Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Deep learning and generative artificial intelligence methods in enzyme and cell engineering

Steffen Docter[1,*], Benoit David[1,*] and Holger Gohlke[1,2,3]

Efficient enzymes and microbial factories are essential to promote the transition toward a sustainable bioeconomy. This review focuses on the progress of artificial intelligence (AI) methods in accelerating the development of optimized biocatalysts and genetic networks in cells. Recent advances in AI in the field of enzyme discovery, engineering, and *de novo* design are discussed. Additionally, we highlight examples of successful applications of AI in optimizing different components in cells, from gene expression regulation to metabolic pathway optimization and design. Finally, this review emphasizes the challenges limiting the reliability and generalizability of current AI methods.

Addresses
[1] Institute of Bio- and Geosciences (IBG-4: Bioinformatics), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
[2] Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany
[3] Bioeconomy Science Center (BioSC), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

Corresponding author: Gohlke, Holger (h.gohlke@fz-juelich.de, gohlke@uni-duesseldorf.de)
[*] These authors contributed equally to this work.

Current Opinion in Biotechnology 2026, **97**:103393

This review comes from a themed issue on **Chemical Biotechnology**

Edited by **Jochen Förster** and **Stephan Noack**

For complete overview of the section, please refer to the article collection, "Chemical Biotechnology (2026)"

Available online 4 December 2025

https://doi.org/10.1016/j.copbio.2025.103393

Introduction

The development of circular, bio-based processes that yield fine chemicals, enzyme biocatalysts, and engineered organisms, or remove environmental waste, is crucial for the global endeavor to move toward a sustainable bioeconomy [1]. Artificial intelligence (AI) methods can accelerate this transition by reducing the experimental effort for the development of optimized strains and biocatalysts. Here, we highlight recent advances in the applications of AI methods in designing and optimizing enzymes as biocatalysts and in cell engineering.

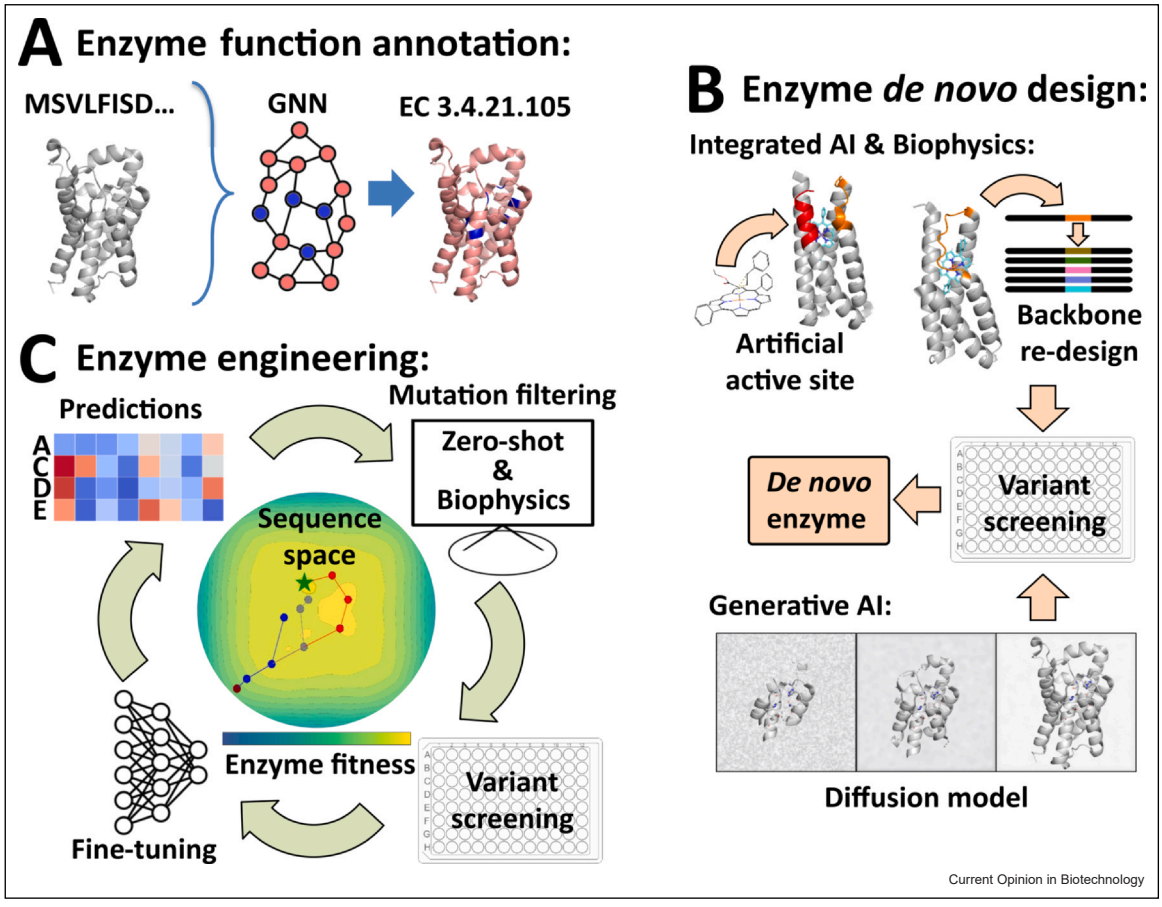
Artificial intelligence methods in enzyme discovery and engineering

Accelerating enzyme discovery

While millions of protein sequences are known, only a fraction have experimental function annotations such as Enzyme Commission (EC) numbers. In 2023, CLEAN (contrastive learning-enabled enzyme annotation) [2], a model trained on enzyme sequences, outperformed the then state-of-the-art, sequence homology-based BLASTp [3] method on the Price-149 benchmark [4] of 149 challenging enzyme sequence-function pairs, achieving an F1 score of 0.495 versus 0.385. This demonstrated that machine learning can surpass traditional similarity searches in EC prediction. In 2024, CLEAN was upgraded to CLEAN-contact [5], which uses structure-based residue contact maps interpreted via the ResNet-50 computer vision model [6] and the protein language model (PLM) ESM-1b [7], improving prediction performance (F1 score of 0.525 on Price-149).

