Data-Driven Protein Engineering for Improving Catalytic Activity and Selectivity

Yu-fei Ao, Mark Dörr, Marian J Menke, Stefan Born, Egon Heuson, Uwe T Bornscheuer

ChemBioChem

Concept
doi.org/10.1002/cbic.202300754

Chemistry Europe
European Chemical Societies Publishing

www.chembiochem.org

VIP *Very Important Paper*

Data-Driven Protein Engineering for Improving Catalytic Activity and Selectivity

Yu-Fei Ao,*[a, b, e] Mark Dörr,[a] Marian J. Menke,[a] Stefan Born,[c] Egon Heuson,[d] and Uwe T. Bornscheuer*[a]

Protein engineering is essential for altering the substrate scope, catalytic activity and selectivity of enzymes for applications in biocatalysis. However, traditional approaches, such as directed evolution and rational design, encounter the challenge in dealing with the experimental screening process of a large protein mutation space. Machine learning methods allow the approximation of protein fitness landscapes and the identification of catalytic patterns using limited experimental data, thus providing a new avenue to guide protein engineering campaigns. In this concept article, we review machine learning models that have been developed to assess enzyme-substrate-catalysis performance relationships aiming to improve enzymes through data-driven protein engineering. Furthermore, we prospect the future development of this field to provide additional strategies and tools for achieving desired activities and selectivities.

Introduction

Protein engineering is essential to improve the catalytic function of enzymes including substrate scope, catalytic activity and selectivity to produce pharmaceuticals, flavors and fragrances, other fine and specialty chemicals as well as biofuels to establish green chemistry processes.[1] Currently, two classical approaches for protein engineering, directed evolution and rational design, are mainly employed to design enzyme variants with increased activity, broader substrate scope and improved selectivity for specific substrates.[2] However, both approaches still encounter significant obstacles due to the expensive, time-consuming and difficult to handle experimental screening process of the huge protein sequence space. As a result, only a small fraction of this can be practically explored experimentally, even using current high-throughput screening and computational techniques.

To overcome these limitations, machine learning (ML) has been employed to assist protein engineering, which led to many successful protein variants with enhanced catalytic activity and/or selectivity.[3] This data-driven strategy enables the approximation of protein fitness landscapes from sparsely sampled experimental data, reducing the portion of the sequence space that is potentially valuable to explore. Moreover, it can identify catalytic patterns, especially nonlinear epistatic effects in the collected data, thereby predicting previously unnoticed but promising variants like novel combinations of substitutions. As a result, machine learning is believed to have the potential to considerably reduce the computational and experimental effort required by traditional methods.

The typical process of machine learning involves several components, including data collection and pre-processing, data curation, feature extraction and selection, model training, validations and iterations.[3] The success of a ML predictor crucially depends on the data selection representation and feature extraction. To date, the available biocatalysis-related open-source databases, such as UniProt[4] and BRENDA,[5] contain a large amount of protein sequence/structure information as well as annotated information about enzyme functions such as the corresponding enzyme commission (EC) numbers and reaction kinetics parameters. Machine learning algorithms trained by these data have been used to characterize the correlation between sequence/structure and function,[6,30] EC numbers,[7] protein-ligand binding affinity[8] or k_{cat} values.[9] However, these databases lack specific information about reactivity, selectivity and details for a particular substrate. This means that a predictor may not be able to establish a relationship between reaction performance and enzyme/substrate structure. As a result, it may not be able to effectively guide protein engineering.

[a] Dr. Y.-F. Ao, Dr. M. Dörr, M.Sc. M. J. Menke, Prof. Dr. U. T. Bornscheuer
Department of Biotechnology and Enzyme Catalysis
Institute of Biochemistry, University of Greifswald
Felix-Hausdorff-Str. 4, 17487 Greifswald (Germany)
E-mail: aoyufe@iccas.ac.cn
uwe.bornscheuer@uni-greifswald.de

[b] Dr. Y.-F. Ao
Beijing National Laboratory for Molecular Sciences
CAS Key Laboratory of Molecular Recognition and Function
Institute of Chemistry, Chinese Academy of Sciences
Zhongguancun North First Street 2, Beijing, 100190 (China)

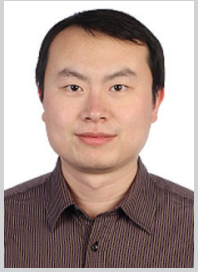
[c] Dr. S. Born
Technische Universität Berlin, Chair of Bioprocess Engineering
Ackerstraße 76, 13355 Berlin (Germany)

[d] Dr. E. Heuson
Univ. Lille, CNRS, Centrale Lille, Univ. Artois, UMR 8181
UCCS, Unité de Catalyse et Chimie du Solide
59000 Lille (France)

[e] Dr. Y.-F. Ao
University of Chinese Academy of Sciences
Yuquan Road 19(A), Beijing, 100049 (China)

In addition, when it comes to the application of enzymes in biocatalytic synthesis, the information contained in these databases is generally non-standardized, often resulting from experiments carried out under very different experimental conditions, with a significant proportion of the metadata missing. It is therefore very difficult to assess their individual veracity, and to remove those that appear to be exceptions, or even invalid, from the training sets. In order to provide access to more homogeneous, and above all, more reproducible and verifiable data sets, several initiatives are ongoing, notably around standardizing the measurement of enzyme kinetic parameters using common conditions,[10] and building more complete databases, requiring a minimal set of parameters for the addition of new entries, such as STRENDA DB.[11] In fact, the main difference between this new database and the more general databases mentioned above is that the latter is oriented towards the use of enzymes in synthesis, rather than towards general enzymology or even the global study of proteins. This difference is crucial when it comes to use them to predict new mutations, as the data they contain *de facto* incorporate the catalytic dimension, with the associated metadata, which remain optional in more general databases. It then appears that the development of new prediction algorithms will be concomitant with the development of this new type of specialized database. It should be noted that database problems are just a few of the many factors that still limit the possibility of predicting enzyme activity accurately from its protein sequence.

