Machine Learning in Enzyme Engineering
Stanislav Mazurenko, Zbyněk Prokop, and Jiří Damborský

ACSCatal., Just Accepted Manuscript • DOI: 10.1021/acscatal.9b04321 • Publication Date (Web): 13 Dec 2019

Just Accepted

“Just Accepted” manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides “Just Accepted” as a service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. “Just Accepted” manuscripts appear in full in PDF format accompanied by an HTML abstract. “Just Accepted” manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are citable by the Digital Object Identifier (DOI®). “Just Accepted” is an optional service offered to authors. Therefore, the “Just Accepted” Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the “Just Accepted” Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these “Just Accepted” manuscripts.
Machine Learning in Enzyme Engineering

Stanislav Mazurenko\textsuperscript{1,*}, Zbyněk Prokop\textsuperscript{1,2}, Jiri Damborsky\textsuperscript{1,2}

\textsuperscript{1}Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic

\textsuperscript{2}International Centre for Clinical Research, St. Ann’s Hospital, 602 00 Brno, Czech Republic

*Author for correspondence: stan.mazurenko@gmail.com

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 analyse 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

Graphical abstract:
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 co-solvents, 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 the in-depth computational modelling of enzymes by accounting for the physico-chemical 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 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 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 neighbours classifiers to predict enzyme function (19) and mechanisms (20), various scoring and clustering algorithms for rapid functional sequence annotation (21, 22). The main attractiveness of ML in enzyme engineering stems from its generalisability: once 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 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 in biocatalyst design due to necessary strict quality control for data collection and reporting, difficult standardisation of data format, lack of sizeable homogenous 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, analysis in systems metabolic engineering, implementation of biocatalysts in systems in the industry, and biosystems design, we refer our readers to following reviews (10, 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, physico-chemical 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 based on their descriptors, using a labelled training dataset. Quite often, these two ML types–supervised and unsupervised–are combined, e.g. to reduce the input data dimensionality of a predictor or to fill in missing labels in training data, which is called semi-supervised 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 towards 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 over-optimistic 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 sub-split 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 dataset compared 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, e.g. excessive noise, irrelevant or missing features, data bias or sparseness. They can also occur
due to poor application of an algorithm, e.g. excessive or insufficient flexibility in parameter selection, improper training protocol, contamination of the training dataset with the test dataset. In the following section, we summarize the state of the art and challenges related to both databases used for training and applications of ML algorithms in enzyme engineering.

Figure 1. Schematic workflow of constructing an ML predictor and associated challenges. Step 1 – The data is 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, hence, must be accounted for. Step 2 – The predictor is trained on the training data set, e.g. 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 dataset, e.g. calculating true and false positives and negatives and the associated measures or calculating the root mean squared error (RMSE) 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
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 dataset 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, co-factors, conditions, wide ranges of substrate specificities and enantioselectivities. Yet, several excellent databases are addressing this challenge from various perspectives.

Table 1. Examples of databases used for ML in enzyme engineering grouped by properties and highlighted accordingly.