Graph neural networks (GNNs) further integrate structural data into AI training, as exemplified by GraphEC [8] and TopEC [9] (Figure 1). Both models follow similar approaches of using protein structures predicted from sequence to build graph representations. Initially, GraphEC predicts the enzyme's active site within its model framework, while TopEC employs P2Rank [10]. Both then predict EC numbers based on the feature importance assigned to the active site residues. GraphEC further aims to predict the optimum pH value of the identified enzymatic reaction; however, due to low data availability, this was only evaluated on roughly 4000 enzymes, with 52% of optimum pH values at 7–8. GraphEC outperforms CLEAN on the Price-149 benchmark (F1 score of 0.613), while TopEC performs on par with CLEAN on the custom, larger benchmark TopEnzyme [11] developed by the TopEC authors. This is notable, as TopEC was evaluated with both a 30% sequence similarity cutoff and a structure-based data split using FoldSeek clusters [12] to prevent data leakage between training and test sets. GraphEC and CLEAN use more permissive cutoffs at 40% and 50% sequence identity, respectively. This increases the likelihood of homologous proteins occurring in the training and test sets, artificially boosting the models' prediction performance [13,14]. Besides the increasing availability of training data, progress in EC number predictions can also be expected from the refinement of protein embeddings used for model training. A recent preprint shows that simple multilayer-perceptron models trained

Figure 1



Workflows using AI methods in enzyme discovery and engineering. **(a)** Enzymes suitable for specific biocatalytic applications can be identified via AI-based function predictions for unannotated sequences. **(b)** They can also be designed *de novo* by inserting artificial active sites into protein scaffolds or by designing novel backbones around defined active site residues and ligands. **(c)** For the optimization of enzyme properties for biocatalysis, AI can be integrated into directed evolution workflows to learn from experimental screening results and predict new mutations, or be applied for computational filtering of predicted mutations to minimize experimental efforts.

on embeddings from the PLM ESM2 [7] or the multimodal models ProTrek [15] and OneProt can outperform CLEAN and TopEC in enzyme classification [16].

De novo design of novel enzymes

De novo designed enzymes are artificial proteins, tailored to catalyze a specific target reaction. Recent progress in the development of AI methods and protein design workflows has enabled the creation of 'new-to-nature' enzymes able to efficiently produce high-value compounds, inspiring several recent review articles [17–21].

Recent successful enzyme designs follow one of two main strategies. The first starts from a computational active site model (theozyme) and uses 3D structure diffusion followed by inverse folding to generate stable and soluble protein backbones around it [21,22]. The second strategy designs a minimal active site into known enzyme scaffolds. Although both methods can yield soluble and stable enzymes, the enzymes generally show low initial activities [21]. Hou et al. recently demonstrated that highly stable and active designs can be achieved [18]. Building on a four-helix bundle scaffold designed in 2013 and refined for more than a decade [23], they combined predictions from biophysical modeling and the message-passing neural network LigandMPNN (Ligand Message-Passing Neural Network) [24] with X-ray crystallography and minimal experimental screening to design variants of a small (< 13 kDa) porphyrin binding protein able to catalyze the cyclopropanation of styrene. The top variant achieves a 99% yield, with high substrate turnover and a diastereomeric ratio of 98.5:1.5 for the trans-(*S*,*S*) product. It retains its activity in up to 70% ethanol and exhibits a melting temperature of > 90°C, highlighting the biocatalytic potential of expert-designed enzymes.

Recent research in the field now focuses on integrating additional sources of valuable information into the

design process. For example, generative graph models such as the Distributional Graphormer framework [25] and BioEmu [26] predict structural ensembles of enzymes without relying on computationally expensive molecular dynamics simulations.

Optimizing enzyme properties

Enzyme engineering typically relies on directed evolution (DE), which consists of iterative cycles of variant generation, experimental screening, and recombination of beneficial mutations [27]. Along with efficient experimental screening methods and computational modeling, AI methods have long been applied to learn how mutations affect enzyme properties [28].

AI methods in enzyme engineering follow two principal frameworks. Zero-shot predictions commonly rely on unsupervised training of models on large datasets of protein sequences, potentially enhanced with structural data [29]. A standout model is the PLM ESM3 [30], trained to generate numerical embeddings from amino acid patterns and epistatic relationships, containing information about a protein's 'fitness'. Fitness scores of many recent zero-shot predictors, while not directly predicting measurable enzyme parameters, can be correlated with experimental properties, including stability and activity, as highlighted in the continuously updated ProteinGym benchmark dataset (proteingym.org, last accessed 06.10.2025) [31]. Importantly, these zero-shot predictions vary widely in performance across enzyme datasets, limiting their reliability in iterative enzyme engineering workflows. Zero-shot methods trained to predict specific enzyme properties, such as K_m or k_{cat} values, exemplify this limitation. Here, predictions on substrates outside the training data remain challenging, as highlighted for the model ESP: "Hence, we conclude that ESP only achieves high accuracies for new enzyme-small molecule pairs if the small molecule was present among the ~1400 substrates of our training set" [32]. Additionally, models trained for application on broad enzyme and substrate classes often disregard important factors such as temperature, solvent composition, and pH, which affect K_m and k_{cat} values and vary widely across biocatalytic processes and enzyme engineering campaigns [29]. However, recent models like MPEK [33] are starting to address this challenge.

When high-quality training data is available for a target enzyme, supervised AI models present a viable alternative to zero-shot predictions [34]. Traditional regression models continue to be employed due to their ability to encode smaller but complex datasets and their high interpretability. Landwehr et al. successfully applied an augmented ridge regression model in combination with zero-shot selection methods and a cell-free, high-throughput screening (HTS) platform to rapidly evolve a promiscuous amide synthetase into nine variants, optimized to specifically produce different amide products with improved activity [35]. Similarly, large PLMs can also be integrated into enzyme engineering workflows by using fine-tuning to increase their accuracy in engineering iterations. Notably, a recent evaluation of eight PLMs fine-tuned for different protein engineering tasks revealed that, while fine-tuning significantly improved stability predictions regarding resilience to protease degradation, only five out of the eight models also showed improved predictive ability in replicating experimentally assessed mutational landscapes of three nonenzyme proteins [36]. Despite rapid progress, accurate AI predictions of mutation effects remain challenging. The most reliable predictions can be expected from workflows that integrate multidimensional experimental data to train AI models for enzyme-specific predictions, followed by rigorous variant filtering with biophysics-based modeling and AI-based zero-shot predictions, to reduce experimental screening efforts, while broadening sequence-space exploration.