As this field is in full expansion, with numerous advances over the last few years, we therefore review in this article the strategies and applications of data-driven protein engineering



Yu-Fei Ao obtained his doctoral degree (supervised by Mei-Xiang Wang) in 2014 at the Institute of Chemistry, Chinese Academy of Sciences (ICCAS) and then joined the group of De-Xian Wang at ICCAS. In 2021, he went to Germany and worked in the group of Prof. U. T. Bornscheuer at the University of Greifswald as a visiting scientist, In 2023, he returned to ICCAS and continued his research on biocatalysis as an associate professor.

Marian J. Menke studied biochemistry at the University of Greifswald and obtained his Master of Science degree in 2020. In 2021 he started his Ph.D. under the supervision of Prof. U.T. Bornscheuer. His current research topic focuses on the work on methyltransferases and transaminases in the context of machine learning.

Mark Dörr studied chemistry and biology at the Universities of Mainz and Jena. He received his PhD at the Institute of Inorganic Chemistry at the University in Jena, Germany (2004). He then joined research projects in the groups of Peter Nielsen in Copenhagen, Denmark (2006–2009) and Steen Rasmussen in Odense, Denmark (2009–2012) working on fundamental bio-/inorganic-chemical questions as a postdoctoral research associate. Since 2012 he has been a staff scientist in the Bornscheuer group working on high-throughput robotic screening, lab automation, molecular modelling and the combination of machine learning with protein engineering.

Stefan Born obtained his PhD in mathematics at Gießen University in 2010 and is working as a lecturer and researcher at TU Berlin. He has collaborated with ISMLL Hildesheim, PIK Potsdam and the Bornscheuer group at the University of Greifswald on a range of modelling and machine learning problems. From 2019 to 2022 he was head of the machine learning section of KIWI-biolab, a joined research project with the University of Greifswald, TU Berlin and the University of Hildesheim on the development of novel machine learning models for bioprocess development and protein design.

Egon Heuson is a researcher at the Unité de Catalyse et Chimie du Solide (UCCS) in Centrale Lille (Lille, France). He obtained his PhD in 2015 in Chemical Biology – Biocatalysis at the Institute de Chimie de Clermont-Ferrand. His PhD research topic was the discovery of new transaminases for the chiral synthesis of valuable amines building blocks. He is now developing new hybrid multicatalytic materials (i.e., catalysts that combine chemo- and biocatalysts on the same material in a "one-pot/one-step" process) for the efficient synthesis of renewable building blocks, and especially bio-based polymers.

Uwe T. Bornscheuer studied chemistry and received his Ph.D. in 1993 at Hannover University followed by postdoctoral research at Nagoya University (Japan). In 1998, he completed his Habilitation at the University of Stuttgart. He has been Professor at the Institute of Biochemistry at the University of Greifswald since 1999. Among other honors he was awarded the Enzyme Engineering Award in 2022. His current research interests are in the discovery and protein engineering of biocatalysts including computational and high-throughput screening tools. Applications of these enzymes are in organic synthesis, for flavors and fragrances, in lipid modification, and the degradation of plastics or complex marine polysaccharides.

with a specific focus on enhancing catalytic activity and selectivity as well as a list of the main problems still to be solved. Generally speaking, current ML models rely on the supervised regression ML models and quantitatively map catalytic performance levels to protein sequence/structure, which enables to predict catalytic performance values for unseen input mutation sequences and efficient design of protein mutations to improve the targeted catalytic properties.

Altering activity and selectivity

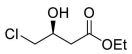
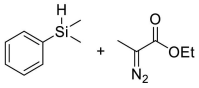
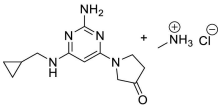
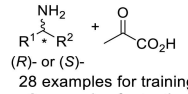
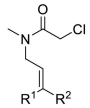
Due to the lack of diversity, sufficient and well-prepared samples in the existing datasets, efficient access to high-quality data has become a major challenge in training models, which limits the application of ML to protein engineering. Commonly, the data provided as a training set for ML needs to be obtained from literature or experiments, and thus the amount of standardized and comparable data is very limited. Therefore, simple ML algorithms (support vector machines or random forest algorithms, etc.) are commonly used for such small training datasets and typical examples are summarized in Table 1.

In 2007, an early outstanding publication by the Codexis team[12] described a statistical approach called ProSAR (protein sequence activity relationships). Through the construction of a multiple linear regression algorithm and the capture of additional information contained in the sequence-activity data, this approach evaluates the contribution of amino acid substitutions for the improvement of catalytic activity in each round of directed evolution. Positive substitutions were iteratively added and continuously increased catalytic activity. This approach was used to engineer a halohydrin dehalogenase for the synthesis of an intermediate of the drug Lipitor, with 35 amino acid substitutions leading to a 4,000-fold increased volumetric productivity without impairing the stereoselectivity of the enzyme (Table 1).

An alternative concept for protein engineering has been the design of 'small but smart' mutant libraries. This can be guided by the commercial software platform 3DM, which in brief identifies the most likely set of amino acid substitutions at a given position in the protein of interest guided by a sequence-structured based alignment. This can also identify correlated mutation networks (CorNet).[13] This concept was first demonstrated to be useful for the design of a more thermostable[14] or a more stereoselective[15] esterase. Another concept is the computational tool FRESCO developed in the Janssen group, which was demonstrated to be useful to increase the thermostability of a limonene epoxide hydrolase from 50 to 85 °C.[16]

In 2018, Cadet and Reetz[17] used a sequence-activity relationship (innov'SAR) methodology, which can be applied to build ML predictive models using protein sequence information and the fitness of variants measured in the wet-lab. The model permits to find the resulting property of the protein when *n* single point mutations are permuted (2^n combinations). Unlike the above-mentioned ProSAR approach, which rather assumes that mutations are additive in nature, innov'SAR is more

Table 1. Summary of data-driven protein engineering for improving activity and/or selectivity.

