

Machine learning-enabled retrobiosynthesis of molecules

Received: 30 May 2022

Accepted: 22 December 2022

Published online: 16 February 2023



Tianhao Yu^{1,2,3}, Aashutosh Girish Boob^{1,2,4}, Michael J. Volk^{1,2,4}, Xuan Liu^{1,2,3}, Haiyang Cui^{1,2,3} & Huimin Zhao^{1,2,3,4}✉

Retrobiosynthesis provides an effective and sustainable approach to producing functional molecules. The past few decades have witnessed a rapid expansion of biosynthetic approaches. With the recent advances in data-driven sciences, machine learning (ML) is enriching the retrobiosynthesis design toolbox and being applied to each step of the synthesis design workflow, including retrosynthesis planning, enzyme identification and engineering, and pathway optimization. The ability to learn from existing knowledge, recognize complex patterns and generalize to the unknown has made ML a promising solution to biological problems. In this Review, we summarize the recent progress in the development of ML models for assisting with molecular synthesis. We highlight the key advantages of ML-based biosynthesis design methods and discuss the challenges and outlook for the further development of ML-based approaches.

Functional molecules play a critical role in addressing many of the problems facing society today, such as energy, sustainability and health. Moreover, the synthesis of large and complex molecules with multiple stereocentres remains a great challenge. To address this challenge, both chemo- and biocatalysis have been explored extensively¹. Compared with chemical catalysts, enzymes usually pose several advantages, including high catalytic activity and selectivity, as well as the ability to perform reactions under mild conditions. Thus, enzymes are often preferable in chemo-, regio- or stereoselective reactions² and can achieve more sustainable production processes on laboratory and industrial scales³. Although many studies have successfully applied enzymes in large-scale organic synthesis⁴, biocatalysts still have limitations in routine synthetic reactions⁵. Even though many enzyme sequences are available in various databases, for example, UniProt⁶, only a small fraction have been annotated due to difficulties in experimental characterization⁷. The collected substrate scope in enzymatic databases is also limited, which raises challenges for the design of retrobiosynthesis planning tools and the selection of the corresponding enzymes in the designed biosynthetic pathways.

Various machine learning (ML) models have been proposed to address these limitations. Generally, the ML models can be divided

into two categories: supervised and unsupervised. Supervised ML models use labelled datasets such as enzyme–function pairs to train an ML model in a biological context. They learn relationships between input samples and labels and generalize them to make predictions for unlabelled inputs. Such models have been designed to assist the learning of reaction rules and enzyme engineering to improve characteristics such as activity, stability and substrate specificity^{8–11}. By contrast, unsupervised ML models take unlabelled datasets as input. They extract and recognize complex features and patterns from input information alone. Such models have enabled the exploration of the existing protein universe and have assisted in enzyme design efforts^{12,13}.

In this Review, we discuss how ML is enabling the realization of synthesis of molecules by accelerating the retrobiosynthesis workflow, including retrosynthesis planning, enzyme identification and selection, and the engineering of enzymes and pathways (Fig. 1). Finally, we provide a summary of standardized databases for readers who are interested in developing ML models.

Retrobiosynthesis planning

With the ever-increasing number of synthetic biology tools¹⁴, it has become possible to engineer biochemical pathways to synthesize

¹Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA. ²Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. ³NSF Molecule Maker Lab Institute, University of Illinois at Urbana-Champaign, Urbana, IL, USA. ⁴DOE Center for Advanced Bioenergy and Bioproducts Innovation, University of Illinois at Urbana-Champaign, Urbana, IL, USA. ✉e-mail: zhao5@illinois.edu

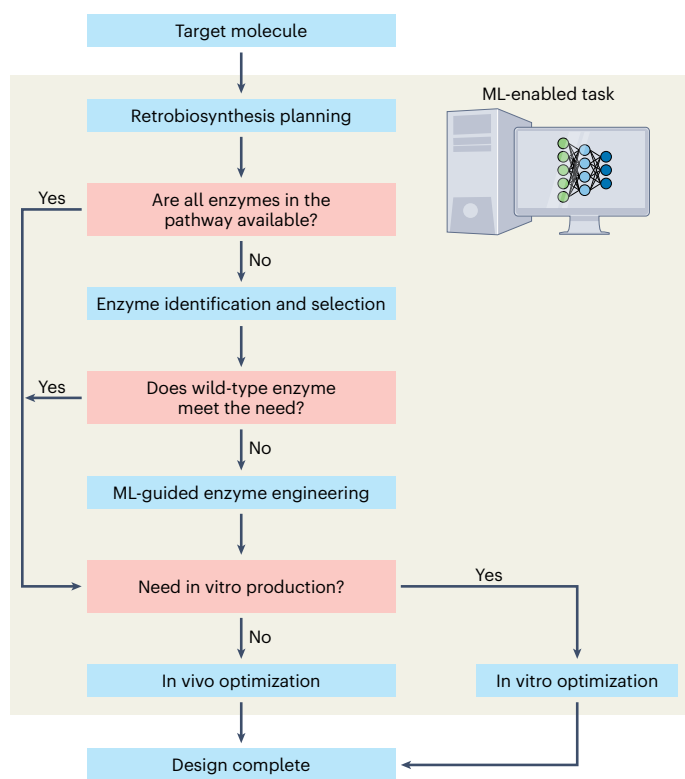


Fig. 1 | An overview of an ML-enabled retrobiosynthesis workflow. The design of target molecule synthesis can be initiated by proposing pathways using retrobiosynthesis planning tools. If enzymes are missing from the proposed pathway, ML can assist in the identification and selection of potential enzymes to catalyse the proposed reaction. In the case where the wild-type enzyme requires further improvement, ML can guide the enzyme engineering campaign. Finally, ML can also accelerate the optimization process for the production of molecules in vivo or in vitro. Tasks that sit within the beige box can be ML-enabled.

molecules with complex stereochemistry and wide structural diversity^{15,16}. However, difficulties lie in synthesizing molecules for which the routes are unknown and in decreasing pathway complexity while increasing yields. Recently, insights have been drawn from the use of ML in organic chemistry to guide efforts to predict biosynthetic pathways¹⁶. Here we discuss the principles of applying ML to retrosynthesis planning tools^{9,16,17} and provide a detailed comparison of template-free and template-based methods^{18–20} with an emphasis on checking reaction feasibility²¹ and novel pathway discovery.

A retrobiosynthesis-based pathway design workflow can be generalized into three modules (Fig. 2): reaction database, precursor inference approach and pathway search. The first module includes the enzymatic reaction corpus and methods for translating reactions into a machine-readable language. The second module deduces the biosynthetic pathway of a target molecule by template-based and/or template-free approaches. The template-based approach infers precursors by matching reaction templates and finding disconnection sites, while the template-free approach predicts precursors using a trained ML model without predefined reaction templates. In the third module, the ranked precursors are checked by terminating conditions, such as commercial availability or presence in the microbial strain of choice¹⁵. If the precursor satisfies one of the terminating conditions, the system outputs the synthetic pathway. Otherwise, the model iteratively uses the precursor as input to search for its precursors.

Most of the existing retrobiosynthesis tools are template-based, using reaction templates in a reverse manner to obtain precursors. For example, novoStoic extracts a set of reaction rules from the MetRxn dataset based on the reaction centre¹⁹, and RetroPath2.0 extracts reaction rules from MetaNetX based on atom-neighbourhood hopping²⁰. While automatically extracting reaction rules from a reaction database may cause rule sets to become redundant, Finnigan et al. developed RetroBioCat using expertly encoded reaction rules consisting of 99 reaction templates¹⁸. However, these template-based tools ignore the effect of long-range substituents on the reaction centre. Additionally, reaction rules that are too specific or too general will lead to the predicted routes being overly conservative or unrealistic, respectively, thereby necessitating laborious and time-consuming optimization by experts.

By contrast, template-free retrosynthesis tools use a database of reactions to train an ML model to predict precursors by inputting a molecule of interest. The formulation of the task is to translate a molecule (input) into another molecule (precursor), which is similar to a widely studied task in the computer science community known as natural language processing (NLP). One of the powerful models applied to language translation tasks in NLP, sequence-to-sequence (Seq2Seq) approaches, is widely used in retrosynthesis due to the ease of representing molecules in text form using a simplified molecular-input line-entry system (SMILES)²². Therefore, using the SMILES of the target molecule to infer the SMILES of precursors can be treated as a machine translation problem. Although the Seq2Seq method has achieved remarkable success in chemical retrosynthesis, it is not advisable to directly apply it to retrobiosynthesis. Unlike chemical reaction databases containing several hundred thousand reactions, the relatively small number of enzymatic reactions, around two orders of magnitude smaller¹, will more likely lead to overfitting. Zheng et al. developed BioNavi-NP to address this issue by combining 62,000 natural product-like organic reactions and 33,000 biochemical reactions¹⁶. The expanded dataset enhanced the performance of their single-step prediction model, and the accuracy is substantially improved compared with using the two datasets separately. Similarly, Probst et al. developed a template-free retrobiosynthesis tool that uses a curated enzymatic reaction dataset, named ECREACT, consisting of 62,000 enzymatic reactions and the US Patent Office (USPTO) dataset consisting of one million organic chemical reactions to train a molecular transformer using multi-task transfer learning (Fig. 3)⁹. In their tool, enzymatic reactions from ECREACT and chemical reactions from USPTO are split into reactants and products. The reactants for enzymatic reactions are associated with the reaction Enzyme Commission (EC) number. Then their SMILES are tokenized using a method in which each character is represented by a token. A prefix letter (v, u or t) is added to each EC level (1, 2 or 3) during tokenization. The forward prediction uses reactants with an EC number as input and products as output, and the backward prediction goes in the opposite direction. Transfer learning divides a batch into two parts based on the assigned weight to determine the number of inputs from ECREACT and the number of inputs from USPTO in the batch, and they are trained simultaneously using a transformer model²³. This approach can compensate for the relatively small datasets of enzymatic reactions that do not provide enough information for models to learn general chemistry and SMILES grammar. As a result, the forward-reaction prediction model achieved a top-1 accuracy of 49.6%, and the backward retrosynthetic prediction model achieved a top-1 single-step round-trip accuracy (the percentage of precursor sets leading to the initial target molecule when the forward model evaluates the precursor sets) of 39.6%.