Artificial intelligence methods in cell engineering

Cell engineering involves targeted genetic modifications to engineer specific aspects of cellular function via the control of gene expression or the design of novel genetic circuits (Figure 2). AI-powered algorithms can aid in this process and guide cell engineering for industrial purposes (Table 1).

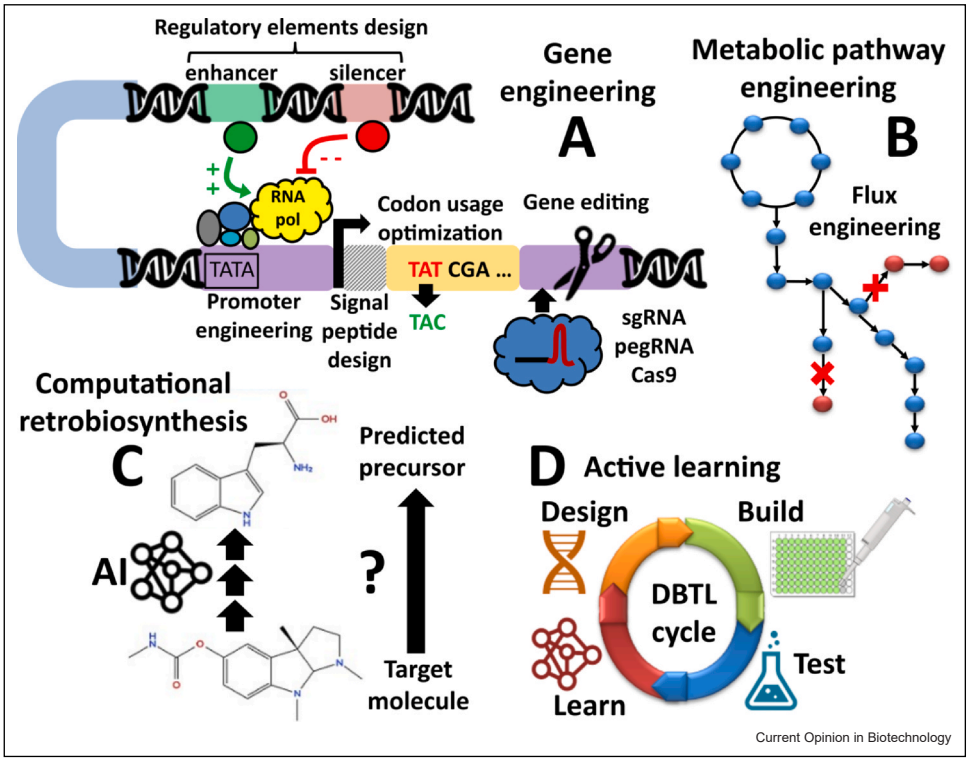
Promoter engineering

Promoters are short DNA sequences that recruit RNA polymerases to initiate gene transcription. Engineering them can boost protein expression. Several deep learning methods have been developed for this purpose. DeepSEED [40] combines a convolutional neural network (CNN) model with a generative adversarial network (GAN) to design synthetic promoters in *Escherichia coli* and mammalian cells, 83% of which showed up to 1.6-fold higher activities. However, GANs instability and convergence issues [45] have prompted the need for alternative models. Variational autoencoders and diffusion models paired with CNNs or transformers have been applied for designing efficient promoters in cyanobacteria [41] and *E. coli* [45]. Notably, Du et al.'s diffusion model [45] outperformed GANs in creating functional synthetic promoters in *E. coli*.

Optimizing DNA/RNA sequences

Codon optimization boosts protein expression but must find a balance between host codon bias, GC content, and mRNA folding. CodonTransformer [48], a generative transformer trained on >1 M DNA-protein pairs, optimizes sequences by reducing the number of negative regulatory elements. While CodonTransformer remains untested, other deep learning models have been

Figure 2



Engineering targets and approaches used for cell optimization. **(a)** Cells can be engineered by optimizing DNA sequences involved in the expression or the regulation of specific genes. **(b)** By engineering BGCs via sequence optimization or gene knock-in/knock-out, metabolic fluxes can be modified to optimize biosynthetic yields or cell growth. **(c)** *De novo* metabolic pathway design via RBS can be used to introduce new metabolic pathways in cells. **(d)** Active learning uses DBTL cycles that integrate predictions from theoretical models with iterative HTS to guide strain optimization.

experimentally validated. As an example, Ravi et al.'s recurrent neural network (RNN) model [43] predicts sequences with reduced rare codon usage, raising mRNA expression by 30–40-fold in HeLa cells. CUSTOM [49], a random forest model, predicts optimal codon usage to optimize tissue-specific protein production, which was experimentally validated in two human cell lines. DeepCodon [50], a transformer model, focuses on preserving important rare codons in *E. coli* genes to boost gene transcription. Experimental validation showed that DeepCodon preserved ~90% of functional rare codons and increased the expression levels of two enzymes in *E. coli*. GEMORNA, a large language model (LLM) trained on over 1 million natural protein sequences from 115 mammalian species, optimizes protein expression by 15–41-fold [44] in human cells. Another RNA language model predicts mutations that increase rRNA thermostability, generating ribosomal units that stayed active *in vitro* after a pre-incubation at 65°C [51].