Enzyme	Substrate	Datasets for training	Algorithm	Catalytic performance improvement	Reference number
Halohydrin dehalogenase		~60,000 datapoints	Multiple linear regression (ProSAR approach)	Activity: 4000-fold improved volumetric productivity	12
Epoxide hydrolase		37 measured datapoints	Partial least square regression (innov'SAR approach)	Stereoselectivity: E-value up to 253	17
Nitric oxide dioxygenase		445 measured datapoints	Linear, kernel, neural network, ensemble methods	Stereoselectivity: up to 93%ee	18
Imine reductase		~8,000 measured datapoints	Random forest	Activity: conversion up to 72%; Stereoselectivity: up to > 99%ee	19
PETase	amorphous PET film	Over 19,000 sequence-balanced protein structures from the PDB	Three-dimensional convolutional neural network (MutCompute approach)	Activity: improved up to 38-fold at 50 °C	22
Amine transaminase	 (*R*)- or (*S*)- 28 examples for training 2 examples for testing	1,948 measured datapoints	Gradient boosting regression tree	Improved activity and stereoselectivity toward new substrates	25
Ene-reductase	 4 examples for training 2 examples for testing	50 measured datapoints	Multivariate linear regression	Improved enantioselectivity toward new substrates	27

suitable for reflecting the effect of epistatic interactions between protein mutations. Finally, this method was used to predict the enantioselectivity of 512 variants ($n=9$) of an epoxide hydrolase from *Aspergillus niger* using data from 37 characterized variants (9 single variants and 28 multiple point mutation variants). Five of the predicted variants were experimentally verified and the best one showed an improved E-value (E = 253), nearly 2.2-fold higher compared to the best existing variant (E = 118).

In 2019, the Arnold group[18] applied machine learning concepts to the directed evolution workflow to explore the full combinatorial space of mutations at selected positions. In this approach, data from a random sample of the combinatorial library were used to train machine learning models used to predict a smaller set of variants, which can be encoded with degenerate codons for experimental evaluations. They applied this to a putative nitric oxide dioxygenase (NOD) from *Rhodothermus marinus* (*Rma*) to produce each of the two product enantiomers via a new-to-nature carbene Si–H insertion reaction. Starting from a parent variant with low enantiopreference (76% ee, (*S*)-enantiomer), the approach predicted libraries enriched in functional enzymes and fixed seven mutations using 805 variants (445 for model training and 360 for model testing) to identify variants enabling stereoselective catalysis with 93% ee ((*S*)-enantiomer) and 79% ee ((*R*)-enantiomer).

A team at Novartis[19] assessed ML-assisted directed evolution to improve the stereoselectivity of an imine reductase. The study discovered that this approach produced a library of highly active and stereoselective variants with a significantly shifted activity distribution compared to variants obtained by deep mutational scanning or error-prone PCR. Utilising data from approximately 8,000 unique variants, two random forest models were trained using UniRep descriptors to score the *in silico* generated variants for either activity or stereoselectivity. The top 89 variants were experimentally verified revealing a significantly improved cumulative activity distribution and stereoselectivities with very good (*R*)- or (*S*)-selective enzymes. The highest ee value obtained was 81% ee, much higher than the 30%ee achieved with the starting scaffold. In addition to this study, two other teams have also demonstrated the application of machine learning-assisted protein engineering to improve an *in vivo* fatty alcohol production by engineering an acyl-ACP reductase[20] and for the asymmetric late-stage functionalization of soraphens by engineering an aliphatic halogenase WelO5*.[21]

In 2022, the Alper group[22] used a three-dimensional self-supervised convolutional neural network, MutCompute,[23] to identify stabilizing mutations in an esterase for the hydrolysis of the plastic polymer (poly)ethylene terephthalate (PET). This algorithm learns the local chemical microenvironments of amino acids by training based on 19,000 protein sequence-structure data obtained from the Protein Data Bank and generates a discrete probability distribution for the fit of all 20 amino acids at each position in the wild-type (WT) PETase, which was accomplished by conducting an extensive *in silico* mutagenesis scan. The predicted distributions were examined versus the protein crystal structure to identify positions where WT amino acid residues fitted less well than potential substitutions. After ranking the predictions according to their predicted probabilities and further characterizations, the generated 159 mutations showed improved catalytic activity and thermostability in the hydrolysis of PET, of which the best variant showed 2.4- and 38-fold improved activity at 40 and 50 °C, respectively.

Broadening substrate scope

The aforementioned examples focused on the relationship between protein variants and the catalytic performance of the enzymes studied. However, for the application of biocatalysts, it is often necessary to take advantage of the substrate promiscuity of proteins to explore novel reactions and chemistries. Thus, it is of great practical importance to develop ML models for predicting the substrate scope. For this purpose, it is vital to evaluate not only the complex protein sequence/structure space, but also the potential substrate space. It is also important to collect data for substrate promiscuity, as well as to design substrate descriptors. However, the combination of such strategies has rarely been published. It is worth mentioning that although many ML models have been built for predicting the relationship between homologous proteins sequence/structure and their substrate specificity,[24] these models were not designed for protein engineering, therefore they will not be discussed here.

We recently reported the structure- and data-driven protein engineering of an amine transaminase (ATA) for improving its activity and stereoselectivity as well as broadening its substrate scope.[25] First, variants of the ATA from *Ruegeria* sp. (3FCR) with improved catalytic activity and reversed stereoselectivity were created by a structure-dependent rational design and a high-quality dataset (high-diversity for catalytic stereoselectivity and activity) was collected in this process. Subsequently, a modified one-hot encoding was designed to describe steric and electronic effects of substrates and residues within ATAs. Finally, a gradient boosting regression tree predictor was built for catalytic activity and stereoselectivity. This was then applied to the data-driven design of optimized variants which indeed showed improved activity.

To better assess the relationship between enzyme variants, substrates and catalytic activity as well as stereoselectivity using a limited amount of data, the data-collection experiments were deliberately designed to increase the diversity of ATA variants and to obtain experimental data for a set of 15 substrates (Figure 1A and 1B). First, a structure-dependent rational design strategy was used to rapidly reverse ATA's enantioselectivity and to improve their activity, so as to cover the comprehensive catalytic performance space; secondly, to cover the complex substrate space as comprehensively as possible.

