Biotechnology Advances 82 (2025) 108603

Contents lists available at ScienceDirect

Biotechnology Advances

journal homepage: www.elsevier.com/locate/biotechadv

AI-driven *de novo* enzyme design: Strategies, applications, and future prospects

Xi-Chen Cui [a,b], Yan Zheng [a,b], Ye Liu [a,b,c], Zhiguang Yuchi [a,b,c,*], Ying-Jin Yuan [a,b,*]

[a] State Key Laboratory of Synthetic Biology, Tianjin University, Tianjin 30072, PR China
[b] Frontiers Science Center for Synthetic Biology(Ministry of Education), School of Synthetic Biology and Biomanufacturing, Tianjin University, Tianjin 300072, PR China
[c] School of Pharmaceutical Science and Technology, Tianjin University, Tianjin 300072, PR China

ARTICLE INFO

Keywords:
Enzyme
De novo design
Artificial intelligence
Enzyme function validation
Enzyme optimization

ABSTRACT

Enzymes are indispensable for biological processes and diverse applications across industries. While top-down modification strategies, such as directed evolution, have achieved remarkable success in optimizing existing enzymes, bottom-up *de novo* enzyme design has emerged as a transformative approach for engineering novel enzymes with customized catalytic functions, independent of natural templates. Recent advancements in artificial intelligence (AI) and computational power have significantly accelerated this field, enabling breakthroughs in enzyme engineering. These technologies facilitate the rapid generation of enzyme structures and amino acid sequences optimized for specific functions, thereby enhancing design efficiency. They also support functional validation and activity optimization, improving the catalytic performance, stability, and robustness of *de novo* designed enzymes. This review highlights recent advancements in AI-driven *de novo* enzyme design, discusses strategies for validation and optimization, and examines the challenges and future prospects of integrating these technologies into enzyme development.

1. Introduction

Enzyme design is a transformative approach for addressing a wide array of scientific, industrial, and medical challenges by engineering enzymes to perform specific functions, through which the limitations of natural enzymes, such as the lack of stability, specificity, or activity required for novel applications or non-native conditions, can be addressed. However, the vast space for enzyme design, comprising countless possible amino acid combinations, renders exhaustive exploration, even making it infeasible (Cobb et al., 2013).

Traditional enzyme engineering methods predominantly rely on top-down strategies, including rational design, semi-rational design, and directed evolution (Chen and Arnold, 1993; Lerner et al., 1964; Tian et al., 2024). These approaches have achieved significant milestones, such as improving the efficiency of enzymes like P450 and alcohol dehydrogenase and engineering enzymes to catalyze new reactions (Brandenberg et al., 2019; Jensen et al., 2021; Kim et al., 2019; Li et al., 2024; Liu et al., 2019; Xu et al., 2024).

Theoretically, every specific reaction has an optimal enzyme sequence that yields maximal activity. While natural evolution could eventually reach this sequence given infinite time, directed evolution accelerates the process by artificially enhancing the rate of sequence variation and selection. Whether there is a fundamentally different strategy that can achieve comparable or superior results within a much shorter timescale by efficiently sampling a broad design space from scratch rather than gradually modifying existing sequences is worth exploring.

De novo enzyme design is the computational creation of novel protein sequences and structures from first principles or learned models, rather than modifying natural enzymes. Unlike traditional rational design or directed evolution, which explore sequence space locally around existing scaffolds, *de novo* design enables access to novel folds and functions absent in nature. By starting from a desired function or structure and leveraging generative models and physics-based simulations, it can efficiently search vast regions of sequence–structure space. This global exploration allows it to bypass local optima and more directly identify high-performance solutions, accelerating the discovery of enzymes with enhanced activity, specificity, and stability.

Early *de novo* designs relied on physicochemical principles (Hodges et al., 1981). For instance, DeGrado et al. (1987) laid the foundation for *de novo* design of enzymes by using geometric parameters of their tertiary and quaternary structures, and Kuhlman et al. (2003) advanced the

* Corresponding authors at: State Key Laboratory of Synthetic Biology, Tianjin University, Tianjin 30072, PR China.
E-mail addresses: yuchi@tju.edu.cn (Z. Yuchi), yjyuan@tju.edu.cn (Y.-J. Yuan).

https://doi.org/10.1016/j.biotechadv.2025.108603
Received 2 February 2025; Received in revised form 22 April 2025; Accepted 10 May 2025
Available online 12 May 2025

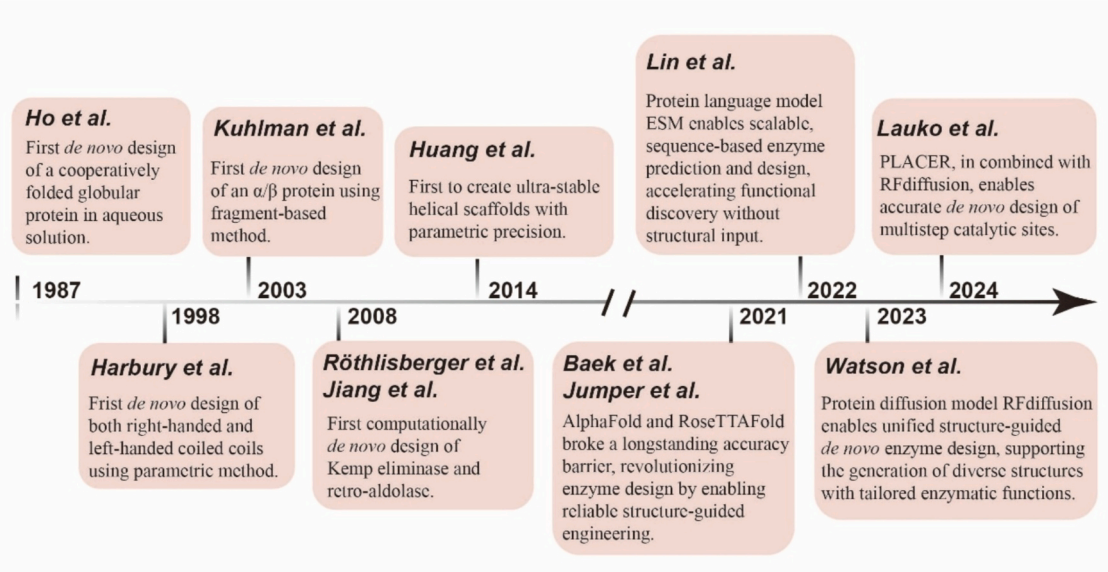

Fig. 1. Evolution of *de novo* enzyme design. Key studies are shown in chronological order, highlighting the transformation of enzyme design from physics-based methods reliant on expert knowledge to statistics-based approaches. Modern methodologies can autonomously infer design constraints, reducing dependence on prior experience or detailed mechanistic understanding of target enzymes.

field with fragment-based methods that enabled backbone generation and sequence optimization. Despite some successes in physicochemical principle-based methods, including high thermodynamically stable helical bundles and water-soluble α-helical barrels (Huang et al., 2014; Thomson et al., 2014), they require extensive knowledge of target enzyme structures and face challenges in parameterizing diverse enzyme families.

Recent advances in machine learning (ML) have revolutionized the landscape of *de novo* enzyme design, providing powerful tools to address the limitations of those traditional approaches. ML-driven strategies can design enzymes by targeting known functional sites or directly addressing specific design requirements (Ingraham et al., 2023; Munsamy et al., 2024; Watson et al., 2023), which often serve as starting points, enabling rapid *in silico* design and functional validation thereafter. Moreover, ML plays a pivotal role in subsequent refining, experimental validation, and optimization, which are essential for achieving enhanced enzyme performance. Fig. 1 highlights the chronological milestones for this transformative approach.

In this review, we highlight recent advances in AI-driven bottom-up *de novo* enzyme design, focusing on three key areas: 1) the application of ML techniques for *de novo* enzyme design, 2) the use of ML for rapid *in silico* functional and interaction validation, and 3) leveraging ML for efficient enzyme modification and optimization. We conclude by discussing the challenges and future directions of AI-driven *de novo* enzyme design, emphasizing its transformative potential in advancing the field.

2. AI/ML data curation and algorithm architectures

2.1. Data curation

The foundation of reliable AI/ML-driven enzyme design lies in access to high-quality, well-curated datasets that reflect the complexity of biological systems. Collaborative efforts have resulted in specialized databases that catalog amino acid sequences (Bateman et al., 2023), protein structures (Berman et al., 2000; Sillitoe et al., 2021), enzymatic functions (Chang et al., 2021), biochemical reactions (Bansal et al., 2022), and metabolic pathways (Caspi et al., 2016; Ogata and Goto, 2000), which are crucial inputs for model training and validation. As the saying goes, "garbage in, garbage out", underscoring the importance of rigorous data quality control in building accurate and trustworthy models.

Sequence dataset curation requires a careful balance between diversity and quality. Current methods often rely on clustering algorithms, such as MMseqs2 (Steinegger and Söding, 2017), to select representative sequences while minimizing redundancy that could bias learning. This process involves a trade-off: stricter clustering reduces redundancy but limits dataset coverage, whereas looser thresholds preserve diversity but risk overrepresentation of homologous sequences. Hierarchical approaches like UniRef (UniRef50/90/100) optimize this balance through empirical evaluation of model performance across clustering levels (Bateman et al., 2023). Later, ESM3 employs an optimized multi-stage reclustering workflow. It begins with stringent clustering to remove redundant sequences, followed by progressively relaxed reclustering steps that reintroduce sequence diversity in a controlled and systematic manner (Hayes et al., 2025).

Structural biology databases, such as the Protein Data Bank (PDB) and CATH database (Berman et al., 2000; Sillitoe et al., 2021), contain a heterogeneous collection of protein structures by experimental determination, like cryo-electron microscopy (cryo-EM), X-ray crystallography (XRC), Nuclear magnetic resonance (NMR), X-ray free-electron lasers (XFELs), and Hydrogen-deuterium exchange (HDX), which vary widely in both resolution and overall quality. Resolution has traditionally served as a primary criterion for data curation, though its optimal threshold often depends on the specific modeling objective.

Similar to the trade-offs observed in sequence clustering, structural filtering faces the challenge of balancing fidelity with dataset breadth, particularly important given the limited number of experimentally resolved structures (~233,000 as of 2024). Moreover, structural databases exhibit inherent biases, particularly the overrepresentation of thermostable or easily crystallizable proteins. Such biases can skew model generalizability, posing a significant challenge for AI-driven enzyme design, which depends critically on diverse, high-quality training data. This scarcity and bias underscore the growing need for effective structural augmentation strategies to expand and diversify training datasets.

To address these limitations, computationally generated structural datasets have become an essential complement. Recent advances in protein structure prediction, such as AlphaFold and ESMFold, have enabled large-scale dataset expansion with atomic-level accuracy across many protein families (Abramson et al., 2024; Lin et al., 2023).

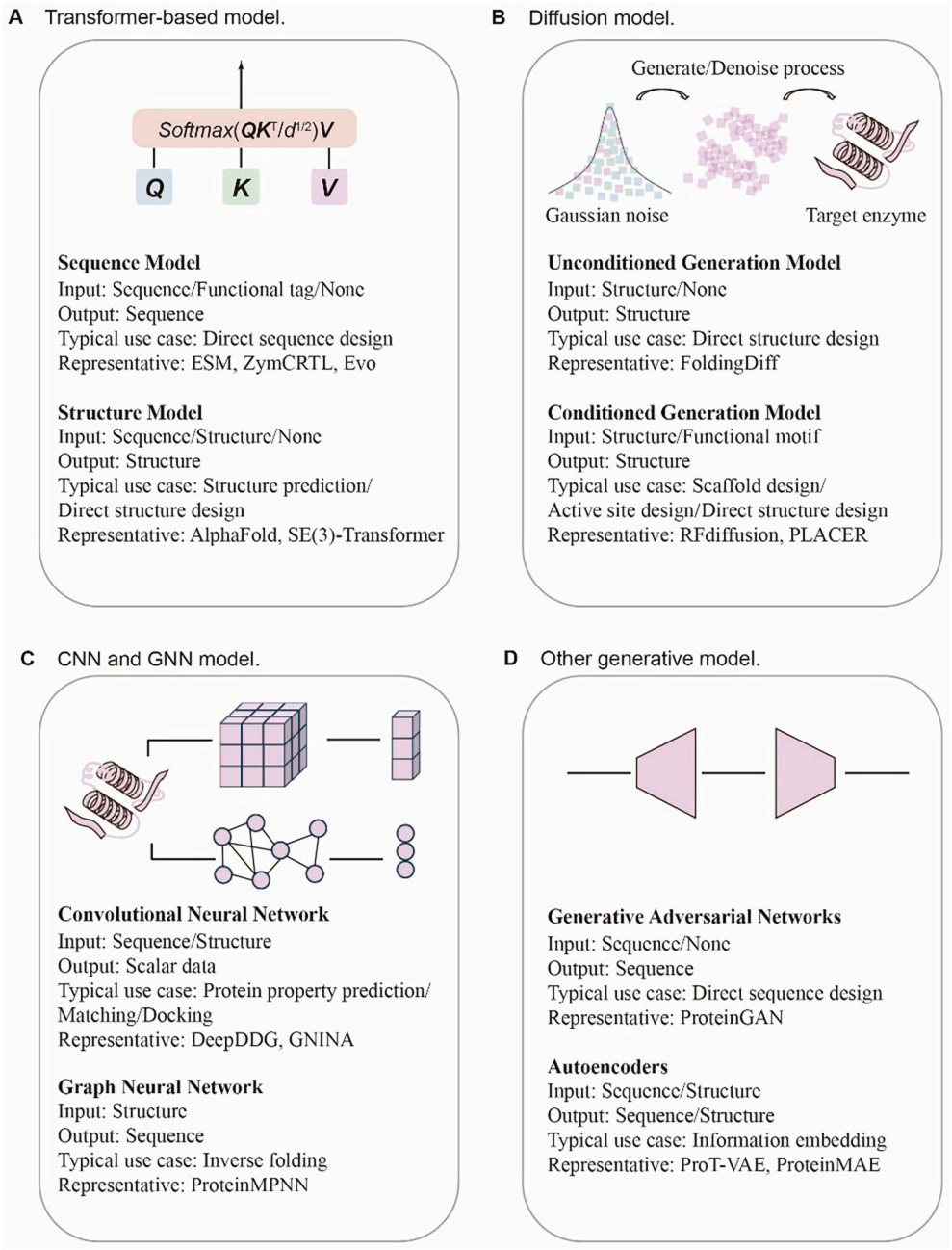

Fig. 2. Overview of algorithm architectures. Key algorithms are categorized into four types, with a comparative summary of their inputs, outputs, typical applications, and representative methods.

Resources like AlphaFoldDB and ESM Atlas offer high-throughput structural predictions, typically validated using the predicted Local Distance Difference Test (pLDDT) score, a per-residue confidence measure where higher values correspond to increased structural reliability (Munsamy et al., 2024; Varadi et al., 2023). Standard curation pipelines apply pLDDT thresholds to extract topologically trustworthy regions, facilitating the integration of predicted and experimental structures under consistent quality standards (Jumper et al., 2021).

In parallel, these expanded structural datasets can also be leveraged to diversify protein sequence space through inverse folding. When natural sequences are insufficient for large-scale training, structure-guided sequence generation provides an effective means of synthetic augmentation. This process typically involves generating sequences from PDB-derived structural templates, followed by validation *via* 3D prediction tools such as AlphaFold3. TM-score metrics are then used to quantify topological congruence between predicted and template structures, ensuring that only structurally faithful sequences are included in the final dataset (Krishna et al., 2024; Varadi et al., 2023).

In summary, rigorous data curation is the cornerstone of *de novo* enzyme design, encompassing high-quality sequence, structure, and functional annotations. Careful selection and validation of these datasets are essential to avoid the pitfalls of poor input, ensuring model accuracy and enabling the successful development of innovative and application-ready enzymes.

