Biotechnology Advances 73 (2024) 108376

Contents lists available at ScienceDirect

Biotechnology Advances

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

Research review paper

Unlocking the potential of enzyme engineering *via* rational computational design strategies

Lei Zhou [1], Chunmeng Tao [1], Xiaolin Shen, Xinxiao Sun, Jia Wang [*], Qipeng Yuan [*]

State Key Laboratory of Chemical Resource Engineering, Beijing University of Chemical Technology, Beijing 100029, China

ARTICLE INFO

Keywords:
Enzyme engineering
Rational computational design
Structure-based computational design
Sequence-based computational design
Data-driven machine learning

ABSTRACT

Enzymes play a pivotal role in various industries by enabling efficient, eco-friendly, and sustainable chemical processes. However, the low turnover rates and poor substrate selectivity of enzymes limit their large-scale applications. Rational computational enzyme design, facilitated by computational algorithms, offers a more targeted and less labor-intensive approach. There has been notable advancement in employing rational computational protein engineering strategies to overcome these issues, it has not been comprehensively reviewed so far. This article reviews recent developments in rational computational enzyme design, categorizing them into three types: structure-based, sequence-based, and data-driven machine learning computational design. Case studies are presented to demonstrate successful enhancements in catalytic activity, stability, and substrate selectivity. Lastly, the article provides a thorough analysis of these approaches, highlights existing challenges and potential solutions, and offers insights into future development directions.

1. Introduction

Remarkably efficient metabolic enzymes play a pivotal role in converting readily available, simple starting materials into valuable products within microbial cells. Furthermore, these enzymes facilitate a diverse array of chemical reactions, typically operating under mild reaction conditions and exhibiting high selectivity (Hollmann and Fernandez-Lafuente, 2021). These advantages have led to the incorporation of enzymes in various industries, including food (Flynn et al., 2021; Lin et al., 2022; Punia, 2020; Rentschler et al., 2015; van Donkelaar et al., 2016; Xu et al., 2020), agricultural (Costa-Silva et al., 2021; Sijinamanoj et al., 2021; Tingley et al., 2021), cosmetics (Almeida et al., 2021; Fournière et al., 2021) and pharmaceuticals (Meghwanshi et al., 2020; Park et al., 2017; Rosenthal and Lütz, 2018). Regrettably, enzymes come with certain drawbacks such as low turnover rates, limited stability and a narrow substrate scope. Consequently, there is a necessity for enzyme engineering to tailor them for diverse industrial applications (Li et al., 2018b).

Utilizing both directed evolution and semi-rational approaches, enzyme engineering involves iterative steps and mutation libraries, with the process repeated until the desired variant is obtained (Dinmukhamed et al., 2021; Gargiulo and Soumillion, 2021). However, the unknown structure-function relationship introduces uncertainty in discovering desired protein variants through a few rounds of iterative mutagenesis. It's more likely that the researchers may never come across it even with a comprehensive library, leading them to rely on insights from rational design and evolutionary analysis (Blazeck et al., 2022). In contrast, rational enzyme design stands out for its ability to provide higher predictive accuracy and streamline the screening library (Jumper et al., 2021; Kuhlman and Bradley, 2019). Proposed mutations are introduced after evaluation and design, increasing the likelihood of beneficial mutations while saving time and labor. When high-throughput screening is not feasible, this approach stands out as particularly useful (Cui et al., 2022; Steiner and Schwab, 2012).

Based on the design principles, rational computational enzyme design is divided into three classes, structure-based computational design, sequence-based computational design, and data-driven machine learning computational design. Here, we assess the recent advancements in rational computational enzyme design, emphasizing the effectiveness and versatility of different approaches in enhancing catalytic activity, stability and substrate selectivity (Table 1). Successful cases are cited to illustrate these improvements. Lastly, the article provides a thorough analysis of these approaches, highlights existing challenges and potential solutions, and offers insights into future development directions.

* Corresponding authors.
E-mail addresses: wangjia@buct.edu.cn (J. Wang), yuanqp@buct.edu.cn (Q. Yuan).
[1] Lei Zhou and Chunmeng Tao contributed equally.

https://doi.org/10.1016/j.biotechadv.2024.108376
Received 27 December 2023; Received in revised form 27 April 2024; Accepted 8 May 2024
Available online 11 May 2024

2. Structure-based computational enzyme engineering and design

Structure-based design necessitates the identification of critical residues within the substrate binding pocket, understanding the enzyme-catalyzed reaction's chemical mechanism and the key amino acid residues involved. This is achieved through chemical intuition and computational methods to precisely tailor the interplay between key amino acid residues and the substrate in the active pocket, potentially enabling *de novo* computational enzyme design. Although traditional methods including cryo-electron microscopy or X-ray diffraction have significantly advanced in obtaining enzyme crystal structures, with 170,000 structures available out of 190 million proteins with known amino acid sequences, the process remains time-consuming and costly (Consortium, 2020) (Burley et al., 2020). To overcome these limitations, AlphaFold2 and RoseTTAfold employing machine learning algorithms to accurately predict protein structures using extensive protein information data (Baek et al., 2021; Jumper et al., 2021). AlphaFold2, in particular, has demonstrated high accuracy comparable to experimental results, exhibiting a root-mean-squared deviation (RMSD) of about 1.5 Å

Table 1
Examples of strategies for enzyme engineering *via* rational computational design.

Type	Enzyme/Recation	Effect/Strategy	Targeted Property	Reference
Structure-based	Phosphoserine aminotransferase	Calculation of relative binding free energy	Substrate specificity	(Zhang et al., 2019)
Structure-based	Ferulic acid decarboxylase	Calculation of affinity between enzymes and substrates	Activity	(Mori et al., 2021)
Structure-based	Fatty acid photodecarboxylase	Modulating the steric hindrance of the binding pocket	Activity and selectivity	(Xu et al., 2022)
Structure-based	Fatty acid photodecarboxylase	Enhancing the electronic interaction between enzymes and substrates	Selectivity	(Li et al., 2021a)
Structure-based	4-hydroxyphenylpyruvate dioxygenase	Manipulating the hydrogen bond network	Product profile	(Lin et al., 2021)
Structure-based	Santalene synthase	Reducing the steric hindrance to increase pocket space	Product profile	(Zha et al., 2022)
Structure-based	The short-chain dehydrogenase/reductase	Tuning the size of the active pockets	Selectivity	(Su et al., 2020)
Structure-based	Amidase	Modifying electrostatic interactions	Activity and stability	(Galmés et al., 2022)
Structure-based	Formolase	Redesign of the binding pocket	Activity and specificity	(Siegel et al., 2015)
Structure-based	Limonene epoxide hydrolase	Redesign of the binding pocket	Selectivity	(Wijma et al., 2015)
Structure-based	C-N lyases	Redesign of the binding pocket	Activity and selectivity	(Cui et al., 2021)
Structure-based	Lipase	Design disulfide bonds	Stability	(Li et al., 2018a)
Structure-based	Glucuronidase	Truncating C-terminal region	Stability	(Han et al., 2018)
Structure-based	ω-transaminase	Calculating folding energy to identify stable point mutations	Stability	(Meng et al., 2020)
Structure-based	Diels-Alder	*De novo* design of enzyme by "inside-out"	New enzyme	(Siegel et al., 2010)
Structure-based	Kemp elimination	*De novo* design of enzyme by "inside-out"	New enzyme	(Röthlisberger et al., 2008)
Structure-based	Retro-Aldol	*De novo* design of enzyme by "inside-out"	New enzyme	(Jiang et al., 2008)
Structure-based	Ester hydrolysis	*De novo* design of enzyme by "inside-out"	New enzyme	(Richter et al., 2012)
Structure-based	Morita-Baylis-Hillman	*De novo* design of enzyme by "inside-out"	New enzyme	(Bjelic et al., 2013)
Sequence-based	Glycosyltransferases	Consensus design	Substrate specificity	(Teze et al., 2021)
Sequence-based	Triosephosphate isomerase	Consensus design	Stability	(Sullivan et al., 2012)
Sequence-based	Glucose-1-dehydrogenase	Consensus design	Stability	(Vázquez-Figueroa et al., 2007)
Sequence-based	Haloalkane dehalogenase	Ancestral sequence reconstruction	Stability	(Babkova et al., 2020)
Sequence-based	Cyclohexadienyl dehydratase	Ancestral sequence reconstruction	Activity	(Kaczmarski et al., 2020)
Sequence-based	β-lactamase/Kemp elimination	Ancestral sequence reconstruction	New enzyme	(Risso et al., 2017)
Data-driven	Halohydrin dehalogenase	Protein sequence activity relationships	Activity	(Fox et al., 2007)
Data-driven	Proteinase K	Linear regression algorithm	Activity and stability	(Liao et al., 2007)
Data-driven	Epoxide hydrolase	Support vector regression	Selectivity	(Zaugg et al., 2017)
Data-driven	Limonene epoxide hydrolase	Innovative sequence-activity relationship	Stability	(Li et al., 2021b)
Data-driven	Chorismate mutase	Evolution-based statistical approach	Activity	(Russ et al., 2020)
Data-driven	PET hydrolase	Three-dimensional self-supervised convolutional neural network	Activity and stability	(Lu et al., 2022)
Data-driven	Carbonic anhydrase II and Δ^5–3-ketosteroid isomerase	Constrained hallucination and inpainting	New enzyme	(Wang et al., 2022)
Data-driven	Luciferase	Family-wide hallucination	New enzyme	(Yeh et al., 2023)

for backbone atoms, meeting the requirements for analyzing catalytic mechanisms and facilitating rational enzyme design (AlQuraishi, 2021). Structure-based rational computational design can be further classified into two subclasses based on whether it involves the computational redesign of existing enzymes or creating new enzymes with novel functions (Dinmukhamed et al., 2021).

