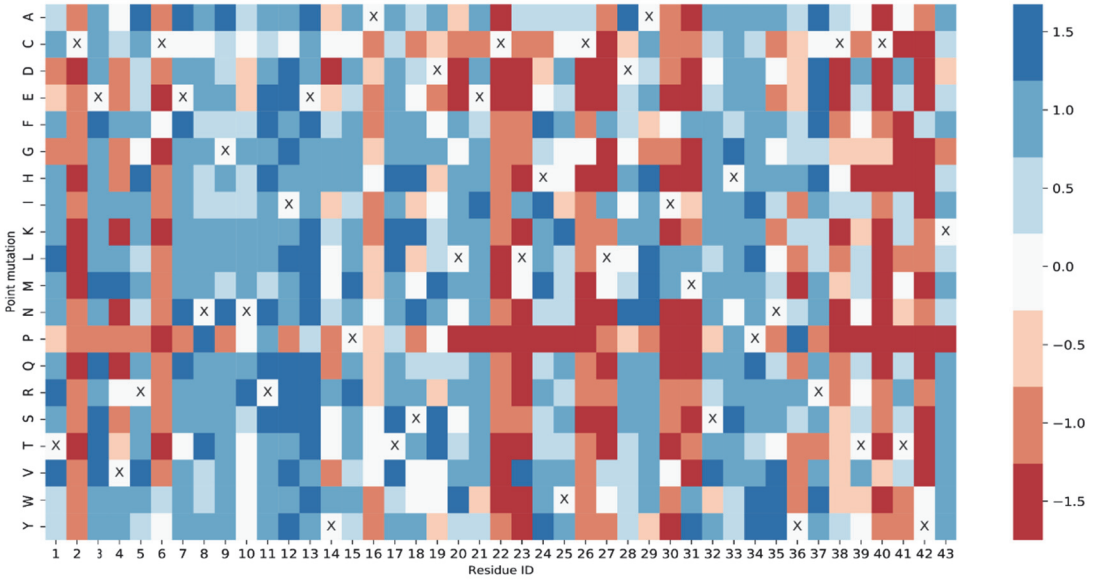

**Supplementary Figure S 3.26 : Electron density map of the crystalized DBL1_03 and DBL2_02. A.** Crystal structure of DBL1_03 (green) in complex with PD-L1 (gray). Refined 2mFo-mFc electron density map of the binder, contoured at 1.0$\sigma$ , is rendered as a white surface. **B.** Crystal structure of DBL2_02 (blue) in complex with PD-L1 (gray). Refined 2mFo-mFc electron density map of the binder, contoured at 1.0$\sigma$ , is rendered as a white surface



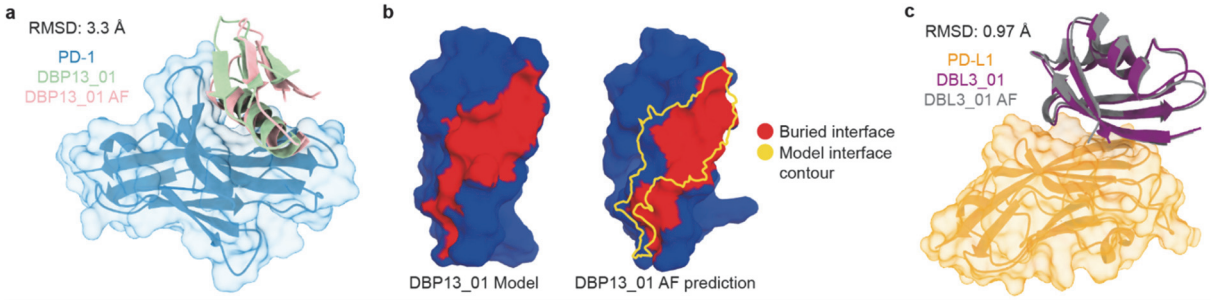

**Supplementary Figure S 3.27 : Overview and comparison between PD-1 binders. A.** PD-1 surface (blue) with the region targeted by PD-L1 (red) and the overlapping region targeted by DBP13_01 (yellow contour). **B.** Histograms of the binding signal (PE) measured on 3 yeast clones displaying designed binders against PD-1. Yeast cells were labeled with 500 nM PD-1-Fc (coloured) or secondary antibodies only (gray, negative control). **C-D.** Overview and close-up of DBP40_01 (a, pink) and DBP52_01 (b, blue) models in complex with PD-1 (gray). Interface seed residues similar to DBP13_01 are highlighted in red, while residues that are different are highlighted in orange. **E.** Seeds used to design DBP13_01 (green), DBP40_01 (pink) and DBP52_01 (blue) aligned with interface residues numbered. **F.** Sequence logo of the seed interface residues for the three PD-1 binders as numbered in e.


a
b
c
Nivolumab
DBP13_01
PD-1
DBP13_01
Neg Ctrl
PD1-Fc + Nivo.
PD1-Fc
Neg ctrl
Pig PD1-Fc
Hm PD1-Fc
Normalized counts
Binding (PE)
Nivolumab
DBP40_01
PD-1
DBP40_01
PD-1
C13
C15
DBL3_01
PD-L1
DBL3_01
Neg Ctrl
PDL1-Fc
+ HA-PD1
PDL1-Fc
Neg ctrl
CTLA4-Fc
PDL1-Fc
DBL4_01
PD-L1
DBL4_01
Neg Ctrl
PDL1-Fc
+ HA-PD1
PDL1-Fc
Neg ctrl
RBD-Fc
PDL1-Fc
Tremelimumab
DBC2_01
CTLA-4
DBC2_01
Neg Ctrl
CTLA4-Fc
+Trem.
CTLA4-Fc
Neg ctrl
PDL1-Fc
CTLA4-Fc

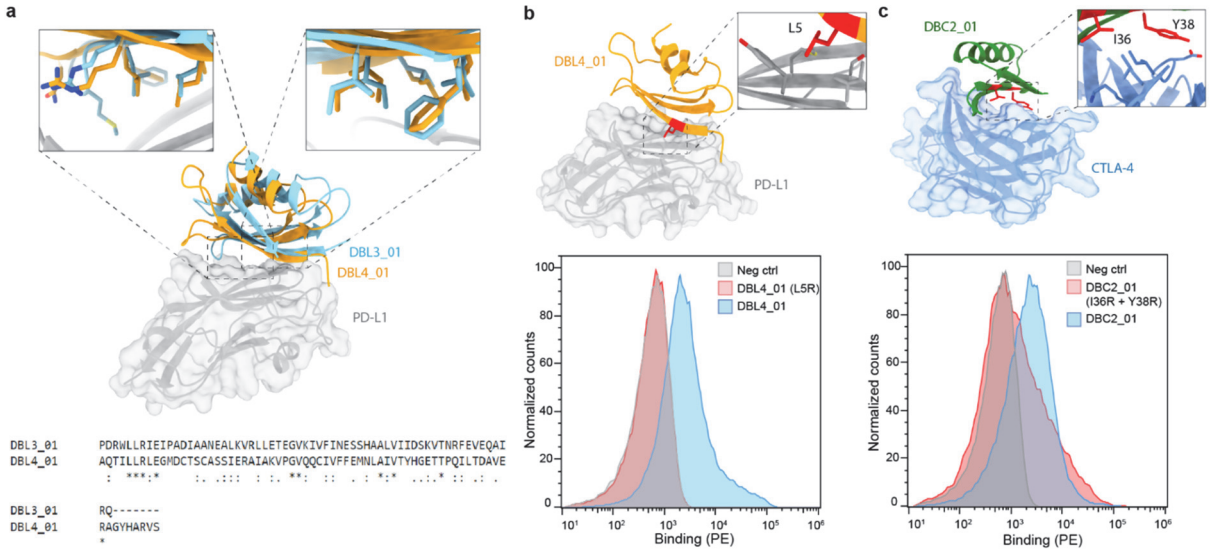
**Supplementary Figure S 3.28 : Competition and specificity binding assay of the different optimized binders on the surface of yeast. A.** Competition between designed binders and a known protein binder (native binder or monoclonal Fab) in complex with the target structure. **B.** Flow cytometry histograms showing fluorescence signals on the surface of yeast displaying the different binders. Yeast were labeled with 500 nM or their respective ligand (blue), 500 nM of blocked ligand pre-incubated with 10-fold molar excess of Fab or high-affinity PD-1 (HA-PD-1) (orange) or labeled with secondary antibodies only (gray, Neg Ctrl). **C.** Flow cytometry histograms showing fluorescence signal on the surface of yeast displaying the different binders and labeled with 500 nM of unrelated protein ligand (red) or labeled with secondary antibodies only (gray, Neg Ctrl).



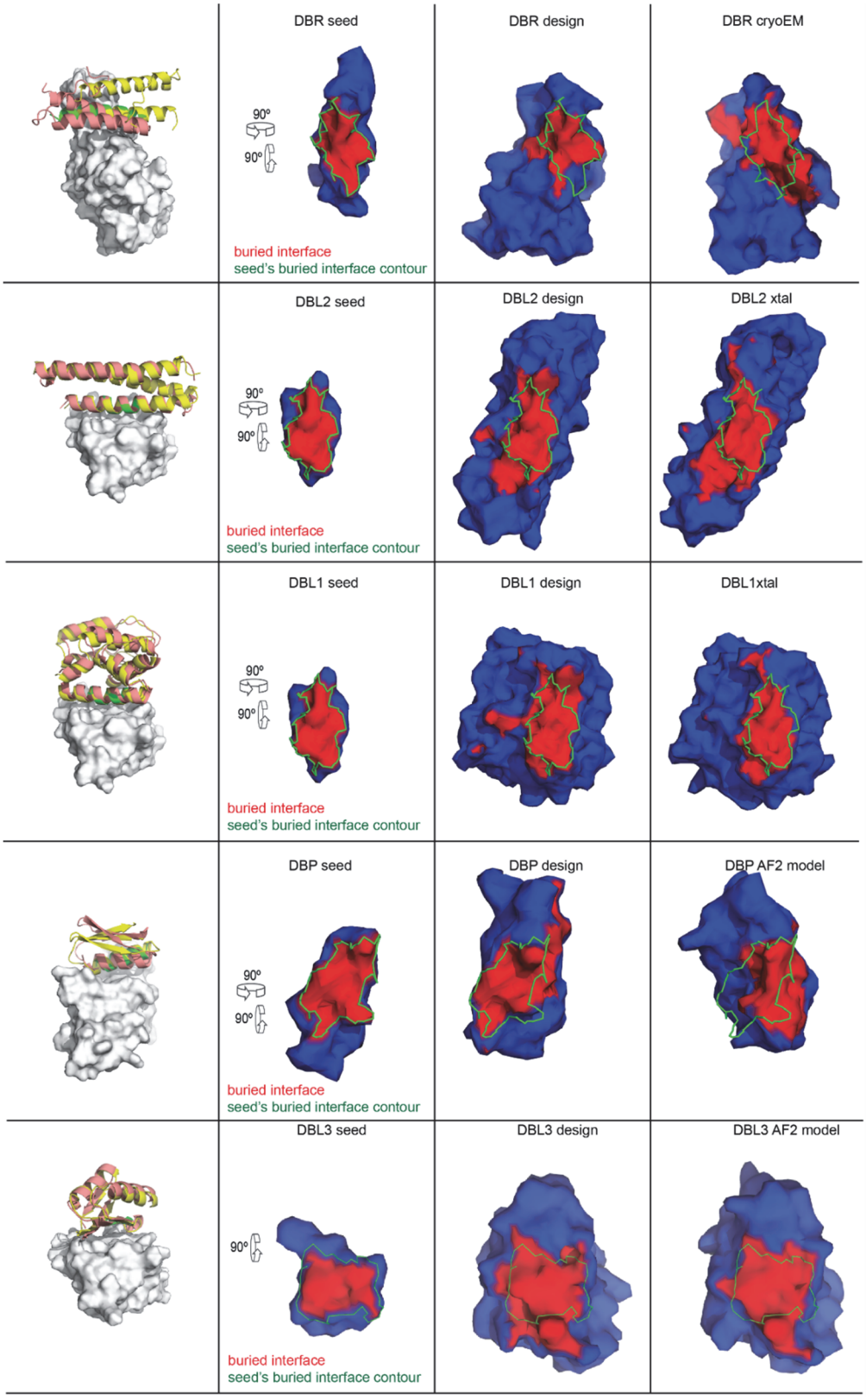

**Supplementary Figure S 3.29 : SSM of DBP13_01. Heatmap covering all positions of DBP13_01**. Yeast displaying point mutants were analyzed by flow cytometry and subsequently binding and non- binding populations were sorted. For each mutation the log-ratio between the enrichment in binding versus non-binding populations was computed. Mutations in red highlight a deleterious effect on binding, while mutations in blue indicate an enrichment on the binding population.



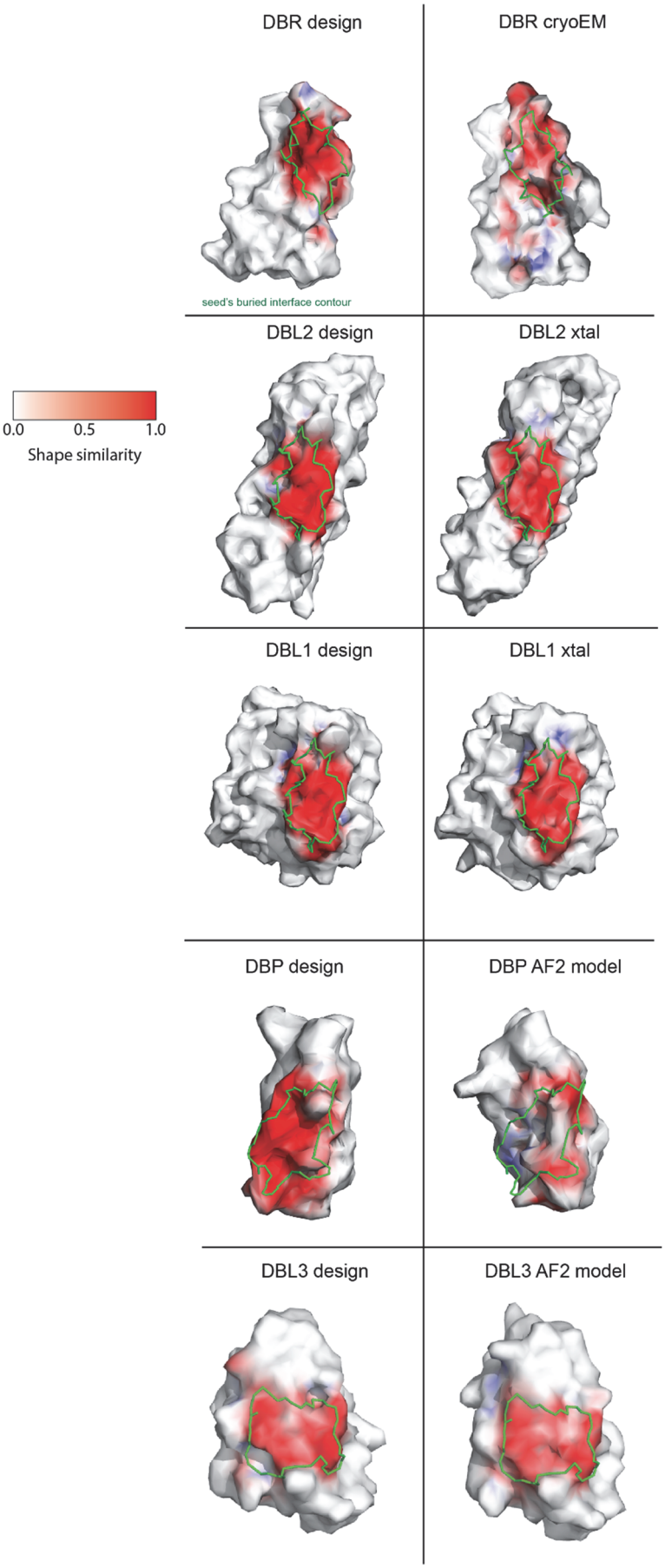

**Supplementary Figure S 3.30 : AF structure prediction of DBP13_01 in complex with PD-1. A.** Comparison of the DBP13_01 computational model (green) and the AlphaFold multimer (AF) prediction (red) on the surface of PD-1 (blue). **B.** Buried interfaces in both DBP13_01 model (left) and AF prediction (right) are shown in red with an overlap yellow, a yellow contour of the footprint of the original model is shown for ease of comparison. **C.** Comparison of the DBL3_01 computational model (purple) and the AlphaFold Multimer (AF) prediction (gray) on the surface of PD-L1 (orange).




**Supplementary Figure S 3.31 : DBL3_01 and DBL4_01 comparison and DBL4_01 and DBC2_01 knock-out mutants. A.** Superposition between DBL3_01 (cyan) and DBL4_01 (orange) in complex with PD-L1 (gray). Multiple sequence alignment of the two designs is shown at the bottom. **B.** DBL4_01 (orange) in complex with PD-L1 (gray) with knock-out mutant highlighted in red. Flow cytometry histograms showing fluorescence signals on the surface of yeast displaying DBL4_01 or the knock-out mutant, compared to unlabeled yeast (Neg Ctrl). **C.** DBC2_01 (green) in complex with CTLA-4 (blue) with two knock-out mutants highlighted in red. Flow cytometry histograms showing fluorescence signals on the surface of yeast displaying DBC2_01 or the knock-out mutants, compared to unlabeled yeast (Neg Ctrl).




**Supplementary Figure S 3.32 : Surface comparison between seeds, designs and final/predicted structures**. Buried interfaces of models/structures when in complex with their target are colored in red, while non-buried regions colored in blue. The contour of the buried interface of the initial binding seed is drawn in green and is shown for the initial seed, for the designs and for the final/predicted structures.


DBR design
DBR cryoEM
seed's buried interface contour
DBL2 design
DBL2 xtal
0.0
0.5
1.0
Shape similarity
DBL1 design
DBL1 xtal
DBP design
DBP AF2 model
DBL3 design
DBL3 AF2 model

**Supplementary Figure S 3.33 : Surface similarity of the computational designs, experimentally solved structures or AF models relative to initial binding seeds**. Each complex was aligned to the target protein (RBD, PD-L1 and PD-1), and the surface similarity of the computational design, the experimental structure or AF model to the binding seed is shown in a gradient from white to red. The buried surface area of the initial binding seed is shown by a green contour. The surface similarity was calculated in the same way as shape complementarity but normal vectors are not inverted during the process, i.e. the normal vectors for both surfaces point outwards of the molecular surface. Briefly, pairs of nearest vertices between the surface of the design or structure/model and the initial binding seed were computed based on the nearest neighbor of the aligned model. The shape similarity was evaluated by computing the dot product of the vertex pairs normal vectors yielding the enclosed angle and scaling it with the distance of the vertex pair. The resulting values are colored in a gradient from white to red and range from 0 (colored in white) indicating no similarity, to 1 (colored in red) indicating high similarity.

**Supplementary Table S 3.1 : Extended Benchmark of MaSIF-seed against other docking methods.** Recovering the native binder in the correct conformation from co-crystal structures for 31 helix-receptor complexes or 83 non-helix seed-receptor complexes, discriminating between 1000 decoys. [a]Benchmarked method. [b-d]Number of receptors for which the method recovered the native binding motif (<3 Å iRMSD) within the [b]top 1, [c]top 10, and [d]top 100 results. [e]Number of receptors for which the method did not recover the native binding motif in the top 100 results. [f]Average running time in minutes, excluding pre-computation time.

| | **Method[a]** | **# in top 1[b]** | **# in top 10[c]** | **# in top 100[d]** | **>100[e]** | **Avg time (m)[f]** |
|---|---|---|---|---|---|---|
| **Helical seeds** | MaSIF-seed | 18 | 18 | 20 | 11 | 15 |
| | PatchDock +MaSIF-site | 3 | 5 | 11 | 20 | 86 |
| | ZDOCK | 3 | 4 | 8 | 23 | 2715 |
| | ZDOCK+ MaSIF-site | 1 | 6 | 10 | 21 | 2485 |
| | ZDOCK+ ZRank2 | 6 | 12 | 21 | 10 | 2946 |
| | ZDOCK+ ZRank2+ MaSIF-site | 5 | 11 | 19 | 12 | 2710 |
| **Non-helical seeds** | MaSIF-seed | 41 | 47 | 49 | 34 | 118 |
| | ZDock | 7 | 9 | 22 | 61 | 2206 |
| | ZDock+ ZRank2 | 21 | 33 | 45 | 38 | 2400 |

**Supplementary Table S 3.2 : Sequences of the designed proteins.**

| Design | Sequence | # of mutations from WT | # of mutations from design_01 | Mutations |
|---|---|---|---|---|
| DBL1 native scaffold (PDB ID: 3S0D) | MTIEELKTRLHTEQSVCKTETGI DQQKANDVIEGNIDVEDKKVQL YCECILKNFNILDKNNVFKPQGI KAVMELLIDENSVKQLVSDCSTIS EENPHLKASKLVQCVSKYKTMK SVDFL | | | |
| DBL1_01 | TSIESLKWTLIVEQILCQLDTGID QQKANDVIEGNIDVEDKKVQLY CECILKAFHILDKNNVFKPQGIK AVMELLIDENSVKQLVSDCSTISE ENPHLKASKLVQCVSKYKTMKS VDFL | 14 | | M1T, T2S, E5S, T8W, R9T, H11I, T12V, S15I, V16L, K18Q, T19L, E20D, N53A, N55H |
| DBL1_02 | SSIESLKWSLIVQQILCQLETGID QQKANDVIEGNIDVEDKKVQLY CECILKQFHILDKNNVFKPQGIK AVMELLIDENSVKQLVSDCSTISE ENPHLKASKLVQCVSKYKTMKS VDFL | 14 | 5 | T1S, T9S, E13Q, D20E, A53Q |
| DBL1_03 | SSIESMKWSMIVQQILCQLETGI DQQKANDVIEGNIDVEDKKVQL YCECILKQFHILDKNNVFKPQGI KAVMELLIDENSVKQLVSDCSTIS EENPHLKASKLMQCISKYKTWK SFDFL | 20 | 11 | L6M, L10M, V104M, V107I, M113W, V116F |
| DBL1_04 | SSIESMKWSMIRQQILCQLETGI DQQKANDVIEGNIDVEDKKVQL YCECILKQFHILDKNNVFKPQGI KAVMELLIDENSVKQLVSDCSTIS EENPHLKASKLMQCISKYKTWK SFDFL | 21 | 14 | E4T, W8N, Q18R |
| DBL2 native scaffold (PDB ID: 3ONJ) | SLLISYESDFKTTLEQAKASLAEA PSQPLSQRNTTLKHVEQQQDEL FDLLDQMDVEVNNSIGDASERA TYKAKLREWKKTIQSDIKRPLQS LVDSGD | | | |
| DBL2_01 | SLLESYEWSFIVQLILAKLELAYA PSQPLSQRNEQLKRVEQQQDQL FDLLDQMDVEVNNSIGDASERA TYKAKLREWKKTIQSDIKRPLQS LVDSGD | 15 | | I4E, S8W, D9S, K11I, T12V, T13Q, E15I, Q16L, A19L, S20E, E23Y, T34E, T35Q, H38R, E45Q |
| DBL2_02 | NLLESYEWSFKVQLILAKLELAK APSQPLSQRNEELKRVEQRQDR LFDLLDQMDVEVNNSIGDASER ATYKAKLREWKKTIQSDIKRPLQ SLVDSGD | 16 | 6 | S1N, I11K, Y23K, Q35E, Q42R, Q45R |

| | | | | |
|---|---|---|---|---|
| DBL2_03 | NLLTSYEGSFKIQLILAKLELAKA PSQPLSQRNEELKRVEQRQDRLF DLLDQMDVEVNNSIGDASERAT YKAKLREWKKTIQSDIKRPLQSL VDSGD | 16 | 9 | E4T, W8G, V12I |
| DBL2_04 | NLLRSYENSFKIQLILAKLELAHA PSQPLSQRNEELKRVEQRQDRLF DLLDQMDVEVNNSIGDASERAT YKAKLREWKKTIQSDIKRPLQSL VDSGD | 16 | 9 | T4R, G8N, K23H |
| DBR_01 | STNMLEALQQRLHKYAAVVSRA ALENNSGKARRFGRIVKQYEDAI KLYKAGKPVPYDELPVPPGFG | 8 | | E13H, Q16A, S17A, E19V, A20S, A21R, K23A, A24L |
| DBR_02 | STNMLEALQQRLQFYFGVVSRA ALENNSGKARRFGRIVKQYEDAI KLYKAGKPVPYDELPVPPGFG | 9 | 4 | H13Q, K14F, A16F, A17G |
| DBR_03 | STNMLEALQQRLQFYHGQVARA ALENNSGKARRFGRIVKQYEDAI KLYKAGKPVPYDELPVPPGFG | 9 | 6 | F16H, V18Q, S20A |
| DBR_03_KO | STNMLEALQQRLQFYHRQVRRA ALENNSGKARRFGRIVKQYEDAI KLYKAGKPVPYDELPVPPGFG | 9 | | G17R, A20R |
| DBP13_01 | TCEVRCENGNRIEYPATSDLECL HWCLDAIMSHPNYRCTCTHK | 10 | | Q10N, E20L, E23L, R24H, R27L, K28D, K30I, K31M, E32S, F33H |
| DBP13_01 (native scaffold) | TCEVRCENGQRIEYPATSDEECE RWCRKAKKEFPNYRCTCTHK | | | |
| DBP40_01 | SQVTWNGVTVTNDNPSQSAM WADLIALLYQGEVRVKDGRWEI H | 12 | | I1S, F12N, E16S, E17Q, A18S, E19A, K20M, Y21W, K23D, K24L, K27L, E28L |
| DBP40_01 (native scaffold) | IQVTWNGVTVTFDNPEEAEKYA KKIAKEYQGEVRVKDGRWEIH | | | |
| DBP52_01 | QKETRHCSGRSCDWWATLWCL LCAMKGKRVRCRQHGQQVEVQ CDK | 13 | | Q10R, R11S, E13D, Q14W, E15W, R17T, R18L, E21L, E22L, K24A, K25M, K34Q, N37Q |
| DBP52_01 (native scaffold) | QKETRHCSGQRCEQEARRWCE ECKKKGKRVRCRKHGNQVEVQ CDK | | | |

| | | | | |
|---|---|---|---|---|
| DBL3_01 | AQTILLRLEGMDCTSCASSIERAI AKVPGVQQCIVFFEMNLAIVTY HGETTPQILTDAVERAGYHARVS | 11 | | N5L, Q7R, S32Q, Q34I, N36F, A38E, L39M, E40N, Q41L, V43I, S45T |
| DBL3_02 | AQTILLRLEGMDSTSSASSIERAI AKVPGVQQCIVFFEMNLAIVTY HGETTPQILTDAVERAGYHARVS | 13 | 2 | N5L, Q7R, C13S, C16S, S32Q, Q34I, N36F, A38E, L39M, E40N, Q41L, V43I, S45T |
| DBL3_01 (Native scaffold) | AQTINLQLEGMDCTSCASSIERA IAKVPGVQSCQVNFALEQAVVSY HGETTPQILTDAVERAGYHARVL | | | |
| DBL4_01 | PDRWLLRIEIPADIAANEALKVRL LETEGVKIVFINESSHAALVIIDSK VTNRFEVEQAIRQ | 12 | | Y2D, V3R, S4W, S5L, E32I, L34F, A36N, E38S, E39S, S41A, Y43L, K45I |
| DBL4_01 (Native scaffold) | PYVSSLRIEIPADIAANEALKVRLL ETEGVKEVLIAEEEHSAYVKIDSK VTNRFEVEQAIRQA | | | |
| DBC2_01 | AFITIMDGEEKARKYAKMLKKQ NLKVIVLMANGKWIIYAK | 11 | | K1A, T3I, T5I, E25K, H27I, R29L, V30M, E31A, V36I, T38Y, E40K |
| DBC2_01 (Native scaffold) | KFTTTMDGEEKARKYAKMLKK QNLEVHVRVENGKWVITAE | | | |

**Supplementary Table S 3.3 : SARS-CoV-2 variant mutations.**

| Variant (graph label in bold) | Mutations | $EC_{50}$ of DBR3_03 with variant: |
|---|---|---|
| D614G / **WT** | D614G | 6.61e-8 g/mL |
| B.1.1.7 / **Alpha** | Δ69-70, Δ144, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H | 6.56e-8 g/mL |
| B.1.351 / **Beta** | L18F, D80A, D215G, Δ242-244, R246I, K417N, E484K, N501Y, D614G, A701V | 6.11e-7 g/mL |
| B.11.28.1 / **Gamma** | L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I, V1176F | 6.76e-7 g/mL |
| B.1.526 / **Iota** | L5F, T95I, D253G, E484K, D614G, A701V | 4.13e-7 g/mL |
| B.1.617.1 / **Kappa** | E154K, L452R, E484Q, D614G, P681R, Q1071H | NA |
| B.1.617.2 / **Delta** | T19R, Δ156-157, R158G, L452R, T478K, D614G, P681R, D950N | NA |
| **Lambda** | G75V, T76I, R246N, Δ247-253, L452Q, F490S, D614G, T859N | NA |
| **Omicron BA.1** | A67V, Δ69-70, T95I, G142D, Δ143-145, Δ211, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493K, G496S, Q498R, N501Y, Y505H, T547K,D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F | 5.68e-8 g/mL |
| **Omicron BA.2** | T19I, Δ24-26, A27S, G142D, V213G G339D, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, S477N, T478K, E484A, Q493R, Q498R, N501Y, Y505H, D614G, H655Y, N679K, P681H, N764K, D796Y, Q954H, N969K | 4.35e-8 g/mL |

**Supplementary Table S 3.4 : Summary of binding candidates obtained after deep sequencing with the optimized design pipeline**. Binding seeds were helical (H) or strand (E). Deep sequencing data comprises reads from the non-binding (Neg reads) and binding population (Pos reads). The enrichment score is calculated based on the logarithm of the ratio between positive and negative reads. Computational models of the complexes were predicted by AlphaFold Multimer (AF) and aligned with respect to the target. Binding signals detected on the surface of yeast (at 500 nM ligand) were categorized as negative(-), marginal (-/+), weak (+), moderate (++) or high (+++). Marginal and weak binding signals were not further characterized (Competition assay, knock-out mutants and negative controls).

| Design | Target | Scaffold origin | Scaffold name | Motif | Neg reads | Pos reads | Enrichment | AF RMSD [Å] | Binding | Competition | Knock-out | Neg Ctrl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DBP13_01 | PD-1 | Miniprotein | EEHE_2.1_02 | H | 776 | 277011 | 2,5526 | 3,3 | +++ | OK | OK | OK |
| DBP40_01 | PD-1 | Miniprotein | EEHEE_rd4_0499 | H | 15 | 18596 | 3,0933 | 7,1 | +++ | OK | Not tested | OK |
| DBP48_01 | PD-1 | Miniprotein | EHEE_rd4_0510 | H | 225 | 2354 | 1,0196 | 16,9 | - | N/A | N/A | N/A |
| DBP52_01 | PD-1 | Miniprotein | EHEE_1.7_09 | H | 12 | 934 | 1,8912 | 15,5 | + | Not tested | Not tested | Not tested |
| DBC1_01 | CTLA-4 | Miniprotein | EEHE_2.1_06 | E | 3 | 1107 | 2,567 | 6,2 | - | N/A | N/A | N/A |
| DBC2_01 | CTLA-4 | Miniprotein | EHEE_rd4_0923 | E | 172 | 39697 | 2,3632 | 34,6 | ++ | OK | OK | OK |
| DBC3_01 | CTLA-4 | Miniprotein | EHEE_rd4_0042 | E | 31 | 3797 | 2,0881 | 30,4 | -/+ | N/A | N/A | N/A |
| DBC4_01 | CTLA-4 | Miniprotein | EHEE_rd4_0924 | E | 55 | 4018 | 1,8636 | 34,9 | -/+ | N/A | N/A | N/A |
| DBC5_01 | CTLA-4 | Miniprotein | EHEE_rd4_0448 | E | 35 | 1203 | 1,5362 | 24,3 | - | N/A | N/A | N/A |
| DBC6_01 | CTLA-4 | Miniprotein | EHEE_rd4_0357 | E | 109 | 1281 | 1,0701 | 22,4 | - | N/A | N/A | N/A |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DBC7_01 | CTLA-4 | Miniprotein | EHEE_rd4_0924 | E | 114 | 1333 | 1,0679 | 38,4 | - | N/A | N/A | N/A |
| DBC8_01 | CTLA-4 | Miniprotein | EHEE_rd4_0636 | E | 3 | 2162 | 2,8577 | 33,6 | - | N/A | N/A | N/A |
| DBC9_01 | CTLA-4 | Miniprotein | EHEE_rd4_0811 | E | 44 | 1166 | 1,4232 | 32,2 | - | N/A | N/A | N/A |
| DBL3_01 | PD-L1 | PDB | 4A48 (B) | E | 654 | 44391 | 1,8317 | 1,2 | ++ | OK | OK | OK |
| DBL4_01 | PD-L1 | PDB | 4Q2M (A) | E | 306 | 10238 | 1,5245 | 9,1 | ++ | OK | OK | OK |
| DBL5_01 | PD-L1 | Miniprotein | EHEE_rd4_0017 | E | 340 | 5443 | 1,2044 | 15,6 | - | N/A | N/A | N/A |

**Supplementary Table S 3.5 : Antibodies used in flow cytometry experiments**.

| **Antibody** | **Catalog number** | **Supplier** | **Dilution** |
|---|---|---|---|
| Anti-HA, FITC | A190-138F | Bethyl | 1:100 |
| Anti-V5 mouse | MA5-15253 | Invitrogen | 1:333 |
| Anti-mouse, FITC | F0257 | Sigma | 1:100 |
| Anti-His, PE | 130-120-787 | Miltenyi Biotec | 1:50 |
| Anti-Myc, FITC | SAB4700448 | Sigma | 1:100 |
| Anti-human IgG, PE | 12-4998-82 | Invitrogen | 1:100 |
| Anti-Human IgG, R-PE | 109-117-008 | Jackson ImmunoResearch | 1:100 |