2.2. Algorithm architectures

Designing efficient model architectures is a critical step in ML-based *de novo* enzyme design. Once high-quality datasets have been curated, attention shifts to developing models capable of accurately capturing the

complex and multiscale relationships underlying enzymatic function (Fig. 2).

2.2.1. Transformer-based model

A range of specialized architectures has been developed to handle the demands of enzyme modeling from both sequence and structural angles. Among these, transformer-based models have emerged as one of the most influential approaches. Their self-attention mechanism enables the modeling of long-range dependencies and co-evolutionary patterns in amino acid sequences, which are critical for capturing the global properties of enzymes (Vaswani et al., 2017). Furthermore, self-supervised training paradigms, including masked language modeling tasks like BERT and autoregressive objectives like GPT, enable models to acquire knowledge from extensive unannotated datasets while substantially improving data utilization efficiency (Devlin et al., 2019).

Transformer has demonstrated outstanding performance in sequence prediction, structural inference, and *de novo* enzyme generation (Ferruz et al., 2022; Lin et al., 2023; Yu et al., 2023). However, transformers also face inherent limitations when applied directly to structural data, where spatial and geometric information must be encoded in a fundamentally different way than in one-dimensional sequences.

To overcome these challenges, architectural innovations have adapted transformer frameworks for 3D protein modeling. For instance, axial attention in the ESM framework enables independent aggregation along the dimensions of multiple sequence alignments (MSAs), preserving both evolutionary and positional context (Lin et al., 2023). AlphaFold2's Evoformer module introduces triangular attention mechanisms to model spatial relationships among residues, embedding geometric constraints such as distance complementarity and contact patterns directly into the learning process (Jumper et al., 2021). These adaptations allow transformer-based models to transition effectively from sequence to structure prediction, bridging the gap between linear sequence data and the multidimensional geometry of enzyme function.

Another key consideration in structural modeling is the enforcement of SE(3) symmetry, which is invariance to rotation and translation in three-dimensional space. This is particularly important in protein modeling, where absolute orientation should not affect prediction outcomes. Architectures such as SE(3)-Transformer and AlphaFold2 implement strategies to ensure symmetry preservation, using spherical harmonics, reference frame anchoring, and invariant point attention to embed geometric constraints directly into their computation (Fuchs et al., 2020; Jumper et al., 2021). The ESM3 framework further refines this approach by constructing local residue-centric reference frames, ensuring SE(3)-equivariant spatial descriptors that are independent of global orientation (Hayes et al., 2025).

2.2.2. Diffusion model

In parallel with transformer-based advances, diffusion models have emerged as a powerful generative paradigm for structure-based enzyme design. These models learn to generate complex 3D conformations by iteratively denoising random inputs, guided by learned protein structural priors (Ho et al., 2020). Diffusion models implicitly enforce critical biophysical constraints such as backbone geometry, secondary structure motifs, and hydrophobic core packing. When combined with SE(3)-equivariant neural networks, they can generate physically realistic and chemically meaningful protein structures with atomic precision, making them particularly well-suited for *de novo* enzyme design under structural constraints (Watson et al., 2023).

2.2.3. Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs)

Beyond these dominant paradigms, convolutional neural networks (CNNs) continue to contribute, particularly in detecting local structural motifs and functional sites. While their limited receptive fields constrain their ability to model long-range interactions, CNNs are effective for analyzing spatially local features such as catalytic residues or short-range sequence motifs (McNutt et al., 2021). In contrast, graph neural networks (GNNs) excel at modeling the complex topologies of proteins by explicitly representing amino acids as nodes and their interactions as edges. This framework captures detailed geometric relationships and has driven progress in inverse folding, although scalability remains a bottleneck for modeling large proteins (Dauparas et al., 2022).

2.2.4. Other generative models

Finally, several other generative models also play roles in enzyme design. Autoencoders, particularly variational forms (VAEs), provide compact representations of protein sequences or structures for efficient sampling and reconstruction (Sevgen et al., 2023; Yuan et al., 2023). Extensions such as vector quantized VAEs (VQ-VAEs) enable discrete latent modeling that aligns well with categorical biological data (Hayes et al., 2025). Generative adversarial networks (GANs), once popular for molecular generation, have seen reduced usage due to issues like mode collapse, with many recent efforts shifting toward more stable diffusion-based frameworks (Gui et al., 2023; Maziarka et al., 2020).

Together, these architectural innovations reflect a broader trend toward multimodal and hybrid modeling strategies. By integrating transformers for global sequence modeling, diffusion models for structure generation, CNNs for local feature extraction, GNNs for geometric encoding, and autoencoders for latent representation and sampling, researchers are building comprehensive systems that capture the hierarchical nature of enzymes, from primary sequence to tertiary structure and functional dynamics. These advances lay the groundwork for rational enzyme design driven by AI.

2.3. Method evaluation

Following the construction of a model, the next crucial step is the rapid evaluation of its effectiveness. *In silico* metrics enable fast computational assessments and facilitate fair comparisons between different models, making them integral to evaluating *de novo* enzyme design outcomes. However, while traditional NLP metrics such as bilingual evaluation understudy (BLEU) and perplexity are often adapted, they lack clear physical relevance, limiting their correlation with actual enzyme functions or experimental results (Blagec et al., 2022; Papineni et al., 2002).

To address this issue, current efforts focus on developing metrics that are more practically relevant to *de novo* enzyme design. These include the following three key considerations: 1) Novelty of design: Since the primary goal of *de novo* enzyme design is to create enzymes with novel functions not found in nature, evaluation of sequence and structural uniqueness is critical. Tools such as BLAST can be used to perform sequence or structural alignment and confirm divergence from existing natural enzymes (Altschul et al., 1990). 2) Structural validity and foldability: Designed enzymes must exhibit realistic folding and catalytic potential. Prediction software can compare key features of the generated datasets, including sequence lengths, secondary structures, and other attributes, to those of natural datasets, ensuring the rationality of the design (Munsamy et al., 2024). High-precision structural prediction tools, such as AlphaFold3 and RoseTTAFold All-Atom, provide confidence scores like pLDDT values to assess foldability (Abramson et al., 2024; Krishna et al., 2024; Tunyasuvunakool et al., 2021). Higher confidence values generally indicate more promising and stable structures. 3) Functional site integrity: The accuracy of functional regions is critical for enzyme activity. Metrics such as the root-mean-square deviation (RMSD) of functional or catalytic sites ensure that predicted structures remain within acceptable thresholds of precision (Watson et al., 2023).

While these *in silico* metrics offer rapid and cost-effective evaluation, it is important to recognize their limitations. These metrics do not always correlate perfectly with experimental results, as computational predictions may fail to capture the full complexity of enzymatic functions under real conditions. Therefore, experimental validation remains a gold standard for confirming the accuracy, activity, and stability of

Table 1
Summary of recent studies in *de novo* enzyme design.

Computational tool	Method	Main architecture	Experimental validation case	Similarity to natural enzyme	Structural validation	k_{cat}/K_M ($M^{-1}\cdot s^{-1}$)	Ref
Family-wide hallucination	Structure-based scaffold design	Markov chain Monte Carlo sampling	Luciferase	N.A.	1.35 Å (AF2)	1.0×10^6	(Yeh et al., 2023)
RFam	Structure-based scaffold design	Flow	Metallohydrolase	N.A.	0.46 Å (AF2)	2.3×10^4	(Kim et al., 2024)
ChemNet	Structure-based active site design	Diffusion	Retroaldolase	N.A.	0.53 Å (XRC)	1.9×10^2	(Anishchenko et al., 2024)
PLACER	Structure-based active site design	Diffusion	Serine hydrolase	N.A.	0.83 Å (XRC)	2.2×10^5	(Lauko et al., 2025)
ZymCTRL	Sequence-based protein-trained models design	Transformer	Carbonic anhydrase / Lactate dehydrogenase	39.2 % / no hits found	N.A.	N.A.	(Munsamy et al., 2024)
ESM3	Sequence-based protein-trained models design	Transformer	Luciferase	51.4 %	N.A.	N.A.	(Hayes et al., 2025)
Evo	Sequence-based DNA-trained models design	StripedHyena	Cas9	79.9 %	N.A.	N.A.	(Nguyen et al., 2024)

N.A.: not available.
Structural validation: root-mean-square deviation of design structures and validated structures.
AF2: structure is validated by AlphaFold2.
XRC: structure is validated by X-ray crystallography.

designed enzymes (Bennett et al., 2023). By combining *in silico* assessments with experimental testing, researchers can ensure robust and reliable outcomes for *de novo* enzyme design.

3. *De novo* enzyme design

De novo enzyme design enables the creation of novel enzymes with defined structures and functions through two primary approaches: structure-based strategies, which use physical energy functions and spatial pattern algorithms to derive stable conformations from 3D constraints, and sequence-based strategies, which employ deep generative models to learn co-evolutionary patterns from protein datasets and generate functional sequences from data-driven principles. Unlike template-dependent methods such as directed evolution, which are limited to local exploration around natural proteins, *de novo* design allows access to vast and previously unexplored regions of sequence space. This capability facilitates the discovery of entirely new folds and catalytic mechanisms, as demonstrated by recent tools that have produced functional enzymes with minimal similarity to any known natural sequences (Table 1 and Fig. 3).

3.1. Structure-based design

Protein function fundamentally arises from the precise spatial organization of amino acid residues within an enzyme's 3D structure. Early physics-based studies demonstrated the feasibility of designing functional proteins from structural principles. As the field has advanced, structural modeling has become a powerful tool in *de novo* enzyme design, enabling the rational construction of proteins with tailored conformations and functional properties. Design strategies typically focus on engineering the scaffold and active site either independently or in an integrated manner.

3.1.1. Scaffold design

Scaffolds maintain the structural framework necessary to support the enzymatic active sites. Natural scaffolds are evolutionarily optimized to stabilize catalytic residues and reaction environments. For example, in cytochrome P450 systems, the organization of α-helical bundles is critical for heme coordination and substrate accessibility (Shaik et al., 2009). Initial scaffold design methods included parametric approaches and fragment assembly strategies. Parametric approaches, such as coiled-coil design, produced successful constructs like the α-helical tetramer 1RH4 using predefined geometric parameters, though structural diversity was limited (Harbury et al., 1998). The computational phase, beginning with zinc finger redesign in 1997, saw Rosetta-based methods apply backbone-sequence co-optimization and fragment assembly (*e.g.*, ERMS, SEWING), culminating in milestone designs like Top7 (Kuhlman et al., 2003; Liu et al., 1997). These frameworks, however, faced challenges in forcefield accuracy and conformational sampling.

Modern AI-driven approaches differ from traditional models by learning implicit sequence-structure-function relationships, expanding the designable protein space without requiring prior knowledge of active-site geometry. The Hallucination method pioneered the co-design of sequence and structure using trRosetta and Monte Carlo sampling. Though only 3 of 129 candidates were experimentally resolved by XRC/NMR and lacked functional activity, this validated the translatability of AI-predicted designs (Anishchenko et al., 2021). Similarly, SCUBA, based on kernel density estimation, also demonstrated folding accuracy comparable to Hallucination, but without functional outcomes (Huang et al., 2022). Family-wide hallucination extended this by preserving conserved domains within enzyme families and sampling flexible regions (Yeh et al., 2023). It yielded a 13.9 kDa artificial luciferase with catalytic efficiency. However, the generalizability of this approach is limited by its heavy reliance on comprehensive high-quality datasets detailing the sequence–structure–function relationships of the target enzyme family, resources which are often unavailable for many enzyme classes.

A breakthrough came with RFdiffusion, a RoseTTAFold2-based diffusion model that incorporates structure evaluation metrics (*e.g.*, backbone RMSD <1 Å, pAE < 5) for efficient enzyme design across EC 1–5 classes (Watson et al., 2023). However, this study lacked experimental confirmation of enzymatic activity and is computationally intensive during training and inference as a diffusion model fine-tuned from extensive structure prediction architectures.

To address computational load, newer models eliminate pretraining phases. SMCdiff allows conditional generation for functional site integration but is limited to 80-residue designs and lacks experimental validation (Trippe et al., 2023). FoldingDiff uses bond angle/dihedral parameterization with transformer architectures devoid of SE(3)-equivariance, achieving TM-scores of 0.83 ± 0.07 in unconditional generation tasks while improving scalability (Wu et al., 2024). However, the absence of experimental validation limits assessment of its real-world accuracy. RFam (Rosetta Flow Atomic Motif) enhances the RFdiffusion framework by implementing flow matching algorithms for scaffold generation that optimally position functional motifs, replacing

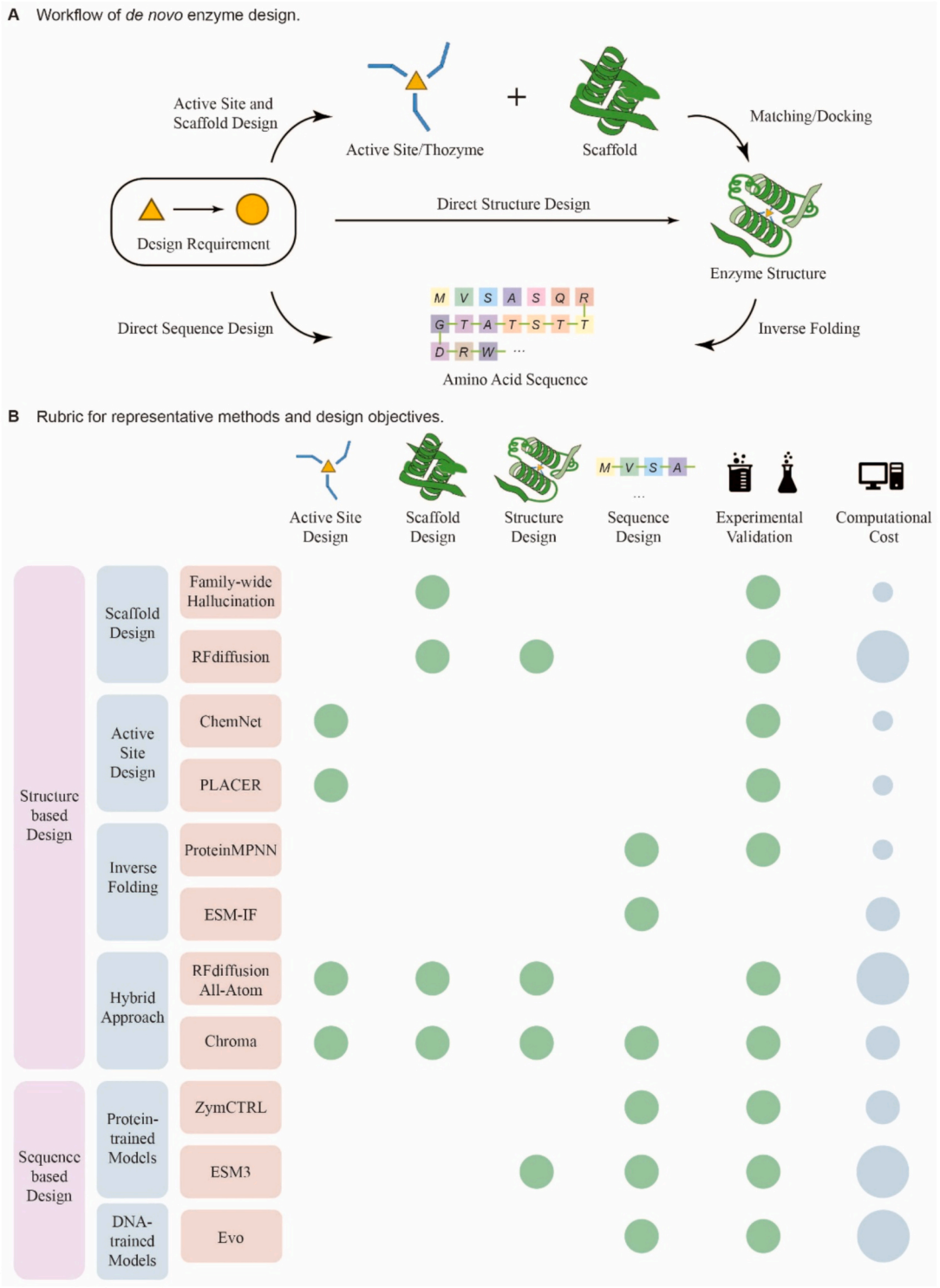

Fig. 3. Workflow and rubric of *de novo* enzyme design approaches. (A) Schematic overview of the *de novo* enzyme design process. The process begins with defining functional requirements, which guide two main design strategies: structure-based and sequence-based. In the structure-based approach, one route involves designing active sites or thozymes and embedding them into compatible protein scaffolds, followed by structural refinement using docking or matching algorithms. Alternatively, enzyme structures can be generated directly through *de novo* structure design. In both cases, the resulting 3D structures require inverse folding to derive compatible amino acid sequences. In contrast, sequence-based approaches generate functional protein sequences directly from the design objectives, without relying on explicit structural templates. (B) Rubric summarizing representative enzyme design tools and their objectives. Methods are grouped by design paradigm, including structure-based and sequence-based approaches. Each row represents a specific method, evaluated across up to five key steps in the enzyme design process: active site design, scaffold design, structure design, sequence design, and experimental validation. Filled green circles indicate which design steps are supported by each method. Computational cost is represented by a blue circle, with its size reflecting the relative computational demand. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

conventional diffusion approaches (Kim et al., 2024). This architecture implicitly samples sequence space and rotamer configurations during inference through transition-state complex modeling and catalytic group optimization. When applied to metallohydrolase design, experimental characterization of 96 constructs demonstrated catalytic efficiencies up to $k_{cat}/K_M = 23{,}000\ M^{-1}{\cdot}s^{-1}$ in top-performing variants.