2.1. Computational engineering of existing enzymes

In 1993, Arnold introduced the concept of directed evolution for enzymes, promoting the idea of restructuring the structure and function of natural enzymes through the random replacement of amino acid residues (Chen and Arnold, 1993). Presently, computational strategies based on virtual screening have emerged as effective means to obtain target mutants. For instance, Zhang et al. utilized both molecular dynamics (MD) simulations and binding free energy calculation in their investigation, building upon the structure and function of phosphoserine aminotransferase (SerC). They incorporated site-directed engineering strategy for virtual screening. Finally, the mutant SerC R42W:R77W exhibiting a minimum ΔΔG ($\Delta\Delta G_{binding} = \Delta G_{binding(mutant)} - \Delta G_{binding\ (wild\ type)}$) of −0.6 kcal/mol, was experimentally obtained. L-phosphoserine was replaced by L-homoserine as the preferred substrate for SerC, representing a 4.2-fold increase in activity compared to the wildtype enzyme (Zhang et al., 2019). Similarly, Mori et al. applied a comparable strategy to tailor ferulic acid decarboxylase (FDC), targeting the unnatural substrate *cis*,*cis*-muconic acid (ccMA) to efficiently produce 1,3-butadiene. The best mutant, AnFDC Y394H:T395Q, displayed a remarkable 1002-fold increase in 1,3-butadiene titer (Mori et al., 2021). Recently, focused rational iterative site-specific mutagenesis (FRISM) was developed, drawing inspiration from the iterative saturation mutagenesis and the combined-active site saturation test. Differing from previous methods, the FRISM introduces only 3–5 amino acids in the hotspots, each with notable characteristics, including distinct steric, electrical, hydrophilic, and hydrophobic properties. This approach eliminates the need for saturation mutagenesis libraries and large-scale screening, thereby reducing computational resources and time while ensuring positive results (Li et al., 2021a; Li et al., 2021c). In line with the FRISM strategy, Xu et al. achieved mutants of fatty acid photodecarboxylase (*Cv*FAP) capable of efficiently converting fatty acids with diverse chain lengths into corresponding hydrocarbons (Xu et al., 2022). Taking into account steric hindrance, they substituted ten amino acids in hotspots with A/L/F/Y to adjust the steric hindrance of the substrate binding pocket. This led to *Cv*FAP mutants (*Cv*FAP I398L and *Cv*FAP P460A/G462A) exhibiting a substantial 29 to 552 times increase in enzyme activity (k_{cat}/K_m) for decarboxylation of short- and medium-fatty acids (C3-C14). The FRISM strategy also demonstrated effectiveness in designing the stereoselectivity of the enzyme. Li et al. applied rational design using FRISM to stabilize elaidic acid binding in CvFAP by introducing p-π interactions with the substrate's double bond. The V453E mutant exhibited a remarkable 1000 times enhancement in trans-over-cis selectivity than the natural enzyme (Li et al., 2021a). Additionally, several high-throughput computational enzyme engineering approaches have been reported. CADEE (computer-aided directed evolution of enzymes) employs *in silico* enzyme directed evolution to reduce the need for extensive screening and enhance the efficiency of identifying desired mutants (Amrein et al., 2017). EnzyHTP software is able to automatically handle the simulation flow, covering tasks such as structural model construction, MD simulation, quantum mechanics (QM)/molecular mechanics (MM) calculation, and modeling data analysis. This automation streamlines the identification of successful enzyme mutants in a high-throughput manner (Shao et al., 2022, 2023; Yang et al., 2023).

The rational design methods discussed above, relies on the idea of directed evolution, often require virtual iterative mutation of hot spots. In contrast, it is more resource-efficient to directly modify the local chemical micro-environment, involving factors such as the hydrogen bond network, steric hindrance of amino acid residues, electrostatic interactions, *etc.*, which significantly impact enzyme performance (Fig. 1). Effective hydrogen bonding within the active site and protein structure plays a vital role in stabilizing enzyme-substrate complexes, transition states, and protein folding processes. Bulky groups such as large amino acid side chains can obstruct or restrict access to the enzyme's active site, impacting substrate binding and catalytic activity. Additionally, electrostatic interactions involving charged residues and substrates can enhance catalysis by stabilizing charged transition states or forming specific binding interactions. Lin and coworkers employed 4-hydroxyphenylpyruvate dioxygenase (HPPD) as an example, manipulated its hydrogen bond network in the amide-rich zone, resulting in 100% conversion of the substrate to the desired intermediate 4-hydroxyphenylacetic acid (HPA)(Lin et al., 2021). In addition, the size of the active pocket can influence substrate conversion, particularly when amino acid residues with steric effects in the active center may impede the movement of the substrate to the reactive position. For example, the catalytic mechanism of santalene synthases was explored through multiscale simulations. The results shown that the restricted space within the active pocket of the enzyme SanSyn, caused by the key residue F441, specifically led to the specific production of the α-santalene. To enhance the active pocket space, mutant F441V was created, resulting in increased conformational change of the reaction intermediate and the generation of both α-and β-santalene (Zha et al., 2022). Unlike the previously mentioned FRISM method, this approach provides a more targeted and efficient strategy by directly identifying the crucial amino acid. This reduces the need for extensive screening and experimental validation. Moreover, Su et al. introduced a strategy for enzyme redesign towards unnatural substrates, identifying key residues with distinct conformations through restricted MD simulations. This approach was employed to construct a dehydrogenase/reductase variant, showcasing apparent conformational differences in simulations upon modifying the substrate-binding pocket. The relative sizes of the C1 and C2 pockets were tuned to enhance stereoselectivity, resulting in mutants with efficient asymmetric reduction of various substrates (Su et al., 2020). In addition, modifying the electrostatic potential or electric field applied by the enzyme to essential atoms of the substrate plays an important role in catalytic activity. This impact occurs through various mechanisms such as stabilizing the transition state of the reaction, altering the activation energy required for the transition state, promoting or preventing proton transfers during catalysis and changing the flexibility and dynamics of the enzyme's active site (Warshel et al., 2006). A quantum QM/MM dynamics strategy was used to redesign a promiscuous esterase Bs2 from *Bacillus subtilis*, aiming to improve its amidase activity. They transferred a spatially relevant aspartate residue from *Candida antarctica* lipase B (CALB) to Bs2, improving the electrostatics of transition state formation. This mutation exhibited a 1.3-fold improvement in catalytic efficiency (Galmés et al., 2022). Furthermore, integrating electric field optimization into enzyme engineering processes can yield substantial catalytic enhancements. An electric field optimization scheme was developed to improve the activity of a synthetic Kemp elimination enzyme KE15. By systematically considering the impact of strategic mutations on the local electric field along the reaction axis, researchers created a new Kemp eliminase with a nearly 50-fold increase in the k_{cat} value (Bhowmick et al., 2016; Vaissier et al., 2018). Except the electric field, substrate positioning dynamics (SPD) also serves as a significant factor for enzyme catalysis *via* aligning the substrate in a reactive conformation. The noteworthy contribution of the non-electrostatic component of SPD to regulating enzyme kinetics was recently validated by employing Kemp eliminase as an illustrative example. A distal mutant R154W created by high-throughput enzyme modeling, displayed favorable SPD. This led to an increased proportion of reactive conformations and consequently yielded the lowest activation free energy (Jiang et al., 2023).