Besides adopting ML for precursor inference, ML can also be used in other modules of the retrobiosynthesis-based pathway design workflow, such as molecule representation, reaction feasibility check, precursor ranking and heuristic pathway search. Molecule representations determine how to encode reactions in a database and can play an

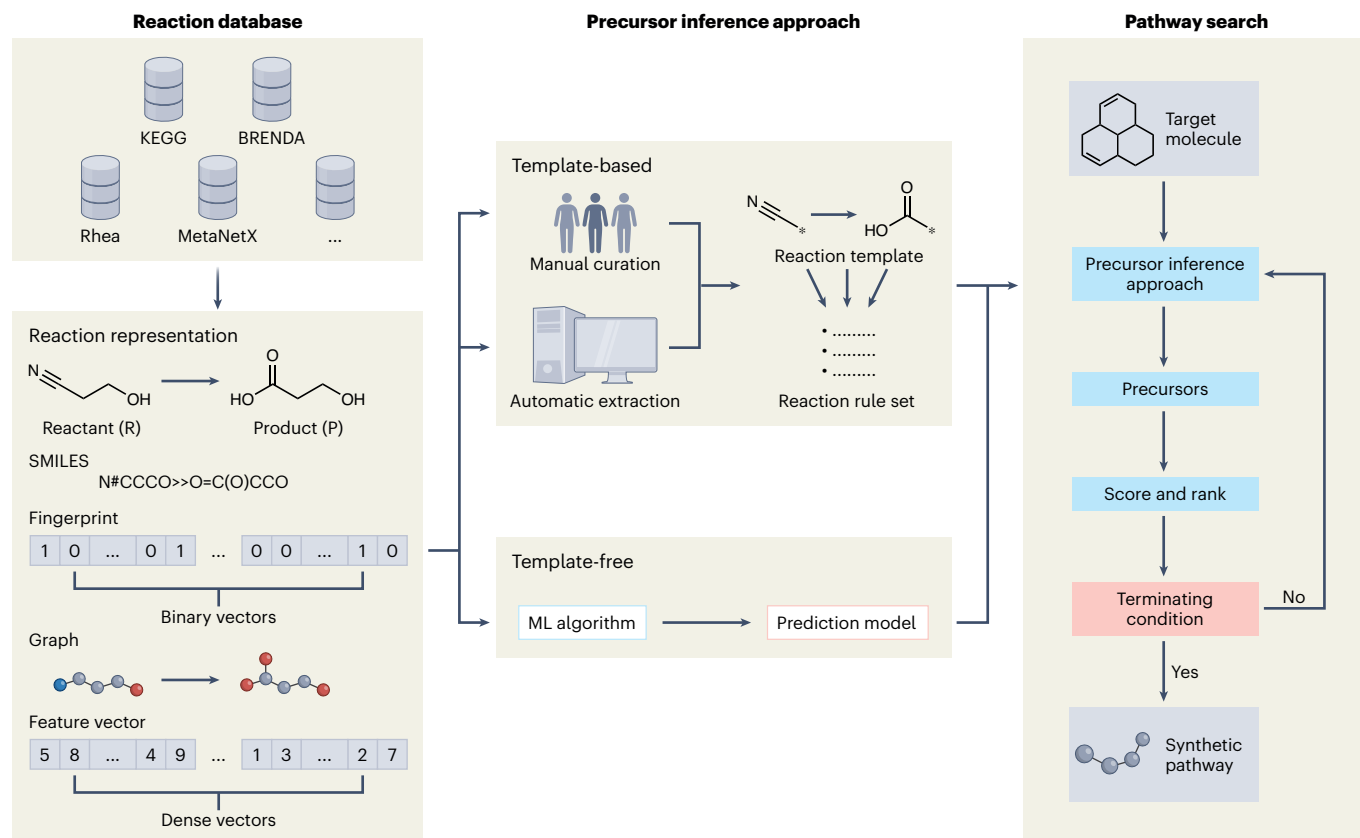


Fig. 2 | A conceptualized pathway design workflow based on a retrobiosynthesis tool. The enzymatic reaction corpus is represented by a machine-readable descriptor, such as SMILES, fingerprints, graphs and feature vectors. Reaction rule sets consist of reaction templates from either manual

curation or automatic extraction, while an ML algorithm uses the reaction corpus to train a prediction model. Then the pathway search algorithm builds synthetic pathways by iterating precursors predicted by inference approaches and ranked by a scoring function until a precursor satisfies the terminating conditions.

important role in how the ML model is formulated to solve the retrobiosynthesis task. In addition to using SMILES to represent molecules, molecular fingerprints, graphs and feature vectors have also been used for retrosynthesis, and will soon find their way to retrobiosynthesis. Molecular fingerprints represent a molecule in vector format based on its fragments/substructure. Hasic et al. developed a template-free fingerprint-based approach for retrosynthesis using a neural network to predict the disconnection sites and suggest the chemical transformations²⁴. A molecular graph represents a molecule with nodes corresponding to atoms and edges corresponding to atom–atom bonds. By identifying synthetic building blocks of a target molecule, Somnath et al. built a graph-based, semi-template-based model that transforms building blocks into valid reactants²⁵. Molecular feature vectors, also known as embeddings, represent molecules as dense vectors with an ML model and can benefit a wide range of downstream tasks, such as retrosynthesis, drug discovery and molecule generation²⁶. One example is a variational autoencoder (VAE), which can project molecules or chemical reactions to a latent space^{27,28}. According to this method, an enzymatic reaction could be regarded as a difference vector between the main substrate and the main product in the latent chemical space¹⁷. Therefore, the forward and backward reaction predictions can be simplified to simpler vector operations.

One particular disadvantage associated with the predictions made by ML models is that they may contain thermodynamically unstable structures, syntax errors in SMILES or unfeasible reactions, while the template-based retrobiosynthesis tools are more likely to infer the stable structure of precursors through the breaking and formation of chemical bonds. Such disadvantages directly affect the accuracy

and reliability of ML-based retrobiosynthesis tools. Therefore, it is necessary to incorporate a filter module to detect and rectify these errors²⁹. In addition, the synthetic complexity and reaction feasibility can serve as a scoring function to rank precursors and facilitate retrosynthesis tools to find optimum synthetic pathways. Synthetic complexity score (SCScore) learnt from a reaction corpus is an ML-based scoring function that evaluates the difficulty of synthesizing a molecule³⁰. For template-based approaches, an ML model learnt from a pair of a product and its corresponding template can predict and rank the best templates given a target molecule³¹. In terms of reaction feasibility, Deep learning-based Reaction Feasibility Checker (DeepRFC) of reactant pairs evaluates the feasibility of enzymatic reactions, providing an efficient means to classify reaction feasibility²¹. Chemical similarity between predicted reactants and reactants of an existing reaction can also be used as a potential metric of reaction feasibility¹⁵. If the ML model of the precursor inference approach includes both a forward and backward prediction model, the prediction results can use a metric related to the forward confidence of precursors to rank³². These prioritized results can be adopted by a heuristic search algorithm to find retrosynthetic pathways. For example, Monte Carlo tree search (MCTS), which consists of selection, expansion, simulation and backpropagation, can be used to effectively perform synthesis planning^{15,33}. Additionally, retrosynthetic pathway search algorithms can also apply beam search (for example, hypergraph exploration strategy³²), which considers multiple best options based on beam size to expand the pathway search tree, and A* search (for example, Retro* (ref. ³⁴)), which is a best-first search that expands the most promising precursor.

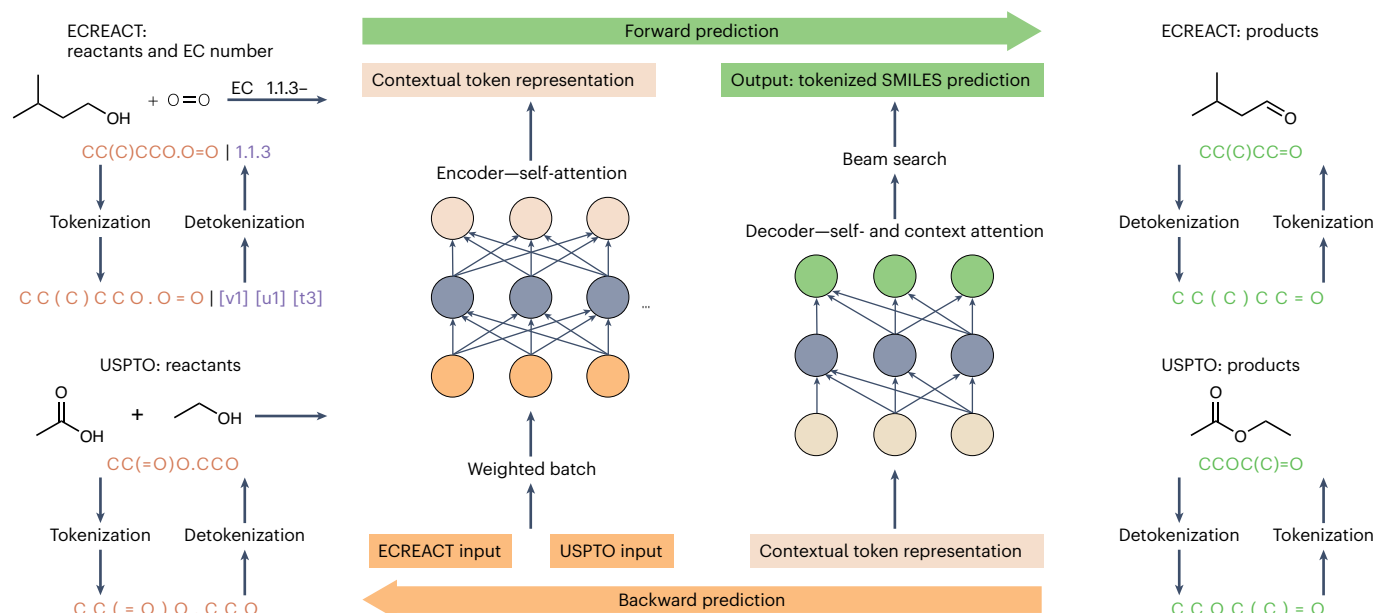


Fig. 3 | Multi-task transfer learning for retrosynthesis. Enzymatic reactions from ECREACT and chemical reactions from USPTO are used to train two molecular transformers. Reactions are represented by SMILES and tokenized SMILES, in which each token is separated by a space, are used for model input. The forward prediction model predicts reaction products based on input

reactants with EC number, while the backward prediction model predicts reactants for a given product. In the transfer learning process, a transformer uses the batch based on the given weights of the ECREACT and USPTO inputs for training. The output of the transformer is detokenized to obtain the result.

Enzyme identification and selection

After a retrosynthesis route has been proposed, enzymes need to be identified and selected to fill the missing links between each step of the pathway. In the case where such enzymes are missing from the enzymatic databases, ML models have been developed to predict enzyme functions, reactivities and other properties, aiming to accelerate the identification and selection process.