<table>
<thead>
<tr>
<th>Database</th>
<th>Property</th>
<th>Size</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Location</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>InterPro</td>
<td>A consortium of databases of protein families and domains</td>
<td>35 020 entries based on 88 938 signatures</td>
<td>Actively maintained; UniProtKB sequences covered by 81%; includes homologous superfamilies; discontinuous domains supported</td>
<td>Inhomogeneous output due to combining entries from 14 databases</td>
<td><a href="http://www.ebi.ac.uk/interpro/">http://www.ebi.ac.uk/interpro/</a></td>
<td>(27)</td>
</tr>
<tr>
<td>UniProtKB</td>
<td>Protein sequences</td>
<td>&gt; 60 000 000 automatically annotated; &gt; 550 000 manually curated</td>
<td>Actively maintained; largest database of sequences; numerous organisms represented; evidence scores provided</td>
<td>Redundant annotations; limited reliability of TrEMBL; many proteins with unknown function</td>
<td><a href="https://www.uniprot.org/">https://www.uniprot.org/</a></td>
<td>(28)</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein structures</td>
<td>&gt; 140 000 manually curated</td>
<td>Actively maintained; largest database of 3D structures; numerous organisms represented; extensive annotations</td>
<td>Missing or inaccurate structures of many membrane-bound and difficult-to-express proteins; poor standardization of PDB format; missing transition states</td>
<td><a href="https://www.rcsb.org/">https://www.rcsb.org/</a></td>
<td>(29)</td>
</tr>
<tr>
<td>Brenda</td>
<td>Functional data: reaction, specificity, kinetic parameters and profiles, genomic sequences, structures, stability</td>
<td>~ 84 000 enzymes manually annotated; ~ 1 600 000 entries based on text mining</td>
<td>Actively maintained; comprehensive coverage; numerous organisms represented (11 000); includes mutants</td>
<td>Inhomogeneous data; numerous data sources used require stricter quality control</td>
<td><a href="https://www.brenda-enzymes.org/">https://www.brenda-enzymes.org/</a></td>
<td>(30)</td>
</tr>
<tr>
<td>BKMS-react</td>
<td>Functional data: reaction, kinetic parameters and profiles, experimental conditions, pathways</td>
<td>59 981 unique reactions from BRENDA, KEGG, MetaCyc, and SABIO-RK</td>
<td>Actively maintained; comprehensive coverage; reaction-oriented; provides overviews on all associated pathways</td>
<td>Inhomogeneous data; numerous data sources used require stricter quality control</td>
<td><a href="http://bkms-react.hs-mittweidt.de/">http://bkms-react.hs-mittweidt.de/</a></td>
<td>(30)</td>
</tr>
<tr>
<td>EzCatDB</td>
<td>Enzyme Catalytic-mechanism Database</td>
<td>871 enzymes</td>
<td>High quality curating of reaction intermediates; wide coverage of enzyme classes; overall reactions rather than individual steps are classified</td>
<td>Inhomogeneous data; numerous data sources used, hence, stricter quality control needed; small dataset; rare updates</td>
<td><a href="http://ezcatdb.cbc-pj.jp/EzCatDB/">http://ezcatdb.cbc-pj.jp/EzCatDB/</a></td>
<td>(31)</td>
</tr>
<tr>
<td>M-CSA© Mechanism and Catalytic Site Atlas</td>
<td>Annotation of catalytic residues, cofactors, and the reaction mechanisms</td>
<td>361 manually curated entries: 423 with detailed mechanism and 38 with catalytic site residues only</td>
<td>Actively maintained; 81% coverage of the 3rd level of EC numbers; Rating of redundant mechanisms based on evidence; reaction steps are annotated in detail</td>
<td>Inhomogeneous data; numerous data sources require stricter quality control</td>
<td><a href="https://www.ebi.ac.uk/thornton-srv/m-csa/">https://www.ebi.ac.uk/thornton-srv/m-csa/</a></td>
<td>(32)</td>
</tr>
<tr>
<td>FireProt DB</td>
<td>Thermostability change upon mutations</td>
<td>1 329 single-point mutations from 79 proteins, manually curated</td>
<td>Actively maintained; pre-processed for the machine learning applications</td>
<td>the amount of data is not sufficient for the application of advanced ML methods</td>
<td><a href="https://schmidtche.mi.muni.cz/fireprotdb/">https://schmidtche.mi.muni.cz/fireprotdb/</a></td>
<td>(5)</td>
</tr>
</tbody>
</table>
### 3.2 Current challenges related to databases

In general, if the sought-for dependency 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 physico-chemical 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, datasets 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 are yet to be developed (38). The lack of established reporting standards often results in missing or, even worse, erroneous values for some descriptors, e.g., 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 clean-up procedures, which is no longer implemented for some databases as those are not actively maintained anymore. Apart from manual double-checking, those procedures may also involve “data tidying” to render datasets ML friendly, i.e., 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

<table>
<thead>
<tr>
<th>Database</th>
<th>Thermostability change upon mutations</th>
<th>Solubility based on in vitro protein translation and centrifugation</th>
<th>Solubility change upon mutations</th>
<th>Solubility</th>
<th>All types of protein mutational data from protein engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProTherm</td>
<td>new data from recent publications added</td>
<td>Largest database of single-point mutants with stability data</td>
<td>Not actively maintained; erroneous entries; mishandling multistep unfolding</td>
<td>eSOL</td>
<td>manual curation only at the initial stage; numerous data sources used and very heterogeneous data, hence, stricter quality control and reliability scoring needed</td>
</tr>
<tr>
<td>SoluProtMut DB</td>
<td>3,464 single-point mutations from 135 proteins; 1,564 entries from 99 proteins after a clean-up</td>
<td>9,147 proteins highly uniform and consistent dataset</td>
<td>Not actively maintained; single organism (E. coli/ ORF library); a low number of negative samples</td>
<td>N.A</td>
<td></td>
</tr>
<tr>
<td>TargetTrack®</td>
<td>297,404 proteins, 961,548 trials</td>
<td>Recent compilation of mutations changing protein solubility</td>
<td>The amount of data not sufficient for the application of advanced ML methods; inhomogeneous data collected using different techniques</td>
<td>N.A</td>
<td></td>
</tr>
<tr>
<td>ProtaBank</td>
<td>&gt;700 unique proteins and &gt;1,800,000 mutants</td>
<td>Recent and actively maintained; provides detailed descriptions of experimental protocols used</td>
<td>Not actively maintained; low-quality annotations of trials and expression systems; different extraction methods used; suboptimal database size</td>
<td>N.A</td>
<td></td>
</tr>
</tbody>
</table>