The computational engineering of natural proteins with new functions holds the potential to revolutionize sustainable biological

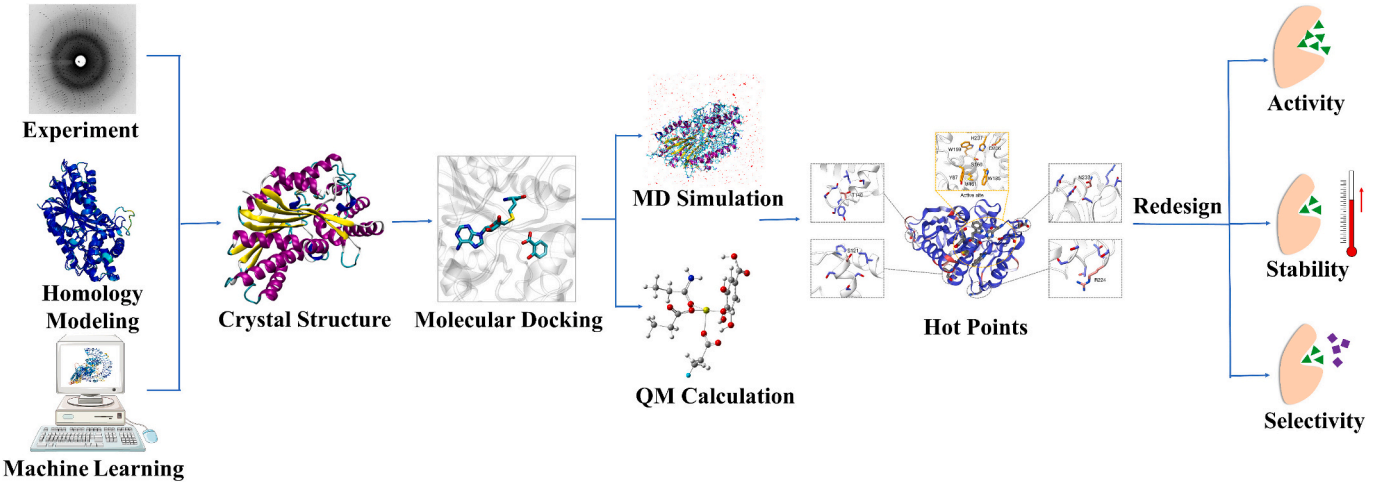

Fig. 1. Flowchart depicting the systematic process for enzyme redesign based on protein structure. Initially, protein crystal structures are acquired using experimental techniques, homologous modeling, or machine learning methodologies. Subsequently, the structural analysis is conducted to scrutinize the substrate-enzyme interaction by employing molecular docking, molecular dynamics (MD) simulations, and quantum mechanics (QM) calculations. This comprehensive examination aims to pinpoint key hot spot residues crucial for enzyme function. Finally, the target enzyme is achieved by strategically designing and incorporating modifications to the identified hot spot residues.

manufacturing. However, the strategies mentioned earlier are mainly utilized to boost or control the selectivity or activity of enzymatic reactions, proving less effective for engineered enzymes with aspirations to introduce entirely new functions. This limitation stems from the difficulty in transitioning the sequence space of the active site to entirely new protein fitness landscapes (Cui et al., 2022). Nonetheless, Rosetta redesign, which considers amino acid mechanical constraints to redesign suitable active pockets, has shown promise in creating efficient mutants with novel functions (Zanghellini et al., 2006). Rosetta redesign was first applied in the re-computational engineering of the benzaldehyde binding pocket of formolase (FLS) to enhance the catalytic efficiency for formaldehyde and the formose reaction. Following four cycles of computational design and subsequent experimental assessment, one of the 121 designed mutants, called Des1, exhibited a 26-fold increase in catalytic activity than the wildtype (Siegel et al., 2015). Despite these promising results, Rosetta redesign faces challenges that need to be addressed. First, the challenges associated with Rosetta's computational enzyme design include its random and often insufficient sampling of enzyme reaction conformations (Osuna, 2021; Tyka et al., 2012; Tyka et al., 2011; Wijma and Janssen, 2013). Additionally, the scoring function used by Rosetta compromises calculation accuracy to enhance speed, a trade-off coupled with limitations in calculation time and starting models (Bender et al., 2016). Therefore, the energy function cannot accurately describe the relevant biological state of the enzyme-catalyzed reaction. To address this, the Rosetta redesign strategy has integrated multi-state conformation design and high-throughput parallel MD simulations to enlarge the sample pool (Löffler et al., 2017). The second computational enzyme design challenge pertains to arranging catalytic residues optimally, especially in the near attack conformation (NAC), which is disrupted during dynamic catalytic processes (Ruscio et al., 2009). By relying on the NAC strategy, the catalytic selectivity by computational design (CASCO) framework addresses this challenge through the design of substrate binding sites in a predetermined orientation. This process includes generating steric hindrance to inhibit undesired substrate binding modes, and the subsequent ranking of results is carried out through high-throughput MD simulations. The application of this strategy resulted in the successful acquisition of a highly stereoselective limonene epoxide hydrolase, achieved with relatively minimal experimental work (Wijma et al., 2015). Recently, Wu et al. redesigned an aspartase AspB to enable the cross-addition of a diverse of nucleophilic amines to unsaturated acids. This process starting by docking ligands with the protein crystal structure through QM/MM calculations. Subsequently, different docking results were simulated using MD simulations. The outcomes of the MD simulations were then utilized in the Rosetta enzyme design. The mutants produced in this process were sorted according to NAC frequency, penalty scores of constraints and total energy scores. Mutants ranked highest were chosen for experimental validation (Fig. 2). Finally, they achieved C—N lyases that exhibit cross-compatibility for unnatural nucleophiles and electrophiles, demonstrating the ability to produce diverse non-canonical amino acids with outstanding catalytic efficiency, regioselectivity, and enantioselectivity (Cui et al., 2021).

Except for enzyme activity, the enzyme stability that can maintain its functionality and structural integrity across diverse and challenging conditions is also paramount for ensuring its sustained effectiveness in industrial applications. To improve the thermostability of designed enzymes, manipulation of the electric charge distribution on the enzyme surface (Wang et al., 2020), disulfide bonds (Yang et al., 2019), rational truncation (Reich et al., 2014) and flexible loop replacement (Damnjanović et al., 2014) are the most commonly used strategies. These strategies are primarily employed to decrease the entropy of unfolded proteins and improve the stability of protein conformation (Dagan et al., 2013; Kumar et al., 2004; Vieille and Zeikus, 1996). By analyzing the three-dimensional structure of *Rhizomucor miehei* lipase (RML), predictions for enhanced thermostability in RML variants can be made using tools such as Rosetta ddg_monomer, FoldX, and I-Mutant. Rosetta ddg_monomer utilizes energy-based calculations to predict mutation-induced changes in free energy, offering detailed physics-based modeling albeit requiring more expertise. FoldX employs an empirical force field to estimate protein stability changes, striking a balance between accuracy and user-friendliness. I-Mutant specializes in predicting mutations' effects on protein stability, particularly concerning protein unfolding and thermostability. If enhancing enzyme thermal stability is the goal, I-Mutant is a focused option to consider. However, for broader considerations such as structural or functional changes, or for more intricate physics-based modeling, Rosetta or FoldX may be more suitable choices. Additionally, Disulfide by Design 2 (DbD2), SSBOND, MODIP, and BridgeD were employed to identify potential disulfide regions and residues capable of forming disulfide bonds. DbD2 primarily serves protein design and engineering needs by optimizing both stability and function. SSBOND predicts and analyzes disulfide bonds crucial for stability and structure. MODIP focuses on protein-protein interactions and their impact on stability and function, while BridgeD identifies and analyzes bridging water molecules affecting

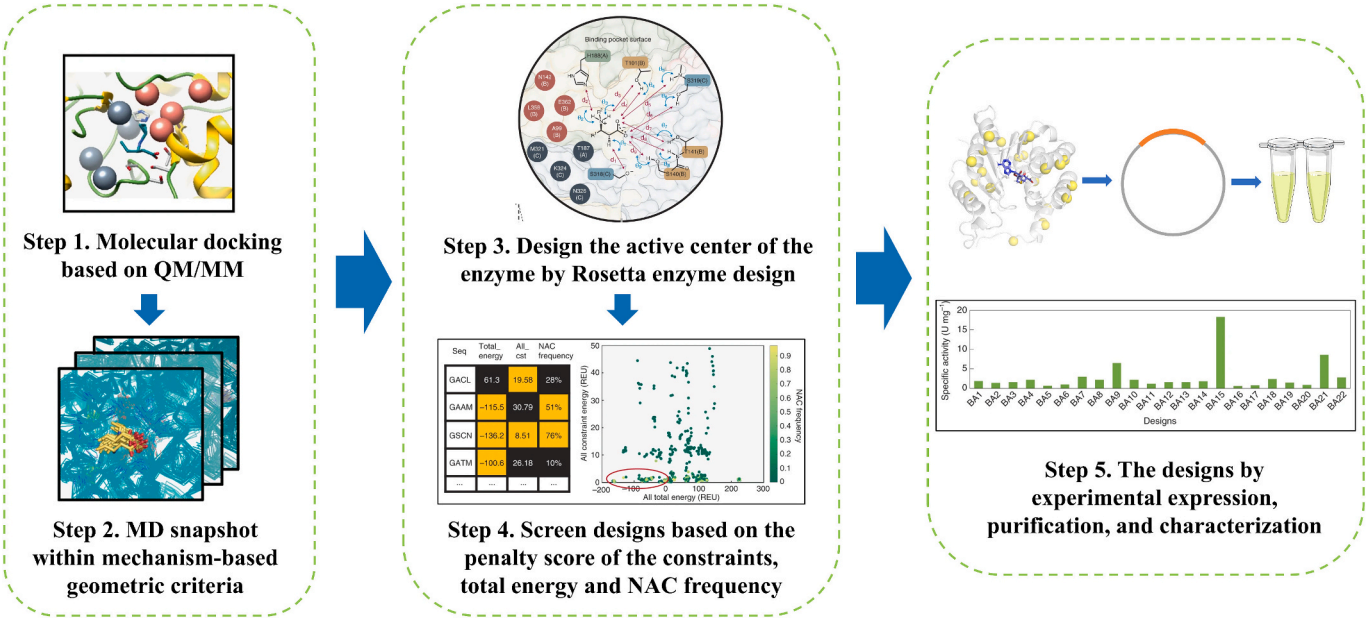

Fig. 2. Schematic representation of the enzyme redesign process employing Rosetta redesign methodology.

