ACS Catalysis

Perspective

Cite This: ACS Catal. 2020, 10, 1210−1223

pubs.acs.org/acscatalysis

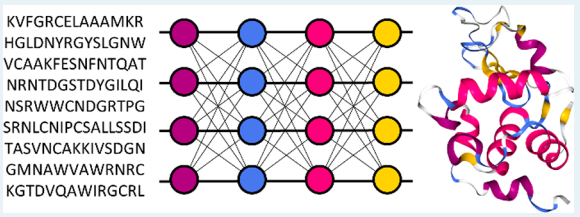
Machine Learning in Enzyme Engineering

Stanislav Mazurenko,*[,†] Zbynek Prokop,[†,‡] and Jiri Damborsky[†,‡]

[†]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic

[‡]International Centre for Clinical Research, St. Ann's Hospital, 602 00 Brno, Czech Republic

ABSTRACT: Enzyme engineering plays a central role in developing efficient biocatalysts for biotechnology, biomedicine, and life sciences. Apart from classical rational design and directed evolution approaches, machine learning methods have been increasingly applied to find patterns in data that help predict protein structures, improve enzyme stability, solubility, and function, predict substrate specificity, and guide rational protein design. In this Perspective, we analyze the state of the art in databases and methods used for training and validating predictors in enzyme engineering. We discuss current limitations and challenges which the community is facing and recent advancements in experimental and theoretical methods that have the potential to address those challenges. We also present our view on possible future directions for developing the applications to the design of efficient biocatalysts.



KEYWORDS: *artificial intelligence, enantioselectivity, function, mechanism, protein engineering, structure−function, solubility, stability*

1. INTRODUCTION

Enzyme engineering is the process of customizing new biocatalysts with improved properties by altering their constituting sequences of amino acids. Despite the immensity of possible alterations, this procedure has already yielded remarkable results in new designs and optimization of enzymes for chemical and pharmaceutical biosynthesis, regenerative medicine, food production, waste biodegradation and biosensing.[1−4] Enzymes are typically, but not exclusively, engineered for catalytic activity, substrate specificity, enantioselectivity, thermodynamic stability, stability in cosolvents, expressibility, and solubility.

The two established and widely used enzyme engineering strategies are rational design[5,6] and directed evolution.[7,8] The former approach is based on the structural analysis and in-depth computational modeling of enzymes by accounting for the physicochemical properties of amino acids and simulating their interactions with the environment. The latter approach takes after the natural evolution in using mutagenesis for iterative production of mutant libraries, which are then screened for enzyme variants with the desired properties. These two strategies may naturally complement each other: e.g., site-directed or saturation mutagenesis may be applied on the rationally chosen hotspots.[9] While both strategies show remarkable results, they require a substantial amount of computational and/or experimental effort in each particular case of a biocatalyst optimization.

Machine learning (ML) is a third approach to designing new biocatalysts that has been gaining attention in the past few decades. Unlike the model-driven rational design, this strategy is data-driven in that it identifies patterns in the existing data to predict properties of the previously unseen but similar input. Unlike iterative selecting of the existing mutants in directed evolution, ML-based design can generate new, previously unseen but promising variants, based on the patterns in the collected data. And similarly to the complementarity of the rational design and directed evolution, ML is being used in combination with the two.[5,10] Its increasing popularity stems from its spectacular performance in some tasks previously deemed impossible or extremely hard algorithmically: natural language processing, handwriting and facial recognition, fraud and spam detection, web search, etc.[11] Recent advances in the analysis of human genetic variation data in biomedicine and healthcare further increase the appeal of this approach for the design of beneficial mutations.[12−14]

Multiple ML algorithms have already been applied to enzyme engineering. Some notable examples include random forests used to predict protein solubility,[15] support vector machines[16,17] and decision trees[18] to predict enzyme stability changes upon mutations, *K*-nearest-neighbor classifiers to predict enzyme function[19] and mechanisms,[20] and various scoring and clustering algorithms for rapid functional sequence annotation.[21,22] The main attractiveness of ML in enzyme engineering stems from its generalizability: once it is trained on the known input, called a training set, an ML algorithm can potentially make predictions about new variants almost instantly. In contrast, the rational design approach often requires the construction of a new model, which might take months of intensive calculations and processing, and the directed evolution approach will most likely involve months of intensive experimentation. However, the success of an ML predictor for previously unseen data crucially depends both on

Received: October 7, 2019
Revised: December 13, 2019
Published: December 13, 2019

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

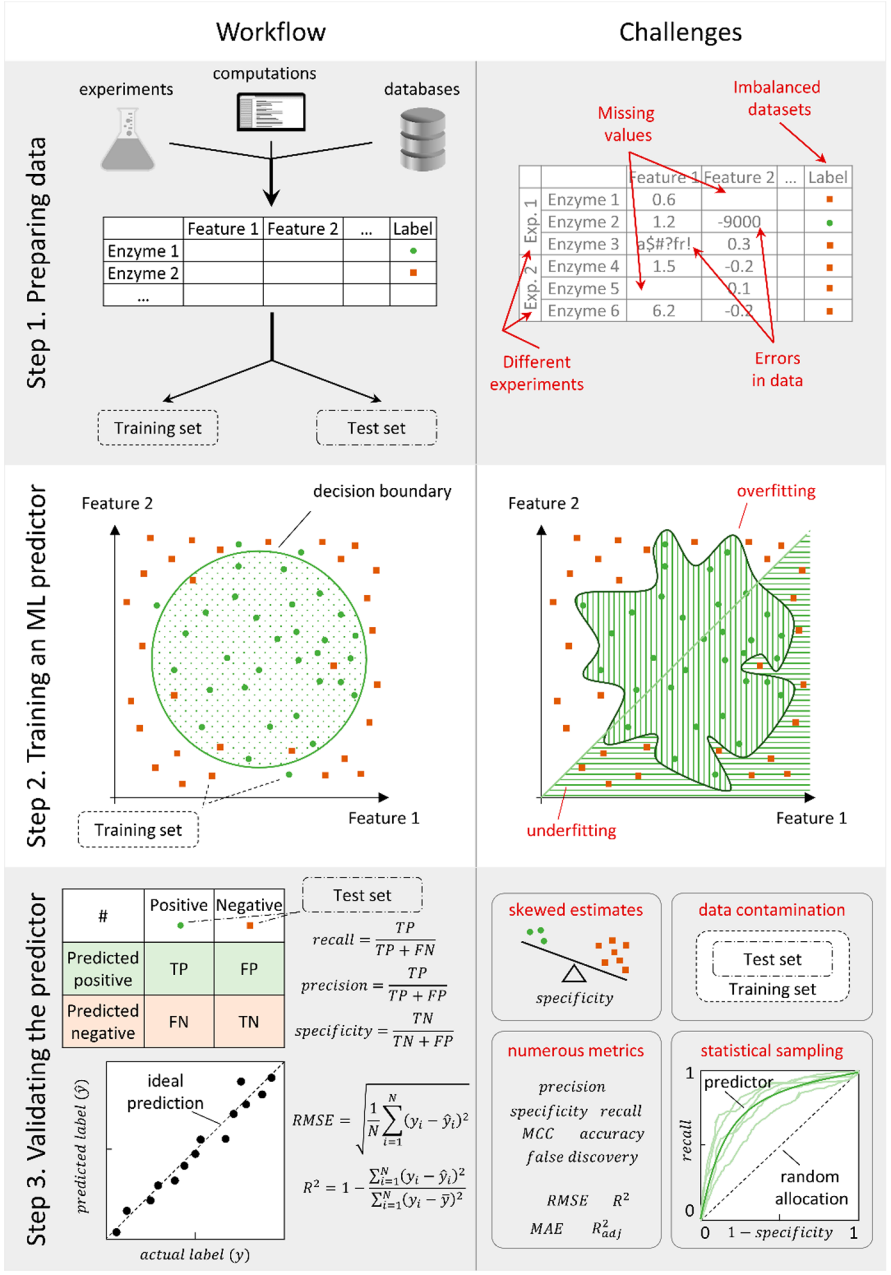


Figure 1. Schematic workflow of constructing an ML predictor and associated challenges. Step 1: the data are usually turned into a table format and split into the training and test parts. Any errors, biases, or imbalances will be translated to the predictor's performance and, hence, must be accounted for. Step 2: the predictor is trained on the training data set. For example, a decision boundary is derived that allows classifying future input based on whether data points are inside or outside the boundary. This is a balancing act between two extremes: explaining noise rather than fundamental dependencies (overfitting) or failure to account for complex dependencies in the data (underfitting). Step 3: the performance of the predictor is evaluated based on the test data set. For example, true and false positives and negatives and the associated measures are calculated or the root mean square error (RMSE) is calculated for continuous labels. The random nature of the initial data split as well as data imbalances might skew the evaluation, and numerous metrics used for evaluation vary in their robustness to different data skews. Even partial inclusion of the test set at any stage of ML predictor training is called data contamination and usually invalidates the final evaluation.

the quality of the data used for training and the efficiency of the underlying algorithm. The great diversity of enzyme mechanisms, reactions, and experimental conditions presents the major challenge for applying ML to biocatalyst design due to necessary strict quality control for data collection and reporting, difficult standardization of data format, lack of sizable homogeneous data sets for model training, and slow new data collection for model testing.