While parametric and fragment-based methods have enabled robust scaffold engineering, AI-based scaffold design offers greater conformational diversity and novel fold exploration. Nonetheless, the predictive success of these ML methods requires rigorous experimental validation across diverse systems to establish reliability comparable to physics-based paradigms.

3.1.2. Active site design and matching

Active site design and matching constitute a core element of *de novo* enzyme engineering, aiming to recapitulate or enhance catalytic activity by integrating theoretical chemistry with structural biology. This process typically begins with the construction of quantum chemistry-derived theoretical enzyme models ("thozymes") that identify transition-state stabilizing motifs for a given chemical reaction. These motifs are then computationally embedded into structurally compatible natural or artificial scaffolds using platforms like RosettaMatch, followed by local microenvironment optimization *via* RosettaDesign or similar tools (Lovelock et al., 2022).

Notable successes include Kemp elimination and retro-aldolase designs. For example, a TIM-barrel scaffold derived from HisF protein achieved over 200-fold catalytic enhancement after seven rounds of directed evolution in the Kemp elimination system (Röthlisberger et al., 2008). Similarly, the integration of hashing algorithms into Rosetta enabled rapid screening and yielded 44.4 % functional activity across 72 theoretical models targeting retro-aldol reactions (Jiang et al., 2008). Additional applications of this strategy include the *de novo* design of FPP biosensors, digoxiginin and COMBS/heme binders, Diels-Alderases, formolases, serine hydrolases, zinc metalloenzymes, and heme enzymes, among others (Glasgow et al., 2019; Kalvet et al., 2023; Khare et al., 2012; Polizzi and DeGrado, 2020; Rajagopalan et al., 2014; Siegel et al., 2015; Siegel et al., 2010; Tinberg et al., 2013).

Although these methods remain foundational for their systematic treatment of transition-state stabilization and active-site optimization, current computational designs often exhibit notable performance gaps compared to natural enzymes. Key challenges persist, including the combinatorial explosion of hotspot configurations and binding poses, the complexity of concurrently designing active sites and selecting compatible scaffolds, and the reliance on limited structural databases for scaffold sourcing.

Recent efforts leveraging deep learning and geometric complementarity have begun to address these limitations. Lucas and Kortemme (2020) introduced a Rosetta-based side-chain reconstruction method for generating stable complex interfaces across thousands of targets, but systematic experimental validation remains lacking. Cao et al. (2022) employed the RIF (Rotamer Interaction Fields) framework to explore diverse binding modalities across targeted surface regions, implementing focused sampling around energetically favorable interactions, ultimately enabling the successful design of 14 high-affinity binders, although its application is still constrained by computational cost. Gainza et al. (2023) developed a geometric deep learning framework capable of mapping molecular surfaces and producing interaction fingerprints that capture spatial and chemical complementarity. While both studies represent major methodological progress, experimental success rate during validation remains modest.

Unlike small-molecule binding, enzymatic catalysis involves dynamic transition states and conformational changes that challenge geometric complementarity approaches. Addressing this, EnzymeFlow employs flow matching and co-evolution strategies to generate active pockets with dynamic properties, demonstrating computational advantages over conventional algorithms (Hua et al., 2024). However, limitations persist, including the untested reliability of its CLEAN classifier and the lack of experimental validation. In contrast, ChemNet achieves *in vitro* validation using diffusion model-driven binding site inverse design (Anishchenko et al., 2024). The best-performing variant reached a catalytic efficiency of 11,249.4 M^{-1} min^{-1}, surpassing early-stage directed evolution outputs by several orders of magnitude. Still, both ChemNet and EnzymeFlow do not explicitly model dynamic interactions or provide mechanistic insight into transition-state coordination.

In the realm of conformational engineering, two distinct methodologies have advanced dynamic enzyme design through specialized strategies. Meta-multistate design employs polymorphic landscape prediction to identify sequences capable of spontaneous transitioning between predefined states. This approach was exemplified by DANCER, which validated conformational exchange between two novel global folding states in the β1 domain of streptococcal protein G (Davey et al., 2017). However, the broader applicability of this method remains to be fully established. In contrast, PLACER adopts a statistically driven approach that achieves catalytic precision through stepwise simulation, integrating active-site modeling, scaffold generation, and conformational screening (Lauko et al., 2025). This framework successfully enabled the design of a serine hydrolase, with its four-step enzymatic process confirmed by XRC. Although the engineered enzyme exhibits lower catalytic efficiency compared to its natural counterpart, it provides valuable insights into conformation-aware design paradigms.

ML approaches offer promising avenues to overcome the computational inefficiencies, energy function limitations, and vast search space inherent in conventional computational chemistry methods. However, successful integration requires rigorous validation within robust biophysical and mechanistic frameworks. A persistent challenge lies in achieving precise quantum chemical characterization of reaction dynamics, particularly in accurately modeling electronic interactions and transition states that govern catalytic efficiency.

3.1.3. Inverse folding

Inverse folding is a computational protein design approach that systematically generates amino acid sequences capable of adopting predetermined 3D protein structures. This is fundamentally challenged by the vast conformational space that must be explored. This bottleneck was partially addressed by the Dead-End Elimination (DEE) algorithm, which systematically prunes suboptimal rotamer combinations using mathematical criteria, thereby reducing the dimensional complexity of the design space (Georgiev et al., 2008). By incorporating energy minimization through force field optimization, DEE formed the computational foundation for early sequence design tools such as Orbit, which implemented the DREIDING force field and remains a reference in benchmarking efforts (Bolon and Mayo, 2001; Mayo et al., 1990).

Subsequent innovations, including Rosetta, have advanced *de novo* protein design by integrating sequence optimization with structural prediction (Leaver-Fay et al., 2011). However, practical applications reveal ongoing limitations: despite success in selected cases, the overall sequence recovery rates remain low, indicating insufficient predictive power for robust and generalizable protein design.

To address these shortcomings, hybrid frameworks integrating statistical principles into physics-based models have emerged. For example, Xiong et al. (2014) introduced statistically-derived energy functions (SEFs) using an SSNAC-based framework, successfully validating the structural foldability of four designed sequences within the TEM1-β-lactamase system. Nevertheless, this approach lacks atomic-level refinement precision and has limited applicability to short-chain proteins.

Recent advances in sequence design predominantly rely on data-driven methods. ProteinMPNN, which leverages backbone geometry and Gaussian noise augmentation, achieved a 52.4 % sequence recovery rate, significantly outperforming Rosetta (Dauparas et al., 2022). However, its accuracy is sensitive to structural resolution, limiting its

robustness. In contrast, Frame2Seq employs invariant point attention to integrate multilevel representations derived from 1D dihedral embeddings, 2D inter-residue distances, and 3D Cartesian coordinates (Akpinaroglu et al., 2023). It achieves enhanced sequence recovery and sixfold acceleration on the CATH4.2 benchmark, with experimental validation demonstrating the generation of 22 soluble *de novo* proteins, including zero-homology monomeric scaffolds. However, these validations rely on native crystal backbones, leaving the performance of synthetic novel scaffold designs untested.

Both ProteinMPNN and Frame2Seq depend heavily on high-resolution structural datasets. To mitigate this data scarcity, ESM-IF incorporates AlphaFold2-predicted pseudo-structures, filtered using pLDDT confidence metrics and enhanced through noise injection (Chloe et al., 2022). While this improves model generalizability, its design outputs remain unvalidated experimentally. In contrast, CarbonDesign inverts AlphaFold's Evoformer information flow to extract backbone features using an Inverseformer module, followed by sequence decoding *via* an amortized Markov random field (Ren et al., 2024). This architecture outperforms existing mainstream models in CASP15 benchmarks but also lacks wet-lab corroboration.

Although ML-based inverse folding has significantly improved sequence recovery and design specificity, three persistent challenges remain: (1) Limited availability of high-quality structural data constraining model generalizability; (2) Evaluation frameworks focused on native crystal structures, which do not reflect algorithmic capacity for novel scaffold generation; and (3) Lack of closed-loop experimental verification, hindering validation and refinement of designed sequences. To overcome these limitations, future research could prioritize: (1) Generation of diverse high-resolution datasets; (2) Development of multi-scale assessment metrics from atomic to functional levels; and (3) Integration of computational design with experimental pipelines, enabling iterative refinement and validation. These directions are essential for bridging the gap between *in silico* design and biologically viable proteins.

3.1.4. Hybrid approaches

Hybrid design strategies enable simultaneous optimization of protein scaffolds and functional sites. One notable approach is the TERM (Tertiary Motif) methodology, which defines designable units as unions of local backbone fragments and structural motifs derived from known interaction interfaces (Zhang and Grigoryan, 2013). By constraining the search space to evolutionarily validated geometries, TERM-based approaches allow effective exploration of secondary, tertiary, and quaternary contexts, facilitating both structural and functional design *via* fragment libraries.

Frappier et al. (2019) applied TERM-derived statistical potentials to predict peptide binding affinities within the Bcl-2 protein family, successfully designing high-affinity binders for Bfl-1 and Mcl-1. However, the effectiveness of such template-dependent approaches is contingent on the availability and diversity of structural databases used for motif extraction.

In contrast, recent ML-based methods offer template-independent, full-atom design capabilities. A breakthrough example is RFDiffusion All-Atom, which extends RosettaFold All-Atom to support simultaneous modeling of protein-ligand interactions during generation (Krishna et al., 2024). Despite this advancement, it depends on external tools to design the corresponding sequence and incurs high computational costs.

Subsequent innovations attempt to address these challenges. ProteinGenerator utilizes discrete diffusion processes to co-optimize backbone conformations and residue identities, demonstrating validated applications in stability and activity engineering (Lisanza et al., 2024). However, scalability remains a limitation. Chroma, leveraging graph neural networks and quasilinear algorithms, extends design to complex protein assemblies, with initial validation *via* split-GFP systems (Ingraham et al., 2023). Nonetheless, catalytic enzyme design within this framework remains largely unexplored. Protpardelle introduces a continuous-space diffusion model with hyper-conformation mechanisms, enabling backbone-sidechain co-optimization and conditional generation (Chu et al., 2023). Despite strong theoretical performance, experimental validation of its predictions is still pending.

ML approaches have introduced innovative frameworks for guiding hybrid enzyme design toward specific functional objective. However, three persistent limitations remain across current implementations: (1) the high computational cost of full-atom modeling limits large-scale screening efficiency; (2) structural innovations from design algorithms often conflict with the evolutionary folding constraints of natural enzymes; and (3) experimental validation systems lag behind the rapid iteration cycles of computational model development.

3.2. Sequence-based design

From a sequence-level perspective, amino acid chains can be conceptualized as a biological language, encoding structural and functional information in a hierarchical manner. Analogous to human language, where letters form words and words form sentences, amino acids act as "letters," motifs resemble "words," and domains represent "sentences" that convey specific biochemical functions (Ferruz et al., 2022). This linguistic analogy has paved the way for applying natural language processing (NLP) techniques to protein sequence analysis. Language models (LMs), originally developed for textual data, have become foundational in this field by learning the statistical relationships between amino acids. These models can capture grammatical, syntactic, and semantic features of protein sequences, thereby enabling the prediction, generation, and optimization of functional sequences for enzyme design (Cambria and White, 2014).

3.2.1. Bioinformatics-based methods

Bioinformatics-based strategies for enzyme amino acid sequence design primarily include consensus design and co-evolution-based design. The consensus design approach is based on the premise that evolutionarily conserved residues play a central role in maintaining protein stability (Porebski and Buckle, 2016). This strategy involves replacing variable positions with consensus residues to generate stabilized variants. It has been effectively applied in antibody engineering and later adapted for LM-guided optimization. However, practical implementation remains limited by the labor-intensive identification of consensus sequences and reliance on expert-driven heuristics, often resulting in suboptimal success rates.

By contrast, co-evolution-based design leverages patterns of evolutionary covariation, where residues or subunits that interact tend to co-mutate in a correlated fashion. These reciprocal mutational dependencies, observed across intra- and inter-molecular interfaces, enable prediction of structurally compensatory mutations (Juan et al., 2013). Co-evolutionary data, extracted from multiple sequence alignments, are commonly used as critical input features in modern predictive modeling pipelines and facilitate the evaluation of mutation-induced structural perturbations and long-range conformational effects.

3.2.2. Protein-trained models

Attention-based models have demonstrated strong capabilities in capturing co-evolutionary relationships within protein sequences. One notable example is ProteinGAN, a self-attention–based generative adversarial network validated through redesign of malate dehydrogenase (MDH) (Repecka et al., 2021). Out of 16 generated candidates, 3 exhibited catalytic activity as soluble enzymes. Despite this promising result, validation remains limited to the MDH family, and the method faces challenges such as unstable training convergence and mode collapse during adversarial optimization.

