Probst *Journal of Cheminformatics* (2023) 15:113
https://doi.org/10.1186/s13321-023-00784-y

Journal of Cheminformatics

METHODOLOGY

Open Access

An explainability framework for deep learning on chemical reactions exemplified by enzyme-catalysed reaction classification

Daniel Probst[1]*

Abstract

Assigning or proposing a catalysing enzyme given a chemical or biochemical reaction is of great interest to life sciences and chemistry alike. The exploration and design of metabolic pathways and the challenge of finding more sustainable enzyme-catalysed alternatives to traditional organic reactions are just two examples of tasks that require an association between reaction and enzyme. However, given the lack of large and balanced annotated data sets of enzyme-catalysed reactions, assigning an enzyme to a reaction still relies on expert-curated rules and databases. Here, we present a data-driven explainable human-in-the-loop machine learning approach to support and ultimately automate the association of a catalysing enzyme with a given biochemical reaction. In addition, the proposed method is capable of predicting enzymes as candidate catalysts for organic reactions amendable to biocatalysis. Finally, the introduced explainability and visualisation methods can easily be generalised to support other machine-learning approaches involving chemical and biochemical reactions.

Keywords Machine learning, Enzymatic reactions, Explainable machine learning, Cheminformatics

Introduction

Most chemical reactions occurring within living organisms are catalysed by proteins or protein complexes called enzymes. Enzymes have a high substrate specificity, are not consumed or changed by the reaction, can be produced from renewable sources, and are themselves biodegradable. The identification and classification of associations between enzymes and the reactions they catalyse is of great interest across fields. For biologists, this includes the mapping of an enzyme and its substrates into a metabolic network, which is an integral part of connecting experimental data with established domain knowledge, the creation of genome-scale metabolic models to enable the analysis of omics data, and the computational design of synthetic metabolic pathways [1–4]. For medicinal chemists, the association of enzymes with their substrates is essential to the process of target-based drug discovery and to predict the metabolic fate of a drug candidate in an organism or specific organ [5, 6]. Meanwhile, process chemistry and material science are interested in the discovery and engineering of enzymes that catalyse known or novel chemical reactions to produce new materials or drugs, increase the efficiency of synthetic routes, or replace existing synthetic routes with more environmentally friendly enzyme-catalysed alternatives [7–9]. To support these efforts, a multitude of computational models have recently been proposed that enable the prediction of a required enzyme for a given reaction—generally relying on complex deep neural network architectures or expert-curated rules, both presupposing the existence of large data sets of biochemical reactions annotated with

*Correspondence:
Daniel Probst
daniel.probst@epfl.ch
[1] Signal Processing Laboratory 2, Institute of Electrical and Micro Engineering, School of Engineering, EPFL, Rte Cantonale, 1015 Lausanne, Vaud, Switzerland

correct enzyme classifications [10–14]. However, compared to other data in biology and chemistry, data sets containing annotated enzyme-catalysed reactions are exceedingly small and imbalanced in regard to enzyme-class distribution, complicating the application of data-hungry deep learning techniques, ultimately resulting in low predictive power—especially for underrepresented enzyme classes [12, 15–17].

The classification of enzymes and enzyme-catalysed reactions, and therefore their association, using EC numbers (see Methods for a detailed description of the EC number classification scheme) remains a manual task carried out by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) [18, 19]. With Rhea, the expert-curated knowledgebase of chemical and transport reactions of biological interest, an effort was started to annotate enzyme-catalysed reactions catalysed by enzymes found in UniProtKB with ChEBI identifiers and EC numbers [17]. While this effort started to provide much-needed additional data, the requirement for extensive involvement of experts in the curation and classification of enzyme–reaction associations causes slow growth of annotated enzyme-catalysed reaction data sets. The resulting lack of data shows a need for automation in order to enable further progress in

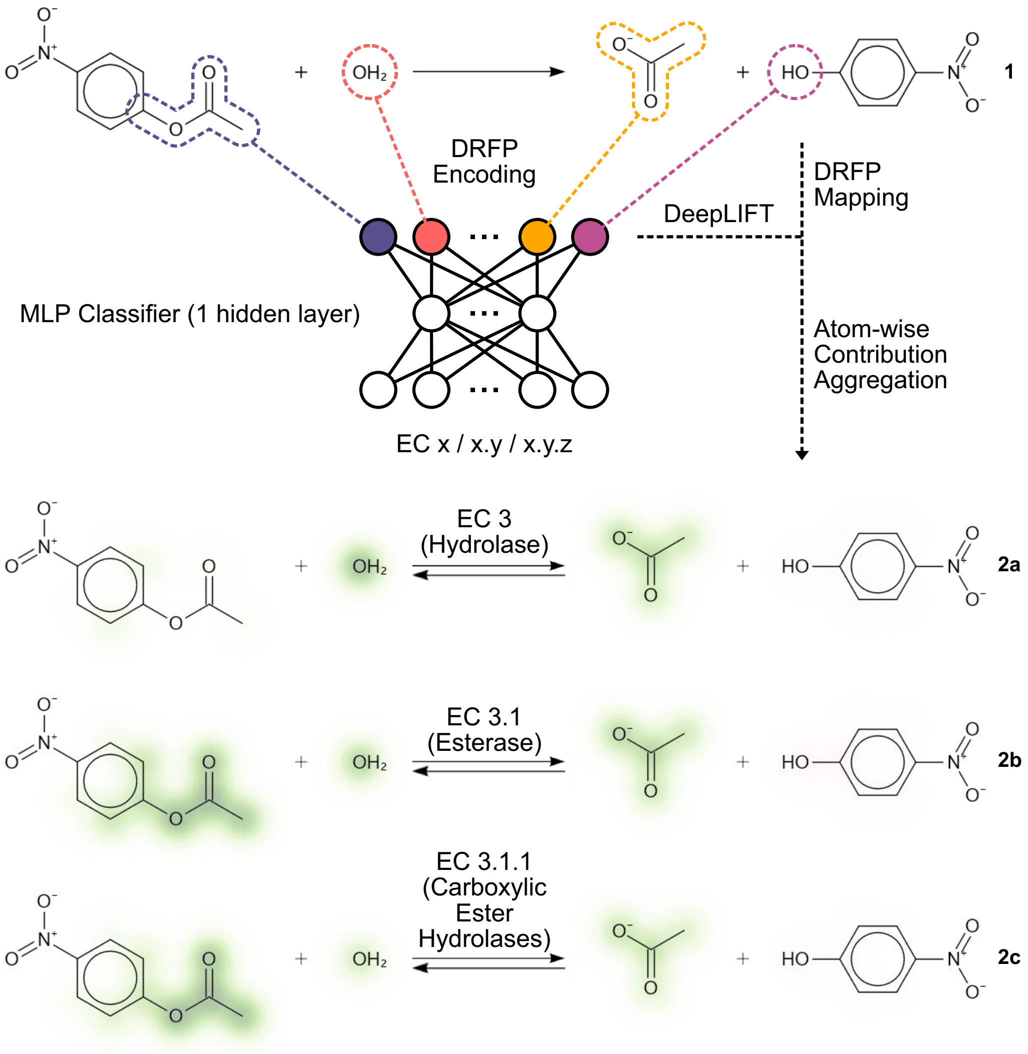

1
DRFP Encoding
DRFP Mapping
DeepLIFT
MLP Classifier (1 hidden layer)
Atom-wise Contribution Aggregation
EC x / x.y / x.y.z
EC 3 (Hydrolase)
2a
EC 3.1 (Esterase)
2b
EC 3.1.1 (Carboxylic Ester Hydrolases)
2c

Fig. 1 Explaining enzyme-catalysed reaction classifications. Enzyme-catalysed reactions (**1**) are classified using three different models. Model ECX predicts the class (**2a**), model ECXY predicts the class and subclass (**2b**), and model ECXYZ predicts the class, the subclass, and the sub-subclass (**2c**). The input to all models is a binary DRFP encoded reaction SMILES that allows a mapping between input and molecular fragment. Using Shapley additive explanations (SHAP), the inputs' influence on the classification can be traced back to a molecular fragment (**2a–c**)

machine learning involving enzyme-catalysed reactions so as to reach the impact and utility of comparable approaches such as recent advancements in computer-assisted synthesis planning (CASP) that can draw from data sets containing millions of organic reactions [20, 21].