**Supplementary Table S 3.6 : Primer sequences**.

| Library: | Primer name: | Primer sequence: |
|---|---|---|
| DBR_01 | 5vny_rev_1 | CAGACGTTGCTGTAGGGCCTCAAGCATGTTCGTGCTGCTAGCAGCGTA GTCTGGAACG |
| DBR_01 | 5vny_fwd_2a | GAGGCCCTACAGCAACGTCTGCWMARATACKYCRBRGTABNARSCNNS GCGGSACTTGAGAATAATAGTGGAAAAGCAAGAAGATTTGGCAGGATC |
| DBR_01 | 5vny_fwd_2b | GAGGCCCTACAGCAACGTCTGCWMYWCTACKYCRBRGTABNARSCNN SGCGGSACTTGAGAATAATAGTGGAAAAGCAAGAAGATTTGGCAGGAT C |
| DBR_01 | 5vny_rev_3 | ACAGGTTTTCCAGCTTTATACAACTTAATTGCGTCCTCGTATTGTTTAAC GATCCTGCCAAATCTTCTTGCTTT |
| DBR_01 | 5vny_fwd_4 | ATTAAGTTGTATAAAGCTGGAAAACCTGTACCATACGACGAACTACCTG TCCCGCCAGGATTCGGCGGATCCCAGGAACTGACAACTATATG |
| DBL1_L1 | 3S0D_fw1 | GCCTTAGCTCAACCGGTTATTTCTACTACCGTCGGTTCCGCTGCAGAA GGCTCTTTGGACAAGAG |
| DBL1_L1 | 3S0D_rev1 | GCTAGCAGCGTAGTCTGGAACGTCGTATGGGTAAGCTTCTCTCTTGTC CAAAGAGCCTTCT |
| DBL1_L1 | 3S0D_fw3a | CCAGACTACGCTGCTAGCHHCTMCATTSWAAGTTTGAAGTGGAVCTTA WTCRTASAACAAATTVTATGTCAACTTBWCACGGGGATTGACCAGCA |
| DBL1_L1 | 3S0D_ fw3b | CCAGACTACGCTGCTAGCHHCTMCATTSWAAGTTTGAAGTGGAVCTTA WTCRTASAACAAATTVTATGTCAACTTGAAACGGGGATTGACCAGCA |
| DBL1_L1 | 3S0D_ fw3c | CCAGACTACGCTGCTAGCHHCTMCATTSWAAGTTTGAAGTGGAVCTTA WTCRTATGGCAAATTVTATGTCAACTTBWCACGGGGATTGACCAGCA |
| DBL1_L1 | 3S0D_ fw3d | CCAGACTACGCTGCTAGCHHCTMCATTSWAAGTTTGAAGTGGAVCTTA WTCRTATGGCAAATTVTATGTCAACTTGAAACGGGGATTGACCAGCA |
| DBL1_L1 | 3S0D_ fw3e | CAGACTACGCTGCTAGCHHCTMCATTSWAAGTTTGAAGTGGAVCTTAW TCRTAWWCCAAATTVTATGTCAACTTBWCACGGGGATTGACCAGCA |
| DBL1_L1 | 3S0D_ fw3f | CCAGACTACGCTGCTAGCHHCTMCATTSWAAGTTTGAAGTGGAVCTTA WTCRTAWWCCAAATTVTATGTCAACTTGAAACGGGGATTGACCAGCA |
| DBL1_L1 | 3S0D_rev4 | CATTCGCAATATAGTTGGACTTTTTTGTCCTCCACGTCAATGTTCCCCT CAATCACGTCATTCGCCTTCTGCTGGTCAATCCCCGT |
| DBL1_L1 | 3S0D_fw5a | GTCCAACTATATTGCGAATGTATACTAAAASMATTCTGGATACTTGATAR AAATAATGTTTTTAAGCCCCAGGGAATTAAAGC |
| DBL1_L1 | 3S0D_fw5b | GTCCAACTATATTGCGAATGTATACTAAAASMATTCYWCATACTTGATAR AAATAATGTTTTTAAGCCCCAGGGAATTAAAGC |
| DBL1_L1 | 3S0D_rev6 | GATATAGTGCTACAGTCGGAGACAAGCTGTTTAACGCTATTTTCATCTAT TAACAGTTCCATCACAGCTTTAATTCCCTGGGGCTT |
| DBL1_L1 | 3S0D_fw7 | CTCCGACTGTAGCACTATATCAGAAGAGAACCCACATCTTAAGGCCAGT AAACTGGTTCAGTGCGTGAGTAAATACAAAACCATGAAAAGCGTGG |
| DBL1_L1 | 3S0D_rev8 | GAGTACGGCGTCGATTCTAAAGTTGGTGAGGGGATTTGCTCGCATATA GTTGTCAGTTCCTGGGATCCCAAGAAGTCCACGCTTTTCATGGTTTTG |

| | | |
|---|---|---|
| DBL1_L2 | 3S0D_core_fw 3a | CCAGACTACGCTGCTAGCTCCTCCATTGAAAGTVTGAAGTGGAGCMTG ATCGTACAACAAATTCTATGTCAACT |
| DBL1_L2 | 3S0D_core_fw 3b | CCAGACTACGCTGCTAGCTCCTCCATTGAAAGTVTGAAGTGGAGCTWC ATCGTACAACAAATTCTATGTCAACT |
| DBL1_L2 | 3S0D_core_re v4 | CGTCAATGTTCCCCTCAATCACGTCATTCGCCTTCTGCTGGTCAATCCC CGTTTCAAGTTGACATAGAATTTGTTGTACGAT |
| DBL1_L2 | 3S0D_core_fw 5 | TGATTGAGGGGAACATTGACGTGGAGGACAAAAAAGTCCAAC |
| DBL1_L2 | 3S0D_core_re v6a | GGCTTAAAAACATTATTTTTATCAAGTATGTGCCATTGTTTGWRCCAAC ATTCGCAATATAGTTGGACTT |
| DBL1_L2 | 3S0D_core_re v6b | GGCTTAAAAACATTATTTTTATCAAGTATGTGGWRTTGTTTGWRCCAAC ATTCGCAATATAGTTGGACTT |
| DBL1_L2 | 3S0D_core_re v6c | GGCTTAAAAACATTATTTTTATCAAGTATGTGCCATTGTTTGWRGWWAC ATTCGCAATATAGTTGGACTT |
| DBL1_L2 | 3S0D_core_re v6d | GGCTTAAAAACATTATTTTTATCAAGTATGTGGWRTTGTTTGWRGWWA CATTCGCAATATAGTTGGACTT |
| DBL1_L2 | 3S0D_core_fw 7a | CACATACTTGATAAAAATAATGTTTTTAAGCCCCAGGGAATTAAAGCTRT GATGGAACTGACTATAGATGAAAATAGCGTTAAACAGCTT |
| DBL1_L2 | 3S0D_core_fw 7b | CACATACTTGATAAAAATAATGTTTTTAAGCCCCAGGGAATTAAAGCTRT GATGGAACTGMTAATAGATGAAAATAGCGTTAAACAGCTT |
| DBL1_L2 | 3S0D_core_re v8 | CAGTTTACTGGCCTTAAGATGTGGGTTCTCTTCTGATATAGTGCTACAG TCGGAGACAAGCTGTTTAACGCTATTTTCATC |
| DBL1_L2 | 3S0D_core_fw 9a | CACATCTTAAGGCCAGTAAACTGRYGCAGTGCVTRTMCAAGTACAAGA CCTWCAAAAGCKKGGATTTCCTTGGATCCCAGGA |
| DBL1_L2 | 3S0D_core_fw 9b | CACATCTTAAGGCCAGTAAACTGRYGCAGTGCVTRTMCAAGTACAAGA CCTWCAAAAGCTWCGATTTCCTTGGATCCCAGGA |
| DBL1_L2 | 3S0D_core_fw 9c | CACATCTTAAGGCCAGTAAACTGRYGCAGTGCVTRTMCAAGTACAAGA CCWKGAAAAGCTWCGATTTCCTTGGATCCCAGGA |
| DBL1_L2 | 3S0D_core_re v10 | GAGTACGGCGTCGATTCTAAAGTTGGTGAGGGGATTTGCTCGCATATA GTTGTCAGTTCCTGGGATCCAAGGAAATC |
| | | |
| DBL2_L1 | 3ONJ_fw1 | GCCTTAGCTCAACCGGTTATTTCTACTACCGTCGGTTCCGCTGCAGAA GGCTCTTTGGACAAGAG |
| DBL2_L1 | 3ONJ_rev2 | GCTAGCAGCGTAGTCTGGAACGTCGTATGGGTAAGCTTCTCTCTTGTC CAAAGAGCCTTCTG |
| DBL2_L1 | 3ONJ_fw3a | CAGACTACGCTGCTAGCWMTCTTSDAGAGAGTTATGAATGGASCTTTR WAGTCCRATTGAWATTGGCTAAGTTGGAMCTGGCCMRGGCGCCATCA CAGCC |
| DBL2_L1 | 3ONJ_fw3b | CAGACTACGCTGCTAGCWMTCTTSDAGAGAGTTATGAATGGASCTTTR WAGTCCRATTGAWATTGGCTAAGTTGGAMCTGGCCTATGCGCCATCAC AGCC |

| | | |
|---|---|---|
| DBL2_L1 | 3ONJ_fw3c | CAGACTACGCTGCTAGCWMTCTTSDAGAGAGTTATGAATWTASCTTTR WAGTCCRATTGAWATTGGCTAAGTTGGAMCTGGCCMRGGCGCCATCA CAGCC |
| DBL2_L1 | 3ONJ_fw3d | CAGACTACGCTGCTAGCWMTCTTSDAGAGAGTTATGAATWTASCTTTR WAGTCCRATTGAWATTGGCTAAGTTGGAMCTGGCCTATGCGCCATCAC AGCC |
| DBL2_L1 | 3ONJ_rev4a | CATCTGGTCCAGTAAATCGAATAATYGATCTTGACGCTGTTCAACACGT TTAAGKTSCTCATTACGTTGAGACAAAGGCTGTGATGGCGC |
| DBL2_L1 | 3ONJ_rev4b | CATCTGGTCCAGTAAATCGAATAATYGATCTTGCTBCTGTTCAACACGT TTAAGKTSCTCATTACGTTGAGACAAAGGCTGTGATGGCGC |
| DBL2_L1 | 3ONJ_rev4c | CATCTGGTCCAGTAAATCGAATAATYGATCTTGACGCTGTTCAACCTBTT TAAGKTSCTCATTACGTTGAGACAAAGGCTGTGATGGCGC |
| DBL2_L1 | 3ONJ_rev4d | CATCTGGTCCAGTAAATCGAATAATYGATCTTGCTBCTGTTCAACCTBTT TAAGKTSCTCATTACGTTGAGACAAAGGCTGTGATGGCGC |
| DBL2_L1 | 3ONJ_fw5 | TTATTCGATTTACTGGACCAGATGGATGTGGAGGTTAATAACAGCATCG GGGACGCATCAGAACGCGCCACTTATAAAG |
| DBL2_L1 | 3ONJ_rev6 | GCTTGATGTCGGACTGGATCGTTTTTTTCCACTCGCGTAACTTTGCTTT ATAAGTGGCGCGTTCT |
| DBL2_L1 | 3ONJ_fw7 | CCAGTCCGACATCAAGCGCCCGCTTCAGAGTTTGGTTGATAGTGGCGA TGGATCCCAGGAACTGACAA |
| DBL2_L1 | 3ONJ_rev8 | GAGTACGGCGTCGATTCTAAAGTTGGTGAGGGGATTTGCTCGCATATA GTTGTCAGTTCCTGGGATCC |

**Supplementary Table S 3.7 : Target protein sequences**.

| Protein target: | Sequence: | Notes: |
|---|---|---|
| PD-1 (Uniprot #Q15116) | LDSPDRPWNPPTFSPALLVVTE GDNATFTCSFSDTSESFVLNWY RMSPSDQTDKLAAFPEDRSQPG QDSRFRVTQLPNGRDFHMSVV RARRNDSGTYLCGAISLAPKAQI KESLRAELRVTERRAEVPTAHPS PSPRPAGQFQ | Mutated glycosylation site (N->D) and mutated free cysteines (C->S) underlined |
| CTLA4 (Uniprot #P16410) | KAMHVAQPAVVLASSRGIASFV CEYASPGKATEVRVTVLRQADS QVTEVCAATYMMGNELTFLDD SICTGTSSGNQVNLTIQGLRAM DTGLYICKVELMYPPPYYLGIGD GTQIYVIDPEPCPDSD | Mutated glycosylation site underlined (N->D) |
| RBD WT (Uniprot #P0DTC2) | RVQPTESIVRFPNITNLCPFGEV FNATRFASVYAWNRKRISNCVA DYSVLYNSASFSTFKCYGVSPTK LNDLCFTNVYADSFVIRGDEVR QIAPGQTGKIADYNYKLPDDFT GCVIAWNSNNLDSKVGGNYNY LYRLFRKSNLKPFERDISTEIYQA GSTPCNGVEGFNCYFPLQSYGF QPTNGVGYQPYRVVVLSFELLH APATVCGPKKSTNLVKNKCVNF NFNGLTGTGVLTESNKKFLPFQ QFGRDIADTTDAVRDPQTLEIL DITPCS | |
| PD-L1 (Uniprot #Q9NZQ7) | SFTVTVPKDLYVVEYGSNMTIE CKFPVEKQLDLAALIVYWEMED KNIIQFVHGEEDLKVQHSSYRQ RARLLKDQLSLGNAALQITDVK LQDAGVYRCMISYGGADYKRIT VKVNAPYNKINQRILVVDPVTSE HELTCQAEGYPKAEVIWTSSDH QVLSGKTTTTNSKREEKLFNVT STLRINTTTNEIFYCTFRRLDPEE NHTAELVIPELPLAHPPNERTD | |

**Supplementary Table S 3.8 : Crystallographic data collection and refinement statistics**.

| | DBL1_03-PD-L1 | DBL2_02-PD-L1 |
|---|---|---|
| **Data collection** | | |
| Space group | P $4_2$ $2_1$ 2 | P $2_1$ $2_1$ $2_1$ |
| Cell dimensions | | |
| *a*, *b*, *c* (Å) | 97.93, 97.93, 106.11 | 85.41, 116.08, 149.61 |
| a, b, g(°) | 90.00, 90.00, 90.00 | 90.00, 90.00, 90.00 |
| Wavelength (Å) | 0.97889 | 0.97918 |
| Resolution (Å) | 48.97 - 2.85 (2.95 - 2.85) | 41.06 - 3.00 (3.11 - 3.00) |
| Unique reflections | 12591 (1241) | 30347 (2986) |
| $R_{merge}$ | 0.141 (3.126) | 0.165 (2.911) |
| $I$ / s$I$ | 20.7 (1.1) | 12.8 (1.1) |
| CC1/2 | 0.998 (0.554) | 0.999 (0.436) |
| Completeness (%) | 99.9 (100.0) | 99.4 (99.7) |
| Redundancy | 25.4 (26.9) | 13.0 (13.1) |
| | | |
| **Refinement** | | |
| Resolution (Å) | 48.97 - 2.85 | 41.06 - 3.00 |
| No. reflections | 12582 | 30282 |
| $R_{work}$ / $R_{free}$ | 0.3005/0.3220 | 0.2671/0.2945 |
| No. atoms | | |
| Protein | 2619 | 9316 |
| Ligand/ion | 0 | 0 |

| | | |
|---|---|---|
| Water | 1 | 0 |
| *B*-factors | | |
| Protein | 124.1 | 134.1 |
| Ligand/ion | - | - |
| Water | 83.5 | - |
| R.m.s. deviations | | |
| Bond lengths (Å) | 0.010 | 0.004 |
| Bond angles (°) | 1.250 | 0.700 |
| Ramachandran plot | | |
| Favored (%) | 93.77 | 96.07 |
| Allowed (%) | 6.23 | 3.93 |
| Outliers (%) | 0.00 | 0.00 |

*Values in parentheses are for highest-resolution shell.

**Supplementary Table S 3.9 : Cryo-EM data collection and model validation statistics**.

| Data collection and processing | D614G-binder | Omicron-binder full | Omicron-binder local |
|---|---|---|---|
| Microscope | TFS Titan Krios G4 + E-CFEG | TFS Titan Krios G4 + E-CFEG | |
| Detector | Falcon 4 | Falcon 4 | |
| Magnification (nominal) | 195K | 165K | |
| Pixel size (Å) | 0.40 | 0.726 | |
| Voltage (Kv) | 300 | 300 | |
| Electron exposure (e-/Å$^2$) | 80 | 60 | |
| Dose rate (e-/px/s) | 4.53 | 5.4 | |
| Exposure times (seconds) | 2.82 | 5.85 | |
| Defocus range (um) | 0.8-2.0 | 0.8-2.5 | |
| Micrographs | 20 794 | 22 266 | |
| Initial particle images (No.) | 832 816 | 1 820 333 | |
| Final particle images (No.) | 67 432 | 50 758 | |
| Map resolution (Å) | 2.63 | 2.80 | 3.29 |
| FSC threshold (cutoff) | 0.143 | 0.143 | 0.143 |
| Symmetry | C1 | C1 | C1 |
| Refinement | | | |
| Initial model used | 7BNO | 7QO7 | --- |

| | | | |
|---|---|---|---|
| Map sharpening B factor (Å²) | -33.7 | -33.8 | 52.7 |
| Model composition | | | |
| Non Hydrogen atoms<br>Protein residues / Nucleotide<br>Ligands | 25948<br>3236/0<br>NAG:46 | 28058<br>3429/0<br>BMA:12<br>NAG:76 | 2121<br>261/0<br>NAG:2 |
| B factors (Å²) | | | |
| protein<br>Ligand | 2.00/198.38/88.48<br>31.30/175.52/79.67 | 0.11/126.79/59.39<br>28.80/129.22/79.33 | 33.55/111.45/61.63<br>58.45/64.51/61.48 |
| R.m.s.d deviations | | | |
| Bond lengths (Å) | 0.004(0) | 0.002(3) | 0.003(0) |
| bond angles (°) | 0.687(35) | 0.534 (18) | 0.577(0) |
| Validation | | | |
| MolProbity score | 1.73 | 1.85 | 1.96 |
| Clash score | 6.36 | 9.53 | 9.82 |
| Poor rotamers (%) | 0.00 | 0.00 | 0.00 |
| Ramachandran plot | | | |
| Favored (%) | 94.36 | 95.05 | 93.00 |
| Allowed (%) | 5.45 | 4.77 | 7.00 |
| Disallowed (%) | 0.19 | 0.18 | 0.00 |
| PDB | 7ZSS | 7ZRV | 7ZSD |
| EMDB | 14947 | 14922 | 14930 |

## 3.7 Addendum

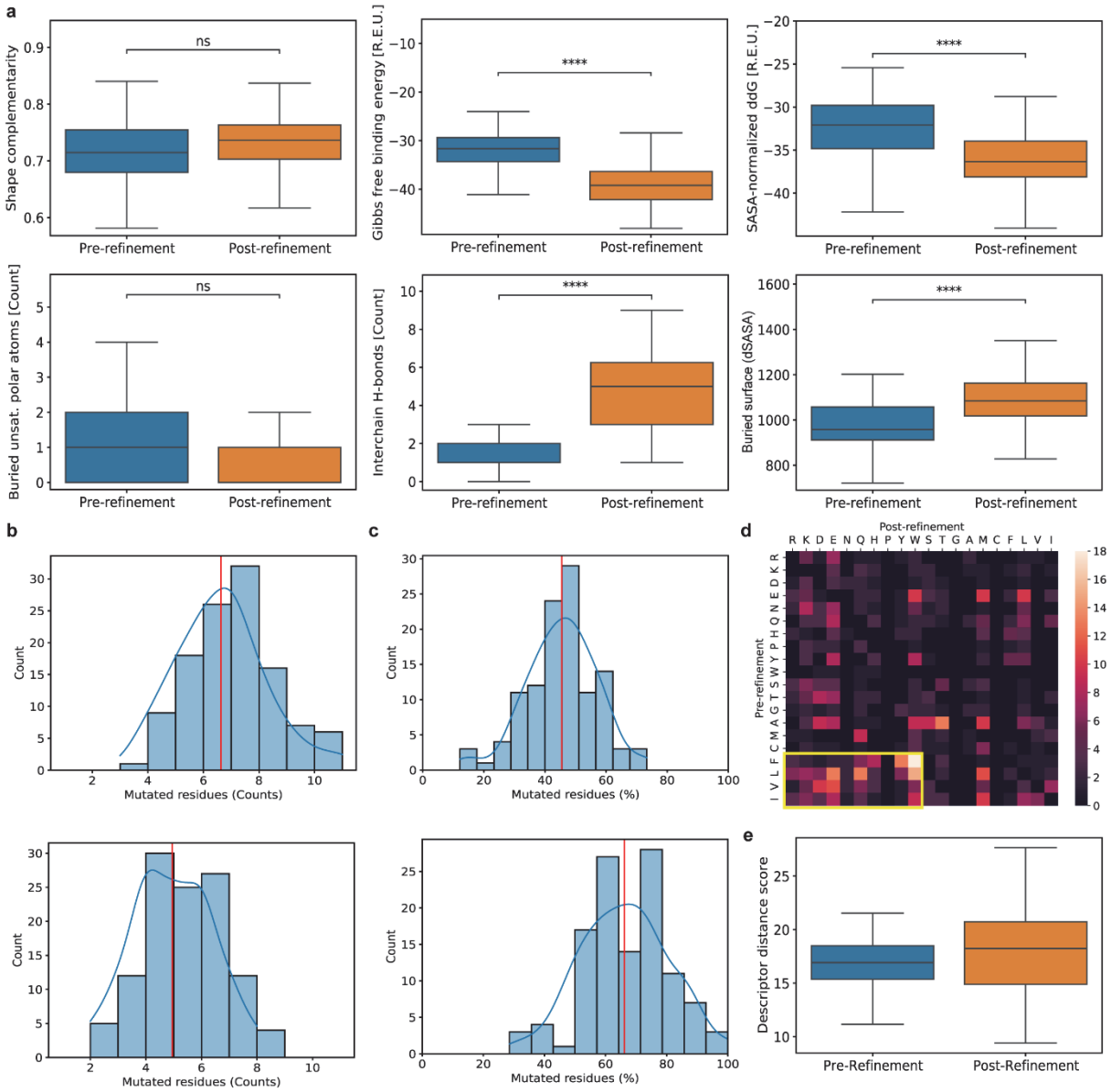
In this additional section, we will briefly discuss one concrete impact and one translational perspective brought by my contribution to the previous work, more precisely: i) the impact of the changes made in the optimized pipeline, and ii) a potential agonist effect of the designed PD-1 binder.

### 3.7.1 Optimized design pipeline and its consequences

In the work presented in this chapter, one key challenge was to obtain better designed binders directly by computational generation. The SARS-CoV-2 binders and PD-L1 binders, presented in section 3.3.1 and 3.3.2 respectively, required extensive *in vitro* optimization to achieve affinities in the range of native PPIs (micromolar and below). The main reasons highlighted by the site-saturation mutagenesis performed in this study were the presence of steric clashes, buried unsatisfied polar atoms and poor polar contacts on the rim of the interface. One of my contribution consisted of bringing better binding seeds by introducing a refinement step prior to seed grafting onto a scaffold protein. Despite its large size (402M surface fingerprints from 640K seed motifs), the MaSIF-seed database does not claim to offer a universal solution for all target patches. Taking into account the limitations that were encountered during the optimization of the first designed binders, I provided a refinement script using Rosetta [223] that combined both a structure relaxation and sequence design protocol with a modified scoring function that introduced a penalty for buried unsatisfied polar atoms [137]. By penalizing these deleterious interactions, the polar interactions on the other hand would be more rewarded and the computed binding energy would be more favorable.

As exemplified for the design of binders for PD-1, the introduction of a refinement step improved the seeds in almost every metrics used in the design process (Supplementary Fig. S 3.34). Indeed, the number of buried unsatisfied polar atoms decreased and, in contrast, the number of hydrogen bonds increased significantly. A closer look at the change of amino acid composition showed that a transition from hydrophobic residues towards hydrophilic has been operated. As highly hydrophilic residues tend to be more elongated and have a broader accessible surface area [224], more molecular contact can be achieved with the protein target which is reflected by the increased of buried solvent-accessible surface area (dSASA). As a consequence of this broader contact and increased polar interactions, the overall change in computed binding free energy ($\Delta\Delta G$) decreased. However, this reduction is not only quantitative – because of the enhanced contact area – but also qualitative as the dSASA-normalized $\Delta\Delta G$ also showed a significant decrease, which attests to the enhanced quality of binding achieved in the refinement process. Altogether, these improvements in every metrics can undoubtedly account for the success of the optimized pipeline and the design of protein binders with native-like affinities straight from the computer.

With the introduction of an additional step in the design pipeline, the role of each component – MaSIF or Rosetta – becomes more convoluted than before. A deeper analysis of the refined helical seeds for PD-1 reveals that each seed has on average 45.5% of its residues undergoing mutation (6.6 residues per seed) (Supplementary Fig. S 3.34). Additionally, 66.2% of its residues that were initially in contact with the target prior to refinement (MaSIF-suggested “hotspots”) were converted to another amino acid




**Supplementary Figure S 3.34 : Pre- and post-refinement metrics of the binding seeds for PD-1.** **A.** Metrics measured for 115 helical seeds with Rosetta showing shape complementarity, Gibbs free binding energy (ΔΔG), the binding energy normalized with the buried surface area, the number of buried unsatisfied polar atoms, the number of hydrogen bonds and the buried surface area (dSASA). Independent t-test with Bonferroni correction; non-significant (n.s.); p-value < $10^{-4}$ (****) **B.** Number of residue refined on each entire seed (top) and among hotspot residues prior refinement (bottom). **C.** Ratio of residues refined among each entire seed (top) and among hotspots residues prior refinement (bottom). **D.** Heatmap of mutated residues prior and after refinement. Residues are ranked from most hydrophilic to most hydrophobic according to Kyte-Doolittle scale. Yellow box indicates a cluster of mutations from hydrophobics towards hydrophilics. Colored bar represents mutation counts. **E.** Descriptor distance score measured by MaSIF-seed prior and after refinement.

(5 residues per seed in average).This suggests that MaSIF contributes to 33.8% of the contact residue of each seed and Rosetta for the rest. Importantly, MaSIF still plays a crucial role for the identification of the binding site (interface propensity prediction by MaSIF-site), the seed selection and its placement on the predicted interface. Rosetta operates a synergetic role by increasing the number of contact residues (+1.6 residue in average) and solving the constraints of the seed database which represents a discrete and finite number of possibilities for MaSIF despite its large size. Finally, an *a posteriori* analysis of the refined seed by MaSIF indicates a moderate increase of the descriptor distance score[3] compared to prior refinement, showing a consensus agreement between Rosetta and MaSIF (Supplementary Fig. S 3.34).

Thus, while Rosetta refinement step bring a substantial contribution to the seed improvement, a collaborative process with MaSIF is still essential to obtain site-specific binding motifs.

### 3.7.2 Translational application of the designed PD-1 binder

On top of bringing a better understanding of PPI design, this work also aimed to provide functional protein binders with translational capabilities. With this rational, a collaboration with the group of Dr. Ricardo A. Fernandes at the University of Oxford was initiated to study the effect of our PD-1 binder, DBP13_02, to the PD-1 receptors found on T cells. PD-1 acts as an immune checkpoint blockade, binding to PD-L1 which is often aberrantly expressed on the surface of cancer cell to escape the T cell immunity [180].

While an antagonist effect was initially expected, such as the one reported with the PD-1 blocking antibody Nivolumab [182] (Supplementary Fig. S 3.35), an agonist effect was observed instead when treating T cells with DBP13_02. Indeed, a downregulation of CD25 (also known as interleukin-2 receptor alpha chain) and CD137 (also known as 4-1BB) was observed on both $CD4^+$ and $CD8^+$ T cell populations, which are two known markers of T cell activation [225,226] (Supplementary Fig. S 3.36). Furthermore, a reduction of the secretion interleukin-2, a cytokine promoting T cell proliferation upon activation [227], was observed in the supernatant of PBMCs (peripheral blood mononuclear cell) treated with an increasing concentration DBP13_02. As a consequence, the percentage of proliferating cells also significantly decreased in both $CD4^+$ and $CD8^+$ T cell populations. Altogether, these data support the agonist effect of DBP13_02, which gives similar outcome as the native binder PD-L1 but more potently.

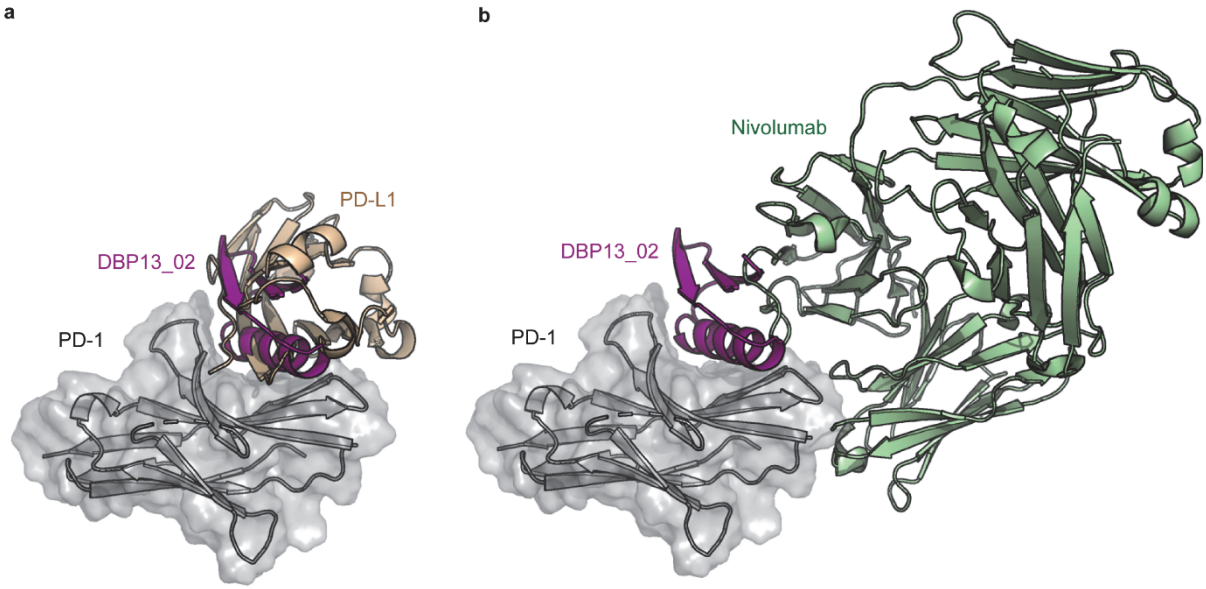
Different hypotheses can explain the observed phenomenon. Firsty, the reported binding mode of Nivolumab is significantly different from the predicted one of DBP13_02. Nivolumab introduces a steric clash with PD-L1 while remaining slightly off PD-L1 binding site [228]. On the other hand, DBP13_02 perfectly correlates with PD-L1 binding site and may therefore be mimicking the same inhibitory stimulus (Supplementary Fig. S 3.35). Secondly, the observations could be explained by the intermediate affinity of DBP13_02 ($K_D$ = 794 nM) which is in line with previous findings demonstrating that low-affinity antibody can increase receptor clustering and therefore the agonist effect [229].

[3] Descriptor distance score (DDS) is an indicator of how close two fingerprints are: $DDS = \sum_{i=0}^{N} 1/d_i^2$ where $d$ is the descriptor distance between a pair of point $i$, with a total pair of contact points $N$.

While synthetic PD-1 agonists using known PD-L1 and PD-L2 motifs as a template were previously proposed [124], the DBP13_02 binder described in our work is, to our current knowledge, the first fully *de novo* protein design showing an agonist effect for PD-1. Overall, these collaboration highlighted the translational capabilities of the *de novo* protein binders designed with our surface-centric approach.

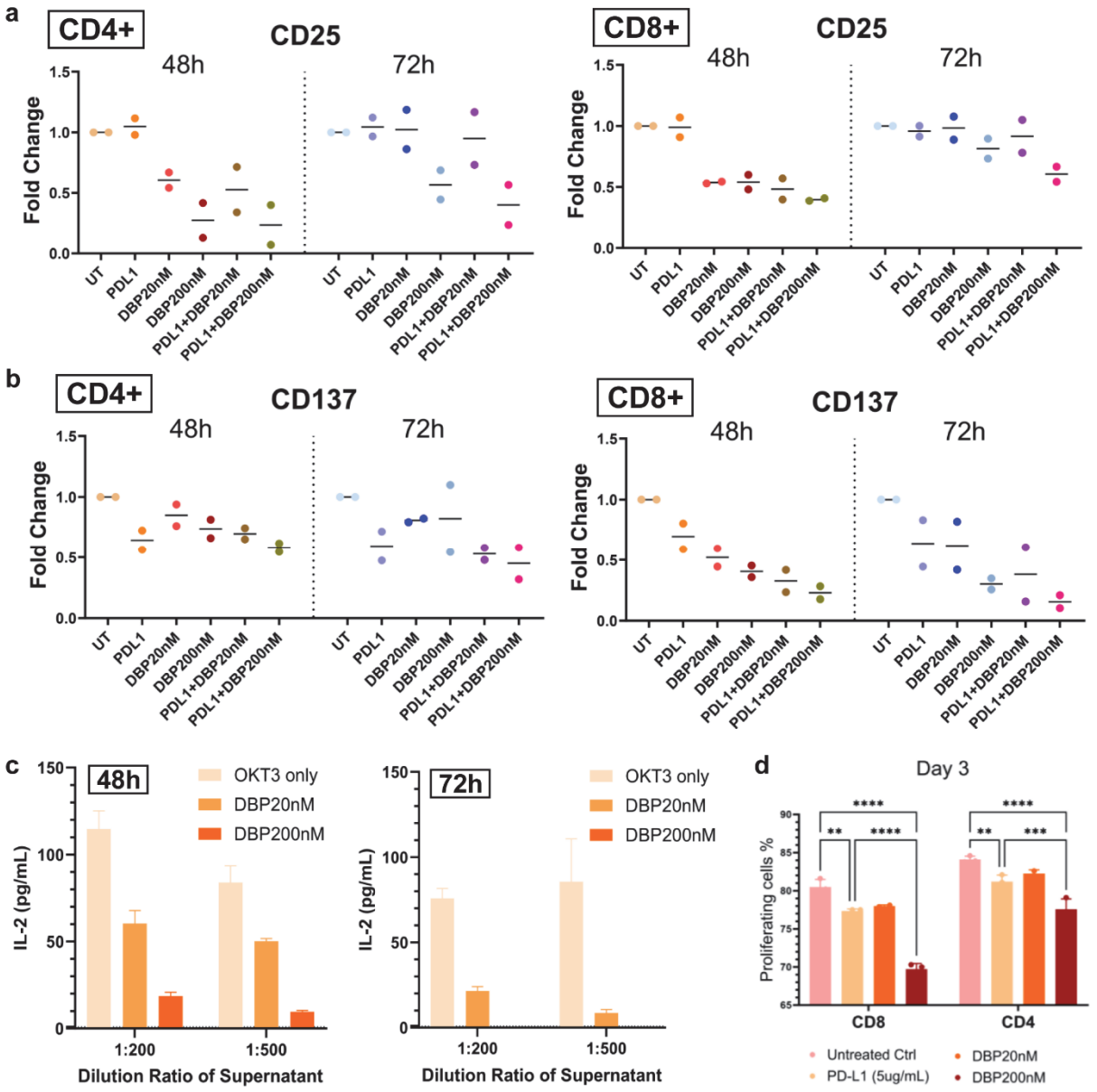
### 3.7.3 Personal contribution

This work involved the contributions of several co-first authors and would not have been possible without the synergetic work between all of them. My colleagues mostly focused on developing the MaSIF-seed pipeline, designing, screening and characterizing the *de novo* protein binders against PD-L1 and SARS-CoV-2. Due to the extensive *in vitro* optimization needed to achieve binders with affinities in the range of native PPIs, my main contribution was the implementation of the optimized design pipeline for one-shot protein binders with enhanced affinities. I personally implemented the improved computational pipeline, designed, screened and characterized binders for PD-1, CTLA-4 and PD-L1[4]. Moreover, I autonomously handled the revisions and edition of the manuscript for its publication in Nature.




**Supplementary Figure S 3.35 Comparison between DBP13_02, PD-L1 and Nivolumab. A.** Binding of PD-L1 (wheat) to PD-1 (gray) (PDB ID: 4ZQK) compared to the modeled binding mode of DBP13_02 (purple). **B.** Binding of Nivolumab (green) to PD-1 (gray) (PDB ID: 5WT9) compared to the modeled binding mode of DBP13_02 (purple)

[4] Together with Dr. Andreas Scheck for the computational designs of CTLA-4 and PD-L1.




**Supplementary Figure S 3.36 : Downregulation of T cell activation and proliferation upon DBP13_02 treatment. A-B.** Fold change in the expression of CD25 (A) and CD137 (B) for both $CD4^+$ (left) and $CD8^+$ (right) T cells upon treatment with PD-L1, DBP13_02 or both combined together. **C.** Concentration of interleukin-2 measured in PBMC supernatant by enzyme-linked immunosorbent assay (ELISA) after 48h (left) or 72h (right). **D.** Percentage of proliferating $CD8^+$ or $CD4^+$ T cell measured by flowcytometry. PBMCs were pre-labelled with CFSE (Carboxyfluorescein succinimidyl ester) and mixed with Dynabeads (anti-CD3 and anti-CD28) to stimulate the proliferation. DBP13_02 or PD-L1 were added to treat PBMCs for 3 days. Experiments and data analysis were performed by Xiaonan Zheng from Dr. Ricardo A. Fernandes Lab.

# Chapter 4

## Targeting protein-ligand neosurfaces using a generalizable deep learning approach

As discussed previously, protein-protein interactions are at the basis of multiple processes in the regular cell homeostasis but also in disease progression. Scientists have therefore leveraged the power of protein interactions to design new biotechnology tools and therapies. Designing novel protein interactions is an efficient way to fight emergent diseases and oncogenic targets as exemplified in the previous chapter. But these novel interactions can also be harnessed for engineering innovative biotechnology tools like chemically-induced switches for cell-based therapies or synthetic biology in general. However, most known chemically-induced dimerization (CID) systems have been exploiting existing systems, and generalizable computational tools for the design of *de novo* CID are missing to the protein science community. In this section, we strived to further improve our MaSIF pipeline for the integration of small molecules and the design of protein binders that are specific to a defined drug-protein complex.

The following section is taken from a preprint manuscript (doi: 10.1101/2024.03.25.585721), that could undergo further modifications.

**Authors**

Anthony Marchand[1‡], Stephen Buckley[1‡], Arne Schneuing[1‡], Martin Pacesa[1], Pablo Gainza[1], Evgenia Elizarova[1], Rebecca Neeser[1,2], Pao-Wan Lee[3], Luc Reymond[4], Maddalena Elia[1], Leo Scheller[1], Sandrine Georgeon[1], Joseph Schmidt[1], Philippe Schwaller[2], Sebastian J. Maerkl[3], Michael Bronstein[5] & Bruno Correia[1]

**Affiliations**

[1] Laboratory of protein design and immunoengineering, Institute of Bioengineering, Ecole polytechnique fédérale de Lausanne, Lausanne (Switzerland)

[2] Laboratory of biological network characterization, Institute of Bioengineering, Ecole polytechnique fédérale de Lausanne, Lausanne (Switzerland)

[3] Laboratory of chemical artificial intelligence, Institute of Chemical Sciences and Engineering , Ecole polytechnique fédérale de Lausanne, Lausanne (Switzerland)

[4] Biomolecular screening core facility, School of Life Sciences, Ecole polytechnique fédérale de Lausanne (Switzerland)

[5] Department of Computer Science, University of Oxford, Oxford, UK

[‡] Equal contribution

**Author contributions**

A.M., S.B. and A.S. contributed equally to this work. A.M. and B.E.C led the project. A.M., S.B., P.W.L. and M.E. performed the experimental work. A.M., S.B., M.P. and L.S. designed the experimental methodology. A.M. and A.S. performed the computational work and protein design. A.M., A.S., P.G., E.E. and R.M.N. contributed to the design of the computational pipeline. M.P. solved the crystal structure. L.R. synthesized the small molecule analogs. S.G. and J.S. participated in the expression and purification of proteins. P.S., S.J.M., M.B. and B.E.C provided supervision and acquired the necessary funding. A.M., S.B., A.S., M.P. and B.E.C wrote the manuscript with inputs from all authors.

**Funding**

This work was supported the Swiss National Science Foundation grant 310030_197724 (B.E.C, A.M., M.E.), TMGC-3_213750 (B.E.C, S.B.), 200020_214843 (P.W.L., S.J.M.); the National Center of Competence in Research in Molecular Systems Engineering grant 182895 (B.E.C and A.M.); the National Center of Competence in Research in Catalysis grant 180544 (P.S.); EPSRC Turing AI World-Leading Research Fellowship No. EP/X040062/1 (M.B.); Microsoft Research AI4Science (B.E.C and A.S.); VantAI (R.M.N.); Huawei Technologies Düsseldorf (B.E.C, L.S.); Reprodivac grant SEFRI 22.00135 (B.E.C, E.E.); the H2020 Marie Sklodowska-Curie EPFL-Fellows grant (P.G.); the “Peter und Traudl Engelhorn Stiftung” (M.P.).

## 4.1 Abstract

Molecular recognition events between proteins drive biological processes in living systems. However, higher levels of mechanistic regulation have emerged, where protein-protein interactions are conditioned to small molecules. Here, we present a computational strategy for the design of proteins that target neosurfaces, i.e. surfaces arising from protein-ligand complexes. To do so, we leveraged a deep learning approach based on learned molecular surface representations and experimentally validated binders against three drug-bound protein complexes. Remarkably, surface fingerprints trained only on proteins can be applied to neosurfaces emerging from small molecules, serving as a powerful demonstration of generalizability that is uncommon in deep learning approaches. The designed chemically-induced protein interactions hold the potential to expand the sensing repertoire and the assembly of new synthetic pathways in engineered cells.

## 4.2 Introduction

Protein-protein interactions (PPIs) play an essential role in healthy cell homeostasis, but are also involved in numerous diseases [9,230]. For this reason, several therapies targeting PPIs have been developed over the last decades and multiple computational tools have been recently proposed to design novel protein interactions [167]. The governing principles determining the propensity of proteins to form interactions are intricate due to the interplay of several contributions, such as geometric and chemical complementarity, dynamics, and solvent interactions. Therefore it remains challenging to predict and design novel PPIs, especially in the absence of evolutionary constraints. Native PPIs can also be controlled by additional regulatory layers such as allostery [231], post-translational modifications [232], or direct ligand binding [233,234] Compound-bound surfaces, which we refer to as neosurfaces, are one

of the most fascinating and challenging types of molecular recognition instances, where relatively minor changes at the protein binding site can have a large impact on binding affinities. The interest in such interactions has been fueled by the development of new drug modalities; specifically molecular glues that form neosurfaces to trigger protein interactions for degradation and other applications [235,236], thus representing a promising route for the development of innovative therapeutics.

In synthetic biology, molecular components that rely on small molecule-induced neosurfaces have been used to engineer chemically-responsive systems with precise spatio-temporal control of cellular activities [237]. Small molecule triggers have been used to both induce and disrupt PPIs, thereby functioning as ON or OFF switches for engineered cellular functions [97,105,237]. There are several practical advantages in using small molecules as triggers due to their simple administration, biodistribution, cell permeability, safety, and high affinity and specificity to their target proteins. Protein-based switches controlled by small molecules have already been applied to regulate transcription [238,239], protein degradation [240–242], and protein localization [243–245], among many other applications. In addition to their use in basic research, engineering molecular switches is becoming a more common mechanism of controlling protein-based and cellular therapeutics, whose activity may have to be regulated to mitigate potentially dangerous side effects [97,246,247]. While several chemically-disruptable heterodimer (OFF-switch) systems have been proposed [97,237,246], computationally designed chemically-induced dimerization (CID, ON-switch) systems remain challenging due to the complexity of modeling neosurfaces. Previous attempts at designing CID systems primarily relied on experimental methods [150,237,238,248] and, despite the emergence of artificial intelligence and numerous computational tools, only few tools can generalize to both proteins and small molecules as a target for protein design, resulting in a lack of suitable approaches for the design of novel chemically-induced PPIs. Computational methods to design novel CIDs mostly relied on transplanting an existing drug binding site to a known heterodimer interface [249] or using docking of putative pre-existing proteins (i.e. scaffolds) followed by interface optimization [169]. However, these approaches can face limitations such as the risk of drug-independent dimerization, the lack of suitable scaffold proteins for design, or the extensive need for in vitro maturation techniques.