Inspired by breakthroughs in NLP, Transformer-based architectures have become dominant in protein modeling. ProtGPT2 adapted the GPT-2 model for unconditional generation of stable protein sequences, though it lacked functional specificity (Ferruz et al., 2022). Later models

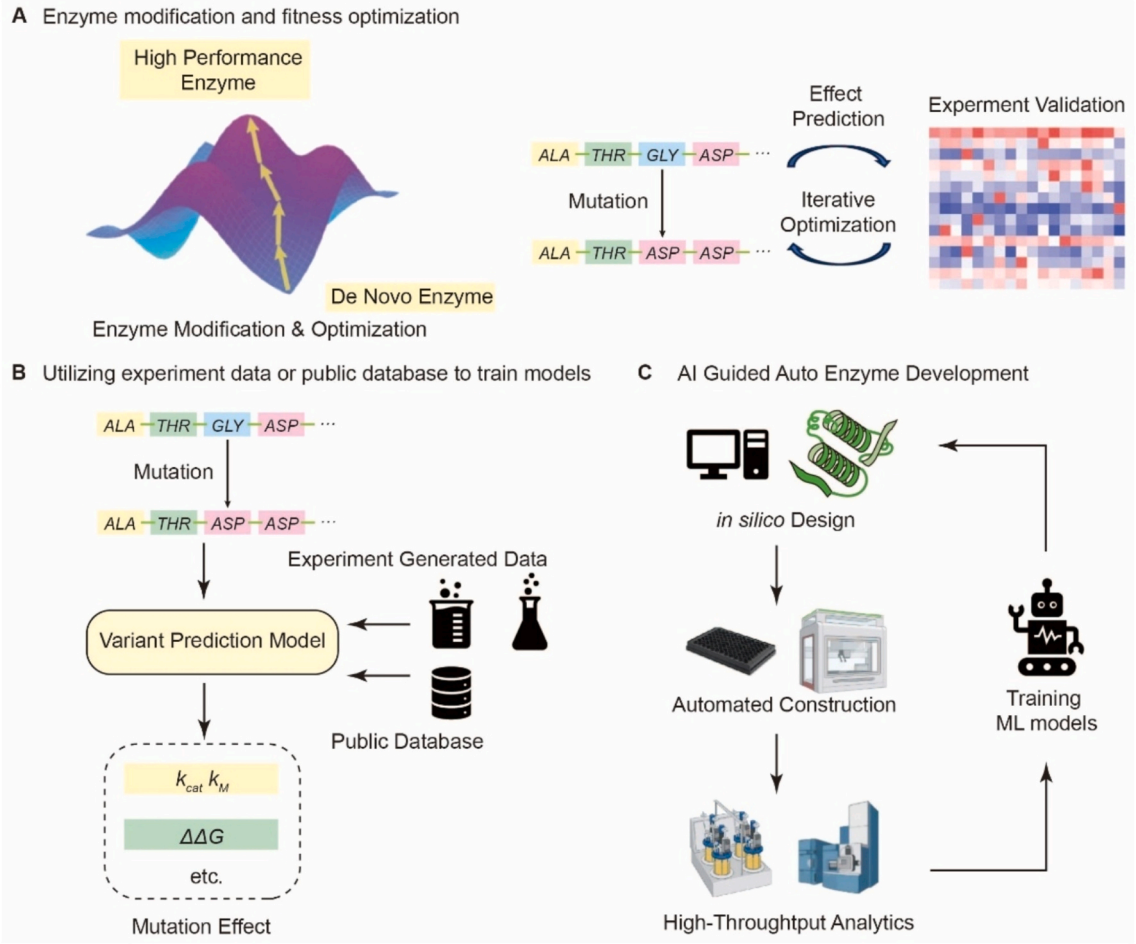

Fig. 4. Enzyme modification and optimization. (A) ML methods can be used to model the relationships between mutational scenarios and their effects, and optimize the enzyme fitness, beneficial modifications will assign higher scores. (B) ML-based variant prediction models can be trained from experimental data or public datasets without additional experimental data. (C) ML-based automation solutions can automatically update experimental strategies online based on experimental results.

introduced target-guided designs. For instance, ProGen, based on the CTRL architecture, achieved a 72 % expression success rate in lysozyme engineering by conditioning sequence generation on descriptive tags (Madani et al., 2023). ZymCTRL extended this approach by guiding sequence design with enzyme commission (EC) codes, yielding experimentally validated results in carbonic anhydrase and lactate dehydrogenase systems (Munsamy et al., 2024).

The most recent innovation, ESM3, introduces cross-modal modeling by jointly tokenizing sequence, structure, and function into a unified representation. It successfully designed novel luciferases with native-level efficiency despite sharing only 53 % sequence identity with their closest natural homologs (Hayes et al., 2025). However, its effectiveness in protease and other enzyme classes remains to be validated, and further studies are needed to assess its generalizability in functional enzyme engineering.

3.2.3. *DNA-trained models*

Recognizing that protein sequences are encoded in DNA sequences, recent efforts have explored generative frameworks trained directly on genomic data. These approaches aim to capture regulatory, evolutionary, and functional signals embedded in nucleotide sequences, thereby offering a novel route for enzyme design.

Outeiral and Deane (2022) demonstrated the potential of self-supervised pretraining on large-scale DNA datasets, showing competitive performance across diverse protein characterization tasks. However, this strategy currently lacks systematic experimental validation, and its direct applicability to enzyme design remains preliminary.

The Evo model introduced a more specialized architecture StripedHyena, which combines rotary attention with Hyena operators to efficiently model long-range dependencies in DNA (Nguyen et al., 2024). Trained on prokaryotic and phage genomes at single-nucleotide resolution and fine-tuned on CRISPR-Cas data, Evo successfully designed novel Cas9 variants with comparable activity to SpCas9 (sharing 79.9 % sequence identity).

Building on this, Evo2 expanded the training scope to cross-species genomic datasets and adopted StripedHyena2 for megabase-scale modeling (Brixi et al., 2025). The incorporation of sparse autoencoders (SAEs) further improved interpretability by highlighting latent regulatory patterns. Experimental validation of Evo2-designed protein variants is ongoing, demonstrating the promise of DNA-based generative models in bridging sequence-to-function relationships across biological scales.

4. Filtering designs and evaluating mutations

Modifying and optimizing *de novo* designed enzymes are crucial. It is popular to utilize well-established enzyme engineering strategies to enhance the stability, efficiency, and other fitnesses of *de novo* designed enzymes, which may lead the enzymes to achieve expected performances for industrial applications. For instance, after obtaining *de novo* designed enzymes, variations can be introduced and screened through techniques such as error-prone PCR, followed by iterative selection. This

Table 2
Summary of recent studies on function prediction tools.

Computational tool	Method	Main architecture	Accuracy	Test dataset	Ref
DEDAL	Integrate ML with alignments	Transformer	F1 score: 0.877	Pfam	(Llinares-López et al., 2023)
DeepBLAST	Integrate ML with alignments	Transformer	TM-score prediction median error: 0.026	SWISS-MODEL	(Hamamsy et al., 2024)
DeepEC	Predict EC	CNN	Precision: 0.920, Recall: 0.455, Prediction time: 13 s	Swiss-Prot	(Ryu et al., 2019)
CLEAN	Predict EC	Transformer	Precision: 0.597, Recall: 0.481, F1 score: 0.499	Swiss-Prot	(Yu et al., 2023)
DeepGOPlus	Predict GO	CNN	F_{max}: 0.544, S_{min}: 8.724, AUPR: 0.487	CAFA3 molecular function	(Kulmanov and Hoehndorf, 2020)
DeepGO-SE	Predict GO	Transformer	F_{max}: 0.554, S_{min}: 11.681, AUPR: 0.552, AUC: 0.874	UniProtKB/Swiss-Prot	(Kulmanov et al., 2024)

F1 score: harmonic mean of precision and recall, indicating overall classification performance.
TM-score: template modeling score, a measure of structural similarity between predicted and native protein structures; values closer to 1 indicate higher accuracy.
Precision: the proportion of true positives among all predicted positives.
Recall: the proportion of true positives identified among all actual positives.
F_{max}: the maximum F1 score across all decision thresholds.
S_{min}: minimum semantic distance between predicted and true annotations in the Gene Ontology graph; lower values indicate better performance.
AUPR: area under precision–recall curve, summarizes the trade-off between precision and recall across thresholds.
AUC: area under the ROC curve, measures the ability of the model to distinguish between classes; higher values reflect better classification.
Prediction time: time required to generate a prediction per input.

method not only boosts the catalytic activity of enzymes but can even endow them with entirely new catalytic functions (Basler et al., 2021; Crawshaw et al., 2022). However, for many enzymes, high-throughput experimental design remains a challenge, and multi-round iterative screening protocols result in excessively long experimental cycles, which pose a significant burden on researchers. In addressing these technical bottlenecks, ML methods demonstrate substantial advantages. After establishing complex functional relationships between input and output data, models can be leveraged to explore sequence spaces beyond the training set, facilitating the enrichment of beneficial mutations as highlighted in Fig. 4 and Table 2 (Mazurenko et al., 2020).

4.1. *Variant predictions*

Variant prediction is one of the major tasks of enzyme optimization. The most direct strategy is collecting data on mutational effects through experiments, and ML methods can be utilized to model the relationships between variants and their effects, thereby exploring novel and beneficial mutational sites. Early on, Fox et al. (2007) validated mutations using experimental techniques such as random or site-directed mutagenesis, and these mutated variants were screened and sequenced to collect data for fitting a linear model *via* Partial Least Squares (PLS) regression. Based on the magnitude and sign of each factor's contribution in the linear model, mutations were classified into beneficial, neutral, and deleterious categories, followed by fixing, re-testing, or removing them accordingly, which resulted in an approximately 4000-fold increase in volumetric productivity for a nitrile-catalyzing protein reaction over 18 rounds of evolution. Wu et al. (2019) employed ML to predict libraries enriched with functional enzymes and fixed seven mutations within two rounds of evolution, identifying selective catalytic variants with 93 % and 79 % enantiomer excess (ee). Huang et al. (2024) evaluated the effect of the short-chain dehydrogenase/reductase BsSDR from *Bacillus subtilis* on promoting the kinetic resolution of (±)-tetraphenazine to dihydrotetraphenazine using both traditional directed evolution and ML-assisted directed evolution methods, and both approaches successfully identified variants with significantly enhanced diastereoselectivity for each isomer of dihydrotetraphenazine. In addition, engineering methods were implemented, achieving an isolated yield of 40.7 % and a diastereoselectivity of 91.3 %.

Other strategies focus on providing a universal prediction scheme for mutational effects across all enzymes. Thanks to the rapid development of NLP, unsupervised learning has been comprehensively applied in the prediction of enzymatic mutational effects. It allows pre-training on unlabeled datasets and can directly perform mutant prediction tasks on specific proteins without additional training, which is known as zero-shot. The model primarily relies on whether the mutant conforms more closely to the rules learned by the model compared to the wild type. This set of rules assesses mutants more from an evolutionary perspective or the perspective of naturally occurring proteins, assigning higher scores to mutation models that adhere to evolutionary rules and resemble naturally occurring proteins.

Meier et al. (2024) developed ESM-1v based on the ESM, which can directly perform unsupervised learning on protein mutants, and the training method is random masking, enabling the model to predict the residue type of the masked part based on the unmasked part. This allows the model to assess the conservativeness of amino acids in proteins, assigning positive scores to mutants that conform more closely to "reasonable" types. Another strategy is supervised learning, which can more accurately predict the properties of protein mutants by learning mutation data for a specific protein compared to unsupervised strategies. Considering that high-quality labeled data is limited and unsupervised models have already learned the evolutionary rules of proteins through training, introducing unsupervised models as encoding modules into supervised models can ensure more accurate predictions of specific protein mutant properties. The ESM-1b model uses a 34-layer transformer pre-trained on the UR50/S database and then fine-tuned using mutation data for specific proteins, achieving higher accuracy compared to previous methods (Rives et al., 2021).

ML exhibits a remarkable proficiency in efficiently recognizing patterns from extensive databases of existing proteins, thereby demonstrating substantial use in exploring protein fitness landscapes. Furthermore, refining the representations of protein variants, incorporating predictive uncertainties, and developing specialized ML models with protein-specific inductive biases can significantly enhance the accuracy of sequence-to-fitness mappings. An intriguing future direction involves the integration of protein fitness information directly into generative models during the *de novo* design phase, potentially eliminating the need for a separate optimization step. However, challenges remain, including the scarcity of high-quality, diverse training datasets and the need for robust model interpretability. Addressing these gaps will involve constructing more comprehensive databases, improving uncertainty quantification, and exploring joint optimization

Table 3
Summary of recent studies on design filtering and evaluating.

Aim	Computational tool	Main architecture	Accuracy	Ref
Integrate ML with docking algorithms	GNINA	CNN	Top-1 accuracy: 73 % for redocking and 37 % for cross-docking when the binding pocket is predefined	(McNutt et al., 2021)
	KarmaDock	GNN	Ligand pose success rate: 89.1 %, Generation time: 0.017 s/complex	(Zhang et al., 2023)
Predict enzyme function from sequence information	ESP	Transformer	Accuracy: 91.5 %, ROC-AUC: 0.956, MCC: 0.78	(Kroll et al., 2023)
	ProSmith	Transformer	Accuracy: 94.2 %, ROC-AUC: 0.972, MCC: 0.85	(Kroll et al., 2024)
Predict enzyme function from structure information	Alphafold3	Diffusion, Transformer	Structural accuracy: 73.1 % structures achieved <2 Å RMSD and passed PB-Valid on the PoseBusters V2 benchmark	(Abramson et al., 2024)
	RoseTTAFold All-atom	Diffusion, Transformer	Blind docking evaluation: 77 % of high-confidence structures had <2 Å RMSD in the CAMEO ligand-docking benchmark	(Krishna et al., 2024)

Top-1 accuracy: percentage of predictions where the top-ranked ligand pose matches the reference pose within a predefined RMSD threshold.
Ligand pose success rate: proportion of ligand poses generated within an acceptable RMSD of the native pose.
Generation time: time required to predict a ligand pose for one protein–ligand complex.
ROC-AUC: receiver operating characteristic – area under curve, measures the ability of a model to distinguish between classes; values closer to 1 indicate better classification performance.
MCC: Matthews correlation coefficient, a balanced metric for binary classification performance; values range from −1 (total disagreement) to +1 (perfect prediction).
RMSD: root-mean-square deviation, measures the average atomic deviation between predicted and reference structures; <2 Å is commonly considered accurate.
PB-Valid: PoseBusters validity, a structural validation score from the PoseBusters benchmark that assesses chemical and geometric plausibility of predicted protein–ligand complexes.
CAMEO: continuous automated model evaluation, a blind benchmarking platform that evaluates the accuracy of protein structure and docking predictions using unseen experimental data.

frameworks. Such efforts could greatly enhance the accuracy, efficiency, and scalability of ML-driven protein engineering, ultimately accelerating the discovery and development of novel proteins with desired functionalities.

4.2. Stability optimization

Stability is also pivotal for engineered enzymes and their biotechnological applications. Although the prediction of enzyme variation effects can, to some extent, guide the stability improvement of enzymes, its unique importance in industrial production requires models specifically dedicated to improving their stability.

Traditional enzyme engineering strategies often rely on force field-based methods, such as FoldX, RosettaDDG, and PROSS, to assess the stability of designed protein variants (Leaver-Fay et al., 2011; Peleg et al., 2021; Schymkowitz et al., 2005). These tools have shown empirical success in improving protein stability through energy minimization (Floor et al., 2014; Luo et al., 2016; Wahab et al., 2012). However, their effectiveness is constrained by a strong dependence on accurate empirical energy functions and the high computational cost associated with exhaustive conformational sampling.

Recent advances have established machine learning frameworks as powerful computational pipelines for protein stability prediction through multi-feature integration. DeepDDG demonstrates this paradigm by combining geometric configurations (secondary structures, residue spatial relationships), evolutionary sequence features (position-specific scoring matrices), and physicochemical descriptors to predict mutational effects (Cao et al., 2019). Subsequent innovations like DDMut employ graph neural networks to model local 3D microenvironments, enabling precise ΔΔG estimation through atomic interaction patterns and residue accessibility parameters (Zhou et al., 2023). The GeoStab suite further advances this field *via* geometric deep learning architectures that systematically quantify stability metrics, including fitness scores (GeoFitness), free energy changes (GeoDDG), and thermal denaturation thresholds (GeoDTm), through unified representations of structural geometry and sequence covariation (Xu et al., 2024a, 2024b). These approaches collectively enhance predictive accuracy while maintaining computational efficiency across diverse protein engineering applications. ThermoMPNN integrates pre-trained ProteinMPNN embeddings as transfer-learning features with sequence recovery data and measured mutational stabilities, achieving optimal performance across benchmarking evaluations *in silico* (Dieckhaus et al., 2023).