For this, a range of substrates containing different steric and electronic substituents was designed, obtained and experimentally verified toward different ATA variants, so that the epistatic effect between substrate and enzyme can be evaluated by this

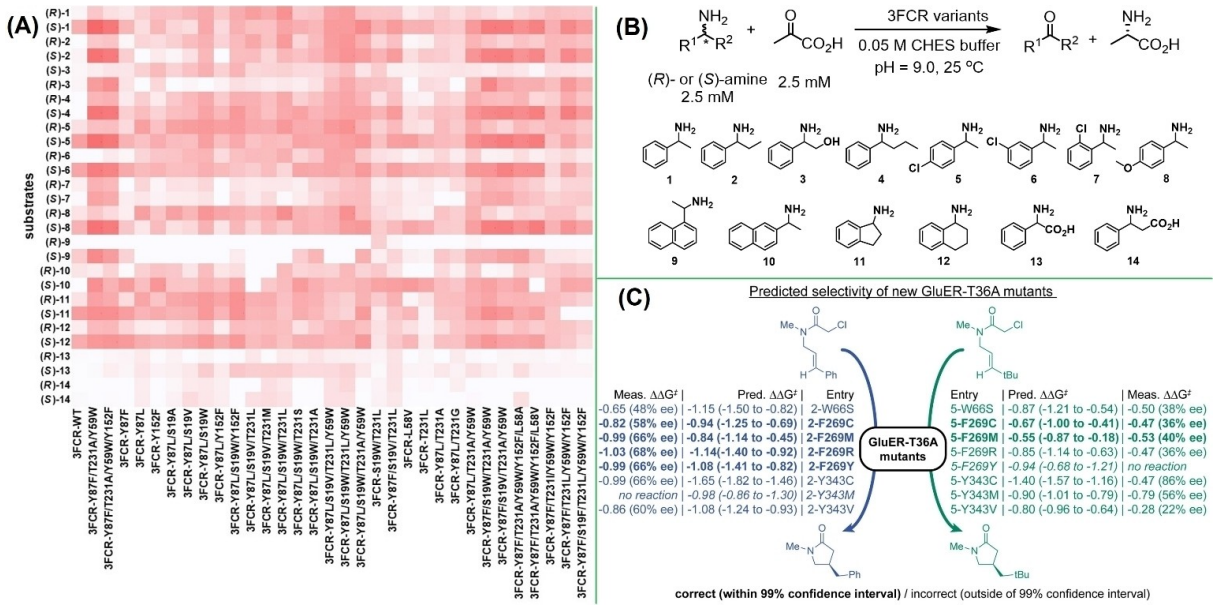


Figure 1. Examples of ML-assisted protein engineering for better catalytic performance toward different substrates. (**A**): Specific activity of 3FCR variants toward (*S*)- and (*R*)-enantiomers of amines. (**B**): Transaminase activity was determined using the acetophenone assay.[26] Both, (*S*)- and (*R*)-amines were assayed with pyruvate as amino acceptor followed by collection of specific activity data for different 3FCR variants.[25] (**C**): Predicted and experimentally verified enantioselectivities of new ene-reductase variants with out-of-sample substrates.[27] The range for the predictions was computed at a 99% confidence interval using bootstrap subsampling.

model; finally a feature extraction method was designed to match the understanding of the underlying catalytic mechanism. For example, each amino acid is encoded into two elements, which respectively represents the electronic and steric properties of a given amino acid residue. Each amine substrate can be divided into a large-pocket binding group and a small-pocket binding group, and each binding group is encoded into four elements to describe its electronic and steric properties.

Recently, the groups of Hyster and Sigman[27] also reported a strategy to investigate the biocatalytic reaction space and to gain an understanding of the molecular mechanisms driving enzymatic transformations. They explored and collected enantioselectivity data (expressed as $\Delta\Delta G^{\neq}$, 50 datapoints) of an ene-reductase. The limited number of datapoints required very informative low-dimensional descriptors, so they employed a more complex approach than the one described in the work by Ao *et al.*[25] to extract protein and ligand structural features: they identified two complementary conformational search platforms (accelerated molecular dynamics and induced fit docking) to acquire the protein-ligand conformational ensembles, and subsequently compute and extract their electronic, steric and dynamic descriptors. Trained by a forward-stepwise multivariate linear regression algorithm, the resultant ML model related structural features of the enzyme and the substrate to enantioselectivity and this information was used to predict $\Delta\Delta G^{\neq}$ values in reactions with out-of-sample substrates and variants (Figure 1C).

Future Directions

Acquisition of large amounts of high-quality data

In the past, it was common for AI experts to collect data from reported publications to train models, and then to validate the models through wet-lab experimentation, therefore data acquisition was usually a one-way and passive process. This causes the problem of a lack of data diversity, especially when a directed evolution strategy is used, the distribution of data in the underlying publications tends to be very imbalanced (too many good results, too few bad results), which affects the generalisation performance of a model derived from it. Enzyme activity data is also very rarely published in a machine understandable form (e.g., with semantic annotations), which impedes automatic data aggregation. We believe that in the future the data used for modelling will be deliberately designed and collected with a more evenly distributed diversity for modelling purposes.[25] The lack of diversity in the data is also found in the overall structure of the enzymes studied. Indeed, while the above example[25] is particularly promising with respect to the possibility of predicting the activity and selectivity of certain enzymes, or even certain scaffolds or subfamilies, it should be noted that one of the main limitations is still that in this example the focus was mainly on certain distinct parts or regions of the protein, and almost exclusively on the first amino acid coordination sphere around the active site. This is largely due to the data sets, particularly experimental data, on which they have been trained, with the proteins tested having a fairly identical overall amino acid structure and sequence, with only a few punctual points of variation located

near the active site. This means that they cannot easily be transposed to other scaffolds within the same enzyme family, or even new families, whose structure is remote from those used to train the algorithms, thus confining them to enzyme engineering approaches and not to the wider exploration of biodiversity or the discovery of enzymes with new activities or other properties.

In order to solve such problems, one solution is to obtain a large amount of high-quality data. Higher reproducibility and availability of enzymatic activity data can be achieved by automating the experimental screening process. Common approaches are microfluidic platforms[28] and classical liquid-handling high-throughput systems,[29a] like the REALCAT platform in Lille[29b] and the LARA system in Greifswald,[29c] offering advantages like high versatility, determination of enzyme kinetics in a high-throughput manner, and providing reliable data similar to HPLC or gas chromatography. Both types of platforms are highly complementary and very promising to collect enzyme activity data in a standardized format for AI and machine learning. This can be achieved by data management systems that (mostly) automatically capture important experimental data and metadata. The vividly developed, open-source project LARAsuite[29d] fills this gap by capturing most of these enumerated aspects and by exploring how far experimentation and semantic annotation can be automated.