The aim of this Perspective is, therefore, to highlight recent advances in data collection and algorithm implementation for ML in enzyme engineering. In particular, we discuss how those developments are affected by available and upcoming experimental techniques and recent advances in mathematical and computational algorithms. We also present our view on the main challenges and possible course of evolution of the ML methods for designing efficient biocatalysts. For recent results and comprehensive articles on ML-guided directed evolution,

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

Table 1. Examples of Databases Used for ML in Enzyme Engineering Grouped by Properties

database	property	size	advantages	disadvantages	location	ref
InterPro	consortium of databases of protein families and domains	35020 entries based on 48938 signatures	actively maintained; UniProtKB sequences covered by 81%; includes homologous superfamilies; discontinuous domains supported	inhomogeneous output due to combining entries from 14 databases	http://www.ebi.ac.uk/interpro/	27
UniProtKB	protein sequences	>120 million automatically annotated; >550000 manually curated	actively maintained; largest database of sequences; numerous organisms represented; evidence scores provided	redundant annotations; limited reliability of TrEMBL; many proteins with unknown function	https://www.uniprot.org/	28
PDB Protein Data Bank	protein structures	>140000 manually curated	actively maintained; largest database of 3D structures; numerous organisms represented; extensive annotations	missing or inaccurate structures of many membrane-bound and difficult to express proteins; poor standardization of PDB format; missing transition states	https://www.rcsb.org/	29
Brenda	functional data: reaction, specificity, kinetic parameters and profiles, genomic sequences, structures, stability	∼84000 enzymes manually annotated; ∼1600000 entries based on text mining	actively maintained; comprehensive coverage; numerous organisms represented (11000); includes mutants	inhomogeneous data; numerous data sources used require stricter quality control	https://www.brenda-enzymes.org/	30
BKMS-react	functional data: reaction, kinetic parameters and profiles, experimental conditions, pathways	69981 unique reactions from BRENDA, KEGG, MetaCyc, and SABIO-RK	actively maintained; comprehensive coverage; reaction oriented; provides overviews on all associated pathways	inhomogeneous data; numerous data sources used require stricter quality control	http://bkms-react.tu-bs.de/	30
EzCatDb Enzyme Catalytic-mechanism Database	functional data: reaction, cofactors, intermediates, catalytic domains, structures	871 enzymes	high-quality curating of reaction intermediates; wide coverage of enzyme classes; overall reactions rather than individual steps are classified	inhomogeneous data; numerous data sources used, hence, stricter quality control needed; small data set; rare updates	http://ezcatdb.cbrc.jp/EzCatDB/	31
M-CSA[b] Mechanism and Catalytic Site Atlas	annotation of catalytic residues, cofactors, and the reaction mechanisms	961 manually curated entries: 423 with detailed mechanism and 538 with catalytic site residues only	actively maintained; 81% coverage of the 3rd level of EC numbers; Rating of redundant mechanisms based on evidence; reaction steps are annotated in detail	inhomogeneous data; numerous data sources require stricter quality control	https://www.ebi.ac.uk/thornton-srv/m-csa/	32
FireProt DB	thermostability change upon mutations	1329 single-point mutations from 79 proteins, manually curated	actively maintained; preprocessed for the machine learning applications new data from recent publications added	the amount of data is not sufficient for the application of advanced ML methods	https://loschmidt.chemi.muni.cz/fireprotdb/	5
ProTherm	thermostability change upon mutations	3464 single-point mutations from 135 proteins; 1564 entries from 99 proteins after a cleanup	largest database of single-point mutants with stability data	not actively maintained; erroneous entries; mishandling multistep unfolding	https://www.iitm.ac.in/bioinfo/ProTherm/	33, 34
eSOL	solubility based on in vitro protein translation and centrifugation	3147 proteins	highly uniform and consistent data set	not actively maintained; single organism (*E. coli* ORF library); a low number of negative samples	http://www.tanpaku.org/tp-esol/index.php?lang=en	35
SoluProtMut DB	solubility change upon mutations	100 proteins >10000 single-point and multiple-point mutations	recent compilation of mutations changing protein solubility	the amount of data not sufficient for the application of advanced ML methods; inhomogeneous data collected using different techniques	https://loschmidt.chemi.muni.cz/soluprotmutdb/	*c*
TargetTrack[a]	solubility	297404 proteins, 961548 trials	largest source of experimental data; description of experimental protocols used	not actively maintained; low-quality annotations of trials and expression systems; different extraction methods used; suboptimal database size	http://dx.doi.org/10.5281/zenodo.821654	36
ProtaBank	all types of protein mutational data from protein engineering	>700 unique proteins and >1800000 mutants	recent and actively maintained; provides detailed descriptions of experiments; stores full mutant sequences; built-in BLAST search for similar mutations	manual curation only at the initial stage; numerous data sources used and very heterogeneous data, hence, stricter quality control and reliability scoring needed	https://protabank.org/	37

[a]Previously known as PepcDB or TargetDB. [b]Merge of MACiE and CSA. [c]Under preparation for publication.

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

analysis in systems metabolic engineering, implementation of biocatalysts in systems in the industry, and biosystems design, we refer our readers to the reviews in refs 10 and 23−25.

2. THE ESSENCE OF MACHINE LEARNING

The essence of most ML algorithms is to find patterns in the available data. These data usually consist of data points with several features or descriptors, e.g. enzyme sequences, their secondary and tertiary structures, substitutions, physicochemical properties of amino acids, etc. That number of features usually varies from dozens to thousands, rendering the problem high-dimensional. The two major types of ML are unsupervised and supervised learning. In *unsupervised learning*, the goal is either to compress the high-dimensional data into a lower number of dimensions or to find data clusters. In *supervised learning* (Figure 1), one or several target properties, such as enzyme activity or stability, are designated as labels, and the goal is to engineer a predictor that will return labels for unseen data points on the basis of their descriptors, using a labeled training data set. Quite often, these two ML types—supervised and unsupervised—are combined, which is called semisupervised learning. In this article, we mainly focus on supervised learning, since enzyme engineers typically aim to improve various enzyme properties.

A schematic workflow of supervised ML algorithms is presented in Figure 1. The most time-consuming stage is usually data collection and preparation for feeding to an ML-based algorithm (step 1). Then the data are split into training and test subsets: the former is used to fine-tune parameters of an ML-based predictor (step 2), whereas the latter is used for final evaluation (step 3). For classification problems with binary labels or labels from a finite number of options, this evaluation is usually based on the confusion matrix: the number of true/false positives and negatives.[26] For regression problems with the labels taking continuous values, the root mean squared error is usually calculated. In either case, the final evaluation is performed on the test data set, which is essential since the ultimate goal is to achieve the predictor's generalizability on the data not used for training. For this reason, in protein engineering, sequence similarities in both data subsets must thus be accounted for. If some protein family is overrepresented in the training set, the resulting predictor might be biased toward discerning patterns valid for this family only. If some sequences in the test set are too close to the training set, the final performance evaluation will yield overoptimistic results.

At the training step 2, fine-tuning a predictor or selecting among several predictors is also possible, usually by means of the K-fold validation. In this case, the training data is further subsplit into K subsets, and the training workflow is repeated K times with each of the K subsets held out for evaluation and the remaining $K - 1$ subsets used for training. The average performance is then used to navigate in the fine-tuning. The main challenge of step 2 in any supervised ML training is to avoid data underfitting (high bias) and overfitting (high variance). *Underfitting* occurs when a predictor fails to find patterns even in the training data: e.g., when a simple linear model is used to explain nonlinear data dependencies. *Overfitting* occurs when the performance of a predictor diminishes dramatically on the test data set in comparison to the training set due to learning too much detail and noise instead of identifying *general patterns*. Both underfitting and overfitting may arise due to insufficient data quality: excessive noise, irrelevant or missing features, data bias, or sparseness. They can also occur due to poor application of an algorithm: excessive or insufficient flexibility in parameter selection, improper training protocol, or contamination of the training data set with the test data set. In the following sections, we summarize the state of the art and challenges related to both databases used for training and applications of ML algorithms in enzyme engineering.

3. DATABASES RELEVANT TO ENZYME ENGINEERING

3.1. The State of the Art in Data Accumulation. Since ML algorithms heavily rely on data, the importance of the *data set quality* used for training can hardly be overestimated. Notable examples of databases often used in enzyme engineering, along with some pros and cons frequently reported by their users, are present in Table 1. The most abundant are databases of protein *sequences* followed by databases of protein *structures*. Protein *stability* and *solubility* are the next two qualities that have routinely been measured for several decades now. A more challenging task is annotating *catalytic properties* of enzymes due to the abundance of reaction types, mechanisms, cofactors, conditions, wide ranges of substrate specificities, enantioselectivities, and promiscuities. However, several excellent databases are addressing this challenge from various perspectives.

3.2. Current Challenges Related to Databases. In general, if the sought-for dependence is simply not in the available data, no amount of new data points will help improve the quality of an ML predictor. In the case of enzyme engineering, however, we expect enzymatic functions to be encoded in the sequences and thus to depend on physicochemical properties of amino acids; hence, both the quantity and quality of data in those databases are of paramount importance for designing an ML predictor. We note that these databases were mostly gathered without any ML application in mind, which causes the following problems. As far as data from a single round of experimentation is concerned, data sets usually cover from dozens to hundreds of variants due to resource and time limitations, which is considered a relatively small sample size in the ML framework. While combining data from different experiments may partially resolve the problem of insufficient data, the issues with data consistency and comparability arise: even when each team's data collection is systematic, it is systematic in its own way, and protocols or consistent dictionaries allowing robust pooling of data have yet to be developed.[38] The lack of established reporting standards often results in missing or, even worse, erroneous values for some descriptors, for instance, opposite signs of the change in some value upon introduction of a mutation.[34] This is further exacerbated by the lack of robust data analysis protocols, such as those used for curve fitting to determine melting temperatures or kinetic rate constants, which often raises doubts about the quality of the reported estimates. In addition to that, recent advancements in the methods for experimental sciences may render some previous results obsolete.