stability and dynamics. DbD2 and SSBOND target specific aspects like mutations or disulfide bonds, while MODIP and BridgeD offer broader insights into interactions and hydration. The most stable mutant identified in the study exhibited a remarkable 12.5 times enhancement in half-life at 70 °C, indicating a substantial enhancement in the thermal stability of RML (Li et al., 2018a). Truncation and replacement of flexible loops have proven to be effective strategies for significantly improving enzyme stability. However, identifying the specific loops that impact enzyme stability remains challenging, primarily because of the significant rearrangement of the tertiary structure. This complexity makes it difficult to pinpoint the exact loops that contribute to or influence the stability of the enzyme (Nagi and Regan, 1997; Yu et al., 2017). Han et al. employed computational-aided design method to examine the influence of the C-terminus on the functionality of GH2 fungal glucuronidase. Through structural analysis, they determined that the C-terminus of the enzyme was redundant for maintaining normal function. The C-terminus deletion induced an optimal conformation in the active site pocket, facilitating substrate binding and catalysis. Modifying the length of the C-terminal through truncation resulted in a PGUS D591:604 mutant that exhibited 3.8 times improvement in half-life at 65 °C, 2.4 times increase in enzyme activity, and 1.8 times increase in expression level. These enhancements signify significant improvements in both kinetic and thermodynamic stability of the enzyme (Han et al., 2018). Wijma and coworkers established a computational libraries framework, FRESCO, for rapid enzyme stabilization. This approach involves initially computing a large number of folding energies ($\Delta\Delta G_{fold} = \Delta G_{fold(mutant)}$-$\Delta G_{fold(wild\ type)}$) to identify potential stable point mutations. Subsequently, high-throughput MD simulations and visual inspection was utilized to screen all potential point mutations. Finally, experimental validation is conducted by combining the identified mutations to significantly enhance the enzyme thermal stability. While computational design methods excel at precise mutation and structural modifications for targeted engineering purposes, FRESCO can utilize available data to inform design decisions, especially when detailed sequence and structural information are accessible. Successful applications of FRESCO in enzyme stability engineering have been demonstrated across various enzymes (Bu et al., 2018; Floor et al., 2014; Fürst et al., 2019; Meng et al., 2020; Wu et al., 2016). Especially, FRESCO was used to improve the thermostability of a homodimeric pyridoxal-5-phosphate (PLP) ω-transaminase. Numerous surface point mutations were predicted by the computational tool. The beneficial mutations, which included P9A, E38Q, A60V, S87D/N, M128F, and I154V, resulted in improved thermostability, co-solvent resistance, isopropylamine compatibility, catalytic activity and the complete retention of enantioselectivity. These improvements were attributed to introduction of salt-bridge interactions or extra hydrogen bonding, increase of hydrophobic interactions, imposition of π-stacking, reduction of hydrophobic surface exposure, altering electrostatic surface charge distribution and the alleviation of steric strain (Meng et al., 2020).

2.2. *Computational design of new enzymes with novel function*

The redesign of enzymes relies on modifications to engineer existing enzymes that have functions similar to the target function or have the appropriate geometry and sufficient stability to tolerate variants needed for incorporating the target functions. However, natural enzymes contain a limited number and type of chemical reactions, which limits the potential of these enzymes to be used by the industry. Creating novel enzymes to catalyze non-natural reactions is a promising solution, which allows for breaking out from the locally optimal configurations found in natural protein sequences, installing amino acids in protein scaffolds with non-classical biological functions to enable new chemical reactions (Lu et al., 2009; Nanda and Koder, 2010).

The creating novel functional proteins with the ability to catalyze unnatural reactions is still in its early stages. Notable advancements were made by Houk and Baker in their cooperative effort to develop an "inside-out" protocol, which involves dividing the enzyme into two distinct components: a catalytic part and a scaffold part. The protocol then separately calculates and designs these two parts, following a series of general steps:1) utilizing QM, an ideal transition state, known as theozyme, is generated within the active site. This theozyme represents the most stable configuration during the enzymatic reaction; 2) the generated transition state is transplanted onto the protein scaffold; 3) The design and optimization of the amino acid residues around the theozyme aim to guarantee favorable stacking and stability in folding (Kiss et al., 2013). Building upon this foundation, electronic and geometric properties of the transition state are incorporated to generate enzymes with novel catalytic functions. The earliest algorithms for grafting theozyme onto protein scaffolds primarily include RosettaMatch and SABER (Fig. 3). The RosettaMatch algorithm is designed to

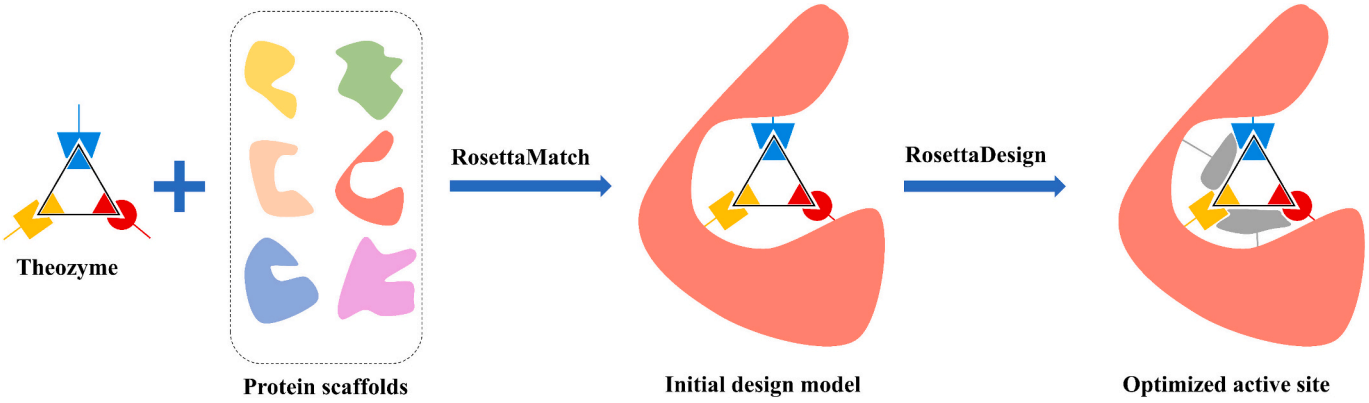

Fig. 3. Key steps in the "inside-out" protocol for enzyme *de novo* design.

identify backbone positions on a given protein scaffold and graft essential catalytic residues onto the scaffold. His process is focused on maintaining and stabilizing the reaction transition state by preserving a corresponding geometry (Richter et al., 2011; Zanghellini et al., 2006). The resulting protein, equipped with grafted catalytic residues, may demonstrate the ability to catalyze the enzymatic reaction. It's noteworthy that the RosettaMatch approach may require a higher number of mutations to adapt to new catalytic functions and substrates, potentially causing significant disruption to enzyme stability and catalytic performance. On the other hand, SABER, developed by Houk and Nosrati, offers an alternative strategy. Instead of inserting theozyme into predefined active sites, SABER searches the Protein Data Bank database to obtain native protein scaffolds with suitable catalytic functions. Once a match is found, SABER selectively mutates amino acid residues that could interfere with the correct binding of the substrate. This mutation process aims to preserve the integrity of the ideal theozyme (Kiss et al., 2013). Importantly, the success of these computational designs is ultimately validated through experimental verification (Richter et al., 2011). Siegel and coworkers employed the "inside-out" enzyme engineering strategy for asymmetric catalysis of bimolecular Diels-Alder reactions. First, the mechanism of the Diels-Alder reaction was elucidated through QM calculations. A transition state model comprising essential amino acid residues and substrates was generated. Second, the transition state model was matched to a group of protein scaffolds with high stability using RosettaMatch. The use of these scaffolds facilitated the positioning of the theozyme without inducing notable steric clashes with the protein backbone. Third, a total of 106 potential active sites were matched on a stable protein scaffold. RosettaDesign was employed to refine each match, maximizing transition state binding without conflicting with bound substrates or products. Finally, a total of 84 variants were chosen, taking into account the satisfaction of shape complementarity, catalytic geometry and binding energy. The experimental verification provided evidence that two enzymes, employing different matched protein scaffolds, were effective in catalyzing the Diels-Alder reaction. Subsequently, 6 site mutations were introduced to the candidate DA_20_00 to improve its activity. This mutational strategy successfully increased its catalytic activity. The experimental results aligned with the expected design, as DA_20_10 demonstrated the ability to catalyze the formation of 3R,4S products with >97% enantiomeric excess (Siegel et al., 2010). In addition to Diels-Alder reactions, *de novo* enzyme design has been developed for other reactions, including Kemp elimination reactions (Röthlisberger et al., 2008), Retro-Aldol reactions (Jiang et al., 2008), ester hydrolases(Richter et al., 2012), and Morita-Baylis-Hillman reactions(Bjelic et al., 2013; Burke et al., 2019; Crawshaw et al., 2022).