Prediction of enzyme classification numbers

The EC number is a hierarchical classification system of enzymes based on the enzymatic reactions that they catalyse. It consists of four levels, separated by periods, where each subsequent level is a subclass of the previous level. The EC number is a useful scheme for enzyme identification and is often included as part of the output from retrosynthesis planning tools. Accurately identifying enzymes with the desired activity is essential to retrosynthesis planning. Several databases have been developed that offer EC number look-up, yet compared with uncharacterized sequence space, only a small portion have been reliably labelled. Many methods have been developed to predict the function of unlabelled sequences. Automatic annotation tools are often used to assign enzyme functions to unknown sequences based on sequence or signature similarity to a known enzyme. Common tools for detecting similarity include the basic local alignment search tool (BLAST) or hidden Markov models (HMMs)^{35,36}. However, similarity-based methods may fail when detecting remote homologues and do not always generate robust predictions³⁷. In fact, one-third of bacterial protein sequences still cannot be labelled by such methods³⁸. Recently, ML-based methods have been reported to predict EC numbers. Compared with traditional methods, ML-based methods not only make predictions based on sequence similarity, but can also infer additional features outside of the similarity to homologues³⁹. The two approaches are compared in Fig. 4.

ProteinInfer, a state-of-the-art model, is a dilated convolutional neural network (CNN) that infers functional annotations from protein

sequences⁴⁰. The implementation of a dilated CNN instead of a regular CNN enables the neural network to explore a longer region of the input amino acid sequences. The model is trained on an annotated enzyme dataset curated from the Swiss-Prot database⁶. The model can achieve a high prediction accuracy even when predicting the fourth level of the EC number. Moreover, the model can link function with specific sequence regions. ProteinInfer is offered as a web tool where users can make rapid predictions and visualize results. Another representative work, DeepEC, is a hybrid approach combining both ML and alignment designed by Ryu et al.³⁷. The ML portion of DeepEC consists of three CNNs predicting enzyme or non-enzyme, EC number up to the third digit and EC number up to the fourth digit, respectively. The alignment portion of DeepEC can be used complementarily to CNNs and is responsible for giving output when the results of the CNNs are inconsistent.

Although ML models have had tremendous success in predicting EC numbers, the apparent accuracy of ML models alone cannot achieve the same performance as sequence alignment methods. However, ProteinInfer has shown that ML models can learn different information from sequence alignment methods. The advantage of ML methods lies in their ability to detect enzymes with low sequence similarity but high functional similarity. A popular approach is to combine both an ML model and sequence alignment as an ensemble model to harvest the merits of both, as demonstrated by both DeepEC³⁷ and ProteinInfer. One potential reason for the accuracy bottleneck might be the imbalanced distribution of EC numbers. Some EC numbers encompass thousands of enzymes, whereas others only encompass one or two characterized enzymes. The biased dataset particularly hinders the learning of under-represented EC numbers. A potential solution to this issue is to use a contrastive learning framework as demonstrated by Heininger et al.⁴¹.

Prediction of enzyme–substrate specificity and promiscuity

Identifying a proper enzyme for each catalytic step proposed by a retrosynthesis tool is critical to the development of a feasible

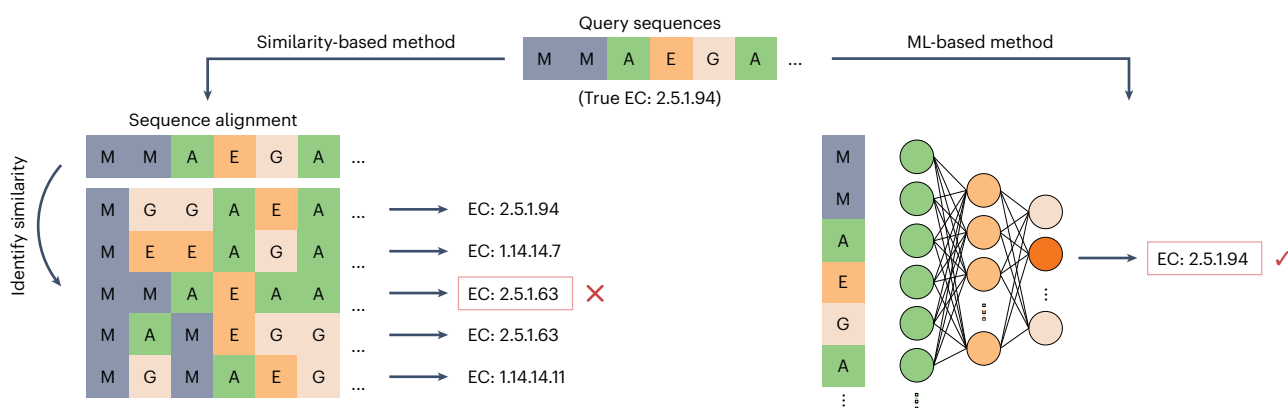


Fig. 4 | Similarity- and ML-based EC number prediction tools. Similarity-based methods align the query sequence with a large known database and identify the closest sequence based on sequence, motif or domain similarity, and assign the EC number accordingly. ML-based methods use known databases to train

classification models tasked to infer and predict the EC numbers of unknown sequences. Both ProteInfer and DeepEC have demonstrated that similarity-based methods can misannotate enzymes that can be correctly identified by ML models.

biosynthetic pathway for a target molecule. Although known for their high specificity, enzymes can exhibit promiscuity, referring to their ability to catalyse reactions with non-native substrates or exhibit new reactivity. Therefore, knowing the substrate specificity of target enzymes is desirable. In the past, enzyme selection was often based on similarity search in databases^{42,43}. Although it remains a challenging task for a generic ML model to make substrate specificity predictions for all enzymes, some recent studies have demonstrated predictive capability on a smaller scale or for family-wide enzymes. For example, Visani et al. framed the promiscuity prediction task as a multi-label classification problem⁴⁴. As promiscuity implies that enzymes possess multiple functions, a multi-label classification model is well suited to the job as it can take a protein sequence as an input and output one or more EC classes. The model can predict which of 983 distinct EC numbers are likely to interact with a given query molecule. The ML model is trained on enzyme–substrate data pairs and it outperforms similarity-based methods. Although the model does not link substrates to enzyme sequences directly, it demonstrates that using enzyme inhibitors as negative training data can boost the model's accuracy. Goldman et al. modelled enzyme–substrate compatibility as a protein–compound interaction task. The study included several families of enzymes covering between 1,000 and 36,000 enzyme–substrate pairs. Although the authors reported that the joint model with both substrate and enzyme failed to outperform single-task models using only substrates or only enzyme sequences, the ML model still outperformed the *k*-nearest neighbours baseline model⁴⁵. More recently, Xu et al. designed an improved substrate encoding and enabled a more accurate substrate specificity⁴⁶.

Prediction of other enzyme parameters

Enzyme characteristics such as solubility, turnover rate and optimum temperature are important parameters for the synthesis of functional molecules. Predicting these parameters in silico can greatly decrease the experimental efforts required for screening or improving these parameters. Many recent studies have used ML models to predict protein solubility and stability. Compared with traditional methods that rely on energy calculations and phylogenetic analysis, ML models are flexible, require no understanding of the mechanistic principles and produce results almost instantaneously⁴⁷. For example, SoluProt predicts protein expression and solubility in *Escherichia coli* using sequence information⁴⁸. The model is trained using the TargetTrack database with the gradient-boosting ML technique. The model achieves almost 60% accuracy, as tested on an independent testing dataset. ML

models are also used to predict the optimal temperature for enzyme activity. Li et al. developed a tool called TOME that predicts the optimum temperature of enzymes from features extracted from enzyme sequences. The model's accuracy outperformed that of the estimate obtained by using the optimal growth temperature of organisms, a commonly used method for estimating the stability of enzymes⁴⁹. In addition, ML models have been developed to predict other properties, such as localization^{50,51}, enzyme loading and yield⁵², kinetics⁵³, and protein–ligand interactions⁵⁴.

Enzyme engineering

Once the biosynthetic pathway has been completely identified in silico, enzymes are expressed to synthesize the functional molecules. However, to increase the titre, rate and yield of the desired molecule, or when the reaction conditions or substrate is not native to the wild-type enzyme, enzymes are engineered by exploring the sequence–function landscape. In this section we summarize two major ML-based approaches for designing variants with improved characteristics and novel attributes.

ML-guided directed evolution

As a powerful protein engineering tool, directed evolution accelerates the process of protein evolution and has been demonstrated to be valuable for increasing the catalytic activity and efficiency of enzymes⁵⁵. The two iterative steps of directed evolution consist of creating a diverse variant library and screening/selecting the library to obtain variants with improved phenotype. However, the enzyme sequence landscape is enormous and using directed evolution to explore all possible mutants is impossible. Moreover, exploration typically follows a greedy search strategy where only the most improved variants of each cycle are selected as parents for the next cycle. The limitation of a greedy search is that it does not guarantee finding the global optimum and in fact it has a high tendency to be trapped in local optima¹¹.

To address these limitations, ML models are used to guide directed evolution experiments, enabling efficient exploration of the sequence landscape^{11,56}. Over the years, ML-guided directed evolution (MLDE) workflow has evolved and now a complete workflow consists of representing proteins using embeddings obtained from a pretrained global language model^{12,13}, predicting fitness using deep learning⁵⁷ or low-*N* models^{58,59}, and exploring the fitness landscape using an optimization model¹⁰ (Fig. 5). Wittmann et al. performed a comprehensive benchmarking of the MLDE workflow in silico⁶⁰. They compared several options for each step of the MLDE, including unsupervised mutational