The use of structure-based descriptors and algorithms

Although sequence information proved to be enough to predict an enzyme structure,[30] enzyme reactions are dynamic in nature: the substrate moves – to the active site (in many cases) through a dynamically changing tunnel[31] – or at least a funnel or groove shaped cavity, it then binds or interacts with residues in the active site, often loops, lids or even the whole enzyme structure moves, then the conversion of the substrate happens – mostly in a directed way, with many electron density changes, finally the product is released, again accompanied with structural changes.[32] Representing enzymes and substrates via structure based descriptors can provide models with the right inductive bias and this reduces the dimensionality compared to sequence-based representations. With the development of computational (machine learning) models that estimate folded 3D structures from amino acid sequences[30] and other computational tools for automated docking, calculation of conformers,[33a] active site detection,[33b] protein binding interfaces prediction[33c] etc., it is now possible to calculate such conformational changes and to use models that operate on this structural information. A next generation of machine learning algorithms is currently developed which can make use of explicit three-dimensional models, molecular dynamics – and ultimately electrodynamics and quantum effects.[34] In general, molecular dynamics and quantum dynamics can be simulated, but at high (computational) cost. This is especially problematic for the virtual screenings of thousands of molecules. Graph Neural Networks (GNN) offer a simpler, yet effective approach to build regressors or classifiers from 3D structures and docked substrates.[35a] This new type of approach could thus make it easier to capture the structural dynamics mentioned above, especially when embedding several intermediates occurring within an enzymatic reaction mechanism, thereby also transcribing the temporal evolution of the system.[35b–c] However, to date, these approaches have not yet been applied to the prediction of enzyme activity or selectivity, and even less so with experimental verification of the latter, providing room for improvement. This or "classical quantum computation approaches" for molecular dynamics simulations might be the ultimate way to calculate and therefore predict enzymatic activity. It has been shown that this approach can be superior to purely sequence-based ML methods.[34,35]

Development of new algorithms

When there is little data, simple ML algorithms such as ensemble algorithms show a clear advantage.[25] However, the number of available protein sequences increased more than twenty-fold in the last five years (2023:[36] $>$ 2,400 Mio.; 2018:[37] ~123 Mio.). The very large amount of available protein sequences with some annotations and the still large amount of resolved crystallized protein structures (~200,000) have allowed the training of deep learning models for the prediction of 3D structures and functions,[6–8,30] and for sequence completion tasks.[38] The internal representations of proteins in such models necessarily capture properties of the evolutional sequence space like preserved motives, protein domains, contacts in folded proteins. Reusing these representations as inputs for regression models that predict targets for protein engineering is a promising approach in protein engineering known as transfer learning.[39] With more effort, one can simultaneously train a model on a large data task and regression models for specific tasks. We would like to mention that the use of trained ML models for 3D structure predictions and the calculation of structure-based descriptors is also a kind of transfer learning. In addition, large language models and deep generative models may contribute to breakthroughs, especially for refining the enzymatic properties where only low-throughput activity assays are feasible. For example, the probability assigned to a sequence by a large language model is positively correlated with the protein fitness for some engineering tasks[40a] and can also be used to generate protein sequences with a predictable function.[40b] Conditional generative models allow to sample candidates for a certain protein property, including enzymatic activity far away from known wild-type enzymes;[38] Diffusion models were applied for molecular docking between proteins and ligands,[41] or for modelling and *de novo* design of proteins and other biomolecules.[42] It is worth noting that because biocatalytic systems are highly complicated, first-principles-based calculations to predict catalytic performance are often difficult to perform. Although the computational process of deep learning is considered an "ultimate black box", the model can be used for virtual screening to guide the design of mutants. The scoring and ranking of the results can summarise the effect patterns of the mutations, so that a deep learning

model with high prediction accuracy can still help to understand the regularity of protein engineering.

Summary

As the availability of high-quality data continues to increase, feature extraction methods continue to be developed, and new algorithms continue to emerge, data-driven protein engineering has been increasingly applied to improve catalytic activity and selectivity of enzymes. In the future, this field will pay more attention to the generalisation ability of models, the establishment of cross-species biocatalytic large models and making the enzyme activity data machine findable and understandable. This is expected to fundamentally change the existing empirically and trial-and-error based directed evolution research strategies and will reshape the knowledge system of biocatalysis.

Acknowledgements

Financial support from the German Research Foundation (NFDI4CAT) to UTB, the National Natural Science Foundation of China (21977098) to YFA, the German Federal Ministry of Education and Research Program (01DD20002A to SB and 01DD20002C to UTB) is gratefully acknowledged. Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interests

The authors declare no conflict of interest.

Keywords: Biocatalysis · catalytic activity · machine learning · protein engineering · selectivity

[1] a) R. Buller, S. Lutz, R. J. Kazlauskas, R. Snajdrova, J. C. Moore, U. T. Bornscheuer, *Science* **2023**, *382*, eadh8615; b) D. C. Miller, S. V. Athavale, F. H. Arnold, *Nat. Synth.* **2022**, *1*, 18–23; c) D. Yi, T. Bayer, C. P. S. Badenhorst, S. Wu, M. Dörr, M. Höhne, U. T. Bornscheuer, *Chem. Soc. Rev.* **2021**, *50*, 8003–8049; d) U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore, K. Robins, *Nature* **2012**, *485*, 185–194.

[2] a) E. J. Hossack, F. J. Hardy, A. P. Green, *ACS Catal.* **2023**, *13*, 12436–12444; b) E. Alfonzo, A. Das, F. H. Arnold, *Curr. Opin. Green Sustain. Chem.* **2022**, *38*, 100701; c) S. Wu, R. Snajdrova, J. C. Moore, K. Baldenius, U. T. Bornscheuer, *Angew. Chem. Int. Ed.* **2020**, *60*, 88–119; *Angew. Chem.* **2020**, *133*, 89–123; d) J. H. Schrittwieser, S. Velikogne, M. Hall, W. Kroutil, *Chem. Rev.* **2018**, *118*, 270–348; e) J. Mangas-Sanchez, S. P. France, S. L. Montgomery, G. A. Aleku, H. Man, M. Sharma, J. I. Ramsden, G. Grogan, N. J. Turner, *Curr. Opin. Chem. Biol.* **2017**, *37*, 19–25.