In recent years, advances in creating enzymes with novel functions has faced challenges, and several factors contribute to this trend. First, many unnatural chemical reactions involve intricate transition states, and the difficulty of *de novo* design is exacerbated by the complexity of these transition states, constrained by both technical and theoretical limitations. Secondly, based on current observations, the catalytic performance of *de novo*-designed enzymes is significantly lower than that of natural enzymes, rendering them less suitable for practical applications. To address this, there is a necessity to enhance the enzyme efficiency of *de novo* enzymes through iterative rounds of directed evolution or rational design (Basler et al., 2021; Broom et al., 2020; Khersonsky et al., 2012; Khersonsky et al., 2011; O'Reilly, 2022; Röthlisberger et al., 2008; Siegel et al., 2010; Völler, 2020). For instance, Weitzner et al. identified a comprehensive network of hydrogen bonds involving catalytic residues in an aldolase designed after directed evolution. Recognizing the limitation of the RosettaMatch method, which primarily considers interactions between protein side chains and the transition state while overlooking hydrogen bonding between residues, they developed a side-chain optimization method called HBNetGen. This method enhanced *de novo* enzyme design by incorporating a network of hydrogen bonds between residues in the transition state model (Obexer et al., 2017; Weitzner et al., 2019). Efficient search for protein scaffolds is crucial for successful *de novo* enzyme design. Addressing this, Zhang et al. developed ProdaMatch, an algorithm that rapidly and accurately matches protein scaffolds throughout the entire protein database (Zhang et al., 2020). To overcome the obstacle of a limited number and variety of protein scaffolds in the current database, Baker and coworkers developed an enumeration algorithm for creating novel proteins with different pocket structures and arbitrary functions, which can generate an almost infinite number of novel proteins. After two rounds of extensive experimental testing and refinement, the enhanced algorithm can produce proteins that maintain their folded structure at elevated temperatures. These proteins also showcase greater pocket diversity compared to natural NTF2-like proteins. This method enables the creation of active site geometries better suited for specific designs and promises to change the status quo of accomplishing *de novo* design of enzymes by reusing limited amounts of naturally occurring NTF2-like proteins (Basanta et al., 2020). Expanding on this foundation, Baker et al. advanced their research by integrating the catalytic site of luciferase into a computationally designed NTF2 protein scaffold, leading to the creation of a luciferase variant with specific targeting capabilities for 2-deoxycoelenterazine (Yeh et al., 2023). Except for structural enhancements in the *de novo* enzyme design process, researchers have proposed novel energy functions to assist in the design. Traditional energy functions often assume independent and additive effects for statistical energy terms, such as backbone dihedral angle, solvent exposure, and secondary structure type. Liu et al. introduced an energy function to design amino acid sequences based on a given backbone structure, named A Backbone-based Amino Acid-Usage-Survey. Derived mainly from statistics in the PDB, the ABACUS model utilizes kernel density estimation and neural network learning to represent multidimensional,

high-order correlations of energy between different structures in known protein structures. This approach introduces neural network-based statistical energy terms that capture multidimensional features not easily described by traditional statistical methods (Huang et al., 2022).

3. Sequence-based enzyme computational engineering and design

Both homology modeling and deep learning-based protein structure prediction approaches are capable of providing accurate and reliable three-dimensional structures for proteins. However, the effectiveness of these methods relies on the presence of resolved crystal structures for homologous proteins in the database. If an appropriate template for homology modeling is unavailable or of poor quality, the prediction may be compromised. Additionally, obtaining the crystal structure experimentally is a time-consuming and labor-intensive process. Furthermore, once the protein structure is acquired, structure-based computational design methods demand a significant understanding of theoretical chemistry and computational techniques to explore chemical mechanisms and transition states. This could pose challenges for researchers lacking expertise in these areas. In contrast, sequence-based computational design methods offer a partial solution to these challenges. These methods allow the direct design of enzymes from protein sequences, bypassing the need for explicit structural information. This approach enables researchers to glean insights from the natural evolution of enzymes over billions of years. With the wealth of sequence data made available by next-generation sequencing methods, researchers can employ phylogenetic analysis to uncover the underlying principles of enzyme evolution (Stephens et al., 2015).

Consensus Design (CD) and Ancestral Sequence Reconstruction (ASR) are two of the most widespread sequence-based computational design methods (Fig. 4) (Musil et al., 2018). CD relies on homologous protein sequences, typically utilizing tens to hundreds of such sequences to generate a Multiple Sequence Alignment (MSA). The distribution frequency of each amino acid at a specific position is then calculated, and a user-defined conservation threshold is applied to identify "consensus" amino acids during evolution. The underlying assumption is that the most common amino acid at a specific position is more likely to be stable and evolutionarily conserved, potentially affecting enzyme catalysis (Magliery, 2015; Porebski and Buckle, 2016; Steipe et al., 1994). Teze et al. introduced a sequence-based method for creating highly efficient glycosyltransferases (GTs) from retaining glycoside hydrolases (GHs) without relying on enzyme structure information. The method involves collecting a substantial number of sequences within the GH family, clustering to reduce redundancy, and iteratively performing MSA to choose sequences with >5% identity. The conserved residues are identified through iterative increases in the sequence identity threshold, and the most conserved residues in GH are replaced with structural analogs. This method has proven to be effective in engineering of enzymes from various GH families (Teze et al., 2021). To address the challenge that conserved amino acid residues may not always be ideal mutation targets, Sullivan et al. enhanced the consensus design by removing nearly invariant positions and sites exhibit high statistical correlations with other positions. This modification improved the identification rate of stable mutations in *Saccharomyces cerevisiae* triosephosphate isomerase from 50% to 90% (Sullivan et al., 2012). Except for consensus design, structure-guided consensus design methods leverage structural information to identify potentially damaging mutations. Vázquez-Figueroa et al. proposed a method incorporating distance between potential mutation sites and active sites, secondary structure information, and the overall count of intramolecular contacts. This method led to a substantial improvement in the stability of mutant glucose-1-dehydrogenase, with a 10^6-fold improvement when combining mutation sites (Vázquez-Figueroa et al., 2007). Another efficient structure-guided consensus design method involves analyzing molecular fluctuations by considering crystallographic B-factors (Parthasarathy and Murthy, 2000). Despite its strengths, consensus design has limitations, such as an inability to explain epistasis between mutation results and noticeable phylogenetic bias when certain subfamilies dominate the MSA (Aerts et al., 2013; Cole and Gaucher, 2011; Hochberg and Thornton, 2017; Lehmann et al., 2000).

Proteins exhibit diverse catalytic functions, often evolving from pre-existing functions. However, the mechanisms underlying the emergence of new enzyme functions and their evolutionary processes remain inadequately understood. ASR employs the Bayesian inference or maximum likelihood method to explore the evolutionary history of homologous sequences from known MSA and appropriate phylogenetic models. This method infers the replacement probability of a defined amino acid at a specific site in the target enzyme and reconstructs the phylogenetic tree containing putative ancestral sequences (Selberg et al., 2021; Spence et al., 2021). Despite the inherent ambiguity in ASR, where even the most likely sequence obtained may not accurately represent the true ancestral sequence in evolutionary history, the method remains valuable. A cytochrome P450 was engineered by the ASR and the optimal mutant exhibited improved solvent stability and thermostability than a human cytochrome P450, despite sharing similar reactivity profiles. Additionally, this work also created a mutant ketol-acid reductoisomerase that showed an 8-fold increase in catalytic efficiency (Gumulya et al., 2018). ASR has also been instrumental in exploring the sequence-activity landscape of a self-sufficient P450

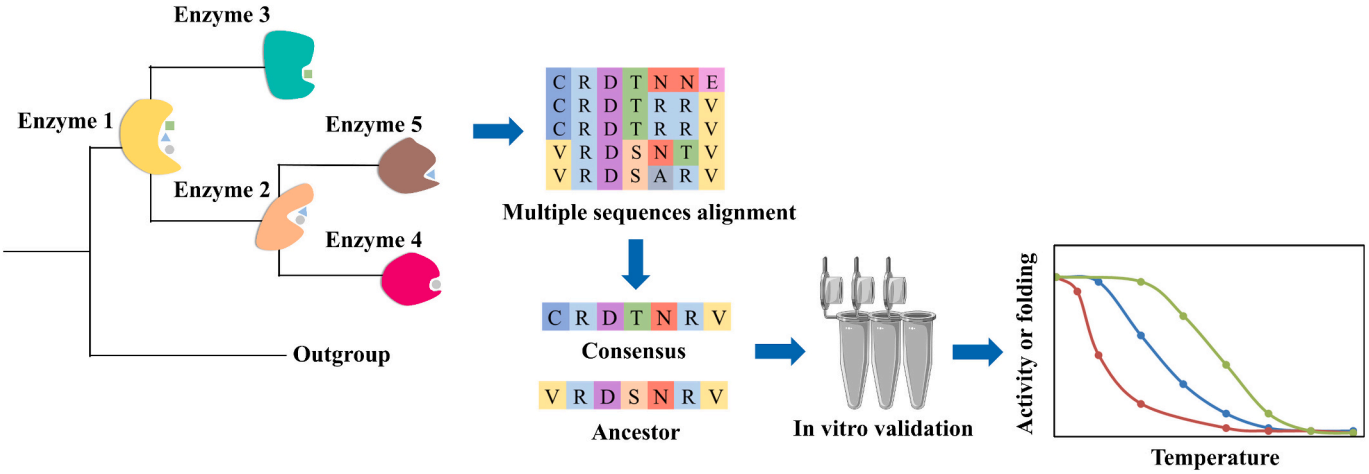

Fig. 4. Overview of consensus design (CD) and ancestral sequence reconstruction (ASR), illustrating the general workflow that encompasses the collection and alignment of homologous sequences, phylogenetic analysis, and experimental characterization.

monooxygenase CYP116B. It identified crucial mutational paths leading to changes in enzyme selectivity, resulting in ancestor variants with shifted regioselectivity from terminal to mid-chain hydroxylation of fatty acids(Jones et al., 2024). However, ASR faces challenges due to the high differentiation among extant proteins, leading to uncertainties in the predicted probabilities of many sites, often falling below 50%. This limitation underscores the difficulty in confidently determining the real ancestral sequence through ASR (Copley, 2021). However, despite these challenges, the protein scaffold obtained through ASR serves as an important starting point for protein design efforts owing to the inherent high thermostability and evolvability. ASR-derived protein scaffolds play a crucial role in studying protein conformational dynamics related to thermostability (Babkova et al., 2020), ligand binding specificity, allosteric regulation and catalytic activity (Jemth et al., 2018; Kaczmarski et al., 2020).