The available methods for the automated, computational classification of enzymes and their association with reactions can be divided into two approaches: Methods predicting an enzyme classification from the amino acid sequence or tertiary structure of an enzyme [22–25], and methods predicting an enzyme classification from the reaction catalysed by an enzyme, where the enzyme classification is the aforementioned EC number. The focus of this article is on the latter of the two approaches, assuming no knowledge of an enzyme's sequence or structure. Many of the methods to computationally predict the class of an enzyme based on the catalysed reaction rely on the explicit mapping or typing of atoms, bonds, or functional groups; a predefined set of physicochemical and topological descriptors; or the balancing of reaction equations, which requires significant preprocessing and manual curation [26–31]. Other approaches include similarity searches based on molecular fingerprints and substructure matching algorithms [32, 33]. While the average accuracy of these data-driven methods has been high, they remain error-prone in edge cases and in predicting the many enzyme subclasses and sub-subclasses where little training data is available. A notable recent approach, setting the state-of-the-art, is the rule-based approach BridgIT by [34]. However, while this method has excellent accuracy and allows for explainability due to its rule-based rather than data-driven nature, the rules have to be created and continuously updated by experts with a deep knowledge of enzyme-catalysed reactions. This superior performance of rule-based compared to data-driven methods, the continuing need for expert curation of databases, and the existence of commission oversight show a need that goes beyond current approaches [17]. While the superior performance of rule-based methods can be attributed to a lack of sufficiently large and balanced data sets, the lack of adaption of an automated annotation process can have multiple possible reasons, including the absence of model explainability or trust, a lack of utility, or a failure to identify the limits of applicability [35, 36]. A potent solution to these is explainable machine learning, which can increase acceptance and usability as well as identify the limits and edge cases of a model, and has been widely used in genetics, healthcare, or education [37–39]. However, recent approaches in explainable machine learning in chemistry remain limited to models trained on single molecular entities rather than reactions [40–42].

Here, we introduce explainability to a multilayer perceptron that predicts the EC number of an enzyme given a reaction without the need for balancing the reaction or any other form of reaction curation. We provide a tool that can support and eventually fully automate enzyme–reaction association and classification by introducing three main advancements. (i) We report multiple models capable of predicting the classes, subclasses, and sub-subclasses of enzymes that catalyse a given reaction with overall accuracies of 98, 97, and 95 per cent, respectively, while requiring minimal training resources, enabling continuous retraining. (ii) By mapping the molecular fragments occurring in the reactions to the vector entries that act as input for a neural network classifier, we enable the use of the DeepLIFT algorithm (implemented as DeepSHAP) to annotate fragments and atoms with their respective classification contributions, providing chemical explainability for all described models (Fig. 1(1)). (iii) We develop and implement a generalised approach for the visualisation of numerical annotations for molecules and reactions, which we use to visualise the classification contributions that explain a model's perception of an input reaction (Fig. 1(2a–c)). Based on these advancements, we introduce an approach that can be used as a human-in-the-loop machine-learning solution for the transition to the fully automated annotation of enzyme-catalysed reactions. Furthermore, our system has the potential to support the prediction of catalysing enzyme candidates for organic reactions amendable to biocatalysis, making it a utility for the exploration of the enzymatic reaction space by chemists and biologists alike. The resulting models, data, and libraries are made accessible as a hosted web application and a locally installable Python package that includes a graphical and command-line interface in addition to a Python API. The modular architecture of the system allows for easy extension or the reuse of specific components in virtually all machine-learning tasks involving chemical or biochemical reactions.

Results

Enzyme classification using differential reaction fingerprints and a simple multilayer perceptron

Enzyme-catalysed reactions are stored as reaction SMILES, a string representation of a chemical reaction based on the molecular graphs of the participant substances [43]. As a first step, we encode these string representations into a binary vector using the differential reaction fingerprint (DRFP), which we recently showed to provide state-of-the-art reaction representations by example of reaction yield predictions, performing at least as well as DFT-derived descriptors or transformer-based methods on a yield prediction task for organic reactions

[44]. A TMAP visualisation shown in Fig. 2 exemplifies the ability of DRFP-encoded enzyme-catalysed reactions, extracted from Rhea ($n = 7010$), to be classified using the EC numbering scheme. Figure 2a is a visualisation of the entirety of the Rhea database coloured by the enzyme classes oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5), and ligases (EC 6). Figure 2b is a detailed view, displaying only transferases, coloured by the nine transferase subclasses found in Rhea: EC 2.1 (transferring one-carbon groups), EC 2.2 (transferring aldehyde or ketonic groups), EC 2.3 (acyltransferases), EC 2.4 (glycosyltransferases), EC 2.5 (transferring alkyl or aryl groups, with the exception of methyl groups), EC 2.6 (transferring nitrogenous groups), EC 2.7 (transperring phosphorus-containing groups), EC 2.8 (transferring sulfur-containing groups), and EC 2.9 (transferring selenium-containing groups). Finally, Fig. 2c shows the three sub-subclasses of glycosyltransferases, namely, hexosyltransferases (EC 2.4.1), pentosyltransferases (EC 2.4.2), and those transferring other glycosyl groups (EC 2.4.99). As these plots illustrate, DRFP is able to separate reactions according to all three levels of the EC classification. Furthermore, the plot shows that the fingerprint is capable of separating oxidoreductases (EC 1) exceedingly well, while there is a relative lack of distinct clustering for isomerases (EC 5). These findings reflect previous observations and are primarily caused by the diversity of reactions within a class [12].

Following the encoding of the enzyme-catalysed reactions extracted from Rhea as DRFP fingerprints, three distinct models were trained on the data: ECX, ECXY, and ECXYZ. While all models share a simple multilayer perceptron (MLP) architecture with a single hidden layer, ECX, ECXY, and ECXYZ were trained with labels representing x.-.- (classes), x.y.- (classes and subclasses), and x.y.z (classes, subclasses, and sub-subclasses), respectively. The specific architecture of the MLPs is described in Methods. In addition to the Rhea-extracted data, the procedure was repeated for our recently released ECREACT data set ($n = 81,205$), which extends the reactions from Rhea with reactions extracted from BRENDA, PathBank, and MetaNetX [12, 45–47]. The accuracies and f-scores of the models are shown in Table 1 together with the training times as well as the training and experimentation energy use. Results of ablation studies on models trained on the Rhea data set, where increasing fractions of the training set labels were shuffled, are shown in Table 2 and indicate robustness of the models to sporadically misclassified training data. Figure 3 shows the confusion matrices for the tests of ECX_{Rhea}, $\text{ECXY}_{\text{Rhea}}$, and $\text{ECXYZ}_{\text{Rhea}}$ in the first row (a-c) and the confusion matrices for $\text{ECX}_{\text{ECREACT}}$, $\text{ECXY}_{\text{ECREACT}}$, and $\text{ECXYZ}_{\text{ECREACT}}$ in the second row (d-f). Given the different sizes of the classes, subclasses, and sub-subclasses, the boxplots in Additional file 1: Fig. S2 yield further insights into the existence of challenging, low-accuracy cases of subclasses and sub-subclasses that have little effect on the overall accuracy when being evaluated together with the larger, better-trained classes but represent important edge-cases. For both Rhea and ECREACT, these challenging cases resulting in low prediction accuracies are generally caused by (sub-)subclasses with a small number of samples, rather than larger (sub-)subclasses with diverse samples (Additional file 1: Figs. S3 and S4). This behaviour follows the examples established in our previous work on biocatalysed synthesis planning [12]. The comparatively poor overall accuracy of isomerases (EC 5) can be explained by their function, which is to carry out modifications within a molecule, that would be assigned to other enzyme classes if they took place between two different molecules. This is shown in Fig. 3a,b where potential intramolecular transferases, which are classified as isomerases (EC 5), have been classified as intermolecular transferases (EC 2).