[3] a) B. Markus, G. Christian C, D. Andreas, K. Arkadij, L. Stefan, O. Gustav, S. Elina, S. Radka, *ACS Catal.* **2023**, *13*, 14454–14469; b) P. Kouba, P. Kohout, F. Haddadi, A. Bushuiev, R. Samusevich, J. Sedlar, J. Damborsky, T. Pluskal, J. Sivic, S. Mazurenko, *ACS Catal.* **2023**, *13*, 13863–13895; c) B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jiang, Y. Zhu, J. Liu, B. Zhang, G.-W. Wei, *Chem. Rev.* **2023**, *123*, 8736–8780; d) M. Wittmund, F. Cadet, M. D. Davari, *ACS Catal.* **2022**, *12*, 14243–14263; e) Y. Jiang, X. Ran, Z. J. Yang, *Protein Eng. Des. Sel.* **2022**, *36*, gzac009; f) N. Sapoval, A. Aghazadeh, M. G. Nute, D. A. Antunes, A. Balaji, R. Baraniuk, C. J. Barberan, R. Dannenfelser, C. Dun, M. Edrisi, R. A. L. Elworth, B. Kille, A. Kyrillidis, L. Nakhleh, C. R. Wolfe, Z. Yan, V. Yao, T. J. Treangen, *Nat. Commun.* **2022**, *13*, 1728; g) B. L. Hie, K. K. Yang, *Curr. Opin. Struct. Biol.* **2022**, *72*, 145–152; h) S. L. Lovelock, R. Crawshaw, S. Basler, C. Levy, D. Baker, D. Hilvert, A. P. Green, *Nature* **2022**, *606*, 49–58; i) Y. Cui, J. Sun, B. Wu, *Trends Chem.* **2022**, *4*, 409–419; j) D. Yi, T. Bayer, C. P. S. Badenhorst, S. Wu, M. Dörr, M. Höhne, U. T. Bornscheuer, *Chem. Soc. Rev.* **2021**, *50*, 8003–8049; k) N. Singh, S. Malik, A. Gupta, K. R. Srivastava, *Emerg. Top. Life Sci.* **2021**, *5*, 113–125; l) Y. Xu, D. Verma, R. P. Sheridan, A. Liaw, J. Ma, N. M. Marshall, J. McIntosh, E. C. Sherer, V. Svetnik, J. M. Johnston, *J. Chem. Inf. Model.* **2020**, *60*, 2773–2790; m) M. J. Volk, I. Lourentzou, S. Mishra, L. T. Vo, C. Zhai, H. Zhao, *ACS Synth. Biol.* **2020**, *9*, 1514–1533; n) S. Mazurenko, Z. Prokop, J. Damborsky, *ACS Catal.* **2021**, *11*, 12433–12445; o) E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, *Nat. Methods* **2019**, *16*, 1315–1322; p) K. K. Yang, Z. Wu, F. H. Arnold, *Nat. Methods* **2019**, *16*, 687–694.

[4] The UniProt Consortium, *Nucleic Acids Res.* **2023**, *51*, D523–D531.

[5] A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitz, I. Schomburg, M. Neumann-Schaal, D. Jahn, D. Schomburg, *Nucleic Acids Res.* **2021**, *49*, D498–D508.

[6] S. Gelman, S. A. Fahlberg, P. Heinzelman, P. A. Romero, A. Gitter, *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2104878118.

[7] T. Yu, H. Cui, J. C. Li, Y. Luo, G. Jiang, H. Zhao, *Science* **2023**, *379*, 1358–1363.

[8] a) H. T. Rube, C. Rastogi, S. Feng, J. F. Kribelbauer, A. Li, B. Becerra, L. A. N. Melo, B. V. Do, X. Li, H. H. Adam, N. H. Shah, R. S. Mann, H. J. Bussemaker, *Nat. Biotechnol.* **2022**, *40*, 1520–1527; b) M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, P. Siedlecki, *Bioinformatics* **2018**, *34*, 3666–3674.

[9] F. Li, L. Yuan, H. Lu, G. Li, Y. Chen, M. K. M. Engqvist, E. J. Kerkhoven, J. Nielsen, *Nat. Catal.* **2022**, *5*, 662–672.

[10] a) S. Lauterbach, H. Dienhart, J. Range, S. Malzacher, J.-D. Spöring, D. Rother, M. F. Pinto, P. Martins, C. E. Lagerman, A. S. Bommarius, A. V. Høst, J. M. Woodley, S. Ngubane, T. Kudanga, F. T. Bergmann, J. M. Rohwer, D. Iglezakis, A. Weidemann, U. Wittig, C. Kettner, N. Swainston, S. Schnell, J. Pleiss, *Nat. Methods* **2023**, *20*, 400–402; b) J. Range, C. Halupczok, J. Lohmann, N. Swainston, C. Kettner, F. T. Bergmann, A. Weidemann, U. Wittig, S. Schnell, J. Pleiss, *FEBS J.* **2022**, *289*, 5864–5874.

[11] a) N. Swainston, A. Baici, B. M. Bakker, A. Cornish-Bowden, P. F. Fitzpatrick, P. Halling, T. S. Leyh, C. O'Donovan, F. M. Raushel, U. Reschel, J. M. Rohwer, S. Schnell, D. Schomburg, K. F. Tipton, M.-D. Tsai, H. V. Westerhoff, U. Wittig, R. Wohlgemuth, C. Kettner, *FEBS J.* **2018**, *285*, 2193–2204; b) R. N. Goldberg, R. T. Giessmann, P. J. Halling, C. Kettner, H. V. Westerhoff, *Beilstein J. Org. Chem.* **2023**, *19*, 303–316.