More recent efforts have integrated machine learning frameworks to tackle the multidimensional challenge of balancing enzyme thermostability with catalytic activity. For example, Cui et al. (2024) developed a hybrid approach that combines protein language models with physics-based algorithms, leading to the design of TurboPETase, a PET hydrolase exhibiting enhanced operational stability under industrial substrate loads (200 g/kg). Despite its success, this integrative strategy still relies on resource-intensive computations for multi-objective optimization during sequence selection, limiting its scalability for large-scale design applications. Sumida et al. (2024) established a computationally efficient ProteinMPNN-based statistical framework integrating evolutionary and structural features to optimize protein stability and catalytic activity, achieving 26-fold functional enhancement in myoglobin designs with an elevated melting temperature of 84 °C.

4.3. Automation

The automation of enzyme engineering has increasingly become interesting. Since the design methods for mutation sites can be intricate, which often require multiple rounds of iterative optimization, long experimental cycles, and challenging data analysis, most current approaches remain low-throughput and labor-intensive (Newton et al., 2018). However, ML-based automation solutions can update experimental strategies online based on experimental results, thereby improving the efficiency of enzyme engineering.

Wang et al. (2023) developed EvoPlay based on the single-player version of the AlphaZero self-play reinforcement learning framework. They treated mutations of individual site residues as actions for optimizing amino acid sequences, analogous to moving pieces on a chessboard. The policy-value neural network collaborates with the lookahead Monte Carlo tree search to guide the optimization agent both broadly and deeply as well. EvoPlay was utilized to design luciferase and discover variants with 7.8-fold higher bioluminescence than the wild type. Orsi et al. (2024) employed ML to guide automated workflows, including library generation, implementation of hypermutation systems, adaptive laboratory evolution, and *in vivo* growth-coupled selection, thereby enabling rapid and automated selection and optimization of experimental conditions to accelerate the development of processes for directed enzyme evolution.

Recent developments by Singh et al. (2025) demonstrate an autonomous protein engineering platform integrating machine learning architectures with large language models and automated instrumentation.

This closed-loop system enabled the successful rational design of methyltransferase variants exhibiting 90-fold enhanced substrate specificity and 16-fold increased activity, concurrent with computational optimization of phytase demonstrating 26-fold catalytic improvement under neutral pH conditions. The four-week engineering cycle employed combinatorial mutagenesis strategies across four iterative rounds, systematically evaluating <500 enzyme variants per target through high-throughput characterization protocols.

The enhancement of enzyme fitness is anticipated to be transferred into a fully automated procedure, with ramifications spanning numerous industrial sectors. These iterative optimization cycles possess the potential to facilitate the continual refinement of enzymes, analogous to the success achieved in small molecule optimization. ML is poised to drive such automated systems, leveraging AI's flexibility and adaptability to execute novel syntheses and screenings through dynamically generated robotic scripts. A key development direction lies in optimizing multiple desirable properties and activities simultaneously during enzyme engineering efforts. ML models capable of integrating multimodal representations such as sequence, structure, and functional data can facilitate this multi-objective optimization.

However, challenges remain in improving interpretability and addressing the trade-offs among competing objectives. Overcoming these challenges will enable the creation of efficient, scalable, and autonomous enzyme engineering workflows, fostering innovations across synthetic biology, industrial biotechnology, and beyond.

5. Prediction of function and functional sites

5.1. Molecule structure and interaction validation

The structure and interaction of the enzyme-substrate complex are crucial for validating the functions and catalytic efficiency of the designed enzyme. Although experimental determination remains the most reliable method currently for defining the complex of structures and interactions, this entails significant experimental costs. Despite the fact that even the most advanced structure prediction methods today cannot accurately predict structures for all enzyme-substrate complexes, nor guarantee complete accuracy in pocket localization, they can still serve as rapid structure assessment tools at low cost, which can be utilized in high-throughput analysis and screening (Table 3). Traditional computational prediction methods primarily rely on physical principles, such as docking methods. Docking employs predefined energy functions to evaluate the structures and interactions (Trott and Olson, 2010). However, this approach is sensitive to structure fluctuations, which struggle to adequately address protein flexibility. Besides, it is challenging to assess the contributions of conformational entropy and solvents (Zheng et al., 2020).

5.1.1. Combining ML with docking algorithm

A strategy that integrates ML with docking methods has been explored. GNINA utilizes an ensemble of convolutional neural networks (CNN) for scoring function evaluation, which not only significantly accelerates the molecular docking process but also surpasses AutoDock Vina in terms of accuracy (McNutt et al., 2021). Zhang et al. (2023) developed KarmaDock, which comprises three components. The first component is an encoder for proteins and ligands, which is designed to learn representations of intramolecular interactions. The second component is an equivariant graph neural network with self-attention, which updates ligand poses based on protein-ligand interactions and intramolecular interactions. The third component is a hybrid density network that scores binding affinity. However, like traditional molecular docking methods, this strategy needs the user to provide the protein structure.

5.1.2. Predictions based on sequence and structure information

AI strategies can also focus on integrating information about small molecules, enzymes, and their interactions into a unified framework (Tsubaki et al., 2019). Small molecules typically contain fewer than 100 heavy atoms and occupy a relatively small structural space. This allows current ML techniques to accurately predict structure properties from their linear representations (Jastrzębski et al., 2016). Similarly, the functions of enzymes can be predicted based on their linear representations of amino acid sequences. Kroll et al. (2023) developed ESP, which employs ESM to encode amino acid sequences, and through supervised training of a binary classification task based on the ESM embeddings, ESP outputs a score representing the possibilities of binding between the enzyme and substrate. Subsequently, the same group further developed a multi-modality transformer network to simultaneously process both amino acid sequence and substrate string from the same input (Kroll et al., 2024). This approach enables a more efficient exchange of all relevant information between the two molecular types during the computation of their latent representations, thereby predicting interactions.

However, the physical nature of the enzyme-substrate interaction is still based on their structures rather than sequences. Predicting the transition from one-dimensional (1D) sequences to 3D structures is challenging, as 1D amino acid sequence representations may not suffice to capture the structural features of 3D space that determine interaction predictions. Therefore, strategies are needed to predict the structures of enzyme-substrate complexes. Both AlphaFold3 and RoseTTAFold All-Atom employ diffusion models as one of their main frameworks, achieving atomic-level accuracy in predicting the structures of complexes (Abramson et al., 2024; Krishna et al., 2024). High-accuracy predictions have brought new tools and insights into enzyme research, which can be utilized to conduct preliminary analyses of interactions within functional sites instead of the costly experimental structural characterization. Their high performance has also become one of the important tools for constructing *in silico* metrics to be employed by many *de novo* design models. However, these models primarily rely on amino acid sequences and evolutionary information to infer structures, which limits their ability to address the heterogeneity of various molecular types.

Recent advances in AI techniques have significantly improved the prediction of structures and the interactions of enzyme-substrate complexes, enhancing the understanding of static binding configurations. However, the actual interactions are inherently dynamic, involving complicated conformational changes, transient states, and energy landscapes that unfold over time. Accurately modeling these processes is critical for understanding enzyme functionality, but remains a significant challenge for ML. Key hurdles include capturing time-dependent behaviors, adapting to large-scale conformational shifts, and integrating physicochemical constraints. Hybrid frameworks integrating physics-based simulations with scalable AI could overcome these obstacles, improving interpretability and accuracy. Addressing these challenges can advance AI-driven enzyme *de novo* design, offering deeper insights into dynamic mechanisms and paving the way for more precise and efficient protein engineering.

5.2. Function validation for enzymes

The function validation of the *de novo* designed enzyme is crucial to identify the effectiveness of the *de novo* design. Besides, it is also helpful to screen the most promising design results and enhance the success of the designs. Traditional enzyme function prediction primarily relies on sequence similarity or structural similarity (Lipman and Pearson, 1985). Sequence-based methods adopt an evolutionary perspective, where enzymes with homologous sequences share similar functions. By constructing phylogenetic trees from sequences to explore the evolutionary relationships, sequence homology and functional similarity can be inferred. However, these methods generally consider a similarity threshold of over 60 % between sequences to achieve a certain level of reliability (Camacho et al., 2009). Furthermore, there is no reliable

correlation between homology and protein function, as sequences with low similarity may still exhibit similar functions (Punta and Ofran, 2008). Structure-based prediction methods mainly assume that enzymes with similar spatial structures often have identical functions. Yet, the precise structures of most proteins remain unknown, and most existing structural alignment tools are computationally intensive, requiring brute-force all-against-all comparisons to search for structurally similar proteins, which hinders large-scale applications (Hamamsy et al., 2024).

5.2.1. Combining ML with alignment algorithm

Recently, efforts have been dedicated to enhancing the accuracy of sequence alignment by integrating ML methods with traditional alignment algorithms. These strategies build upon classic alignment algorithms, such as the Smith-Waterman (SW) algorithm, which rely on a predefined scoring function for dynamic programming to align sequences (Smith and Waterman, 1981). However, these scoring functions face significant limitations as they are not adaptable to all alignment scenarios. Furthermore, in cases of low sequence similarity, scoring functions based solely on sequence similarity may fail to effectively predict function with low sequence similarity. ML strategies, on the other hand, employ data-driven approaches to dynamically predict scoring functions, rather than relying on fixed ones, which explore functional similarities in the functional space, rather than sequence or structural similarities.

Llinares-López et al. (2023) leveraged the transformer architecture to improve the efficiency and effectiveness of the traditional SW alignment algorithm, and the transformer was first used to extract embeddings of the sequences to be aligned, which was then used to determine critical parameters of the SW algorithm: gap open penalty (O), gap extend penalty (E), and substitution scores (S). To enable end-to-end training, the SW algorithm was also modified into a differentiable form to perform backpropagation. Hamamsy et al. (2024) developed the TM-Vec and DeepBLAST suite. Firstly, TM-Vec was utilized to search for structure similarities within large sequence databases, which was trained to directly predict TM scores from sequence pairs as a metric of structure similarity, without the need for intermediate calculations or resolved structures. Subsequently, once structurally similar proteins were identified, DeepBLAST can structurally align proteins using only sequence information, identifying structurally homologous regions between proteins.

ML-based alignment algorithms have significantly enhanced prediction accuracy when dealing with sequences with low similarity. However, akin to traditional alignment algorithms, the prerequisite for employing this strategy to determine functionality is the presence of enzymes with the target function in the current database. This could potentially limit its capacity to predict entirely novel functions.

5.2.2. Predicting function annotations

Apart from improving traditional alignment methods, another strategy is to directly predict enzyme functional annotations. EC numbers and GO (Gene Ontology) annotations are currently prevalent methods for enzyme functional annotation. The EC number system, developed by the Enzyme Commission, classifies enzymes based on the chemical reactions they catalyze, with each EC number corresponding to a specific enzymatic reaction. On the other hand, GO annotations encompass thousands of terms, covering various functions and locations of proteins within cells and organisms. They categorize gene functions into three parts: cellular components, molecular functions, and biological processes, providing a systematic approach to represent and share knowledge about the functions and processes of genes within organisms.

Recently, due to the widespread popularity of transfer learning and pre-training-fine-tuning workflows, LMs can be utilized to perform self-supervised training methods and map amino acid sequences into embeddings in functional space. By constructing shallow networks, or fine-tuning based on these pre-trained embeddings, efficient enzyme function predictions can be made. Yu et al. (2023) developed CLEAN, which computed embeddings of enzyme amino acid sequences using ESM and added additional mapping layers. Subsequently, through contrastive learning, the Euclidean distances between embeddings of the same EC number are minimized, while those between different EC numbers are maximized. This results in different spatial distributions for amino acid sequences with different EC numbers. CLEAN could classify and predict EC numbers of enzymes by calculating the Euclidean distances between the target sequence embeddings and the cluster centers, which avoids the intra-class imbalance issue that may arise when predicting EC numbers as a multi-class classification task.

DeepGOPlus infers GO annotation information based on sequence similarity and homology relationships, combining CNNs with sequence similarity-based predictions (Kulmanov and Hoehndorf, 2020). The CNNs have multiple convolutional kernels of variable sizes, learning patterns similar to structural domain motifs. This method significantly improves upon the traditional BLAST method and performs well in protein subcellular localization. Furthermore, Kulmanov et al. (2024) developed DeepGO-SE, a pre-trained large language model for predicting GO terms based on protein sequences. DeepGO-SE generates multiple approximate models for GO terms and predicts the truth values of protein function statements in these models using neural networks. The truth values from multiple models are then aggregated, improving the accuracy of protein function predictions.

ML has been extensively applied in the prediction of enzyme functions. However, the majority of the efforts still concentrate on natural enzymes, with limited exploration of function prediction for *de novo* designed enzymes. While current functional prediction methods can, to some extent, screen *de novo* designed enzymes, their accuracy and reliability are often unvalidated and require careful evaluation. Addressing this gap necessitates the development of function prediction frameworks tailored to *de novo* enzyme designs. Incorporating unsupervised learning techniques and generative models may offer potential solutions by uncovering latent functional patterns and exploring hypothetical enzyme designs. Additionally, creating benchmark datasets of *de novo* enzymes with experimentally validated functions is crucial for robust evaluation. These advancements would enhance the utility of ML in *de novo* enzyme design, paving the way for predicting novel enzymatic activities and expanding the scope of AI-driven enzyme design.

6. Conclusions and perspectives

De novo enzyme design has made notable advances. However, its applications for creating enzymes with specific catalytic functions have been less successful, with only a few examples, such as luciferase and triosephosphate isomerase. Challenges include low efficiency in design processes, as seen in the “Family Hallucination” strategy, which required extensive screening to yield a few active enzymes. These challenges highlight the inherent complexity of enzyme catalysis, which relies on precise substrate binding, transition state stabilization, and active site geometry to achieve high specificity and efficiency.

One of the major bottlenecks in enzyme design is the lack of standard and comprehensive catalytic data to develop models. High-throughput experimental platforms offer a promising solution by enabling rapid acquisition of large-scale, high-quality data under controlled conditions, which integrate automated reagent dispensing, precise environmental controls, and advanced detection systems, allowing for the measurement of key thermodynamic parameters such as kcat and Km with high accuracy and reproducibility (Faure et al., 2022; Hastings et al., 2023; Hekstra et al., 2016; Kim et al., 2022). Such data are critical for training robust ML models capable of predicting enzyme activity and guiding design processes.

The integration of structural characterization techniques, such as cryo-EM, XRC, NMR, XFELs, HDX, *etc.*, further enhances our understanding of the mechanism underlying enzymatic catalysis (Bhattacharya et al., 2022; Glasgow et al., 2023; Marco et al., 2025; Nakane et al., 2020; Pellegrini, 2020). These techniques, combined with

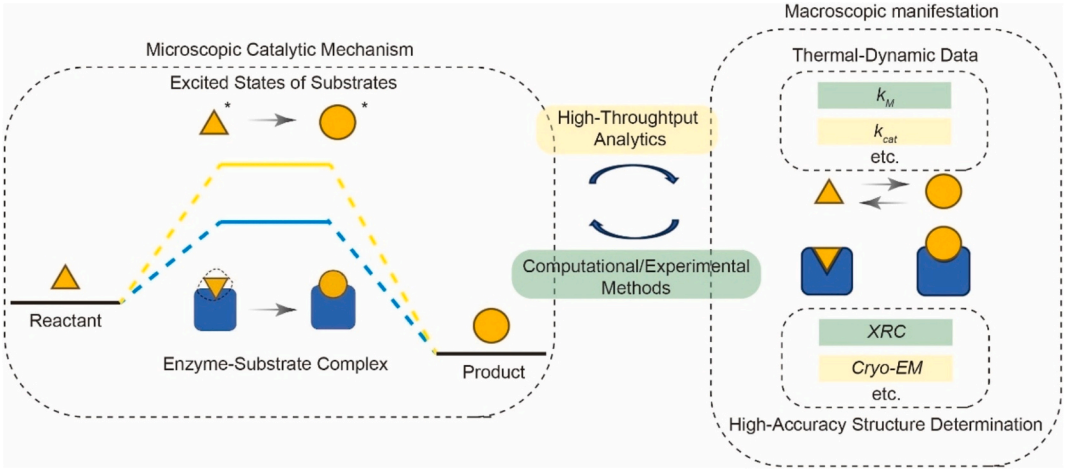

Fig. 5. Macroscopic manifestation and microscopic catalytic mechanism of enzyme reactions.

computational tools like molecular dynamics (MD) simulations and density functional theory (DFT), allow for high-resolution modeling of enzyme-substrate interactions, transition states, and active site dynamics. Such insights not only improve the accuracy of *de novo* enzyme design but also deepen our understanding of fundamental enzymatic principles, offering a roadmap for future innovations, which are highlighted in Fig. 5.