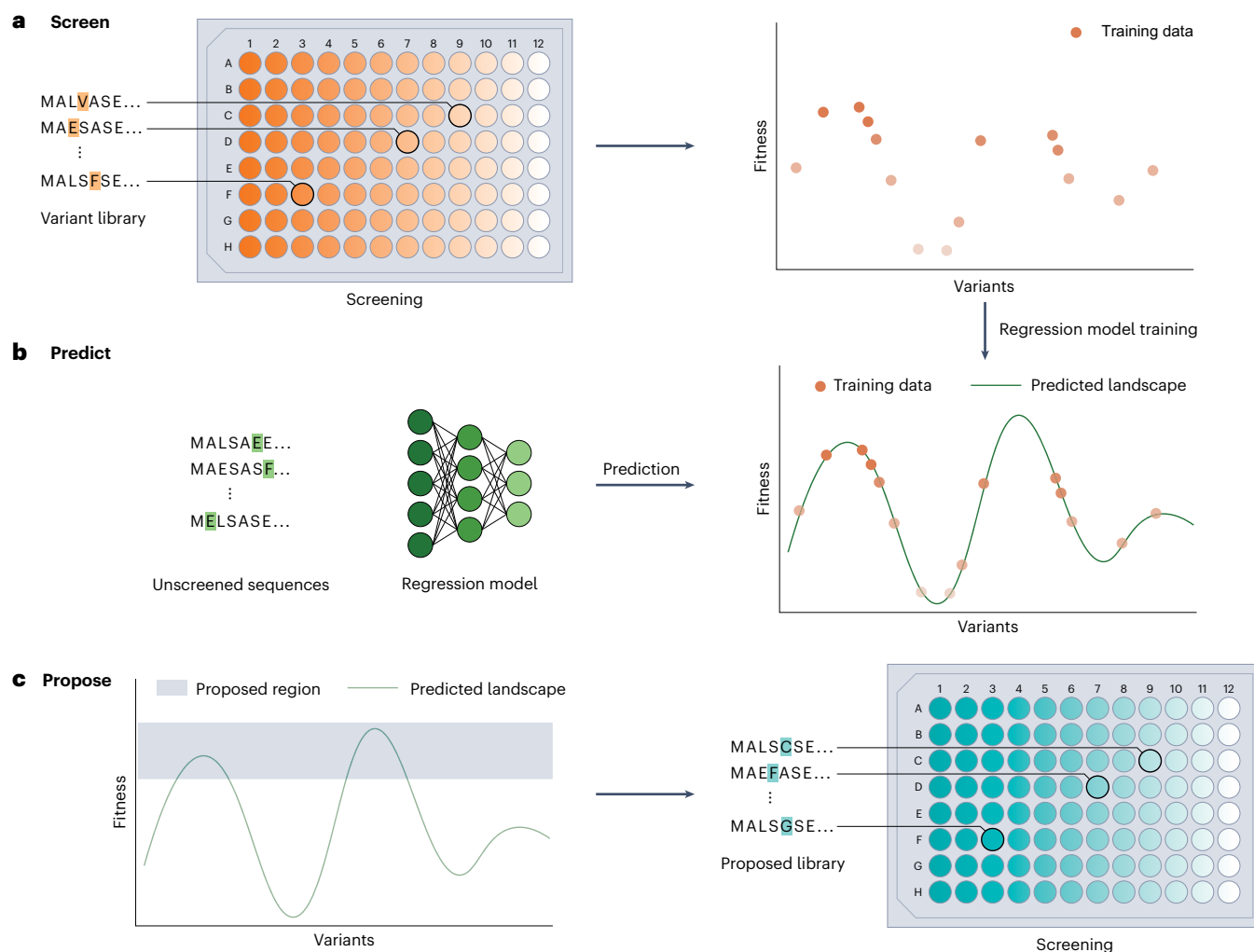


Fig. 5 | An overview of the MLDE workflow. **a**, The first step of the MLDE is to obtain the quantitative fitness of a variant library using an experimental assay. **b**, The obtained variant fitness data can then be used to train a regression model using either a complex neural network architecture or a simple linear regression model. After the model has been trained, it can be used to predict the fitness of

unscreened sequences in silico. **c**, A new variant library is recommended on the basis of the trained ML model. According to the prediction results, the variants with high fitness are prioritized. These steps can be performed iteratively to engineer variants with better activity, stability and so on.

effect predictors, methods for amino acid embedding, ML models for the predictor and different evaluation metrics. This work offers a general guideline on how to apply ML to guide the protein engineering experiment starting with the wild type of the target protein.

Data-driven methods in general require a large amount of training data to be accurate. This is often challenging to obtain because high-throughput screening assays may not be available. To address this limitation, low-*N* ML models were developed to accommodate situations where only few variant data are available. Notably, Hsu et al. developed a low-*N* framework by combining assay-labelled data with evolutionary information⁵⁹. The model embeds each variant's amino acid sequence with the inferred likelihood of its frequency among its homologues augmented with one-hot encoding. To be specific, one-hot encoding is an encoding scheme where each amino acid is represented by a combination of zeros and ones, and all the encoded numerical representations are concatenated to obtain the representation of the full protein sequence. The authors evaluated the model using 19 different protein mutagenesis datasets; the results showed that the model can reach a Spearman correlation of over 0.6 even if trained on only 48 training data points. The work also showed that the augmentation with one-hot encoding improved the model's accuracy.

Directed evolution relies on iterative rounds of library construction and screening⁵⁵. Even with the assistance of ML, it is infeasible to explore all possible variants. Therefore, an efficient method for exploring the variant landscape is desirable to reduce the number of iterations. To this end, Greenhalgh et al. used a Gaussian regressor with the upper confidence boundary (UCB) method to iteratively optimize an acyl-ACP (acyl carrier protein) reductase for improved fatty alcohol production⁶¹. The UCB is a criterion that efficiently explores uncertain regions and rapidly converges to optima. The authors used the UCB to guide the design of the library for each round of experiments. The experimental results showed that variants had a notable improvement in each iteration. In ten rounds of optimization, the engineered reductase leads to a fivefold increase in *in vivo* fatty acid production titre compared with the starting point. Other adaptive learning methods have also been explored in various MLDE studies and are discussed elsewhere¹⁰.

ML for novel enzyme design

MLDE has shown tremendous success in improving enzyme parameters, such as activity, solubility and stability. However, most of the directed evolution studies have only explored the local landscape.

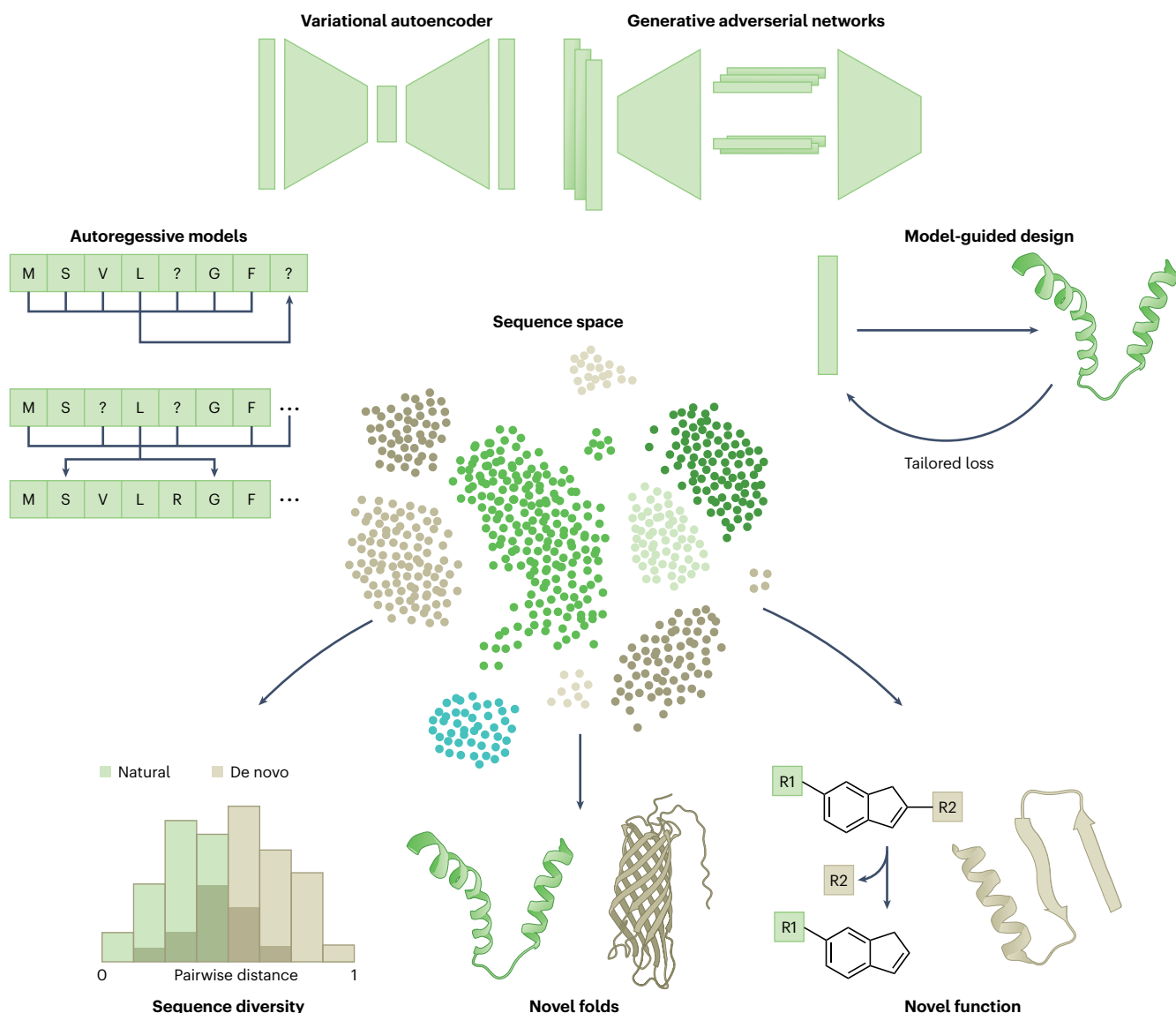


Fig. 6 | Overview of the ML approaches for designing novel enzymes. Popular deep learning frameworks, such as VAEs, generative adversarial networks and autoregressive models, are trained on protein sequences or structures to learn the underlying distribution and sample the sequence space to design distant

functional variants, enzymes with novel folds and new functions. Model-guided design is another approach that leverages ML-based protein structure prediction models for designing enzymes with desired functional sites.

Recent progress in NLP has made it possible to generate functional higher-order mutants with properties resembling those of the native counterparts and even enzymes with novel folds and better attributes. Furthermore, the unsupervised nature of NLP models makes them highly desirable in protein engineering. Generally, there are two categories of ML models: deep generative models and model-guided design.

The initial work in the first category involved the use of a VAE, a deep learning model that maps the training data to an underlying Gaussian distribution. Using the flexibility of the VAE framework, Hawkins-Hooker et al. created two models based on aligned and raw sequence input. A multiple sequence alignment (MSA) VAE is constructed by using MSA as input, while an autoregressive (AR) VAE is a hybrid model that generates functional luciferase-like oxidoreductases with distant variants containing as many as 35 mutations⁶². Repecka et al. used generative adversarial networks (GAN) to design a functional malate dehydrogenase with novel structural domains⁶³. One of the tested variants contained 106 mutations, which corresponds to a

34% change in the protein sequence, thus establishing the approach as a good starting point to test diverse, non-natural sequences for protein engineering. A similar framework, GENhance, was developed to generate highly stable variants of the human angiotensin-converting enzyme 2 protein from less stable variants⁶⁴. The model consists of a generator to sample novel sequences and a discriminator to rank them according to the attribute. It is important to note that such models have been specifically used to generate artificial proteins with a single enzyme parameter such as reactivity or stability. Generative models can also be exploited for the task of enhancing or changing substrate scope. This has recently been demonstrated by training a conditional VAE to design recombinases capable of excising the DNA at novel target sites⁶⁵. Machine translation models provide another such architecture to generate novel enzyme sequences conditioned on the substrate.