Manual curation certainly helps improve the data quality but is also not a panacea, since earlier studies revealed several issues such as misannotations of protein functions and propagating errors from already disproved results.[39] This necessitates intensive quality control and regular cleanup procedures, which is no longer implemented for some

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

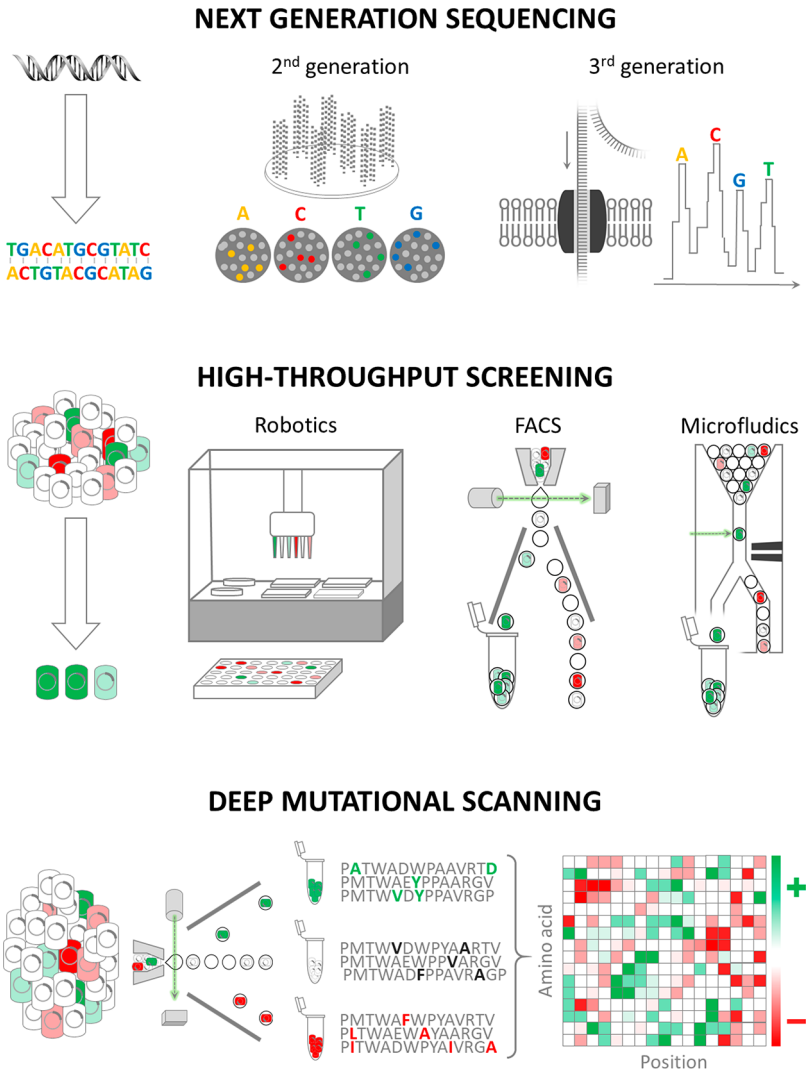


Figure 2. Schematic representation of the methods applicable for collection of robust and reliable data. (top) Next-generation sequencing (NGS) offers the high-throughput analysis of DNA/RNA sequences in the gigabase range per instrument. Second-generation instruments increased the throughput and accuracy by massive parallelization of short (100′ bp) DNA fragments reads after amplification. The third-generation (long-read) methods employ a single-molecule real-time sequencing of long DNA fragments (>1 Mbp). (middle) High-throughput screening (HTS) includes a wide range of different approaches: (i) liquid handling robotics with average throughput of 10^4 variants per day, (ii) fluorescence-activated cell sorting (FACS) enabling screening of up to 10^8 variants per day, and (iii) microfluidics with the production speed of up to 10^8 reaction droplets per day. (bottom) Deep mutational scanning (DMS) coupling high-throughput screening with next-generation sequencing offers a powerful strategy for comprehensively analyzing sequence−function relationships in enzymes.

databases, as those are no longer actively maintained. Apart from manual double-checking, those procedures may also involve "data tidying" to render data sets ML friendly: representing data in a table format with features in columns and observations in rows.[40] Following the increasingly popular FAIR principles—findable, accessible, interoperable, and reusable data—should also enhance the ability of machines to automatically find and use the data.[41] More specifically for enzymes, following the guidelines of the standards for reporting enzyme data (STRENDA) project should increase the data quality, especially in heterogeneous databases collected from multiple sources.[42]

In addition to the data quality, reporting, and organization problems outlined above, experimental designs themselves may become an issue for ML applications. The selection of the protein test variants is usually skewed toward anticipated best performers, and negative results are often not reported. This introduces data biases, which affect the performance of ML-based predictors, since the available parameter space is not sampled uniformly.[43,44] Those biases in databases relevant to protein science have only recently started to attract researchers' attention[45,46] and have yet to be explored and corrected by the community.

Developing new ML predictors is dramatically boosting the demand for improving the existing databases and generating new, more uniform, and representative data sets of higher quality. The former is obstructed by improving scientific methods and, consequently, the pressing need for a review and replication of the results published earlier. Hence, the latter option becomes more attractive, and in our opinion, several up-and-coming experimental techniques promising in this respect have already been presented: (i) next-generation sequencing, (ii) fluorescence-activated cell sorting, (iii) deep mutational scanning, and (iv) microfluidics. These techniques

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

Table 2. Selected Examples of the Application of Machine Learning in the Field of Enzyme Engineering

year	object	target property	data[a]	model and method	ref
1997	haloalkane dehalogenase	function	15 mutants of haloalkane dehalogenase[b]	partial least-squares regression	72
1998	subtilisin, haloalkane dehalogenase, T4 lysozyme, tryptophan synthase	function and stability	19 mutants of subtilisin, 15 mutants of haloalkane dehalogenase, 13 mutants of T4 lysozyme, 18 mutants of tryptophan synthase[b]	partial least-squares regression, principal component analysis	73
1999	human acetylcholinesterase	expression and activity	35 dipeptide mutants	partial least-squares regression, principal component analysis	74
2000	prolyl endopeptidase and thermolysin	fitness landscape: thermostability and enzymatic activity	19 mutants of prolyl endopeptidase and 16 mutants of thermolysin[b]	additive model (a version of linear regression)	75
2001	Kazal protein inhibitors	association equilibrium constants	1146 constants = 191 single-point variants × 6 proteins for training and 398 constants for testing	sequence to reactivity algorithm (a version of linear regression)	76
2005	haloalkane dehalogenase	substrate specificity	116 halogenated compounds[b]	partial least-squares regression, principal component analysis	77
2007	halohydrin dehalogenase	function	∼600000 mutants based on various techniques with ∼280000 used by the ML-based method	partial least-squares regression	78
2010	toluene-4-monooxygenase	function	24 variants[b] (phase I) + 16 variants (phase II)	Gaussian random field	79
2014	various enzymes	binding affinity	1300 protein−ligand complexes from PDBbind[d]	random forest	80
2015	various enzymes	optimal pH range	217 enzymes from BRENDA database[d]	*K*-nearest neighbors, support vector machine, decision tree, artificial neural network, probabilistic neural network	81
2018	various enzymes	lysine malonylation sites	9760 experimentally validated malonylation sites: 1746 sites from 595 *E. coli* proteins, 3435 sites from 1 174 proteins in *M. musculus*, and 4579 sites from 1660 proteins in H. sapiens[d]	random forest, support vector machine, decision trees with gradient boosting, *K*-nearest neighbor, logistic regression	82
2018	glycosyltransferase superfamily 1	function	label-free mass spectroscopy-based assay data: 54 enzymes and 91 substrates[c]	decision trees	66
2018	various enzymes	EC class	63558 enzymes from RCSB PDB[d]	convolutional neural networks	83
2019	various proteins	solubility	chaperone-free reconstituted translation system: ∼500 cytosolic budding yeast proteins and ∼3000 *E. coli* proteins[d]	random forest	84
2019	nitric oxide dioxygenase	function	saturation mutagenesis: 445 variants in 3 stages	linear, kernel models, shallow neural networks, ensemble methods	71

[a]Data availability. [b]In the text. [c]In the supplement. [d]On the web server.

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

provide high-throughput data collection and more uniform sampling of possible combinations in quantities more suitable for ML.

3.3. Emerging Methods for High-Throughput Data Collection. Technological advances toward miniaturization, automation, and parallelization have brought efficient technologies to the novel generation of experimental research methods with incomparable throughput (Figure 2). *Next-generation sequencing (NGS)* has revolutionized genomic research, enabled access to fundamental molecular data, and revealed genomic and transcriptomic signatures.[47,48] The throughput of sequencing in the gigabase range per instrument run enables sequencing whole human genomes in as little as 1 day. The advent of this ultrahigh-throughput sequencing is propelling research that was considered impossible only a few years ago and is becoming widespread in many areas of life sciences and medical research.[49] Multiple commercially available second-generation instruments offer increased throughput and accuracy. Recently introduced third-generation (long-read) methods employing single-molecule real-time[50] or nanopore sequencing[51] resolve the limitations of short reads, such as GC bias or mapping to repetitive elements.