Practical applications of enzyme design also require careful consideration of the expression efficiency of genes, scalability of processes, and production costs, which are often overlooked in current models. Factors such as post-translational modifications, production time, and enzyme stability under different environments influence the feasibility of engineered enzymes in applications. Designing minimized enzymes that retain only essential functions can significantly improve expression levels of encoding genes and reduce production costs, making them more suitable for industrial production.

Another promising area for innovation lies in the design of enzymes for cascaded reaction pathways. Current models typically focus on single catalytic reactions, but industrial processes often involve interconnected steps with thermodynamic and kinetic dependencies. Designing multifunctional enzymes, such as fusion proteins, can optimize these pathways and improve overall process efficiency. Future models should integrate these considerations to better address the complexities of enzyme applications.

Recent advances in AI-driven enzyme design have reshaped resource allocation across the design cycle. Traditional methods like rational design and directed evolution are experimentally intensive but computationally light. In contrast, AI-based approaches, particularly those using deep generative or language models, accelerate early-stage design by prioritizing candidates *in silico*, reducing the experimental burden. This shift, however, introduces higher computational costs. Rather than lowering total costs outright, AI redistributes them, replacing labor-intensive screening with computational demands. As models become more efficient, AI is expected to further shorten design cycles and enable more targeted experimental validation.

The rapid advancements in computational power and the growing availability of high-precision data have significantly improved the potential of *de novo* enzyme design. By integrating these advancements with innovative experimental and computational approaches, researchers can design highly efficient enzymes with precise catalytic functions. As the field evolves, it not only expands the boundaries of green catalysis, industrial production, and therapeutic development but also provides deeper insights into the fundamental mechanisms of enzymatic action. These developments underscore the transformative potential of *de novo* enzyme design in addressing critical challenges across diverse fields.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Central Guidance for Local Science and Technology Development Fund (24ZYCGSY00030) and the National Natural Science Foundation of China (32372580).

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., Bodenstein, S.W., Evans, D.A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A.I., Cowie, A., Figurnov, M., Fuchs, F.B., Gladman, H., Jain, R., Khan, Y.A., Low, C.M.R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E.D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., Jumper, J.M., 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630 (8016), 193–500. https://doi.org/10.1038/s41586-024-07487-w.

Akpinaroglu, D., Seki, K., Guo, A., Zhu, E., Kelly, M.J.S., Kortemme, T., 2023. Structure-conditioned masked language models for protein sequence design generalize beyond the native sequence space. bioRxiv. https://doi.org/10.1101/2023.12.15.571823.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403–410. https://doi.org/10.1006/jmbi.1990.9999.

Anishchenko, I., Pellock, S.J., Chidyausiku, T.M., Ramelot, T.A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A.K., DiMaio, F., Carter, L., Chow, C.M., Montelione, G.T., Baker, D., 2021. *De novo* protein design by deep network hallucination. Nature 600 (7889), 547–552. https://doi.org/10.1038/s41586-021-04184-w.

Anishchenko, I., Kipnis, Y., Kalvet, I., Zhou, G., Krishna, R., Pellock, S.J., Lauko, A., Lee, G.R., An, L., Dauparas, J., DiMaio, F., Baker, D., 2024. Modeling protein-small molecule conformational ensembles with ChemNet. bioRxiv. https://doi.org/10.1101/2024.09.25.614868v1, 2024.09.25.614868v1.

Bansal, P., Morgat, A., Axelsen, K.B., Muthukrishnan, V., Coudert, E., Aimo, L., Hyka-Nouspikel, N., Gasteiger, E., Kerhornou, A., Neto, T.B., Pozzato, M., Blatter, M.-C., Ignatchenko, A., Redaschi, N., Bridge, A., 2022. Rhea, the reaction knowledgebase in 2022. Nucleic Acids Res. 50 (D1), D693–D700. https://doi.org/10.1093/nar/gkab1016.

Basler, S., Studer, S., Zou, Y., Mori, T., Ota, Y., Camus, A., Bunzel, H.A., Helgeson, R.C., Houk, K.N., Jiménez-Osés, G., Hilvert, D., 2021. Efficient Lewis acid catalysis of an abiological reaction in a *de novo* protein scaffold. Nat. Chem. Biol. 13 (3), 231–235. https://doi.org/10.1038/s41557-020-00628-4.

Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., Gonzales, L.J.D., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Joshi, V., Jyothi, D., Kandasaamy, S., Lock, A., Luciani, A., Lugaric, M., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Mishra, A., Moulang, K., Nightingale, A., Pundir, S., Qi, G.Y., Raj, S., Raposo, P., Rice, D.L., Saidi, R., Santos, R., Speretta, E., Stephenson, J., Totoo, P., Turner, E., Tyagi, N., Vasudev, P., Warner, K., Watkins, X., Zellner, H., Bridge, A.J., Aimo, L.,

Argoud-Puy, G.L., Auchincloss, A.H., Axelsen, K.B., Bansal, P., Baratin, D., Neto, T. M.B., Blatter, M.C., Bolleman, J.T., Boutet, E., Breuza, L., Gil, B.C., Casals-Casas, C., Echioukh, K.C., Coudert, E., Cuche, B., de Castro, E., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gaudet, P., Gehant, S., Gerritsen, V., Gos, A., Gruaz, N., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Kerhornou, A., Le Mercier, P., Lieberherr, D., Masson, P., Morgat, A., Muthukrishnan, V., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Poux, S., Pozzato, M., Pruess, M., Redaschi, N., Rivoire, C., Sigrist, C.J.A., Sonesson, K., Arighi, C.N., Armin-ski, L., Chen, C.M., Chen, Y.X., Huang, H.Z., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q.H., Wang, Y.Q., Zhang, J., Bye-A-Jee, H., Zaru, R., Sundaram, S., Wu, C.H., UniProt, C., 2023. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 51 (D1), D523–D531. https://doi.org/10.1093/nar/gkac1052.

Bennett, N.R., Coventry, B., Goreshnik, I., Huang, B., Allen, A., Vafeados, D., Peng, Y.P., Dauparas, J., Baek, M., Stewart, L., DiMaio, F., Munck, S.D., Savvides, S.N., Baker, D., 2023. Improving *de novo* protein binder design with deep learning. Nat. Commun. 14 (1), 2625. https://doi.org/10.1038/s41467-023-38328-5.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28 (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

Bhattacharya, S., Margheritis, E.G., Takahashi, K., Kulesha, A., D'Souza, A., Kim, I., Yoon, J.H., Tame, J.R.H., Volkov, A.N., Makhlynets, O.V., Korendovych, I.V., 2022. NMR-guided directed evolution. Nature 640, 389–393. https://doi.org/10.1038/s41586-022-05278-9.

Blagec, K., Dorffner, G., Moradi, M., Ott, S., Samwald, M., 2022. A global analysis of metrics used for measuring performance in natural language processing. In: Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP, pp. 52–63. https://doi.org/10.18653/v1/2022.nlppower-1.6.

Bolon, D.N., Mayo, S.L., 2001. Enzyme-like proteins by computational design. Proc. Natl. Acad. Sci. U. S. A. 98 (25), 14274–14279. https://doi.org/10.1073/pnas.251555398.

Brandenberg, O.F., Chen, K., Arnold, F.H., 2019. Directed evolution of a cytochrome P450 carbene transferase for selective functionalization of cyclic compounds. J. Am. Chem. Soc. 141 (22), 8989–8995. https://doi.org/10.1021/jacs.9b02931.

Brixi, G., Durrant, M.G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G.A., King, S.H., Li, D.B., Merchant, A.T., Naghipourfar, M., Nguyen, E., Ricci-Tam, C., Romero, D.W., Sun, G., Taghibakshi, A., Vorontsov, A., Yang, B., Deng, M., Gorton, L., Nguyen, N., Wang, N.K., Adams, E., Baccus, S.A., Dillmann, S., Ermon, S., Guo, D., Ilango, R., Janik, K., Lu, A.X., Mehta, R., Mofrad, M.R.K., Ng, M.Y., Pannu, J., Ré, C., Schmok, J.C., John, J.S., Sullivan, J., Zhu, K., Zynda, G., Balsam, D., Collison, P., Costa, A.B., Hernandez-Boussard, T., Ho, E., Liu, M.-Y., McGrath, T., Powell, K., Burke, D.P., Goodarzi, H., Hsu, P.D., Hie, B.L., 2025. Genome modeling and design across all domains of life with Evo 2. bioRxiv. https://doi.org/10.1101/2025.02.18.638918.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST plus : architecture and applications. BMC Bioinformatics 10, 421. https://doi.org/10.1186/1471-2105-10-421.

Cambria, E., White, B., 2014. Jumping NLP curves: A review of natural language processing research. IEEE Comput. Intell. Mag. 9 (2), 48–57.

Cao, H., Wang, J., He, L., Qi, Y., Zhang, J.Z., 2019. DeepDDG: predicting the stability change of protein point mutations using neural networks. J. Chem. Inf. Model. 59 (4), 1508–1514. https://doi.org/10.1021/acs.jcim.8b00697.

Cao, L., Coventry, B., Goreshnik, I., Huang, B., Sheffler, W., Park, J.S., Jude, K.M., Marković, I., Kadam, R.U., Verschueren, K.H.G., Verstraete, K., Walsh, S.T.R., Bennett, N., Ashish Phal, A.Y., Kozodoy, L., DeWitt, M., Picton, L., Miller, L., Strauch, E.-M., DeBouver, N.D., Pires, A., Bera, A.K., Halabiya, S., Hammerson, B., Yang, W., Bernard, S., Stewart, L., Wilson, I.A., Ruohola-Baker, H., Schlessinger, J., Lee, S., Savvides, S.N., Garcia, K.C., Baker, D., 2022. Design of protein-binding proteins from the target structure alone. Nature 605 (7910), 551–560. https://doi.org/10.1038/s41586-022-04654-9.

Caspi, R., Billington, R., Luciana Ferrer, H.F., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Karp, P.D., 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 44 (D1), D471–D480. https://doi.org/10.1093/nar/gkv1164.

Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D., Schomburg, D., 2021. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res. 49 (D1), D498–D508. https://doi.org/10.1093/nar/gkaa1025.

Chen, K., Arnold, F.H., 1993. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. Proc. Natl. Acad. Sci. U. S. A. 90 (12), 5618–5622. https://doi.org/10.1073/pnas.90.12.5618.

Chloe, H., Robert, V., Jason, L., Zeming, L., Brian, H., Tom, S., Adam, L., Alexander, R., 2022. Learning inverse folding from millions of predicted structures. bioRxiv. https://doi.org/10.1101/2022.04.10.487779, 2022.04.10.487779.

Chu, A.E., Kim, J., Cheng, L., Nesr, G.E., Xu, M., Shuai, R.W., Huang, P.-S., 2023. An all-atom protein generative model. Proc. Natl. Acad. Sci. U. S. A. 121 (27), e2311500121. https://doi.org/10.1073/pnas.2311500121.

Cobb, R.E., Chao, R., Zhao, H., 2013. Directed evolution: past, present, and future. AICHE J. 59 (5), 1432–1440. https://doi.org/10.1002/aic.13995.

Crawshaw, R., Crossley, A.E., Johannissen, L., Burke, A.J., Hay, S., Levy, C., Baker, D., Lovelock, S.L., Green, A.P., 2022. Engineering an efficient and enantioselective enzyme for the Morita-Baylis-Hillman reaction. Nat. Chem. Biol. 14 (3), 313–320. https://doi.org/10.1038/s41557-021-00833-9.

Cui, Y., Chen, Y., Sun, J., Zhu, T., Pang, H., Li, C., Geng, W.-C., Wu, B., 2024. Computational redesign of a hydrolase for nearly complete PET depolymerization at industrially relevant high-solids loading. Nat. Commun. 15, 1417. https://doi.org/10.1038/s41467-024-45662-9.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R.J., Milles, L.F., Wicky, B.I. M., Courbet, A., Haas, R.J.D., Bethel, N., Leung, P.J.Y., Huddy, T.F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A.K., King, N.P., Baker, D., 2022. Robust deep learning-based protein sequence design using ProteinMPNN. Science 378 (6615), 49–55. https://doi.org/10.1126/science.add2187.

Davey, J.A., Damry, A.M., Goto, N.K., Chica, R.A., 2017. Rational design of proteins that exchange on functional timescales. Nat. Chem. Biol. 13 (12), 1280–1285. https://doi.org/10.1038/nchembio.2503.

DeGrado, W.F., Regan, L., Ho, S.P., 1987. The design of a four-helix bundle protein. Cold Spring Harb. Symp. Quant. Biol. 52, 521–526. https://doi.org/10.1101/sqb.1987.052.01.059.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Dieckhaus, H., Brocidiacono, M., Randolph, N.Z., Kuhlman, B., 2023. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. Proc. Natl. Acad. Sci. U. S. A. 121 (6), e2314853121. https://doi.org/10.1073/pnas.2314853121.

Faure, A.J., Domingo, J., Schmiedel, J.M., Hidalgo-Carcedo, C., Diss, G., Lehner, B., 2022. Mapping the energetic and allosteric landscapes of protein binding domains. Nature 604, 175–183. https://doi.org/10.1038/s41586-022-04586-4.

Ferruz, N., Schmidt, S., Höcker, B., 2022. ProtGPT2 is a deep unsupervised language model for protein design. Nat. Commun. 13 (1), 4348. https://doi.org/10.1038/s41467-022-32007-7.

Floor, R.J., Wijma, H.J., Colpa, D.I., Ramos-Silva, A., Jekel, P.A., Szymański, W., Feringa, B.L., Marrink, S.J., Janssen, D.B., 2014. Computational library design for increasing haloalkane dehalogenase stability. ChemBioChem 15 (11), 1660–1672. https://doi.org/10.1002/cbic.201402128.

Fox, R.J., Davis, S.C., Mundorff, E.C., Newman, L.M., Gavrilovic, V., Ma, S.K., Chung, L. M., Ching, C., Tam, S., Muley, S., Grate, J., Gruber, J., Sheldon, J.C.W.R.A., Huisman, G.W., 2007. Improving catalytic function by ProSAR-driven enzyme evolution. Nat. Biotechnol. 25 (3), 338–344. https://doi.org/10.1038/nbt1286.

Frappier, V., Jenson, J.M., Zhou, J., Grigoryan, G., Keating, A.E., 2019. Tertiary structural motif sequence statistics enable facile prediction and design of peptides that bind anti-apoptotic Bfl-1 and Mcl-1. Structure 27 (4), 606–617. https://doi.org/10.1016/j.str.2019.01.008.

Fuchs, F.B., Worrall, D.E., Fischer, V., Welling, M., 2020. SE(3)-transformers: 3D roto-translation equivariant attention networks. Adv. Neural Inf. Proces. Syst. 33, 1970–1981.

Gainza, P., Wehrle, S., Hall-Beauvais, A.V., Marchand, A., Scheck, A., Harteveld, Z., Buckley, S., Ni, D., Tan, S., Sverrisson, F., Goverde, C., Turelli, P., Raclot, C., Teslenko, A., Pacesa, M., Rosset, S., Georgeon, S., Marsden, J., Petruzzella, A., Liu, K., Xu, Z., Chai, Y., Han, P., Gao, G.F., Oricchio, E., Fierz, B., Trono, D., Stahlberg, H., Bronstein, M., Correia, B.E., 2023. *De novo* design of protein interactions with learned surface fingerprints. Nature 617 (7959), 176–184. https://doi.org/10.1038/s41586-023-05993-x.