Overall, the observations show that a lack of training samples in under-represented classes, subclasses, and sub-subclasses leads to low accuracies in data-driven machine-learning approaches. Furthermore, the poorer performance of the models trained and tested on ECREACT compared to Rhea can likely be explained by the distribution of samples across classes as shown in Additional file 1: Figs. S3 and S4, following the observations made within each data set. The following sections will introduce the methods and tools necessary to facilitate the bridging of expert curation and data-driven learning of enzyme-catalysed reactions to speed up data curation and explain a model's behaviour to scientists.

Explaining classification using DeepLIFT

Using the differential reaction fingerprint to embed the reaction as a binary vector for input into the MLP allows for a mapping between binary input feature v_i and molecular fragment f_i. This enables the use of an arbitrary approach for explainable machine learning capable of determining or estimating the contribution of an input

(See figure on next page.)

Fig. 2 TMAPs of DRFP-encoded biochemical reactions extracted from the Rhea database coloured by the associated EC number. **a** All reactions coloured by enzyme class. **b** Transferase-catalysed reactions coloured by subclass. **c** Glycosyltransferase-catalysed reactions coloured by sub-subclass

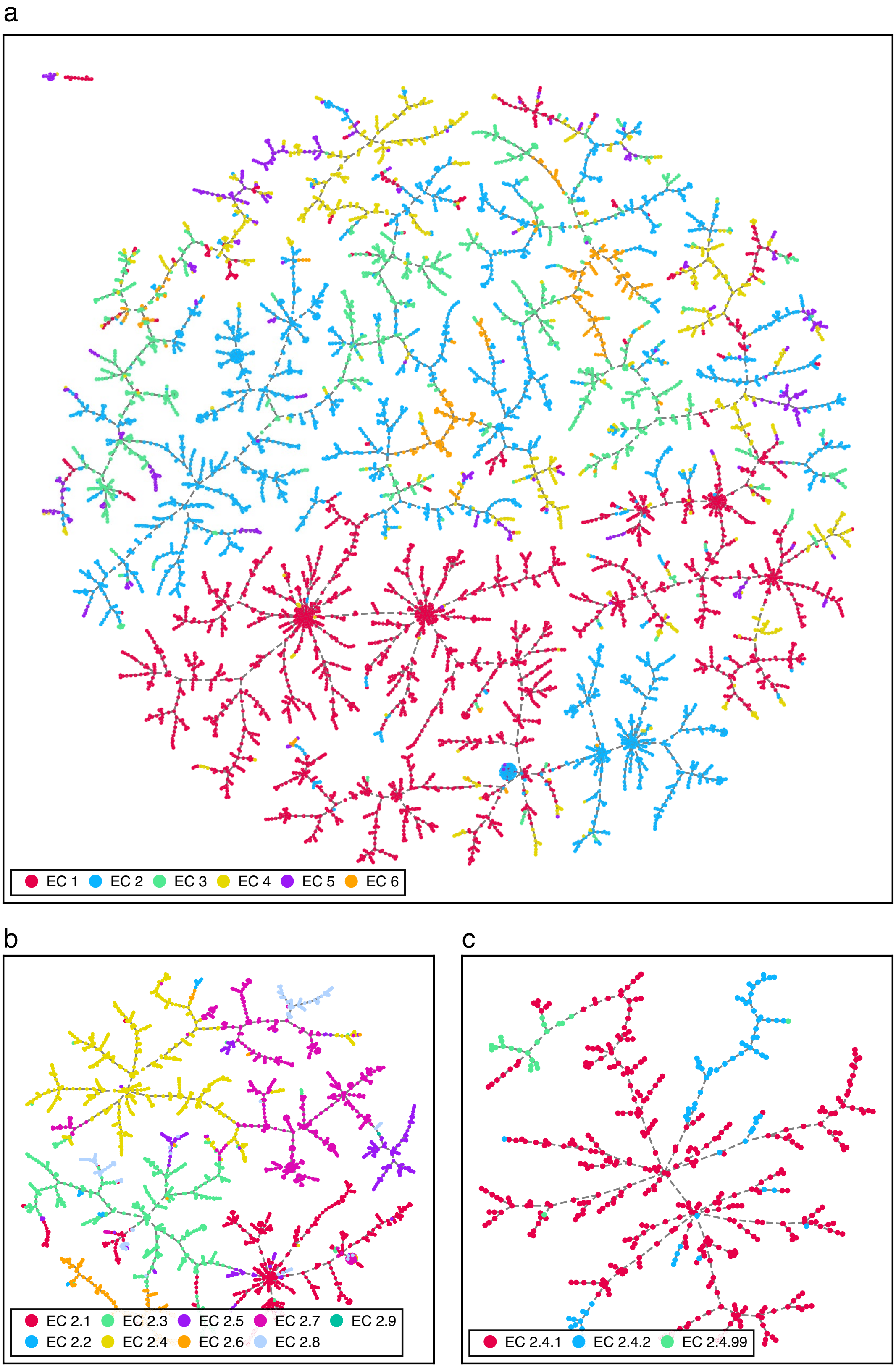

a
● EC 1 ● EC 2 ● EC 3 ● EC 4 ● EC 5 ● EC 6
b
● EC 2.1 ● EC 2.3 ● EC 2.5 ● EC 2.7 ● EC 2.9
● EC 2.2 ● EC 2.4 ● EC 2.6 ● EC 2.8
c
● EC 2.4.1 ● EC 2.4.2 ● EC 2.4.99

Fig. 2 (See legend on previous page.)

Table 1 Accuracies and F-scores of the different models trained on Rhea and ECREACT

Model	Accuracy	F-Score	Training Time	Energy Use
ECX_{Rhea}	$0.98 \pm eq0.00$	0.97 ± 0.01	100 s	43 Wh
$ECXY_{Rhea}$	0.96 ± 0.01	0.88 ± 0.02	160 s	70 Wh
$ECXYZ_{Rhea}$	0.95 ± 0.00	0.87 ± 0.02	190 s	82 Wh
$ECX_{ECREACT}$	0.98 ± 0.00	0.96 ± 0.00	2090 s	904 Wh
$ECXY_{ECREACT}$	0.95 ± 0.00	0.82 ± 0.01	2170 s	936 Wh
$ECXYZ_{ECREACT}$	0.93 ± 0.00	0.77 ± 0.01	2810 s	1,216 Wh

The energy use is calculated based on the energy use of the device (Dell XPS 15, i7-12700 H CPU, NVIDIA GeForce RTX 3050 Ti Laptop GPU) and includes the power usage of models trained for 4x cross-validation and four experiments with fingerprint variations. The hyperparameter values were taken from the previous work on organic reactions [44]. The total resulting energy consumption for model experimentation, training, and validation for this project was 3.25 kWh. Energy mix (2022): 65% hydro, 23% solar, and 12% other renewables

feature v_i to a resulting classification, to also quantify the influence of each molecular fragment f_i. Based on its ease of use and remarkable performance, we selected the DeepSHAP implementation of DeepLIFT to estimate the contributions of input features to the classification [48, 49]. Given the fragment contributions w_{f_i}, the atom contributions w_{a_j} are calculated by summing w_{f_i} of a given reaction ($v_i = 1$) that include atom a_j.

$$w_{a_j} = \sum_{i}^{N_f} w_{f_i}, \text{ if } a_j \in f_i \text{ and } v_i = 1 \quad (1)$$

This atom-wise weighing enables later visualisation of overlapping or contained fragments. The result of this operation can be seen in Fig. 4, where fragments with positive weights that contribute towards a certain classification are coloured green, while fragments with negative weights that contributed against a certain classification are shown in magenta. However, not only fragments present in the reaction can have an effect on the classification. The absence of a certain fragment can influence the classification as much as the presence of another. Therefore, the information on the contributions of absent fragments is retained to provide a more complete picture of the model's decision at a later point (Fig. 5).