Autoregressive models are another popular framework. These models are trained to predict either the next possible amino acid after a

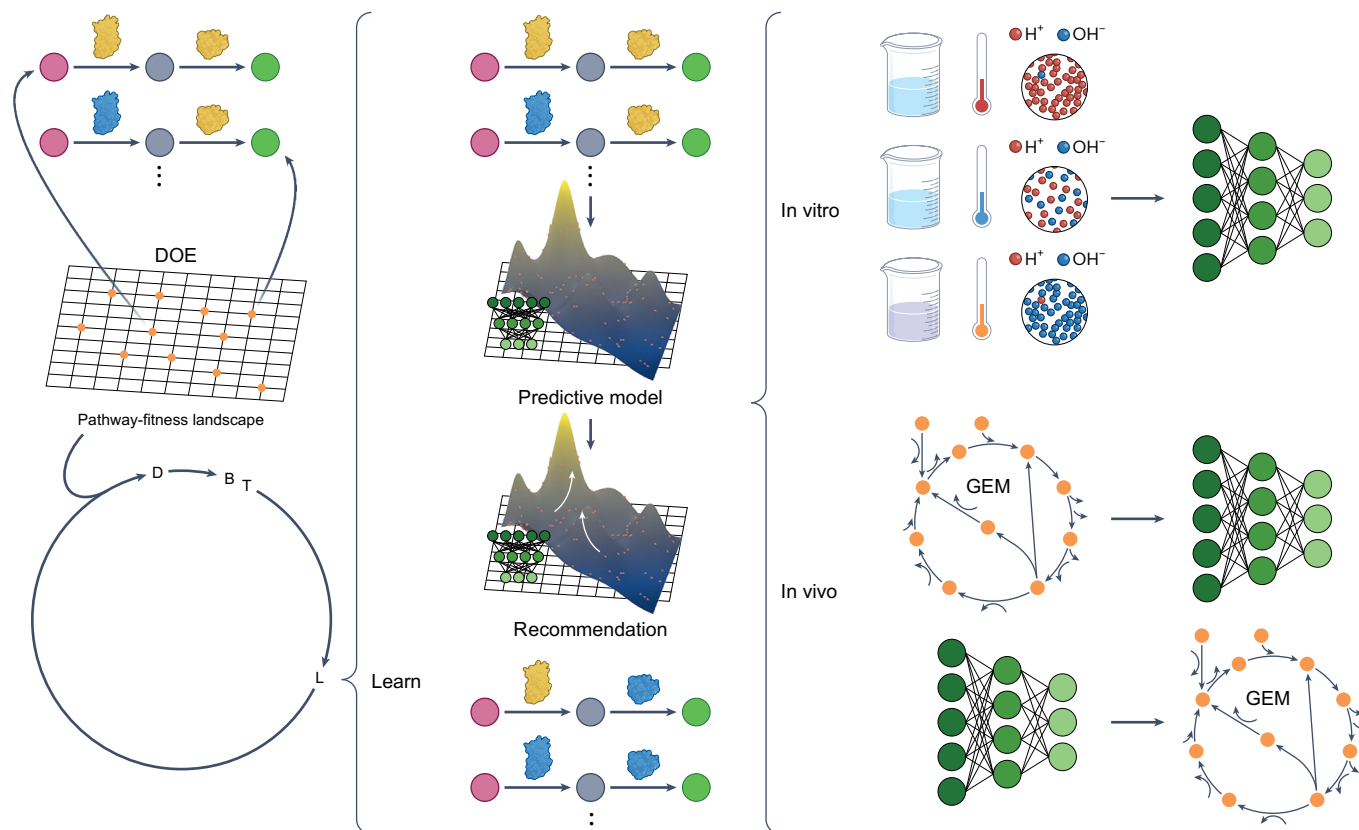


Fig. 7 | Reaction network optimization in vivo and in vitro. Pathway engineering relies on the DBTL cycle to find improved variants. The cycle is initiated by the design (D) of experiment (DOE), variants are built (B) and tested (T), and then learning (L) takes place where a predictive model estimates how variants fall on the pathway fitness landscape. Predictive models can be

further improved with in vitro or in vivo context-specific data. These data create opportunities for other prediction tasks, such as in vitro prediction of reaction conditions and in vitro estimation of k_{cat} . The learn phase of the DBTL cycle finishes with global optimization and recommendation of new pathway designs.

given sequence or the masked or perturbed amino acids in the remaining protein sequence. Madani et al. developed a conditional language model for de novo protein generation⁶⁶. ProGen, a 1.2 billion parameter decoder-style transformer variant, can generate any artificial protein of interest given the tag that specifies the protein metadata. During the training, the model is prepended with metadata, such as protein family, taxonomy and localization. The idea was inspired by the use of input control tags to generate English sentences of a particular style and sentiment. The authors evaluated the model's capabilities by testing artificial enzymes generated with the 'lysozymes' tag: 72/100 proteins expressed well through cell-free protein synthesis with sequence identity to any known natural proteins in the range of 40–90%. A random set of 90 well-expressed artificial proteins were further tested for activity and 73% (66/90) were found to be functional and exhibited high levels of lysozyme activity across families compared with 59% (53/90) of their natural counterparts. Similarly, the state-of-the-art Generative Pre-trained Transformer 2 (GPT2) model was adopted to learn the language of proteins. ProtGPT2 trained on UniRef50, a clustering of UniProt sequences with 50% identity, can generate non-natural protein sequences similar to natural ones⁶⁷. The model can easily generate artificial proteins with versatile folds, multifaceted surfaces and difficult-to-design de novo structures. The most remarkable feature of ProtGPT2 is its ability to expand the sequence space of superfamilies by sampling sequences from the dark proteome, the evolutionarily unexplored regions of the protein space. Models trained on structural information have also been reported in the literature to successfully produce novel folds^{68,69}.

By contrast, model-guided design uses structure prediction models such as RoseTTAFold⁷⁰ to create new-to-nature proteins. Anishchenko et al. developed the concept of protein hallucinations where random amino acid sequences are optimized to fold into distinct three-dimensional (3D) structures by iterative updating based on the gradient of the loss or by Markov chain Monte Carlo sequence optimization⁷¹. Recent work by Wang et al. further constrained the loss function to generate candidates tailored with desired functional sites⁷².

Given the aforementioned examples, ML-guided novel protein design can be used to synthesize protein libraries with huge diversity, novel folds and potentially novel functions to navigate the protein landscape thoroughly (Fig. 6). Detailed reviews of the use of generative models for protein design have recently been published^{73,74}.

Pathway engineering

Once all of its component enzymes are identified, the target biosynthetic pathway must be constructed in vitro or in vivo. To optimize its function in either setting, the pathway itself is typically modelled along with the surrounding environment, that is, the cellular metabolism in the case of in vivo and the reaction conditions in the case of in vitro (Fig. 7). In this section we will discuss and compare the implementation of biosynthetic pathways in vitro and in vivo.

In vivo metabolic pathway optimization

In vivo pathway implementation is typically selected for fermentative processes to take advantage of cell machinery or to pursue a sustainable synthesis process^{75,76}. Most knowledge-driven pathway

optimization in vivo is performed with mechanistic models, specifically constraint-based models (CBMs) such as genome-scale models (GEMs) or kinetic models⁷⁷. High-dimensional omics data are typically used as additional constraints for GEMs and their inclusion has created the greatest improvements in terms of predicting cellular phenotypes and chemical concentrations, but they are still neglected as tools for in vivo pathway design and optimization. At present, the CBM paradigm is not fully compatible with the modern deep learning paradigm in that it is difficult to incorporate backpropagation for end-to-end learning in GEMs and deep learning alone offers little explanatory value for reaction mechanisms. Recently developed methods have integrated CBM and deep learning models by connecting them in series, with GEMs producing the input for the deep learning model or vice versa^{77–79}.

A major scientific challenge in modelling metabolic reactions is the estimation in vivo of the Michaelis–Menten kinetic parameters k_{cat} (the catalytic rate constant) and K_{m} (the Michaelis constant) as they do not typically correlate with in vitro measurements⁸⁰. Accurate prediction of in vivo enzyme parameters would allow for more principled design and engineering of microbial cell factories. Heckmann et al. showed how a model trained on enzyme network context, enzyme structure and enzyme biochemistry with a limited labelled dataset of measured in vivo k_{cat} can be used to extrapolate to 3,000 in vivo k_{cat} values, thereby allowing parameterization of an *E. coli* GEM. When estimating gene expression, the authors showed that ML-extrapolated k_{cat} values could improve $\log_{10}$ root mean squared error of gene expression prediction by 38% compared with 10% for median imputed in vitro k_{cat} values⁸¹. Another method for parameterizing GEMs uses the substrate structure of metabolic enzymes and the enzyme sequence to predict in vivo k_{cat} . Li et al. used a combination of a graph neural network to encode substrates and a CNN to encode proteins to predict in vivo k_{cat} (ref. ⁵³). Model predictions were further improved through Bayesian learning from experimental data. Attention from the model identified amino acid residues that affect enzyme activity, thereby generating targets for protein engineering⁵³. More fundamental work aims to reconstruct the metabolism of under-characterized microbes, which can then be used with the existing CBM framework for pathway optimization^{14,82}. While the reconstruction of metabolism using ML is a promising application, mechanistic models must maintain a high standard of integrity for them to be useful as in vivo pathway optimization tools⁸³.