Georgiev, I., Lilien, R.H., Donald, B.R., 2008. The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. J. Comput. Chem. 29 (10), 1527–1698.

Glasgow, A.A., Huang, Y.-M., Mandell, D.J., Thompson, M., Ritterson, R., Loshbaugh, A. L., Pellegrino, J., Krivacic, C., Pache, R.A., Barlow, K.A., Ollikainen, N., Jeon, D., Kelly, M.J.S., Fraser, J.S., Kortemme, T., 2019. Computational design of a modular protein sense-response system. Science 366 (6468), 1024–1028. https://doi.org/10.1126/science.aax8780.

Glasgow, A., Hobbs, H.T., Perry, Z.R., Wells, M.L., Marqusee, S., Kortemme, T., 2023. Ligand-specific changes in conformational flexibility mediate long-range allostery in the lac repressor. Nat. Commun. 14, 1179. https://doi.org/10.1038/s41467-023-36798-1.

Gui, J., Sun, Z., Wen, Y., Tao, D., Ye, J., 2023. A review on generative adversarial networks: algorithms, theory, and applications. IEEE Transactions on Knowledge and Data Engineerin 35 (4), 3313–3332. https://doi.org/10.1109/TKDE.2021.3130191.

Hamamsy, T., Morton, J.T., Blackwell, R., Berenberg, D., Carriero, N., Gligorijevic, V., Strauss, C.E.M., Leman, J.K., Cho, K., Bonneau, R., 2024. Protein remote homology detection and structural alignment using deep learning. Nat. Biotechnol. 42 (6), 975–985. https://doi.org/10.1038/s41587-023-01917-2.

Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T., Kim, P.S., 1998. High-resolution protein design with backbone freedom. Science 282 (5393), 1462–1467. https://doi.org/10.1126/science.282.5393.1462.

Hastings, R., Aditham, A., DelRosso, N., Suzuki, P., Fordyce, P.M., 2023. High-throughput thermodynamic and kinetic measurements of transcription factor/DNA mutations reveal how conformational heterogeneity can shape motif selectivity. bioRxiv. https://doi.org/10.1101/2023.11.13.566946.

Hayes, T., Rao, R., Akin, H., Sofroniew, N.J., Oktay, D., Lin, Z., Verkuil, R., Tran, V.Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R.S., Thomas, N., Khan, Y.A., Mishra, C., Kim, C., Bartie, L.J., Nemeth, M., Hsu, P.D., Sercu, T., Candido, S., Rives, A., 2025. Simulating 500 million years of evolution with a language model. Science 387 (6736), 850–858. https://doi.org/10.1126/science.ads0018.

Hekstra, D.R., White, K.I., Socolich, M.A., Henning, R.W., Šrajer, V., Ranganathan, R., 2016. Electric-field-stimulated protein mechanics. Nature 560, 400–405. https://doi.org/10.1038/nature20571.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, 6840-6851. https://doi.org/10.5555/3495724.3496298.

Hodges, R.S., Saund, A.K., Chong, P.C., St-Pierre, S.A., Reid, R.E., 1981. Synthetic model for two-stranded alpha-helical coiled-coils. Design, synthesis, and characterization of an 86-residue analog of tropomyosin. J. Biol. Chem. 256 (3), 1214–1224. https://doi.org/10.1016/S0021-9258(19)69951-5.

Hua, C., Liu, Y., Zhang, D., Zhang, O., Luan, S., Yang, K.K., Wolf, G., Precup, D., Zheng, S., 2024. EnzymeFlow: generating reaction-specific enzyme catalytic pockets through flow matching and co-evolutionary dynamics. arXiv. https://doi.org/10.48550/arXiv.2410.00327, 2410.00327.

Huang, P.-S., Oberdorfer, G., Xu, C., Pei, X.Y., Nannenga, B.L., Rogers, J.M., DiMaio, F., Gonen, T., Luisi, B., Baker, D., 2014. High thermodynamic stability of parametrically designed helical bundles. Science 346 (6208), 481–485. https://doi.org/10.1126/science.1257481.

Huang, B., Xu, Y., Hu, X., Liu, Y., Liao, S., Zhang, J., Huang, C., Hong, J., Chen, Q., Liu, H., 2022. A backbone-centred energy function of neural networks for protein design. Nature 602, 523–528. https://doi.org/10.1038/s41586-021-04383-5.

Huang, C., Zhang, L., Tang, T., Wang, H., Jiang, Y., Ren, H., Zhang, Y., Fang, J., Zhang, W., Jia, X., You, S., Qin, B., 2024. Application of directed evolution and machine learning to enhance the diastereoselectivity of ketoreductase for dihydrotetrabenazine synthesis. JACS Au 4 (7), 2547–2556. https://doi.org/10.1021/jacsau.4c00284.

Ingraham, J.B., Baranov, M., Costello, Z., Barber, K.W., Wang, W., Ismail, A., Frappier, V., Lord, D.M., Ng-Thow-Hing, C., Vlack, E.R.V., Tie, S., Xue, V., Cowles, S. C., Leung, A., Rodrigues, J.V., Morales-Perez, C.L., Ayoub, A.M., Green, R., Puentes, K., Oplinger, F., Panwar, N.V., Obermeyer, F., Root, A.R., Beam, A.L., Poelwijk, F.J., Grigoryan, G., 2023. Illuminating protein space with a programmable generative model. Nature 623 (7989), 1070–1078. https://doi.org/10.1038/s41586-023-06728-8.

Jastrzębski, S., Leśniak, D., Czarnecki, W.M., 2016. Learning to SMILE(S). arXiv. https://doi.org/10.48550/arXiv.1602.06289, 1602.06289.

Jensen, S.B., Thodberg, S., Parween, S., Moses, M.E., Hansen, C.C., Thomsen, J., Sletfjerding, M.B., Knudsen, C., Giudice, R.D., Lund, P.M., Castaño, P.R., Bustamante, Y.G., Velazquez, M.N.R., Jørgensen, F.S., Pandey, A.V., Laursen, T., Møller, B.L., Hatzakis, N.S., 2021. Biased cytochrome P450-mediated metabolism via small-molecule ligands binding P450 oxidoreductase. Nat. Commun. 12 (1), 2260. https://doi.org/10.1038/s41467-021-22562-w.

Jiang, L., Althoff, E.A., Clemente, F.R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J.L., Betker, J.L., Tanaka, F., Barbas, C.F., Hilvert, D., Houk, K.N., Stoddard, B.L., Baker, D., 2008. *De novo* computational design of retro-aldol enzymes. Science 319 (5868), 1387–1391. https://doi.org/10.1126/science.1152692.

Juan, D.D., Pazos, F., Valencia, A., 2013. Emerging methods in protein co-evolution. Nat. Rev. Genet. 14, 249–261. https://doi.org/10.1038/nrg3414.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kalvet, I., Ortmayer, M., Zhao, J., Crawshaw, R., Ennist, N.M., Levy, C., Roy, A., Green, A.P., Baker, D., 2023. Design of heme enzymes with a tunable substrate binding pocket adjacent to an open metal coordination site. J. Am. Chem. Soc. 145 (26), 14307–14315. https://doi.org/10.1021/jacs.3c02742.

Khare, S.D., Kipnis, Y., Greisen, P.J., Takeuchi, R., Ashani, Y., Goldsmith, M., Song, Y., Gallaher, J.L., Silman, I., Leader, H., Sussman, J.L., Stoddard, B.L., Tawfik, D.S., Baker, D., 2012. Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. Nat. Chem. Biol. 8, 294–300. https://doi.org/10.1038/nchembio.777.

Kim, G., Azmi, L., Jang, S., Jung, T., Hebert, H., Roe, A.J., Byron, O., Song, J.-J., 2019. Aldehyde-alcohol dehydrogenase forms a high-order spirosome architecture critical for its activity. Nat. Commun. 10, 4527. https://doi.org/10.1038/s41467-019-12427-8.

Kim, T.-E., Tsuboyama, K., Houliston, S., Martell, C.M., Phoumyvong, C.M., Lemak, A., Haddox, H.K., Arrowsmith, Cheryl H., Rocklin, G.J., 2022. Dissecting the stability determinants of a challenging *de novo* protein fold using massively parallel design and experimentation. Proc. Natl. Acad. Sci. U. S. A. 119 (41), e2122676119. https://doi.org/10.1073/pnas.2122676119.

Kim, D., Woodbury, S.M., Ahern, W., Kalvet, I., Hanikel, N., Salike, S., Pellock, S.J., Lauko, A., Hilvert, D., Baker, D., 2024. Computational design of metallohydrolases. bioRxiv. https://doi.org/10.1101/2024.11.13.623507.

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G.R., Morey-Burrows, F.S., Anishchenko, Ivan, Humphreys, I.R., McHugh, R., Vafeados, D., Li, X., Sutherland, G.A., Hitchcock, A., Hunter, C.N., Kang, A., Brackenbrough, E., Bera, A.K., Baek, M., DiMaio, F., Baker, D., 2024. Generalized biomolecular modeling and design with RoseTTAFold all-atom. Science 384 (6693), eadl2528. https://doi.org/10.1126/science.adl2528.

Kroll, A., Ranjan, S., Engqvist, M.K.M., Lercher, M.J., 2023. A general model to predict small molecule substrates of enzymes based on machine and deep learning. Nat. Commun. 14 (1), 2787. https://doi.org/10.1038/s41467-023-38347-2.

Kroll, A., Ranjan, S., Lercher, M.J., 2024. A multimodal transformer network for protein-small molecule interactions enhances predictions of kinase inhibition and enzyme-substrate relationships. PLoS Comput. Biol. 20 (5), e1012100. https://doi.org/10.1371/journal.pcbi.1012100.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., Baker, D., 2003. Design of a novel globular protein fold with atomic-level accuracy. Science 302 (5649), 1364–1368. https://doi.org/10.1126/science.1089427.

Kulmanov, M., Hoehndorf, R., 2020. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 36 (2), 422–429. https://doi.org/10.1093/bioinformatics/btz595.

Kulmanov, M., Guzmán-Vega, F.J., Roggli, P.D., Lane, L., Arold, S.T., Hoehndorf, R., 2024. Protein function prediction as approximate semantic entailment. Nat. Mach. Intell. 6 (2), 220–228. https://doi.org/10.1038/s42256-024-00795-w.

Lauko, A., Pellock, S.J., Sumida, K.H., Anishchenko, I., Juergens, D., Ahern, W., Jeung, J., Shida, A., Hunt, A., Kalvet, I., Norn, C., Humphreys, I.R., Jamieson, C., Krishna, R., Kipnis, Y., Kang, A., Brackenbrough, E., Bera, A.K., Sankaran, B., Houk, K.N., Baker, D., 2025. Computational design of serine hydrolases. Science 0, eadu2454. https://doi.org/10.1126/science.adu2454.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K. W., Renfrew, P.D., Smith, C.A., Sheffler, W., Davis, I.W., Cooper, S., Treuille, A., Mandell, D.J., Richter, F., Ban, Y.-E.A., Fleishman, S.J., Corn, J.E., Kim, D.E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J.J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J.J., Kuhlman, B., Baker, D., Bradley, P., 2011. Chapter nineteen - Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 487, 545–574. https://doi.org/10.1016/B978-0-12-381270-4.00019-6.

Lerner, S.A., Wu, T.T., Lin, E.C., 1964. Evolution of a catabolic pathway in bacteria. Science 146 (3649), 1313–1315. https://doi.org/10.1126/science.146.3649.1313.

Li, Y., Li, J., Chen, W.-K., Li, Y., Xu, S., Li, L., Xia, B., Wang, R., 2024. Tuning architectural organization of eukaryotic P450 system to boost bioproduction in Escherichia coli. Nat. Commun. 15 (1), 10009. https://doi.org/10.1038/s41467-024-54259-1.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Costa, A.D., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A., 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379 (6637), 1123–1130. https://doi.org/10.1126/science.ade2574.

Lipman, D.J., Pearson, W.R., 1985. Rapid and sensitive protein similarity searches. Science 227 (4693), 1435–1441. https://doi.org/10.1126/science.2983426.

Lisanza, S.L., Gershon, J.M., Tipps, S.W.K., Sims, J.N., Arnoldt, L., Hendel, S.J., Simma, M.K., Liu, G., Yase, M., Wu, H., Tharp, C.D., Li, X., Kang, A., Brackenbrough, E., Bera, A.K., Gerben, S., Wittmann, B.J., McShan, A.C., Baker, D., 2024. Multistate and functional protein design using RoseTTAFold sequence space diffusion. Nat. Biotechnol. https://doi.org/10.1038/s41587-024-02395-w.

Liu, Q., Segal, D.J., Ghiara, J.B., Barbas, C.F., 1997. Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. Proc. Natl. Acad. Sci. U. S. A. 94 (11), 5525–5530. https://doi.org/10.1073/pnas.94.11.5525.

Liu, B., Qu, G., Li, J.-K., Fan, W., Ma, J.-A., Xu, Y., Nie, Y., Sun, Z., 2019. Conformational dynamics-guided loop engineering of an alcohol dehydrogenase: capture, turnover and enantioselective transformation of difficult-to-reduce ketones. Adv. Synth. Catal. 361 (13), 3182–3190. https://doi.org/10.1002/adsc.201900249.

Llinares-López, F., Berthet, Q., Blondel, M., Teboul, O., Vert, J.-P., 2023. Deep embedding and alignment of protein sequences. Nat. Methods 20 (1), 104–111. https://doi.org/10.1038/s41592-022-01700-2.

Lovelock, S.L., Crawshaw, R., Basler, S., Levy, C., Baker, D., Hilvert, D., Green, A.P., 2022. The road to fully programmable protein catalysis. Nature 606 (7912), 49–58. https://doi.org/10.1038/s41586-022-04456-z.

Lucas, J.E., Kortemme, T., 2020. New computational protein design methods for *de novo* small molecule binding sites. PLoS Comput. Biol. 16 (10), e1008178. https://doi.org/10.1371/journal.pcbi.1008178.

Luo, X.-J., Zhao, J., Li, C.-X., Bai, Y.-P., Reetz, M.T., Yu, H.-L., Xu, J.-H., 2016. Combinatorial evolution of phosphotriesterase toward a robust malathion degrader by hierarchical iteration mutagenesis. Biotechnol. Bioeng. 113 (11), 2350–2357. https://doi.org/10.1002/bit.26012.

Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton Jr., J.M., Xiong, C., Sun, Z.Z., Socher, R., Fraser, J.S., Naik, N., 2023. Large language models generate functional protein sequences across diverse families. Nat. Biotechnol. 41 (8), 1099–1106. https://doi.org/10.1038/s41587-022-01618-2.

Marco, M.D., Rai, S.R., Scietti, L., Mattoteia, D., Liberi, S., Moroni, E., Pinnola, A., Vetrano, A., Iacobucci, C., Santambrogio, C., Colombo, G., Forneris, F., 2025. Molecular structure and enzymatic mechanism of the human collagen hydroxylysine galactosyltransferase GLT25D1/COLGALT1. Nat. Commun. 16, 3624. https://doi.org/10.1038/s41467-025-59017-5.

Mayo, S.L., Olafson, B.D., Goddard, W.A., 1990. DREIDING: a generic force field for molecular simulations. J. Phy. Chem. B 94 (26), 8897–8909. https://doi.org/10.1021/j100389a010.

Maziarka, Ł., Pocha, A., Kaczmarczyk, J., Rataj, K., Danel, T., Warchoł, M., 2020. Mol-CycleGAN: a generative model for molecular optimization. J. Cheminformatics 12 (2). https://doi.org/10.1186/s13321-019-0404-1.