*Previously known as PepDB or TargetDB; °merge of MAcIE and CSA; N.A. – not applicable
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 towards 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 started to attract researchers’ attention recently (45, 46) and are 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 datasets of higher quality. The former is obstructed by improving scientific methods and consequently pressing the need for a review and replication of the results published earlier. Hence, the latter option becomes more attractive, and in our opinion, several promising in this respect up-and-coming experimental techniques have already been presented: (i) next-generation sequencing, (ii) fluorescence-activated cell sorting, (iii) deep mutational scanning, and (iv) microfluidics. These techniques 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 towards 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 a day. The advent of this ultra-high-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, e.g., recently introduced third-generation (long-read) methods employing single-molecule real-time (50) or nanopore sequencing (51) resolve the limitations of short reads, e.g., 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 micro-capillaries, further miniaturization and parallelization are possible (100,000 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 micron-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 compared 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 $10^6$–$10^8$ members) at low cost. Screening of enzyme mutants in picoliter compartments, generated at a kilohertz speed in microfluidic devices, is coming of age (54).

When 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.
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 sequence 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 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.
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 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 145 thousand 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 dataset size compared to millions of data points usually available for this method, deep neural networks showed most 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 the 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 nowadays. Computational methods for the protein function prediction range from sequence- to structure-based; 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 based on 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 non-uniformity 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 datasets of enzyme structures and activities accumulated to date already allow using deep learning in the engineering of catalytic activity. Nevertheless, the problems with the datasets 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 earlier. 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. logP,
molecular area, 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
retrained on the data for another family.

Predictors of protein solubility usually exploit the eSol database for the entire ensemble of
*Escherichia coli* proteins (67). In their recent paper (35), Xi Han and co-authors 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. 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 datasets
of protein solubility. Moreover, a modest best-achieved R² 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 the 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, and
enantioselectivity can also be targeted if sufficient data is available. Protein stability predictors are
perhaps the ones with the most abundant datasets 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 co-evolution scores for mutated positions, amino acid substitutions and their
physicochemical properties, neighbourhood 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 the data set of
higher quality, predicting protein stability remains an extremely challenging task.

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 catalysed 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 enantimeric excess (ee) of 76% for (S)-enantiomer was sequentially improved to 93% by several rounds of ML-guided evolution experiments, and a new variant with 79% ee for (R)-enantiomer was discovered. The authors also compared two standard directed evolution approaches with the one assisted by shallow neural networks. They used previously published 149 361 measurements of a total of 160 000 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 the variants tested.

**Historical perspective on the applications of ML to enzyme engineering**

Table 2. Selected examples of the application of machine learning in the field of enzyme engineering.