We recently reported a geometric deep learning-based framework called MaSIF (Molecular Surface Interaction Fingerprinting) [138] for the study of protein surface features, and for the design of novel protein-protein interactions [250]. In this study, we aim to test whether our surface-centric approach can generalize to non-protein ligands without additional training data by using a higher-level representation, namely the geometric and chemical features found on the molecular surface. To do so, we designed site-specific binders that target neosurfaces composed of a small molecule ligand and protein surface moieties, resulting in de novo ligand-dependent protein interactions. We successfully designed and characterized novel protein binders recognizing the B-cell lymphoma 2 (Bcl2) protein in complex with the clinically-approved inhibitor Venetoclax [251], the progesterone-binding antibody DB3 in complex with its ligand [252], and finally the peptide deformylase 1 (PDF1) protein from Pseudomonas aeruginosa in complex with the antibiotic Actinonin [253]. Lastly, we show that such ligand-controlled

systems can be utilized in both in vitro and cellular contexts for a range of synthetic biology applications, unlocking possibilities for the development and regulation of novel therapeutic approaches.

## 4.3 Results

### 4.3.1 MaSIF captures interaction propensity of neosurfaces

Within our geometric deep learning framework, MaSIF [138], we previously developed two applications: i) MaSIF-site to accurately predict regions of the protein surface with a high propensity to form an interface with another protein and, ii) MaSIF-search to rapidly find and dock protein partners based on complementary surface patches. In MaSIF-search, we extract surface patch descriptors ("fingerprints"), so that patches with complementary geometry and chemistry have similar fingerprints, whereas non-interacting patches have low fingerprint similarity. Surface fingerprints allow to perform an initial ultra-fast search in an alignment-free manner using the Euclidean distances between them. Patches with fingerprint distances below a threshold are then further aligned in 3D and scored with an interface-post alignment (IPA) score to refine the selection.

In its initial conception, MaSIF only considered canonical amino acids as part of the protein molecular surface and was not compatible with small molecules, glycans, and other ligands. Thus, we present here MaSIF-neosurf to incorporate small molecules as part of the molecular surface representation of the target protein to predict interfaces and partners based on the neosurface fingerprints (Fig. 4.1A, see Methods). MaSIF was initially trained to operate on general chemical and geometric surface properties of biomolecules, while abstracting the underlying structure. Thus, it is not restricted to only protein surfaces, but should in principle also capture the surface patterns arising from other non-protein surfaces. Upon generation of the molecular surface of the protein-drug complex, MaSIF-neosurf computes the two geometric features: shape index [192] and distance-dependent curvature [193]. In addition, three chemical features are also used: Poisson-Boltzmann electrostatics, which can be computed directly from the small molecule, and hydrogen bond donor/acceptor propensity [196] and hydrophobicity [194,254,255], for which we developed new featurizers tailored to capture the chemical properties of the small molecules (see Methods and Supplementary Fig. S4.1).

To assess the capabilities of MaSIF-neosurf, we benchmarked its performance on several ternary complexes whose interface is composed of protein and ligand surfaces. We aimed to recover known binding partners for proteins with small molecules at the binding interface. After assembling a list of 14 ligand-induced protein complexes, we split each of them into two subunits, resulting in 28 independent benchmarking cases, and processed them with and without the small molecule bound. The ligand-free protein surfaces, together with 200 decoy proteins, constitute our database, which we query with surface patches from all 28 protein-ligand complexes. Since each of the 228 protein candidates is decomposed into almost 4000 patches on average, the database represents a large search space with more than 900'000 potential binding sites. We then evaluated whether the correct binding partner is retrieved and docked in the correct rigid-body orientation. When considering the protein-ligand complex as a docking partner, MaSIF-neosurf recovers more than 70% of the correct binding partners and their binding poses (Fig. 4.1B). Only a small subset of test cases could be recovered in the absence

of the ligand and the general trend is that in such cases the protein surface is a large contributor towards the overall protein interaction (Supplementary Fig. S4.2). The ability to capture the neosurface properties is further supported by an increased descriptor distance score between interacting partners (i.e. an increased complementarity between interacting fingerprints, see Methods) and an increased interface post-alignment (IPA, see Methods) score in the presence of the small molecule compared to the case without (Fig. 4.1C-D). Overall, MaSIF-neosurf captures, in many instances, features that are determinant for ligand-mediated protein interactions and, to further test its capabilities, we sought to de novo design this type of interactions.

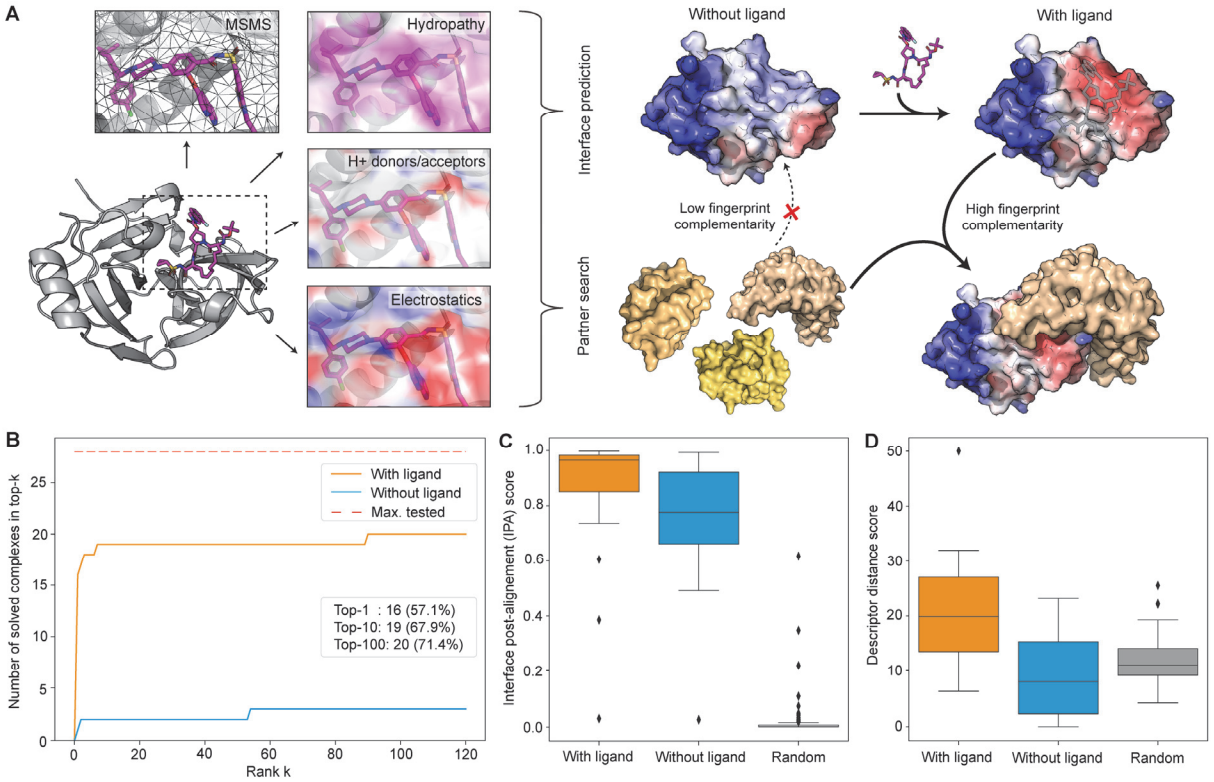



**Figure 4.1: Neosurface properties are captured to identify interface sites and binding partners**. **A.** Geometric and chemical features of the ligand-protein complexes are computed, including the molecular surface representation (MSMS), hydropathy score, proton donors/acceptors and Poisson-Boltzmann electrostatics. Surface features are vectorized in a descriptor (also referred to as "fingerprint") and used by MaSIF-neosurf for interface propensity prediction or protein partner search. The ligand-containing fingerprint is then used to find complementary fingerprints in a patch database. **B.** Ranking predictions using MaSIF-neosurf on a benchmark dataset of known ternary complexes and a set of 200 decoys. Complementary partner search was performed in the presence (orange) and absence (blue) of the respective small molecule ligand. **C-D.** Interface post-alignment score (IPA; C) and descriptor distance score (see Methods; D) of the interacting complexes in the presence (orange) and absence (blue) of the drug compared to a set of random patch alignments (gray)

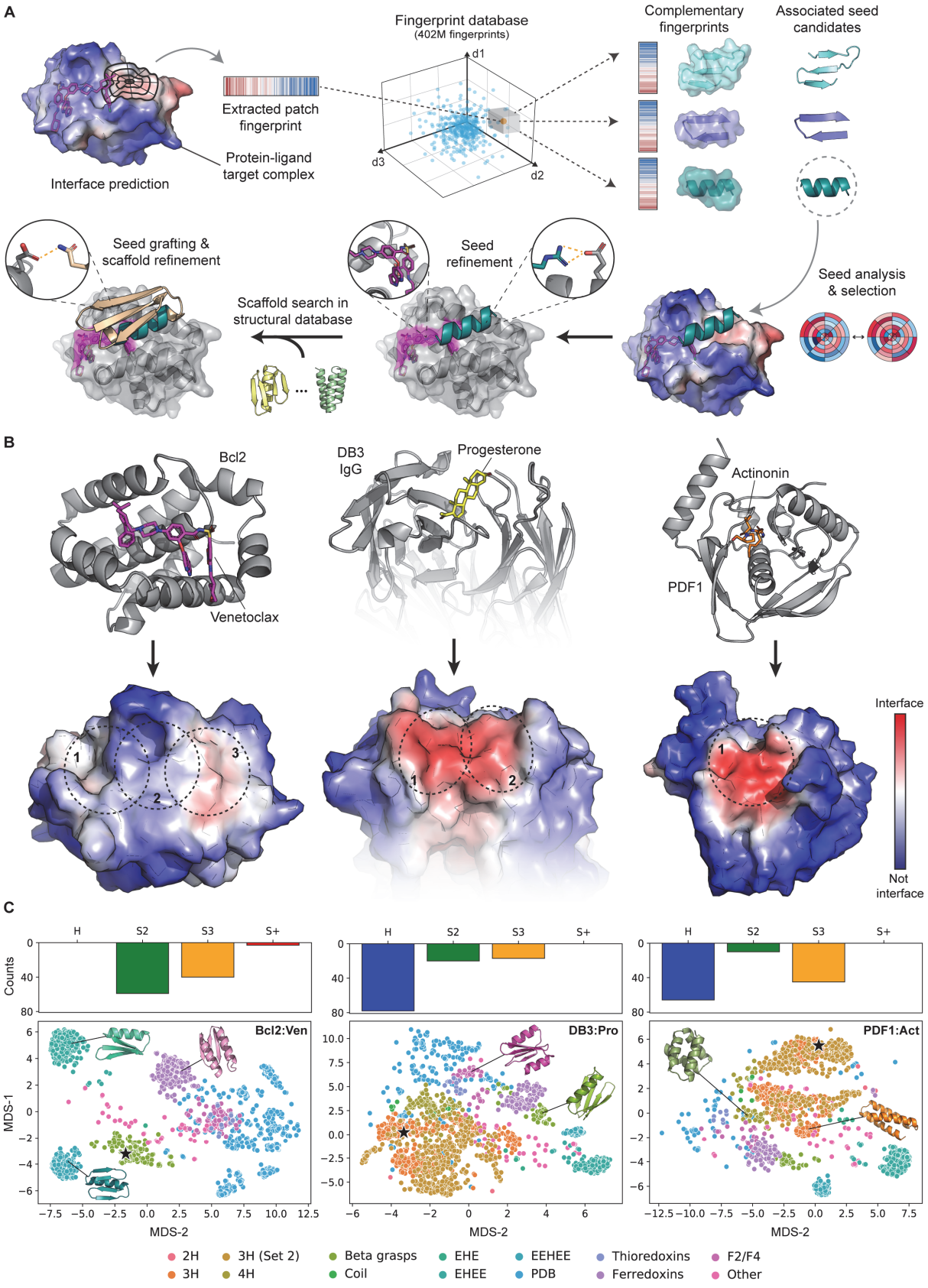



**Figure 4.2: Design of ligand-induced protein interactions with MaSIF-neosurf. A**. To design novel ligand-induced protein interactions, potential interface sites are first identified on the target protein-ligand complex. The corresponding patches are then used to find complementary fingerprints in a patch database. The top patches are aligned and scored to refine the selection. Associated binding motifs (seeds) undergo sequence optimization with an emphasis on designing new hydrogen bond networks with the target protein and small

molecule. Seeds are then grafted on suitable scaffolds from a structural database, and the rest of the scaffold interface is redesigned using Rosetta. Finally, the top ~2000 designs, according to different structural metrics, are selected and screened experimentally. **B.** Target candidates in complex with their respective small molecules (top row). Neosurfaces displaying their protein binding propensity (bottom row). Sites selected for binder design are highlighted with dashed circles. **C.** Seed structural diversity (top row) includes motifs that are: helical (H); two-strand beta sheets (S2); three-strand beta sheets (S3); and more complex beta sheet motifs (S+). Diversity of the ~2000 computational designs (bottom row) mapped using multidimensional scaling (MDS) of pairwise RMSDs between all designs. Experimentally confirmed binders are highlighted with a star.

### 4.3.2 Designing novel ligand-induced protein interactions

Recently, we proposed the MaSIF-seed pipeline for the design of de novo site-specific protein binders [250]. Given the performance of MaSIF-seed against multiple therapeutically relevant targets, we sought to test whether such an approach could generalize to design site-specific binders to neosurfaces composed of ligand and protein atoms. By doing so, we tackle the challenge of designing chemically controlled protein interactions and test our understanding of molecular recognition events mediated by neosurfaces. We therefore adapted our MaSIF-seed pipeline to our newly proposed MaSIF-neosurf framework (Fig. 4.2A). Once neosurfaces are computed for a given protein-ligand complex, we first take advantage of MaSIF-site to identify the regions most likely to become buried in an interface. Then an extensive fingerprint search identifies complementary structural motifs (i.e. binding seeds) from a database of ~640'000 structural fragments (402 million surface patches/fingerprints). Therefore, by focusing on the predicted buried regions of the interface and searching for highly complementary motifs, the vast space of patches and binding motifs is quickly reduced to the most promising candidates. Finally, the top seeds are refined by sequence optimization and grafted with Rosetta [223] on recipient proteins (i.e. scaffolds) to stabilize the binding motif. Lastly, a final round of sequence design is performed to improve atomic contacts at the interface.

We designed ligand-dependent protein binders targeting ligand-bound proteins from different families: Bcl2 in complex with the clinically approved drug Venetoclax; an anti-progesterone antibody (DB3) in complex with its ligand; and peptide deformylase 1 (PDF1) from P. aeruginosa in complex with the antibiotic Actinonin (Fig 4.2B). We first identified a moderate to high interface propensity of these neosurfaces with MaSIF-neosurf, selected 1 to 3 relevant interface patches depending on the solvent-accessible surface area exposed by the ligand (Fig. 4.2B), and searched for complementary fingerprints in our seed database. Top-ranking seeds were selected, refined, and grafted onto recipient scaffolds, and approximately 2000 designs per target complex were selected with computational filters (Fig. 4.2C and Supplementary Table S4.1, see Methods). Our pipeline generated designs with diverse helical and beta sheet-based binding motifs, as well as various protein folds, thus sampling a wide space of sequences and topologies (Fig. 4.2C). All selected designs were predicted to favorably engage the neosurface by showing increased interface structural metrics in the presence of the ligand, such as the predicted binding energy, the buried surface area and the number of atomic contacts (Supplementary Fig. S4.3).

### 4.3.3 Experimental validation of ligand-induced PPIs

The computational designs were screened by yeast display [111] and, after two rounds of fluorescent-activated cell sorting (FACS), enriched clones were deep sequenced (Supplementary Fig. S4.4 and Supplementary Table S4.2). We show one binder targeting each of the selected test cases (Fig. 4.3A). The best designs show no binding in the absence of the corresponding small molecules, whereas modest to high binding signals were observed with the ligands in yeast display experiments (Fig. 4.3B). These changes of binding signal upon small molecule addition are consistent with the expected behavior of a chemically induced PPI. Interestingly, small molecules contributed about 10-12% of the predicted target buried surface area, but they improved the predicted binding energy ($\Delta\Delta G$) of the interface compared to the ligand unbound form by 17% to 27.7%. This result demonstrates a small, yet critical contribution that each ligand plays in the binding event, highlighting the difficulty of the design problem (Supplementary Table S4.3).

Moreover, point mutants at the interface hotspot residues abrogated binding to the target complex, which further supports the designed binding mode (Fig. 4.3C). No binding was observed with the native scaffolds used for the seed grafting and interface design, underlying the critical role of the interface design pipeline (Fig. 4.3C). Finally, specificity towards the desired ligand was confirmed by using control compounds: S55746 for Bcl2, 19-O-Benzoyl-Progesterone (OBz-Pro) for DB3 IgG and Tertbutyldimethylsilyl-Actinonin (TBDMS-Act) for PDF1 (Fig. 4.3D and Supplementary Fig. S4.5). These analogs retained binding to the protein target (Supplementary Fig. S4.5). However, no binding to the designs was observed, confirming that the correct interface on the target complex is engaged with high ligand specificity (Fig. 4.3D).

### 4.3.4 Biochemical characterization and structural validation

To map the binding site with high confidence and identify potential beneficial mutations, we performed a site-saturation mutagenesis (SSM) study (Supplementary Fig. S4.6). To assess the effect of the different mutations over the designed ligand-dependent interaction, we computed the average enrichment score of each mutation when comparing binding versus non-binding populations on yeast display experiments, similar to other deep saturation mutagenesis studies [78,166]. Globally, we observed that such interactions have exquisite sensitivity to single-point mutants and that residues with high sensitivity mapped very closely to the designed interfaces, supporting the accuracy of our computational models (Fig. 4.4A).

The initial successful designs were expressed and purified for further biophysical characterization. All designs were monomeric, folded and highly stable in solution (Supplementary Fig. S4.7). All three designs showed binding affinities in the range of native transient PPIs [22], from mid-nanomolar to low-micromolar, after pure in silico generation (Supplementary Fig. S4.8). Specifically, DBAct553_1 showed a binding affinity ($K_D$) of 542 nM, DBVen1619_1 and DBPro1156_1 showed affinities of 4 μM and >10 μM, respectively.

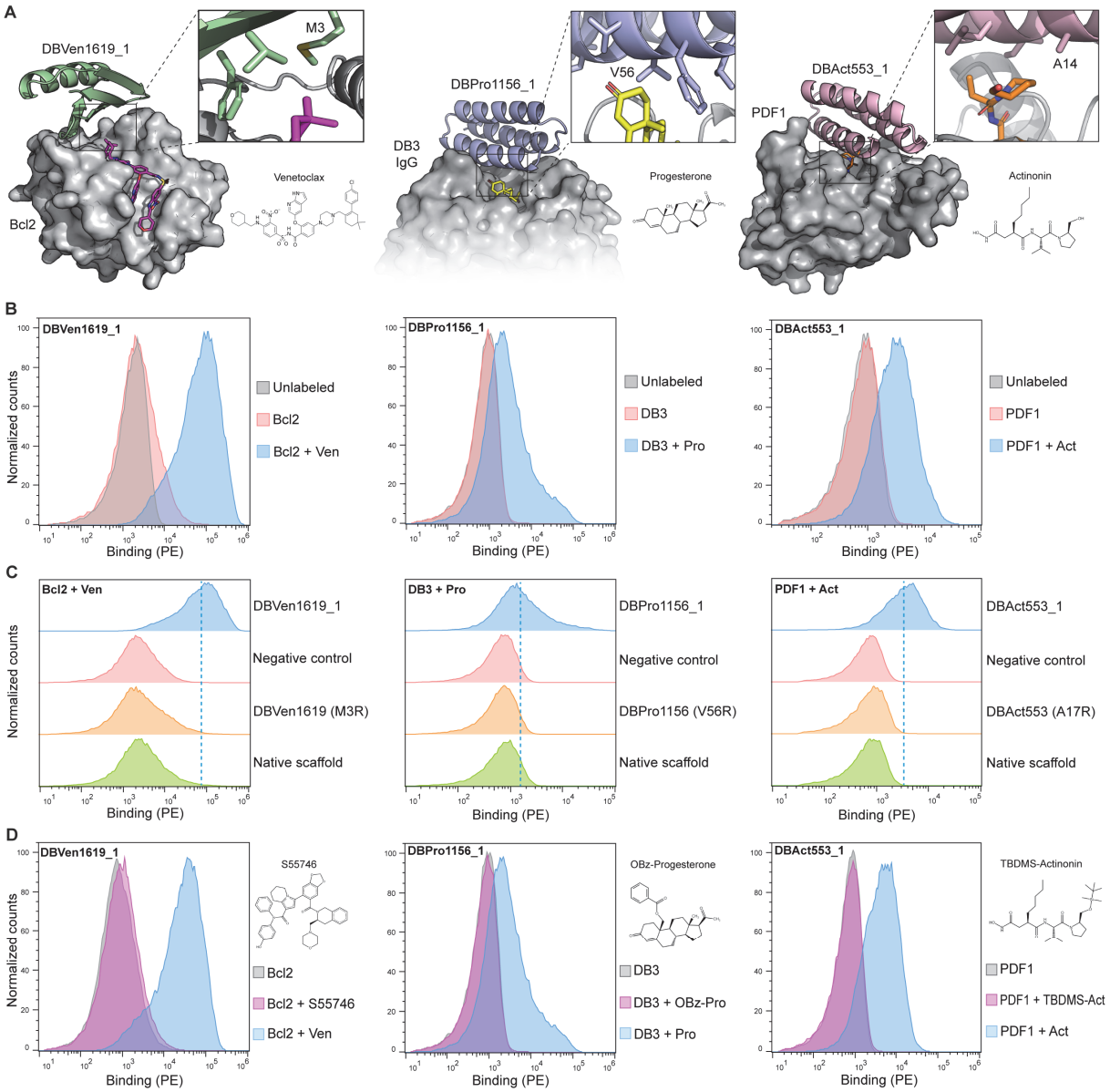



**Figure 4.3: De novo design and screening of small molecule-dependent binders. A.** Models of the designed binders in complex with their respective target complexes: Bcl2:Venetoclax, DB3:Progesterone and PDF1:Actinonin **B.** Histograms of the binding signal (PE, phycoerythrin) measured by flow cytometry on yeast displaying the designed binders. Yeast were either unlabeled or labeled with 500 nM of their respective target protein preincubated with the ligand, or with the target protein alone. **C.** Histograms of the binding signal (PE, phycoerythrin) measured by flow cytometry on yeast displaying designed binders, a mutated version with a single point mutant at the predicted interface and the starting scaffold used for the design process. Yeast cells were labeled with 500 nM of their respective drug:protein complex. Dashed lines represent the geometric mean of the designed binder signal. **D.** Binding measured on yeast displaying DBVen1619_1, DBPro1156_1 or DBAct553_1 labeled with the target protein alone (gray), the target protein in complex with the original small molecule (blue), or the target protein in complex with the small molecule analog (magenta). Control analogs tested were S55746, Progesterone-19-O-Benzoyl (OBz-Pro) and Tertbutyldimethylsilyl-Actinonin (TBDMS-Act).

In the SSM scan, some mutations suggested potential improvements in affinity (Supplementary Fig. S4.6). Due to the large number of beneficial mutation candidates for DBVen1619_1, we created a combinatorial library covering 6 residues, sampling a set of favorable amino acids identified by SSM (Supplementary Fig. S4.9). Three of the six positions converged into single mutations (K1Q, M3L, I13K) while the remaining three residues did not converge. We engineered a variant, DBVen1619_2, with the three beneficial mutations and confirmed the binding improvement on yeast display (Supplementary Fig. S4.9). Among the favorable mutations, M3L in the core of the interface between Bcl2:Venetoclax and DBVen1619_2 plays a crucial role (Fig. 4.4B). The conformational rigidity of a leucine is likely to be preferred to the rotameric flexibility of a methionine [256], reducing the entropic cost of the binding interaction [257]. On the other hand, the second beneficial mutation (I13K) is likely to provide a favorable electrostatic interaction with a glutamate nearby. Overall, the incorporation of the three mutations resulted in a 42-fold improvement of the affinity ($K_D$ = 96 nM, Fig. 4.4C)

For the progesterone-dependent binder, DBPro1156_1, four favorable mutations were identified by SSM and showed an increased binding on yeast display (Supplementary Fig. S4.10). Two mutations (Y12W and S16G) significantly improved the binding signal and showed an additive effect in the resulting design, DBPro1156_2. Modeling of the two mutations suggested increased interface packing (Y12W) and the removal of a steric clash (S16G) (Fig. 4B, middle panel). DBPro1156_2 showed a binding affinity of 18 nM, which represents an improvement of three orders of magnitude, relative to the parent design, solely with two mutations (Fig. 4.4C).

Several mutations were found to slightly improve binding of DBAct553_1 to Actinonin-bound PDF1 (Supplementary Fig. S4.11). Most of these mutations were hypothesized to result in a more elaborate hydrogen bond network across the interface (e.g. R7N or A8R) (Fig. 4.4B). Of note, the combination of I3E with R7N was found to be deleterious for binding (Supplementary Fig. S4.11), most probably because of their spatial proximity that might trigger unwanted side chain rearrangement. A combination of the beneficial mutations (R7N and A8R) gave rise to DBAct553_2, which bound with an affinity of 446 nM for the Actinonin-bound PDF1 (Fig. 4.4C).

To evaluate the structural accuracy of our computational design approach, we co-crystalized the ternary complex of Actinonin-bound PDF1 with DBAct553_1 (PDB: 8S1X, Fig. 4.4D). The crystal structure closely resembled the computational model with a $C_\alpha$ RMSD (Root Mean Square Deviation) of 2.33 Å and a full-atom interface RMSD (iRMSD) of 2.26 Å, which demonstrates the accuracy of our design pipeline. The deviation from our initial model can to a large extent be attributed to a misplaced residue (Y2) in the model of the design scaffold which induced a slight shift of the N-terminal helix (Supplementary Fig. S4.12). Consequently, the $C_\alpha$ RMSD of our model deviates 0.93 Å from that of the experimental structure (Supplementary Fig. S4.12). Of note, the AlphaFold2 [31] prediction of the monomeric designed binder aligned perfectly with our structure with a Cα RMSD of 0.49 Å by placing residue Y2 with the correct orientation. Overall, this observation together with previous findings suggests that an increased use of deep learning tools like AlphaFold should significantly increase the model accuracy and therefore success rate [85]. Finally, we solved a cryo-electron microscopy structure (3.23 Å local resolution) of the DBPro1156_2 in complex with DB3 Fab and progesterone that confirmed the designed binding mode and interface engagement with the small molecule (Fig. 4.4E,

Supplementary Fig. S4.13). Despite the absence of structural data for the remainder of the designs, the mutational sensitivity assessed by the SSM (Fig. 4.4A) and the lack of binding with the small molecule analogs (Fig. 4.3D) suggests that the binders engage the target interface with a binding mode in agreement with our computational models.

### 4.3.5 Functionalization in cell-based assays

Chemically controllable components have important applications in synthetic biology and have been shown to be useful in modulating the activity of emerging cell-based therapies [97,237,258]. To test whether our computationally designed CIDs would assemble in a more complex cellular context, we engineered reporter proximity-based systems that were expressed in cell-free system or mammalian cells and that in the presence of the small molecule could activate a signaling pathway or lead to the reconstitution of a reporter protein. The most natural functional logic for chemically-induced protein interactions is to function as ON-switch systems.

We first repurposed a previously described heterodimerization-based reporter system [259] to test the DB3 antibody as a single-chain variable fragment (scFv) binding to DBPro1156_2. Here, DB3 was fused to a zinc finger 438 transcription factor and DBPro1156_2 to a T7 RNA polymerase (Fig. 4.5A), and tested in a cell-free reporter system. The heterodimerization in presence of the drug induces proximity between the T7 RNA polymerase and the transcription factor, thus leading to the transcription of a reporter linear DNA template and its translation into a red fluorescent protein (mCherry). While only baseline fluorescence was observed in absence of progesterone, a 15.8-fold increase was observed after addition of progesterone (Fig. 4.5B). Similarly, a titration of progesterone demonstrated a dose-response curve, suggesting possible utilization as a novel cell-free biosensor (Fig. 4.5C).

To test the chemically-induced activity of the designed modules in mammalian cells, we used a previously described system called generalized extracellular molecule sensor (GEMS) [100]. Briefly, the target protein and the designed binder are both fused to an erythropoietin receptor (EpoR) linked to an intracellular domain of a human interleukin 6 receptor subunit B (IL6RB) (Fig. 4.5D). Transcription of a reporter gene (NanoLuc luciferase) [260] will be triggered upon a conformational change induced by the heterodimerization in presence of the drug. By incorporating Bcl2 and DBVen1619_2 in the GEMS system, we observed a 26.8-fold change in luminescence in the presence of Venetoclax, while minimal background was observed in the absence of the drug (Fig. 4.5E). These results show the desired behavior of an ON-switch system. Additionally, our modified GEMS system displayed a heightened sensitivity to the drug, with a half maximal effective concentration ($EC_{50}$) of 0.31 nM, which is likely due to the co-localization of the sensing modules in the cell membrane (Fig. 4.5F).

Next, we designed a cytoplasmic system to respond to Actinonin and fused PDF1 and DBAct553_1 to two moieties of a split NanoLuc (Fig. 4.5G). In this system we also observed a significant increase in signal (19.1-fold) upon dosing of the cells with Actinonin (Fig. 4.5H). This novel ON-switch system was also highly sensitive to the presence of the drug, as shown by the titration reporting an $EC_{50}$ of 27 nM (Fig. 4.5I). Overall, we showed that our computationally designed CIDs can be used to functionalize molecular components in cellular systems, suggesting a promising route for the development of new modules for synthetic biology including a wide range of biosensors and cell-based applications.

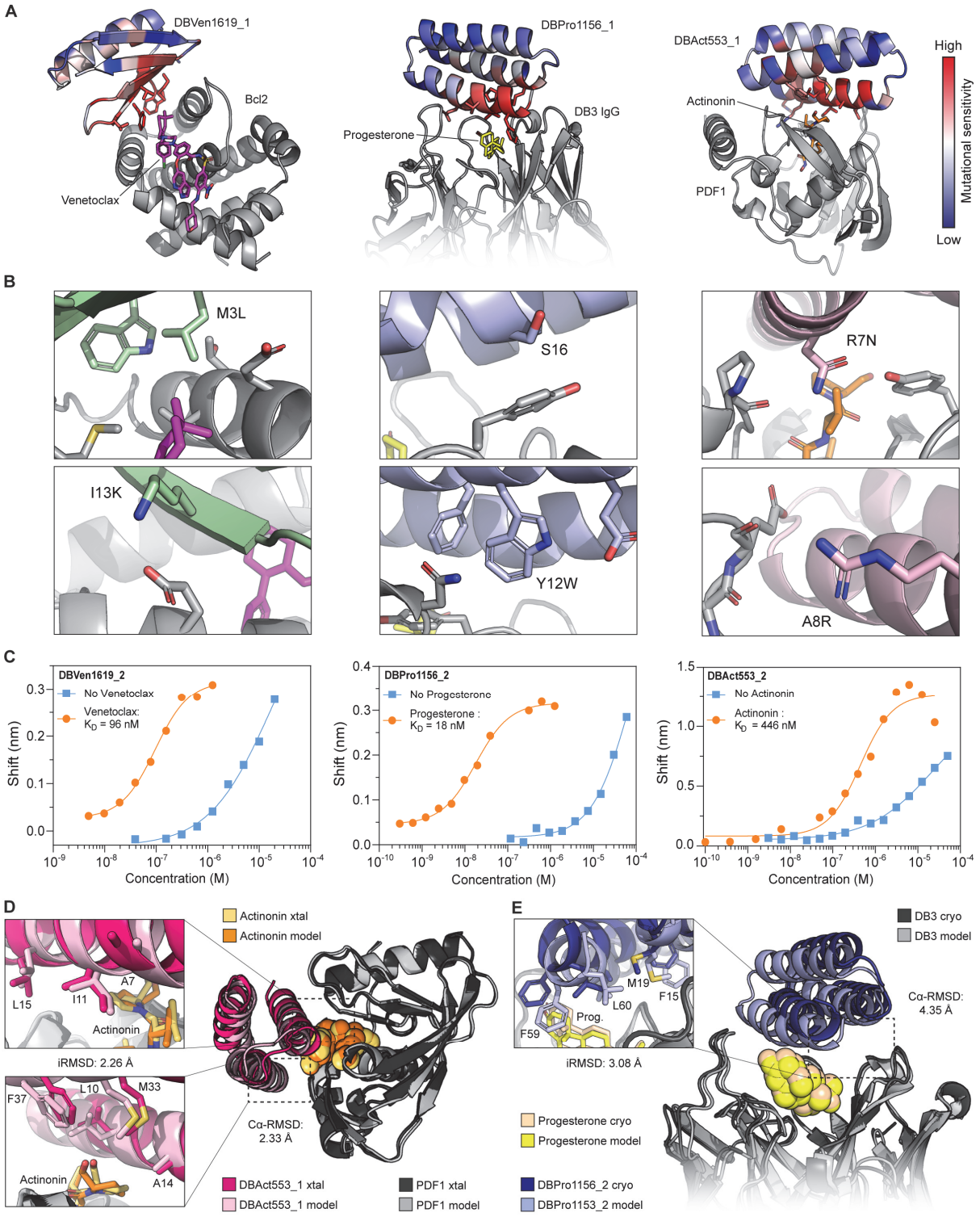



**Figure 4.4: Optimization, characterization and functionalization of the designed binders.** **A**. Computational model colored with the average enrichment score of the site-saturation mutagenesis of each amino acid position of the designed binder. Red color suggests that an amino acid position is sensitive to mutations, while blue color highlights a more tolerant amino acid position. Target proteins are shown in gray. **B.** Computational models incorporating the beneficial mutations that improved the affinity of the designed binders. Target proteins are shown in gray and designed binders in their respective color. **C.** Affinity measurement for DBVen1619_2, DBPro1156_2 and DBAct553_2 performed by biolayer interferometry. Each measurement was performed in presence (orange) or absence (blue) of the respective small molecule. The fits were calculated using a nonlinear four-parameter curve fitting analysis. **D.** Crystal (xtal) structure of DBAct553_1 in complex with

Actinonin-bound PDF1 (PDB: 8S1X). The computational model (light pink) is aligned with the crystal structure (magenta). Inset: the alignment of the residues at the interface. **E.** Cryo-electron microscopy structure obtained for DBPro1156_2 in complex with progesterone-bound DB3.The computational model (light blue) is aligned with the cryo-EM structure (dark blue). Inset: the alignment of the residues at the interface.

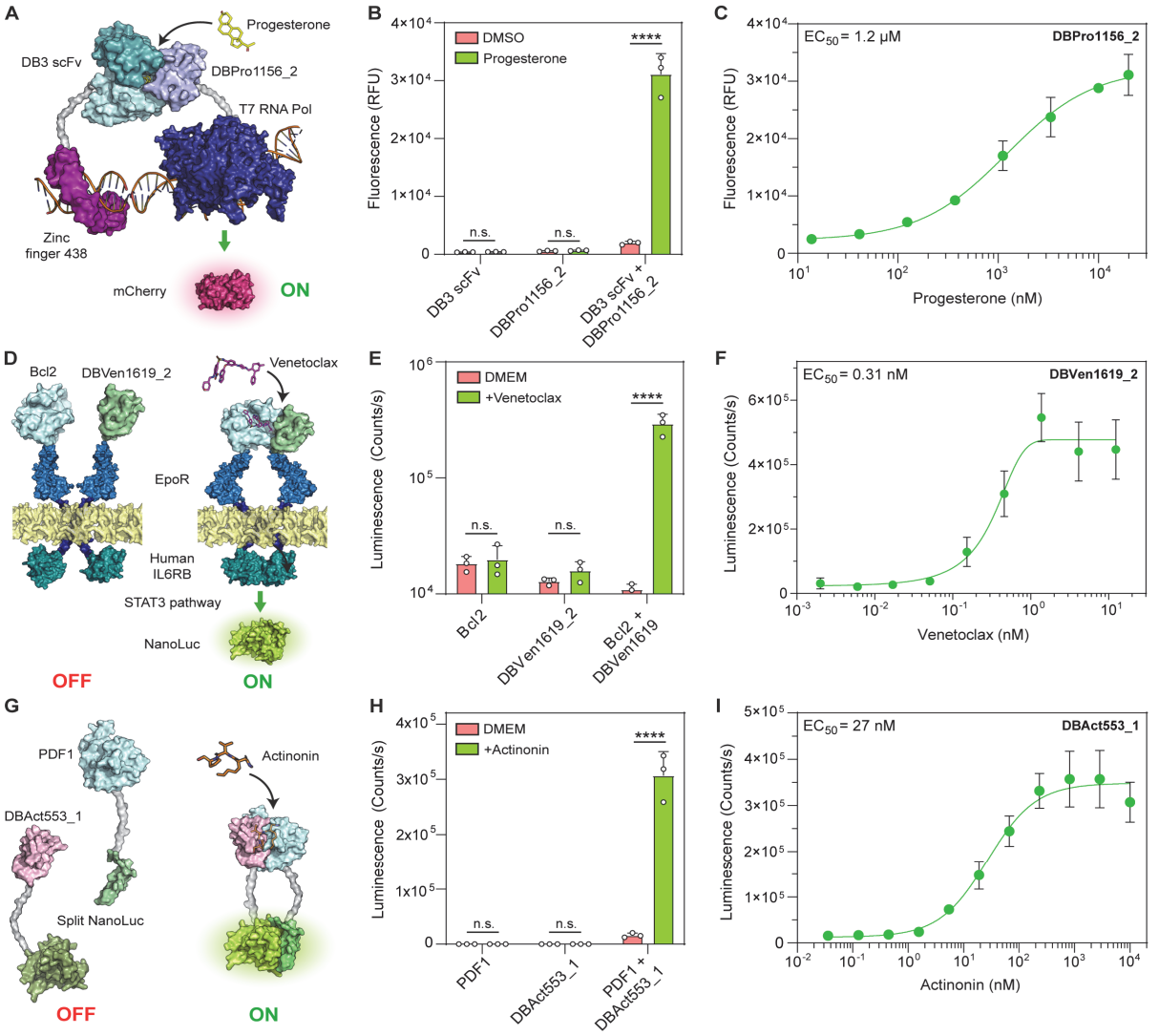



**Figure 4.5: Computationally designed CIDs are functional in cell-based systems. A.** Schematic of the cell free-expression system with single chain variable fragment (scFv) DB3-fused to a zinc finger transcription factor and DBPro1156_2 fused to T7 RNA polymerase. **B.** Fluorescence (Relative fluorescence unit; RFU) measured in wells containing each monomeric component or mixed, without or with 20 µM progesterone. **C.** Progesterone dose-dependent responses performed in a cell free system containing both components. **D.** Schematic of the GEMS reporter system functionalizing Bcl2-based CID. Both protein components of the CID are individually fused to erythropoietin receptor (EpoR) chains linked to an intracellular human IL6RB domain, which induces the expression of a reporter gene (secreted NanoLuc luciferase) when activated. **E.** NanoLuc luminescence of HEK293 cells transfected with Bcl2-GEMS only, DBVen1619_2 only or both together without or with 1 µM Venetoclax. **F.** Venetoclax dose-dependent responses performed on HEK293 transfected with Bcl2 and DBV1619 GEMS receptors. **G.** Schematic of the split NanoLuc system functionalizing DBAct553_1 and PDF1. **H.** Intracellular NanoLuc luminescence of HEK293 transfected with C-term split NanoLuc-fused PDF1 only, N-term split NanoLuc-fused DBAct553_1 only or both together without or with 25µM Actinonin. **I.** Actinonin dose-dependent responses performed on HEK293 transfected with split-NanoLuc PDF1 and DBAct553_1. $p<0.0001$ (****), non-significant (ns).