Engineering protein secretion and translocation

Signal peptides (SPs) are short N-terminal sequences that direct proteins to the secretory pathway or across the plasma membrane. Engineering SPs can boost recombinant protein secretion and improve protein translocation. Deep learning models [52,53], such as TSignal and USPNet**,** are currently the most accurate methods for predicting SPs across all domains of life. AlphaFold's ability to model SPs in specific conformations can also be leveraged to filter out false positives predicted by the former methods [54]. Predictions from generative models were also experimentally validated in several studies. For example, optimized SPs generated by the SPgo generative model [55] increased the secretion of recombinant peptides by ~150-fold, while Wu et al. [56] used gradient boosting to design novel SPs with up to 2.9 times enhanced secretion efficiency in *Yarrowia lipolytica*. Additionally, ProtGPS [57] can generate protein sequences with high subcellular localization specificity, as confirmed by cell imaging experiments.

Engineering gene transcription regulation

Cis-regulatory elements (CREs) regulate gene transcription and can be optimized via deep learning. The CREATor transformer model [39] captures distal CRE patterns up to 2 Mb from target genes across cell types, outperforming prior methods. In addition to improved

Table 1

Deep learning architectures used for enzyme and cell engineering.

Architecture	Example of application	Advantages	Drawbacks
Artificial neural networks (ANN)	Metabolic flux modeling [38]	Straightforward training, lower complexity, less risk of overfitting	Poor at modeling positional/long-range correlations in sequences
CNN	Enhancer detection [39]	Good at detecting spatial hierarchies and patterns between features	Computationally expensive, limited modeling of long-range correlations
GAN	Promoter engineering [40]	Can be combined with a CNN for pretraining, generative application	Training instability of GANs, convergence issues, biases in generated data
Variational autoencoders	Promoter engineering [41]	Dimensionality reduction, generative application, robust to model perturbation	Biases in generated data due to compressed latent space representation
Reinforcement learning/ active learning	Metabolic pathway optimization [42]	Efficient exploration of design space with experimental feedback	Experiment costs, convergence risk to local optimum, definition of a reward criterion
RNN	Codon optimization [43]	Modeling of sequential and time-dependent information in data	Similar issues as for ANN, gradient vanishing problem
Transformer models/LLM	*De novo* RNA design [44]	Good at modeling long-range correlations, strong generalization, generative tasks	Computationally expensive, risk of overfitting, requires diverse and large datasets
Diffusion models	Promoter design [45]	Enhanced generative sampling and increased stability compared to GANs	Computationally expensive, can generate biologically unrealistic sequences
GNN	Function prediction [9] Metabolic flux optimization [46]	Good at modeling spatial features and relational information between multimodal data, high interpretability	Memory-intensive training, risk of overfitting, over-squashing problem [47], scalability issues

Source: Adapted from ref. [37].

detection, recent generative models have demonstrated excellent performance in producing functional CREs with *in vivo* activity. DREAM [58], combining a CNN model trained on millions of *Drosophila* enhancers, generated enhancers 3.9 times more active than the strongest natural *Drosophila* enhancer. Another approach [59] used transfer learning to design synthetic enhancers in *Drosophila*, 78% of which showed *in vivo* tissue-specific activity, while CODA [60], a CNN model paired with a simulated annealing algorithm, was developed to design tissue-specific CREs active in mice and zebrafish. Finally, iterative retraining of generative CNNs was successfully used to create ultra-short human enhancers showing improved cell-specificity when tested in different human cell lines [61].

Engineering gene editing systems

Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-associated protein 9 (CRISPR/Cas9) enables precise genome editing through single guide RNA (sgRNA)-guided Cas9 nucleases. Data-driven models can predict sgRNA binding specificity and design sgRNA with superior activity. DeepCRISTL [62] uses transfer learning to predict on-target activity across cellular contexts and showed consistently higher Spearman correlation coefficients than other competing methods across 13 cellular contexts. For predicting off-target sites, CRISPR-DIPOFF [63] (a RNN model) achieves comparable prediction accuracy to CRISPR-IP [64] (a CNN model), with the latter offering a superior precision-recall balance. For generating novel sgRNAs, sgRNAGen [65], a transformer model, learns from AlphaFold3 ipTM scores to create synthetic sgRNAs that showed *in vivo* genome editing activity in *Bacillus subtilis*. Besides this, LLMs [66,67] have also been used to design Cas9 nucleases with improved binding properties. ProGen2 [66], a language model retrained on 1 M CRISPR operon sequences, generated an active Cas9 variant with improved on-target specificity, with a 95% off-target activity reduction when tested in a human cell line.

Gene discovery and retrobiosynthesis for metabolic pathway design

Identifying biosynthetic gene clusters (BGCs) is crucial for metabolic engineering. Deep learning has improved the detection and classification of BGCs in metagenomes [68]. Large BGCs and enzyme reaction databases provide reaction templates and rules [69] that can be used for computational retrobiosynthesis (RBS). RBS refers to methods that predict plausible enzymatic routes from a target molecule to a precursor. However, rule-based RBS is limited by conservative predictions and convergence issues. Deep learning-based RBS, learning from reaction data without predefined rules, offers better generalization and predictive capabilities [70]. BioRetro [71], combining a generative model with AND-OR tree search, reached 81.6% top-10 accuracy on single-step retrosynthesis prediction tasks and successfully predicted 19 biosynthetic pathways reported in literature. Comparatively, READRetro [72], a dual-representation model, achieved 85.3% prediction success with 72.8%

and 26.6% hit rates for building blocks and pathways, respectively. As an example, READRetro correctly identified L-phenylalanine as a building block of benzylacetone. It also predicted the biosynthetic pathways of several classes of secondary metabolites and suggested a possible synthetic pathway for menisdaurilide, a metabolite without a known biosynthetic route [72]. Notably, a recent prompt-driven model [73] showed an improved performance on reaction class diversity, making it easier to explore alternative pathways beyond biases inherited from training data.