Recently, an escalating number of researchers have recognized the significant implications of conformational dynamics for the catalytic promiscuity and evolution of enzymes (Babkova et al., 2020; Campitelli et al., 2020; Crean et al., 2020; Gardner et al., 2020; James and Tawfik, 2003; Maria-Solano et al., 2018; Spence et al., 2021; St-Jacques et al., 2023; Zamora et al., 2020). James and Tawfik were among the first to propose that conformational plasticity contributes in enzyme evolution, enabling enzymes to bind new ligands and adopt conformations that catalyze novel chemical reactions (James and Tawfik, 2003). These conformations were generally rare in ancestral enzymes, but natural evolutionary pressures favored the prevalence of such rare conformations, leading to the creation of new catalytic reactions for previously unexplored substrates. Laboratory-based directed evolution can facilitate the generation of these new conformations more readily than in natural evolution, as the lower level of catalytic promiscuity can be detected and amplified in a controlled laboratory environment. Currently, two methods have been developed to fine-tune conformational dynamics for enhancing enzyme activity: 1) expanding the sampling of states that can impart new catalytic activity; 2) decreasing the sampling of non-productive conformations (Campbell et al., 2018). Recent research has demonstrated that ASR can provide a starting scaffold for modifying catalytic properties, with results that can be explained by conformational dynamics. Risso et al. reconstructed Precambrian β-lactamase using ASR as a scaffold for protein engineering. They designed an artificial Kemp eliminase with catalytic activity up to 10^7 higher than the uncatalyzed reaction, along with significant ester hydrolysis activity. Both experimental and computational analyses revealed that despite substantial differences in amino acid sequences and protein scaffolds across evolutionary time, the overall tertiary structure remained relatively unchanged (Risso et al., 2017). The "Shortest Path Map (SPM)" method is a notable advancement in computational enzyme engineering. This method utilizes a deep understanding of conformational dynamics and distal mutations, complemented by long-time-scale molecular dynamics (MD) simulations, to explore a wide range of conformations and predict mutations that can induce an allosteric impact on protein function. For instance, the direct evolution of a retro-aldolase enzyme led to a variant exhibiting a remarkable $>10^9$ rate enhancement. Through the analysis of enzyme active site conformational dynamics, it was discovered that long-range mutations played a crucial role in shifting populations of conformational states towards active states. The key amino acids identified through the SPM method were consistent with those mutated in the direct evolution process (Romero-Rivera et al., 2017). The SPM method also has proven successful in understanding of the catalytic mechanism of enzyme reactions. An engineered cytochrome P450 monooxygenase exhibited regio- and stereoselective hydroxylation activities towards steroid. The mechanism was investigated by SPM and the results revealed that the epistatic effects and conformational dynamics are influenced by distal interactions in loops, β-strands and helices, which control the substrate access tunnel. This regulation ultimately facilitates optimal catalysis (Acevedo-Rocha et al., 2021). Similarly, the molecular mechanism of tryptophan synthase for stand-alone activity (Maria-Solano et al., 2019; Maria-Solano et al., 2021) and monoamine oxidase for broadening substrate scope (Curado-Carballada et al., 2019) were also elucidated using this approach. Studying the evolutionary history of native proteins enhances our understanding of protein engineering, emphasizing the importance of considering protein dynamics, neofunctionalization, and epistatic interactions in the design and engineering of proteins. Recently, several user-friendly web-based tools have been developed for improving enzyme performance without requiring extensive expertise or installation. For instance, Caver is a specialized tool used for the detailed analysis and visualization of substrate access tunnels and channels in protein structures (Stourac et al., 2019). HotSpot Wizard is designed to automatically identify "hotspots" in proteins, facilitating the engineering of substrate specificity, enzyme activity, or enantioselectivity (Pavelka et al., 2009). The Rosetta-based PROSS is specifically used for designing enzymes to ensure stability and high expression levels (Gomez de Santos et al., 2023). FuncLib is designed to generate multipoint mutations, specifically honing in on the active sites of enzymes (Khersonsky et al., 2018).

4. Data-driven enzyme engineering and design

Indeed, the structure-based and sequence-based computational design approaches can offer powerful and complementary strategies for engineering of enzymes with desired properties. Despite the promising results shown by both methods, our knowledge of the sequence and structure of native enzymes is still in its infancy. Consequently, the computational design of enzymes, based on limited sequence or structural information, necessitates a substantial computational or experimental investment, thereby amplifying the overall cost of enzyme engineering. In the vast realm of nature, there exists substantial, untapped sequence and structural space. Efficient exploration of these uncharted territories within protein sequence space holds the potential for the discovery of novel and valuable enzymes. Integrating machine learning (ML) into enzyme engineering becomes pivotal, as it empowers the modeling of intricate sequence-function and structure-function relationships using existing data. ML facilitates the prediction of new, valuable enzymes that may prove challenging to attain through conventional methods. This approach not only enhances our understanding of these relationships but also guides the design of highly efficient enzymes, closely mirroring the efficacy observed in natural systems. Various ML algorithms have been employed for enzyme engineering, showcasing their versatility. For instance, random forests have been employed for predicting protein solubility (Yang et al., 2016; Yang et al., 2021). Support vector machines and decision trees have found utility in predicting changes in enzyme stability after mutation (Folkman et al., 2016; Teng et al., 2010) (Huang et al., 2007). K-nearest neighbor methods, such as the K-Nearest Neighbor Classifier, have been utilized for predicting enzyme functions and catalytic mechanisms (De Ferrari and Mitchell, 2014; Koskinen et al., 2015). Moreover, diverse scoring and clustering algorithms have been utilized for the swift annotation of functional sequences (Cozzetto et al., 2013; Falda et al., 2012). The main allure of ML in protein engineering lies in its ability to provide rapid predictions once trained on a suitable dataset. In contrast, rational design necessitates the construction of new models and extended periods of intensive computation, while directed evolution methods often require months of experimental works. The efficacy of ML relies on the quality of the training dataset and the efficiency of the underlying algorithms. Challenges such as the need for rigorous control over data collection and reporting, difficulties in standardizing data formats, the lack of large homogeneous datasets for training, slow establishment of new datasets for model testing, and the diversity of reactions, catalytic mechanisms and experimental conditions, significantly limit the widespread application of ML in designing biocatalysts (Fig. 5).

Supervised ML are commonly employed in enzyme engineering, where researchers represent enzyme fitness with a given label. The

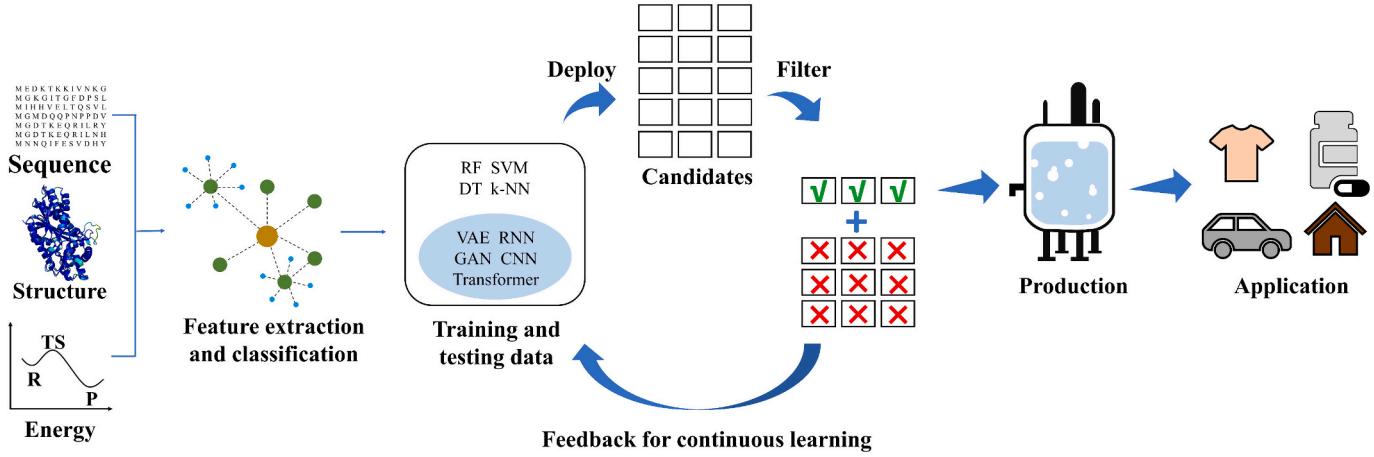

Fig. 5. Employ diverse machine learning techniques to categorize and train the various crucial parameters of enzymes, thereby obtaining corresponding models. Subsequently, apply these models to the systematic and rational design of enzymes, ultimately resulting in the delineation of the process for obtaining target enzymes.