## 4.4 Discussion

Current deep learning-based protein design pipelines are primarily conditioned on the natural amino acid repertoire [37,38,80] and therefore lack generalization to the design of interactions involving small molecules. This gap is mainly due to the scarcity of protein-ligand structural data, and especially ternary complexes, within the training sets based on the PDB, where such complexes are limited [261–263]. Geometric deep learning approaches principled in the physical and chemical features of the molecular surface can overcome these limitations, and provide joint representations for protein and small molecule complexes. The resulting neosurfaces capture and present generalizable molecular features that enable the challenging task of designing protein binders targeting these hybrid interfaces. Utilizing the MaSIF-neosurf framework, we successfully designed three specific binders against Bcl2:Venetoclax, DB3:Progesterone and PDF1:Actinonin complexes. All designed binders showed high stability, specificity and native-like affinity for their target complexes by pure in silico generation. The affinities were experimentally optimized to nanomolar range and their binding mode was confirmed through mutational and structural characterization, showcasing the accuracy of our design pipeline. Notably, our pipeline managed to capture the subtle, yet crucial contributions of each ligand (10-12% of the buried SASA only; Supplementary Table S4.3) to induce protein interactions. This sensitivity represents an additional layer of complexity to the task of designing highly sensitive CIDs, compared to previous attempts targeting large ligand interfaces [169].

To demonstrate the functionality of our designed CID systems, we probed their efficiency and specificity in the context of a complex cellular environment. They exhibited robust ON-switch behavior in both cytoplasmic and membrane-bound circuits, showcasing their potentially wide applicability in mammalian systems as logic gates, synthetic circuits, or new biosensors for detecting specific metabolites [100,237]. This relevance is further underscored by our use of the FDA-approved drug Venetoclax for treating leukemia [251] the natural product Actinonin with potentially chemotherapeutic effects [253] or the endogenous hormone progesterone [264]. These can be utilized for combined anti-cancer therapies with chimeric antigen receptor (CAR) T cells, which are often hindered by off-target toxicities [97,265]. The addition of synthetic small molecule activators could allow finer control of their activity and elevate their safety profile.

While the design of specific protein-ligand interactions remains challenging, the presented results lay a strong foundation for further innovations. Incorporation of deep learning-based structure validation methods, such as AlphaFold2 [31,266] (Supplementary Fig. S4.14) or RoseTTAFold [57], or generative models complemented with surface fingerprints [38] could improve design success rates. Additionally, accounting for conformational flexibility and dynamics at the surface could pave the way for more complex interaction types, such as intrinsically disordered proteins. Overall, we envision that surface-based representation can contribute to solving molecular design problems in low-data regimens, such as the design of protein-based molecules with non-natural amino acids. The capability of extracting expressive fingerprints from protein:ligand complexes opens up the tantalizing possibility of rationally designing innovative drug modalities, such as on-command cell-based therapies [97,242], controllable biologics [150,246], or molecular glues, which thus far remains an outstanding challenge in drug development [235,236].

## 4.5 Methods

*Incorporation of small molecules in MaSIF-seed*

Molecular surface meshes were triangulated using the MSMS program [141] and radial patches (geodesic radius 12Å) computed following the original MaSIF preprocessing scripts [138]. Before applying MaSIF's geodesic convolutional layers, five input features are computed for each patch: shape index [192], distance-dependent curvature [193], Poisson-Boltzmann continuum electrostatics, hydrogen bond donor and acceptor potential [196], and hydropathy [194,254,255]. The first two features are purely geometric and are calculated analogously to protein surfaces alone. Moreover, the APBS program [195] used for computing the Poisson-Boltzmann electrostatics on the surface supports small molecules in the MOL2 file format and hence does not require us to treat them in a conceptually different way. The remaining two chemical input features are computed as described below.

*Hydrogen bond donors and acceptors*

The hydrogen bond propensity feature assigns a positive value to points on the molecular surface near the optimal direction in which a hydrogen could be formed with an acceptor atom. It is determined by the direction of the covalent bond between a donor atom and its hydrogen (Supplementary Fig. S4.1B-C). Likewise, a negative value is assigned to points corresponding to hydrogen bond acceptors. For different acceptor types, the theoretically optimal position for forming a hydrogen bond can either lie on a cone (Supplementary Fig. S4.1D-F) or in a small number of specific directions that can be derived from the molecular geometry. We assign different magnitudes of the donor/acceptor feature based on the angular deviation from the ideal hydrogen bond geometry according to a quadratic function.

The optimal direction of the hydrogen bond is determined using the RDKit software package [267] and surface points are assigned positive (donor) or negative (acceptor) values between -1 and +1 based on their angular deviation from the ideal direction. For potential acceptors, RDKit was also used to determine whether the idealized location of the hydrogen bond lies on a cone or in one or more discrete directions.

*Hydropathy*

MaSIF's hydrophobicity feature makes use of the Kyte-Doolittle scale [194] which is exclusively defined for amino acids. Equivalent values for small molecules thus need to be approximated based on a more general hydrophobicity measure that can be estimated computationally, such as the logarithm of the octanol–water partition coefficient (LogP) [254]. To this end, we develop a nonlinear function that maps LogP values to the KD scale. We fit the parameters of this function to find an optimal match for the KD and LogP values of all twenty amino acids. Since the best functional form of this mapping is not immediately obvious from the raw values (Supplementary Fig. S4.1L), we experimented with different hydrophobicity scales as intermediates and found that the Eisenberg scale [255] has approximately linear and exponential relationships with LogP and KD-values of amino acids, respectively. We first compute the optimal parameters of the mappings from LogP to Eisenberg scale (Supplementary Fig.

S4.1G) and Eisenberg scale to Kyte-Doolittle scale (Supplementary Fig. S4.1H) and then compose these two functions to establish the desired relationship between LogP and KD values (Supplementary Fig. S4.1H). Finally, we also restrict the outputs to the valid interval of KD values [-4.5, 4.5] to ensure that the feature does not leave the domain MaSIF was trained on.

Furthermore, since some ligands can cover large surface patches, we aim to respect local variations of the hydrophobicity by fragmenting the molecules prior to calculating their hydrophobicity score. We employ the BRICS algorithm [268] to decompose molecules, and compute estimates of each fragment's logP value with RDKit. The resulting fragments are more similar in size to amino acids and tend to have less extreme hydrophobicity scores than whole ligands, moving the distribution of this feature closer to that expected on protein surfaces (Supplementary Fig. S4.1K-L). To translate from logP to the Kyte-Doolittle scale, we parameterize a function so that it approximates the relationship between these hydrophobicity values for the 20 amino acids. KD and Eisenberg values of all amino acids are available in tabular form, whereas we compute their LogP with RDKit to fit the curves. The final function is

$$KD = clip\,(-6.2786 + exp\,(0.4772 * logP + 1.8491), min = -4.5, max = 4.5).$$

After computing equivalent KD values for all small molecule fragments, we assign the resulting hydrophobicity score of the closest fragment to each surface vertex.

To create the histograms in Supplementary Fig. S4.1G, we extracted 20,363 unique small molecule ligands from the Binding MOAD [269] dataset, fragmented each, and removed duplicates. This resulted in 9,362 unique fragments that are compared to the set of ligands and the twenty standard amino acids.

*Binding site identification*

MaSIF-site [138] was trained on a dataset of known PPIs to predict regions on protein surfaces with high propensity for forming a buried interface. The neural network takes a protein-ligand complex decomposed into 12 Å (geodesic radius) overlapping patches as input and generates a per-vertex regression score, indicating the propensity of each point to become a buried surface area within a protein interaction. In this study, we employed MaSIF-site to predict interfaces and guide the selection of target patches both in our computational benchmark and for all three target complexes for design (Bcl2:Venetoclax, DB3:Progesterone and PDF1:Actinonin). In the computational benchmark, we conducted the search only for the three patches with the highest interface propensity near the center of the binding site. For design, the number of targeted sites overlapping with the protein-ligand neosurface depended on the solvent-accessible surface area of each ligand to ensure that all the ligand-exposed surface was covered during the complementary motif search: 1 for PDF1:Actinonin, 2 for DB3:Progesterone and 3 for Bcl2:Venetoclax.

*Binding seed identification*

The fingerprints of the predicted 12 Å (geodesic radius) patches comprising both protein target and bound drug were used to find a complementary fingerprint in the MaSIF-seed database [250] which consists of ~640'000 continuous structural fragments (seeds) amounting to 402 million surface

patches/fingerprints. The seed database covers distinct secondary structures with approximately 390'000 sheet-based and 250'000 helical motifs respectively. The MaSIF-search algorithm was trained to make patch fingerprints similar for interacting patches and dissimilar for non-interacting patches. Seeds with interface propensity scores above the defined threshold and with fingerprint distances (Euclidean distance between target and seed fingerprint) below the defined thresholds were selected. In a second-stage alignment and scoring using the RANSAC algorithm, seeds were selected based on interface post-alignment (IPA) score. Cutoffs used for the seed selection are summarized in Supplementary Table S4.1.

*Scoring aligned structures*

We consider two descriptor-based post-alignment scores. The descriptor distance score (DDS) is a simple heuristic that aggregates descriptor distances across the predicted binding interface. DDS is based on the squared Euclidean distances between interacting patches on both sides of the interface. Two patches are considered interacting with each other if their center points are less than 1.5Å apart. The descriptor distance score is computed according to the following formula

$$DDS = \sum_{i} \frac{1}{|| \, binder\_desc(i) \, - \, target\_desc(NN(i)) \, ||^2}$$

where *i* indexes interacting patches of the first protein and *NN(i)* returns the index of the spatially nearest neighbor on the other protein. Higher scores mean higher complementarity.

The interface post-alignment score (IPA) is computed by a neural network that was trained to discriminate between near-native and high-RMSD poses of docked proteins [138]. The inputs of this predictor are 3D Euclidean distances, descriptor distances and dot products between surface normals of up to 200 pairs of corresponding patches at the predicted interface. It outputs values between zero and one where larger values indicate higher confidence in the presented interface.

*Computational binder recovery benchmark*

The binder recovery experiment was performed for 14 known ligand-induced protein complexes, where both proteins involved in the interaction are considered as separate items, resulting in 28 search queries. Additionally, we included 200 decoys (based on 100 PPIs) in the database. All benchmark complexes and decoys are listed in Supplementary Table S4.4. After triangulating and featurizing all protein surfaces with and without ligands, we screen the database and dock candidates analogously to the binding seed search. Here, we assume the location of the binding site on the target protein is known and select the three surface vertices with the largest predicted surface propensity within 10Å of the center of this site as input patches. The center of the binding site was approximated with a simple heuristic. We first identify interface atoms as those within 4Å of any atom from the binding protein in the original complex structure. This can and typically will include atoms belonging to the small molecule. Then we define the average of the coordinates of all interface atoms of the target protein as the center of the binding site. Furthermore, we declare a binder correctly recovered if its interface

RMSD (iRMSD) compared to the ground truth structure of the same protein is less than 5Å, where iRMSD considers only heavy atoms in the immediate vicinity of the target protein (<5Å).

*Seed and interface refinement*

To optimize binding energy of the seed for the target complex, seeds were refined using a FastDesign protocol on Rosetta [223] with a penalty for buried unsatisfied polar atoms in the scoring function [137]. Refined seeds were then selected based on the computed binding energy (ddG), shape complementarity, number of interface hydrogen bonds, number of buried unsatisfied polar atoms and number of atoms in contact with the small molecule. Beta sheet-based motifs making >33% contact with the target complex using loop regions were discarded. Moreover, uniqueness of each seed was assessed by doing a pairwise alignment of the hotspot residues. For seeds showing >70% hotspot identity with another seed, only the one with the best surface-normalized ddG was kept.

*Seed grafting and computational design*

Selected seeds were subsequently grafted with a Rosetta MotifGraft [64] protocol for stabilizing the binding motif and bringing additional contacts with the target complex. Each seed was match with a database of ~6500 small protein scaffolds (<90 amino acids) originating from small globular monomeric protein from the protein data bank (PDB) [117] and four computationally designed miniprotein databases that were experimentally validated [74–76,79]. Prior grafting, seeds were cropped to the minimum number of residues making contact with the target, and loop motifs were removed from beta sheet-based seeds for optimizing the grafting success rate. Once grafting was performed, scaffolds underwent sequence optimization using a FastDesign protocol on Rosetta with a penalty for buried unsatisfied polar atoms in the scoring function. Final designs were selected based on the computed binding energy (ddG), shape complementarity, number of interface hydrogen bonds and count of buried unsatisfied polar atoms. A similar number of designs per seed was ensured by setting dynamic cutoffs of these metrics adjusted for each seed.

*Library screening*

For each target complex, around ~2000 designs were reverse-translated into DNA and purchased from Twist Bioscience as oligo pools with 18bp homology overhangs. Oligo pools underwent two rounds of PCR : i) for the amplification of the library using the 18bp overhangs and ii) for adding 45bp homology with the yeast display vector (57.5 °C annealing for 30 s, 72 °C extension time for 1 min, 15 cycles). EBY-100 yeast were transformed by electroporation using the amplified inserts and linearized HA-tagged pCTcon2 vector as described previously [111]. A similar approach was done for site-saturation mutagenesis (SSM) library of single designs. Transformed yeast cells were grown in minimal glucose medium (SDCAA) medium at 30°C and induced with minimal galactose medium (SGCAA) medium overnight prior sorting.

*Yeast surface display of single designs*

Genes encoding for single designs were purchased from Twist Bioscience with a ~25bp homology overhang for cloning. Each design was cloned into HA-tagged pCTcon2 plasmid using Gibson assembly and transformed into XL10-Gold or HB101 bacteria for DNA production. The purified and sequence-approved DNA was then used to transform competent EBY-100 yeast using the Frozen-EZ Yeast Transformation II Kit (Zymo Research). As for libraries, transformed yeast cells were grown in minimal glucose medium (SDCAA) medium at 30°C and induced with minimal galactose medium (SGCAA) medium overnight prior flow cytometry analysis.

*Flow cytometry analysis and sorting*

Induced yeast cells were washed with PBS supplemented with 0.1% BSA and then labeled with the respective binding target for 2 hours at 4°C. Prior to labeling, protein-drug complexes were pre-incubated at room temperature for 5 min with a 1:5-10 ratio. Cells were then washed and labeled with a FITC-conjugated goat anti-HA tag antibody (Bethyl; ref: A190-138F; display tag; 1:100 dilution) and a PE-conjugated goat anti-human Fc antibody (Invitrogen; ref: 12-4317-87; binding tag; 1:100 dilution) for 30 min at 4°C. Cells were washed, resuspended in an appropriate volume of buffer and analyzed on a Gallios flow cytometer (Beckman Coulter), or sorted with a Sony SH800 cell sorter. Kaluza software (Beckman Coulter, v.1.1.20388.18228) and LE-SH800SZFCPL Cell Sorter (Sony, v.2.1.5) were respectively used for the data acquisition. In the case of cell sorting, each designed library was sorted for binding and non-binding populations separately. Flow cytometry data were then analyzed using FlowJo (BD Biosciences, v.10.8.1).

*Library sequencing*

Sorted yeast were cultured and plasmids encoding protein designs were extracted using the Zymoprep Yeast Plasmid Miniprep II (Zymo Research) following the manufacturer's protocol.The sequence of interest was then amplified by PCR with vector-specific primers flanking the protein design gene. A second PCR was performed to add Illumina adapters and Nextera barcodes, and the PCR product was desalted and purified using the Qiaquick PCR purification kit (Qiagen). Illumina MiSeq system with 500 cycles was used for the next generation sequencing. Around 0.8-1.2 millions reads per sample were obtained, translated into the appropriate reading frame and matched with expected input sequences from the libraries. The enrichment of each design was calculated by normalizing the counts in the binding population with the counts in the non-binding populations. Hits were identified if the enrichment was >10-fold and the number of counts in the binding population was >10'000.

*Protein expression and purification*

A list of protein sequences can be found in Supplementary Table S4.5. Genes encoding the 6xHis-tagged and/or human Fc-tagged protein of interest were purchased to Twist Bioscience cloned into pET11 (bacteria vector) or pHLSec (mammalian vector) by Gibson assembly and transformed into XL10-Gold or HB101 bacteria. Plasmids were extracted using a GeneJET plasmid Miniprep kit

(Thermofisher, for bacteria vector) or a PureLink Fast Low-Endotoxin Midi plasmid purification kit (Invitrogen, for mammalian vector) and checked by Sanger sequencing. Proteins were purified by bacteria or mammalian expression systems. Mammalian expressions were performed using the Expi293 expression system (ThermoFisher; ref: A14635). Supernatants were collected after 6 days, filtered and purified as explained below. For bacteria expression, BL21(DE3) or T7 Express Competent E. coli were transformed with the plasmid of interest and grown as a pre-culture overnight. Pre-cultures were inoculated 1:50 in Terrific Broth medium and incubated at 37°C until they reached a density ~0.7 at OD600. Then, bacteria were induced with 1mM IPTG and incubated overnight at 18-20°C. Cells were collected by centrifugation at 4000g for 10 min, resuspended in lysis buffer (50 mM Tris, pH 7.5, 500 mM NaCl, 5% glycerol, 1 mg ml$^{-1}$ lysozyme, 1 mM PMSF and 1 µg ml$^{-1}$ DNase) and lysed by sonication. Lysates were then clarified by centrifugation at 30'000g for 30 min and filtered.

All 6xHis-tagged protein are purified using the ÄKTA pure system (GE healthcare) Ni-NTA HisTrap affinity column followed by a size exclusion chromatography on a Superdex HiLoad 16/600 75pg or 200pg depending on the size of the protein. All proteins were concentrated in PBS as a final buffer.

*Surface plasmon resonance*

Affinity measurements were done on a Biacore 8K (GE Healthcare) using HBS-EP+ as a running buffer (10 mM HEPES at pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.005% v/v Surfactant P20, GE Healthcare). All proteins were immobilized on a CM5 chip (GE Healthcare #29104988) via amine coupling to reach 500–1000 response units (RU). Analytes were then injected in serial dilutions using the running buffer. The flow rate was 30 µL/min for a contact time of 120 s followed by 400 s of dissociation time. SPR data were either fit with a 1:1 Langmuir binding model within the Biacore 8K analysis software (GE Healthcare #29310604) or done in steady-state affinity mode by reporting the relative RU for each concentration.

*Biolayer interferometry*

Biolayer Interferometry (BLI) measurements were performed on the Gator BLI system using the GatorOne software (Gator Bio, v.2.7.3.0728). Running buffer consisted of 500mM NaCl and 50mM Tris pH 7.5 or HPS-P+ buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 1µM NiSO4, 0.005% v/v Surfactant P20, GE Healthcare), supplemented by 100nM Venetoclax or 5µM Actinonin if needed. Fc-tagged proteins were immobilized at a concentration of 7µg/ml on protein A probes (1.5 to 2.5nm immobilized) and dipped into serial dilutions of the ligand. Steady state responses were normalized with the maximum value and plotted using a nonlinear four-parameter curve fitting analysis.

*Grating-Coupled Interferometry*

Grating-Coupled Interferometry (GCI) measurements were performed on a Creoptix WAVE system (Malvern Panalytical) using the Creoptix WAVE control software (Malvern Panalytical, v. 4.5.18). Running buffer consisted of HPS-P+ buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 0.005% v/v Surfactant P20, GE Healthcare). All protein targets were immobilized on a 4PCH chip (Malvern Panalytical) via amine coupling to reach 7'000-10'000 pg/mm2 . An intermediate injection with 1µM NiSO4 was used

for PDF1 protein. S55746, 19-O-Benzoyl-Progesterone (OBz-Progesterone) and Tertbutyldimethylsilyl-Actinonin (TBDMS-Actinonin) were then injected sequentially as analytes at a concentration of 2, 2.5 and 5 μM respectively using the waveRAPID (Repeated Analyte Pulses of Increasing Duration) kinetic assay [270]. The flow rate was 100 μL/min for an injection duration of 25s followed by 300s of dissociation time for TBDMS-Actinonin, while an injection duration of 50s followed by 600s of dissociation time was used for S55746 and OBz-Progesterone. Measurements were either fitted with a 1:1 model for Bcl2:S55746 and PDF1:TBDMS-Actinonin, or with a mass transport model for BD3:OBz-Progesterone.

*Size-exclusion chromatography–multi-angle light scattering*

Size exclusion chromatography combined to multiangle light scattering device (miniDAWN TREOS, Wyatt) was performed to determine the molecular weight of the purified designs. The final concentration was approximately 1 mg/ml in PBS (pH 7.4), and 100 μl of the sample was injected into a Superdex 75 10/300 GL column (GE Healthcare) with a flow rate of 0.5 ml/min. UV280, refractive index (dRI) and light scattering signals were recorded. Molecular weight was determined using the ASTRA software (version 6.1, Wyatt).

*Circular dichroism*

Far-ultraviolet circular dichroism spectra were carried with a Chirascan spectrometer (AppliedPhotophysics). Protein samples were prepared diluted in PBS at a protein concentration 300 μg/ml and placed in 1 mm path-length cuvette. Wavelengths between 200 nm and 250 nm were recorded with a scanning speed of 20 nm min–1 and a response time of 0.125 s. All spectra were corrected for buffer absorption. Temperature ramping melts were performed from 20 to 90 °C with an increment of 2 °C $min^{-1}$. Thermal denaturation curves were plotted by the change of ellipticity at the global curve minimum. If possible, melting temperature ($T_m$) were determined after fitting the data with a sigmoid curve equation on GraphPad Prism.

*Cell transfection and induction*

Human Embryonic Kidney (HEK293T; RRID: CVCL_0063) cells were cultured in Dulbecco's Modified Eagle Medium (41966-029, Gibco) supplemented with 10% (v/v) FBS (A5256701, Gibco) and 1%(v/v) antibiotic penicillin/streptomycin (15140-122, Gibco). Cells were maintained at 37°C with 5% CO2 and passaged every two to three days at around 80% confluency. Cells were seeded into the inner 60 wells of a 96 well plate, at 10'000 cells per well, 24 hours prior to transfection. Cells were transfected by layering 50uL from a mixture of 330uL DMEM, 825ng-850ng total DNA, and 4.125ug PEI (24765-1, Polysciences) on top of the media in each well, enough for each 6 well column with a 10% extra margin, as described previously [100]. Cells were left to incubate overnight, for a minimum of 12 hours. The next morning media was replaced with fresh media including the respective dilutions of the inducing agent.

*Cellular detection assay*

For secreted NanoLuc assays, cells were plated on clear 96 well cell culture plates (655-180, Greiner Bio-One). The next day cells were transfected with STAT3 (100ng), STAT3-NanoLuc reporter (150ng), and either a single GEMS receptor chain containing Bcl2 or DBVen 1619_V2 (600ng) or both chains together (300ng each). The following day cells were induced with different dilutions of the inducing agent Venetoclax. After 24 hours of induction, 5uL media was transferred to a black 384 well plate (3820, Corning) and mixed with 5uL diluted substrate from the Nano-Glo Luciferase Assay kit (N1120, Promega). After gentle shaking, plates were measured on a Tecan Spark plate reader with an integration time of 1000ms.

For intracellular NanoLuc assays, cells were plated in black 96 well cell culture plates (655086, Greiner). The next day cells were transfected with either a single chain of PDF1-C-term-NanoLuc or DBAct553_1-N-term-NanoLuc (825ng) or both chains together (412.5ng each). The following day cells were induced with different dilutions of the inducing agent Actinonin. After 24 hours of induction, intracellular Nanoluciferase activity was measured using the Nano-Glo Live Cell Assay kit (N2012, Promega). Media was aspirated and replaced with 24uL RPMI Medium (52400-025, Gibco) containing 10% v/v FBS and 6uL diluted substrate was added to each well. After gentle shaking, plates were measured on a Tecan Spark plate reader with an integration time of 1000 ms. All cell-based fits presented in Figure 4.5 were calculated from technical replicates (n = 3) using a nonlinear four-parameter curve fitting analysis. All statistical analyses are based on a two way ANOVA with multiple comparisons. Data points represent technical replicates (n = 3) with mean and standard deviation.

*Cell-free reporter system*

The gene encoding the 6xHis-DBPro1156_2 protein fused to T7 RNA Polymerase (T7RNAP) was cloned into a pQE30 plasmid using Gibson assembly. The plasmid was then transformed into NEBExpress Iq competent E. coli (NEB; ref: C3037I) for protein expression. Bacteria were pre-cultured overnight and inoculated to a 500 ml LB-medium culture, grown until the OD600~0.7, and then induced with 0.1 mM IPTG for 3 hours. The cells were collected by centrifugation at 4000 g and lysed by sonication. Proteins were purified using Ni-NTA IMAC sepharose gravity columns.

The ZF438-DB3 scFv (VH/VL) fusion protein was expressed using a PURExpress kit from NEB (E6800S) with the addition of a disulfide bond enhancer (E6820S). The reaction volume was 10 µl, containing 4 µl of solution A, 3 µl of solution B, 0.4 µl of NEB disulfide bond enhancer 1, 0.4 µl of NEB disulfide bond enhancer 2, 2 µl of DNA template (10 ng/µl), and 0.2 µl of water. The reaction was incubated at 34 ˚C for 3 hours and used for the following reporter reaction.

PURExpress kit from NEB (E6800S) with disulfide bond enhancer (E6820S) was used to set up the mCherry reporter expression as well. The reporter-expressing reaction additionally includes 100 nM purified DBPro1156_2-T7RNAP and ZF438-DB3 scFv pre-expressed with PURExpress. DNA template for the mCherry gene is set to 4 nM, the mCherry gene is transcribed under the regulation of a truncated T7 promoter downstream of the zinc finger 438 protein binding site, which requires a zinc finger protein

for activating transcription. Progesterone was dissolved in 2% DMSO. 10 µl reactions with different conditions are loaded into a 384-well plate. The mCherry fluorescent intensity is measured on a BioTek Synergy H1 Multimode Reader (Agilent) with an excitation wavelength of 565 nm and an emission wavelength of 615 nm at 34 ˚C for 8 hours with 2-minute intervals. All cell-based fits presented in Figure 4.5 were calculated from technical replicates (n = 3) using a nonlinear four-parameter curve fitting analysis. All statistical analyses are based on a two way ANOVA with multiple comparisons. Data points represent technical replicates (n = 3) with mean and standard deviation.

*Protein purification for crystallography*

6xHis-tagged PDF1 from Pseudomonas aeruginosa and DBAct553_1 were expressed in E. coli (BL21 T7 Express). Amino acid sequences of both proteins are shown in Supplementary Table S4.5. For PDF1, cells were grown in Luria-Bertani (LB) medium supplemented with 100mM NiSO4 up to an OD600 of 0.7 at 37°C, induced with 1mM IPTG and continued growing overnight at 18°C. For DBAct553_2, cells were grown in auto-induction medium (AIM) up to OD600 of 0.7 at 37°C and then overnight at 18°C. Cells were collected by centrifugation at 4000g for 10 min, resuspended in lysis buffer (50 mM Tris, pH 7.5, 500 mM NaCl, 5% glycerol, 1 mg ml−1 lysozyme, 1 mM PMSF and 1 µg ml−1 DNase) and lysed by sonication. Lysates were then clarified by centrifugation at 30'000g for 30 min and filtered. Proteins were purified using the ÄKTA pure system (GE healthcare) Ni-NTA HisTrap affinity column followed by a size exclusion chromatography on a Superdex HiLoad 16/600 75pg with TBS (50mM Tris pH 7.5, 250mM NaCl, 10µM NiSO4) as a final buffer. PDF1, DBAct553_2 and Actinonin were mixed at a final concentration of 35µM, 105µM and 300µM respectively and incubated on ice for 1 hour. Proteins were then concentrated by centrifugation prior to crystallization.

*Crystallographic data collection and structure determination*

The Actinonin-bound PDF1:DBAct553_1 complex (5 mg/ml) was crystallized using the sitting drop vapor diffusion setup at 18°C with 200nl of protein and 200nl crystallization solution consisting of 0.2M sodium formate, 0.1M sodium phosphate pH 6.2, 20% (v/v) PEG and 10% (v/v) glycerol. Crystals were cryoprotected with 25% glycerol and flash-cooled in liquid nitrogen. Diffraction data were collected at a temperature of 100K at the European Synchrotron Radiation Facility (ESRF Grenoble, France). Raw data were processed and scaled with XDS, and then processed using the autoPROC package [212]. Phases were obtained by molecular replacement using the Phaser module of the Phenix package and a model from PDB 1LRY in complex with our designed binder DBAct553_1 [213]. Atomic model adjustment and refinement was completed using COOT and Phenix.refine [214,215]. Finally, MolProbity [216] was used to assess the quality of the refined model. Details of data collection and refinement statistics are shown in Supplementary Table S4.6.

*Cryo-EM preparation and data acquisition*

A chimeric DB3 Fab (see Supplementary Table S4.5) was produced using the Expi293 expression system from Thermo Fisher Scientific (A14635). An anti-kappa light chain Fab [271] (see Supplementary Table S4.5) was produced using the ExpiCHO-S cells (Thermo Fisher Scientific, ref: A29127) growing in a

ProCHO-5 medium (Lonza) supplemented with 2% DMSO. Supernatants were collected 6 and 7 days respectively after transfection, filtered and purified by Ni-NTA affinity chromatography followed by a size exclusion chromatography on a Superdex HiLoad 16/600 75pg. All proteins were concentrated in PBS as a final buffer. DBPro1156_2 was purified as indicated previously in the "protein expression and purification" section.

DB3 Fab, anti-kappa light chain Fab, DBPro1156_2 and progesterone were mixed with a molar ratio of 1 : 0.9 : 3 : 2 respectively, supplemented with 0.1% n-dodecyl-ß-D-maltoside (DDM) and concentrated to 3.87 mg/ml. Proteins were applied to a glow discharged 300-mesh holey carbon grid (Au 1.2/1.3 Quantifoil Micro Tools), blotted for 4 s at 95% humidity, 10 °C, plunge frozen in liquid ethane (Vitrobot, Thermo Fisher Scientific(TFS)) and stored in liquid nitrogen. Data collection was performed on a 300 kV FEI Titan Krios G4 microscope equipped with a FEI Falcon IV detector. Micrographs were recorded at a calibrated magnification of 120'000× with a pixel size of 0.658 Å and a nominal defocus ranging from -1.0 µm to -1.7 µm.

*Cryo-EM image processing*

Acquired cryo-EM data was processed (Supplementary Fig. S4.13) using cryoSPARC v4.4.1. Gain-corrected micrographs were imported, and micrographs with a resolution estimation worse than 5.5 Å were discarded after patch CTF estimation. Initial particles were picked using blob picker with 90-150 Å particle size. Particles were extracted with a box size of 360 × 360 pixels, downsampled to 140 x 140. After 2D classification, clean particles were used for *ab initio* 3D reconstruction. After several rounds of 2D and 3D classification, the class with most detailed features was re-extracted using full box size and subjected to non-uniform and local refinement to generate high-resolution reconstructions. The local resolution was calculated and visualized using ChimeraX [222].

For structure building, we used ColabFold [266] re-predictions of the anti-Kappa and DB3 Fabs, as well as the designed binder. Subsequent manual model adjustment and refinement was completed using Coot [214]. Atomic model refinement was performed using Phenix.real_space_refine [215]. quality of the refined model was assessed using MolProbity [216].

*Chemical synthesis*

All chemical reagents and solvents for synthesis were purchased from commercial suppliers (Sigma-Aldrich, Fluka, Acros) and were used without further purification or distillation. The composition of mixed solvents is given by the volume ratio (v/v). $^{1}$H nuclear magnetic resonance (NMR) spectra were recorded on a Bruker DPX 400 (400 MHz for 1H) with chemical shifts (δ) reported in ppm relative to the solvent residual signals (7.26 ppm for of CDCl3; 3.31 ppm for MeOD) (Supplementary Fig. S4.15). Coupling constants are reported in Hz. LC-MS was performed on a Shimadzu MS2020 connected to a Nexerra UHPLC system equipped with a Waters ACQUITY UPLC BEH Phenyl 1.7µm 2.1x50mm column. Buffer A: 0.05% HCOOH in $H_2O$ Buffer B: 0.05% HCOOH in acetonitrile. LC gradient: 10% to 90% B within 6.0 min with 0.5 ml/min flow. Preparative HPLC was performed on a Dionex system equipped with an UltiMate 3000 diode array detector for product visualization on a Waters

SymmetryPrep C18 column (7 µm, 7.8 x 300 mm). Buffer A: 0.1% v/v TFA in $H_2O$; Buffer B: acetonitrile. Gradient was from 25% to 90% B within 30 min with 3 ml/min flow.

*19-O-benzoylprogesterone*

19-hydroxyprogesterone (2.0 mg, 6.1 µmol, 1 eq.) was dissolved in pyridine (0.5 ml) and benzoyl chloride (0.9 ul, 7.9 µmol, 1.3 eq) was added. The reaction mixture was stirred for 3h. LC-MS analysis showed reaction completion and 10 µl methanol were added. After 30 minutes, the solvents were evaporated under reduced pressure. The residue was dissolved in a minimum of acetonitrile and subjected to preparative HPLC. The fractions containing the product were pooled and lyophilized. Yield: 1.1 mg (41%). $^1$H NMR (400 MHz, CDCl3) δ 7.89 (d, J = 8.4 Hz, 2H), 7.56 (t, J = 7.4 Hz, 1H), 7.42 (t, J = 7.8 Hz, 2H), 5.98 (s, 1H), 4.81 (d, J = 11.3 Hz, 1H), 4.46 (d, J = 11.3 Hz, 1H), 2.68 (ddd, J = 17.0, 13.8, 5.9 Hz, 1H), 2.57 – 2.32 (m, 4H), 2.26 – 2.06 (m, 5H), 2.03 – 1.63 (m, 6H), 1.55 – 1.37 (m, 2H), 1.36 – 1.06 (m, 5H), 0.69 (s, 3H). HRMS (ESI/QTOF) m/z: $[M+H]^+$ Calcd for $C_{28}H_{35}O_4^+$ 435.2530; Found 435.2528.

*TBDMS-Actinonin*

Actinonin (2.0 mg, 5.2 µmol, 1 eq.) and 4-dimethylaminopyridine (3.8 mg, 31.2 µmol, 6 eq.) were suspended in DCM (0.5 ml). TBDMS-Cl (2.5 mg, 16.6 µmol, 3.2 eq.) was added and the reaction was stirred for 5h at r.t. The solvent was evaporated under reduced pressure, the residue was dissolved in MeOH (0.5 ml), water (50 µl) was added and the reaction was heated to 60°C for 5h. The solvents were evaporated again, the residue was dissolved in a minimum of DCM and subjected to preparative TLC using DCM/MeOH 9:1 as the eluent. Yield: 2.0 mg (77%). $^1$H NMR (400 MHz, MeOD) δ 4.38 (d, J = 8.5 Hz, 1H), 4.13 (s, 1H), 3.89 (dt, J = 10.0, 6.8 Hz, 1H), 3.79 (dd, J = 9.9, 5.3 Hz, 1H), 3.68 (dd, J = 9.9, 2.8 Hz, 1H), 3.63 – 3.42 (m, 1H), 2.83 – 2.75 (m, 1H), 2.34 (dd, J = 14.5, 8.0 Hz, 1H), 2.24 – 1.84 (m, 6H), 1.67 – 1.48 (m, 1H), 1.46 – 1.18 (m, 6H), 1.02 – 0.94 (m, 7H), 0.93 – 0.86 (m, 12H), 0.07 (s, 3H), 0.05 (s, 3H). HRMS (ESI/QTOF) m/z: $[M+Na]^+$ Calcd for $C_{25}H_{49}N_3NaO_5Si^+$ 522.3334; Found 522.3342.

*Data and material availability*

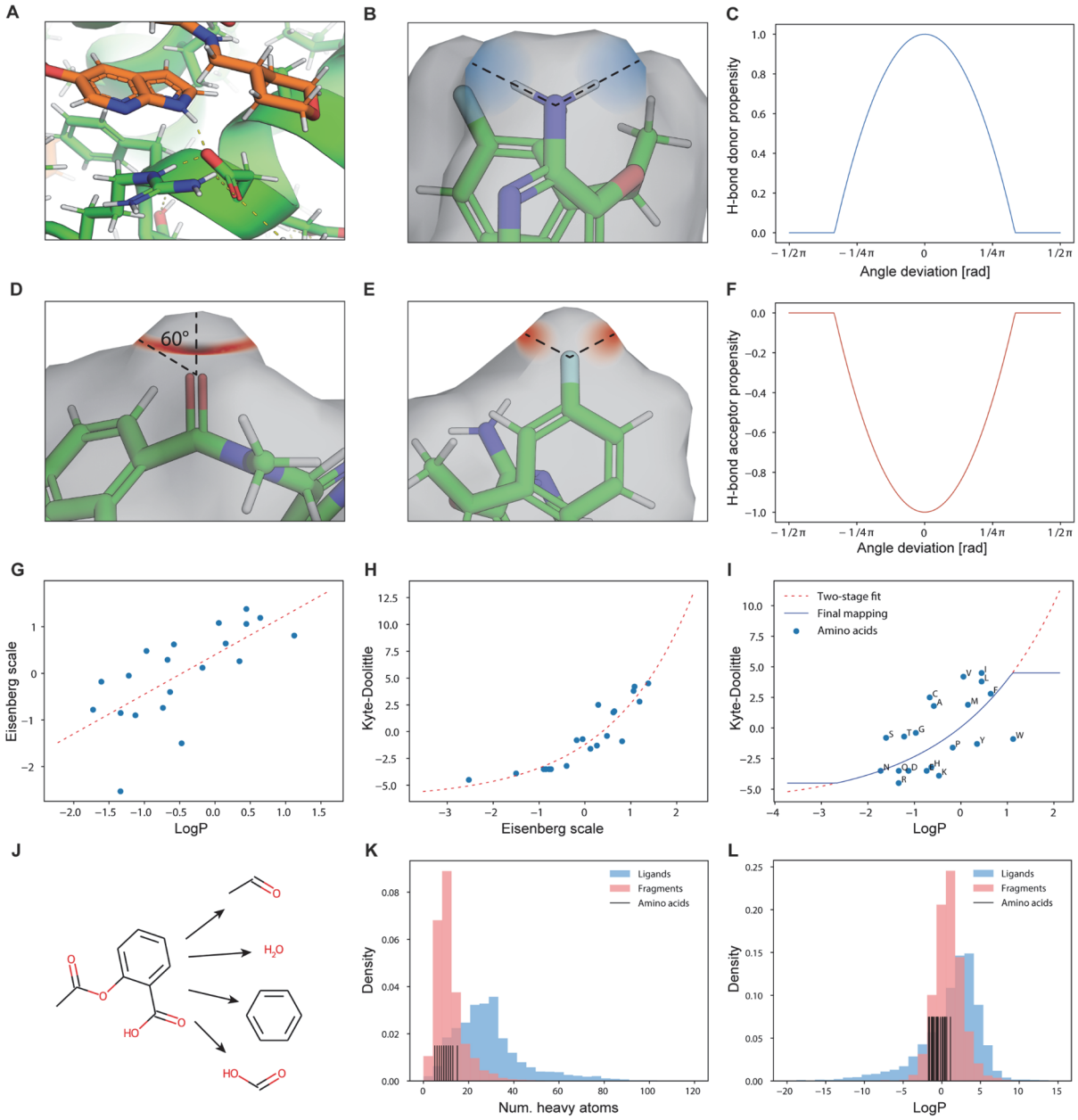
Crystal structure of DBAct553_2 in complex with Actinonin-bound PDF1 has been deposited at the PDB under the accession code 8S1X (DOI: https://doi.org/10.2210/pdb8S1X/pdb). MaSIF-neosurf and the Rosetta design scripts are available on GitHub (https://github.com/LPDI-EPFL/masif-neosurf). The scaffold database generated for grafting the seed provided by MaSIF-neosurf is partly available at Zenodo (https://zenodo.org/records/7643697#.Y-z533ZKhaQ) and partly on Github (https://github.com/strauchlab/DBP and https://github.com/strauchlab/scaffold_design/). All other data needed to evaluate the conclusions in this paper are present either in the main text or the supplementary materials.