Optimizing metabolic pathways

Metabolic modeling traditionally uses flux balance analysis (FBA) with constrained mechanistic models under pseudo-steady-state assumptions. AI methods can integrate FBA with omics datasets, improving generalization beyond mechanistic constraints. Recent machine learning models [74,75] trained on FBA data identified detrimental reactions in *E. coli* and *Synechocystis*, enabling flux redirection toward growth or CO_2 fixation with strong experimental validation. An enzyme-constrained FBA model [76] (ecMTM) trained on machine-learned k_{cat} values accurately predicted *Myceliophthora thermophila*'s substrate utilization hierarchy between five carbon sources, in agreement with experiments. ecMTM also identified potential enzyme targets that could be engineered to enhance the yields of key products, such as malate and ethanol. While these methods are designed for fine-tuning a predefined combination of known parameters, metabolic engineering often requires identifying new parameters to improve over iterative optimization trials. Active learning [77] can address this challenge by using iterative design-build-test-learn (DBTL) cycles that combine strain mutagenesis, HTS, and machine learning classification to identify critical features to optimize until convergence is met. A recent approach [78] successfully increased *p*-coumaric acid yields in *Saccharomyces cerevisiae* by 68%. Multi-agent reinforcement learning [42] offers an alternative for efficient design-space exploration and was used to optimize L-tryptophan production in two microbial strains.

Limitations of artificial intelligence methods

AI methods face significant limitations despite their potential. Models trained on limited datasets specific to particular organisms or conditions may lack transferability. Unmodeled factors like competing pathway kinetics, regulatory constraints, and toxicity effects can reduce the scalability of metabolic models. As exemplified in one study [79], experimental validation of engineering predictions may fail due to ignored parameters in the models. Limited training data availability also affects generative models, which, unless constrained by well-trained classifiers, may generate biologically unrealistic designs or suffer from inadequate sampling due to mode or posterior collapse [45,80]. Active learning's experimental costs limit its feasibility, favoring methods using small datasets over large neural networks for efficient convergence. Here, however, improper data splitting, such as random splits [38], and class imbalance can limit generalizability and predictive capabilities. The FlowGAT model exemplifies class imbalance issues, showing strong prediction bias toward essential genes, which compromises nonessential gene predictions [46]. These limitations underscore the importance of large, diverse, and well-balanced datasets to ensure reliable biological predictions.

Conclusion

AI methods are rapidly advancing biotechnology, accelerating enzyme and cell engineering for various use-cases, ranging from high-value fine chemical production with optimized microbial factories to potential applications in biomedicine. Still, the successful integration of AI methods into biotechnological processes remains challenging. More standardization in handling data sparsity and imbalance is crucial to enable the fair evaluation of model reliability and generalizability. Overcoming these challenges is a prerequisite to effectively using AI for identifying or designing novel enzymes, optimizing biocatalytic processes, and building efficient cell factories. Future developments in these fields will be key to facilitating the transition from fossil fuel dependence toward a sustainable global bioeconomy.

CRediT authorship contribution statement

Steffen Docter: Conceptualization, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Benoit David:** Conceptualization, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Holger Gohlke:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Data Availability

No data were used for the research described in the article.

Declaration of Competing Interest

The authors declare that there is no potential conflict of interest.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used the BLABLADOR large language model server provided by the Helmholtz AI Artificial Intelligence Cooperation Unit for editorial purposes. After using this

service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgements

This study was funded, in part, by the CasCAR SEED FUND 3.0 project of the Bioeconomy Science Center, which is financially supported by the Ministry of Culture and Science, North-Rhine Westfalia (NRW), Germany within the framework of the NRW Strategieprojekt BioSC (No. 005-2012-0107) and the Helmholtz HFMI project PROFOUND. We are grateful for the financial support from evoxx technologies GmbH, Germany. We gratefully acknowledge the support of our work by the Gauss Centre for Supercomputing e.V. through the John von Neumann Institute for Computing (NIC) at the Jülich Supercomputing Centre (JSC) in providing computing time for projects FOUND and TAM.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Kim GB, Choi SY, Cho IJ, Ahn DH, Lee SY: **Metabolic engineering for sustainability and health**. *Trends Biotechnol* 2023, **41**:425-451.

2. Yu T, Cui H, Li JC, Luo Y, Jiang G, Zhao H: **Enzyme function prediction using contrastive learning**. *Science* 2023, **379**:1358-1363.

3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403-410.

4. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, *et al*.: **Mutant phenotypes for thousands of bacterial genes of unknown function**. *Nature* 2018, **557**:503-509.

5. Yang Y, Jerger A, Feng S, Wang Z, Brasfield C, Cheung MS, *et al*.: **Improved enzyme functional annotation prediction using contrastive learning with structural inference**. *Commun Biol* 2024, **7**:1690.

6. He K, Zhang X, Ren S, Sun J: **Deep residual learning for image recognition.** In *Proceedings of the 2016 Conference on Computer Vision and Pattern Recognition (CVPR)*; Las Vegas, NV, USA: IEEE; 2016:770–778.

7. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, *et al*.: **Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences**. *Proc Natl Acad Sci USA* 2021, **118**:e2016239118.

8. Song Y, Yuan Q, Chen S, Zeng Y, Zhao H, Yang Y: **Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures**. *Nat Commun* 2024, **15**:8180.

9. van der Weg K, Merdivan E, Piraud M, Gohlke H: **TopEC: prediction of Enzyme Commission classes by 3D graph neural networks and localized 3D protein descriptor**. *Nat Commun* 2025, **16**:2737.

10. Krivák R, Hoksza D: **P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure**. *J Chemin* 2018, **10**:39.

11. van der Weg KJ, Gohlke H: **TopEnzyme: a framework and database for structural coverage of the functional enzyme space**. *Bioinformatics* 2023, **39**:btad116.

12. Van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, *et al*.: **Fast and accurate protein structure search with Foldseek**. *Nat Biotechnol* 2024, **42**:243-246.

13. Urban G, Torrisi M, Magnan CN, Pollastri G, Baldi P: **Protein profiles: biases and protocols**. *Comput Struct Biotechnol J* 2020, **18**:2281-2289.

14. •• Joeres R, Blumenthal DB, Kalinina OV: **Data splitting to avoid information leakage with DataSAIL**. *Nat Commun* 2025, **16**:3337.
This study presents the Python package DataSAIL to assist AI researchers in biochemistry and related fields in creating dataset splits with minimal data similarity, reducing the negative impact of data leakage for fair model evaluation. This addresses a core challenge for the performance evaluation and subsequent application of AI models in biotechnology.