A potential caveat in mapping the contribution values to a fragment is collisions. Collisions can happen at two stages of the method: (i) When hashing the SMILES representation of the substructures from a reaction to a set

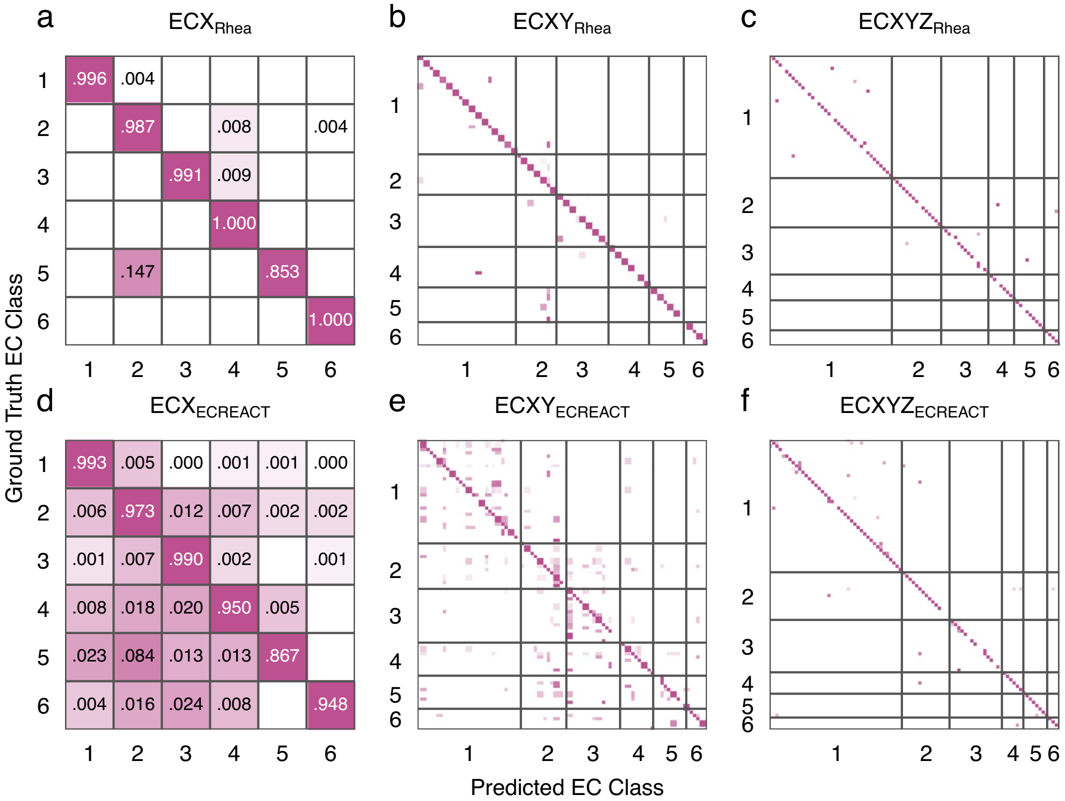


Fig. 3 Confusion matrices for reaction-based enzyme classification. **a**, **d** Enzyme class-level confusion matrices for models trained on the Rhea and ECREACT data sets, respectively. **b**, **e** Subclass-level confusion matrices for models trained on the Rhea and ECREACT data sets, respectively. **c**, **f** Sub-subclass-level confusion matrices for models trained on the Rhea and ECREACT data sets, respectively

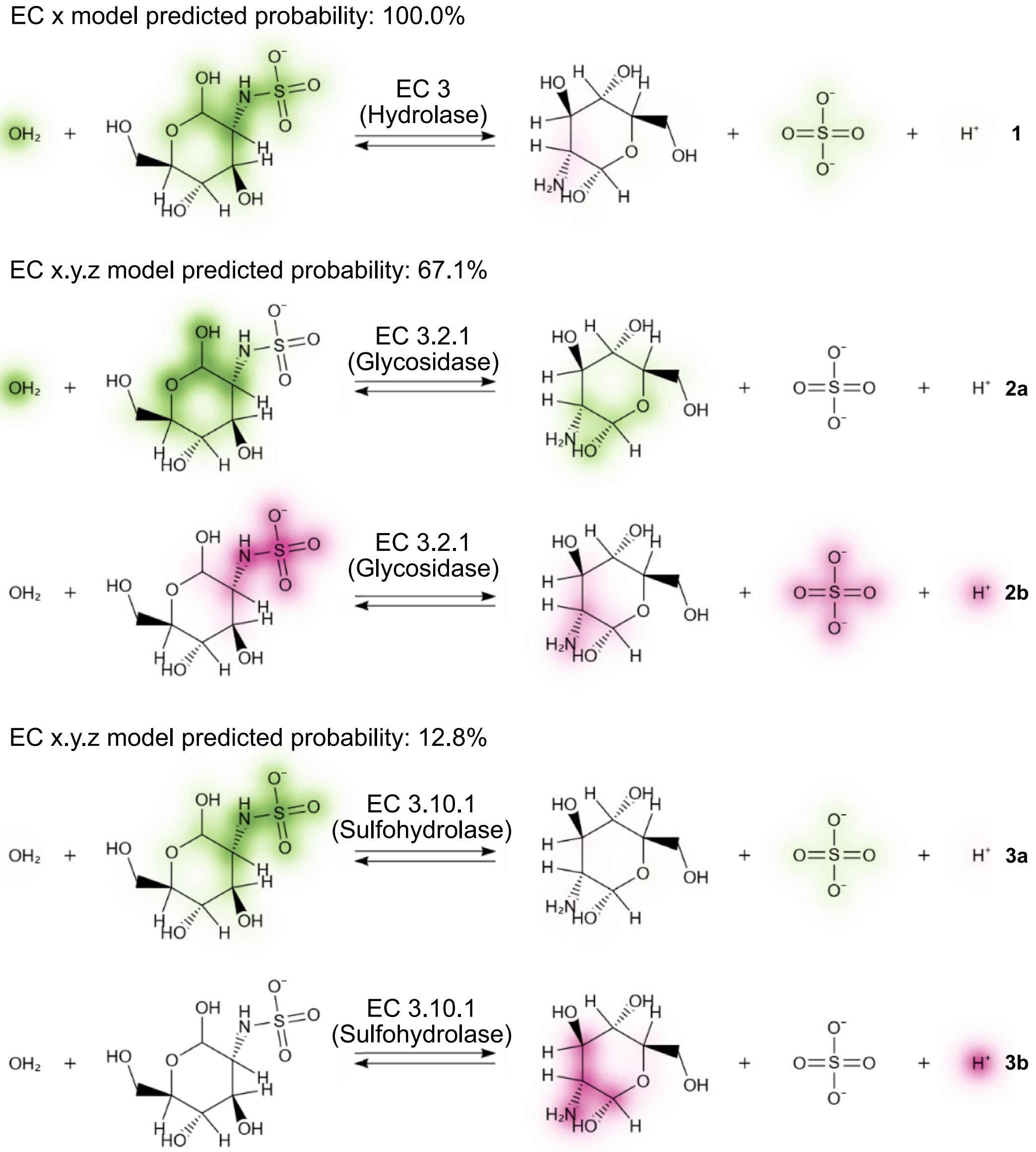
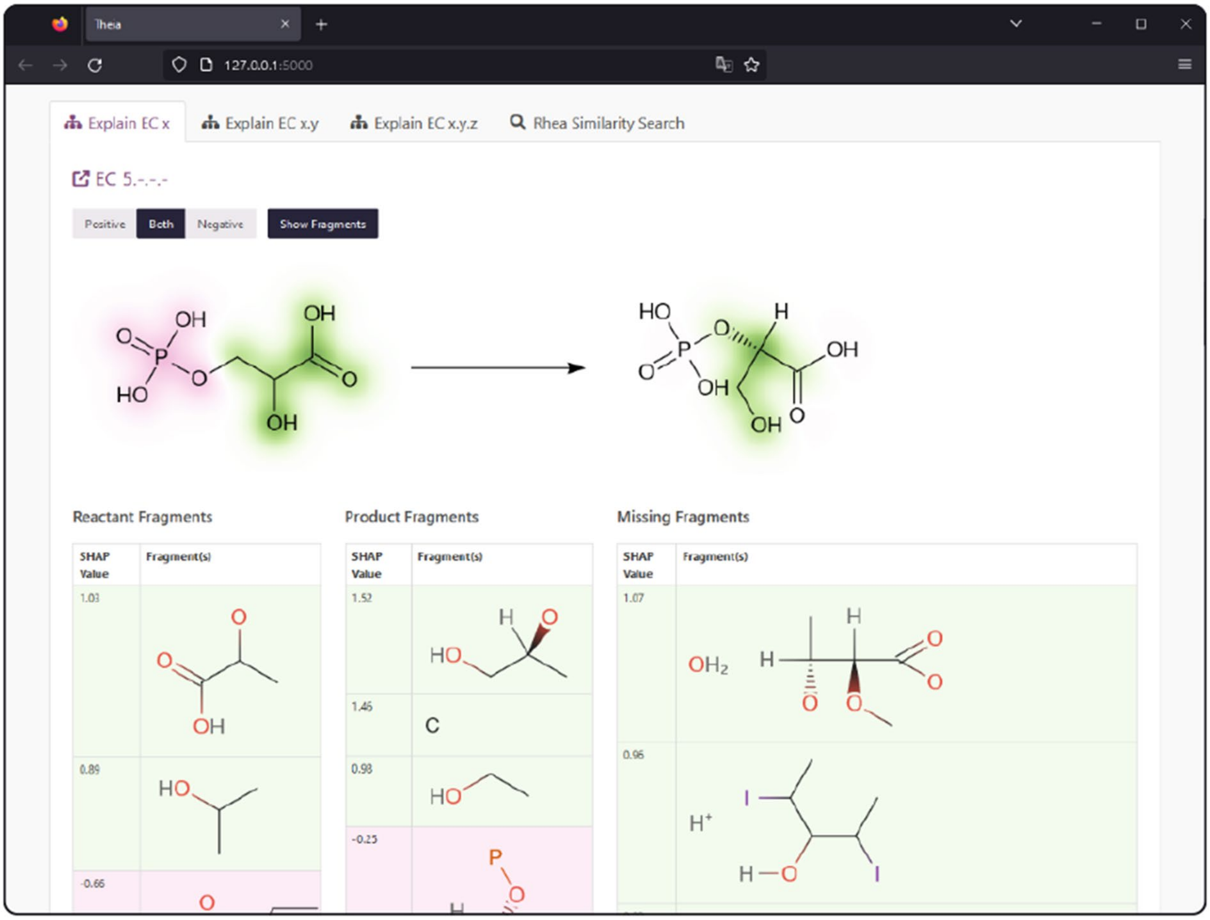