Mazurenko, S., Prokop, Z., Damborsky, J., 2020. Machine learning in enzyme engineering. ACS Catal. 10 (2), 1210–1223. https://doi.org/10.1021/acscatal.9b04321.

McNutt, A.T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., Sunseri, J., Koes, D.R., 2021. GNINA 1.0: molecular docking with deep learning. J. Cheminformatics 13 (1), 43. https://doi.org/10.1186/s13321-021-00522-2.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., Rives, A., 2024. Language models enable zero-shot prediction of the effects of mutations on protein function. In: NIPS'21:

Proceedings of the 35th International Conference on Neural Information Processing Systems, pp. 29287–29303. https://doi.org/10.5555/3540261.3542504.
Munsamy, G., Illanes-Vicioso, R., Funcillo, S., Nakou, I.T., Lindner, S., Ayres, G., Sheehan, L.S., Moss, S., Eckhard, U., Lorenz, P., Ferruz, N., 2024. Conditional language models enable the efficient design of proficient enzymes. bioRxiv. https://doi.org/10.1101/2024.05.03.592223v1.
Nakane, T., Kotecha, A., Sente, A., McMullan, G., Masiulis, S., Brown, P.M.G.E., Grigoras, I.T., Malinauskaite, L., Malinauskas, T., Miehling, J., Uchański, T., Yu, L., Karia, D., Pechnikova, E.V., Jong, E.D., Keizer, J., Bischoff, M., McCormack, J., Tiemeijer, P., Hardwick, S.W., Chirgadze, D.Y., Murshudov, G., Aricescu, A.R., Scheres, S.H.W., 2020. Single-particle cryo-EM at atomic resolution. Nature 587, 152–156. https://doi.org/10.1038/s41586-020-2829-0.
Newton, M.S., Arcus, V.L., Gerth, M.L., Patrick, W.M., 2018. Enzyme evolution: innovation is easy, optimization is complicated. Curr. Opin. Struct. Biol. 48, 110–116. https://doi.org/10.1016/j.sbi.2017.11.007.
Nguyen, E., Poli, M., Durrant, M.G., Kang, B., Katrekar, D., Li, D.B., Bartie, L.J., Thomas, A.W., King, S.H., Brixi, G., Sullivan, J., Ng, M.Y., Lewis, A., Lou, A., Ermon, S., Baccus, S.A., Hernandez-Boussard, T., Ré, C., Hsu, P.D., Hie, B.L., 2024. Sequence modeling and design from molecular to genome scale with Evo. Science 386 (6723), eado9336. https://doi.org/10.1126/science.ado9336.
Ogata, H., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28 (1), 27–30. https://doi.org/10.1093/nar/28.1.27.
Orsi, E., Borzyskowski, L.S.V., Noack, S., Nikel, P.I., Lindner, S.N., 2024. Automated in vivo enzyme engineering accelerates biocatalyst optimization. Nat. Commun. 15 (1), 3447. https://doi.org/10.1038/s41467-024-46574-4.
Outeiral, C., Deane, C.M., 2022. Codon language embeddings provide strong signals for protein engineering. bioRxiv. https://doi.org/10.1101/2022.12.15.519894.
Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311-318.
Peleg, Y., Vincentelli, R., Collins, B.M., Chen, K.-E., Livingstone, E.K., Weeratunga, S., Leneva, N., Guo, Q., Remans, K., Perez, K., Bjerga, G.E.K., Larsen, Ø., Vaněk, O., Skořepa, O., Jacquemin, S., Poterszman, A., Kjær, S., Christodoulou, E., Albeck, S., Dym, O., Ainbinder, E., Unger, T., Schuetz, A., Matthes, S., Bader, M., Marco, A.D., Storici, P., Semrau, M.S., Stolt-Bergner, P., Aigner, C., Suppmann, S., Goldenzweig, A., Fleishman, S.J., 2021. Community-wide experimental evaluation of the PROSS stability-design method. J. Mol. Biol. 433 (13), 166964. https://doi.org/10.1016/j.jmb.2021.166964.
Pellegrini, C., 2020. The development of XFELs. Nat. Rev. Phys. 2, 330–331. https://doi.org/10.1038/s42254-020-0197-1.
Polizzi, N.F., DeGrado, W.F., 2020. A defined structural unit enables *de novo* design of small-molecule–binding proteins. Science 369 (6508), 1227–1233. https://doi.org/10.1126/science.abb8330.
Porebski, B.T., Buckle, A.M., 2016. Consensus protein design. Protein Eng. Des. Sel. 29 (7), 245–251. https://doi.org/10.1093/protein/gzw015.
Punta, M., Ofran, Y., 2008. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comput. Biol. 4 (10), e1000160. https://doi.org/10.1371/journal.pcbi.1000160.
Rajagopalan, S., Wang, C., Yu, K., Kuzin, A.P., Richter, F., Lew, S., Miklos, A.E., Matthews, M.L., Seetharaman, J., Su, M., Hunt, J.F., Cravatt, B.F., Baker, D., 2014. Design of activated serine–containing catalytic triads with atomic-level accuracy. Nat. Chem. Biol. 10, 386–391. https://doi.org/10.1038/nchembio.1498.
Ren, M., Yu, C., Bu, D., Zhang, H., 2024. Accurate and robust protein sequence design with CarbonDesign. Nat. Mach. Intell. 6 (5), 536–547. https://doi.org/10.1038/s42256-024-00838-2.
Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Poviloniene, S., Laurynenas, A., Viknander, S., Abuajwa, W., Savolainen, O., Meskys, R., Engqvist, M.K.M., Zelezniak, A., 2021. Expanding functional protein sequence spaces using generative adversarial networks. Nat. Mach. Intell. 3, 324–333. https://doi.org/10.1038/s42256-021-00310-5.
Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., Fergus, R., 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. U. S. A. 118 (15), e2016239118. https://doi.org/10.1073/pnas.2016239118.
Röthlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O., Albeck, S., Houk, K.N., Tawfik, D.S., Baker, D., 2008. Kemp elimination catalysts by computational enzyme design. Nature 453, 190–195. https://doi.org/10.1038/nature06879.
Ryu, J.Y., Kim, H.U., Lee, S.Y., 2019. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc. Natl. Acad. Sci. U. S. A. 116 (28), 13996–14001. https://doi.org/10.1073/pnas.1821905116.
Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., Serrano, L., 2005. The FoldX web server: an online force field. Nucleic Acids Res. 33 (suppl_2), W382–W388. https://doi.org/10.1093/nar/gki387.
Sevgen, E., Moller, J., Lange, A., Parker, J., Quigley, S., Mayer, J., Srivastava, P., Gayatri, S., Hosfield, D., Korshunova, M., Livne, M., Gill, M., Ranganathan, R., Costa, A.B., Ferguson, A.L., 2023. ProT-VAE: protein transformer variational autoencoder for functional protein design. bioRxiv. https://doi.org/10.1101/2023.01.23.525232.
Shaik, S., Cohen, S., Wang, Y., Chen, H., Kumar, D., Thiel, W., 2009. P450 enzymes: their structure, reactivity, and selectivity—modeled by QM/MM calculations. Chem. Rev. 110 (2), 949–1017. https://doi.org/10.1021/cr900121s.
Siegel, J.B., Zanghellini, A., Lovick, H.M., Kiss, G., Lambert, A.R., St. Clair, J.L., Gallaher, J.L., Hilvert, D., Gelb, M.H., Stoddard, B.L., Houk, K.N., Michael, F.E., Baker, D., 2010. Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science 329 (5989), 309–313. https://doi.org/10.1126/science.1190239.
Siegel, J.B., Smith, A.L., Poust, S., Wargacki, A.J., Bar-Even, A., Louw, C., Shen, B.W., Eiben, C.B., Tran, H.M., Noor, E., Gallaher, J.L., Bale, J., Yoshikuni, Y., Gelb, M.H., Keasling, J.D., Stoddard, B.L., Lidstrom, M.E., Baker, D., 2015. Computational protein design enables a novel one-carbon assimilation pathway. Proc. Natl. Acad. Sci. U. S. A. 112 (12), 3704–3709. https://doi.org/10.1073/pnas.1500545112.
Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Cornu, S.L., Lam, S.D., Berka, K., Varekova, I.H., Svobodova, R., Lees, J., Orengo, C.A., 2021. CATH: increased structural coverage of functional space. Nucleic Acids Res. 49 (D1), D266–D273. https://doi.org/10.1093/nar/gkaa1079.
Singh, N., Lane, S., Yu, T., Lu, J., Ramos, A., Cui, Haiyang, Zhao, H., 2025. A generalized platform for artificial intelligence-powered autonomous protein engineering. bioRxiv. https://doi.org/10.1101/2025.02.12.637932.
Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147 (1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5.
Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35, 1026–1028. https://doi.org/10.1038/nbt.3988.
Sumida, K.H., Núñez-Franco, R., Kalvet, I., Pellock, S.J., Wicky, B.I.M., Milles, L.F., Dauparas, J., Wang, J., Kipnis, Y., Jameson, N., Kang, A., Cruz, J.D.L., Sankaran, B., Bera, A.K., Jiménez-Osés, G., Baker, D., 2024. Improving protein expression, stability, and function with ProteinMPNN. J. Am. Chem. Soc. 146 (3), 2054–2061. https://doi.org/10.1021/jacs.3c10941.
Thomson, A.R., Wood, C.W., Burton, A.J., Bartlett, G.J., Sessions, R.B., Brady, R.L., Woolfson, D.N., 2014. Computational design of water-soluble α-helical barrels. Science 346 (6208), 485–488. https://doi.org/10.1126/science.1257452.
Tian, R., Rehm, F.B.H., Czernecki, D., Gu, Y., Zürcher, J.F., Liu, K.C., Chin, J.W., 2024. Establishing a synthetic orthogonal replication system enables accelerated evolution in E. Coli. Science 383 (6681), 421–426. https://doi.org/10.1126/science.adk1281.
Tinberg, C.E., Khare, S.D., Dou, J., Doyle, L., Nelson, J.W., Schena, A., Jankowski, W., Kalodimos, C.G., Johnsson, K., Stoddard, B.L., Baker, D., 2013. Computational design of ligand-binding proteins with high affinity and selectivity. Nature 501, 212–216. https://doi.org/10.1038/nature12443.
Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., Jaakkola, T., 2023, Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. arXiv. Doi: 10.48550/arXiv.2206.04119.
Trott, O., Olson, A.J., 2010. Software news and update AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. 31 (2), 455–461. https://doi.org/10.1002/jcc.21334.
Tsubaki, M., Tomii, K., Sese, J., 2019. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics 35 (2), 309–318. https://doi.org/10.1093/bioinformatics/bty535.
Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G.J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S.A.A., Potapenko, A., Ballard, A.J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A.W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J., Hassabis, D., 2021. Highly accurate protein structure prediction for the human proteome. Nature 596 (7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1.
Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., Birney, E., Steinegger, M., Hassabis, D., Velankar, S., 2023. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. 52 (D1), D368–D375.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000–6010. https://doi.org/10.5555/3295222.3295349.
Wahab, R.A., Basri, M., Rahman, R.N.Z.R.A., Salleh, A.B., Rahman, M.B.A., Chor, L.T., 2012. Manipulation of the conformation and enzymatic properties of T1 lipase by site-directed mutagenesis of the protein core. Appl. Biochem. Biotechnol. 167, 612–620. https://doi.org/10.1007/s12010-012-9728-2.
Wang, Y., Tang, H., Huang, L., Pan, L., Yang, L., Yang, H., Mu, F., Yang, M., 2023. Self-play reinforcement learning guides protein engineering. Nat. Mach. Intell. 5 (8), 845–860. https://doi.org/10.1038/s42256-023-00691-9.
Watson, J.L., Juergens, D., Bennett, N.R., Trippe, B.L., Yim, J., Eisenach, H.E., Ahern, W., Borst, A.J., Ragotte, R.J., Milles, L.F., Wicky, B.I.M., Hanikel, N., Pellock, S.J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S.V., Lauko, A., Bortoli, V.D., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T.S., DiMaio, F., Baek, M., Baker, D., 2023. *De novo* design of protein structure and function with RFdiffusion. Nature 620 (7976), 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.
Wu, Z., Kan, S.B.J., Lewis, R.D., Wittmann, B.J., Arnold, F.H., 2019. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc. Natl. Acad. Sci. U. S. A. 116 (18), 8852–8858. https://doi.org/10.1073/pnas.1901979116.
Wu, K.E., Yang, K.K., Berg, R., v. d., Alamdari, S., Zou, J. Y., Lu, A. X., Amini, A. P., 2024. Protein structure generation via folding diffusion. Nat. Commun. 15 (1), 1059. https://doi.org/10.1038/s41467-024-45051-2.
Xiong, P., Wang, M., Zhou, X., Zhang, T., Zhang, J., Chen, Q., Liu, H., 2014. Protein design with a comprehensive statistical energy function and boosted by experimental

selection for foldability. Nat. Commun. 5, 5330. https://doi.org/10.1038/ncomms6330.
Xu, Y., Liu, D., Gong, H., 2024a. Improving the prediction of protein stability changes upon mutations by geometric learning and a pre-training strategy. Nat. Comput. Sci. 4 (11), 840–850. https://doi.org/10.1038/s43588-024-00716-2.
Xu, H., Yuan, Z., Yang, S., Su, Z., Hou, X.-D., Deng, Z., Zhang, Y., Rao, Y., 2024b. Discovery of a fungal P450 with an unusual two-step mechanism for constructing a Bicyclo[3.2.2] nonane skeleton. J. Am. Chem. Soc. 146 (12), 8716–8726. https://doi.org/10.1021/jacs.4c01284.
Yeh, A.H.-W., Norn, C., Kipnis, Y., Tischer, D., Pellock, S.J., Evans, D., Ma, P., LeeJ, G.R., Zhang, A.Z., Anishchenko, I., Coventry, B., Cao, L., Dauparas, J., Halabiya, S., DeWitt, M., Carter, L., Houk, K.N., Baker, D., 2023. *De novo* design of luciferases using deep learning. Nature 614 (7949), 774–780. https://doi.org/10.1038/s41586-023-05696-3.
Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G., Zhao, H., 2023. Enzyme function prediction using contrastive learning. Science 379 (6639), 1358–1363. https://doi.org/10.1126/science.adf2465.
Yuan, M., Shen, A., Fu, K., Guan, J., Ma, Y., Qiao, Q., Wang, M., 2023. ProteinMAE: masked autoencoder for protein surface self-supervised learning. Bioinformatics 39 (12), btad724. https://doi.org/10.1093/bioinformatics/btad724.
Zhang, J., Grigoryan, G., 2013. Chapter two - mining tertiary structural motifs for assessment of designability. Methods Enzymol. 523, 21–40. https://doi.org/10.1016/B978-0-12-394292-0.00002-3.
Zhang, X., Zhang, O., Shen, C., Qu, W., Chen, S., Cao, H., Kang, Y., Wang, Z., Wang, E., Zhang, J., Deng, Y., Liu, F., Wang, T., Du, H., Wang, L., Pan, P., Chen, G., Hsieh, C.-Y., Hou, T., 2023. Efficient and accurate large library ligand docking with KarmaDock. Nat. Comput. Sci. 3 (9), 789–804. https://doi.org/10.1038/s43588-023-00511-5.
Zheng, S., Li, Y., Chen, S., Xu, J., Yang, Y., 2020. Predicting drug-protein interaction using quasi-visual question answering system. Nat. Mach. Intell. 2 (2), 134–140. https://doi.org/10.1038/s42256-020-0152-y.
Zhou, Y., Pan, Q., Pires, D.E.V., Rodrigues, C.H.M., Ascher, D.B., 2023. DDMut: predicting effects of mutations on protein stability using deep learning. Nucleic Acids Res. 51 (W1), W122–W128. https://doi.org/10.1093/nar/gkad472.