*Acknowledgments*

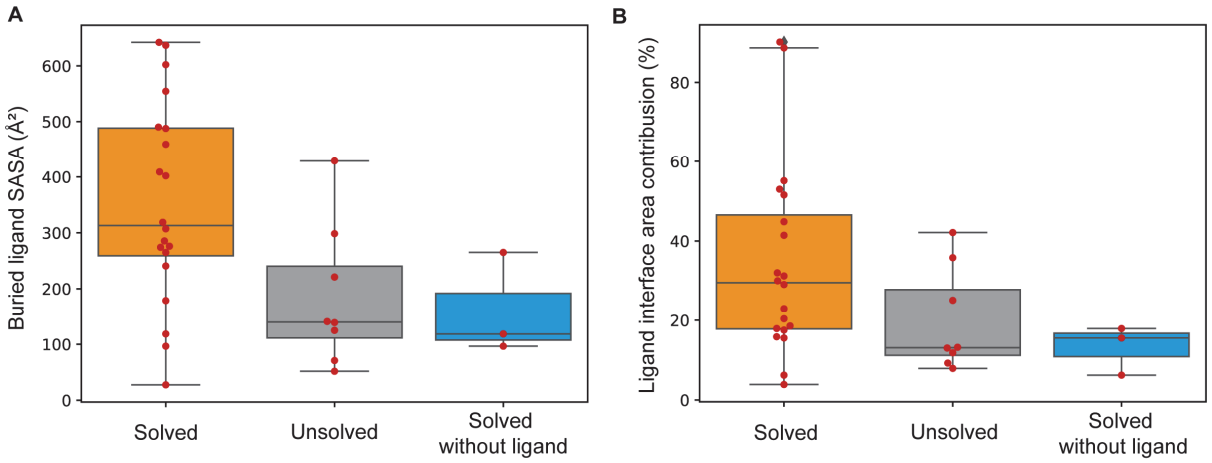
We would like to thank the staff at PTPSP at EPFL, Florence Pojer, Kelvin Lau, Amédé Larabi, Laurence Durrer and Soraya Quinche for their advice on the biophysical characterization of proteins and their work for the structural validations; the staff at the Dubochet Center for Imaging (DCI) in Lausanne for

cryo-EM data collection and processing; SCITAS at EPFL for support in the computational simulations; the staff at GECF for assistance with deep sequencing and members of FCCS for assistance in FACS. We also thank Dr. Andrea Moretti from Malvern Panalytical for his support to run the Grating-Coupled Interferometry on Creoptix Wave and giving access to the instrument. We express our gratitude to Prof. Eva-Maria Strauss for providing de novo hyperstable proteins for our scaffold database. Finally, we thank Prof. Nicolas Thomä and Dr. Kelvin Lau for their feedback on the manuscript.

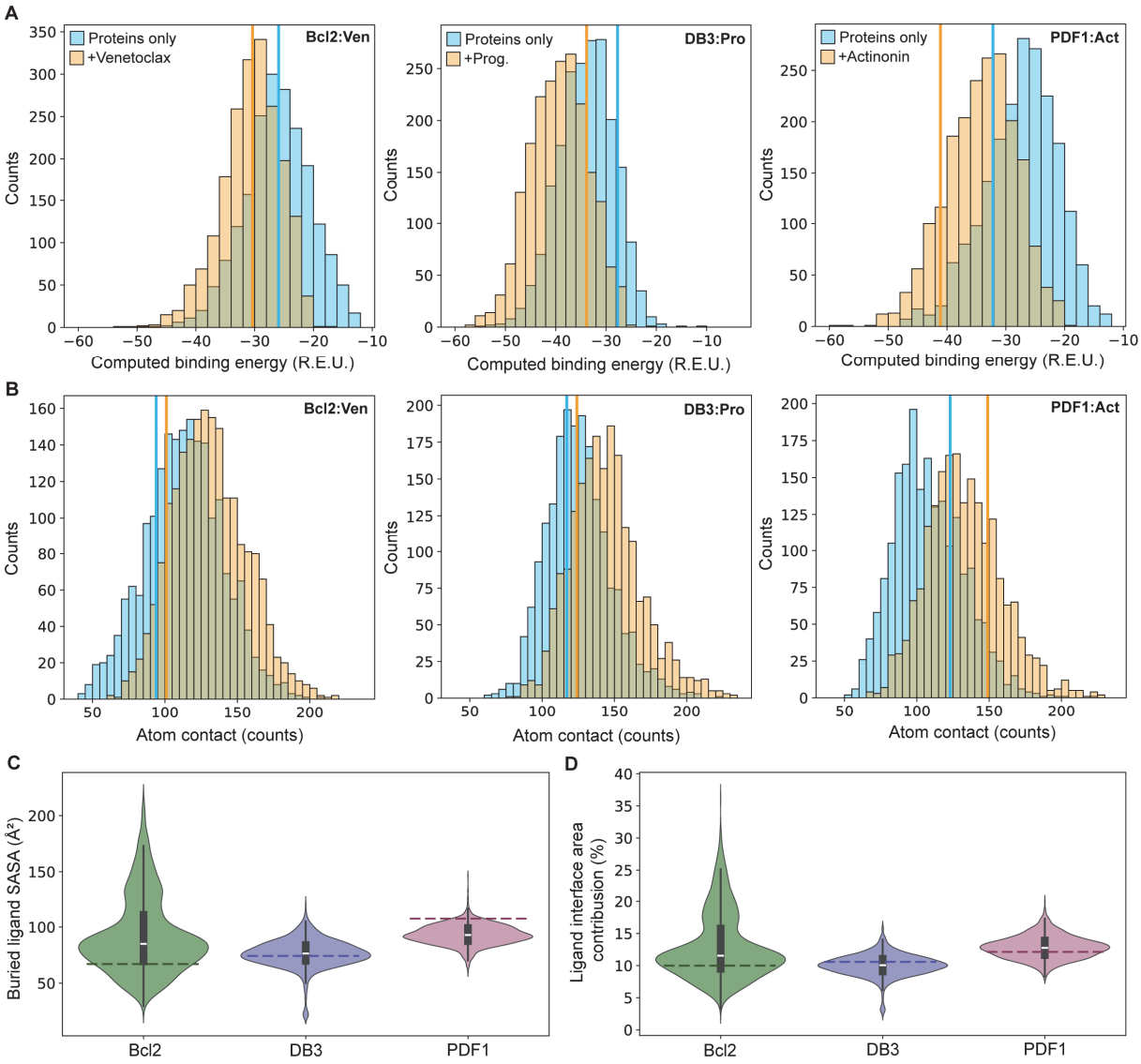
## 4.6 Supplementary materials



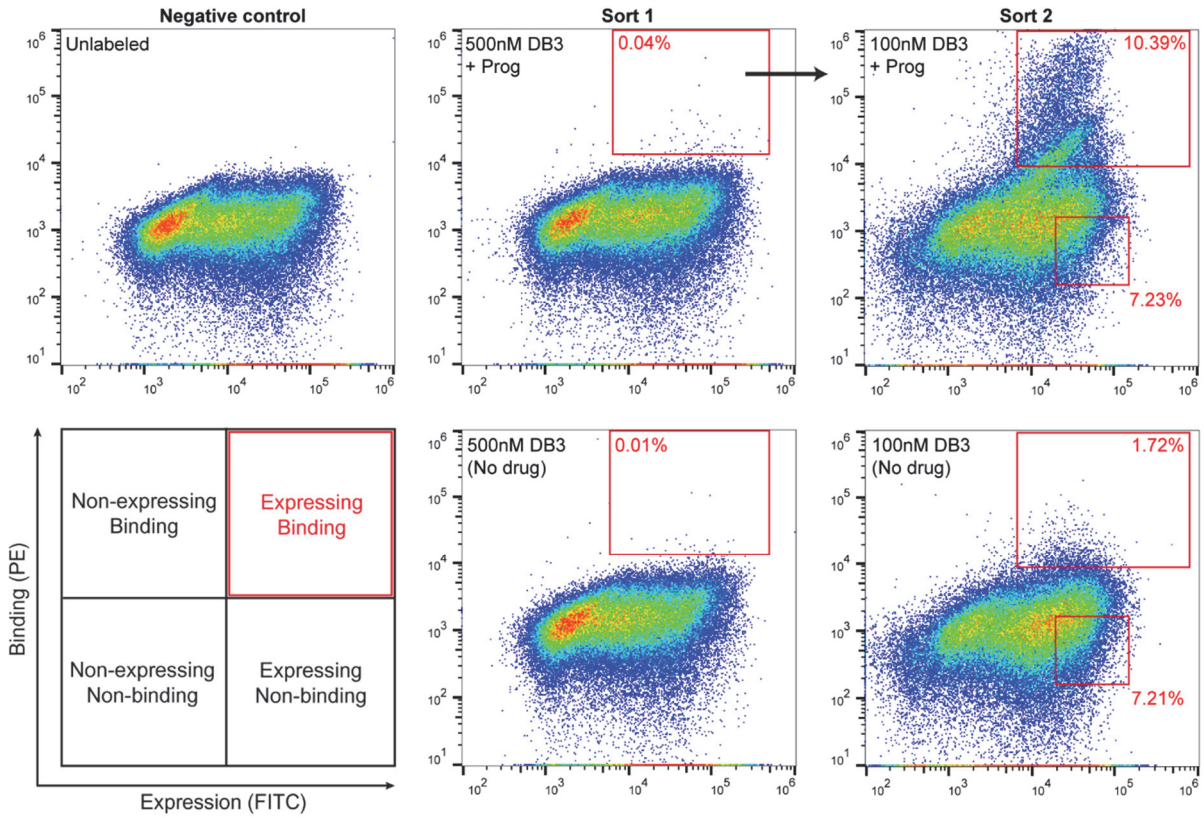

**Supplementary Figure S 4.1: MaSIF feature computation for small molecule ligands. A-F.** Hydrogen bond propensity is assigned in a direction-dependent manner [196]. Surface points are assigned positive (donor) or negative (acceptor) values based on their distance to the ideal direction (C, F). The optimal position for an acceptor can either lie anywhere on a cone (D) or in one or more unique directions (E). **G-I.** For hydropathy, we convert computational LogP values to the protein-specific Kyte-Doolittle (KD) scale required by MaSIF using the Eisenberg scale as an intermediate. We also restrict the outputs to be between the minimum and maximum KD values after the mapping leading to the relationship shown in panel (I). **J-L.** Ligands were fragmented into smaller objects (J). The reason is that most ligands are significantly larger than any amino acid (K) and exhibit more extreme LogP values (L). We therefore compute hydrophobicity values based on fragments. This procedure ensures that the new hydropathy feature remains "in-distribution" of the pre-trained MaSIF model.



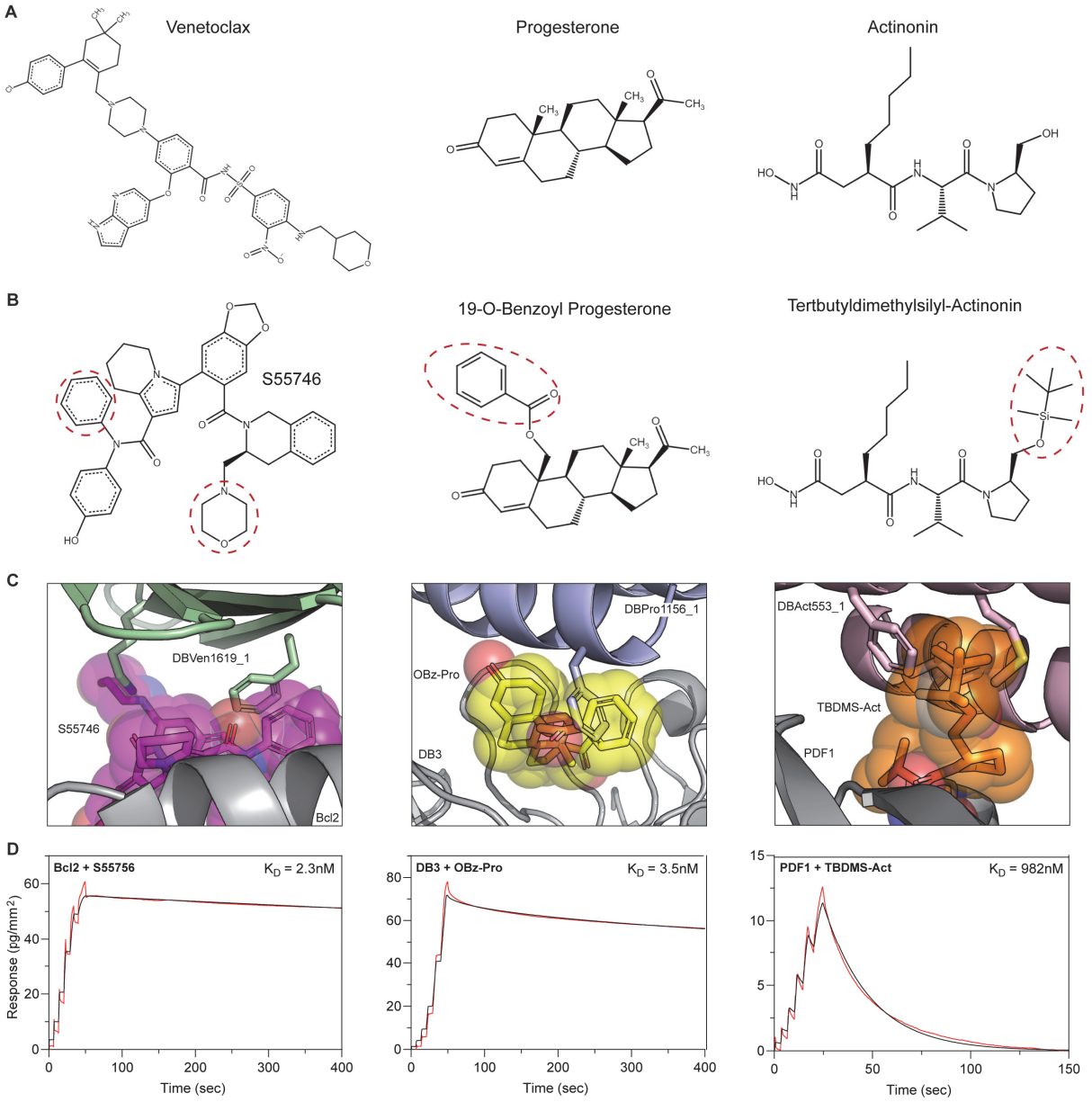

**Supplementary Figure S 4.2: Ligand interface area contribution in the benchmark dataset. A-B.** Buried solvent-accessible surface area (SASA) contribution of the ligand in absolute value (A) or in percentage of the total interface surface area (B) for the 28 protein-ligand complexes used in the benchmark dataset of known ternary complexes. Complexes were categorized based on the benchmark outcome, namely successfully solved complexes with the ligand (orange), unsolved complexes with the ligand (gray) or solved without the ligand (blue).



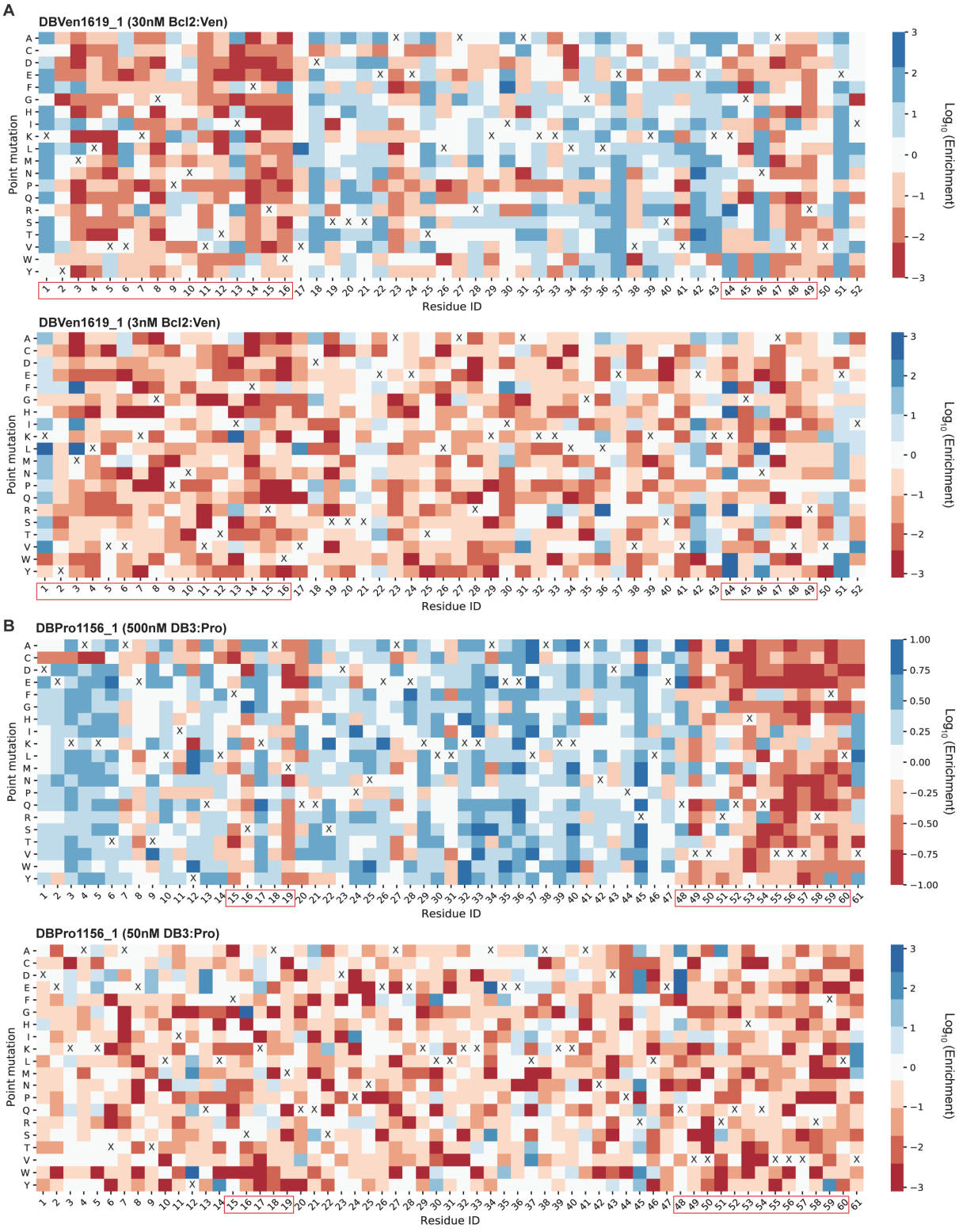

**Supplementary Figure S 4.3: Neosurface properties captured by the designed binders. A-B.** Computed binding energy (ΔΔG, A) and number of atomic contacts (B) for all designs targeting Bcl2:Venetoclax (left), DB3:Progesterone (middle) and PDF1:Actinonin (right) complexes. Calculations were done in absence (blue) and presence (orange) of the respective small molecules. Atom contacts were defined based on the Van der Waals radii (rVdW + 0.2Å tolerance) of each pair of atoms. Vertical lines represent the identified binder for each targeted complex. **C-D.** Buried solvent-accessible surface area (SASA) of the ligand (C) and ligand contribution with respect to the total buried SASA (D) for each target protein-ligand complex. Dashed lines represent the identified binder for each targeted complex.

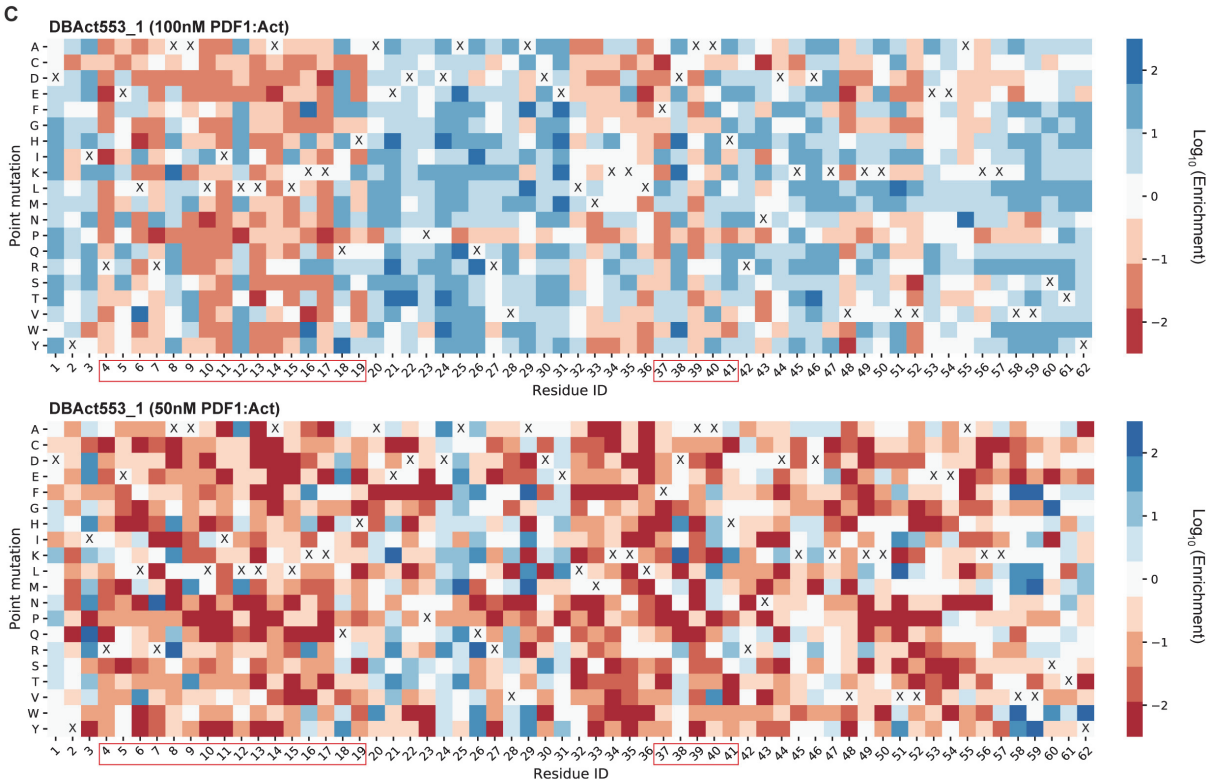



**Supplementary Figure S 4.4: Representative flow cytometry graphs of the binder screening.** Yeast surface display screening of the first and second sort of the library against DB3:Progesterone in presence (+Prog) and absence of the small molecule. Yeasts labeled with secondary antibodies but without any ligand were used as a negative control to set the gates. In sort 1, binding population from the selected gates was used for a second sort. In sort 2, yeasts were sorted for both binding (upper gate) and non-binding (lower gate) populations and used for next-generation sequencing.

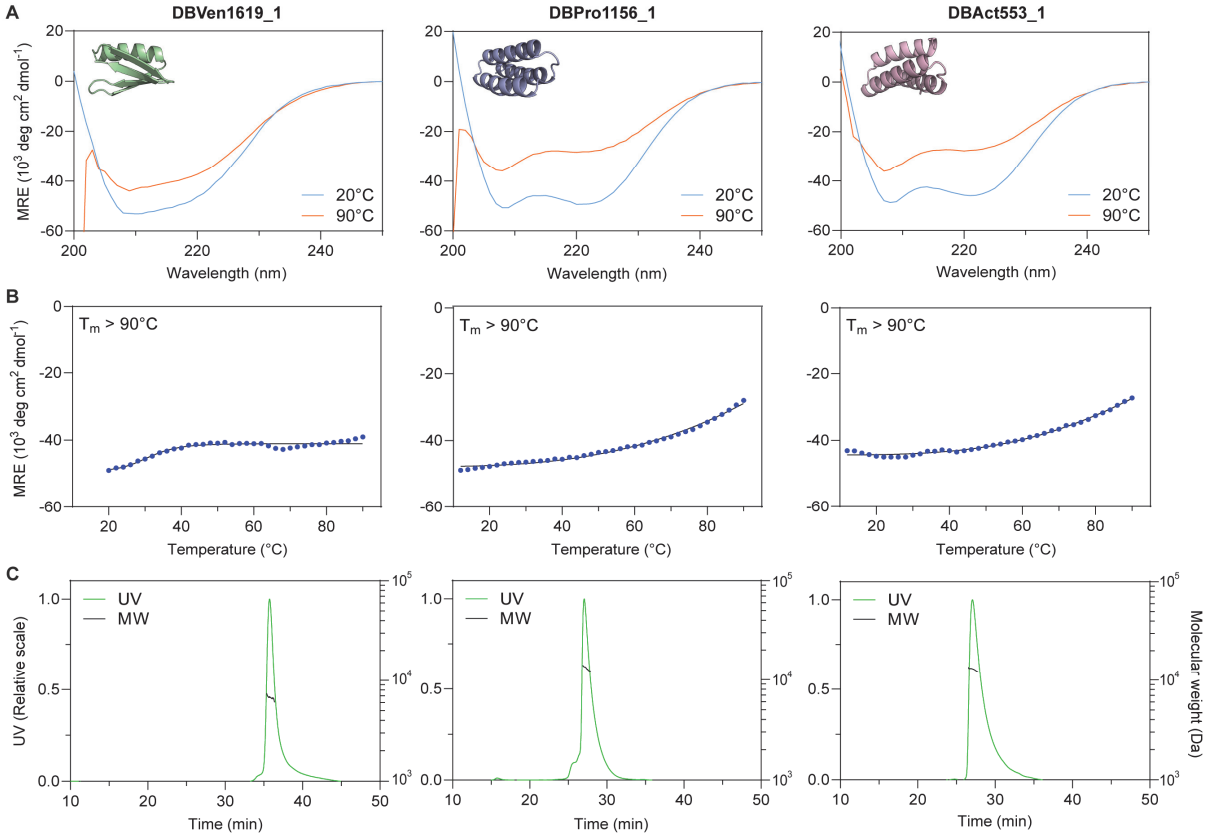



**Supplementary Figure S 4.5: Binding control of small molecules analogs. A.** Original small molecule used for each drug:protein complex. **B.** Small molecule derivative binding to the same target but introducing a clash with the designed binder. Steric clashes are indicated with red dashed circles. **C.** Crystal structure (S55746; PDB: 6GL8) or computational models (19-O-Benzoyl Progesterone and Tertbutyldimethylsilyl-Actinonin) of the small molecule analogs relaxed by Gnina(80) with their respective target protein in absence of the designed binders. Complexes were then overlapped with the computational model of the designed binders to identify clashing regions. **D.** WaveRAPID binding kinetics of each small molecule analog to their respective target protein performed by Grating-Coupled Interferometry (GCI). Measurements are indicated in red and fit curves in black.

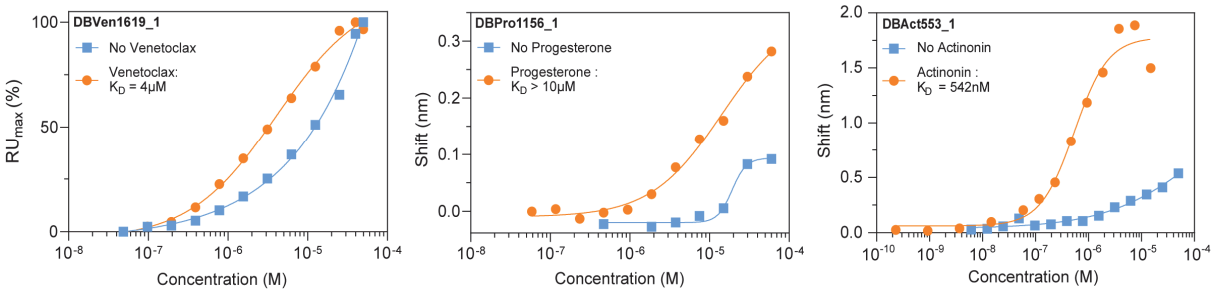

A
DBVen1619_1 (30nM Bcl2:Ven)
DBVen1619_1 (3nM Bcl2:Ven)
B
DBPro1156_1 (500nM DB3:Pro)
DBPro1156_1 (50nM DB3:Pro)
Point mutation
Residue ID
Log10 (Enrichment)



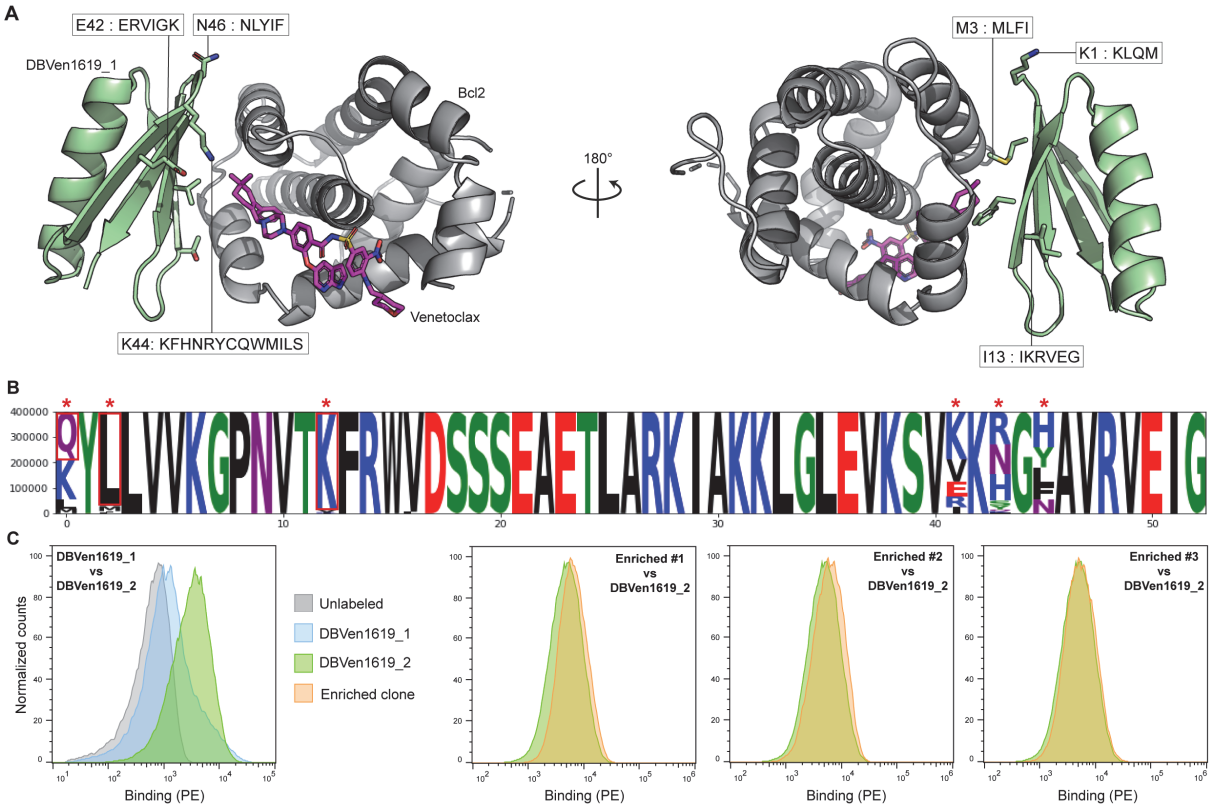

**Supplementary Figure S 4.6: Full data of the site-saturation mutagenesis. A-C.** Heatmaps represent the logarithmic value of the enrichment score (Counts in binding population divided by counts in non-binding population) of each mutation at every position. Sorting has been performed following the gating strategy presented in fig. S2. Native amino acids are marked with a cross (X) and near-interface positions are highlighted with a red box. Sortings were performed with a high concentration and a low concentration of target complex to focus on deleterious and beneficial mutations respectively. Site saturation mutagenesis were performed on DBVen1619_1 for the the Bcl2:Venetoclax complex (A), DBPro1156_1 for the DB3:Progesterone complex (B) and DBAct553_1 for the PDF1:Actinonin complex (C).



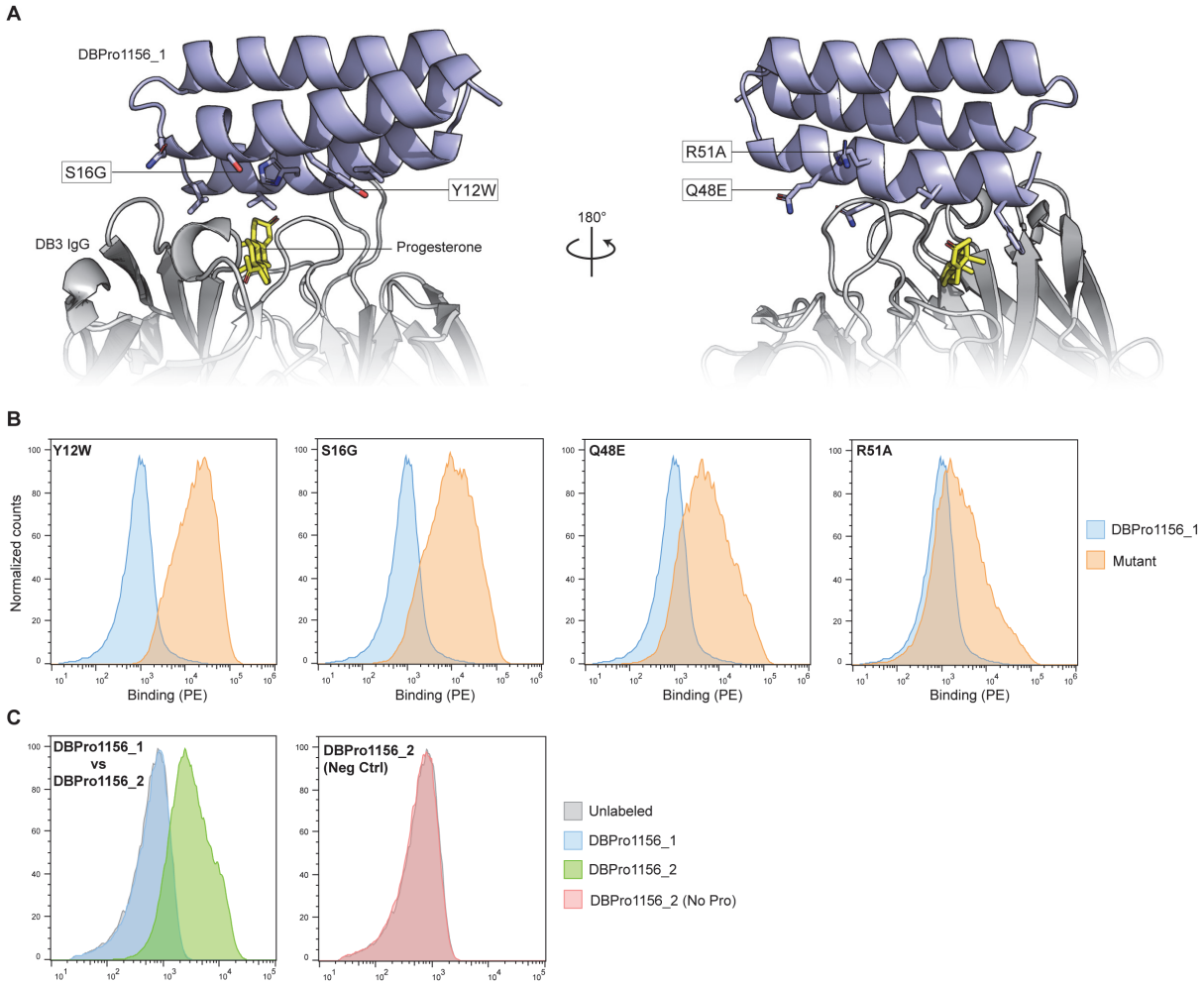

**Supplementary Figure S 4.7: Biophysical characterization of purified binders. A.** Protein folding of the purified binder measured by circular dichroism at 20°C (blue) or 90°C (orange). **B**. Thermal stability determined by measuring the ellipticity at 218nm at increasing temperature. **C.** Oligomeric state determined by size-exclusion multi-angle light scattering (SEC-MALS)



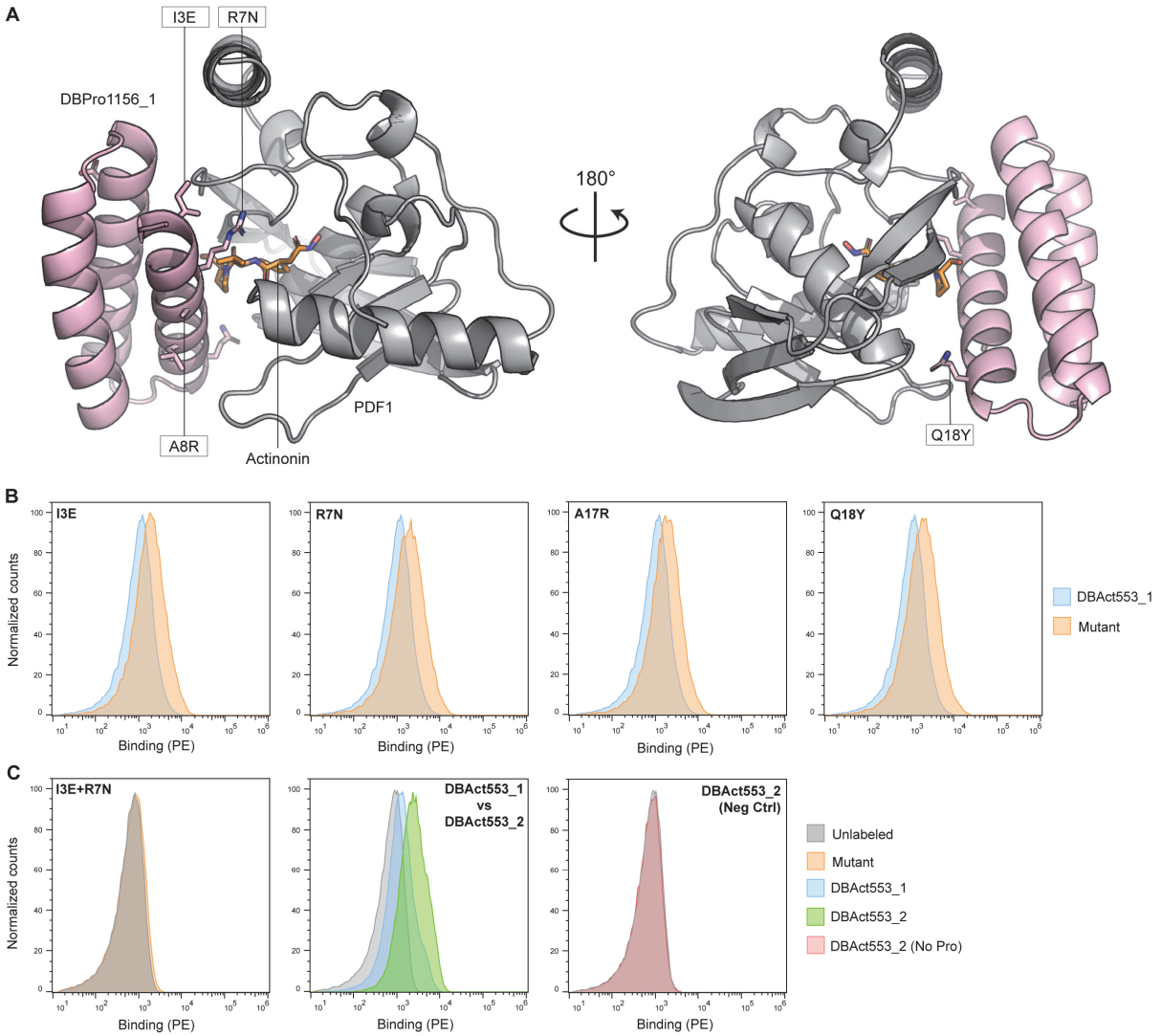

**Supplementary Figure S 4.8: Affinity measurements of first-generation purified binders.** Affinity measurement for DBVen1619_1, DBPro1156_1 and DBAct553_1 performed by surface plasmon resonance (DBVen1619_1) or biolayer interferometry (DBPro1156_1 and DBAct1156_1). Each measurement was performed in presence (orange) or absence (blue) of the respective small molecule. The fits were calculated using a nonlinear four-parameter curve fitting analysis.