choice of features and models distinguishes various ML methods. Generally, selected features in enzyme engineering can include protein sequences (Foroozandeh Shahraki et al., 2021; Greenhalgh et al., 2021; Vasina et al., 2022), structures(Feehan et al., 2021; Krivák and Hoksza, 2018; Unger et al., 2020), or energy data(Li et al., 2022; Pan et al., 2021; Xie et al., 2022). Sequence-based features have been predominantly chosen for enzyme engineering, allowing for the rapid engineering of sequence features and the generation of predictable feature models *in silico*. Despite the advantages of sequence-based ML in acquiring a large amount of data for building regression models, numerical transformation of sequence information is still required for accurate enzyme engineering predictions (Bonetta and Valentino, 2020). Fox et al. advanced directed evolution by introducing a statistical analysis strategy called ProSAR (Protein Sequence-Activity Relationship), which assigns specific weights (regression coefficients) to positions based on a simple transformation of the absence (0) or presence (1) of substitution. These weights are determined through statistical analysis of the initial dataset and the impact of each substitution. The ProSAR method effectively identifies beneficial mutations even in mutants with reduced enzyme function. By employing ProSAR, the volumetric productivity of the cyanation process increased approximately 4000-fold in a library of aldol dehalogenase mutants (Fox et al., 2007). Liao et al. proposed a similar method considering amino acid mutation weighting and recombination effect summarization. They tested eight different linear regression algorithms on the dataset and, after two rounds of design, achieved a mutant with enzyme activity 20 times higher than the wildtype among 95 proteinase K mutants (Liao et al., 2007). However, linear models may not be as effective in describing higher-order epistasis. Zaugg et al. constructed a model using support vector regression (SVR) on 136 mutant data of epoxide hydrolase enantioselectivity. Exploring various SVR kernels and utilizing kernel tricks, they found that nonlinear models outperformed linear models in predicting protein function (Zaugg et al., 2017). Li et al. efficiently identified highly robust limonene epoxide hydrolase variants by identifying epimutation interactions through the innov'SAR (innovative sequence-activity relationship) algorithm. Using Fourier transform (FT) to capture nonlinear interactions between the order and position of protein sequences, they obtained variants with higher unfolding stability and tolerance to aggregation (Li et al., 2021b). Russ et al. described a method for generating new artificial sequences with protein family properties solely from evolutionary sequence data. Considering the conservation of amino acid positions and the relatedness of amino acid pairs in evolution, they experimentally demonstrated that the predicted new artificial sequences exhibit similar catalytic functions to chorismate mutase enzymes. This work suggests that evolution-based statistical models are sufficient to describe the extensive functional sequence space of a specific enzyme, laying the foundation for evolution-based artificial protease design (Russ et al., 2020).

Deep learning simulates the human brain for analysis and learning *via* establishing neural networks. It offers several advantages over traditional methods. Firstly, deep learning stands out for its capacity to process vast amounts of data, yielding more precise predictions. Secondly, deep learning eliminates the need for intricate feature engineering, requiring only sequence, structure, or energy data to be directly input into the neural network for effective predictive performance. This eradicates the challenges associated with extensive feature engineering in the modeling process. Lastly, adaptive deep learning can seamlessly transition across different types of datasets, and transfer learning enables pre-trained deep neural networks to be applied to various applications within the same domain. However, deep learning has its limitations. It necessitates very large datasets, and obtaining enzyme data for enzyme engineering is both expensive and time-consuming. In cases where the dataset is small, classical machine learning tends to outperform deep learning. As biological information databases rapidly expand, an increasing number of deep learning algorithms are being applied to the computational design of enzymes, including Variational Autoencoders (VAE) (Doersch, 2016; Hawkins-Hooker et al., 2021; Lobzaev et al., 2022), Recurrent Neural Network (RNN) (Alley et al., 2019; Hawkins-Hooker et al., 2021; Lipton et al., 2015), Generative Adversarial Network (GAN) (Repecka et al., 2021), Convolutional Neural Networks (CNN) (Lu et al., 2022) and Transformer (Dubourg-Felonneau et al., 2021; Vaswani et al., 2017). Alley et al. employed a RNN to acquire statistical characterizations of proteins using 24 million UniRef50 sequences. This approach condenses arbitrary protein sequences into fixed-length backbones, estimating essential protein properties independently of structural or evolutionary data (Alley et al., 2019). By leveraging the original protein sequence, this approach addressed data scarcity issues in protein informatics and showcased exceptional performance in crucial engineering tasks, encompassing stability, function, and design. UniRep, based on a relatively small sequence dataset, demonstrated effectiveness in constructing precise models representing a protein's fitness landscape. Even with a limited number of functionally characterized mutants, this unsupervised deep learning model facilitated large-scale exploration of sequence space and achieved protein design optimization comparable to mutants obtained in previous high-throughput studies (Biswas et al., 2021). Lu and coworkers employed a self-supervised convolutional neural network operating in three dimensions for training nearly twenty thousand protein structures. The model acquired knowledge of the local chemical microenvironment surrounding amino acids, enabling the prediction of

locations in proteins where wild-type amino acids were not present. Through a computer-synthesized mutation scan and a stepwise combination strategy, this approach generated mutants with significant catalytic improvement (Lu et al., 2022). A generative maximum entropy model was used to optimize the Renilla luciferase enzyme. Their work illustrates that leveraging natural evolutionary information enables predictive improvement of enzyme stability and activity through the design of active sites and protein scaffolds (Xie et al., 2023). A deep learning model EnzyKR was developed to predict biocatalysts with high enantioselectivity. In contrast to existing k_{cat} predictors, EnzyKR integrates substrate chirality by utilizing dihedral angles, geometric features, and atomic distance maps derived from hydrolase-substrate pairs. The capability of EnzyKR to efficiently identify enzymes with high stereoselectivity is attributed to its consideration of variations in k_{cat} values between different enantiomers, leading to distinct stereoselectivities (Ran et al., 2023). For creating novel proteins, Norn et al. reversed the supervised model used for protein structure prediction into a model based on structural design sequences (Norn et al., 2020). Anishchenko et al. utilized deep neural networks to optimize randomly given amino acid sequences, generating new proteins with diverse sequences and predicted structures (Anishchenko et al., 2021). These methods, based on inverting deep networks trained to predict the natural structures of proteins, offer insights into *de novo* protein design, complementing traditional physics-based models. Recently, Wang et al. introduced two deep learning approaches for creating proteins with prespecified functional sites. "Constrained hallucination" refers to a method that begins with a desired or predicted structure or functional site and optimizes sequences to match or contain these features. This approach imposes specific constraints or requirements during sequence optimization to ensure that certain features or properties are maintained or achieved. "Inpainting" starts from a known or predicted functional site and expands sequences or structures to fill in missing information or enhance specific elements using a RoseTTAFold network trained on the PDB database. It focuses on complementing or enhancing existing features, often without strict constraints on other parts of the sequence or structure. The authors successfully designed proteins with specific functions, including immunogens, receptor traps, metalloproteins, and enzymes, showcasing the capability to obtain desired protein scaffolds with functional sites (Wang et al., 2022). Using data-driven methods for enzyme engineering presents challenges and limitations. One primary issue is that the accuracy and robustness of these methods are limited by the quality and quantity of available data. To tackle this challenge, developing robust data-cleaning and data augmentation techniques can be beneficial. These strategies aim to remove errors, duplicates, and inconsistencies from the data, while also enhancing data diversity and representativeness, thus improving the overall reliability of data-driven approaches. Moreover, processing large-scale datasets and training complex models can be a time-consuming and resource-intensive task, which makes data-driven methods challenging for broad application. Developing parallel processing techniques and distributed computing infrastructure can improve the efficiency and scalability of data-driven approaches in enzyme engineering, making them more practical for real-world applications.