[12] R. J. Fox, S. C. Davis, E. C. Mundorff, L. M. Newman, V. Gavrilovic, S. K. Ma, L. M. Chung, C. Ching, S. Tam, S. Muley, J. Grate, J. Gruber, J. C. Whitman, R. A. Sheldon, G. W. Huisman, *Nat. Biotechnol.* **2007**, *25*, 338–344.

[13] T. van den Bergh, G. Tamo, A. Nobili, Y. Tao, T. Tan, U. T. Bornscheuer, R. R. K. P. Kuipers, B. Vroling, R. de Jong, V. Subramanian, P. J. Schaap, T. Desmet, B. Niedetzky, G. Vriend, H. J. Joosten, *PLoS One* **2017**, *12*, e0176427.

[14] H. Jochens, D. Aerts, U. T. Bornscheuer, *Prot. Eng. Des. Sel.* **2010**, *23*, 903–909.

[15] H. Jochens, U. T. Bornscheuer, *ChemBioChem* **2010**, *11*, 1861–1866.

[16] H. J. Wijma, R. J. Floor, P. A. Jekel, D. Baker, S. J. Marrink, D. B. Janssen, *Protein Eng. Des. Sel.* **2014**, *27*, 49–58.

[17] F. Cadet, N. Fontaine, G. Li, J. Sanchis, M. N. F. Chong, R. Pandjaitan, I. Vetrivel, B. Offmann, M. T. Reetz, *Sci. Rep.* **2018**, *8*, 16757.

[18] Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 8852–8858.

[19] E. J. Ma, E. Siirola, C. Moore, A. Kummer, M. Stoeckli, M. Faller, C. Bouquet, F. Eggimann, M. Ligibel, D. Huynh, G. Cutler, L. Siegrist, R. A. Lewis, A.-C. Acker, E. Freund, E. Koch, M. Vogel, H. Schlingensiepen, E. J. Oakeley, R. Snajdrova, *ACS Catal.* **2021**, *11*, 12433–12445.

[20] J. C. Greenhalgh, S. A. Fahlberg, B. F. Pfleger, P. A. Romero, *Nat. Commun.* **2021**, *12*, 5825.

[21] J. Büchler, S. H. Malca, D. Patsch, M. Voss, N. J. Turner, U. T. Bornscheuer, O. Allemann, C. L. Chapelain, A. Lumbroso, O. Loiseleur, R. Buller, *Nat. Commun.* **2022**, *13*, 371.

[22] H. Lu, D. J. Diaz, N. J. Czarnecki, C. Zhu, W. Kim, R. Shroff, D. J. Acosta, B. R. Alexander, H. O. Cole, Y. Zhang, N. A. Lynd, A. D. Ellington, H. S. Alper, *Nature* **2022**, *604*, 662–667.

[23] R. Shroff, A. W. Cole, D. J. Diaz, B. R. Morrow, I. Donnell, A. Annapareddy, J. Gollihar, A. D. Ellington, R. Thyer, *ACS Synth. Biol.* **2020**, *9*, 2927–2935.

[24] a) Q, Zhang, W. Zheng, Z. Song, Q. Zhang, L. Yang, J. Wu, J. Lin, G. Xu, H. Yu, *ACS Synth. Biol.* **2023**, *12*, 2403–2417; b) S. Goldman, R. Das, K. K.

Yang, C. W. Coley, *PLoS Comput. Biol.* **2022**, *18*, e1009853; c) Z. Mou, J. Eakes, C. J. Cooper, C. M. Foster, R. F. Standaert, M. Podar, M. J. Doktycz, J. M. Parks, *Proteins* **2021**, *89*, 336–347; d) S. L. Robinson, M. D. Smith, J. E. Richman, K. G. Aukema, L. P. Wackett, *Synth. Biol.* **2020**, *5*, ysaa004; e) M. Yang, C. Fehl, K. V. Lees, E.-K. Lim, W. A. Offen, G. J. Davies, D. J. Bowles, M. G. Davidson, S. J. Roberts, B. G. Davis, *Nat. Chem. Biol.* **2018**, *14*, 1109–1117.

[25] Y.-F. Ao, S. Pei, C. Xiang, M. J. Menke, L. Shen, C. Sun, M. Dörr, S. Born, M. Höhne, U. T. Bornscheuer, *Angew. Chem. Int. Ed.* **2023**, *62*, e202301660; *Angew. Chem.* **2023**, *136*, e202301660.

[26] S. Schätzle, M. Höhne, E. Redestad, K. Robins, U. T. Bornscheuer, *Anal. Chem.* **2009**, *81*, 8244–8248.

[27] H. D. Clements, A. R. Flynn, B. T. Nicholls, D. Grosheva, S. J. Lefave, M. T. Merriman, T. K. Hyster, M. S. Sigman, *J. Am. Chem. Soc.* **2023**, *145*, 17656–17664.

[28] a) M. Gantz, S. Neun, E. J. Medcalf, L. D. van Vliet, F. Hollfelder, *Chem. Rev.* **2023**, *123*, 5571–5611; b) T. Yang, K. J. Buholzer, A. Sottini, X. Cao, A. deMello, D. Nettels, B. Schuler, *Nat. Methods* **2023**, *20*, 1479–1482; c) M. Vasina, D. Kovar, J. Damborsky, Y. Ding, T. Yang, A. deMello, S. Mazurenko, S. Stavrakis, Z. Prokop, *Biotechnol. Adv.* **2023**, *66*, 108171.

[29] a) T. Yu, A. G. Boob, N. Singh, Y. Su, H. Zhao, *Cell Sys.* **2023**, *14*, 633–644; b) S. Paul, S. Heyte, B. Katryniok, C. Garcia-Sancho, P. Maireles-Torres, F. Dumeignil, *Oil Gas Sci. Technol.* **2015**, *70*, 455–462; c) M. Dörr, M. P. C. Fibinger, D. Last, S. Schmidt, J. Santos-Aberturas, D. Böttcher, A. Hummel, C. Vickers, M. Voss, U. T. Bornscheuer, *Biotechnol. Bioeng.* **2016**, *113*, 1421–1432; d) http://www.tib-op.org/ojs/index.php/CoRDI/article/view/359.