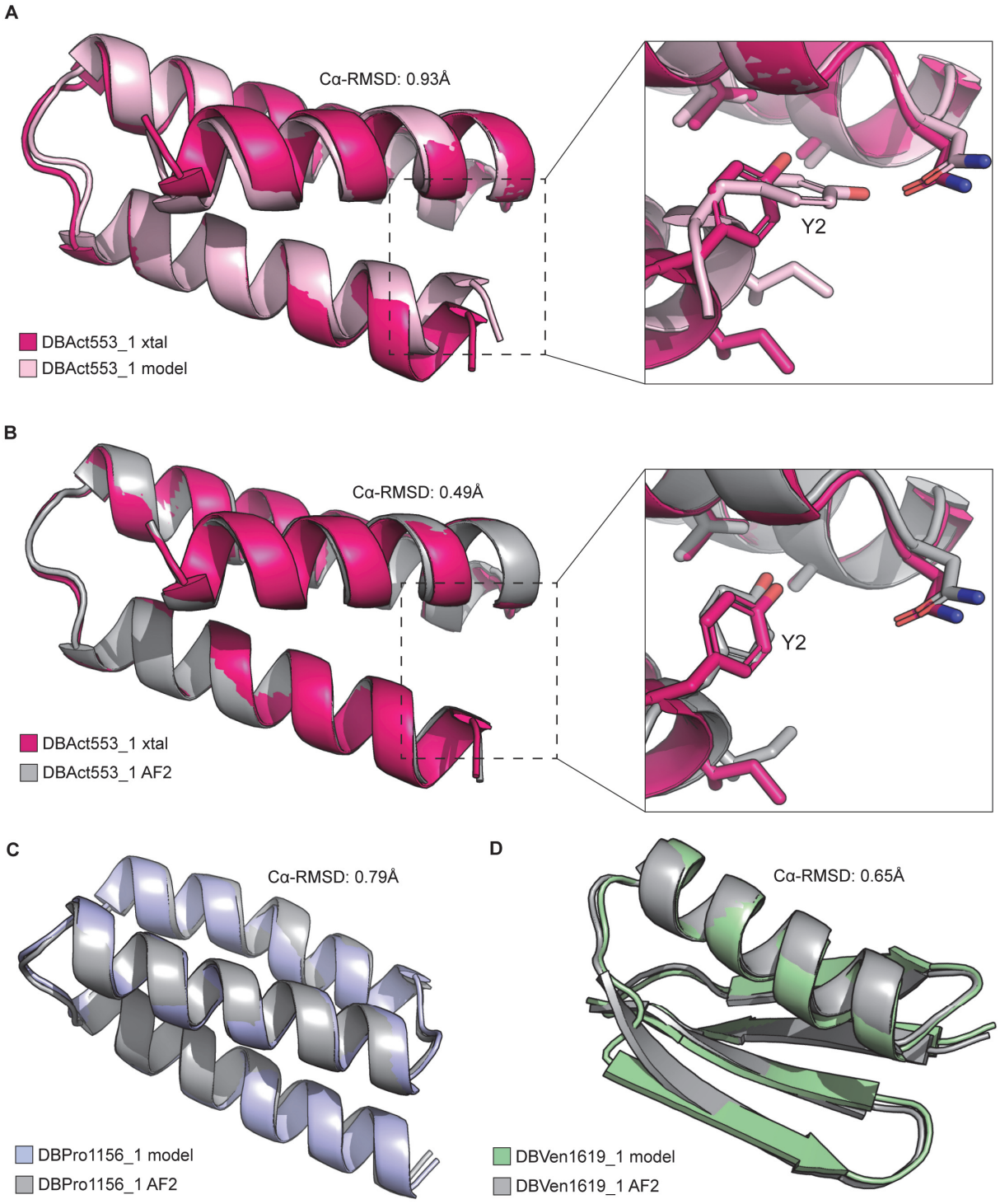

**Supplementary Figure S 4.9: Experimental optimization of DBVen1619. A.** Computational model of DBVen1619_1 (green) in complex with Bcl2 (gray) and Venetoclax (Magenta). Potential beneficial mutations obtained from site-saturation mutagenesis (SSM) data and subsequent degenerate codons are found in black boxes for each mutated position. **B.** Sequence logo plot of the combinatorial library sorted twice with yeast display. Mutated positions are highlighted with a red asterisk. Mutations selected to constitute DBVen1619_2 are highlighted with a red square. **C.** Comparison of unlabeled yeast (gray), or yeasts displaying DBVen1619_1 (blue), DBVen1619_2 (green, K1Q+M3L+I13K) or the top 3 most enriched sequences of the combinatorial library (orange). All yeasts were labeled with 3nM Bcl2:Venetoclax complex.



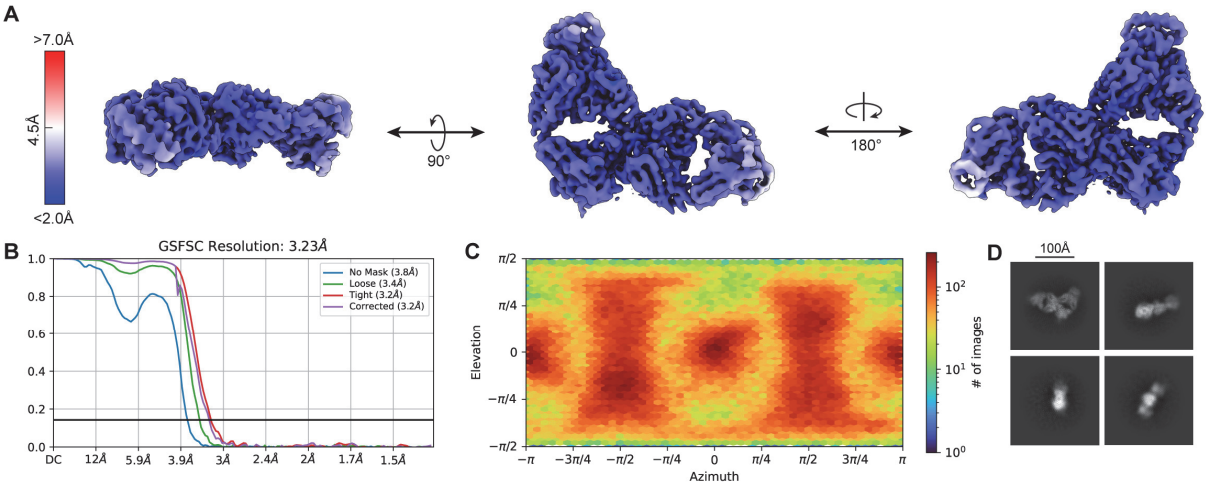

**Supplementary Figure S 4.10: Experimental optimization of DBPro1156. A.** Computational model of DBPro1156_1 (blue) in complex with DB3 IgG (gray) and Progesterone (yellow). Potential beneficial mutations obtained from site-saturation mutagenesis (SSM) data are found in black boxes for each mutated position**. B.** Binding signal measured by flowcytometry for yeasts displaying DBPro1156_1 (blue) or the corresponding mutant (orange). All yeasts were labeled with 50nM DB3:Progesterone complex. **C.** Binding signal measured by flowcytometry measured for yeasts displaying DBPro1156_1 (blue) compared to DBPro1156_2 (green, Y12W+S16G). Yeasts were labeled with 3nM DB3:Progesterone complex. Negative control performed on yeast displaying DBPro1156_2 labeled with DB3 only.



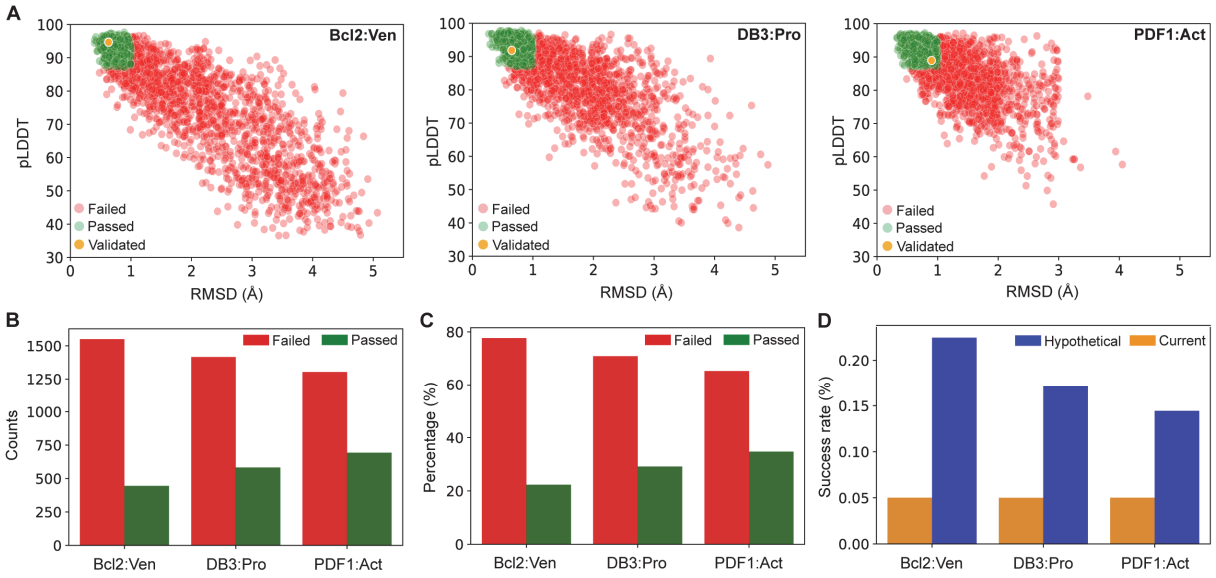

**Supplementary Figure S 4.11: Experimental optimization of DBAct553. A.** Computational model of DBAct553_1 (pink) in complex with PDF1 (gray) and Actinonin (orange). Potential beneficial mutations obtained from site-saturation mutagenesis (SSM) data are found in black boxes for each mutated position. **B.** Binding signal measured by flowcytometry for yeasts displaying DBAct553_1 (blue) or the corresponding mutant (orange). All yeasts were labeled with 10nM PDF1:Actinonin complex. **C.** Binding signal measured by flowcytometry measured for yeasts displaying DBAct553_1 (blue) compared to DBAct553_2 (green, R7N+A8R). Yeasts were labeled with 10nM PDF1:Actinonin complex. Negative control performed on yeast displaying DBAct553_2 labeled with DB3 only. Combination of mutants I3E+R7N was deleterious for binding as compared to unlabeled yeasts.



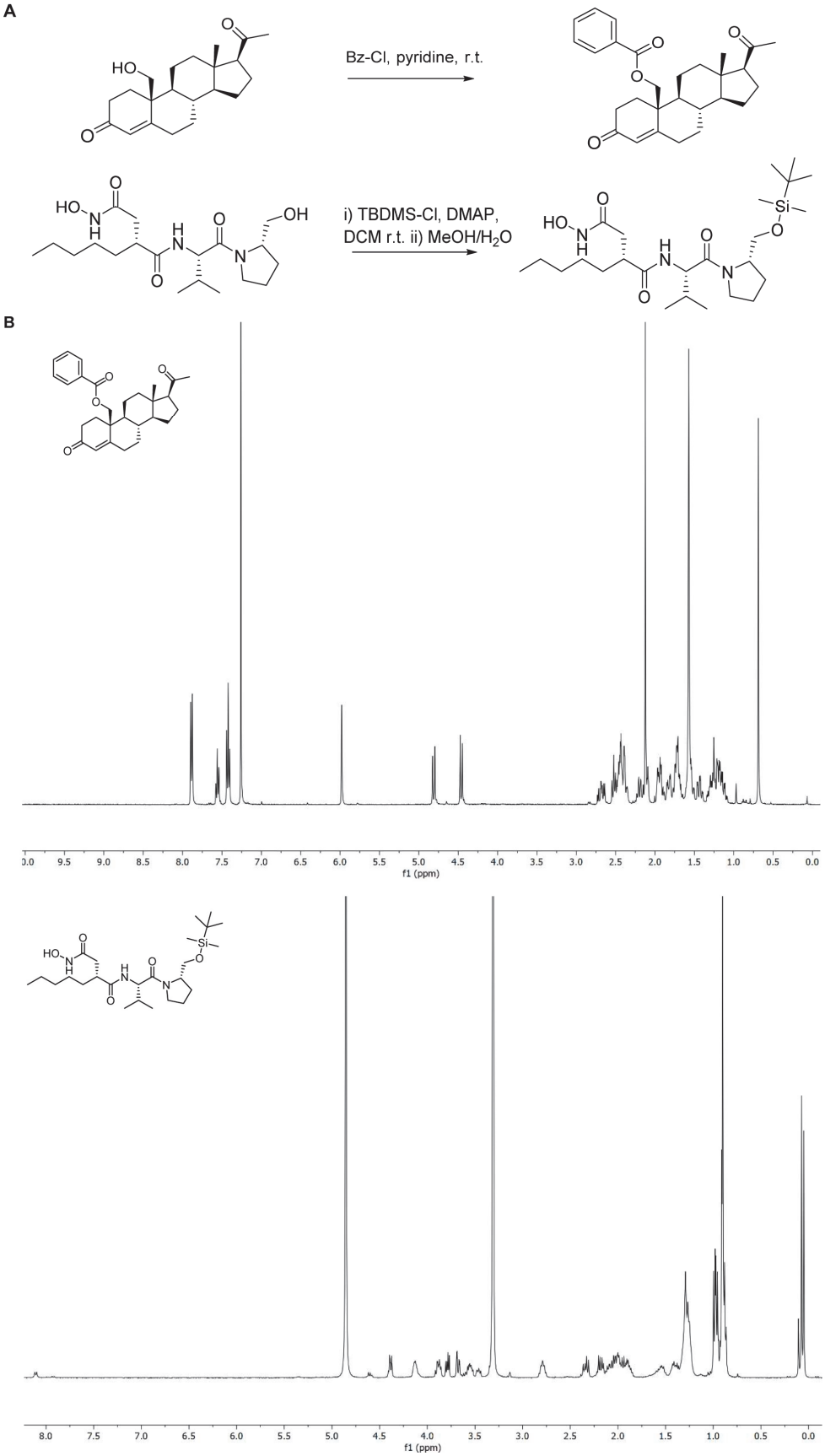

**Supplementary Figure S 4.12: Comparison between crystallographic data and AlphaFold2 predictions. A.** Computational model of DBAct553_1 (light pink) aligned with its crystal structure (magenta) with a close-up on tyrosine-2. **B.** AlphaFold2 (AF2) prediction of DBAct553_1 (gray) aligned with its crystal structure (magenta) with a close-up on tyrosine-2. **C-D.** Comparison between computational models of DBPro1156_1 (C) and DBVen1619_1 (D) and their respective AlphaFold2 prediction as monomers




**Supplementary Figure S 4.13: Details of Cryo-EM data processing for DBPro1153_2 in complex with DB3.** **A.** The cryo-EM map of anti-Kappa IgG:DB3:DBPro1156_2 used for model building. Views of the unsharpened cryo-EM density maps colored by local resolution. **B.** Gold-standard FSC curve with resolution cutoff indicated at 0.143. **C.** Particle distribution heatmap of the final reconstruction. **D.** Representative 2D classes of the anti-Kappa IgG:DB3:DBPro1156_2 complex.




**Supplementary Figure S 4.14: AlphaFold prediction and post-filtering of generated designs. A.** AlphaFold monomer prediction (single sequence mode) of the ~2000 designs generated against each drug:protein complex. Prediction confidence (pLDDT) and root mean square deviation (RMSD) from the computational models are plotted. Designs that would pass a strict filtering (RMSD ≤ 1Å and pLDDT ≥ 87) are colored in green, while ones that failed filtering are colored in red. Validated binders are colored in orange. **B-C.** Counts (B) and percentage (C) of generated designs that failed (red) or passed (green) the strict AlphaFold2 filtering. **D.** Experimental success rate obtained with the current data (orange) compared to the hypothetical success rate (blue) if a strict filtering with AlphaFold2 (RMSD ≤ 1Å and pLDDT ≥ 87) was used prior screening.




**Supplementary Figure S 4.15: Chemical synthesis and $^{1}$H NMR spectra validation. A.** Chemical synthesis reaction of 19-O-Benzoyl-Progesterone (OBz-Progesterone, top) and Tertbutyldimethylsilyl-Actinonin (TBDMS-Actinonin, bottom). **B.** $^{1}$H NMR spectra of OBz-Progesterone (top) and TBDMS-Actinonin (bottom).

**Supplementary Table S 4.1: Metrics and cutoffs for binder design with MaSIF-seed.**

| Target | Motif | Site | Interface cutoff | NN score cutoff | Descriptor distance cutoff | #seeds | #selected seeds | #designs (#grafted seeds) | #select designs (#seeds) | Total design |
|---|---|---|---|---|---|---|---|---|---|---|
| Bcl2: Ven | S | 1 | 0.65 | 0.9 | 2.0 | 1743 | 78 | 28396 (69) | 1456 (67) | 1995 |
| | | 2 | 0.65 | 0.87 | 2.2 | 1048 | 33 | 7485 (29) | 464 (27) | |
| | | 3 | 0.6 | 0.85 | 2.3 | 1012 | 11 | 1073 (8) | 75 (8) | |
| DB3: Pro | H | 1 | 0.75 | 0.9 | 1.8 | 995 | 49 | 160488 (49) | 975 (44) | 1998 |
| | | 2 | 0.65 | 0.9 | 2.1 | 1046 | 36 | 147940 (36) | 548 (34) | |
| | S | 1 | 0.8 | 0.9 | 1.7 | 1775 | 98 | 10097 (39) | 475 (37) | |
| PDF1: Act | H | 1 | 0.65 | 0.87 | 2.2 | 1272 | 74 | 56813 (67) | 1447 (66) | 1997 |
| | S | 1 | 0.65 | 0.85 | 2.3 | 1373 | 98 | 3711 (56) | 550 (55) | |

**Supplementary Table S 4.2: Deep sequencing analysis of FACS-enriched populations.**

| Target | Design | Original scaffold | Scaffold set | Topology | Binding counts | Non-binding counts | Enrichment |
|---|---|---|---|---|---|---|---|
| Bcl2 | DBVen1619 | 3hC_242_0001 | Ref30 | EEHEE | 459449 | 15561 | 1.47 |
| DB3 | DBPro1156 | bGC_85 | Ref29 | HHH | 134116 | 652 | 2.31 |
| PDF1 | DBAct553 | 3hC_605_0001 | Ref30 | HHH | 35195 | 557 | 1.80 |

**Supplementary Table S 4.3: Computational analysis of ligand contributions.** Summary of different metrics in absence or presence of ligands. The buried solvent-accessible surface area (SASA) of the protein-ligand target complex, the percentage of ligand contribution to this buried SASA, computed binding energy (ddG) in Rosetta Energy Unit (R.E.U.) and the number of atoms in contact with the target complex have been measured. Atom contacts were calculated based on the Van der Waals radii ($r_{VdW}$ + 0.2Å tolerance) of each pair of atoms. N/A: Not applicable.

| Design | Ligand | Target buried SASA [Å$^2$] | Ligand contribution | ddG [R.E.U] | ddG shift | Atom contact [Counts] | Atom cont. shift |
|---|---|---|---|---|---|---|---|
| DBVen1619_1 | - | N/A | N/A | -25.94 | -17.04% | 94 | +7 |
| | + | 672 | 9.98% | -30.36 | | 101 | |
| DBPro1156_1 | - | N/A | N/A | -28.59 | -18.46% | 117 | +7 |
| | + | 702 | 10.57% | -33.87 | | 124 | |
| DBAct553_1 | - | N/A | N/A | -32.18 | -27.72% | 123 | +26 |
| | + | 886 | 12.13% | -41.10 | | 149 | |

**Supplementary Table S 4.4: Docking benchmark complexes.** List of 14 protein-ligand complexes and 200 decoys used in the binding partner recovery experiment. Search parameters: interface cutoff = 0.0; NN score cutoff = 0.8; descriptor distance cutoff = 3.0; #sites = 3; selection radius = 10Å.

| **Protein-protein-ligand complex** | | | |
|---|---|---|---|
| **PDB ID** | **Protein 1** | **Protein 2** | **Ligand** |
| 1A7X | FKBP12 (chain A) | FKBP12 (chain B) | BENZYL-CARBAMIC ACID [8-DEETHYL-ASCOMYCIN-8-YL]ETHYL ESTER (FKA) |
| 1S9D | ADP-Ribosylation Factor 1 (chain A) | Arno (chain E) | 1,6,7,8,9,11A,12,13,14,14A-DECAHYDRO-1,13-DIHYDROXY-6-METHYL-4H-CYCLOPENT[F]OXACYCLOTRIDECIN-4-ONE (AFB) |
| 1TCO | SERINE/THREONINE PHOSPHATASE B2 (chains A and B) | FK506-BINDING PROTEIN (chain C) | 8-DEETHYL-8-[BUT-3-ENYL]-ASCOMYCIN (FK5) |
| 3QEL | NMDA glutamate receptor subunit (chain A) | Glutamate [NMDA] receptor subunit epsilon-2 (chain B) | 4-[(1R,2S)-2-(4-benzylpiperidin-1-yl)-1-hydroxypropyl]phenol (QEL) |
| 4DRI | Peptidyl-prolyl cis-trans isomerase FKBP5 (chain A) | Serine/threonine-protein kinase mTOR (chain B) | RAPAMYCIN IMMUNOSUPPRESSANT DRUG (RAP) |
| 4MDK | Ubiquitin-conjugating enzyme E2 R1 (chain A) | Ubiquitin (chain E) | 4,5-dideoxy-5-(3',5'-dichlorobiphenyl-4-yl)-4-[(methoxyacetyl)amino]-L-arabinonic acid (U94) |
| 6ENG | DNA gyrase subunit B (chain A) | DNA gyrase subunit B (chain B) | Coumermycin A1 (BHW) |
| 6H0F | Protein cereblon (chain B) | DNA-binding protein Ikaros (chain C) | S-Pomalidomide (Y70) |
| 6N4N | NS3 protease (chain A) | Rosetta-designed danoprevir/NS3a complex reader 2 (chain F) | (2R,6S,12Z,13aS,14aR,16aS)-6-[(tert-butoxycarbonyl)amino]-14a-[(cyclopropylsulfonyl)carbamoyl]-5,16-dioxo-1,2,3,5,6,7,8 ,9,10,11,13a,14,14a,15,16,16a-hexadecahydrocyclopropa[e]pyrrolo[1,2-a][1,4]diazacyclopentadecin-2-yl 4-fluoro-2H-isoindole-2-carboxylate (TSV) |
| 6OB5 | Maltodextrin-binding protein (chain B) | Ankyrin Repeat Domain (AR), S3-2D variant (chain D) | FARNESYL DIPHOSPHATE (FPP) |
| 6QTL | VHH (chain A) | VHH (chain C) | CAFFEINE (CFF) |

| | | | |
|---|---|---|---|
| 6SJ7 | DDB1- and CUL4-associated factor 15 (chain A) | RNA binding protein 39 (chain C) | N~1~-(3-chloro-1H-indol-7-yl)benzene-1,4-disulfonamide (EF6) |
| 7DC8 | Switch Ab Fab light & heavy chain (chains A and B) | Interleukin-6 receptor subunit alpha (chains C and F) | ADENOSINE-5'-TRIPHOSPHATE (ATP) |
| 7TE8 | DB21 (chain A) | CA14 (chain C) | cannabidiol (P0T) |
| **Decoys** | | | |
| 1A2K_AB, 1A2K_C, 1AVX_A, 1AVX_B, 1BRS_A, 1BRS_D, 1ERN_A, 1ERN_B, 1H6K_B, 1H6K_Y, 1I07_A, 1I07_B, 1I4O_B, 1I4O_D, 1ID5_H, 1ID5_L, 1JKG_A, 1JKG_B, 1JZO_A, 1JZO_B, 1LQM_E, 1LQM_F, 1NPO_A, 1NPO_C, 1O9Y_A, 1O9Y_D, 1PXV_A, 1PXV_C, 1Q5H_A, 1Q5H_B, 1SHY_A, 1SHY_B, 1SOT_A, 1SOT_C, 1T0F_A, 1T0F_B, 1TQ9_A, 1TQ9_B, 1UGH_E, 1UGH_I, 1UUG_A, 1UUG_B, 1XDT_R, 1XDT_T, 1XPJ_A, 1XPJ_D, 1XT9_A, 1XT9_B, 1XUA_A, 1XUA_B, 1YC0_A, 1YC0_I, 1YLQ_A, 1YLQ_B, 1YY9_A, 1YY9_D, 1Z0K_A, 1Z0K_C, 1ZR0_A, 1ZR0_B, 1ZVN_A, 1ZVN_B, 2A2L_B, 2A2L_C, 2AQX_A, 2AQX_B, 2B3Z_C, 2B3Z_D, 2B42_A, 2B42_B, 2FE8_A, 2FE8_C, 2G2W_A, 2G2W_B, 2GD4_B, 2GD4_C, 2GKW_A, 2GKW_B, 2HDP_A, 2HDP_B, 2HEK_A, 2HEK_B, 2I32_A, 2I32_E, 2J12_A, 2J12_B, 2JI1_C, 2JI1_D, 2LBU_D, 2LBU_E, 2O8Q_A, 2O8Q_B, 2P45_A, 2P45_B, 2P47_A, 2P47_B, 2QLC_B, 2QLC_C, 2WAM_A, 2WAM_C, 2WQ4_A, 2WQ4_C, 2Y32_B, 2Y32_D, 2YZJ_A, 2YZJ_C, 2Z0P_C, 2Z0P_D, 2Z29_A, 2Z29_B, 2Z7F_E, 2Z7F_I, 3AXY_B, 3AXY_D, 3B5U_J, 3B5U_L, 3BTV_A, 3BTV_B, 3CDW_A, 3CDW_H, 3CEW_C, 3CEW_D, 3CG8_B, 3CG8_C, 3CHW_A, 3CHW_P, 3E2U_A, 3E2U_E, 3ECY_A, 3ECY_B, 3EYD_C, 3EYD_D, 3F74_A, 3F74_B, 3FJS_C, 3FJS_D, 3HCG_A, 3HCG_C, 3HN6_B, 3HN6_D, 3HRD_E, 3HRD_H, 3IBM_A, 3IBM_B, 3ISM_A, 3ISM_B, 3K3C_A, 3K3C_B, 3KMT_A, 3KMT_B, 3KZH_A, 3KZH_B, 3M85_B, 3M85_E, 3OGF_A, 3OGF_B, 3P71_C, 3P71_T, 3P8B_C, 3P8B_D, 3PGA_1, 3PGA_4, 3Q0Y_B, 3Q0Y_C, 3Q87_A, 3Q87_B, 3Q9U_A, 3Q9U_C, 3QWN_I, 3QWN_J, 3QWQ_A, 3QWQ_B, 3RDZ_A, 3RDZ_C, 3S8V_A, 3S8V_X, 3S9C_A, 3S9C_B, 3SGB_E, 3SGB_I, 3SLH_A, 3SLH_B, 3TND_B, 3TND_D, 3WN7_A, 3WN7_B, 4AG2_A, 4AG2_C, 4CJ0_A, 4CJ0_B, 4KGG_A, 4KGG_C, 4M5F_A, 4M5F_B, 4TQ1_A, 4TQ1_B, 4YDJ_G, 4YDJ_HL, 5GPG_A, 5GPG_B | | |

**Supplementary Table S 4.5**: **Target protein and binder sequences.**

| Design | Sequence | Mutations from native |
|---|---|---|
| DBVen1619_1 | KYMLVVKGPNVTIFRWVDSSSEAETLARKIAKKLGLEVKSVEKKGNAVRVEIG | |
| DBVen1619_2 | QYLLVVKGPNVTKFRWVDSSSEAETLARKIAKKLGLEVKSVEKKGNAVRVEIG | K1Q, M3L, I13K |
| DBPro1156_1 | DEKAKTAETLIYQLFSKAMQQSDPNEAEKLLKKAEELAKKANDPRLEQVVRQ HQVVVRFLV | |
| DBPro1156_2 | DEKAKTAETLIWQLFGKAMQQSDPNEAEKLLKKAEELAKKANDPRLEQVVR QHQVVVRFLV | Y12W, S16G |
| DBAct553_1 | DYIRELRAALILLALKKQHAEDPDAQRVADELMKKLFDAAHRNDKDKVKKV VEEAKKVVSTY | |
| DBAct553_2 | DYIRELNRALILLALKKQHAEDPDAQRVADELMKKLFDAAHRNDKDKVKKV VEEAKKVVSTY | R7N, A8R |
| DB3_H | QIQLVQSGPELKKPGETVKISCKASGYAFTNYGVNWVKEAPGKELKWMGWI NIYTGEPTYVDDFKGRFAFSLETSASTAYLEINNLKNEDTATYFCTRGDYVNW YFDVWGAGTTVTVSSAKTTPPSVYPLAPGSAAQTNSMVTLGCLVKGYFPEPV TVTWNSGSLSSGVHTFPAVLQSDLYTLSSSVTVPSSPRPSETVTCNVAHPASST KVDKKIVPR | |
| DB3_L | DVVMTQIPLSLPVNLGDQASISCRSSQSLIHSNGNTYLHWYLQKPGQSPKLL MYKVSNRFYGVPDRFSGSGSGTDFTLKISRVEAEDLGIYFCSQSSHVPPTFGG GTKLEIKRADAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKDINVKWKIDGSE RQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYERHNSYTCEATHKTSTSPIV KSFNR | |
| DB3_H (Chimeric) | QIQLVQSGPELKKPGETVKISCKASGYAFTNYGVNWVKEAPGKELKWMGWI NIYTGEPTYVDDFKGRFAFSLETSASTAYLEINNLKNEDTATYFCTRGDYVNW YFDVWGAGTTVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVT VSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNT KVDKKVEPKSCDKTHT | |
| DB3_L (Chimeric) | DVVMTQIPLSLPVSLGEQASISCRSSQSLIHSNGNTYLHWYLQKPGQSPKLLM YKVSNRFYGVPDRFSGSGSGTDFTLKISRVEAEDLGIYFCSQSSHVPPTFGGGT KLEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQ SGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKS FNRGEC | |
| Anti-kappa_H | EVKLLESGGGLVQPGRSLRLSCIASGFDFSGYWMTWVRQAPGKGLEWIGDIN PDSSTINSTPSLKDKVIISRDNAKNTLFLQMSKVRSEDTALYYCAQRGNYVPFP YWGQGTLVTVSAAKTTPPSVYPLAPGSAAQTNSMVTLGCLVKGYFPEPVTVT WNSGSLSSGVHTFPAVLQSDLYTLSSSVTVPSSTWPSETVTCNVAHPASSTKV DKKIVPRDCGCK | |
| Anti-kappa_L | SIVMTQTPKFLFVSAGDRVTITCKASQSVSNDVEWYQQKPGQSPKLMIYFASK RYNGVPDRFTGSGFGTEFTFTISTVQAEDLAVYFCQQDYSSPWTFGGGTKLEI KRADAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKDINVKWKIDGSERQNGV LNSWTDQDSKDSTYSMSSTLTLTKDEYERHNSYTCEATHKTSTSPIVKSFNRG EC | |
| Bcl2 | MAHAGRTGYDNREIVMKYIHYKLSQRGYEWDAGDDAEENRTEAPEGTESEV VHRALRDAGDDFERRYRRDFAEMSSQLHLTPDTARQRFETVVEELFRDGVN WGRIVAFFEFGGVMCVESVNREMSPLVDNIAEWMTEYLNRHLHTWIQDNG GWDAFVELYGPSMR | |
| PDF1 | AILNILEFPDPRLRTIAKPVEVVDDAVRQLIDDMFETMYEAPGIGLAATQVNV HKRIVVMDLSEDKSEPRVFINPEFEPLTEDMDQYQEGCLSVPGFYENVDRPQ KVRIKALDRDGNPFEEVAEGLLAVCIQHECDHLNGKLFVDYLSTLKRDRIRKK LEKQHRQQA | |

**Supplementary Table S 4.6 : Crystallographic data collection and refinement statistics**.

| DBAct553_1:Actinonin:PDF1 (PDB: 8S1X) | |
|---|---|
| Data collection | |
| Space group | $P2_12_12_1$ |
| Cell dimensions | |
| *a*, *b*, *c* (Å) | 49.44, 75.01, 83.16 |
| a, b, g(°) | 90.0, 90.0, 90.0 |
| Wavelength (Å) | 0.87313 |
| Resolution (Å) | 55.7 - 1.88 (1.96 - 1.88) |
| Unique reflections | 24990 (1266) |
| $R_{merge}$ | 0.044 (1.125) |
| $I$ / s$I$ | 15.0 (1.3) |
| CC1/2 | 0.999 (0.426) |
| Completeness (%) | 96.9 (99.6) |
| Redundancy | 4.3 (4.4) |
| Refinement | |
| Resolution (Å) | 55.7 - 1.88 |
| No. reflections | 24982 (2801) |
| $R_{work}$ / $R_{free}$ | 0.1838/0.2030 |
| No. atoms | 1999 |
| Protein | 1850 |
| Ligand/ion | 73 |
| Water | 76 |
| *B*-factors (Å$^2$) | 57.2 |
| Protein | 56.9 |
| Ligand/ion | 66.7 |
| Water | 55.3 |
| R.m.s. deviations | |
| Bond lengths (Å) | 0.007 |
| Bond angles (°) | 0.730 |
| Ramachandran plot | |

| | |
|---|---|
| Favored (%) | 99.11 |
| Allowed (%) | 0.89 |
| Outliers (%) | 0.00 |

# Chapter 5

## Conclusions and perspectives

This work presented some of my contributions made to the field of computational protein design and more specifically for the engineering of novel protein interactions. In addition to introducing novel computational tools approaches enhancing our comprehension of PPI engineering, a particular emphasis was placed on ensuring translational applications for all the cases that were discussed. From my perspective, biomedical research should strive to address unsolved scientific questions, while maintaining a constant awareness of how its progress can ultimately benefit to a broad range of the population, including the patients. In the following sections, we will discuss the conclusions that were made in the three projects discussed hereinbefore and bring an overall perspective to this work (Figure 5.1).

### 5.1 Controlling protein therapeutics with a drug-responsive switch

Monoclonal antibodies, and other protein-based therapeutic such as cytokines, represent a growing market with promising outcomes in the clinics [143,144,147]. However, their development is hindered by the presence of systemic toxicities that can induce deleterious side effects, namely cytokine storm syndromes leading to organ failures [145,146]. Several engineering strategies using ON-switch systems were proposed to control their localization and activity *in vivo,* but these approaches are depending on internal stimuli that lacks external monitoring [149], or have a slow OFF turnover following the cessation of the stimulus [150]. Chemically-disruptable heterodimers can act as OFF-switch systems upon the addition of a small molecule for a rapid stop of the therapeutic effect. Yet, the number of successful examples of soluble protein therapeutic using OFF-switch systems is limited. Numerous attempts of engineering novel chemically-responsive switches were made in the past years [237], but a number of them were not using proteins of human origin [155] or were shown to have a low affinity [152]. Though, to design a soluble therapeutic for use in human patients, there are two parameters that are crucial to ensure safety and efficacy: i) The lack of immunogenicity and ii) the binding stability of the switchable moieties. Another consideration would be to use safe, well-characterized and deliverable molecules as a switch trigger, such as a clinically approved drug.

In chapter 2, we leveraged a high-affinity CDH – previously developed in our lab for CAR-T cell therapy [165] – composed of the human Bcl2 protein and the computationally designed LD3 that originates from a human globular protein. This CDH was incorporated into a protein therapeutic such as an antibody or an Fc-fused cytokines to act as a drug-responsive OFF-switch system. We hypothesized that the loss of the Fc moiety upon switching can lead to a rapid OFF-state due to the decreased half-life *in vivo*, the reduced avidity and the loss of effector function [156–158].

However, this CDH complex was initially designed for a cell-based therapy and a significantly different switchability was observed when the chemical switch was tested in solution. Firstly, molecular crowding in cells can lead to very different behaviors and protein dynamic in solutions compared to cell assays [272]. Secondly, proteins have a continuous turnover and are repeatedly expressed and degraded to maintain cell homeostasis [273]. The successful switchability of Bcl2:LD3 system in CAR T cells could be explained by a pre-blocking of Bcl2 by Venetoclax directly after its synthesis but prior complexation with LD3, thus locking Bcl2 in an LD3-unbound state. However, once Bcl2 is bound to LD3, the probabilities for the drug to compete are very low. Indeed, an extremely low dissociation rate was measured *in vitro*, which highly restricts the opportunity of the drug to displace the LD3 binder. We therefore hypothesized that slightly reducing the affinity by increase the dissociation rate could lead to a better switchability.

Similarly to alanine scanning performed experimentally to assess binding contribution at the single amino acid level [274], we used a computational alanine scan tool on Rosetta modelling suite to predict mutations that can decrease the computed binding energy. With this approach, we then focused our screening on 5 residue positions, which considerably reduced the experimental work. After experimental validation, we selected a mutation that slightly increased the dissociation rate ($k_{off}$) while maintaining the association rate ($k_{on}$) as similar as possible to ensure a good LD3:Bcl2 complex stability. Of note, no clear correlation between the predicted $\Delta\Delta G$ decrease and changes of binding kinetic was observed, which further supports the low energetic resolution of physics-based scoring function and the need of computational tools with better sensitivity in future works [275,276].

Overall, the integration of the predicted mutation lead to a better switchability *in vitro* and *in vivo*, while maintaining the complex stability and efficacy in absence of the small molecule. However, the overall molecular weight of the switchable complex (250 kDa) could represent a potential limitation. Although this size might be suitable for targets within the bloodstream, it could impede tissue penetration, for example in solid tumors [160]. Future engineering could incorporate smaller CDH systems or directly incorporating some binding motifs within the protein therapeutic moieties by motif grafting and design [277].

In summary, the first aim of this dissertation proposed a rational blueprint to design drug-responsive protein therapeutics in order to increase safety and control of protein-based therapies. We exemplified the translation of protein interactions into tangible biomedical applications, achieved through the utilization of computational tools for a better switchability.