EC x model predicted probability: 100.0%
EC 3
(Hydrolase)
1
EC x.y.z model predicted probability: 67.1%
EC 3.2.1
(Glycosidase)
2a
EC 3.2.1
(Glycosidase)
2b
EC x.y.z model predicted probability: 12.8%
EC 3.10.1
(Sulfohydrolase)
3a
EC 3.10.1
(Sulfohydrolase)
3b

Fig. 4 Explaining a misclassification of a N-sulfoglucosamine sulfohydrolase (EC 3.10.1) as a glycosidase (EC 3.2.1). The model ECX_{Rhea} correctly predicts the reaction (**1**) to be catalysed by a hydrolase (EC 3), primarily focusing on the water (OH_2) and the hydrolysed bond, both with a positive contribution towards EC 3. In addition, there is a small negative contribution against EC3 shown on the amine group. Unlike **1**, where positive and negative contributions are shown in one reaction drawing, positive and negative contributions are split into separate depictions for **2** and **3** for visualization purposes. For the top prediction (67.1%) of model $ECXYZ_{Rhea}$ the focus of the model shifts to a non-reactive site including a hydroxy group in the N-sulfo-D-glucosamine as a major positive contribution (**2a**), while the sulfur-nitrogen bond is the major negative contribution (**2b**). For the correct prediction (top-2, 12.8%), the model remains focused on the hydrolised sulfur-nitrogen bond with a positive contribution (**3a**) as the negative contributions (**3b**) can be found on the D-glucosamine and the proton

of 32-bit integers and (ii) when folding the 32-bit integers into a fixed-size binary vector using a modulo operation. The number of collisions of a 32-bit hashing function can be estimated based on the maximum hash value and the number of unique fragments using a generalisation of the birthday problem [50]. For a maximum hash value of $2^{32} - 1$ and 9509 and 16,983 unique fragments extracted from Rhea and ECREACT, respectively, this results in 0.01 expected hash collisions for Rhea and 0.03 for ECREACT. However, when folding the sets of 32-bit integers into 10,240-dimensional binary vectors using a modulo operation, 2439 and 5060 of the entries represent more than one fragment for Rhea and ECREACT, respectively. While the models still perform well given these



Reactant Fragments

SHAP Value	Fragment(s)
1.03	
0.89	
-0.66	

Product Fragments

SHAP Value	Fragment(s)
1.52	
1.46	C
0.98	
-0.25	

Missing Fragments

SHAP Value	Fragment(s)
1.07	OH_2
0.96	H^+

Fig. 5 Visualising the contributions to the classification of a reaction to being catalysed by an isomerase (phosphoglycerate mutase, EC 5.4.2) using a web interface. Fragments that are not present in a participating molecule yet still contribute heavily towards the predicted class are displayed with their respective contribution. In addition, for entries in the binary input vector that represent multiple fragments, all the associated structures are represented

collisions, as shown by our previous work [44, 50], this could potentially negatively impact the interpretability of the model using DeepLIFT values, especially if multiple fragments of a given reaction are represented by the same entry in the binary vector. For the visualisation of fragments that are part of a given reaction, this is a minor concern, as two fragments occupying the same entry only occur 147 times (2.1 %) in Rhea ($n = 7,010$) and 2025 times (2.5 %) in ECREACT ($n = 81,205$). For fragments that are not part of a given reaction yet still contribute to the decision of the model, however, this is not the case. We solve this problem and introduce a generalisable approach to the visualisation of explainable machine learning for reactions in the following section.

Visualising explanations using reaction depictions

The visualisation of the classification contributions calculated using the DeepSHAP implementation of DeepLIFT as shown in Fig. 4 is enabled by implementing an upgrade to the previously released SmilesDrawer JavaScript library [51]. The script takes the reaction SMILES and the atom contributions as computed in Eq. 1 as an input. As a first step, the atom contribution values w_{a_j} are normalised across all molecules in the reaction to show the relative contributions. Next, the contributions w_{a_j} are assigned the coordinates of the respective atoms $v_j = \left[x_{a_j}\ y_{a_j}\right]$ in each molecules 2D drawing space. Finally, the colour values for the pixel grid are assigned by the summatory function G of all bivariate gaussian distributions centred at v_j with an arbitrary σ.

$$G(x,y) = \sum_{j}^{N_a} w_{a_j} e^{-\frac{(x-x_{a_j})^2-(y-y_{a_j})^2}{2\sigma^2}} \tag{2}$$

Choosing a diverging colour scale to represent the contribution values at each coordinate in the pixel grid produces an intuitive representation of the contribution values. The resulting visualisation can be seen in Fig. 4. However, this only enables the visualisation of the contributions of fragments that are present in a given reaction. In order to also visualise the contribution of fragments that are not present in a reaction, these fragments are listed with their respective value as part of a web service or a locally run program. Figure 5 shows an example of a phosphoglycerate mutase (EC 5) catalysed reaction as explained in the web application on which the introduced approach is being made available. The contributions towards being classified as an isomerase come as much from the absent fragments, a proton (H^+) and water (OH_2), as from the fragments found in the participating molecules. The example of the interface shown in Fig. 5 also introduces how the occurrence of multiple fragments assigned to the same entry of the binary input vector can be handled by displaying the colliding substructures. In this case, it is trivial to determine the influential missing fragments, water and proton, from the context of the given reaction and the predicted class (EC 5).

To facilitate access to the presented visualisations, the presented models are deployed as a web application and a Jupyter notebook, as well as graphical and command line interfaces installable via Python's pip package manager.