[30] a) Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos S. Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, *Science* **2023**, *379*, 1123–1130; b) M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, *Science* **2021**, *373*, 871–876; c) J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583–589; d) A. H.-W. Yeh, C. Norn, Y. Kipnis, D. Tischer, S. J. Pellock, D. Evans, P. Ma, G. R. Lee, J. Z. Zhang, I. Anishchenko, B. Coventry, L. Cao, J. Dauparas, S. Halabiya, M. DeWitt, L. Carter, K. N. Houk, D. Baker, *Nature* **2023**, *614*, 774–780.

[31] a) E. Chovancova, A. Pavelka, P. Benes, O. Strnad, J. Brezovsky, B. Kozlikova, A. Gora, V. Sustr, M. Klvana, P. Medek, L. Biedermannova, J. Sochor, J. Damborsky, *PLoS Comput. Biol.* **2012**, *8*, e1002708; b) A. Jurcik, D. Bednar, J. Byska, S. M. Marques, K. Furmanova, L. Daniel, P. Kokkonen, J. Brezovsky, O. Strnad, J. Stourac, A. Pavelka, M. Manak, J. Damborsky, B. Kozlikova, *Bioinform.* **2018**, *34*, 3586–3588.

[32] a) R. M. Daniel, R. V. Dunn, J. L. Finney, J. C. Smith, *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 69–92; b) K.-W. Chen, T.-Y. Sun, Y.-D. Wu, *Chem. Rev.* **2023**, *123*, 9940–9981.

[33] a) A. D. St-Jacques, J. M. Rodriguez, M. G. Eason, S. M. Foster, S. T. Khan, A. M. Damry, N. K. Goto, M. C. Thompson, R. A. Chica, *Nat. Commun.* **2023**, *14*, 6058; b) F. Pazos, *Adv. Protein Chem. Struct. Biol.* **2022**, *130*, 39–57; c) L. F. Krapp, L. A. Abriata, F. C. Rodriguez, M. D. Peraro, *Nat. Commun.* **2023**, *14*, 2175.

[34] a) Z. Yang, W. Zhong, Q. Lv, T. Dong, C. Y.-C. Chen, *J. Phys. Chem. Lett.* **2023**, *14*, 2020–2033; b) P. J. Ollitrault, A. Miessen, I. Tavernelli, *Acc. Chem. Res.* **2021**, *54*, 4229–4238; c) P. A. Romero, A. Krause, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2012**, *110*, E193–E201.

[35] a) C. Lu, J. H. Lubin, V. V. Sarma, S. Z. Stentz, G. Wang, S. Wang, S. D. Khare, *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2303590120; b) A. Longa, V. Lachi, G. Santin, M. Bianchini, B. Lepri, P. Lio, F. Scarselli, A. Passerini, *arXiv preprint* **2023**, DOI: 10.48550/arXiv.2302.01018; c) V. Gligorijevic, P. D. Renfrew, T. Kosciolek, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis, R. J. Xavier, R. Knight, K. Cho, R. Bonneau, *Nat. Commun.* **2021**, *12*, 3168.

[36] L. Richardson, B. Allen, G. Baldi, M. Beracochea, M. L. Bileschi, T. Burdett, J. Burgin, J. Caballero-Pérez, G. Cochrane, L. J. Colwell, T. Curtis, A. Escobar-Zepeda, T. A. Gurbich, V. Kale, A. Korobeynikov, S. Raj, A. B. Rogers, E. Sakharova, S. Sanchez, D. J. Wilkinson, R. D. Finn, *Nucleic Acids Res.* **2023**, *51*, D753–D759.

[37] M. Steinegger, J. Söding, *Nat. Commun.* **2018**, *9*, 2542.

[38] Z. Wu, K. E. Johnston, F. H. Arnold, K. K. Yang, *Curr. Opin. Chem. Biol.* **2021**, *65*, 18–27.

[39] a) E. Fenoy, A. A. Edera, G. Stegmayer, *Brief Bioinform.* **2022**, *23*, bbac232; b) A. Kroll, S. Ranjan, M. K. M. Engqvist, M. J. Lercher, *Nat. Commun.* **2023**, *14*, 2787; c) B. J. Wittmann, Y. Yue, F. H. Arnold, *Cell Sys.* **2021**, *12*, 1026–1045.e7.

[40] a) A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos Jr., C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, N. Naik, *Nat. Biotechnol.* **2023**, *41*, 1099–1106; b) C. Hsu, H. Nisonoff, C. Fannjiang, J. Listgarten, *Nat. Biotechnol.* **2022**, *40*, 1114–1122.

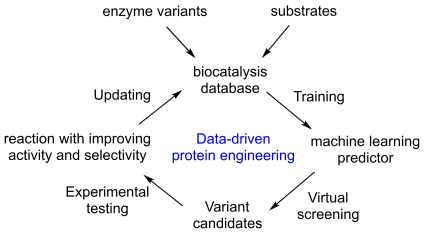
[41] G. Corso, H. Stärk, B. Jing, R. Barzilay, T. Jaakkola, *arXiv preprint* **2022**, DOI: 10.48550/arXiv.2210.01776.

[42] a) J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. D. Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, *Nature* **2023**, *620*, 1089–1100; b) R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G. R. Lee, F. S. Morey-Burrows, I. Anishchenko, I. R. Humphreys, R. McHugh, D. Vafeados, X. Li, G. A. Sutherland, A. Hitchcock, C. N. Hunter, M. Baek, F. DiMaio, D. Baker, *bioRxiv preprint* **2023**, DOI: 10.1101/2023.10.09.561603.

Manuscript received: November 3, 2023
Revised manuscript received: November 28, 2023
Accepted manuscript online: November 29, 2023
Version of record online: November 29, 2023

CONCEPT

Machine learning approaches allow the creation of protein adaptive landscapes and the identification of catalytic modes using limited experimental data to establish relationships between enzyme, substrate and catalytic performance, and have been used for data-driven protein engineering to improve their catalytic activity and selectivity.



Dr. Y.-F. Ao, Dr. M. Dörr, M.Sc. M. J. Menke, Dr. S. Born, Dr. E. Heuson, Prof. Dr. U. T. Bornscheuer**

VIP