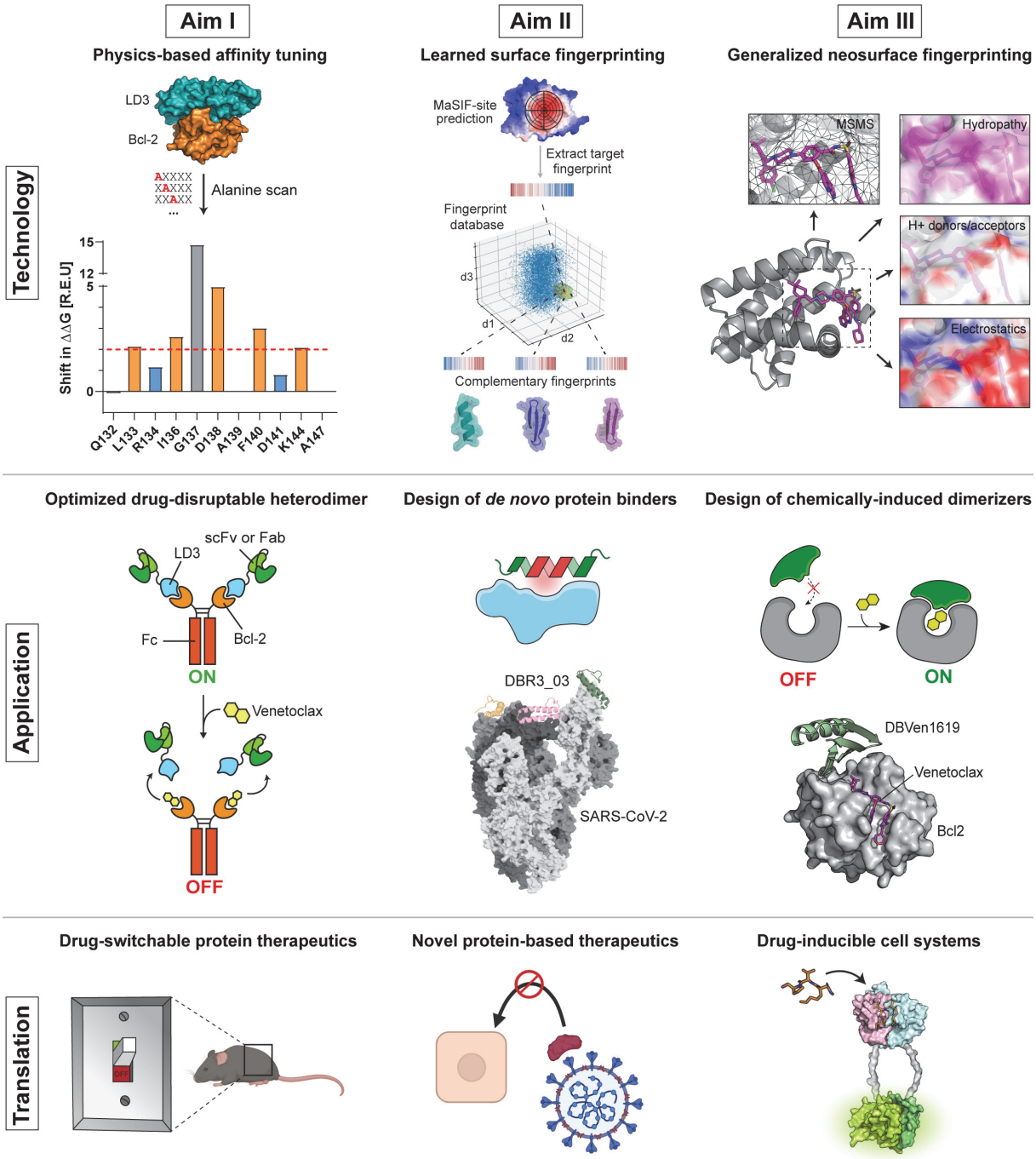



**Figure 5.1 : Summary of the different technologies, applications and translations.** The three aims presented in this dissertation are summarized with the computational technologies that were used or developed, the experimental applications and their translation into potential biomedical systems. Overall, this work aimed to take advantage of existing tools or to develop novel computational methods for the design of novel protein-protein interactions with translational capabilities, such as protein-based therapeutics or chemical-responsive switches.

## 5.2 Designing novel protein interactions straight from a computer

The design of novel protein-protein interaction has remained a challenge for the field of protein science. Most engineering strategies have relied on template-based approaches that repurpose existing interfaces – such as in chapter 2 [97] and in other works [104,116,118] – but this method is inefficient for targeting sites where no experimental or structural data are available. Scientists have therefore generated monoclonal antibodies [14] or screened among libraries with millions of variants [278], but the outcomes are often agnostic to where and how the respective targets are engaged. Among the *de novo* design methods, the dock-&-optimize approach demonstrated only low affinity and specificity [129–131], while the hotspot-centric approach has so far been dependent on known hotspots or extensive experimental screening and optimization [132,162,166]. Of note, in order to design *de novo* PPIs, two key aspects need to be solved: i) the binding site, and ii) the binding partner.

With the advent of machine learning, novel tools are available to study protein interfaces and interactions. Among them, MaSIF has been proposed to use vectorized chemical and geometrical features found on the protein molecular surface to predict PPI site and partners based solely on these fingerprints, without any co-evolutionary signature. In chapter 3, we took advantage of MaSIF capabilities to find high-propensity interfaces and search for binding motifs displaying highly complementary fingerprints for four therapeutically-relevant targets: SARS-CoV-2, PD-L1, PD-1 and CTLA-4. With our surface-centric approach, we successfully obtained highly specific binders for each of these “undruggable” targets with affinities ranging from nanomolar to low-micromolar.

However, in the first attempts, we faced limitations in terms of i) binding affinity and ii) scaffold stability. Firstly, the early designs always required extensive *in vitro* maturation to obtain binding affinities in the range of native PPIs. The causes were mostly originating from a suboptimal polar interactions at the rim of the interface, the presence of buried unsatisfied polar atoms and steric clashes at the core of the interface, all contributing to inadequate protein interactions also reported in previous findings [130,132,135,279,280]. To address these downsides, a seed refinement step was added prior grafting with the aim to relax the structure and perform sequence optimization emphasizing the establishment of a more robust polar network thanks to a penalization of buried unsatisfied polar atoms [137]. Secondly, as exemplified by DBL1_02, some designs originating from natural protein scaffolds were not readily expressible and stable, and required some further optimization. Natural proteins can often be unstable, difficult to express and less tolerant to mutations [281,282]. Nowadays, evolution-based or deep learning tools can greatly enhance expressibility and solubility of natural proteins [283,284]. However, to avoid the addition of another layer of computational design in the current pipeline, we opted for the direct integration of hyperstable *de novo* miniproteins in our scaffold database [74,75]. With this optimized pipeline in hands, we successfully obtained another set of binders with native-like affinities by pure *in silico* generation and with high accurate prediction proved by mutational characterizations. One of these binders, DBP13_02 binding to PD-1, demonstrated translational capabilities as a potent agonist and T cell inhibitor.

However, some limitations are inherent to our current pipeline, such as the need for an external database for the search of complementary binding seeds and scaffolds. Despite their large size, they do not represent a universal answer to all design cases. Nowadays, new deep learning tools, such as hallucination or diffusion models, are emerging as a solution for scaffolding binding motifs [38,87] or for a full binder design [38,285]. Considering the success of surface fingerprinting for PPI prediction and design, we can anticipate that these novel deep learning approaches could benefit from the incorporation of fingerprint descriptors in their loss function for an enhanced success rate. Moreover, the MaSIF-seed framework and its synergy with other tools would greatly benefit from the transition to a lighter architecture, such as dMaSIF (differentiable molecular surface interaction fingerprinting) [286], that removes input precomputation and reduces computational time or memory requirements. This adaptation would enable an on-the-fly utilization without the necessity of heavy pre-computation steps.

An additional challenge persists in addressing sites characterized by high flexibility or polarity. MaSIF's input only use a snapshot of the protein conformational space that might not be representative of highly flexible region in experimental conditions (e.g. loops). A synergistic combination of MaSIF with other machine learning-based conformation predictors [287,288] would greatly enhance the designability of flexible sites, however with an associated computational cost. The second main challenge to address involves polar sites that lack large hydrophobic patches. Indeed, MaSIF has been primarily trained on a set of PPI characterized by a hydrophobic patch at the core of the interface, in line with the classical representation of these interactions [18,110]. Success rates are also highly correlated with the hydrophobicity of the target site as it contributes to a large part of the binding energy [166,170]. However, to generalize to a broader range of sites, akin to antibodies, overcoming challenges associated with accurate polar network [289], water molecule modelling [53,110,290] and interface desolvation energy [132,291] becomes imperative.

Finally, the surface fingerprinting approach focuses on a 12Å-radius patch assuming that most crucial biomolecular interactions will occur within this defined area. Nevertheless, this oversimplification towards local interactions will often neglect long-range electrostatics which play a crucial role in protein binding, notably for the association phase ($k_{on}$) [292]. A novel framework leveraging a multimodal architecture with inputs from structure, sequence and surfaces features together, could potentially address this concern.

Altogether, the second aim of this thesis proposed to leverage surface fingerprints to successfully design *de novo* protein binders against four therapeutically-relevant targets, with applications as protein-based therapeutics. By using a higher-level representation, namely the protein molecular surface, our approach represents a new paradigm and is the only one to leverage surface features for the design of novel PPIs. While some limitations became evident, the success rate and accuracy achieved in this work reflect a crucial milestone achieved in the field compared to previous experimental and computational methods. Therefore, this approach still represents an exciting route to design novel protein interactions and will benefit from a synergy with novel deep learning tools.

## 5.3 Generalizing surface fingerprinting towards drug-induced protein interactions

Protein-protein interactions is often regulated by various chemical stimuli such as post-translational modification [293] and ligand binding [105,237], or other biophysical stimuli (pH, light, temperature etc.) [148,149,294,295]. Small molecules rapidly became attractive candidates in the field of synthetic biology for the spatiotemporal control of protein interactions thanks to their fast responsiveness and delivery. Several research groups successfully reported OFF-switch system leveraging chemically-disruptable heterodimer with numerous applications [165], such as the switchable protein therapeutics exemplified in chapter 2. On the other hand, only a limited number of computational approaches for the development of chemically-induced dimerization that can be functionalized into ON-switch system have been proposed. Indeed, the wide majority of CID systems elaborated so far were engineered using extensive experimental methods. The avenue of new methods that incorporate small molecules in the protein design pipeline would expand the engineering landscape for novel synthetic biology tools.

In chapter 4, we took advantage of the surface fingerprinting approach developed in chapter 3 for the design of protein binders that target neosurfaces, i.e. surfaces arising from the protein-ligand complex. We hypothesized that small molecules bound on protein surfaces may display similar features and fingerprints as any canonical protein amino acids. The MaSIF prediction framework was therefore adapted to incorporate small molecules in the molecular surface representation and descriptors of the target protein. After demonstrating that this new tool can successfully generalize to small molecules, but also predict and recover known CID systems, we successfully validated three designed binders for three small molecule-bound protein complexes: Bcl-2:Venetoclax, DB3-IgG:Progesterone and PDF1:Actinonin. All designs showed highly accurate prediction by mutational characterization, reached native-like affinities after pure *in silico* generation and were readily optimized to nanomolar affinities with only few mutations. Ultimately, our designs were functionalized as drug-inducible ON-switch systems in cell-based assays which opens the door to a broader range of therapeutic applications like CAR-T cells.

Nevertheless, some of the limitations encountered previously, namely the use of constrained databases, the challenges of flexible sites or polar interfaces, were not yet solved in this new framework. The success rate (0.05%) is still a major drawback, but a deeper analysis of machine-learning and physics-based scores provided in this work will allow a better discrimination and filtering in future applications. Of note, the affinities and accuracies exhibited by the successful binders counterbalance the modest success rate, which remains encouraging considering the inherent difficulty of the chosen task.

As for previous computational methods, our approach partially took advantage of well-characterized drug binding sites to design novel CIDs [169,249]. Targeting existing drug-protein complexes to design a binding partner significantly reduces complexity. Yet it hinders the diversity of potential partners involved in the CID system and constrains the approach to complexes where the ligand is solvent-accessible. Innovative approaches involving deep learning – and more specifically generative

modelling – could constitute a new paradigm for developing fully *de novo* CIDs in the near future. Apart from a couple previous cases [169,249], our generalizable framework is one of the rare specifically tailored to design novel CIDs computationally. Structure prediction tools integrating small molecules [56,57] or generative models including nucleic acid components are about to be released [93], but a gap persists for the design of multimeric protein complexes binding to defined small molecules [296,297]. Message passing neural networks (MPNN) for the sequence design of drug-binding protein were recently suggested, but still need the placement of a putative backbone structure [298]. This strategy could however constitute a new way to consider the dock-and-optimize approach (see chapter 1.3) whose early attempts were limited by the absence of accurate energy force fields and the limited number of scaffolds available at that time [129–131]. Our surface-centric approach, could also greatly benefit from MPNN tools that includes small molecules for the seed and scaffold refinement steps.

Altogether, we showed that our surface fingerprinting approach was generalizable to small molecule-bound interfaces without any new training required, which is uncommon in deep learning methods. The subsequent designed binders demonstrated a drug-specific ON-switch behavior that was functionalized in cells. We anticipate that this work could have some applications in the field of cell-based therapies, for instance, CAR-T cell therapies, which possess a great potential to fight various types of cancers [299], yet numerous off-target and deleterious effects have been reported [265]. Molecular switches upon internal and external stimuli can bring a better spatiotemporal control and safety [148,149,165]. Therefore, we foresee that our approach could be used for the design of CIDs involving tumor microenvironment-specific ligands (e.g. ATP) or punctual triggers (e.g. drug) and lead to safer and more specific treatment in the future.

## 5.4 Overall outlook and perspectives

Taken together, this dissertation provided new computational tools for the design of novel protein-protein interactions and their applications in various therapeutic strategies. While experimental methods – and notably the generation of monoclonal antibodies – are still preferred for the development of novel protein-based therapeutics, computational protein design is emerging as new avenue for pharmaceutical industry with multiple assets in terms of i) specificity, ii) applicability, iii) modularity, and iv) costs. First, computational methods, such as the ones proposed in this work, can rationally design protein binders for a specific site while experimentally-generated antibodies are often agnostic to where and how they bind their target immunogen. Secondly, computational protein design can provide molecules which have been optimized for solubility and stability [283], like the hyperstable miniprotein databases used in this work [74–76,79], which would facilitate manufacturing and handling processes. Third, a broader landscape of proteins in terms of folds, biochemical compositions, and sizes are possible and can be tailored for specific applications. Notably, the generation of low molecular weight binders could greatly enhance the tissue penetration of certain protein-based therapeutics [160]. Finally, the generation and screening of a limited number of protein designs could reduce R&D costs. Moreover, most protein designs were tailored for bacteria expression, which makes

their manufacturing more cost-effective than antibodies which need mammalian cell expression systems [300].

However, two major challenges persist for a greater expansion of the field: the immunogenicity of the engineered proteins and a better data-driven energetic resolution of the machine learning algorithms. As showed throughout this dissertation, *de novo* protein designs can be functionalized in a wide range of therapeutic applications. However, major concerns remain for these proteins which are by definition *de novo* and potentially recognized as "non-self" by the immune system. Immunogenicity is also a concern for more traditional biologics [301], but the lack of wide *in vivo* studies with *de novo* protein designs restrain their use in the clinics. As of today, and to our current knowledge, only two *in vivo* experiments demonstrated only little or no immunogenicity of these *de novo* protein binders, most probably due to their small size and hyperstability [77,163]. Nevertheless, further investigation about safety should be conducted for a broader application of these innovative biologics in pharmaceutical industry.

Despite the improvement of the predictions made in the field protein design, most computational models fail to accurately capture crucial energetic information at the single amino acid level [302,303]. Therefore, the experimental screening of hundreds of protein variants is often required to perform full characterization when engineering enzymes or PPIs. Low-throughput processes represents a limiting factor for the generation of experimental data that could benefit to the training of energetically high-resolution algorithm (e.g. point mutation binding energy prediction). Therefore, while lots of efforts have already been made for increasing computational resources, lab automation currently in progress should be pursued [304]. The generation of data at the single amino acid level, coupled with experimental metrics, will greatly enhance the training of next-generation algorithms.

Altogether, this work provided new insights for the design of novel protein-protein interactions with therapeutic potentials using cutting-edge physics-based and machine learning tools. While the number of deep learning tools available is expanding every month, this dissertation took part in this coming era of protein design and marked a step towards the generation of better and safer therapeutics.

# Chapter 6

## Appendix

## 6.1 Bibliography


1. Amaral P, Carbonell-Sala S, De La Vega FM, Faial T, Frankish A, Gingeras T, Guigo R, Harrow JL, Hatzigeorgiou AG, Johnson R, et al.: The status of the human gene catalogue. *Nature* 2023, 622:41–47.
2. Crick FH: On protein synthesis. *Symp Soc Exp Biol* 1958, 12:138–163.
3. Ramachandran GN, Ramakrishnan C, Sasisekharan V: Stereochemistry of polypeptide chain configurations. *J Mol Biol* 1963, 7:95–99.
4. Anfinsen CB: Principles that Govern the Folding of Protein Chains. *Science* 1973, 181:223–230.
5. Dill KA: Dominant forces in protein folding. *Biochemistry* 1990, 29:7133–7155.
6. Dill KA, Ozkan SB, Shell MS, Weikl TR: The Protein Folding Problem. *Annu Rev Biophys* 2008, 37:289–316.
7. Fleming PJ, Rose GD: Do all backbone polar groups in proteins form hydrogen bonds? *Protein Sci* 2005, 14:1911–1917.
8. Morris R, Black KA, Stollar EJ: Uncovering protein function: from classification to complexes. *Essays Biochem* 2022, 66:255–285.
9. Gonzalez MW, Kann MG: Chapter 4: Protein Interactions and Disease. *PLOS Comput Biol* 2012, 8:e1002819.
10. Bonetta L: Interactome under construction. *Nature* 2010, 468:851–852.
11. Stumpf MPH, Thorne T, De Silva E, Stewart R, An HJ, Lappe M, Wiuf C: Estimating the size of the human interactome. *Proc Natl Acad Sci* 2008, 105:6959–6964.
12. Mabonga L, Kappo AP: Protein-protein interaction modulators: advances, successes and remaining challenges. *Biophys Rev* 2019, 11:559–581.
13. Robertson NS, Spring DR: Using Peptidomimetics and Constrained Peptides as Valuable Tools for Inhibiting Protein–Protein Interactions. *Molecules* 2018, 23.
14. Lu R-M, Hwang Y-C, Liu I-J, Lee C-C, Tsai H-Z, Li H-J, Wu H-C: Development of therapeutic antibodies for the treatment of diseases. *J Biomed Sci* 2020, 27:1.
15. Quijano-Rubio A, Ulge UY, Walkey CD, Silva D-A: The advent of de novo proteins for cancer immunotherapy. *Curr Opin Chem Biol* 2020, 56:119–128.
16. Janin J, Bahadur RP, Chakrabarti P: Protein–protein interaction and quaternary structure. *Q Rev Biophys* 2008, 41:133–180.
17. Clackson Tim, Wells James A.: A Hot Spot of Binding Energy in a Hormone-Receptor Interface. *Science* 1995, 267:383–386.

18. Nilofer C, Sukhwal A, Mohanapriya A, Kangueane P: Protein-protein interfaces are vdW dominant with selective H-bonds and (or) electrostatics towards broad functional specificity. *Bioinformation* 2017, 13:164–173.

19. Xu D, Tsai CJ, Nussinov R: Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng Des Sel* 1997, 10:999–1012.

20. Richards FM: AREAS, VOLUMES, PACKING, AND PROTEIN STRUCTURE. *Annu Rev Biophys Bioeng* 1977, 6:151–176.

21. Lawrence MC, Colman PM: Shape Complementarity at Protein/Protein Interfaces. *J Mol Biol* 1993, 234:946–950.

22. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C: Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure* 2010, 18:1233–1243.

23. Meenan NAG, Sharma A, Fleishman SJ, MacDonald CJ, Morel B, Boetzel R, Moore GR, Baker D, Kleanthous C: The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proc Natl Acad Sci* 2010, 107:10080–10085.

24. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, et al.: Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* 2012, 490:556–560.

25. Lu L, Lu H, Skolnick J: MULTIPROSPECTOR: An algorithm for the prediction of protein–protein interactions by multimeric threading. *Proteins Struct Funct Bioinforma* 2002, 49:350–364.

26. Davis FP: Protein complex compositions predicted by structural similarity. *Nucleic Acids Res* 2006, 34:2943–2952.

27. Moore GE: Cramming more components onto integrated circuits. *IEEE Solid-State Circuits Soc Newsl* 2006, 11:33–35.

28. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC: GROMACS: Fast, flexible, and free. *J Comput Chem* 2005, 26:1701–1718.

29. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA: Development and testing of a general amber force field. *J Comput Chem* 2004, 25:1157–1174.

30. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, et al.: CHARMM: The biomolecular simulation program. *J Comput Chem* 2009, 30:1545–1614.

31. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al.: Highly accurate protein structure prediction with AlphaFold. *Nature* 2021, 596:583–589.

32. Baek Minkyung, DiMaio Frank, Anishchenko Ivan, Dauparas Justas, Ovchinnikov Sergey, Lee Gyu Rie, Wang Jue, Cong Qian, Kinch Lisa N., Schaeffer R. Dustin, et al.: Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021, 373:871–876.

33. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, et al.: Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins Struct Funct Bioinforma* 2009, 77:89–99.

34. Rohl CA, Strauss CEM, Misura KMS, Baker D: Protein Structure Prediction Using Rosetta. In *Methods in Enzymology*. . Elsevier; 2004:66–93.

35. Pierce BG, Wiehe K, Hwang H, Kim B-H, Vreven T, Weng Z: ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics* 2014, 30:1771–1773.

36. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D: Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. *J Mol Biol* 2003, 331:281–299.

37. Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, De Haas RJ, Bethel N, et al.: Robust deep learning–based protein sequence design using ProteinMPNN. *Science* 2022, 378:49–56.

38. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, et al.: De novo design of protein structure and function with RFdiffusion. *Nature* 2023, 620:1089–1100.

39. Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, Aprahamian M, Baker D, Barlow KA, Barth P, et al.: Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* 2020, 17:665–680.

40. Liu Y, Kuhlman B: RosettaDesign server for protein design. *Nucleic Acids Res* 2006, 34:W235–W238.

41. Kuhlman B, Bradley P: Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol* 2019, 20:681–697.

42. Anfinsen CB, Haber E, Sela M, White FH: THE KINETICS OF FORMATION OF NATIVE RIBONUCLEASE DURING OXIDATION OF THE REDUCED POLYPEPTIDE CHAIN. *Proc Natl Acad Sci* 1961, 47:1309–1314.

43. Roy A, Kucukural A, Zhang Y: I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 2010, 5:725–738.

44. Webb B, Sali A: Protein Structure Modeling with MODELLER. In *Functional Genomics*. Edited by Kaufmann M, Klinger C, Savelsbergh A. Springer New York; 2017:39–54.

45. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, et al.: SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018, 46:W296–W303.

46. Kawai H, Kikuchi T, Okamoto Y: A prediction of tertiary structures of peptide by the Monte Carlo simulated annealing method. *Protein Eng Des Sel* 1989, 3:85–94.

47. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E: Equation of State Calculations by Fast Computing Machines. *J Chem Phys* 1953, 21:1087–1092.

48. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG: Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins Struct Funct Bioinforma* 1995, 21:167–195.

49. Wolynes PG: Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* 2015, 119:218–230.

50. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, et al.: The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* 2017, 13:3031–3048.

51. König R, Dandekar T: Solvent entropy-driven searching for protein modeling examined and tested in simplified models. *Protein Eng Des Sel* 2001, 14:329–335.

52. Bhowmick A, Sharma SC, Honma H, Head-Gordon T: The role of side chain entropy and mutual information for improving the de novo design of Kemp eliminases KE07 and KE70. *Phys Chem Chem Phys* 2016, 18:19386–19396.

53. Pavlovicz RE, Park H, DiMaio F: Efficient consideration of coordinated water molecules improves computational protein-protein and protein-ligand docking discrimination. *PLOS Comput Biol* 2020, 16:e1008103.

54. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Topf M: Critical assessment of techniques for protein structure prediction, fourteenth round. 2020,

55. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al.: Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023, 379:1123–1130.

56. Google Deepmind: A glimpse of the next generation of AlphaFold. 2023,

57. Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, Lee GR, Morey-Burrows FS, Anishchenko I, Humphreys IR, et al.: *Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom*. Biochemistry; 2023.

58. Yue K, Dill KA: Inverse protein folding problem: designing polymer sequences. *Proc Natl Acad Sci* 1992, 89:4163–4167.

59. Huang P-S, Boyken SE, Baker D: The coming of age of de novo protein design. *Nature* 2016, 537:320–327.

60. Dahiyat BI, Mayo SL: De Novo Protein Design: Fully Automated Sequence Selection. *Science* 1997, 278:82–87.

61. Desjarlais JR, Handel TM: De novo design of the hydrophobic cores of proteins. *Protein Sci* 1995, 4:2006–2018.

62. Benson DE, Haddy AE, Hellinga HW: Converting a Maltose Receptor into a Nascent Binuclear Copper Oxygenase by Computational Design. *Biochemistry* 2002, 41:3262–3269.

63. Looger LL, Dwyer MA, Smith JJ, Hellinga HW: Computational design of receptor and sensor proteins with novel functions. *Nature* 2003, 423:185–190.

64. Silva D-A, Correia BE, Procko E: Motif-Driven Design of Protein–Protein Interfaces. In *Computational Design of Ligand Binding Proteins*. Edited by Stoddard BL. Springer New York; 2016:285–304.

65. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, et al.: Kemp elimination catalysts by computational enzyme design. *Nature* 2008, 453:190–195.

66. Kuhlman Brian, Dantas Gautam, Ireton Gregory C., Varani Gabriele, Stoddard Barry L., Baker David: Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* 2003, 302:1364–1368.

67. Pan X, Kortemme T: Recent advances in de novo protein design: Principles, methods, and applications. *J Biol Chem* 2021, 296:100558.

68. Jacobs TM, Williams B, Williams T, Xu X, Eletsky A, Federizon JF, Szyperski T, Kuhlman B: Design of structurally distinct proteins using strategies inspired by evolution. *Science* 2016, 352:687–690.

69. Harteveld Z, Bonet J, Rosset S, Yang C, Sesterhenn F, Correia BE: A generic framework for hierarchical de novo protein design. *Proc Natl Acad Sci* 2022, 119:e2206111119.

70. Yang C, Sesterhenn F, Bonet J, van Aalen EA, Scheller L, Abriata LA, Cramer JT, Wen X, Rosset S, Georgeon S, et al.: Bottom-up de novo design of functional proteins with complex structural features. *Nat Chem Biol* 2021, 17:492–500.

71. Sesterhenn F, Yang C, Bonet J, Cramer JT, Wen X, Wang Y, Chiang C-I, Abriata LA, Kucharska I, Castoro G, et al.: De novo protein design enables the precise induction of RSV-neutralizing antibodies. *Science* 2020, 368:eaay5051.

72. Huang P-S, Feldmeier K, Parmeggiani F, Fernandez Velasco DA, Höcker B, Baker D: De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat Chem Biol* 2016, 12:29–34.

73. Pan X, Thompson MC, Zhang Y, Liu L, Fraser JS, Kelly MJS, Kortemme T: Expanding the space of protein geometries by computational design of de novo fold families. *Science* 2020, 369:1132–1136.

74. Bhardwaj G, Mulligan VK, Bahl CD, Gilmore JM, Harvey PJ, Cheneval O, Buchko GW, Pulavarti SVSRK, Kaas Q, Eletsky A, et al.: Accurate de novo design of hyperstable constrained peptides. *Nature* 2016, 538:329–335.

75. Rocklin Gabriel J., Chidyausiku Tamuka M., Goreshnik Inna, Ford Alex, Houliston Scott, Lemak Alexander, Carter Lauren, Ravichandran Rashmi, Mulligan Vikram K., Chevalier Aaron, et al.: Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* 2017, 357:168–175.

76. Linsky TW, Noble K, Tobin AR, Crow R, Carter L, Urbauer JL, Baker D, Strauch E-M: Sampling of structure and sequence space of small protein folds. *Nat Commun* 2022, 13:7151.

77. Chevalier A, Silva D-A, Rocklin GJ, Hicks DR, Vergara R, Murapa P, Bernard SM, Zhang L, Lam K-H, Yao G, et al.: Massively parallel de novo protein design for targeted therapeutics. *Nature* 2017, 550:74–79.

78. Cao L, Goreshnik I, Coventry B, Case JB, Miller L, Kozodoy L, Chen RE, Carter L, Walls AC, Park Y-J, et al.: De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* 2020, 370:426.

79. Tobin AR, Crow R, Urusova DV, Klima JC, Tolia NH, Strauch E: Inhibition of a malaria host–pathogen interaction by a computationally designed inhibitor. *Protein Sci* 2023, 32:e4507.

80. Khakzad H, Igashov I, Schneuing A, Goverde C, Bronstein M, Correia B: A new age in protein design empowered by deep learning. *Cell Syst* 2023, 14:925–939.

81. Scarselli F, Gori M, Ah Chung Tsoi, Hagenbuchner M, Monfardini G: The Graph Neural Network Model. *IEEE Trans Neural Netw* 2009, 20:61–80.

82. Rosenfeld R: Two decades of statistical language modeling: where do we go from here? *Proc IEEE* 2000, 88:1270–1278.

83. Ingraham J, Garg V, Barzilay R, Jaakkola T: Generative Models for Graph-Based Protein Design. *Adv Neural Inf Process Syst 32 NeurIPS 2019* 2019,

84. Goverde CA, Pacesa M, Dornfeld LJ, Goldbach N, Georgeon S, Rosset S, Dauparas J, Schellhaas C, Kozlov S, Baker D, et al.: *Computational design of soluble analogues of integral membrane protein structures*. Bioinformatics; 2023.

85. Bennett NR, Coventry B, Goreshnik I, Huang B, Allen A, Vafeados D, Peng YP, Dauparas J, Baek M, Stewart L, et al.: Improving de novo protein binder design with deep learning. *Nat Commun* 2023, 14:2625.

86. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, Olmos JL, Xiong C, Sun ZZ, Socher R, et al.: Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* 2023, 41:1099–1106.

87. Wang J, Lisanza S, Juergens D, Tischer D, Watson JL, Castro KM, Ragotte R, Saragovi A, Milles LF, Baek M, et al.: Scaffolding protein functional sites using deep learning. *Science* 2022, 377:387–394.

88. Eguchi RR, Choe CA, Huang P-S: Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. *PLOS Comput Biol* 2022, 18:e1010271.

89. Guo X, Du Y, Tadepalli S, Zhao L, Shehu A: Generating tertiary protein structures via interpretable graph variational autoencoders. *Bioinforma Adv* 2021, 1:vbab036.

90. Anand N, Huang P-S: Generative Modeling for Protein Structures. *32nd Conf Neural Inf Process Syst NeurIPS 2018* 2018,

91. Rahman T, Du Y, Zhao L, Shehu A: Generative Adversarial Learning of Protein Tertiary Structures. *Molecules* 2021, 26:1209.

92. Ho J, Jain A, Abbeel P: Denoising Diffusion Probabilistic Models. 2020, doi:10.48550/ARXIV.2006.11239.

93. Morehead A, Ruffolo J, Bhatnagar A, Madani A: Towards Joint Sequence-Structure Generation of Nucleic Acid and Protein Complexes with SE(3)-Discrete Diffusion. 2024, doi:10.48550/ARXIV.2401.06151.

94. Yim J, Campbell A, Mathieu E, Foong AYK, Gastegger M, Jiménez-Luna J, Lewis S, Satorras VG, Veeling BS, Noé F, et al.: Improved motif-scaffolding with SE(3) flow matching. 2024, doi:10.48550/ARXIV.2401.04082.

95. Smith MC, Gestwicki JE: Features of protein-protein interactions that translate into potent inhibitors: topology, surface area and affinity. *Expert Rev Mol Med* 2012, 14:e16–e16.

96. Zhang G, Andersen J, Gerona-Navarro G: Peptidomimetics Targeting Protein-Protein Interactions for Therapeutic Development. *Protein Pept Lett* 2018, 25:1076–1089.

97. Giordano-Attianese G, Gainza P, Gray-Gaillard E, Cribioli E, Shui S, Kim S, Kwak M-J, Vollers S, Corria Osorio ADJ, Reichenbach P, et al.: A computationally designed chimeric antigen receptor provides a small-molecule safety switch for T-cell therapy. *Nat Biotechnol* 2020, 38:426–432.

98. Wu C-Y, Roybal KT, Puchner EM, Onuffer J, Lim WA: Remote control of therapeutic T cells through a small molecule-gated chimeric receptor. *Science* 2015, 350:aab4077–aab4077.

99. Zajc CU, Dobersberger M, Schaffner I, Mlynek G, Pühringer D, Salzer B, Djinović-Carugo K, Steinberger P, De Sousa Linhares A, Yang NJ, et al.: A conformation-specific ON-switch for controlling CAR T cells with an orally available drug. *Proc Natl Acad Sci* 2020, 117:14926.

100. Scheller L, Strittmatter T, Fuchs D, Bojar D, Fussenegger M: Generalized extracellular molecule sensor platform for programming cellular behavior. *Nat Chem Biol* 2018, 14:723–729.

101. Quijano-Rubio A, Yeh H-W, Park J, Lee H, Langan RA, Boyken SE, Lajoie MJ, Cao L, Chow CM, Miranda MC, et al.: De novo design of modular and tunable protein biosensors. *Nature* 2021, 591:482–487.

102. Langan RA, Boyken SE, Ng AH, Samson JA, Dods G, Westbrook AM, Nguyen TH, Lajoie MJ, Chen Z, Berger S, et al.: De novo design of bioactive protein switches. *Nature* 2019, 572:205–210.

103. Correia BE, Bates JT, Loomis RJ, Baneyx G, Carrico C, Jardine JG, Rupert P, Correnti C, Kalyuzhniy O, Vittal V, et al.: Proof of principle for epitope-focused vaccine design. *Nature* 2014, 507:201–206.

104. Correia BE, Ban Y-EA, Holmes MA, Xu H, Ellingson K, Kraft Z, Carrico C, Boni E, Sather DN, Zenobia C, et al.: Computational Design of Epitope-Scaffolds Allows Induction of Antibodies Specific for a Poorly Immunogenic HIV Vaccine Epitope. *Structure* 2010, 18:1116–1126.

105. Shui S, Gainza P, Scheller L, Yang C, Kurumida Y, Rosset S, Georgeon S, Di Roberto RB, Castellanos-Rueda R, Reddy ST, et al.: A rational blueprint for the design of chemically-controlled protein switches. *Nat Commun* 2021, 12:5754.

106. Chen Zibo, Kibler Ryan D., Hunt Andrew, Busch Florian, Pearl Jocelynn, Jia Mengxuan, VanAernum Zachary L., Wicky Basile I. M., Dods Galen, Liao Hanna, et al.: De novo design of protein logic gates. *Science* 2020, 368:78–84.

107. Edgell CL, Smith AJ, Beesley JL, Savery NJ, Woolfson DN: De Novo Designed Protein-Interaction Modules for In-Cell Applications. *ACS Synth Biol* 2020, 9:427–436.

108. Lajoie MJ, Boyken SE, Salter AI, Bruffey J, Rajan A, Langan RA, Olshefsky A, Muhunthan V, Bick MJ, Gewe M, et al.: Designed protein logic to target cells with precise combinations of surface antigens. *Science* 2020, 369:1637–1643.

109. Shaul Y, Schreiber G: Exploring the charge space of protein–protein association: A proteomic study. *Proteins Struct Funct Bioinforma* 2005, 60:341–352.

110. Schreiber G: CHAPTER 1 Protein–Protein Interaction Interfaces and their Functional Implications. In *Protein–Protein Interaction Regulators*. . The Royal Society of Chemistry; 2021:1–24.

111. Chao G, Lau WL, Hackel BJ, Sazinsky SL, Lippow SM, Wittrup KD: Isolating and engineering human antibodies using yeast surface display. *Nat Protoc* 2006, 1:755–768.

112. Packer MS, Liu DR: Methods for the directed evolution of proteins. *Nat Rev Genet* 2015, 16:379–394.

113. Kieke MC, Cho BK, Boder ET, Kranz DM, Wittrup KD: Isolation of anti-T cell receptor scFv mutants by yeast surface display. *Protein Eng Des Sel* 1997, 10:1303–1310.

114. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y: The I-TASSER Suite: protein structure and function prediction. *Nat Methods* 2015, 12:7–8.

115. Schreiber G, Fleishman SJ: Computational design of protein–protein interactions. *Catal Regul Protein-Protein Interact* 2013, 23:903–910.

116. Liu S, Liu S, Zhu X, Liang H, Cao A, Chang Z, Lai L: Nonnatural protein–protein interaction-pair design by key residues grafting. *Proc Natl Acad Sci* 2007, 104:5330.

117. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. *Nucleic Acids Res* 2000, 28:235–242.

118. Kortemme T, Baker D: A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* 2002, 99:14116–14121.

119. Ofek G, Guenaga FJ, Schief WR, Skinner J, Baker D, Wyatt R, Kwong PD: Elicitation of structure-specific antibodies by epitope scaffolds. *Proc Natl Acad Sci U S A* 2010, 107:17880–17887.

120. McLellan JS, Correia BE, Chen M, Yang Y, Graham BS, Schief WR, Kwong PD: Design and Characterization of Epitope-Scaffold Immunogens That Present the Motavizumab Epitope from Respiratory Syncytial Virus. *J Mol Biol* 2011, 409:853–866.

121. Azoitei ML, Ban Y-EA, Julien J-P, Bryson S, Schroeter A, Kalyuzhniy O, Porter JR, Adachi Y, Baker D, Pai EF, et al.: Computational Design of High-Affinity Epitope Scaffolds by Backbone Grafting of a Linear Epitope. *J Mol Biol* 2012, 415:175–192.

122. Azoitei Mihai L., Correia Bruno E., Ban Yih-En Andrew, Carrico Chris, Kalyuzhniy Oleksandr, Chen Lei, Schroeter Alexandria, Huang Po-Ssu, McLellan Jason S., Kwong Peter D., et al.: Computation-Guided Backbone Grafting of a Discontinuous Motif onto a Protein Scaffold. *Science* 2011, 334:373–376.

123. Procko E, Berguig GY, Shen BW, Song Y, Frayo S, Convertine AJ, Margineantu D, Booth G, Correia BE, Cheng Y, et al.: A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells. *Cell* 2014, 157:1644–1656.

124. Bryan CM, Rocklin GJ, Bick MJ, Ford A, Majri-Morrison S, Kroll AV, Miller CJ, Carter L, Goreshnik I, Kang A, et al.: Computational design of a synthetic PD-1 agonist. *Proc Natl Acad Sci* 2021, 118:e2102164118.

125. Bonet J, Wehrle S, Schriever K, Yang C, Billet A, Sesterhenn F, Scheck A, Sverrisson F, Veselkova B, Vollers S, et al.: Rosetta FunFolDes - A general framework for the computational design of functional proteins. *PLoS Comput Biol* 2018, 14:e1006623–e1006623.

126. Linsky TW, Vergara R, Codina N, Nelson JW, Walker MJ, Su W, Barnes CO, Hsiang T-Y, Esser-Nobis K, Yu K, et al.: De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. *Science* 2020, 370:1208.