Conclusions

The approach presented in this work introduces a way forward in enzyme–reaction classification beyond expert curation. The introduced models and software will initially support the growth and balancing of databases containing annotated enzyme-catalysed reactions such as Rhea through human-in-the-loop machine learning. The utility of this approach is illustrated in Fig. 4, where the explanation of a misclassification of the model lends insight into the underlying causes of the inaccuracy such as a lack of training data of certain classes, in this case, Sulfohydrolase-catalysed reactions. Based on such information, an expert curator can modify either the architecture of the model or the composition of the training data set. The fast and efficient training of the classifiers (below 10 min using approximately 25 Wh on a consumer laptop) then allows for continuous retraining with an adjusted architecture or on newly annotated or balanced data. Eventually, as the classes with sufficient examples show, the presented solution will be able to take over reaction-based enzyme and enzyme function classification from humans. In their current iteration, the trained models can already be used to predict a candidate catalysing enzyme for an arbitrary chemical reaction, while the software allows for easy human evaluation of the predictions. In addition, the generalisable method to visualise explainable machine learning for chemical and biochemical reactions can be adapted by current and future machine learning tools involving arbitrary explainability techniques and machine learning architectures. Documentation, code, and notebooks showcasing the described functionality as well as instruction to install the tools locally from PyPI can be found at https://github.com/daenuprobst/theia.

Methods

EC Classification

Once identified and associated with a biochemical reaction through experimentation, enzymes are classified according to their function and the reactions they catalyse using the hierarchical EC Number (Enzyme Commission Number) scheme based on the reaction they catalyse. This hierarchical classifier, in the form x.y.z.sn, where x is the class, y is the subclass, z is the sub-subclass, and sn is an incremental serial number assigned to an enzyme. The class (x) encompasses seven categories: (1) Oxidoreductases, (2) Transferases, (3) Hydrolases, (4) Lyases, (5) Isomerases, (6) Ligases, and (7) Translocases. While classes 1 through 6 catalyse a chemical modification of the substrate, translocases are limited to catalyse the movement of molecules or ions across membranes. Translocases are, therefore, not within the scope of this study. The subclass (y) of an enzyme specifies the group or bond on which the enzyme acts. For example, hydrolases of subclasses 3.4 and 3.7 act on peptide and carbon-carbon bonds, respectively. The sub-subclass (z) of an enzyme further specifies the reaction. Peptidases (3.4) with the sub-subclass 3.4.13 are dipeptidases with dipeptides as a substrate, while peptidases with the subsubclass 3.4.22 are cysteine endopeptidases that hydrolyse peptide bonds after non-terminal cysteines. Finally, the serial number (sn) does not convey learnable information on the reaction type but distinguishes different enzymes that catalyse the same type of reaction on specific substrates.

Data processing

The ECREACT (Version 1.0) data set does not require any preprocessing as it is available as a .csv file containing reaction SMILES and the associated EC number. For Rhea (Release 123), the file containing the reactions annotated with ChEBI identifiers is downloaded and then processed with a ChEBI export to match the molecular identifiers with the respective SMILES to generate reaction SMILES. The SMILES in both data sets contain stereochemistry information. The processed data, as well as a shell script to download the required data and a Python script to process the raw Rhea data are included in the GitHub repository.

Differential reaction fingerprint (DRFP)

The successful representation of enzyme-catalysed reactions by the differential reaction fingerprint (DRFP) has already been shown by [52]. In contrast to this previous approach, the parameters of the DRFP PyPI package (drfp = 0.3.6) were selected to maximise accuracy while minimising the probability of collisions, in order to enhance explainability. The folded length (dimensionality) of the DRFP fingerprint was chosen as 10,240 (default 2,048) and the radius as 2 (default 3). The DRFP encoding function was adapted to produce non-centred canonicalised SMILES, further reducing the number of potential collisions. Whereas the original function would produce multiple SMILES rooted at each atom for each fragment (e.g. COC, OCC, and CCO for dimethyl ether), the adapted version produces only a single SMILES per fragment (e.g. COC for dimethyl ether). This change can be toggled in the updated DRFP package using the argument root_central_atom in the static function encode. Furthermore, the function was adapted to explicitly include hydrogens in the SMILES encoding of the fragments (e.g. [H]C([H])([H])OC([H])([H])[H] instead of COC). This change can also be toggled in the updated DRFP package using the argument include_hydrogens in the static function encode. DRFP preserves the stereochemistry information in the processed SMILES.

Multilayer perceptrons

The multilayer perceptrons used for all models in this work were implemented using PyTorch (version 1.13.0). The initial hyperparameters were taken from our previous publication on DRFP [44]. The MLP consists of a linear input layer with 10,240 nodes, a hidden layer with 1664 nodes, and a linear output layer with a number of nodes that is equal to the number of classes (unique EC numbers). Cross entropy with default parameters is chosen as the loss function (criterion), and Adam with a learning rate of 0.001 as optimiser. PyTorch's exponential learning rate scheduler with a gamma of 0.9 is set as the scheduler. Finally, early stopping is implemented by monitoring the mean validation loss of the 5 most recent epochs. The training is stopped if the improvement of the current loss drops below 0.001. Training and validation losses for all models are shown in Additional file 1: Fig. S1. Note that, as the models were trained and run inference independently from each other, their respective predictions may not follow the EC hierarchy. This can result in seemingly contradictory results where the prediction of the class (x) from the ECX model may differ from the class predicted by the ECXYZ model.

DeepLIFT explanations

DeepSHAP is an implementation of DeepLIFT, an additive feature attribution method, based on the assumption that DeepLIFT approximates SHAP values. A detailed description of the method can be found in the section *DeepSHAP (DeepLIFT + Shapley values)* of [49]. Using the described method, DeepSHAP assigns each feature (molecular fragment) an importance value for a given prediction based on a baseline value. The baseline value is calculated from a set of samples—100 reaction SMILES in the presented implementation—and represents an approximation of the average of all predictions. The SHAP (SHapley Additive exPlanation) values that measure the contributions of a feature based on the baseline are then the summed Shapley values of a conditional expectation function of the original model [49].

TMAP visualisation

The TMAP (Tree map) visualisations shown in Fig. 2 were generated using the PyPI package tmap-viz. The parameters sl_repeats = 2, mmm_repeats = 2, and n_trees = 50 remained constant for all three subfigures, while while node_size was set to 3, 2, and 10 for subfigure a, b, and c, respectively. The script to generate the TMAPs is available in the project's GitHub repository.

Reaction visualisation

The reaction visualisations are based on the Smiles-Drawer JavaScript library that, compared to other available libraries, allows the depictions of molecules and reactions in web applications without the need for server-side image rendering [51]. To enable the visualisation of numeric attributes on a per-atom level, the library was extended with the ability to draw arbitrary pixel values on a background layer. While the reactions can be rendered as rasterised images (HTML canvas, or image elements) or vector images (HTML SVG elements), the background layer is always rendered as a rasterised image and scaled without interpolation for performance reasons. Finally, a wrapper for the display of explainable reactions with the SmilesDrawer JavaScript library in Jupyter notebooks is available on PyPI in the package faerun-notebook.

Web application

In order to make the models and visualisations easily accessible, the presented approach is deployed as a Flask-based web application. In addition to the EC number predictions and the visualisation of the molecular contributions, the application also performs a nearest neighbour search of the DRFP-encoded reaction on the Rhea data using an Annoy index [53]. In addition to the hosted version, the application is available as Docker and PyPI packages for on-premise deployment or local use.

Scientific contribution statement

The speed of recent advances in machine learning on (bio)chemical reactions has been unprecedented; however, most new methods lack explainability and rely on large data sets. The methodology presented in this work not only enables the prediction of catalysing enzymes from reactions but also, for the first time, provides explanations that are directly visualised on the reaction depiction, allowing for researchers and data curators to evaluate the results and potential biases in the underlying data set, respectively.

Appendix

See Table 2.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-023-00784-y.

Additional file 1: Fig. S1. Training and validation losses for the modelspresented in this work. Early stopping is implemented bymonitoring the mean validation loss of the 5 most recentepochs. The training is stopped if the improvement of thecurrent loss drops below 0.001. **Fig. S2.** Boxplots showing the distribution of accuraciesamong classes, subclasses, and sub-subclasses. The filledorange circle represents the mean. **Fig. S3.** Scatter plots showing the dependence of accuracyon sample size for models trained on Rhea data. **Fig. S4.** Scatter plots showing the dependence of accuracyon sample size for models trained on ECREACT data.