While the advanced sequencing technology provides a large amount of sequence data, for most of these entries, the structural and functional annotations are still missing. As the next step, the development of novel effective experimental methods is being focused on the collection of functional and structural information. *Liquid handling robotics* coupled with the miniaturization of the reaction chambers (up to 1536-well plates) has become a common technology for high-throughput screening of enzymatic reactions. By replacing wells with microcapillaries, further miniaturization and parallelization are possible (100000 capillaries per standard-sized plate).[52] Although wells and capillaries are conceptually simple, the assay throughput typically does not exceed 10^6 variants. Higher numbers can be analyzed if screening is moved from a solid support to fluids. *Fluorescence-activated cell sorting (FACS)* is a widely available technology enabling screening of up to 10^8 enzyme variants per day.[52,53] FACS requires fluorogenic substrates to be trapped inside or at the surface of the cell to link genotype and phenotype. Alternatively, sorting enzymes encapsulated together with their encoding DNA and a fluorogenic substrate in hydrogel beads is used.

A different approach to miniaturization relies on in vitro compartmentalization of libraries and reagents within surfactant-stabilized micrometer-sized droplets in emulsions.[52] The water−oil−water double emulsion droplets serve as the reaction chamber, which can be sorted by conventional FACS instruments. The utility of this approach was greatly expanded by microfluidic technologies. *Droplet-based microfluidics* enables the production of large numbers (10^8) of monodisperse droplets at very high rates by a continuous flow on a chip.[52,53] Sophisticated manipulations, such as droplet fusion, incubation, mixing, splitting, and sorting, are also possible. Droplet-based microfluidics has become a powerful tool combining the versatility of traditional microtiter plate screening with the high throughput achieved by FACS. The tiny volumes involved reduce the costs of screening a single clone by as much as million-fold in comparison to automated microtiter-plate screening.[52] Ultrahigh-throughput screening in microfluidic single water-in-oil droplets has emerged as a new tool with the potential to identify even very rare events from the large libraries (with 10^6−10^8 clones) at low cost. Screening of enzyme mutants in picoliter compartments, generated at a kilohertz speed in microfluidic devices, is coming of age.[54]

When they are coupled with the next-generation sequencing, high-throughput assays represent a powerful strategy for comprehensively analyzing sequence−function relationships in enzymes.[52,55] This approach, called *deep mutational scanning (DMS)*, links genotype to phenotype without the need for laborious processes involving protein purification and characterization. During the process, a large library of mutant sequences is synthesized, followed by selection for expressed phenotypes. Then sequencing the library before and after the selection quantifies the fitness of each mutant. DMS thus provides a rapid and facile method to infer sequence determinants of protein stability and function.[52,56,57] DMS has been employed as an alternative experimental strategy for the determination of protein fold. The pairs of sequence positions with strong positive epistasis are overwhelmingly close in 3D and can be systematically identified by mutation scans with sufficient coverage to determine high-resolution (1.8 Å) three-dimensional structure of a protein.[58,59] Still, several computational and experimental challenges must be addressed to generalize the use of genetic experiments for structure determination and application to larger proteins.

4. MACHINE LEARNING APPLICATIONS TO ENZYME ENGINEERING

4.1. The State of the Art in ML-Aided Biocatalyst Design. Despite being a relatively new field of study, machine learning for enzyme engineering has already been applied for several challenging predictions. In this section, we first consider predictors aimed at elucidating the structure−function relationships crucial for enzymes on both sides: predicting the structure for a known sequence, and predicting the catalytic activity or substrate specificity for a known sequence/structure. We then touch upon two other important properties, namely solubility and stability, especially from the point of view of amino acid substitutions, which is critical for successful protein engineering. We then present examples in another active area of application focused on ML-guided directed evolution. We conclude this section by providing a

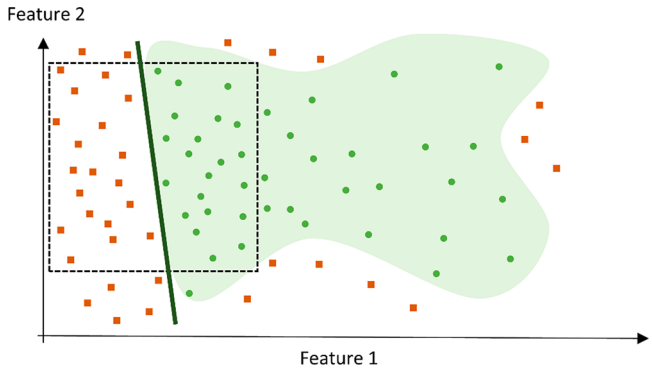


Figure 3. Comparison of decision boundaries for a hypothetical linear predictor with a more flexible nonlinear predictor. While the flexible predictor (shaded green area), which could be a neural network, is better at capturing the patterns in the whole feature space, the linear predictor (green line) also performs reasonably well, especially locally in the selected square area. In addition, the linear predictor provides a much more straightforward interpretation of the coefficients and easier guidelines in the design of new data points, at least locally.

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

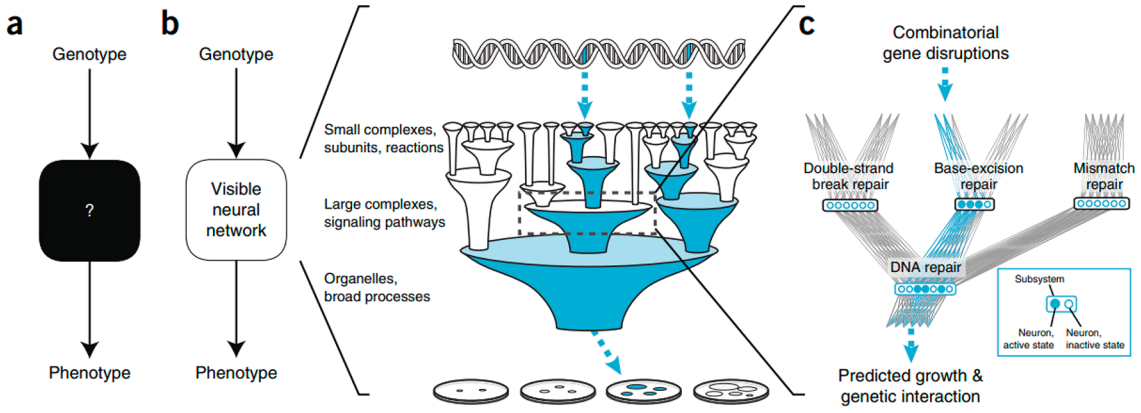


Figure 4. Visible neural networks to model the hierarchical structure and function of a living cell. (a) A conventional neural network translates the input to output as a black box without any knowledge of system structure. (b) In a visible neural network, input−output translation is based on prior knowledge. In DCell, gene-disruption genotypes (top) are translated to cell-growth predictions (bottom) through a hierarchy of cell subsystems (middle). (c) A neural network is embedded in the prior structure using multiple neurons per subsystem. Reproduced with permission from Ma and co-workers.[102] Copyright 2018, Springer Nature.

short historical excursion on the development of ML-based predictors for enzymes.

The *protein structure prediction* is arguably one of the longest-standing challenges in biochemistry, as the number of resolved structures is dramatically lagging behind the number of known sequences. Over 145000 structures have been released in the Protein Data Bank, but this is still nowhere near over 215 million publicly available protein sequences.[28] Nevertheless, even despite a relatively small data set size in comparison to millions of data points usually available for this method, deep neural networks showed most the notable results in the latest biennial assessment of protein structure prediction methods, CASP13. The AlphaFold network was trained on the PDB entries to predict the distances between C-beta atoms of residues using multiple sequence alignments[60] and received the highest score at the competition. Out of 124 targets, around two-thirds of AlphaFold predictions had a GDT_TS score above 50, which is indicative of a topologically correct structure.[61] Despite showing a tremendous improvement on the CASP12 results, this still indicates enough room for further improvement of protein structure predictors.

Apart from predicting protein structures, *predicting catalytic activities* is another active field of research currently. Computational methods for the protein function prediction range from sequence- to structure-based and from gene- to genome- and interactome-based.[62] Several initiatives similar to the CASP competition have already been proposed to address the functional annotation of enzymes, namely Enzyme Function Initiative (EFI), the Computational Bridges to Experiments initiative (COMBREX), and the Critical Assessment of Function Annotation community-driven experiment (CAFA). Certain successful attempts to apply ML to assign enzyme EC numbers using predicted 3D structures[63] or exploiting sequence similarities[64] have already been made. Recently, deep learning was also applied to predict EC numbers on the basis of a protein sequence using both sequence-length-dependent features, such as raw sequence one-hot encoding, and sequence-length-independent features, such as functional domain encoding.[65] The former type of features introduced nonuniformity in feature dimensionality, and the authors presented a framework to perform simultaneously dimensionality uniformization, feature selection, and classification model training. As the validation for their predictor, activities of three isoforms of glutaminase and five isoforms of Aurora kinases B were predicted in good correspondence with the experimental data available. Thus, the large data sets of enzyme structures and activities accumulated to date already allow using deep learning in the engineering of catalytic activity. Nevertheless, the problems with the data sets mentioned earlier are aggravated in the case of recording enzyme activity profiles due to both complex nomenclature and the abundance of possible mechanisms.