<table>
<thead>
<tr>
<th>Year</th>
<th>Object</th>
<th>Target property</th>
<th>Data</th>
<th>Model and method</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>haloalkane dehalogenase</td>
<td>Function</td>
<td>15 mutants of haloalkane dehalogenase©</td>
<td>partial least squares regression</td>
<td>(72)</td>
</tr>
<tr>
<td>1998</td>
<td>subtilisin, haloalkane dehalogenase, T4 lysozyme, tryptophan synthase</td>
<td>Function and stability</td>
<td>19 mutants of subtilisin, 15 mutants of haloalkane dehalogenase, 13 mutants of T4 lysozyme, 18 mutants of tryptophan synthase©</td>
<td>partial least squares regression, principal component analysis</td>
<td>(73)</td>
</tr>
<tr>
<td>1999</td>
<td>human acetylcholinesterase</td>
<td>Cell surface located activity</td>
<td>35 dipeptide mutants</td>
<td>principal component analysis, partial least squares regression</td>
<td>(74)</td>
</tr>
<tr>
<td>2000</td>
<td>prolyl endopeptidase and thermolysin</td>
<td>Fitness landscape: thermostability and enzymatic activity</td>
<td>19 mutants of prolyl endopeptidase and 16 mutants of thermolysin©</td>
<td>additive model (a version of linear regression)</td>
<td>(75)</td>
</tr>
<tr>
<td>2001</td>
<td>Kazal protein inhibitors</td>
<td>Association equilibrium constants</td>
<td>1146 constants = 191 single-point variants x 6 proteins for training and 398 constants for testing</td>
<td>sequence to reactivity algorithm (a version of linear regression)</td>
<td>(76)</td>
</tr>
<tr>
<td>2005</td>
<td>haloalkane dehalogenase</td>
<td>Substrate specificity</td>
<td>116 halogenated compounds©</td>
<td>partial least squares regression, principal component analysis</td>
<td>(77)</td>
</tr>
<tr>
<td>2007</td>
<td>halohydrin dehalogenase</td>
<td>Function</td>
<td>~600 000 mutants based on various techniques with ~280 000 used by the ML-based method</td>
<td>partial least squares regression</td>
<td>(78)</td>
</tr>
<tr>
<td>2010</td>
<td>toluene-4-monoxygenase</td>
<td>Function</td>
<td>24 variants© (phase I) + 9 variants (phase II)</td>
<td>Gaussian random field</td>
<td>(79)</td>
</tr>
<tr>
<td>2014</td>
<td>various enzymes</td>
<td>Scoring functions for binding affinity</td>
<td>1 300 protein–ligand complexes from PDBbind®</td>
<td>random forest</td>
<td>(80)</td>
</tr>
<tr>
<td>2015</td>
<td>various enzymes</td>
<td>Discrimination of acidic and alkaline enzymes</td>
<td>217 enzymes from BRENDA database®</td>
<td>K-nearest neighbour, support vector machine, decision tree, artificial neural network, probabilistic neural network</td>
<td>(81)</td>
</tr>
<tr>
<td>2018</td>
<td>various enzymes</td>
<td>Lysine malonylation sites</td>
<td>9 760 experimentally validated malonylation sites: 1 746 sites from 595 E. coli proteins, 3 435 sites from 1 174 proteins in <em>M. musculus</em>, and 4 579 sites from 1 660 proteins in <em>H. sapiens</em>©</td>
<td>random forest, support vector machine, decision trees with gradient boosting, K-nearest neighbour, logistic regression</td>
<td>(82)</td>
</tr>
<tr>
<td>2018</td>
<td>glycosyltransferase superfamily 1</td>
<td>Function</td>
<td>label-free mass spectroscopy-based assay data: 54 enzymes and 91 substrates©</td>
<td>decision trees</td>
<td>(86)</td>
</tr>
<tr>
<td>2018</td>
<td>various enzymes</td>
<td>EC class prediction</td>
<td>63 558 enzymes from RCSB PDB®</td>
<td>convolutional neural networks</td>
<td>(83)</td>
</tr>
<tr>
<td>2019</td>
<td>various proteins</td>
<td>Solubility</td>
<td>chaperone-free reconstituted translation system: ~500 cytosolic</td>
<td>random forest</td>
<td>(84)</td>
</tr>
</tbody>
</table>
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 standardise the training of an ML algorithm for non-specialists, 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, the 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 towards new and more complex ML methods, namely aggregating several algorithms into hybrid meta-predictors, hyperparameter optimization with many training sub-cycles, feature learning, and the fusion of ML-based and classical bioinformatics tools in a single predictor, will further challenge the cross-talk 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 standardised protocols for comparison and new datasets for testing. The lack of benchmark datasets, discrepancies in the performance measurements used, inaccurate or insufficient disclosure in publications, and the difficulty 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 earlier. 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 noticing 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 impacts 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 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 analyse learning curves. Why a particular predictor has 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 analyses, 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 datasets and unified performance measurements is only a matter of time. Recently, an intriguing algorithm based on semi-supervised 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 dataset 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 stereo-divergent 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 usage 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 modelling of data sets (generative models), and using 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 only recently been made to apply some of those advanced techniques to proteins, i.e. using generative models to create soluble and functional malate dehydrogenase variants (90) or predict mutational effects with high correlation with those actually observed in 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 usage 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 nowadays. This should further boost predictive power since 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 towards 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 based on the structure of a predictor, e.g. using adaptive sampling (98). Such an approach favours 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 based on the coefficients (99); in contrast, sophisticated predictors are usually prone to pathological behaviour, i.e. sudden misclassification after a slight and almost imperceptible perturbation of input (100).

![Figure 3. A comparison of decision boundaries for a hypothetical linear predictor with a more flexible non-linear predictor.](image)

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. And 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.

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 (102) simulates cellular growth and allows in silico investigations of the molecular mechanisms underlying genotype-phenotype associations based on the analysis of different parts of the neural network (Figure 4). Since enzymes can also be represented hierarchically based on the annotation of their secondary, tertiary, and quaternary structure elements and their interactions, a shift towards applying interpretable visible neural networks may provide new insights into the mechanism in addition to better predictors.
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.

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 datasets 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. Several databases with millions of 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 biocatalysts’ design is far from being fully explored. The community has still many challenges to face. The lack of homogeneous and consistent datasets of high quality for training and validation, classical data imbalances and biases, intrinsic multidisciplinarity of the approach, difficulties in explaining, interpreting, and comparing the results of predictors are among one 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.
Acknowledgements

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). Stanislav Mazurenko 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