Acknowledgements

I would like to thank Prof. Pierre Vandergheynst for hosting me in his research group as well as the Swiss-Prot group at the Swiss Institute of Bioinformatics for their fruitful feedback and discussions.

Author contributions

Conceptualization—DP, data curation—DP; formal analysis—DP; investigation—DP, methodology—DP; writing, original draft—DP, writing, review and editing—DP.

Funding

DP thanks the Pierre Vandergheynst and the EPFL for funding.

Availability of data and materials

All used data sets are available as .csv files on GitHub. The training, validation, and test sets can be recreated using Python scripts available in the same GitHub repository https://github.com/daenuprobst/theia. Furthermore, the original data used has been deposited on Zenodo under Version 1.0 (10.5281/zenodo.8318231, https://zenodo.org/record/8318231). All implementations and used software are available under open-source licences. The code for the web application as well as the scripts necessary to reproduce this work can be found at https://github.com/daenuprobst/theia as release V1.0.0. The repositories of the adapted DRFP and SmilesDrawer code are https://github.com/reymond-group/drfp and https://github.com/reymond-group/smilesDrawer, respectively.

Declarations

Ethics approval and consent to participate

This study did not require ethics approval nor participation consents.

Consent for publication

This study does not require consents for publication.

Competing interests

The author declares no competing interests.

Table 2 Ablation study on models trained on Rhea. For each run, a fraction (0.01 to 0.5) of the labels has been shuffled in order to simulate real world conditions of non-curated data containing misclassifications

Shuffled fraction/Model	ECX_{Rhea}	$ECXY_{Rhea}$	$ECXYZ_{Rhea}$
–	0.98 ± 0.00 (0.97 ± 0.01)	0.96 ± 0.01 (0.88 ± 0.02)	0.95 ± 0.00 (0.87 ± 0.02)
0.01	0.98 ± 0.01 (0.96 ± 0.01)	0.95 ± 0.00 (0.86 ± 0.00)	0.94 ± 0.01 (0.86 ± 0.02)
0.05	0.96 ± 0.01 (0.94 ± 0.01)	0.91 ± 0.00 (0.83 ± 0.02)	0.91 ± 0.01 (0.81 ± 0.01)
0.10	0.93 ± 0.00 (0.91 ± 0.01)	0.88 ± 0.00 (0.77 ± 0.02)	0.88 ± 0.01 (0.75 ± 0.01)
0.20	0.88 ± 0.01 (0.86 ± 0.01)	0.88 ± 0.01 (0.86 ± 0.01)	0.82 ± 0.01 (0.68 ± 0.02)
0.50	0.74 ± 0.02 (0.66 ± 0.01)	0.64 ± 0.01 (0.49 ± 0.03)	0.60 ± 0.03 (0.44 ± 0.01)

In addition, the statistics for the model trained on the non-shuffled data is shown (–). The metrics reported are the accuracies and, shown in parentheses, the F-Scores. Runtimes and energy usage is identical to those reported in the main text

Received: 22 May 2023 Accepted: 13 November 2023
Published online: 23 November 2023

References

1. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45(D1):353–361. https://doi.org/10.1093/nar/gkw1092
2. Lee D-S, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabási A-L (2008) The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci 105(29):9880–9885. https://doi.org/10.1073/pnas.0802208105
3. Lu H, Li F, Sánchez BJ, Zhu Z, Li G, Domenzain I, Marcišauskas S, Anton PM, Lappa D, Lieven C, Beber ME, Sonnenschein N, Kerkhoven EJ, Nielsen J (2019) A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat Commun 10(1):3586. https://doi.org/10.1038/s41467-019-11581-3
4. Kumar A, Wang L, Ng CY, Maranas CD (2018) Pathway design using de novo steps through uncharted biochemical spaces. Nat Commun 9(1):184. https://doi.org/10.1038/s41467-017-02362-x
5. Harrigan JA, Jacq X, Martin NM, Jackson SP (2018) Deubiquitylating enzymes and drug discovery: emerging opportunities. Nat Rev Drug Discov 17(1):57–78. https://doi.org/10.1038/nrd.2017.152
6. Kazmi SR, Jun R, Yu M-S, Jung C, Na D (2019) In silico approaches and tools for the prediction of drug metabolism and fate: A review. Comput Biol Med 106:54–64. https://doi.org/10.1016/j.compbiomed.2019.01.008
7. Slagman S, Fessner W-D (2020) Biocatalytic routes to anti-viral agents and their synthetic intermediates. Chem Soc Rev 50(3):1968–2009. https://doi.org/10.1039/d0cs00763c
8. Sheldon RA, Woodley JM (2018) Role of biocatalysis in sustainable chemistry. Chem Rev 118(2):801–838. https://doi.org/10.1021/acs.chemrev.7b00203
9. Wu S, Snajdrova R, Moore JC, Baldenius K, Bornscheuer UT (2021) Biocatalysis: enzymatic synthesis for industrial applications. Angew Chem Int Ed 60(1):88–119. https://doi.org/10.1002/anie.202006648
10. Delépine B, Duigou T, Carbonell P, Faulon J-L (2018) RetroPath2.0: a retrosynthesis workflow for metabolic engineers. Metab Eng 45:158–170. https://doi.org/10.1016/j.ymben.2017.12.002
11. Peyhani HM, Hafner J, Sveshnikova A, Viterbo V, Hatzimanikatis V (2022) Expanding biochemical knowledge and illuminating metabolic dark matter with ATLASx. Nat Commun 13(1):1560. https://doi.org/10.1038/s41467-022-29238-z
12. Probst D, Manica M, Teukam YGN, Castrogiovanni A, Paratore F, Laino T (2022) Biocatalysed synthesis planning using data-driven learning. Nat Commun 13(1):964. https://doi.org/10.1038/s41467-022-28536-w
13. Kreutter D, Schwaller P, Reymond J-L (2021) Predicting enzymatic reactions with a molecular transformer. Chem Sci 12(25):8648–8659. https://doi.org/10.1039/d1sc02362d
14. Karp PD, Weaver D, Latendresse M (2018) How accurate is automated gap filling of metabolic models? BMC Syst Biol 12(1):73. https://doi.org/10.1186/s12918-018-0593-7
15. Lowe D (2017) Chemical reactions from US patents (1976–Sep2016). *figshare* https://doi.org/10.6084/M9.FIGSHARE.5104873.V1. https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873/1. Accessed 16 Dec 2022
16. ...Bateman A, Martin M-J, Orchard S, Magrane M, Alpi E, Bely B, Bingley M, Britto R, Bursteinas B, Busiello G, Bye-A-Jee H, Silva AD, Giorgi MD, Dogan T, Castro LG, Garmiri P, Georghiou G, Gonzales D, Gonzales L, Hatton-Ellis E, Ignatchenko A, Ishtiaq R, Jokinen P, Joshi V, Jyothi D, Lopez R, Luo J, Lussi Y, MacDougall A, Madeira F, Mahmoudy M, Menchi M, Nightingale A, Onwubiko J, Palka B, Pichler K, Pundir S, Qi G, Raj S, Renaux A, Lopez MR, Saidi R, Sawford T, Shypitsyna A, Speretta E, Turner E, Tyagi N, Vasudev P, Volynkin V, Wardell T, Warner K, Watkins X, Zaru R, Zellner H, Bridge A, Xenarios I, Poux S, Redaschi N, Aimo L, Argoud-Puy G, Auchincloss A, Axelsen K, Bansal P, Baratin D, Blatter M-C, Bolleman J, Boutet E, Breuza L, Casals-Casas C, de Castro E, Coudert E, Cuche B, Doche M, Dornevil D, Estreicher A, Famiglietti L, Feuermann M, Gasteiger E, Gehant S, Gerritsen V, Gos A, Gruaz N, Hinz U, Hulo C, Hyka-Nouspikel N, Jungo F, Keller G, Kerhornou A, Lara V, Lemercier P, Lieberherr D, Lombardot T, Martin X, Masson P, Morgat A, Neto TB, Paesano S, Pedruzzi I, Pilbout S, Pozzato M, Pruess M, Rivoire C, Sigrist C, Sonesson K, Stutz A, Sundaram S, Tognolli M, Verbregue L, Wu CH, Arighi CN, Arminski L, Chen C, Chen Y, Cowart J, Garavelli JS, Huang H, Laiho K, McGarvey P, Natale DA, Ross K, Vinayaka CR, Wang Q, Wang Y, Yeh L-S, Zhang J (2018) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47(Database):1049. https://doi.org/10.1093/nar/gky1049
17. Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L, Hyka-Nouspikel N, Gasteiger E, Kerhornou A, Neto TB, Pozzato M, Blatter M-C, Ignatchenko A, Redaschi N, Bridge A (2021) Rhea, the reaction knowledgebase in 2022. Nucleic Acids Res 50(D1):693–700. https://doi.org/10.1093/nar/gkab1016
18. McDonald AG, Boyce S, Tipton KF (2009) ExplorEnz: the primary source of the IUBMB enzyme list. Nucleic Acids Res 37(Suppl–1):593–597. https://doi.org/10.1093/nar/gkn582
19. Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res 28(1):304–305. https://doi.org/10.1093/nar/28.1.304
20. Meuwly M (2021) Mach Learn Chem React. Chemical Rev 121(16):10218–10239. https://doi.org/10.1021/acs.chemrev.1c00033
21. Schwaller P, Vaucher AC, Laplaza R, Bunne C, Krause A, Corminboeuf C, Laino T (2022) Machine intelligence for chemical reaction space. Wiley Interdiscip Rev Comput Mol Sci. https://doi.org/10.1002/wcms.1604
22. Zou Z, Tian S, Gao X, Li Y (2019) mlDEEPre: Multi-Functional Enzyme Function Prediction With Hierarchical Multi-Label Deep Learning. Front Genet 9:714. https://doi.org/10.3389/fgene.2018.00714
23. Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H, Xavier RJ, Knight R, Cho K, Bonneau R (2021) Structure-based protein function prediction using graph convolutional networks. Nat Commun 12(1):3168. https://doi.org/10.1038/s41467-021-23303-9
24. Dalkiran A, Rifaioglu AS, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T (2018) ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. BMC Bioinf 19(1):334. https://doi.org/10.1186/s12859-018-2368-y
25. Ryu JY, Kim HU, Lee SY (2019) Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci 116(28):13996–14001. https://doi.org/10.1073/pnas.1821905116
26. Kotera M, Okuno Y, Hattori M, Goto S, Kanehisa M (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. J Am Chem Soc 126(50):16487–16498. https://doi.org/10.1021/ja0466457
27. Yamanishi Y, Hattori M, Kotera M, Goto S, Kanehisa M (2009) E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. Bioinformatics 25(12):179–186. https://doi.org/10.1093/bioinformatics/btp223
28. Rahman SA, Cuesta SM, Furnham N, Holliday GL, Thornton JM (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. Nat Methods 11(2):171–174. https://doi.org/10.1038/nmeth.2803
29. Latino DARS, Aires-de-Sousa J (2009) Assignment of EC numbers to enzymatic reactions with MOLMAP reaction descriptors and random forests. J Chem Inf Model 49(7):1839–1846. https://doi.org/10.1021/ci900104b
30. Egelhofer V, Schomburg I, Schomburg D (2010) Automatic assignment of EC numbers. PLoS Comput Biol 6(1):1000661. https://doi.org/10.1371/journal.pcbi.1000661
31. Hu Q-N, Zhu H, Li X, Zhang M, Deng Z, Yang X, Deng Z (2012) Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints. PLoS ONE 7(12):52901. https://doi.org/10.1371/journal.pone.0052901
32. Carbonell P, Wong J, Swainston N, Takano E, Turner NJ, Scrutton NS, Kell DB, Breitling R, Faulon J-L (2018) Selenzyme: enzyme selection tool for pathway design. Bioinformatics 34(12):2153–2154. https://doi.org/10.1093/bioinformatics/bty065
33. Matsuta Y, Ito M, Tohsato Y (2013) ECOH: an Enzyme Commission number predictor using mutual information and a support vector machine. Bioinformatics 29(3):365–372. https://doi.org/10.1093/bioinformatics/bts700
34. Hadadi N, MohammadiPeyhani H, Miskovic L, Seijo M, Hatzimanikatis V (2019) Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. Proc Natl Acad Sci 116(15):7298–7307. https://doi.org/10.1073/pnas.1818877116