A more precise *functional prediction* is possible by restricting ML training to a particular family of enzymes, which comes at the cost of much smaller data sets available for training. This problem may be tackled by applying high-throughput data collection methods mentioned in section 3.3. The authors of the recently released GT-predict[66] selected for their analysis the glycosyltransferase superfamily 1, a group of enzymes with highly diverse substrates. This diversity, combined with the high scaffold conservation, increases the importance of subtle background mutations for the chemical function. Data from the label-free mass spectroscopy-based assay of 91 substrates and 54 enzymes derived from the plant *Arabidopsis thaliana* were used for functional prediction. The authors trained sequence-based decision trees, systematically varying combinations of physicochemical properties, e.g. log *P*, molecular area, and number/type of nucleophilic groups, and structural information, e.g. scaffold type and functional groups. The resulting predictor was successfully tested on four individually selected gene sequences as well as two complete families of enzymes from four different organisms, which highlights the tremendous potential of training ML predictors on the newly acquired data from high-throughput data collection methods. However, caution must be taken when extrapolating the results of this study to other families. It is yet to be seen if a strong predictor for one family will perform well when it is retrained on the data for another family.

Predictors of *protein solubility* usually exploit the eSol database (Table 1) for the entire ensemble of *Escherichia coli* proteins.[67] In their recent paper,[35] Han and coauthors considered seven different binary and continuous ML algorithms: logistic regression, decision tree, support vector machines, Naive Bayes, conditional random forest, XGboost, and artificial neural networks. The support vector machine showed the highest accuracy based on 10-fold cross-validation.

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

Notably, the authors attempted to use generative adversarial networks to synthesize more data. This is a pair of two neural networks competing against each other: one learns to generate artificial examples and the other to distinguish them from real data. However, due to data scarcity, no independent test set was used to evaluate the resulting predictor, implying there is a strong demand for more abundant data sets of protein solubility. Moreover, a modest best-achieved R^2 value of around 0.4 indicates that there is still ample room for designing a more reliable continuous predictor of solubility scores.

Another point of view on protein solubility prediction is studying the effects of *individual mutations*. The recent successes in the application of deep mutational scanning to collect the data on protein solubility changes upon mutations[68] are likely to promote the development of more sophisticated ML-based protein solubility predictors in the nearest future. Predicting the effects of amino acid substitutions is not only limited to solubility: stability, substrate specificity, catalytic activity, and enantioselectivity can also be targeted if sufficient data are available. *Protein stability predictors* are perhaps those with the most abundant data sets of this type available for ML training (Table 1). The recently released PON-tstab[34] stands out due to the impressive work the authors undertook to identify major issues with the widely used ProTherm database. The authors also presented a random forest classifier trained using 1106 features from the following groups: experimental conditions, conservation and coevolution scores for mutated positions, amino acid substitutions and their physicochemical properties, neighborhood features for 11 positions before and after substitution sites, and thermodynamic sequence-based features extracted from ProtDCal.[69] PON-tstab is a three-class predictor (stability increasing, decreasing, unchanged) and achieved the correct prediction ratio of around 0.5 versus the value 0.33 for a random predictor. This implies that, even with a data set of higher quality, predicting protein stability remains an extremely challenging task.[5]

Another intriguing application of ML in protein engineering is to design smart combinatorial libraries for *directed protein evolution*.[70] This has the potential to both reduce the experimental effort and improve the exploration of the sequence space by mutating multiple positions simultaneously. Moreover, it can approximate the empirical fitness landscape to suggest a refined set of variants for the next round of screening. Wu et al.[71] used ML-assisted directed evolution to engineer an enzyme for a new stereodivergent carbon−silicon bond formation. The authors selected the reaction of phenyldimethyl silane with ethyl 2-diazopropanoate catalyzed by a putative nitric oxide dioxygenase from *Rhodothermus marinus*. They tested a variety of ML algorithms such as linear and kernel models, shallow neural networks, and ensemble methods to improve the enzyme enantioselectivity. The starting enantiomeric excess (ee) of 76% for the *S* enantiomer was sequentially improved to 93% by several rounds of ML-guided evolution experiments, and a new variant with 79% ee for the *R* enantiomer was discovered. The authors also compared two standard directed evolution approaches with the one assisted by shallow neural networks. They used 149361 previously published measurements of a total of 160000 variants from saturation mutagenesis of protein G domain B1 at four positions. The ML-guided approach yielded a global optimum twice as often with a 30% reduction in the number of variants tested.

A historical perspective on the applications of ML to enzyme engineering is presented in Table 2. The linear regression and its variants were often used in the first attempts to obtain data-driven guidance, whereas lately there is a tendency to apply artificial neural networks and random forests, in part owing to the increase in data availability and improving high-throughput data collection methods (see section 3.3).

4.2. Current Challenges Related to ML-Aided Methods. One of the main challenges in applications of ML to enzyme engineering stems from the intrinsic multidisciplinarity of the approach. Biochemists, molecular biologists, mathematicians, and computer scientists have to find a common language to clarify goals, carry out rigorous analysis and training, and avoid common pitfalls, wrongful usage of methods, and misinterpretations. Ready to use software packages certainly help standardize the training of an ML algorithm for nonspecialists, but heaping all the available data and running a range of ML algorithms to select the best predictor might not be the optimal strategy. The No Free Lunch theorem[85] claims that no single ML method is superior to others a priori;[86] therefore, a thorough understanding of the data types to be used and problems to be solved is essential in the development of efficient predictors. The current shift toward new and more complex ML methods, namely aggregating several algorithms into hybrid meta-predictors, hyperparameter optimization with many training subcycles, feature learning, and the fusion of ML-based and classical bioinformatics tools in a single predictor, will further challenge the crosstalk between disciplines necessary for the development of efficient and robust predictors in enzyme engineering.

With the continuous growth of ML applications in enzyme engineering, the need for robust comparison of various predictors is of growing importance. This comparison is mainly obstructed by the lack of both standardized protocols for comparison and new data sets for testing. The lack of benchmark data sets, discrepancies in the performance measurements used, inaccurate or insufficient disclosure in publications, and the difficulty in finding reviewers with sufficiently broad expertise[87] are among the most pressing issues. Researchers working on some applications with a long track record in bioinformatics, such as protein structure or function predictions, have already established several platforms that can be used for comparison of the ML predictors, i.e. CASP, CAFA, EFI, and COMBREX mentioned in section 4.1. Other applications have yet to see similar initiatives as, in our opinion, at least three key ingredients are necessary: (i) a sufficiently large community of researchers working on development of such applications, (ii) a sufficient amount of new high-quality data being collected regularly, and (iii) a leader that will take on responsibility and invest time and effort into coordinating this activity. It is also worth noting that competitions of this kind are not flawless themselves, as their appearance led to an unwanted side effect: greater secrecy and an increased time delay before publishing newly developed methods due to the competition deadlines, which negatively affects the speed of knowledge circulation in science. Moreover, while their participation is welcome, industrial participants often have a competitive advantage, i.e. access to private data, and are often not required to make their codes public.

Finally, the excitement about novel applications of ML to enzyme engineering seems to put another critical component of the approach on the back burner. The ultimate goal of

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

science is not only to achieve better predictive power but also to be able to explain the results. Few papers go beyond simple ROC analysis: e.g., resample cross-validation to estimate its statistical significance, explore the reasons for weak predictions, and analyze learning curves. Why does a particular predictor have a better performance? What features are critical for the performance of a predictor on a global scale? What ranges for feature values and what parts of the feature space are most critical for a particular data point to be classified correctly? Many articles on the topic lack this kind of analysis, which limits our understanding of the underlying molecular principles. In the next section, we touch upon modern trends in the ML workflow and architecture and also discuss how interpretable and explainable predictors can possibly provide some answers to the questions above.

4.3. Emerging Trends in ML-Based Methods for Enzyme Engineering. With the accumulation of more data by virtue of the emerging high-throughput experimental methods, the development of benchmark data sets and unified performance measurements is only a matter of time. Recently, an intriguing algorithm based on semisupervised learning has been presented to allow benchmarking in five different prediction tasks related to protein engineering, including secondary structure, fluorescence landscape, and stability landscape predictions.[88] Moreover, as the data generation is streamlined, a data set from a single experiment is starting to have the size large enough for training ML algorithms to guide the design of future experiments, as was the case in the development of stereodivergent carbon−silicon bond formation[71] and the application of Gaussian processes to the directed evolution of cytochromes.[89]

The increase in the available data will prompt more extensive use of deep neural networks. This approach has already shown remarkable potential for complex tasks in genomics and proteomics but still has limited usage in enzyme engineering due to data scarcity. Sophisticated neural network architectures, such as recurrent or graph-based neural networks, simultaneous training of several types of predictors (multitasking), combining structurally different input data (multimodal design), ML-based modeling of data sets (generative models), and retraining predictors used in one area by new data from another area (transfer learning) have only recently been applied in genomics.[14] Several exciting attempts have also recently been made to apply some of those advanced techniques to proteins: using generative models to create soluble and functional malate dehydrogenase variants[90] or predict mutational effects with high correlation with those actually observed in 42 high-throughput deep mutational scanning experiments.[91] More data will also allow improving the existing methods, i.e. learning the optimal architecture of a predictor from the data (hyperparameter optimization),[92] smart aggregation of several predictions from multiple methods,[93] and introducing robust confidence scores for predictions.[94] In enzyme engineering, this new level of algorithmic complexity will further save time and resources wasted on validating misleading predictions but will also require more sophisticated computer architecture, e.g. an increased use of parallel computing and stochastic training methods, which have already become standard techniques for the acceleration of deep neural network training.