In terms of individual pathway optimization, the most common techniques rely on the design of experiments, predictive modelling and the recommendation of future designs. In 2019, Hamedirad et al. integrated an ML model with a robotic system to fully automate the design–build–test–learn (DBTL) loop for pathway optimization and used the resultant platform, named BioAutomata, to optimize the biosynthetic pathway for lycopene production⁸⁴. Specifically, a Gaussian process predictive model and Bayesian optimization were used to suggest the designs of future pathway variants and this platform outperformed random screening with the same number of pathway variants by 77% while evaluating less than 1% of all the possible pathway variants. Building on this idea, Radiojević et al. built an Automated Recommendation Tool (ART) that uses an ensemble predictive modelling approach followed by Bayesian optimization to rank future variants with an estimated probability distribution, quantifying the uncertainty of prediction, which can help to assess the feasibility of future wet-lab experiments⁸⁵. A recent implementation of ART combined a knowledge-driven approach using a GEM with a DBTL loop to engineer tryptophan overproduction. The GEM helped to identify promising gene targets to limit the experimental design space and ART was used for DBTL optimization, which led to improvements of the tryptophan titre by 74% and productivity by 43% compared with the first DBTL cycle⁸⁶. DBTL has also been applied to the overproduction of violacein⁸⁷, 1-dodecanol⁸⁸ and monoterpenoids⁸⁹.

In vitro pathway optimization

In vitro enzymatic pathways, often called multistep or cascading enzymatic reactions, can be implemented for large-scale synthesis of target molecules. Compared with in vivo synthesis, in vitro synthesis benefits from a more controlled environment and fewer constraints, for example, cell survival and energy maintenance, which can allow greater conversion rates. Key design choices include a target synthesis route, typically selected with retrobiosynthesis tools, choices of enzymes, reaction conditions and process design. Enzyme cascades can be performed sequentially, but ideally they are performed in a one-pot reaction to reduce the number of isolation steps. A one-pot reaction faces challenges of inhibitory interactions⁹⁰, incompatible reaction conditions⁹¹ and enzyme promiscuity⁹². Traditionally, optimization schemes have attempted to model system dynamics with mechanistic models such as kinetic models or data-driven methods, for example, support vector machine (SVM), Gaussian process or artificial neural network (ANNs), and then suggest future variants with genetic algorithms or Bayesian optimization^{84,93}. In a recent example, Wan et al. used a quadratic SVM model trained on quantum chemistry reactivity descriptors along with reaction conditions to predict the yield and corresponding reaction conditions⁹⁴. It is worth noting that many of these approaches have largely ignored the representation of enzyme structure, especially if proteins are no longer being engineered in the DBTL pipeline. In principle, any of the reaction condition optimization techniques used for organic synthesis can be used for biosynthesis by representing enzymes with a categorical variable, similar to how solvents are often represented as categorical variables for organic synthesis⁹⁴. There has been recent success in predicting organic synthesis reaction conditions, including catalysts, solvents, reagents and temperature⁹⁵, but the limited data on biocatalytic reaction conditions makes it difficult to adopt these methods in biocatalysis⁹⁶.

While in vitro and in vivo optimization have traditionally been performed separately, the lines are beginning to blur with cell-free systems. In 2020, Karim et al. developed a system called iProbe that uses the flexibility of the cell-free system to test a combinatorial design space that would be infeasible to test in vivo⁹⁰. From the initial screening of 120 pathway combinations, 43 pathways predicted from a group of ANNs proved to give better performance than expert design. Pathways optimized in vitro were then transformed in vivo, retaining the performance with a Pearson correlation of 0.79, showing that under proper conditions, the benefits of in vitro optimization can be ported to the in vivo setting⁹⁶. Furthermore, this method has been used to scale up industrial fermentation of C_1 waste to yield acetone and isopropanol⁷⁵.

Pathway inhibition

Another task of immediate interest to in vitro and in vivo optimization is the prediction of compound–protein interactions (CPIs) and protein–protein interactions (PPIs), which could be inhibitory to the target biosynthetic pathways. The prediction of drug–protein interactions (DPIs) has attracted the most attention in CPI research because DPI models can be used to identify inhibitory ligand binding. Common tasks include the prediction of binding sites, protein–ligand binding affinity and protein–ligand binding conformation⁵⁴. Gainza et al. learnt protein interaction fingerprints with topological and chemical features. The model generating the interaction fingerprint was combined with an application-specific layer for various prediction tasks, including active site classification and ligand or protein interaction prediction⁹⁷. In the in vivo setting, combinations of DPI and PPI data have helped to identify important network effects⁹⁸ and the toxicity of new compounds⁹⁹. Counter to the idea that including all complex interaction data improves performance, Goldman et al. demonstrated that single-task models, deemed less generalizable, can outperform more complex models that attempt to include all interaction data. This result shows a need for new representation learning methods for extrapolation to unseen CPIs⁴⁵.

Table 1 | List of databases commonly used to develop ML models for retrobiosynthesis

Classification	Name	Description	Size	Representative ML models
Enzymes	UniProt ⁶	A comprehensive, high-quality and freely accessible resource of protein sequence and functional information	~120 million proteins with ~570,000 reviewed entries	ProteinInfer ⁴⁰ , DeepEC ³⁷ , UniRep ¹³ , AlphaFold2 ¹²¹ , RoseTTAfold ⁷⁰ , MSA VAE ⁶² , AR VAE ⁶² , ProteinGAN ⁶³ , ProGen ⁶⁶ , ProtGTP2 ⁶⁷ , CAThe ³⁹ , ECNet ⁵⁷ , DeepLoc ⁵⁰ and 3DCNNs ¹²²
	ExPASy-ENZYME ¹⁰⁰	Describes each type of characterized enzyme with the associated EC number	~7,000 active entries	
	ExPASy-PROSITE ¹²⁴	Describes protein domains, families and functional sites	~1,900 entries, ~1,300 patterns, ~1,300 profiles and ~1,300 prorules	
	Protein Data Bank ¹²⁵	Describes the 3D structures of proteins, nucleic acids and complex assemblies	~190,000 biological macromolecular structures	
	ProtaBank ¹⁰¹	A central repository to store, query, analyse and share all types of protein design and protein engineering data	~7.7 million data points covering 1.8 million protein variants and 15,000 assays	
	ProtDataTherm ¹⁰²	A database focusing on analysing and engineering protein thermostability	>14 million protein sequences	
	FireProtDB ¹⁰³	A comprehensive and manually curated database of protein stability information for single mutants	~6,700 mutants covering ~242 proteins and ~16,000 experiments	
Substrates	ChEBI ¹⁰⁴	A database and ontology containing information on small chemical compounds of biological interest	~60,000 fully annotated compounds	DeepCSeqSite ¹²⁶ , DELIA ¹²⁷ , Kalasanty ¹²⁸ , DeepSurf ¹²⁹ and PURESNet ¹³⁰
	HMDB ¹⁰⁵	A freely available database comprising detailed information on small-molecule metabolites found in the human body	~217,000 compounds	
	LMDS ¹⁰⁶	A relational database encompassing structures and annotations of biologically relevant lipids	~47,000 unique lipid structures with 25,457 curated ones	
	SwissLipids ¹⁰⁷	An expert-curated resource of lipids and their biology providing biological information on lipid and lipidomic structures and metabolism	~780,000 lipid species and ~7,000 distinct pieces	
Reactions	BRENDA ¹⁰⁸	A main enzyme and enzyme–ligand information system comprising disease relevant data, enzyme sequences, 3D structures, predicted enzyme locations and genome annotations	~4.3 million data for ~84,000 enzymes belonging to ~7,600 enzyme classes	EPP-HMCNF ⁴⁴ , K_m prediction ¹³¹ and k_{cat} prediction ⁵³
	Rhea ¹⁰⁹	A comprehensive and non-redundant resource of expert-curated biochemical reactions covering the reactions of all EC numbers as well as thousands of additional enzymatic reactions, transport reactions and spontaneously occurring reactions	~14,000 reactions with ~12,000 unique compounds	
	BioCatNet ¹¹⁰	A repository of sequence, structure and biocatalytic experiments for a given enzyme family	12 enzyme families with ~55,000 sequences and ~2,000 3D structures	
	SABIO-RK ¹¹¹	A curated database containing structured information on biochemical reactions, kinetic rate equations with parameters and experimental conditions	~72,000 curated entries with ~56,000 K_m data, ~53,000 velocity constants and ~16,000 inhibition constants	
	BKMS-react ^{108,112}	An integrated non-redundant reaction database containing known enzyme-catalysed and spontaneous reactions collected from BRENDA, KEGG, MetaCyc and SABIO-RK	~41,000 reactions with EC numbers	
	ECREACT ⁹	A simplified enzymatic reaction database with EC numbers; the data were extracted from four databases, namely Rhea, BRENDA, PathBank and MetaNetX	~62,000 reactions with EC numbers	
	KEGG ¹¹³	A knowledge database for systematic analysis of gene functions linking genomic information with higher-structured functional information	~16,000 enzymes with ~2,000 reactions	Metabolic Allele Classifier ¹³² , DeepRFC ²¹ , GC-ANN ¹³³ , RetroPath2.0 ²⁰ , BioNavi-NP ¹⁶ , Evo-DoE ¹³⁴ and feasible-metabolic-pathway exploration ¹⁷

Table 1 (continued) | List of databases commonly used to develop ML models for retrobiosynthesis

Classification	Name	Description	Size	Representative ML models
	enviPath ¹¹⁴	A database and prediction system for the microbial biotransformation of organic environmental contaminants	>1,500 microbial catabolic reactions and ~220 biotransformation pathways	
	BiGG ¹¹⁵	A knowledge base of genome-scale metabolic network reconstructions	~70 published genome-scale metabolic networks	
	Reactome ¹¹⁶	An open-source and peer-reviewed pathway database providing intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge	~21,000 pathways covering 15 species and 88,000 proteins	
	PathBank ¹¹⁷	A comprehensive, visually rich database providing a pathway for every protein and a map for every metabolite	~110,000 machine-readable pathways found in 10 model organisms	
	MetaCyc ¹¹⁸	An evidence-based and richly curated database of metabolic pathways and enzymes from all domains of life	>2,600 pathways	
	MetaNetX ¹¹⁹	A website for accessing, analysing and manipulating genome-scale metabolic networks and biochemical pathways		

Common databases used for ML model development

The development of ML models relies on high-quality databases. As briefly mentioned above and summarized in Table 1, a wide variety of databases have been used to develop ML models for retrobiosynthesis that contain information on enzyme sequences^{6,100–103}, substrates^{104–107}, chemical reactions^{9,108–112}, and metabolic pathways and/or networks^{113–119}. However, most databases are not initially designed to develop ML models and few have incorporated protein mutagenesis data (a notable exception is Protobank¹⁰¹), which is necessary to develop ML models for enzyme engineering. Moreover, an ML-friendly database needs to be actively maintained and should have comprehensive coverage, extensive annotation and evidence scores to ensure that the data are of high quality and contain minimal false positives¹⁰⁷. Therefore, there is a need to develop standardized databases by implementing cross-comparison between databases to reduce redundant or inconsistent information, steady maintenance to capture the ever-growing depth of biology, and easily accessible and user-friendly databases.