15. Su J, He Y, You S, Jiang S, Zhou X, Zhang X, *et al*.: **A trimodal protein language model enables advanced protein searches**. *Nat Biotechnol* 2025, https://doi.org/10.1038/s41587-025-02836-0 ([Epub ahead of print]).

16. • Flöge K, Udayakumar S, Sommer J, Piraud M, Kesselheim S, Fortuin V, *et al*.: **OneProt: towards Multi-Modal Protein Foundation Models**. *PLOS Comp. Biol.* 2024, **21**:e1013679.
The OneProt model represents a unified, large-scale PLM trained across multiple data modalities, enabling accurate prediction of diverse protein properties from sequence alone. By integrating evolutionary, structural, and functional information, OneProt achieves state-of-the-art performance in protein annotation tasks.

17. • Listov D, Vos E, Hoffka G, Hoch SY, Berg A, Hamer-Rogotner S, *et al*.: **Complete computational design of high-efficiency Kemp elimination enzymes**. *Nature* 2025, **643**:1421-1427.
This study addresses the common challenges of high experimental screening efforts and low starting activities in the *de novo* design of enzymes. They developed a fully computational design workflow to create new enzymes based on known protein fragments. The workflow yielded three Kemp eliminase designs with high stability and impressive catalytic efficiency.

18. • Hou K, Huang W, Qi M, Tugwell TH, Alturaifi TM, Chen Y, *et al*.: **De novo design of porphyrin-containing proteins as efficient and stereoselective catalysts**. *Science* 2025, **388**:665-670.
This study demonstrates the successful *de novo* design of small porphyrin-containing enzymes with high thermostability and organic solvent tolerance. The enzymes also show high activity toward a range of substrates for catalyzing cyclopropanation reactions and silicon–hydrogen insertion.

19. Markus B, C GC, Andreas K, Arkadij K, Stefan L, Gustav O, *et al*.: **Accelerating biocatalysis discovery with machine learning: a paradigm shift in enzyme engineering, discovery, and design**. *ACS Catal* 2023, **13**:14454-14469.

20. Winnifrith A, Outeiral C, Hie BL: **Generative artificial intelligence for de novo protein design**. *Curr Opin Struct Biol* 2024, **86**:102794.

21. Vornholt T, Stockinger P, Mutný M, Jeschek M, Nestl B, Oberdorfer G, *et al*.: **Of revolutions and roadblocks: the emerging role of machine learning in biocatalysis**. *ACS Cent Sci* 2025, **11**:1828-1838.

22. Tantillo DJ, Jiangang C, Houk KN: **Theozymes and compuzymes: theoretical models for biological catalysis**. *Curr Opin Chem Biol* 1998, **2**:743-750.

23. Fry HC, Lehmann A, Sinks LE, Asselberghs I, Tronin A, Krishnan V, *et al*.: **Computational de novo design and characterization of a protein that selectively binds a highly hyperpolarizable abiological chromophore**. *J Am Chem Soc* 2013, **135**:13914-13926.

24. Dauparas J, Lee GR, Pecoraro R, An L, Anishchenko I, Glasscock C, *et al*.: **Atomic context-conditioned protein sequence design using LigandMPNN**. *Nat Methods* 2025, **22**:717-723.

25. Zheng S, He J, Liu C, *et al*.: **Predicting equilibrium distributions for molecular systems with deep learning**. *Nat Mach Intell* 2024, **6**:558-567.

26. Lewis S, Hempel T, Jiménez-Luna J, Gastegger M, Xie Y, Foong AYK, *et al*.: **Scalable emulation of protein equilibrium ensembles with generative deep learning**. *Science* 2025, **389**:eadv9817.

27. Packer MS, Liu DR: **Methods for the directed evolution of proteins**. *Nat Rev Genet* 2015, **16**:379-394.

28. Fox RJ, Davis SC, Mundorff EC, Newman LM, Gavrilovic V, Ma SK, *et al*.: **Improving catalytic function by ProSAR-driven enzyme evolution**. *Nat Biotechnol* 2007, **25**:338-344.

29. Liu C, Wu J, Chen Y, Liu Y, Zheng Y, Liu L, *et al*.: **Advances in zero-shot prediction-guided enzyme engineering using machine learning**. *ChemCatChem* 2025, **17**:e202401542.

30. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, *et al.*: **Simulating 500 million years of evolution with a language model**. *Science* 2025, **387**:850-858.

31. Notin P, Kollasch A, Ritter D, van Niekerk L, Paul S, Spinner H et al: **ProteinGym: large-scale benchmarks for protein fitness prediction and design.** In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*; 2023: 64331–64379.

32. Kroll A, Ranjan S, Engqvist MKM, Lercher MJ: **A general model to predict small molecule substrates of enzymes based on machine and deep learning**. *Nat Commun* 2023, **14**:2787.

33. Wang J, Yang Z, Chen C, Yao G, Wan X, Bao S, *et al.*: **MPEK: a multitask deep learning framework based on pretrained language models for enzymatic reaction kinetic parameters prediction**. *Brief Bioinform* 2024, **25**:bbae387.

34. Jiang K, Yan Z, Di Bernardo M, Sgrizzi SR, Villiger L, Kayabolen A, *et al.*: **Rapid in silico directed evolution by a protein language model with EVOLVEpro**. *Science* 2025, **387**:eadr6006.

35. • Landwehr GM, Bogart JW, Magalhaes C, Hammarlund EG, Karim AS, Jewett MC: **Accelerated enzyme engineering by machine-learning guided cell-free expression**. *Nat Commun* 2025, **16**:865.
In this study, a complete enzyme engineering platform to optimize promiscuous enzymes for specific substrates is presented. The authors integrate cell-free variant generation and screening with an active learning AI framework for variant prediction and computational filtering to design amide synthase variants for the production of nine pharmaceuticals.

36. Schmirler R, Heinzinger M, Rost B: **Fine-tuning protein language models boosts predictions across diverse tasks**. *Nat Commun* 2024, **15**:7407.

37. Wang X, Xu K, Huang Z, Lin Y, Zhou J, Zhou L, *et al.*: **Accelerating promoter identification and design by deep learning**. *Trends Biotechnol* 2025,S016777992500174X.