With the increase in computational power, the incorporation of molecular dynamics simulations into ML training will allow accounting for the dynamics of enzyme molecules in contrast to predominantly using static features currently. This should further boost predictive power, since the propensity for catalysis critically depends on the conformational dynamics and the kinetics of the underlying processes. We also envisage the combination of ML models with fundamentally different types of predictors. The development of hybrid methods became very successful, for example, in the prediction of protein stability.[5] Moreover, models targeting several properties of biocatalyst simultaneously, e.g. activity, stability, and solubility, would dramatically reduce the risk of unsuccessful laboratory experiments resulting from in silico design of active but unstable or poorly soluble proteins.

Another noticeable trend in ML is toward interpretable and explainable predictors.[95] Apart from the global importance of features for ML predictors, feature importance scores calculated for each input example[96,97] may help explain why a particular prediction was made for each input data point. In addition to providing mechanistic insights, interpretable algorithms can aid in smart biocatalyst design. For instance, instead of simply screening all the possible mutations with an ML-based tool to improve a target property, researchers can make use of designing variants on the basis of the structure of a predictor using adaptive sampling.[98] Such an approach favors predictors whose parameters can provide such guidance: e.g., linear predictors over more flexible yet harder to interpret artificial neural networks (Figure 3). Linear predictors allow analytical design on the basis of the coefficients;[99] in contrast, sophisticated predictors are usually prone to pathological behavior, i.e. sudden misclassification after a slight and almost imperceptible perturbation of input.[100]

Another promising approach is to use interpretable architectures of predictors already at the design stage, e.g. the visible neural networks.[101] The design of such networks is guided by the knowledge of the underlying biological mechanism, e.g. the choice of layers and the connections between layers may mimic the hierarchical organization of transcriptional regulatory factors in the cell nucleus. For instance, the recently released DCell simulates cellular growth and allows in silico investigations of the molecular mechanisms underlying genotype−phenotype associations on the basis of the analysis of different parts of the neural network (Figure 4). Since enzymes can also be represented hierarchically on the basis of the annotation of their secondary, tertiary, and quaternary structure elements and their interactions, a shift toward applying interpretable visible neural networks may provide new insights into the mechanism in addition to better predictors.

Finally, as the field will be getting more accustomed to ML tools, more stringent requirements for data collection and transparent application of statistical methods are to be expected. This is further encouraged by the pressure from publishers and grant agencies to make scripts and data sets publicly available. The next logical step will be the creation of platforms for rapid exchange and validation of models and their penetration to user communities.

5. SUMMARY

Here we considered recent advancements of ML in enzyme engineering. The range of possible applications is extensive: from predicting protein structures, through improving enzyme stability, solubility, and functional properties, to guiding through the vast expanse of combinatorial libraries in directed evolution experiments. Several databases with millions of

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

protein sequences, hundreds of thousands of structures, thousands of biophysical values, and hundreds of well-annotated catalytic mechanisms already offer practicable means for training ML-based predictors. Yet the ML potential in biocatalyst design is far from being fully explored. The community has still many challenges to face. The lack of homogeneous and consistent data sets of high quality for training and validation, classical data imbalances and biases, intrinsic multidisciplinarity of the approach, and difficulties in explaining, interpreting, and comparing the results of predictors are among some of the most pressing issues today. Those are now being increasingly appreciated and addressed due to growing demands and the increasing number of scientists working in this exciting new domain of enzyme engineering. Some powerful recent experimental techniques, namely next-generation sequencing, high-throughput screening, deep mutational scanning, and microfluidics, already allow collecting data in larger amounts and of better quality and consistency. As more data are collected, more advanced ML methods, such as deep learning, with more involved implementation will take over, necessitating efficient use of computing power and memory allocation. The recent developments in interpretable architectures of artificial neural networks and feature importance scores may provide insights into the internal principles leading to better prediction. Reliable ML tools will provide the best possible starting points for enzyme engineering. They will also create further research opportunities for explaining derived models, interpreting their parameters, and understanding underlying molecular mechanisms, eventually leading to a clearer perception of structure–function relationships of enzymes.

AUTHOR INFORMATION

Corresponding Author

*E-mail for S.M.: stan.mazurenko@gmail.com.

ORCID

Stanislav Mazurenko: 0000-0003-3659-4819

Zbynek Prokop: 0000-0001-9358-4081

Jiri Damborsky: 0000-0002-7848-8216

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by grants from the Czech Ministry of Education (LM2015051, LM2015047, LM2015055, CZ.02.1.01/0.0/0.0/16_026/0008451, CZ.02.1.01/0.0/0.0/16_019/0000868 and CZ.02.1.01/0.0/0.0/16_013/0001761), Technology Agency of Czech Republic (TN01000013), and the European Union (720776 and 814418). S.M. is supported by the Operational Programme Research, Development and Education project MSCAfellow@MUNI (CZ.02.2.69/0.0/0.0/17_050/0008496).

REFERENCES

(1) Bornscheuer, U. T.; Hauer, B.; Jaeger, K. E.; Schwaneberg, U. Directed Evolution Empowered Redesign of Natural Proteins for the Sustainable Production of Chemicals and Pharmaceuticals. *Angew. Chem., Int. Ed.* **2019**, *58*, 36–40.

(2) Arnold, F. H. Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). *Angew. Chem., Int. Ed.* **2019**, *58*, 14420–14426.

(3) Sheldon, R. A.; Pereira, P. C. Biocatalysis Engineering: The Big Picture. *Chem. Soc. Rev.* **2017**, *46*, 2678–2691.

(4) Qu, G.; Li, A.; Sun, Z.; Acevedo-Rocha, C. G.; Reetz, M. T. The Crucial Role of Methodology Development in Directed Evolution of Selective Enzymes. *Angew. Chem., Int. Ed.* **2019**, DOI: 10.1002/anie.201901491.

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

(6) Romero-Rivera, A.; Garcia-Borràs, M.; Osuna, S. Computational Tools for the Evaluation of Laboratory-Engineered Biocatalysts. *Chem. Commun.* **2017**, *53*, 284–297.

(7) Arnold, F. H. The Nature of Chemical Innovation: New Enzymes by Evolution. *Q. Rev. Biophys.* **2015**, *48*, 404–410.

(8) Currin, A.; Swainston, N.; Day, P. J.; Kell, D. B. Synthetic Biology for the Directed Evolution of Protein Biocatalysts: Navigating Sequence Space Intelligently. *Chem. Soc. Rev.* **2015**, *44*, 1172–1239.

(9) Kaushik, S.; Marques, S. M.; Khirsariya, P.; Paruch, K.; Libichova, L.; Brezovsky, J.; Prokop, Z.; Chaloupkova, R.; Damborsky, J. Impact of the Access Tunnel Engineering on Catalysis Is Strictly Ligand-Specific. *FEBS J.* **2018**, *285*, 1456–1476.

(10) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-Learning-Guided Directed Evolution for Protein Engineering. *Nat. Methods* **2019**, *16*, 687–694.

(11) Domingos, P. M. A Few Useful Things to Know about Machine Learning. *Commun. ACM* **2012**, *55*, 78–87.

(12) Libbrecht, M. W.; Noble, W. S. Machine Learning Applications in Genetics and Genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332.

(13) Park, Y.; Kellis, M. Deep Learning for Regulatory Genomics. *Nat. Biotechnol.* **2015**, *33*, 825–826.

(14) Eraslan, G.; Avsec, Ž; Gagneur, J.; Theis, F. J. Deep Learning: New Computational Modelling Techniques for Genomics. *Nat. Rev. Genet.* **2019**, *20*, 389–403.

(15) Yang, Y.; Niroula, A.; Shen, B.; Vihinen, M. PON-Sol: Prediction of Effects of Amino Acid Substitutions on Protein Solubility. *Bioinformatics* **2016**, *32*, 2032–2034.

(16) Folkman, L.; Stantic, B.; Sattar, A.; Zhou, Y. EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. *J. Mol. Biol.* **2016**, *428*, 1394–1405.

(17) Teng, S.; Srivastava, A. K.; Wang, L. Sequence Feature-Based Prediction of Protein Stability Changes upon Amino Acid Substitutions. *BMC Genomics* **2010**, *11*, S5.

(18) Huang, L.; Gromiha, M. M.; Ho, S. iPTREE-STAB: Interpretable Decision Tree Based Method for Predicting Protein Stability Changes Upon Mutations. *Bioinformatics* **2007**, *23*, 1292–1293.

(19) Koskinen, P.; Törönen, P.; Nokso-Koivisto, J.; Holm, L. PANNZER: High-Throughput Functional Annotation of Uncharacterized Proteins in an Error-Prone Environment. *Bioinformatics* **2015**, *31*, 1544–1552.

(20) De Ferrari, L.; Mitchell, J. B. From Sequence to Enzyme Mechanism Using Multi-Label Machine Learning. *BMC Bioinf.* **2014**, *15*, 150.

(21) Falda, M.; Toppo, S.; Pescarolo, A.; Lavezzo, E.; Di Camillo, B.; Facchinetti, A.; Cilia, E.; Velasco, R.; Fontana, P. Argot2: A Large Scale Function Prediction Tool Relying on Semantic Similarity of Weighted Gene Ontology Terms. *BMC Bioinf.* **2012**, *13*, S14.

(22) Cozzetto, D.; Buchan, D. W.; Bryson, K.; Jones, D. T. Protein Function Prediction by Massive Integration of Evolutionary Analyses and Multiple Data Sources. *BMC Bioinf.* **2013**, *14*, S1.

(23) Kim, G. B.; Kim, W. J.; Kim, H. U.; Lee, S. Y. Machine Learning Applications in Systems Metabolic Engineering. *Curr. Opin. Biotechnol.* **2020**, *64*, 1–9.