Future perspectives and conclusion

ML has made a great impact on every aspect of retrobiosynthesis, including synthesis planning, enzyme selection and engineering, and pathway optimization, both in vitro and in vivo¹²⁰. However, opportunities still exist to further develop ML models for retrobiosynthesis. For example, current retrosynthesis tools rarely consider both chemo- and biocatalysis in the design of synthetic routes. Moreover, one of the tougher challenges in the biosynthesis of a target molecule comes from the missing links in the biosynthetic pathway predicted by retrobiosynthesis tools. The missing links can be patched by predicting enzyme–substrate interactions, which may indeed be challenging from sequence information alone. With the recent development of protein structure prediction tools, such as AlphaFold¹²¹ and trRosettaFold⁷⁰, combined with catalytic site prediction tools^{40,122}, new ML models using structural information could achieve the interpretation of enzyme–substrate interactions. Alternatively, the use of semantically rich, conditional language models to sample artificial protein sequences for new reactions and further engineering using MLDE will potentially pave the way for the successful design of desired enzymes. As of now, the integration of existing ML tools assists in GEM reconstruction, the parameterization of GEMs for pathway simulation, the identification of target amino acid residues in individual proteins for site-specific mutagenesis, the identification of interfering small molecules and

DBTL optimization over a combination of pathways. The major pieces supporting in vivo and in vitro biocatalytic vision are in place, and we expect to see incremental improvements in isolated prediction tasks and new efforts in model integration. Learning mappings between the in vitro and in vivo settings would further expedite this vision. To establish a comprehensive database for ML model development, the scientific community needs to continually develop a global collection of enzymes, reactions and pathways. There is much room for improvement in automatic data mining (for example, visualization mining in figures and text mining for various information) and data reliability. With the rapid emergence of ML models, a standardized benchmarking database for model development, evaluation and validation is needed. An initial effort in this direction has been made by Dallago et al., who designed Fitness Landscape Inference for Proteins (FLIP) to benchmark ML models for protein engineering tasks¹²³. Such a platform enables rapid scoring and assessment of models, and similar datasets can be designed for other tasks, such as retrobiosynthesis planning. As researchers continue to explore biocatalysts, ML will become an irreplaceable tool with which to expand the boundary of molecule synthesis.

References

- Lin, G.-M., Warden-Rothman, R. & Voigt, C. A. Retrosynthetic design of metabolic pathways to chemicals not found in nature. *Curr. Opin. Syst. Biol.* **14**, 82–107 (2019).
- Bornscheuer, U. T. et al. Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).
- Sheldon, R. A. & Woodley, J. M. Role of biocatalysis in sustainable chemistry. *Chem. Rev.* **118**, 801–838 (2018).
- de Souza, R. O. M. A., Miranda, L. S. M. & Bornscheuer, U. T. A retrosynthesis approach for biocatalysis in organic synthesis. *Chem. Eur. J.* **23**, 12040–12063 (2017).
- Turner, N. J. & O'Reilly, E. Biocatalytic retrosynthesis. *Nat. Chem. Biol.* **9**, 285–288 (2013).
- The UniProt Consortium UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
- Khan, A. Z., Bilal, M., Rasheed, T. & Iqbal, H. M. N. Advancements in biocatalysis: from computational to metabolic engineering. *Chin. J. Catal.* **39**, 1861–1868 (2018).
- Feehan, R., Montezano, D. & Slusky, J. S. G. Machine learning for enzyme engineering, selection and design. *Protein Eng. Des. Sel.* **34**, gzab019 (2021).

9. Probst, D. et al. Biocatalysed synthesis planning using data-driven learning. *Nat. Commun.* **13**, 964 (2022).
This paper describes the development of a template-free retrobiosynthesis tool by training a molecular transformer with multi-task transfer learning using both enzymatic and chemical reaction databases.
10. Hie, B. L. & Yang, K. K. Adaptive machine learning for protein engineering. *Curr. Opin. Struct. Biol.* **72**, 145–152 (2022).
11. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
12. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
13. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
14. Wang, L., Dash, S., Ng, C. Y. & Maranas, C. D. A review of computational tools for design and reconstruction of metabolic pathways. *Synth. Syst. Biotechnol.* **2**, 243–252 (2017).
15. Koch, M., Duigou, T. & Faulon, J.-L. Reinforcement learning for bioretrosynthesis. *ACS Synth. Biol.* **9**, 157–168 (2020).
16. Zheng, S. et al. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. *Nat. Commun.* **13**, 3342 (2022).
This paper introduces a useful retrobiosynthesis tool for navigating biosynthetic pathways to complex natural products from simple building blocks.
17. Fuji, T., Nakazawa, S. & Ito, K. Feasible-metabolic-pathway-exploration technique using chemical latent space. *Bioinformatics* **36**, i770–i778 (2020).
18. Finnigan, W., Hepworth, L. J., Flitsch, S. L. & Turner, N. J. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. *Nat. Catal.* **4**, 98–104 (2021).
19. Kumar, A., Wang, L., Ng, C. Y. & Maranas, C. D. Pathway design using de novo steps through uncharted biochemical spaces. *Nat. Commun.* **9**, 184 (2018).
20. Delépine, B. et al. RetroPath2.0: a retrosynthesis workflow for metabolic engineers. *Metab. Eng.* **45**, 158–170 (2018).
21. Kim, Y., Ryu, J. Y., Kim, H. U., Jang, W. D. & Lee, S. Y. A deep learning approach to evaluate the feasibility of enzymatic reactions generated by retrobiosynthesis. *Biotechnol. J.* **16**, 2000605 (2021).
22. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **28**, 31–36 (1988).
23. Pesciullesi, G., Schwaller, P., Laino, T. & Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.* **11**, 4874 (2020).
24. Hasic, H. & Ishida, T. Single-step retrosynthesis prediction based on the identification of potential disconnection sites using molecular substructure fingerprints. *J. Chem. Inf. Model.* **61**, 641–652 (2021).
25. Somnath, V. R., Bunne, C., Coley, C., Krause, A. & Barzilay, R. Learning graph models for retrosynthesis prediction. In *Proc. 34th Advances in Neural Information Processing Systems* (eds Ranzato, M. et al.) 9405–9415 (Curran Associates, Inc., 2021).
26. Wang, H. et al. Chemical-reaction-aware molecule representation learning. Preprint at <https://arxiv.org/abs/2109.09888> (2021).
27. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
28. Tempke, R. & Musho, T. Autonomous design of new chemical reactions using a variational autoencoder. *Commun. Chem.* **5**, 40 (2022).
29. Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *J. Chem. Inf. Model.* **60**, 47–55 (2020).
30. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).
31. Segler, M. H. S. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry* **23**, 5966–5971 (2017).
32. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
33. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
34. Chen, B., Li, C., Dai, H. & Song, L. Retro*: learning retrosynthetic planning with neural guided A* search. In *Proc. 37th International Conference on Machine Learning* (eds Daumé, H. III & Singh, A) 1608–1616 (PMLR, 2020).
35. Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531 (1994).
36. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
37. Ryu, J. Y., Kim, H. U. & Lee, S. Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc. Natl Acad. Sci. USA* **116**, 13996–14001 (2019).
38. Price, M. N. et al. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).
39. Nallapareddy, V. et al. CATHe: detection of remote homologues for CATH superfamilies using embeddings from protein language models. *Bioinformatics* **39**, btad029 (2022).
40. Sanderson, T., Bileschi, M. L., Belanger, D. & Colwell, L. J. ProteInfer: deep networks for protein functional inference. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.09.20.461077> (2021).
This paper reports a state-of-the-art ML-based protein annotation tool capable of predicting both EC number and Gene Ontology (GO) from amino acid sequences.
41. Heinzinger, M. et al. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genom. Bioinform.* **4**, lqac043 (2022).
42. Carbonell, P. et al. Selenzyme: enzyme selection tool for pathway design. *Bioinformatics* **34**, 2153–2154 (2018).
43. Cho, A., Yun, H., Park, J. H., Lee, S. Y. & Park, S. Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Syst. Biol.* **4**, 35 (2010).
44. Visani, G. M., Hughes, M. C. & Hassoun, S. Enzyme promiscuity prediction using hierarchy-informed multi-label classification. *Bioinformatics* **37**, 2017–2024 (2021).
45. Goldman, S., Das, R., Yang, K. K. & Coley, C. W. Machine learning modeling of family wide enzyme–substrate specificity screens. *PLoS Comput. Biol.* **18**, e1009853 (2022).
46. Xu, Z., Wu, J., Song, Y. S. & Mahadevan, R. Enzyme activity prediction of sequence variants on novel substrates using improved substrate encodings and convolutional pooling. In *Proc. 16th Machine Learning in Computational Biology meeting* (eds Knowles, D. A. et al) 78–87 (PMLR, 2022).
47. Musil, M., Konegger, H., Hon, J., Bednar, D. & Damborsky, J. Computational design of stable and soluble biocatalysts. *ACS Catal.* **9**, 1033–1054 (2019).