127. Fletcher JM, Horner KA, Bartlett GJ, Rhys GG, Wilson AJ, Woolfson DN: De novo coiled-coil peptides as scaffolds for disrupting protein–protein interactions. *Chem Sci* 2018, 9:7656–7665.

128. Fleishman SJ, Corn JE, Strauch E-M, Whitehead TA, Karanicolas J, Baker D: Hotspot-Centric De Novo Design of Protein Binders. *J Mol Biol* 2011, 413:1047–1062.

129. Jha RK, Leaver-Fay A, Yin S, Wu Y, Butterfoss GL, Szyperski T, Dokholyan NV, Kuhlman B: Computational Design of a PAK1 Binding Protein. *J Mol Biol* 2010, 400:257–270.

130. Procko E, Hedman R, Hamilton K, Seetharaman J, Fleishman SJ, Su M, Aramini J, Kornhaber G, Hunt JF, Tong L, et al.: Computational Design of a Protein-Based Enzyme Inhibitor. *J Mol Biol* 2013, 425:3563–3575.

131. Dang LT, Miao Y, Ha A, Yuki K, Park K, Janda CY, Jude KM, Mohan K, Ha N, Vallon M, et al.: Receptor subtype discrimination using extensive shape complementary designed interfaces. *Nat Struct Mol Biol* 2019, 26:407–414.

132. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch E-M, Wilson IA, Baker D: Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. *Science* 2011, 332:816.

133. Dou J, Vorobieva AA, Sheffler W, Doyle LA, Park H, Bick MJ, Mao B, Foight GW, Lee MY, Gagnon LA, et al.: De novo design of a fluorescence-activating β-barrel. *Nature* 2018, 561:485–491.

134. Cao L, Coventry B, Goreshnik I, Huang B, Park JS, Jude KM, Marković I, Kadam RU, Verschueren KHG, Verstraete K, et al.: Robust de novo design of protein binding proteins from target structural information alone. *bioRxiv* 2021, doi:10.1101/2021.09.04.459002.

135. Stranges PB, Kuhlman B: A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Sci Publ Protein Soc* 2013, 22:74–82.

136. Boyken Scott E., Chen Zibo, Groves Benjamin, Langan Robert A., Oberdorfer Gustav, Ford Alex, Gilmore Jason M., Xu Chunfu, DiMaio Frank, Pereira Jose Henrique, et al.: De novo design of protein homo-oligomers with modular hydrogen-bond network–mediated specificity. *Science* 2016, 352:680–687.

137. Coventry B, Baker D: Protein sequence optimization with a pairwise decomposable penalty for buried unsatisfied hydrogen bonds. *PLOS Comput Biol* 2021, 17:e1008061.

138. Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM, Correia BE: Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 2020, 17:184–192.

139. MSV J: Here Are Three Factors That Accelerate The Rise Of Artificial Intelligence. *Forbes* 2018,

140. Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P: Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Process Mag* 2017, 34:18–42.

141. Sanner MF, Olson AJ, Spehner J-C: Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* 1996, 38:305–320.

142. Krizhevsky A, Sutskever I, Hinton GE: ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017, 60:84–90.

143. Urquhart L: Top drugs and companies by sales in 2018. *Nat Rev Drug Discov* 2019, doi:10.1038/d41573-019-00049-0.

144. Liu JKH: The history of monoclonal antibody development - Progress, remaining challenges and future innovations. *Ann Med Surg* 2014, 3:113–116.

145. Baldo BA: Side Effects of Cytokines Approved for Therapy. *Drug Saf* 2014, 37:921–943.

146. Hansel TT, Kropshofer H, Singer T, Mitchell JA, George AJT: The safety and side effects of monoclonal antibodies. *Nat Rev Drug Discov* 2010, 9:325–338.

147. Bonati L, Tang L: Cytokine engineering for targeted cancer immunotherapy. *Curr Opin Chem Biol* 2021, 62:43–52.

148. Miller IC, Zamat A, Sun L-K, Phuengkham H, Harris AM, Gamboa L, Yang J, Murad JP, Priceman SJ, Kwong GA: Enhanced intratumoural activity of CAR T cells engineered to produce immunomodulators under photothermal control. *Nat Biomed Eng* 2021, 5:1348–1359.

149. Zhao Y, Xie Y-Q, Van Herck S, Nassiri S, Gao M, Guo Y, Tang L: Switchable immune modulator for tumor-specific activation of anticancer immunity. *Sci Adv* 2021, 7:eabg7291.

150. Martinko AJ, Simonds EF, Prasad S, Ponce A, Bracken CJ, Wei J, Wang Y-H, Chow T-L, Huang Z, Evans MJ, et al.: Switchable assembly and function of antibody complexes in vivo using a small molecule. *Proc Natl Acad Sci* 2022, 119:e2117402119.

151. Rivera VM, Wang X, Wardwell S, Courage NL, Volchuk A, Keenan T, Holt DA, Gilman M, Orci L, Cerasoli F, et al.: Regulation of Protein Secretion Through Controlled Aggregation in the Endoplasmic Reticulum. *Science* 2000, 287:826–830.

152. Rollins CT, Rivera VM, Woolfson DN, Keenan T, Hatada M, Adams SE, Andrade LJ, Yaeger D, Van Schravendijk MR, Holt DA, et al.: A ligand-reversible dimerization system for controlling protein–protein interactions. *Proc Natl Acad Sci* 2000, 97:7096–7101.

153. Ran X, Gestwicki JE: Inhibitors of protein–protein interactions (PPIs): an analysis of scaffold choices and buried surface area. *Curr Opin Chem Biol* 2018, 44:75–86.

154. Arkin MR, Tang Y, Wells JA: Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing toward the Reality. *Chem Biol* 2014, 21:1102–1114.

155. Boncompain G, Divoux S, Gareil N, De Forges H, Lescure A, Latreche L, Mercanti V, Jollivet F, Raposo G, Perez F: Synchronization of secretory protein traffic in populations of cells. *Nat Methods* 2012, 9:493–498.

156. Unverdorben F, Richter F, Hutt M, Seifert O, Malinge P, Fischer N, Kontermann RE: Pharmacokinetic properties of IgG and various Fc fusion proteins in mice. *mAbs* 2016, 8:120–128.

157. Oostindie SC, Lazar GA, Schuurman J, Parren PWHI: Avidity in antibody effector functions and biotherapeutic drug design. *Nat Rev Drug Discov* 2022, 21:715–735.

158. Lu LL, Suscovich TJ, Fortune SM, Alter G: Beyond binding: antibody effector functions in infectious diseases. *Nat Rev Immunol* 2018, 18:46–61.

159. Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, Gonzalez R, Robert C, Schadendorf D, Hassel JC, et al.: Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. *N Engl J Med* 2010, 363:711–723.

160. Li Z, Krippendorff B-F, Sharma S, Walz AC, Lavé T, Shah DK: Influence of molecular size on tissue distribution of antibody fragments. *mAbs* 2016, 8:113–119.

161. Lewis GD, Figari I, Fendly B, Lee Wong W, Carter P, Gorman C, Shepard HM: Differential responses of human tumor cell lines to anti-p185HER2 monoclonal antibodies. *Cancer Immunol Immunother* 1993, 37:255–263.

162. Cao L, Goreshnik I, Coventry B, Case JB, Miller L, Kozodoy L, Chen RE, Carter L, Walls AC, Park Y-J, et al.: De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* 2020, 370:426–431.

163. Silva D-A, Yu S, Ulge UY, Spangler JB, Jude KM, Labão-Almeida C, Ali LR, Quijano-Rubio A, Ruterbusch M, Leung I, et al.: De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* 2019, 565:186–191.

164. Marcandalli J, Fiala B, Ols S, Perotti M, De Van Der Schueren W, Snijder J, Hodge E, Benhaim M, Ravichandran R, Carter L, et al.: Induction of Potent Neutralizing Antibody Responses by a Designed Protein Nanoparticle Vaccine for Respiratory Syncytial Virus. *Cell* 2019, 176:1420-1431.e17.

165. Giordano-Attianese G, Gainza P, Gray-Gaillard E, Cribioli E, Shui S, Kim S, Kwak M-J, Vollers S, Corria Osorio ADJ, Reichenbach P, et al.: A computationally designed chimeric antigen receptor provides a small-molecule safety switch for T-cell therapy. *Nat Biotechnol* 2020, 38:426–432.

166. Cao L, Coventry B, Goreshnik I, Huang B, Sheffler W, Park JS, Jude KM, Marković I, Kadam RU, Verschueren KHG, et al.: Design of protein-binding proteins from the target structure alone. *Nature* 2022, 605:551–560.

167. Marchand A, Van Hall-Beauvais AK, Correia BE: Computational design of novel protein–protein interactions – An overview on methodological approaches and applications. *Curr Opin Struct Biol* 2022, 74:102370.

168. DeGrado WF, Wasserman ZR, Lear JD: Protein Design, a Minimalist Approach. *Science* 1989, 243:622–628.

169. Foight GW, Wang Z, Wei CT, Jr Greisen P, Warner KM, Cunningham-Bryant D, Park K, Brunette TJ, Sheffler W, Baker D, et al.: Multi-input chemical control of protein dimerization for programming graded cellular responses. *Nat Biotechnol* 2019, 37:1209–1216.

170. Chothia C, Janin J: Principles of protein–protein recognition. *Nature* 1975, 256:705–708.

171. Conte LL, Chothia C, Janin J: The atomic structure of protein-protein recognition sites. Edited by A. R. Fersht. *J Mol Biol* 1999, 285:2177–2198.

172. Chakrabarti P, Janin J: Dissecting protein-protein recognition sites. *Proteins Struct Funct Genet* 2002, 47:334–343.

173. Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R: Signature Verification Using a "Siamese" Time Delay Neural Network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*. . Morgan Kaufmann Publishers Inc.; 1993:737–744.

174. Pierce BG, Hourai Y, Weng Z: Accelerating Protein Docking in ZDOCK Using an Advanced 3D Convolution Library. *PLOS ONE* 2011, 6:e24657.

175. Pierce B, Weng Z: A combination of rescoring and refinement significantly improves protein docking performance. *Proteins Struct Funct Bioinforma* 2008, 72:270–279.

176. Lensink MF, Velankar S, Wodak SJ: Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins Struct Funct Bioinforma* 2017, 85:359–377.

177. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S, Yim J, et al.: *Protein complex prediction with AlphaFold-Multimer*. Bioinformatics; 2021.

178. Ramaraj T, Angel T, Dratz EA, Jesaitis AJ, Mumey B: Antigen–antibody interface properties: Composition, residue interactions, and features of 53 non-redundant structures. *Biochim Biophys Acta BBA - Proteins Proteomics* 2012, 1824:520–532.

179. Fenwick C, Turelli P, Perez L, Pellaton C, Esteves-Leuenberger L, Farina A, Campos J, Lana E, Fiscalini F, Raclot C, et al.: A highly potent antibody effective against SARS-CoV-2 variants of concern. *Cell Rep* 2021, 37:109814–109814.

180. Francisco LM, Sage PT, Sharpe AH: The PD-1 pathway in tolerance and autoimmunity: PD-1 pathway, Tregs, and autoimmune diseases. *Immunol Rev* 2010, 236:219–242.

181. Zak KM, Grudnik P, Magiera K, Dömling A, Dubin G, Holak TA: Structural Biology of the Immune Checkpoint Receptor PD-1 and Its Ligands PD-L1/PD-L2. *Structure* 2017, 25:1163–1174.

182. Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, Powderly JD, Carvajal RD, Sosman JA, Atkins MB, et al.: Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer. *N Engl J Med* 2012, 366:2443–2454.

183. Maute RL, Gordon SR, Mayer AT, McCracken MN, Natarajan A, Ring NG, Kimura R, Tsai JM, Manglik A, Kruse AC, et al.: Engineering high-affinity PD-1 variants for optimized immunotherapy and immuno-PET imaging. *Proc Natl Acad Sci* 2015, 112:E6506–E6514.

184. Nooren IMA, Thornton JM: Diversity of protein-protein interactions. *EMBO J* 2003, 22:3486–3492.

185. Fleishman SJ, Baker D: Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution. *Cell* 2012, 149:262–273.

186. Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, Bafna K, Norn C, Kang A, Bera AK, et al.: De novo protein design by deep network hallucination. *Nature* 2021, 600:547–552.

187. Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, Nie W, Liu Y, Wang R: PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* 2015, 31:405–412.

188. Zhou Q: PyMesh—Geometry processing library for Python. 2019,

189. Dijkstra EW: A note on two problems in connexion with graphs. *Numer Math* 1959, 1:269–271.

190. Ingwer Borg, Patrick JF Groenen: *Modern Multidimensional Scaling Theory and applications*. Springer Science and Business Media; 2005.

191. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al.: Scikit-Learn: Machine Learning in Python. *J Mach Learn Res* 2011, 12:2825–2830.

192. Koenderink JJ, van Doorn AJ: Surface shape and curvature scales. *Image Vis Comput* 1992, 10:557–564.

193. Yin Shuangye, Proctor Elizabeth A., Lugovskoy Alexey A., Dokholyan Nikolay V.: Fast screening of protein surfaces using geometric invariant fingerprints. *Proc Natl Acad Sci* 2009, 106:16622–16626.

194. Kyte J, Doolittle RF: A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982, 157:105–132.

195. Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE, Brookes DH, Wilson L, Chen J, Liles K, et al.: Improvements to the APBS biomolecular solvation software suite. *Protein Sci* 2018, 27:112–128.

196. Morozov AV, Kortemme T: Potential Functions for Hydrogen Bonds in Protein Structure Prediction and Design. In *Advances in Protein Chemistry*. . Academic Press; 2005:1–38.

197. Monti F, Boscaini D, Masci J, Rodolà E, Svoboda J, Bronstein MM: Geometric deep learning on graphs and manifolds using mixture model CNNs. 2016, doi:10.48550/ARXIV.1611.08402.

198. Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A: PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. *Nucleic Acids Res* 2014, 42:W285–W289.

199. Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jiménez-García B, Bates PA, Fernandez-Recio J, et al.: Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol* 2015, 427:3031–3041.

200. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM: SAbDab: the structural antibody database. *Nucleic Acids Res* 2014, 42:D1140–D1146.

201. Diederik P. Kingma, Jimmy Ba: Adam: A Method for Stochastic Optimization. *ArXiv Prepr* 2014,

202. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, et al.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2016, doi:10.48550/ARXIV.1603.04467.

203. J. Svoboda, J. Masci, M. M. Bronstein: Palmprint recognition via discriminative index learning. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. . 2016:4232–4237.

204. Zhou Q-Y, Park J, Koltun V: Open3D: A Modern Library for 3D Data Processing. 2018, doi:10.48550/ARXIV.1801.09847.

205. Frishman D, Argos P: Knowledge-based protein secondary structure assignment. *Proteins Struct Funct Bioinforma* 1995, 23:566–579.

206. Zhou J, Grigoryan G: Rapid search for tertiary fragments reveals protein sequence–structure relationships. *Protein Sci* 2015, 24:508–524.

207. Duhovny D, Nussinov R, Wolfson HJ: Efficient Unbound Docking of Rigid Molecules. In *Algorithms in Bioinformatics*. Edited by Guigó R, Gusfield D. Springer Berlin Heidelberg; 2002:185–200.

208. Chen R, Li L, Weng Z: ZDOCK: An initial-stage protein-docking algorithm. *Proteins Struct Funct Bioinforma* 2003, 52:80–87.

209. Cox MAA, Cox TF: Multidimensional Scaling. In *Handbook of Data Visualization*. Edited by Chen C, Härdle W, Unwin A. Springer Berlin Heidelberg; 2008:315–347.

210. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al.: AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022, 50:D439–D444.

211. Liu C, Lu J, Tian H, Du W, Zhao L, Feng J, Yuan D, Li Z: Increased expression of PD-L1 by the human papillomavirus 16 E7 oncoprotein inhibits anticancer immunity. *Mol Med Rep* 2017, 15:1063–1070.

212. Kabsch W: *XDS*. *Acta Crystallogr D Biol Crystallogr* 2010, 66:125–132.

213. Otwinowski Z, Minor W: Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzym* 1997, 276:307–326.

214. Emsley P, Cowtan K: Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 2004, 60:2126–2132.

215. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, et al.: PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 2010, 66:213–221.

216. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC: MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 2010, 66:12–21.

217. Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA: cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* 2017, 14:290–296.

218. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: UCSF Chimera—A visualization system for exploratory research and analysis. *J Comput Chem* 2004, 25:1605–1612.

219. Emsley P, Lohkamp B, Scott WG, Cowtan K: Features and development of ıt Coot. *Acta Crystallogr Sect D* 2010, 66:486–501.

220. Afonine PV, Poon BK, Read RJ, Sobolev OV, Terwilliger TC, Urzhumtsev A, Adams PD: Real-space refinement in ıt PHENIX for cryo-EM and crystallography. *Acta Crystallogr Sect D* 2018, 74:531–544.

221. Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB, Croll TI, Hintze B, Hung L-W, Jain S, McCoy AJ, et al.: Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in ıt Phenix. *Acta Crystallogr Sect D* 2019, 75:861–877.

222. Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE: UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* 2018, 27:14–25.

223. Fleishman SJ, Leaver-Fay A, Corn JE, Strauch E-M, Khare SD, Koga N, Ashworth J, Murphy P, Richter F, Lemmon G, et al.: RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PLOS ONE* 2011, 6:e20161.

224. Xie Z, Deng X, Shu K: Prediction of Protein–Protein Interaction Sites Using Convolutional Neural Network and Improved Data Sets. *Int J Mol Sci* 2020, 21:467.

225. Reddy M, Eirikis E, Davis C, Davis HM, Prabhakar U: Comparative analysis of lymphocyte activation marker expression and cytokine secretion profile in stimulated human peripheral blood mononuclear cell cultures: an in vitro model to monitor cellular immune function. *J Immunol Methods* 2004, 293:127–142.

226. Otano I, Azpilikueta A, Glez-Vaz J, Alvarez M, Medina-Echeverz J, Cortés-Domínguez I, Ortiz-de-Solorzano C, Ellmark P, Fritzell S, Hernandez-Hoyos G, et al.: CD137 (4-1BB) costimulation of CD8+ T cells is more potent when provided in cis than in trans with respect to CD3-TCR stimulation. *Nat Commun* 2021, 12:7296.

227. Liao W, Lin J-X, Leonard WJ: IL-2 family cytokines: new insights into the complex roles of IL-2 as a broad regulator of T helper cell differentiation. *Curr Opin Immunol* 2011, 23:598–604.

228. Tan S, Zhang H, Chai Y, Song H, Tong Z, Wang Q, Qi J, Wong G, Zhu X, Liu WJ, et al.: An unexpected N-terminal loop in PD-1 dominates binding by nivolumab. *Nat Commun* 2017, 8:14369.

229. Yu X, Orr CM, Chan HTC, James S, Penfold CA, Kim J, Inzhelevskaya T, Mockridge CI, Cox KL, Essex JW, et al.: Reducing affinity as a strategy to boost immunomodulatory antibody agonism. *Nature* 2023, 614:539–547.

230. Janin J, Bahadur RP, Chakrabarti P: Protein–protein interaction and quaternary structure. *Q Rev Biophys* 2008, 41:133–180.

231. Monod J, Changeux J-P, Jacob F: Allosteric proteins and cellular control systems. *J Mol Biol* 1963, 6:306–329.

232. Seet BT, Dikic I, Zhou M-M, Pawson T: Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* 2006, 7:473–483.

233. Patel D, Kopec J, Fitzpatrick F, McCorvie TJ, Yue WW: Structural basis for ligand-dependent dimerization of phenylalanine hydroxylase regulatory domain. *Sci Rep* 2016, 6:23748.

234. Schlessinger J: Ligand-Induced, Receptor-Mediated Dimerization and Activation of EGF Receptor. *Cell* 2002, 110:669–672.

235. Oleinikovas V, Gainza P, Ryckmans T, Fasching B, Thomä NH: From Thalidomide to Rational Molecular Glue Design for Targeted Protein Degradation. *Annu Rev Pharmacol Toxicol* 2024, 64:291–312.

236. Schreiber SL: The Rise of Molecular Glues. *Cell* 2021, 184:3–9.

237. Shui S, Buckley S, Scheller L, Correia BE: Rational design of small-molecule responsive protein switches. *Protein Sci* 2023, 32:e4774.

238. Scheller L: Synthetic Receptors for Sensing Soluble Molecules with Mammalian Cells. In *Mammalian Cell Engineering*. Edited by Kojima R. Springer US; 2021:15–33.

239. Taylor ND, Garruss AS, Moretti R, Chan S, Arbing MA, Cascio D, Rogers JK, Isaacs FJ, Kosuri S, Baker D, et al.: Engineering an allosteric transcription factor to respond to new ligands. *Nat Methods* 2016, 13:177–183.

240. Békés M, Langley DR, Crews CM: PROTAC targeted protein degraders: the past is prologue. *Nat Rev Drug Discov* 2022, 21:181–200.

241. Wells JA, Kumru K: Extracellular targeted protein degradation: an emerging modality for drug discovery. *Nat Rev Drug Discov* 2023, doi:10.1038/s41573-023-00833-z.

242. Jan M, Scarfò I, Larson RC, Walker A, Schmidts A, Guirguis AA, Gasser JA, Słabicki M, Bouffard AA, Castano AP, et al.: Reversible ON- and OFF-switch chimeric antigen receptors controlled by lenalidomide. *Sci Transl Med* 2021, 13:eabb6295.

243. Ishida M, Watanabe H, Takigawa K, Kurishita Y, Oki C, Nakamura A, Hamachi I, Tsukiji S: Synthetic Self-Localizing Ligands That Control the Spatial Location of Proteins in Living Cells. *J Am Chem Soc* 2013, 135:12684–12689.

244. Gibson WJ, Sadagopan A, Shoba VM, Choudhary A, Meyerson M, Schreiber SL: Bifunctional Small Molecules That Induce Nuclear Localization and Targeted Transcriptional Regulation. *J Am Chem Soc* 2023, 145:26028–26037.

245. Ng CSC, Liu A, Cui B, Banik SM: *Targeted Protein Relocalization via Protein Transport Coupling*. Biochemistry; 2023.

246. Marchand A, Bonati L, Shui S, Scheller L, Gainza P, Rosset S, Georgeon S, Tang L, Correia BE: Rational Design of Chemically Controlled Antibodies and Protein Therapeutics. *ACS Chem Biol* 2023, 18:1259–1265.

247. Mata M, Gerken C, Nguyen P, Krenciute G, Spencer DM, Gottschalk S: Inducible Activation of MyD88 and CD40 in CAR T Cells Results in Controllable and Potent Antitumor Activity in Preclinical Solid Tumor Models. *Cancer Discov* 2017, 7:1306–1319.

248. Spencer DM, Wandless TJ, Schreiber SL, Crabtree GR: Controlling Signal Transduction with Synthetic Ligands. *Science* 1993, 262:1019–1024.

249. Glasgow AA, Huang Y-M, Mandell DJ, Thompson M, Ritterson R, Loshbaugh AL, Pellegrino J, Krivacic C, Pache RA, Barlow KA, et al.: Computational design of a modular protein sense-response system. *Science* 2019, 366:1024–1028.

250. Gainza P, Wehrle S, Van Hall-Beauvais A, Marchand A, Scheck A, Harteveld Z, Buckley S, Ni D, Tan S, Sverrisson F, et al.: De novo design of protein interactions with learned surface fingerprints. *Nature* 2023, 617:176–184.

251. Roberts AW, Davids MS, Pagel JM, Kahl BS, Puvvada SD, Gerecitano JF, Kipps TJ, Anderson MA, Brown JR, Gressick L, et al.: Targeting BCL2 with Venetoclax in Relapsed Chronic Lymphocytic Leukemia. *N Engl J Med* 2016, 374:311–322.

252. Arevalo JH, Stura EA, Taussig MJ, Wilson IA: Three-dimensional Structure of an Anti-steroid Fab′ and Progesterone-Fab′ Complex. *J Mol Biol* 1993, 231:103–118.

253. Chen DZ, Patel DV, Hackbarth CJ, Wang W, Dreyer G, Young DC, Margolis PS, Wu C, Ni Z-J, Trias J, et al.: Actinonin, a Naturally Occurring Antibacterial Agent, Is a Potent Deformylase Inhibitor. *Biochemistry* 2000, 39:1256–1262.

254. Wildman SA, Crippen GM: Prediction of Physicochemical Parameters by Atomic Contributions. *J Chem Inf Comput Sci* 1999, 39:868–873.

255. Eisenberg D, Schwarz E, Komaromy M, Wall R: Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol* 1984, 179:125–142.

256. Najmanovich R, Kuttner J, Sobolev V, Edelman M: Side-chain flexibility in proteins upon ligand binding. *Proteins Struct Funct Genet* 2000, 39:261–268.

257. Bissantz C, Kuhn B, Stahl M: A Medicinal Chemist's Guide to Molecular Interactions. *J Med Chem* 2010, 53:5061–5084.

258. Shui S, Scheller L, Correia BE: Protein-based bandpass filters for controlling cellular signaling with chemical inputs. *Nat Chem Biol* 2023, doi:10.1038/s41589-023-01463-7.

259. Hussey BJ, McMillen DR: Programmable T7-based synthetic transcription factors. *Nucleic Acids Res* 2018, 46:9842–9854.

260. England CG, Ehlerding EB, Cai W: NanoLuc: A Small Luciferase Is Brightening Up the Field of Bioluminescence. *Bioconjug Chem* 2016, 27:1175–1187.

261. Rui H, Ashton KS, Min J, Wang C, Potts PR: Protein–protein interfaces in molecular glue-induced ternary complexes: classification, characterization, and prediction. *RSC Chem Biol* 2023, 4:192–215.

262. Ferreira De Freitas R, Schapira M: A systematic analysis of atomic protein–ligand interactions in the PDB. *MedChemComm* 2017, 8:1970–1981.

263. Orasch O, Weber N, Müller M, Amanzadi A, Gasbarri C, Trummer C: Protein–Protein Interaction Prediction for Targeted Protein Degradation. *Int J Mol Sci* 2022, 23:7033.

264. Nagy B, Szekeres-Barthó J, Kovács GL, Sulyok E, Farkas B, Várnagy Á, Vértes V, Kovács K, Bódis J: Key to Life: Physiological Role and Clinical Implications of Progesterone. *Int J Mol Sci* 2021, 22:11039.

265. Morgan RA, Yang JC, Kitano M, Dudley ME, Laurencot CM, Rosenberg SA: Case Report of a Serious Adverse Event Following the Administration of T Cells Transduced With a Chimeric Antigen Receptor Recognizing ERBB2. *Mol Ther* 2010, 18:843–851.

266. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M: ColabFold: making protein folding accessible to all. *Nat Methods* 2022, 19:679–682.

267. Landrum G: RDKit.org. *RDKit Open-Source Cheminformatics* [date unknown],

268. Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M: On the Art of Compiling and Using "Drug-Like" Chemical Fragment Spaces. *ChemMedChem* 2008, 3:1503–1507.

269. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA: Binding MOAD (Mother Of All Databases). *Proteins Struct Funct Bioinforma* 2005, 60:333–340.

270. Kartal Ö, Andres F, Lai MP, Nehme R, Cottier K: waveRAPID—A Robust Assay for High-Throughput Kinetic Screens with the Creoptix WAVEsystem. *SLAS Discov* 2021, 26:995–1003.

271. Macdonald LE, Meagher KA, Franklin MC, Levenkova N, Hansen J, Badithe AT, Zhong M, Krueger P, Rafique A, Tu N, et al.: Kappa-on-Heavy (KoH) bodies are a distinct class of fully-human antibody-like therapeutic agents with antigen-binding properties. *Proc Natl Acad Sci* 2020, 117:292–299.

272. Ellis RJ: Macromolecular crowding: obvious but underappreciated. *Trends Biochem Sci* 2001, 26:597–604.

273. Ross AB, Langer JD, Jovanovic M: Proteome Turnover in the Spotlight: Approaches, Applications, and Perspectives. *Mol Cell Proteomics MCP* 2021, 20:100016.

274. Morrison KL, Weiss GA: Combinatorial alanine-scanning. *Curr Opin Chem Biol* 2001, 5:302–307.

275. Barlow KA, Ó Conchúir S, Thompson S, Suresh P, Lucas JE, Heinonen M, Kortemme T: Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation. *J Phys Chem B* 2018, 122:5389–5399.

276. Sora V, Laspiur AO, Degn K, Arnaudi M, Utichi M, Beltrame L, De Menezes D, Orlandi M, Stoltze UK, Rigina O, et al.: RosettaDDGPrediction for high-throughput mutational scans: From stability to binding. *Protein Sci* 2023, 32:e4527.

277. Donaldson JM, Zer C, Avery KN, Bzymek KP, Horne DA, Williams JC: Identification and grafting of a unique peptide-binding site in the Fab framework of monoclonal antibodies. *Proc Natl Acad Sci* 2013, 110:17456–17461.

278. Scott JK, Smith GP: Searching for Peptide Ligands with an Epitope Library. *Science* 1990, 249:386–390.

279. Ramachandran S, Kota P, Ding F, Dokholyan NV: Automated minimization of steric clashes in protein structures. *Proteins Struct Funct Bioinforma* 2011, 79:261–270.

280. Chen TS, Keating AE: Designing specific protein–protein interactions using computation, experimental library screening, or integrated methods. *Protein Sci* 2012, 21:949–963.

281. Musil M, Konegger H, Hon J, Bednar D, Damborsky J: Computational Design of Stable and Soluble Biocatalysts. *ACS Catal* 2019, 9:1033–1054.

282. Magliery TJ: Protein stability: computation, sequence statistics, and new experimental methods. *Curr Opin Struct Biol* 2015, 33:161–168.

283. Sumida KH, Núñez-Franco R, Kalvet I, Pellock SJ, Wicky BIM, Milles LF, Dauparas J, Wang J, Kipnis Y, Jameson N, et al.: Improving Protein Expression, Stability, and Function with ProteinMPNN. *J Am Chem Soc* 2024, doi:10.1021/jacs.3c10941.

284. Goldenzweig A, Goldsmith M, Hill SE, Gertman O, Laurino P, Ashani Y, Dym O, Unger T, Albeck S, Prilusky J, et al.: Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol Cell* 2016, 63:337–346.

285. Torres SV, Leung PJY, Venkatesh P, Lutz ID, Hink F, Huynh H-H, Becker J, Yeh AH-W, Juergens D, Bennett NR, et al.: De novo design of high-affinity binders of bioactive helical peptides. *Nature* 2023, doi:10.1038/s41586-023-06953-1.

286. Sverrisson F, Feydy J, Correia BE, Bronstein MM: Fast end-to-end learning on protein surfaces. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. . IEEE; 2021:15267–15276.

287. Janson G, Valdes-Garcia G, Heo L, Feig M: Direct generation of protein conformational ensembles via machine learning. *Nat Commun* 2023, 14:774.

288. Wayment-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Hömberger M, Ovchinnikov S, Colwell L, Kern D: Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* 2023, doi:10.1038/s41586-023-06832-9.

289. Boyken SE, Chen Z, Groves B, Langan RA, Oberdorfer G, Ford A, Gilmore JM, Xu C, DiMaio F, Pereira JH, et al.: De novo design of protein homo-oligomers with modular hydrogen-bond network–mediated specificity. *Science* 2016, 352:680–687.

290. Ahmed MH, Spyrakis F, Cozzini P, Tripathi PK, Mozzarelli A, Scarsdale JN, Safo MA, Kellogg GE: Bound Water at Protein-Protein Interfaces: Partners, Roles and Hydrophobic Bubbles as a Conserved Motif. *PLoS ONE* 2011, 6:e24712.

291. Sheinerman FB, Honig B: On the Role of Electrostatic Interactions in the Design of Protein–Protein Interfaces. *J Mol Biol* 2002, 318:161–177.

292. Schreiber G, Haran G, Zhou H-X: Fundamental aspects of protein-protein association kinetics. *Chem Rev* 2009, 109:839–860.

293. Duan G, Walther D: The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. *PLOS Comput Biol* 2015, 11:e1004049.

294. Reis JM, Burns DC, Woolley GA: Optical Control of Protein–Protein Interactions via Blue Light-Induced Domain Swapping. *Biochemistry* 2014, 53:5008–5016.

295. Dumetz AC, Chockla AM, Kaler EW, Lenhoff AM: Effects of pH on protein–protein interactions and implications for protein phase behavior. *Biochim Biophys Acta BBA - Proteins Proteomics* 2008, 1784:600–610.

296. Lucas JE, Kortemme T: New computational protein design methods for de novo small molecule binding sites. *PLOS Comput Biol* 2020, 16:e1008178.

297. Polizzi NF, DeGrado WF: A defined structural unit enables de novo design of small-molecule–binding proteins. *Science* 2020, 369:1227–1233.

298. Dauparas J, Lee GR, Pecoraro R, An L, Anishchenko I, Glasscock C, Baker D: *Atomic context-conditioned protein sequence design using LigandMPNN*. Biochemistry; 2023.

299. Lim WA, June CH: The Principles of Engineering Immune Cells to Treat Cancer. *Cell* 2017, 168:724–740.

300. Lebozec K, Jandrot-Perrus M, Avenard G, Favre-Bulle O, Billiald P: Quality and cost assessment of a recombinant antibody fragment produced from mammalian, yeast and prokaryotic host cells: A case study prior to pharmaceutical development. *New Biotechnol* 2018, 44:31–40.

301. Baker M, Reynolds HM, Lumicisi B, Bryson CJ: Immunogenicity of protein therapeutics: The key causes, consequences and challenges. *Self/Nonself* 2010, 1:314–322.

302. Dieckhaus H, Brocidiacono M, Randolph N, Kuhlman B: *Transfer learning to leverage larger datasets for improved prediction of protein stability changes*. Bioinformatics; 2023.

303. Thieker DF, Maguire JB, Kudlacek ST, Leaver-Fay A, Lyskov S, Kuhlman B: Stabilizing proteins, simplified: A Rosetta-based webtool for predicting favorable mutations. *Protein Sci Publ Protein Soc* 2022, 31:e4428.

304. Rapp JT, Bremer BJ, Romero PA: Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat Chem Eng* 2024, 1:97–107.

## 6.2 Curriculum Vitae


# Anthony Marchand

Life sciences engineer and doctoral research-assistant

Date of birth: 12.05.1994 Nationality: Swiss Mobile: +41(0)76 450 36 17

E-mail: antho.marchand@gmail.com
LinkedIn: www.linkedin.com/in/anthony-marchand-CH


### Experience

| | | |
|---|---|---|
| Apr 2020 – May 2024 | Doctoral research-assistant<br>Laboratory of protein design and immunoeng. (EPFL) | Lausanne (CH) |
| Sep 2019 – Mar 2020 | Guest Scientist<br>Max Delbrück Center for molecular medicine | Berlin (DE) |
| Jul 2018 – Aug 2018 | Trainee<br>The Institute for Cancer Research | London (UK) |
| Feb 2018 – Jun 2018 | Project student<br>Pharmacology Institute (Bern University) | Bern (CH) |
| Aug 2017 – Jan 2018 | R&D Intern<br>AC Immune SA | Lausanne (CH) |
| Oct 2016 – Jul 2017 | Project student-assistant<br>Laboratory of intestinal immunology (EPFL) | Lausanne (CH) |
| Apr 2016 – Sep 2016 | Laboratory student-assistant<br>Laboratory unit of Prof. Gönczy (EPFL) | Lausanne (CH) |

### Education

| | | |
|---|---|---|
| Apr 2020 – May 2024 | PhD in Bioengineering & Biotechnology<br>Swiss Federal Institute of Technologies (EPFL) | Lausanne (CH) |
| Aug 2017 – Mar 2020 | M.Sc. in Life Sciences and Technologies (GPA of 5.78 / 6)<br>Swiss Federal Institute of Technologies (EPFL) | Lausanne (CH) |
| Sep 2014 – Jul 2017 | B.Sc. in Life Sciences and Technologies (GPA of 5.38 / 6)<br>Swiss Federal Institute of Technologies (EPFL) | Lausanne (CH) |

### Awards & fellowships

| | |
|---|---|
| May 2023 | Finn-Wold Young Investigator Award – The Protein Society |
| Oct 2020 | Mention of Excellence – Ecole polytechnique fédérale de Lausanne |
| May 2018 | Werner-Siemens excellence fellowship – Swiss Study Foundation |
| Dec 2014 | Academic excellence certificate – Swiss Study Foundation |
| Jul 2013 | Price for the best high school matura – Gymnase intercantonal de la Broye |

### Teaching

| | | |
|---|---|---|
| Sep 2020 – Jul 2022 | Biological Chemistry I & II | EPF Lausanne |
| Sep 2017 – Jan 2018 | C++ Programming | EPF Lausanne |
| Sep 2015 – Feb 2016 | Advanced General Chemistry | EPF Lausanne |

## Skills

| | |
|---|---|
| Computer language | C++, Python, Bash, HTML, CSS |
| Language | French (Native), English (C2), German (C1) |
| Software & tools | Microsoft office, FlowJo, GraphPad Prism, PyMOL, ChimeraX, Rosetta, AlpaFold |
| Biology technics | Molecular biology (PCR, Gibson, cloning,...), protein production & purification, FPLC, MALS, CD, SPR, BLI, WB, SDS-PAGE, ELISA, mammalian cell culture, transfection, yeast display library & *in vitro* evolution, flow cytometry, FACS, immunofluorescence, microscopy |

## Communications

| | |
|---|---|
| Sep 2023 | Poster & Talk – European RosettaCon (Leipzig, DE) |
| Sep 2023 | Talk – EPFL Bioengineering Days (Lausanne, CH) |
| Jul 2023 | Poster – The Protein Society Annual Symposium (Boston, USA) |
| May 2022 | Poster & talk – European RosettaCon (Warsaw, PL) – Prize for 2$^{nd}$ best talk |
| Apr 2022 | Poster & talk – NCCR Fellow Retreat (Grindelwald, CH) – Prize for best talk |

## Publications

Marchand A.*, Buckley S.*, Schneuing A.* *et al* (2024). Targeting protein-ligand neosurfaces using a generalizable deep learning approach. BioRxiv (Preprint). DOI: 10.1101/2024.03.25.585721

Gainza P.*, Wehrle S.*, Van Hall-Beauvais A. K.*, Marchand A.*, Scheck A. *et al* (2023). *De novo* design of protein interactions with learned surface fingerprints. Nature 617, 176-184.

Marchand A.* & Gainza P.* (2023). New protein–protein interactions designed by a computer. Nature Briefing.

Marchand A.* & Bonati L.* *et al* (2023). Rational design of chemically controlled antibodies and protein therapeutics. ACS Chem. Bio. 18 (6), 1259–1265.

Schweke, H. *et al* (2023). Discriminating physiological from non-physiological interfaces in structures of protein complexes: A community-wide study. Proteomics, 23(17).

Marchand A. *et al* (2022). Computational design of novel protein–protein interactions – An overview on methodological approaches and applications. Curr. Opinion in Struct. Bio., 74, 102370.

Marchand A. *et al* (2022). Validation of a novel high-throughput fluorescence-based assay for the identification of new compounds targeting the blood-feeding pathway of Hookworms. Pharmaceuticals, 15(6), 669.

## Patents

Marchand A. et al (2022). Chemically disruptable molecule switch and use thereof (European Patent Application Nr. 22215876.8). European Patent Office.

Gainza P. et al. (2022). *De novo* design of protein lnteractions with learned surface fingerprints (European Patent Application Nr. 22177692.5). European Patent Office.

## References

Prof. Dr. Bruno E. Correia (Associate Professor at EPFL) bruno.correia@epfl.ch

Dr. Pablo Gainza (Director at Monte Rosa therapeutics) pgainza@monterosatx.com