(24) Woodley, J. M. Accelerating the Implementation of Biocatalysis in Industry. *Appl. Microbiol. Biotechnol.* **2019**, *103*, 4733–4739.

(25) HamediRad, M.; Chao, R.; Weisberg, S.; Lian, J.; Sinha, S.; Zhao, H. Towards a Fully Automated Algorithm Driven Platform for Biosystems Design. *Nat. Commun.* **2019**, *10*, 1–10.

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210–1223

(26) Niroula, A.; Vihinen, M. Variation Interpretation Predictors: Principles, Types, Performance, and Choice. *Hum. Mutat.* **2016**, *37*, 579−597.
(27) Mitchell, A. L.; Attwood, T. K.; Babbitt, P. C.; Blum, M.; Bork, P.; Bridge, A.; Brown, S. D.; Chang, H.; El-Gebali, S.; Fraser, M. I. InterPro in 2019: Improving Coverage, Classification and Access to Protein Sequence Annotations. *Nucleic Acids Res.* **2018**, *47*, D351−D360.
(28) UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2018**, *47*, D506−D515.
(29) Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J. M.; Dutta, S. RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy. *Nucleic Acids Res.* **2018**, *47*, D464−D474.
(30) Jeske, L.; Placzek, S.; Schomburg, I.; Chang, A.; Schomburg, D. BRENDA in 2019: A European ELIXIR Core Data Resource. *Nucleic Acids Res.* **2019**, *47*, D542−D549.
(31) Nagano, N.; Nakayama, N.; Ikeda, K.; Fukuie, M.; Yokota, K.; Doi, T.; Kato, T.; Tomii, K. EzCatDB: The Enzyme Reaction Database, 2015 Update. *Nucleic Acids Res.* **2015**, *43*, D453−D458.
(32) Ribeiro, A. J. M.; Holliday, G. L.; Furnham, N.; Tyzack, J. D.; Ferris, K.; Thornton, J. M. Mechanism and Catalytic Site Atlas (M-CSA): A Database of Enzyme Reaction Mechanisms and Active Sites. *Nucleic Acids Res.* **2018**, *46*, D618−D623.
(33) Montanucci, L.; Martelli, P. L.; Ben-Tal, N.; Fariselli, P. A Natural Upper Bound to the Accuracy of Predicting Protein Stability Changes upon Mutations. *Bioinformatics* **2019**, *35*, 1513−1517.
(34) Yang, Y.; Urolagin, S.; Niroula, A.; Ding, X.; Shen, B.; Vihinen, M. PON-tstab: Protein Variant Stability Predictor. Importance of Training Data Quality. *Int. J. Mol. Sci.* **2018**, *19*, 1009.
(35) Han, X.; Wang, X.; Zhou, K. Develop Machine Learning-Based Regression Predictive Models for Engineering Protein Solubility. *Bioinformatics* **2019**, *35*, 4640−4646.
(36) Berman, H. M.; Gabanyi, M. J.; Kouranov, A.; Micallef, D. I.; Westbrook, J.; Protein Structure Initiative network of investigators Protein Structure Initiative - TargetTrack 2000−2017 - all data. *Zenodo* **2017**.1
(37) Wang, C. Y.; Chang, P. M.; Ary, M. L.; Allen, B. D.; Chica, R. A.; Mayo, S. L.; Olafson, B. D. ProtaBank: A Repository for Protein Design and Engineering Data. *Protein Sci.* **2018**, *27*, 1113−1124.
(38) Bastian, F. B.; Chibucos, M. C.; Gaudet, P.; Giglio, M.; Holliday, G. L.; Huang, H.; Lewis, S. E.; Niknejad, A.; Orchard, S.; Poux, S. The Confidence Information Ontology: A Step towards a Standard for Asserting Confidence in Annotations. *Database* **2015**, *2015*, bav043.
(39) Holliday, G. L.; Bairoch, A.; Bagos, P. G.; Chatonnet, A.; Craik, D. J.; Finn, R. D.; Henrissat, B.; Landsman, D.; Manning, G.; Nagano, N. Key Challenges for the Creation and Maintenance of Specialist Protein Resources. *Proteins: Struct., Funct., Genet.* **2015**, *83*, 1005−1013.
(40) Wickham, H. Tidy Data. *J. Stat. Softw.* **2014**, *59*, 1−23.
(41) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J. W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S. A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018.
(42) Tipton, K. F.; Armstrong, R. N.; Bakker, B. M.; Bairoch, A.; Cornish-Bowden, A.; Halling, P. J.; Hofmeyr, J.; Leyh, T. S.; Kettner, C.; Raushel, F. M. Standards for Reporting Enzyme Data: The STRENDA Consortium: What It Aims to Do and Why It Should Be Helpful. *Perspect. Sci.* **2014**, *1*, 131−137.
(43) Pucci, F.; Bernaerts, K.; Teheux, F.; Gilis, D.; Rooman, M. Symmetry Principles in Optimization Problems: An Application to Protein Stability Prediction. *IFAC-PapersOnLine* **2015**, *48*, 458−463.
(44) Pucci, F.; Bernaerts, K. V.; Kwasigroch, J. M.; Rooman, M. Quantification of Biases in Predictions of Protein Stability Changes upon Mutations. *Bioinformatics* **2018**, *34*, 3659−3665.
(45) Schnoes, A. M.; Ream, D. C.; Thorman, A. W.; Babbitt, P. C.; Friedberg, I. Biases in the Experimental Annotations of Protein Function and Their Effect on Our Understanding of Protein Function Space. *PLoS Comput. Biol.* **2013**, *9*, e1003063.
(46) Fu, G.; Wang, J.; Yang, B.; Yu, G. NegGOA: Negative GO Annotations Selection Using Ontology Structure. *Bioinformatics* **2016**, *32*, 2996−3004.
(47) Kulski, J. *Next Generation Sequencing: Advances, Applications and Challenges*; InTechOpen: London, 2016.
(48) Straiton, J.; Free, T.; Sawyer, A.; Martin, J. From Sanger Sequencing to Genome Databases and Beyond. *BioTechniques* **2019**, *66*, 60−63.
(49) Goodwin, S.; McPherson, J. D.; McCombie, W. R. Coming of Age: Ten Years of Next-Generation Sequencing Technologies. *Nat. Rev. Genet.* **2016**, *17*, 333−351.
(50) Ardui, S.; Ameur, A.; Vermeesch, J. R.; Hestand, M. S. Single Molecule Real-Time (SMRT) Sequencing Comes of Age: Applications and Utilities for Medical Diagnostics. *Nucleic Acids Res.* **2018**, *46*, 2159−2168.
(51) Kono, N.; Arakawa, K. Nanopore Sequencing: Review of Potential Applications in Functional Genomics. *Dev., Growth Differ.* **2019**, *61*, 316−326.
(52) Bunzel, H. A.; Garrabou, X.; Pott, M.; Hilvert, D. Speeding Up Enzyme Discovery and Engineering with Ultrahigh-Throughput Methods. *Curr. Opin. Struct. Biol.* **2018**, *48*, 149−156.
(53) Jacques, P.; Béchet, M.; Bigan, M.; Caly, D.; Chataigné, G.; Coutte, F.; Flahaut, C.; Heuson, E.; Leclère, V.; Lecouturier, D. High-Throughput Strategies for the Discovery and Engineering of Enzymes for Biocatalysis. *Bioprocess Biosyst. Eng.* **2017**, *40*, 161−180.
(54) Mair, P.; Gielen, F.; Hollfelder, F. Exploring Sequence Space in Search of Functional Enzymes Using Microfluidic Droplets. *Curr. Opin. Chem. Biol.* **2017**, *37*, 137−144.
(55) Wrenbeck, E. E.; Faber, M. S.; Whitehead, T. A. Deep Sequencing Methods for Protein Engineering and Design. *Curr. Opin. Struct. Biol.* **2017**, *45*, 36−44.
(56) Fowler, D. M.; Fields, S. Deep Mutational Scanning: A New Style of Protein Science. *Nat. Methods* **2014**, *11*, 801−807.
(57) Gupta, K.; Varadarajan, R. Insights into Protein Structure, Stability and Function from Saturation Mutagenesis. *Curr. Opin. Struct. Biol.* **2018**, *50*, 117−125.
(58) Schmiedel, J. M.; Lehner, B. Determining Protein Structures Using Deep Mutagenesis. *Nat. Genet.* **2019**, *51*, 1177−1186.
(59) Rollins, N. J.; Brock, K. P.; Poelwijk, F. J.; Stiffler, M. A.; Gauthier, N. P.; Sander, C.; Marks, D. S. Inferring Protein 3D Structure from Deep Mutation Scans. *Nat. Genet.* **2019**, *51*, 1170−1176.
(60) Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D.; Senior, A. W. De Novo Structure Prediction with Deeplearning Based Scoring. In *Thirteenth Critical Assessment of Techniques for Protein Structure Prediction Abstracts*; 2018; pp 11−12.
(61) Kinch, L. N.; Shi, S.; Cheng, H.; Cong, Q.; Pei, J.; Mariani, V.; Schwede, T.; Grishin, N. V. CASP9 Target Classification. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 21−36.
(62) Shehu, A.; Barbará, D.; Molloy, K. A Survey of Computational Methods for Protein Function Prediction. In *Big Data Analytics in Genomics*; Wong, K. C., Ed.; Springer: Cham, 2016; pp 225−298.
(63) Zhang, C.; Freddolino, P. L.; Zhang, Y. COFACTOR: Improved Protein Function Prediction by Combining Structure,