48. Hon, J. et al. SoluProt: prediction of soluble protein expression in *Escherichia coli*. *Bioinformatics* **37**, 23–28 (2021).
49. Li, G., Rabe, K. S., Nielsen, J. & Engqvist, M. K. M. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS Synth. Biol.* **8**, 1411–1420 (2019).
50. Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**, 3387–3395 (2017).
51. Stärk, H., Dallago, C., Heinzinger, M. & Rost, B. Light attention predicts protein location from the language of life. *Bioinform. Adv.* **1**, vbab035 (2021).
52. Chai, M. et al. Application of machine learning algorithms to estimate enzyme loading, immobilization yield, activity retention, and reusability of enzyme–metal–organic framework biocatalysts. *Chem. Mater.* **33**, 8666–8676 (2021).
53. Li, F. et al. Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction. *Nat. Catal.* **5**, 662–672 (2022).
54. Dhakal, A., McKay, C., Tanner, J. J. & Cheng, J. Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. *Brief. Bioinform.* **23**, bbab476 (2022).
55. Wang, Y. et al. Directed evolution: methodologies and applications. *Chem. Rev.* **121**, 12384–12444 (2021).
56. Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol.* **69**, 11–18 (2021).
57. Luo, Y. et al. ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nat. Commun.* **12**, 5743 (2021).
58. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-*N* protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).
59. Hsu, C. et al. Learning protein fitness models from evolutionary and assay-labeled data. *Nat. Biotechnol.* **40**, 1114–1122 (2022).
This paper reports an in depth evaluation and discussion of ML models predicting variant effects under a low-*N* situation.
60. Wittmann, B. J., Yue, Y. & Arnold, F. H. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst.* **12**, 1026–1045.e7 (2021).
61. Greenhalgh, J. C., Fahlberg, S. A., Pflieger, B. F. & Romero, P. A. Machine learning-guided acyl-ACP reductase engineering for improved in vivo fatty alcohol production. *Nat. Commun.* **12**, 5825 (2021).
62. Hawkins-Hooker, A. et al. Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol.* **17**, e1008736 (2021).
63. Repecka, D. et al. Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **3**, 324–333 (2021).
64. Chan, A., Madani, A., Krause, B. & Naik, N. Deep extrapolation for attribute-enhanced generation. In *Proc. 34th Advances in Neural Information Processing Systems* (eds Ranzato, M. et al.) 14084–14096 (Curran Associates, Inc., 2021).
65. Schmitt, L. et al. Prediction of designer-recombinases for DNA editing with generative deep learning. *Nat. Commun.* **13**, 7966 (2022).
66. Madani, A. et al. Deep neural language modeling enables functional protein generation across families. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.07.18.452833> (2021).
Using a global language model trained on protein sequences and annotations, the authors demonstrate a universal generative model capable of generating protein sequences with desired properties with a varying degree of sequence similarity.
67. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
68. Anand, N. et al. Protein sequence design with a learned potential. *Nat. Commun.* **13**, 746 (2022).
69. Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. In *Proc. 32nd Advances in Neural Information Processing Systems* (eds Wallach, H. et al.) (Curran Associates, Inc., 2019).
70. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
71. Anishchenko, I. et al. De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
72. Wang, J. et al. Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
73. Wu, Z., Johnston, K. E., Arnold, F. H. & Yang, K. K. Protein sequence design with deep generative models. *Curr. Opin. Chem. Biol.* **65**, 18–27 (2021).
74. Strokach, A. & Kim, P. M. Deep generative modeling for protein design. *Curr. Opin. Struct. Biol.* **72**, 226–236 (2022).
An insightful review of various deep learning approaches to protein design.
75. Liew, F. E. et al. Carbon-negative production of acetone and isopropanol by gas fermentation at industrial pilot scale. *Nat. Biotechnol.* **40**, 335–344 (2022).
This paper describes the in vitro machine learning screening method iPROBE used to engineer *Clostridium autoethanogenum* for the overproduction of acetone and isopropanol at the industrial scale.
76. Sun, X., Xu, Y. & Huang, H. Thraustochytrid cell factories for producing lipid compounds. *Trends Biotechnol.* **39**, 648–650 (2021).
77. Antonakoudis, A., Barbosa, R., Kotidis, P. & Kontoravdi, C. The era of big data: genome-scale modelling meets machine learning. *Comput. Struct. Biotechnol. J.* **18**, 3287–3300 (2020).
78. Kim, Y., Kim, G. B. & Lee, S. Y. Machine learning applications in genome-scale metabolic modeling. *Curr. Opin. Syst. Biol.* **25**, 42–49 (2021).
79. Zampieri, G., Vijayakumar, S., Yaneske, E. & Angione, C. Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput. Biol.* **15**, e1007084 (2019).
80. Heckmann, D. et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nat. Commun.* **9**, 5252 (2018).
81. Heckmann, D. et al. Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. *Proc. Natl Acad. Sci. USA* **117**, 23182–23190 (2020).
This paper reports fluxomic and proteomic data for estimating in vivo k_{cat} values that were used to parameterize a metabolic model that could then be used for more accurate gene expression prediction.
82. Shah, H. A., Liu, J., Yang, Z. & Feng, J. Review of machine learning methods for the prediction and reconstruction of metabolic pathways. *Front. Mol. Biosci.* **8**, 634141 (2021).
83. Fang, X., Lloyd, C. J. & Palsson, B. O. Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nat. Rev. Microbiol.* **18**, 731–743 (2020).
84. Hamedirad, M. et al. Towards a fully automated algorithm driven platform for biosystems design. *Nat. Commun.* **10**, 5150 (2019).
85. Radivojević, T., Costello, Z., Workman, K. & Garcia Martin, H. A machine learning automated recommendation tool for synthetic biology. *Nat. Commun.* **11**, 4879 (2020).

86. Zhang, J. et al. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nat. Commun.* **11**, 4880 (2020).
87. Zhou, Y. et al. MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in *Saccharomyces cerevisiae*. *Metab. Eng.* **47**, 294–302 (2018).
88. Opgenorth, P. et al. Lessons from two design–build–test–learn cycles of dodecanol production in *Escherichia coli* aided by machine learning. *ACS Synth. Biol.* **8**, 1337–1351 (2019).
89. Jervis, A. J. et al. Machine learning of designed translational control allows predictive pathway optimization in *Escherichia coli*. *ACS Synth. Biol.* **8**, 127–136 (2019).
90. Karim, A. S. et al. In vitro prototyping and rapid optimization of biosynthetic enzymes for cell design. *Nat. Chem. Biol.* **16**, 912–919 (2020).
91. Huffman, M. A. et al. Design of an in vitro biocatalytic cascade for the manufacture of islatravir. *Science* **366**, 1255–1259 (2019).
92. Peters, R. J. R. W. et al. Cascade reactions in multicompartmentalized polymersomes. *Angew. Chem. Int. Ed.* **126**, 150–154 (2014).
93. Nobeli, I., Favia, A. D. & Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**, 157–167 (2009).
94. Wan, Z., Wang, Q.-D., Liu, D. & Liang, J. Accelerating the optimization of enzyme-catalyzed synthesis conditions via machine learning and reactivity descriptors. *Org. Biomol. Chem.* **19**, 6267–6273 (2021).
95. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
96. Gao, H. et al. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* **4**, 1465–1476 (2018).
97. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
98. Morselli Gysi, D. et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc. Natl Acad. Sci. USA* **118**, e2025581118 (2021).
99. Mayr, A., Klambauer, G., Unterthiner, T. & Hochreiter, S. DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* **3**, 80 (2016).
100. Gasteiger, E. ExpASY: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788 (2003).
101. Wang, C. Y. et al. ProtaBank: a repository for protein design and engineering data. *Protein Sci.* **27**, 1113–1124 (2018).
102. Pezeshgi Modarres, H., Mofrad, M. R. & Sanati-Nezhad, A. ProtDataTherm: a database for thermostability analysis and engineering of proteins. *PLoS ONE* **13**, e0191222 (2018).
103. Stourac, J. et al. FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res.* **49**, D319–D324 (2021).
104. Hastings, J. et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–D1219 (2016).
105. Wishart, D. S. et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
106. Sud, M. et al. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **35**, D527–D532 (2007).
107. Aimo, L. et al. The SwissLipids knowledgebase for lipid biology. *Bioinformatics* **31**, 2860–2866 (2015).
108. Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.* **47**, D542–D549 (2018).
109. Lombardot, T. et al. Updates in Rhea: SPARQLing biochemical reaction data. *Nucleic Acids Res.* **47**, D596–D600 (2019).
110. Buchholz, P. C. F. et al. BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments. *ChemBioChem* **17**, 2093–2098 (2016).
111. Wittig, U., Rey, M., Weidemann, A., Kania, R. & Müller, W. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. *Nucleic Acids Res.* **46**, D656–D660 (2018).
112. Lang, M., Stelzer, M. & Schomburg, D. BKM-react, an integrated biochemical reaction database. *BMC Biochem.* **12**, 42 (2011).
113. Kanehisa, M. et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2007).
114. Wicker, J. et al. enviPath—the environmental contaminant biotransformation pathway resource. *Nucleic Acids Res.* **44**, D502–D508 (2016).
115. King, Z. A. et al. BiGG models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* **44**, D515–D522 (2016).
116. Gillespie, M. et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).
117. Wishart, D. S. et al. PathBank: a comprehensive pathway database for model organisms. *Nucleic Acids Res.* **48**, D470–D478 (2020).
118. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* **46**, D633–D639 (2017).
119. Moretti, S., Tran, V. D. T., Mehl, F., Ibberson, M. & Pagni, M. MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models. *Nucleic Acids Res.* **49**, D570–D574 (2021).
120. Mazurenko, S., Prokop, Z. & Damborsky, J. Machine learning in enzyme engineering. *ACS Catal.* **10**, 1210–1223 (2020).
121. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
122. Torng, W. & Altman, R. B. High precision protein functional site detection using 3D convolutional neural networks. *Bioinformatics* **35**, 1503–1512 (2019).
123. Dallago, C. et al. FLIP: benchmark tasks in fitness landscape inference for proteins. Preprint at *bioRxiv* <https://www.biorxiv.org/content/10.1101/2021.11.09.467890v2> (2022).
124. Hulo, N. The PROSITE database. *Nucleic Acids Res.* **34**, D227–D230 (2006).
125. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
126. Cui, Y., Dong, Q., Hong, D. & Wang, X. Predicting protein–ligand binding residues with deep convolutional neural networks. *BMC Bioinform.* **20**, 93 (2019).
127. Xia, C.-Q., Pan, X. & Shen, H.-B. Protein–ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data. *Bioinformatics* **36**, 3018–3027 (2020).
128. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P. & Siedlecki, P. Improving detection of protein–ligand binding sites with 3D segmentation. *Sci. Rep.* **10**, 5035 (2020).
129. Mylonas, S. K., Axenopoulos, A. & Daras, P. DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics* **37**, 1681–1690 (2021).
130. Kandel, J., Tayara, H. & Chong, K. T. PURESNet: prediction of protein–ligand binding sites using deep residual neural network. *J. Cheminform.* **13**, 65 (2021).
131. Kroll, A., Engqvist, M. K. M., Heckmann, D. & Lercher, M. J. Deep learning allows genome-scale prediction of Michaelis constants from structural features. *PLoS Biol.* **19**, e3001402 (2021).

132. Kavvas, E. S., Yang, L., Monk, J. M., Heckmann, D. & Palsson, B. O. A biochemically interpretable machine learning classifier for microbial GWAS. *Nat. Commun.* **11**, 2580 (2020).
133. Ajjolli Nagaraja, A. et al. A machine learning approach for efficient selection of enzyme concentrations and its application for flux optimization. *Catalysts* **10**, 291 (2020).
134. Caschera, F. et al. Coping with complexity: machine learning optimization of cell-free protein synthesis. *Biotechnol. Bioeng.* **108**, 2218–2228 (2021).

Acknowledgements

This work was supported by the Molecule Maker Lab Institute, an AI Research Institutes programme supported by the US National Science Foundation under grant no. 2019897 (H.Z.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the NSF.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Huimin Zhao.

Peer review information *Nature Catalysis* thanks William Finnigan, Pablo Carbonell and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2023