35. Borrego-Díaz J, Galán-Páez J (2022) Explainable artificial intelligence in data science. Minds Mach 32(3):485–531. https://doi.org/10.1007/s11023-022-09603-z
36. Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007
37. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S (2022) Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet. https://doi.org/10.1038/s41576-022-00532-2
38. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR (2022) Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022). Comput Methods Progr Biomed 226:107161. https://doi.org/10.1016/j.cmpb.2022.107161
39. Khosravi H, Shum SB, Chen G, Conati C, Tsai Y-S, Kay J, Knight S, Martinez-Maldonado R, Sadiq S, Gašević D (2022) Explainable artificial intelligence in education. Comput Educ Artifl Intell 3:100074. https://doi.org/10.1016/j.caeai.2022.100074
40. Mastropietro A, Pasculli G, Feldmann C, Rodríguez-Pérez R, Bajorath J (2022) EdgeSHAPer: bond-centric Shapley value-based explanation method for graph neural networks. iScience 25(10):105043. https://doi.org/10.1016/j.isci.2022.105043
41. Heberle H, Zhao L, Schmidt S, Wolf T, Heinrich J (2023) XSMILES: interactive visualization for molecules, SMILES and XAI attribution scores. J Cheminf 15(1):2. https://doi.org/10.1186/s13321-022-00673-w
42. Wellawatte GP, Seshadri A, White AD (2022) Model agnostic generation of counterfactual explanations for molecules. Chem Sci 13(13):3697–3705. https://doi.org/10.1039/d1sc05259d
43. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Model 28(1):31–36. https://doi.org/10.1021/ci00057a005
44. Probst D, Schwaller P, Reymond J-L (2022) Reaction classification and yield prediction using the differential reaction fingerprint DRFP. Digital Discov 1(2):91–97. https://doi.org/10.1039/d1dd00006c
45. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D, Schomburg D (2020) BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res 49(D1):498–508. https://doi.org/10.1093/nar/gkaa1025
46. Wishart DS, Li C, Marcu A, Badran H, Pon A, Budinski Z, Patron J, Lipton D, Cao X, Oler E, Li K, Paccoud M, Hong C, Guo AC, Chan C, Wei W, Ramirez-Gaona M (2019) PathBank: a comprehensive pathway database for model organisms. Nucleic Acids Res 48(D1):470–478. https://doi.org/10.1093/nar/gkz861
47. Moretti S, Tran V, Mehl F, Ibberson M, Pagni M (2020) MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models. Nucleic Acids Res 49(D1):992. https://doi.org/10.1093/nar/gkaa992
48. Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. arXiv. https://doi.org/10.48550/arxiv.1704.02685
49. Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. arXiv. https://doi.org/10.48550/arxiv.1705.07874
50. Probst D, Reymond J-L (2018) A probabilistic molecular fingerprint for big data settings. J Cheminf 10(1):66. https://doi.org/10.1186/s13321-018-0321-8
51. Probst D, Reymond J-L (2018) SmilesDrawer: parsing and drawing SMILES-encoded molecular structures using client-side Javascript. J Chem Inf Model 58(1):1–7. https://doi.org/10.1021/acs.jcim.7b00425
52. Hoyt CT. Rhea differential reaction fingerprints for enzyme classification prediction. https://doi.org/10.5281/zenodo.7591839
53. Bernhardsson E (2017) Annoy: approximate nearest neighbors in c++/python optimized for memory usage and loading/saving to disk. GitHub. https://github.com/spotify/annoy. Accessed 6 Sept 2022

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.