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

Sequence and Protein−Protein Interaction Information. *Nucleic Acids Res.* **2017**, *45*, W291−W299.
(64) Kumar, N.; Skolnick, J. EFICAz2. 5: Application of a High-Precision Enzyme Function Predictor to 396 Proteomes. *Bioinformatics* **2012**, *28*, 2687−2688.
(65) Li, Y.; Wang, S.; Umarov, R.; Xie, B.; Fan, M.; Li, L.; Gao, X. DEEPre: Sequence-Based Enzyme EC Number Prediction by Deep Learning. *Bioinformatics* **2018**, *34*, 760−769.
(66) Yang, M.; Fehl, C.; Lees, K. V.; Lim, E. K.; Offen, W. A.; Davies, G. J.; Bowles, D. J.; Davidson, M. G.; Roberts, S. J.; Davis, B. G. Functional and Informatics Analysis Enables Glycosyltransferase Activity Prediction. *Nat. Chem. Biol.* **2018**, *14*, 1109−1117.
(67) Niwa, T.; Ying, B. W.; Saito, K.; Jin, W.; Takada, S.; Ueda, T.; Taguchi, H. Bimodal Protein Solubility Distribution Revealed by an Aggregation Analysis of the Entire Ensemble of Escherichia Coli Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 4201−4206.
(68) Klesmith, J. R.; Bacik, J. P.; Wrenbeck, E. E.; Michalczyk, R.; Whitehead, T. A. Trade-Offs Between Enzyme Fitness and Solubility Illuminated by Deep Mutational Scanning. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 2265−2270.
(69) Ruiz-Blanco, Y. B.; Paz, W.; Green, J.; Marrero-Ponce, Y. ProtDCal: A Program to Compute General-Purpose-Numerical Descriptors for Sequences and 3D-Structures of Proteins. *BMC Bioinf.* **2015**, *16*, 162.
(70) Li, G.; Dong, Y.; Reetz, M. T. Can Machine Learning Revolutionize Directed Evolution of Selective Enzymes? *Adv. Synth. Catal.* **2019**, *361*, 2377−2386.
(71) Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine Learning-Assisted Directed Protein Evolution with Combinatorial Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 8852−8858.
(72) Damborsky, J. Quantitative Structure-Function Relationships of the Single-Point Mutants of Haloalkane Dehalogenase: A Multivariate Approach. *Quant. Struct.-Act. Relat.* **1997**, *16*, 126−135.
(73) Damborsky, J. Quantitative Structure-Function and Structure-Stability Relationships of Purposely Modified Proteins. *Protein Eng., Des. Sel.* **1998**, *11*, 21−30.
(74) Bucht, G.; Wikström, P.; Hjalmarsson, K. Optimising the Signal Peptide for Glycosyl Phosphatidylinositol Modification of Human Acetylcholinesterase Using Mutational Analysis and Peptide-Quantitative Structure−Activity Relationships. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.* **1999**, *1431*, 471−482.
(75) Aita, T.; Uchiyama, H.; Inaoka, T.; Nakajima, M.; Kokubo, T.; Husimi, Y. Analysis of a Local Fitness Landscape with a Model of the Rough Mt. Fuji-Type Landscape: Application to Prolyl Endopeptidase and Thermolysin. *Biopolymers* **2000**, *54*, 64−79.
(76) Lu, S. M.; Lu, W.; Qasim, M. A.; Anderson, S.; Apostol, I.; Ardelt, W.; Bigler, T.; Chiang, Y. W.; Cook, J.; James, M. N.; Kato, I.; Kelly, C.; Kohr, W.; Komiyama, T.; Lin, T. Y.; Ogawa, M.; Otlewski, J.; Park, S. J.; Qasim, S.; Ranjbar, M.; Tashiro, M.; Warne, N.; Whatley, H.; Wieczorek, A.; Wieczorek, M.; Wilusz, T.; Wynn, R.; Zhang, W.; Laskowski, M., Jr Predicting the Reactivity of Proteins from Their Sequence Alone: Kazal Family of Protein Inhibitors of Serine Proteinases. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 1410−1415.
(77) Kmunícek, J.; Hynková, K.; Jedlicka, T.; Nagata, Y.; Negri, A.; Gago, F.; Wade, R. C.; Damborský, J. Quantitative Analysis of Substrate Specificity of Haloalkane Dehalogenase LinB from Sphingomonas Paucimobilis UT26. *Biochemistry* **2005**, *44*, 3390−3401.
(78) Fox, R. J.; Davis, S. C.; Mundorff, E. C.; Newman, L. M.; Gavrilovic, V.; Ma, S. K.; Chung, L. M.; Ching, C.; Tam, S.; Muley, S. Improving Catalytic Function by ProSAR-Driven Enzyme Evolution. *Nat. Biotechnol.* **2007**, *25*, 338−344.
(79) Brouk, M.; Nov, Y.; Fishman, A. Improving Biocatalyst Performance by Integrating Statistical Methods into Protein Engineering. *Appl. Environ. Microbiol.* **2010**, *76*, 6397−6403.
(80) Ballester, P. J.; Schreyer, A.; Blundell, T. L. Does a More Precise Chemical Description of Protein−Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *J. Chem. Inf. Model.* **2014**, *54*, 944−955.
(81) Khan, Z. U.; Hayat, M.; Khan, M. A. Discrimination of Acidic and Alkaline Enzyme Using Chou's Pseudo Amino Acid Composition in Conjunction with Probabilistic Neural Network Model. *J. Theor. Biol.* **2015**, *365*, 197−203.
(82) Zhang, Y.; Xie, R.; Wang, J.; Leier, A.; Marquez-Lago, T. T.; Akutsu, T.; Webb, G. I.; Chou, K.; Song, J. Computational Analysis and Prediction of Lysine Malonylation Sites by Exploiting Informative Features in an Integrative Machine-Learning Framework. *Briefings Bioinf.* **2018**, *5*, bby079.
(83) Amidi, A.; Amidi, S.; Vlachakis, D.; Megalooikonomou, V.; Paragios, N.; Zacharaki, E. I. EnzyNet: Enzyme Classification Using 3D Convolutional Neural Networks on Spatial Representation. *PeerJ* **2018**, *6*, e4750.
(84) Hou, Q.; Kwasigroch, J. M.; Rooman, M.; Pucci, F. Solart: A Structure-Based Method to Predict Protein Solubility and Aggregation. *Bioinformatics* **2019**, btz773.
(85) Wolpert, D. H.; Macready, W. G. No Free Lunch Theorems for Optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67−82.
(86) Wolpert, D. H. The Lack of a Priori Distinctions between Learning Algorithms. *Neural Comput.* **1996**, *8*, 1341−1390.
(87) Walsh, I.; Pollastri, G.; Tosatto, S. C. Correct Machine Learning on Protein Sequences: A Peer-Reviewing Perspective. *Briefings Bioinf.* **2016**, *17*, 831−840.
(88) Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, X.; Canny, J.; Abbeel, P.; Song, Y. S. Evaluating Protein Transfer Learning with TAPE. *arXiv preprint arXiv:1906.08230*, 2019.
(89) Romero, P. A.; Krause, A.; Arnold, F. H. Navigating the Protein Fitness Landscape with Gaussian Processes. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E193−E201.
(90) Repecka, D.; Jauniskis, V.; Karpus, L.; Rembeza, E.; Zrimec, J.; Poviloniene, S.; Rokaitis, I.; Laurynenas, A.; Abuajwa, W.; Savolainen, O.; Meskys, R.; Engqvist, M. K. M.; Zelezniak, A. Expanding Functional Protein Sequence Space Using Generative Adversarial Networks. *bioRxiv* **2019**, DOI: 10.1101/789719.
(91) Riesselman, A. J.; Ingraham, J. B.; Marks, D. S. Deep Generative Models of Genetic Variation Capture the Effects of Mutations. *Nat. Methods* **2018**, *15*, 816−822.
(92) Thornton, C.; Hutter, F.; Hoos, H. H.; Leyton-Brown, K. Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*; 2013; pp 847−855.
(93) Polikar, R. Ensemble based systems in decision making. *IEEE Circuits and systems magazine* **2006**, *6*, 21−45.
(94) Gammerman, A.; Vovk, V. Hedging Predictions in Machine Learning. *Comput. J.* **2007**, *50*, 151−163.
(95) Samek, W.; Wiegand, T.; Müller, K. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITU Journal: ICT Discoveries* **2017**, 39−48.
(96) Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning*; 2017; Vol. *70*, pp 3145−3153.
(97) Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv preprint arXiv:1312.6034* 2013.
(98) Brookes, D. H.; Park, H.; Listgarten, J. Conditioning by Adaptive Sampling for Robust Design. In *Proceedings of the 36th International Conference on Machine Learning*; 2019; Vol. *97*, pp 773−782.
(99) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*; 2016; pp 1135−1144.
(100) Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing Properties of Neural Networks. *arXiv preprint arXiv:1312.6199* 2013.

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223

(101) Yu, M. K.; Ma, J.; Fisher, J.; Kreisberg, J. F.; Raphael, B. J.; Ideker, T. Visible Machine Learning for Biomedicine. *Cell* **2018**, *173*, 1562−1565.
(102) Ma, J.; Yu, M. K.; Fong, S.; Ono, K.; Sage, E.; Demchak, B.; Sharan, R.; Ideker, T. Using Deep Learning to Model the Hierarchical Structure and Function of a Cell. *Nat. Methods* **2018**, *15*, 290−298.

DOI: 10.1021/acscatal.9b04321
ACS Catal. 2020, 10, 1210−1223