38. Oulia F, Charton P, Lo-Thong-Viramoutou O, Acevedo-Rocha CG, Liu W, Huynh D, *et al.*: **Metabolic fluxes using deep learning based on enzyme variations: application to glycolysis in *Entamoeba histolytica***. *IJMS* 2024, **25**:13390.

39. Li Y, Ju F, Chen Z, Qu Y, Xia H, He L, *et al.*: **CREaTor: zero-shot cis-regulatory pattern modeling with attention mechanisms**. *Genome Biol* 2023, **24**:266.

40. Zhang P, Wang H, Xu H, Wei L, Liu L, Hu Z, *et al.*: **Deep flanking sequence engineering for efficient promoter design using DeepSEED**. *Nat Commun* 2023, **14**:6309.

41. Seo E, Choi YN, Shin YR, Kim D, Lee JW: **Design of synthetic promoters for cyanobacteria with generative deep-learning model**. *Nucleic Acids Res* 2023, **51**:7071-7082.

42. Sabzevari M, Szedmak S, Penttilä M, Jouhten P, Rousu J: **Strain design optimization using reinforcement learning**. *PLoS Comput Biol* 2022, **18**:e1010177.

43. Ravi S, Sharma T, Yip M, Yang H, Xie J, Gao G, *et al.*: **A deep learning model trained on expressed transcripts across different tissue types reveals cell-type codon-optimization preferences**. *Nucleic Acids Res* 2025, **53**:gkaf233.

44. •• Zhang H, Liu H, Xu Y, Huang H, Liu Y, Wang J, *et al.*: **Deep generative models design mRNA sequences with enhanced translational capacity and stability**. *Science* 2025, **390**:eadr8470.
This study highlights the development of a large language generative model that can design synthetic mRNAs with significantly enhanced expression and stability. The RNA sequences generated by this model achieved a remarkable 41-fold increase in expression compared to an optimized benchmark. Moreover, they resulted in a 15-fold increase in the expression of human erythropoietin and higher antibody titers of the COVID vaccine in mice.

45. • Du Q, Poon MN, Zeng X, Zhang P, Wei Z, Wang H, *et al.*: **Synthetic promoter design in *Escherichia coli* based on multinomial diffusion model**. *iScience* 2024, **27**:111207.
This study presents a diffusion model for efficient promoter design in *E. coli*. The model shows high robustness, superiority to GANs in decoupling weak sequence signals, and was able to generate synthetic promoters with superior *in vivo* activity compared to natural counterparts.

46. Hasibi R, Michoel T, Oyarzún DA: **Integration of graph neural networks and genome-scale metabolic models for predicting gene essentiality**. *npj Syst Biol Appl* 2024, **10**:24.

47. Li H, Li C, Zhang J, Ouyang Y, Rong W: **Addressing over-squashing in GNNs with graph rewiring and ordered neurons.** In *Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN)*; Yokohama, Japan: IEEE; 2024:1–8.

48. Fallahpour A, Gureghian V, Filion GJ, Lindner AB, Pandi A: **CodonTransformer: a multispecies codon optimizer using context-aware neural networks**. *Nat Commun* 2025, **16**:3205.

49. Hernandez-Alias X, Benisty H, Radusky LG, Serrano L, Schaefer MH: **Using protein-per-mRNA differences among human tissues in codon optimization**. *Genome Biol* 2023, **24**:34.

50. Han X, Shao X, Liu S, Shi Z, Huang R, Chu H, *et al.*: **DeepCodon: a deep learning codon-optimization model to enhance protein expression**. *BioDesign Res* 2025, **7**:100042.

51. Shulgina Y, Trinidad MI, Langeberg CJ, Nisonoff H, Chithrananda S, Skopintsev P, *et al.*: **RNA language models predict mutations that improve RNA function**. *Nat Commun* 2024, **15**:10627.

52. Dumitrescu A, Jokinen E, Paatero A, Kellosalo J, Paavilainen VO, Lähdesmäki H: **TSignal: a transformer model for signal peptide prediction**. *Bioinformatics* 2023, **39**:347-356.

53. Shen J, Yu Q, Chen S, Tan Q, Li J, Li Y: **Unbiased organism-agnostic and highly sensitive signal peptide predictor with deep protein language model**. *Nat Comput Sci* 2023, **4**:29-42.

54. Sanaboyana VR, Elcock AH: **Improving signal and transit peptide predictions using AlphaFold2-predicted protein structures**. *J Mol Biol* 2024, **436**:168393.

55. • Dai X-p, Meng X-c, Zhou Y-j, Li Z-m, Ji Y, Schwaneberg U, *et al.*: ***De novo* design of high-performance sec-type signal peptide via a hybrid deep learning architecture**. *JACS Au* 2025, **5**:4669-4674.
This work introduces the SPgo generative model, which integrates biophysical constraints to design optimized Sec-type SP sequences. Experimental validation on various protein and peptide targets showed the model can effectively boost secretory production of optimized sequences up to 150-fold in comparison to intracellular expression.

56. Wu Z, Chen W, Hong Y, Wang Y, Xu P: **Machine learning-assisted rational design and evolution of novel signal peptides in *Yarrowia lipolytica***. *Synth Syst Biotechnol* 2025, **10**:1275-1283.

57. •• Kilgore HR, Chinn I, Mikhael PG, Mitnikov I, Van Dongen C, Zylberberg G, *et al.*: **Protein codes promote selective subcellular compartmentalization**. *Science* 2025, **387**:1095-1101.
This study introduces ProtGPS, a PLM that accurately predicts the localization of human proteins beyond those included in the training. ProtGPS successfully generated new protein sequences that selectively gather in the nucleolus and identified mutations leading to altered subcellular localization.

58. Li Z, Zhang Y, Peng B, Qin S, Zhang Q, Chen Y, *et al.*: **A novel interpretable deep learning-based computational framework designed synthetic enhancers with broad cross-species activity**. *Nucleic Acids Res* 2024, **52**:13447-13468.

59. •• De Almeida BP, Schaub C, Pagani M, Secchia S, Furlong EEM, Stark A: **Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo**. *Nature* 2024, **626**:207-211.
The authors created a deep learning model using single-cell transcriptome data and enhancer activity data. They designed 40 synthetic enhancers for Drosophila tissues and found that 78% were active and 68% drove tissue-specific expression.

60. •• Gosai SJ, Castro RI, Fuentes N, Butts JC, Mouri K, Alasoadura M, *et al.*: **Machine-guided design of cell-type-targeting cis-regulatory elements**. *Nature* 2024, **634**:1211-1220.
This study showcases CODA, a deep learning model that predicts and creates DNA sequences with high cell-specific regulatory activity. CODA was used to generate thousands of synthetic regulatory elements that outperformed natural ones in driving cell-type-specific gene expression across three cell lines.

61. Yin C, Castillo-Hair S, Byeon GW, Bromley P, Meuleman W, Seelig G: **Iterative deep learning design of human enhancers exploits**

condensed sequence grammar to achieve cell-type specificity. *Cell Syst* 2025, **16**:101302.

62. Elkayam S, Tziony I, Orenstein Y: **DeepCRISTL: deep transfer learning to predict CRISPR/Cas9 on-target editing efficiency in specific cellular contexts**. *Bioinformatics* 2024, **40**:btae481.

63. Toufikuzzaman M, Hassan Samee MA, Sohel Rahman M: **CRISPR-DIPOFF: an interpretable deep learning approach for CRISPR Cas-9 off-target prediction**. *Brief Bioinform* 2024, **25**:bbad530.

64. Zhang ZR, Jiang ZR: **Effective use of sequence information to predict CRISPR-Cas9 off-target**. *Comput Struct Biotechnol J* 2022, **20**:650-661.

65. Xia Y, Liang Z, Du X, Cao D, Li J, Sun L, *et al*.: **Design of function-regulating RNA via deep learning and AlphaFold 3**. *Brief Bioinform* 2025, **26**:bbaf419.

66. •• Ruffolo JA, Nayfach S, Gallagher J, Bhatnagar A, Beazer J, Hussain R, *et al*.: **Design of highly functional genome editors by modeling CRISPR–Cas sequences**. *Nature* 2025, **645**:518-525.
The authors retrained a protein model (ProGen2) on over a million CRISPR operons, generating a million new Cas9 proteins. Using HTS, they identified a variant with improved on-target specificity and a 95% reduction in off-target activity compared to conventional SpCas9 nucleases.

67. Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, *et al*.: **Sequence modeling and design from molecular to genome scale with Evo**. *Science* 2024, **386**:eado9336.

68. Rios-Martinez C, Bhattacharya N, Amini AP, Crawford L, Yang KK: **Deep self-supervised learning for biosynthetic gene cluster detection and product classification**. *PLoS Comput Biol* 2023, **19**:e1011162.

69. Udwary DW, Doering DT, Foster B, Smirnova T, Kautsar SA, Mouncey NJ: **The secondary metabolism collaboratory: a database and web discussion portal for secondary metabolite biosynthetic gene clusters**. *Nucleic Acids Res* 2025, **53**:717-723.

70. Gricourt G, Meyer P, Duigou T, Faulon JL: **Artificial intelligence methods and models for retro-biosynthesis: a scoping review**. *ACS Synth Biol* 2024, **13**:2276-2294.

71. Zhang X, Liu J, Yang F, Zhang Q, Yang Z, Shah HA: **Planning biosynthetic pathways of target molecules based on metabolic reaction prediction and AND-OR tree search**. *Comput Biol Chem* 2024, **111**:108106.

72. Kim T, Lee S, Kwak Y, Choi M, Park J, Hwang SJ, *et al*.: **READRETRO: natural product biosynthesis predicting with retrieval-augmented dual-view retrosynthesis**. *New Phytol* 2024, **243**:2512-2527.

73. Thakkar A, Vaucher AC, Byekwaso A, Schwaller P, Toniato A, Laino T: **Unbiasing retrosynthesis language models with disconnection prompts**. *ACS Cent Sci* 2023, **9**:1488-1498.

74. • Woo H, Kim Y, Kim D, Yoon SH: **Machine learning identifies key metabolic reactions in bacterial growth on different carbon sources**. *Mol Syst Biol* 2024, **20**:170-186.
This study trained two deep learning models on gene-deletion and flux data to predict metabolic reactions that enhance or hinder *E. coli* growth across 30 carbon sources. Both models surpassed traditional methods by identifying key non-essential reactions with strong experimental agreement. Key predictions, including carbon-source-dependent effects on several metabolic pathways, were experimentally validated.

75. Kugler A, Stensjö K: **Machine learning predicts system-wide metabolic flux control in cyanobacteria**. *Metab Eng* 2024, **82**:171-182.

76. Wang Y, Mao Z, Dong J, Zhang P, Gao Q, Liu D, *et al*.: **Construction of an enzyme-constrained metabolic network model for *Myceliophthora thermophila* using machine learning-based kcat data**. *Micro Cell Fact* 2024, **23**:138.

77. Van Lent P, Schmitz J, Abeel T: **Simulated design–build–test–learn cycles for consistent comparison of machine learning methods in metabolic engineering**. *ACS Synth Biol* 2023, **12**:2588-2599.

78. • Moreno-Paz S, Van Der Hoek R, Eliana E, Zwartjens P, Gosiewska S, Martins Dos Santos VAP, *et al*.: **Machine learning-guided optimization of p-coumaric acid production in yeast**. *ACS Synth Biol* 2024, **13**:1312-1322.
This study introduces a machine-learning-driven approach to optimize the production of p-coumaric acid in yeast, using feature importance to explore the design space. This strategy enabled a 68% increase in p-coumaric acid production over two optimization cycles.

79. Bernstein DB, Akkas B, Price MN, Arkin AP: **Evaluating *E. coli* genome-scale metabolic model accuracy with high-throughput mutant fitness data**. *Mol Syst Biol* 2023, **19**:e11566.

80. Yan C, Yang J, Ma H, Wang S, Huang J: **Molecule sequence generation with rebalanced variational autoencoder loss**. *J Comput Biol* 2023, **30**:82-94.