5. Conclusions and perspectives

In the review, we have delved into three major approaches to rational computational enzyme design, offering insights into their recent advancements. While each method provides a robust strategy for enzyme engineering, it is crucial to acknowledge the challenges that persist in the field. The foundation of structure-based computational design, which posits that protein structure dictates function, has led to significant progress. Through meticulous simulations and algorithmic analyses, this approach guides the targeted modification of natural enzymes for desired outcomes. Despite its success, challenges arise from its limited ability to explore the vast mutation sample space comprehensively. The nascent field of *de novo* design faces hurdles such as constrained conformational sampling, insufficient protein scaffold diversity, and the need for improved accuracy in transition state descriptions. Sequence-based rational computational design starts from protein amino acid sequences, employing methods like consensus design and ancestral sequence reconstruction to leverage nature's evolutionary skills for enzyme transformation. However, this approach may struggle to explain epistasis between mutations, especially those occurring at a distance. The bias towards natural substrates in evolutionary optimization poses limitations when aiming to improve the catalytic efficiency of non-natural substrates. To overcome these challenges, combining sequence and structural information is essential. The emergence of data-driven artificial intelligence and machine learning leverages vast amounts of sequence, structural, and energy data for efficient exploration of the combinatorial space of diverse enzyme sequences. Notably, deep learning methods, which can predict protein structure directly from sequence information, enhance retrofitting accuracy. However, challenges persist, particularly in *de novo* designing proteins with functional sites. Overcoming hurdles, such as embedding functional sites into designed protein scaffolds and predicting correctly folded amino acid sequences with functional sites, remains a priority. While data-driven machine learning methods hold promise for the future of enzyme design, current structure- and sequence-based methods remain crucial for elucidating catalytic mechanisms and rational enzyme design. Other challenges encountered in computational enzyme engineering have been thoroughly reviewed in recent articles (Kouba et al., 2023; Yang et al., 2023). In the future, the integration of machine learning with these traditional methods is vital to address these issues. This interdisciplinary approach aims to design biocatalysts with entirely new catalytic reactions, breaking through the bottleneck of limited reaction types and fostering advancements in synthetic biology. In conclusion, the synergistic application of these rational computational design approaches will likely lead to transformative breakthroughs in enzyme engineering, ultimately contributing to the flourishing fields of synthetic biology and metabolic engineering.

Declaration of competing interest

None.

Data availability

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

Acknowledgments

This work was funded by National Key Research and Development Program of China (2022YFA2106100) and National Natural Science Foundation of China (22378016, 22078011 and 22238001).

References

Acevedo-Rocha, C.G., Li, A., D'Amore, L., Hoebenreich, S., Sanchis, J., Lubrano, P., Ferla, M.P., Garcia-Borràs, M., Osuna, S., Reetz, M.T., 2021. Pervasive cooperative mutational effects on multiple catalytic enzyme traits emerge via long-range conformational dynamics. Nat. Commun. 12 (1), 1–13. https://doi.org/10.1038/s41467-021-21833-w.

Aerts, D., Verhaeghe, T., Joosten, H.J., Vriend, G., Soetaert, W., Desmet, T., 2013. Consensus engineering of sucrose phosphorylase: the outcome reflects the sequence input. Biotechnol. Bioeng. 110 (10), 2563–2572. https://doi.org/10.1002/bit.24940.

Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., Church, G.M., 2019. Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods 16 (12), 1315–1322. https://doi.org/10.1038/s41592-019-0598-1.

Almeida, T., Silvestre, A.J., Vilela, C., Freire, C.S., 2021. Bacterial nanocellulose toward green cosmetics: recent progresses and challenges. Int. J. Mol. Sci. 22 (6), 2836. https://doi.org/10.3390/ijms22062836.

AlQuraishi, M., 2021. Protein-structure prediction revolutionized. Nature 596 (7873), 487–488. https://doi.org/10.1038/d41586-021-02265-4.

Amrein, B.A., Steffen-Munsberg, F., Szeler, I., Purg, M., Kulkarni, Y., Kamerlin, S.C.L., 2017. CADEE: computer-aided directed evolution of enzymes. IUCrJ 4 (1), 50–64. https://doi.org/10.1107/S2052252516018017.

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

Babkova, P., Dunajova, Z., Chaloupkova, R., Damborsky, J., Bednar, D., Marek, M., 2020. Structures of hyperstable ancestral haloalkane dehalogenases show restricted conformational dynamics. Comput. Struct. Biotechnol. J. 18, 1497–1508. https://doi.org/10.1016/j.csbj.2020.06.021.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373 (6557), 871–876. https://doi.org/10.1126/science.abj8754.

Basanta, B., Bick, M.J., Bera, A.K., Norn, C., Chow, C.M., Carter, L.P., Goreshnik, I., Dimaio, F., Baker, D., 2020. An enumerative algorithm for *de novo* design of proteins with diverse pocket structures. Proc. Natl. Acad. Sci. USA 117 (36), 22135–22145. https://doi.org/10.1073/pnas.2005412117.

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

Bender, B.J., Cisneros, A., Duran, A.M., Finn, J.A., Fu, D., Lokits, A.D., Mueller, B.K., Sangha, A.K., Sauer, M.F., Sevy, A.M., Sliwoski, G., Sheehan, J.H., DiMaio, F., Meiler, J., Moretti, R., 2016. Protocols for molecular modeling with Rosetta3 and rosettascripts. Biochemistry 55 (34), 4748–4763. https://doi.org/10.1021/acs.biochem.6b00444.

Bhowmick, A., Sharma, S.C., Honma, H., Head-Gordon, T., 2016. The role of side chain entropy and mutual information for improving the *de novo* design of Kemp eliminases KE07 and KE70. PCCP 18 (28), 19386–19396. https://doi.org/10.1039/C6CP03622H.

Biswas, S., Khimulya, G., Alley, E.C., Esvelt, K.M., Church, G.M., 2021. Low-N protein engineering with data-efficient deep learning. Nat. Methods 18 (4), 389–396. https://doi.org/10.1038/s41592-021-01100-y.

Bjelic, S., Nivón, L.G., Çelebi-Ölçüm, N., Kiss, G., Rosewall, C.F., Lovick, H.M., Ingalls, E. L., Gallaher, J.L., Seetharaman, J., Lew, S., Montelione, G.T., Hunt, J.F., Michael, F. E., Houk, K.N., Baker, D., 2013. Computational design of enone-binding proteins with catalytic activity for the Morita-Baylis-Hillman reaction. ACS Chem. Biol. 8 (4), 749–757. https://doi.org/10.1021/cb3006227.

Blazeck, J., Karamitros, C.S., Ford, K., Somody, C., Qerqez, A., Murray, K., Burkholder, N. T., Marshall, N., Sivakumar, A., Lu, W.-C., 2022. Bypassing evolutionary dead ends and switching the rate-limiting step of a human immunotherapeutic enzyme. Nat. Catal. 5 (10), 952–967. https://doi.org/10.1038/s41929-022-00856-6.

Bonetta, R., Valentino, G., 2020. Machine learning techniques for protein function prediction. Proteins Struct. Funct. Bioinf. 88 (3), 397–413. https://doi.org/10.1002/prot.25832.

Broom, A., Rakotoharisoa, R.V., Thompson, M.C., Zarifi, N., Nguyen, E., Mukhametzhanov, N., Liu, L., Fraser, J.S., Chica, R.A., 2020. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution *in silico*. Nat. Commun. 11 (1), 1–10. https://doi.org/10.1038/s41467-020-18619-x.

Bu, Y., Cui, Y., Peng, Y., Hu, M., Tian, Y.E., Tao, Y., Wu, B., 2018. Engineering improved thermostability of the GH11 xylanase from Neocallimastix patriciarum via computational library design. Appl. Microbiol. Biotechnol. 102 (8), 3675–3685. https://doi.org/10.1007/s00253-018-8872-1.

Burke, A.J., Lovelock, S.L., Frese, A., Crawshaw, R., Ortmayer, M., Dunstan, M., Levy, C., Green, A.P., 2019. Design and evolution of an enzyme with a non-canonical organocatalytic mechanism. Nature 570 (7760), 219–223. https://doi.org/10.1038/s41586-019-1262-8.

Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G.V., Christie, C.H., Dalenberg, K., Di Costanzo, L., Duarte, J.M., Dutta, S., Feng, Z., Ganesan, S., Goodsell, D.S., Ghosh, S., Green, R.K., Guranović, V., Guzenko, D., Hudson, B.P., Lawson, Catherine L., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Persikova, I., Randle, C., Rose, A., Rose, Y., Sali, A., Segura, J., Sekharan, M., Shao, C., Tao, Y.-P., Voigt, M., Westbrook, John D., Young, J.Y., Zardecki, C., Zhuravleva, M., 2020. RCSB protein data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. 49 (D1), D437–D451. https://doi.org/10.1093/nar/gkaa1038.

Campbell, E.C., Correy, G.J., Mabbitt, P.D., Buckle, A.M., Tokuriki, N., Jackson, C.J., 2018. Laboratory evolution of protein conformational dynamics. Curr. Opin. Struct. Biol. 50, 49–57. https://doi.org/10.1016/j.sbi.2017.09.005.

Campitelli, P., Modi, T., Kumar, S., Ozkan, S.B., 2020. The role of conformational dynamics and allostery in modulating protein evolution. Annu. Rev. Biophys. 49, 267–288. https://doi.org/10.1146/annurev-biophys-052118-115517.

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. USA 90 (12), 5618–5622. https://doi.org/10.1073/pnas.90.12.5618.

Cole, M.F., Gaucher, E.A., 2011. Exploiting models of molecular evolution to efficiently direct protein engineering. J. Mol. Evol. 72 (2), 193–203. https://doi.org/10.1007/s00239-010-9415-2.

Consortium, T.U., 2020. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49 (D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100.

Copley, S.D., 2021. Setting the stage for evolution of a new enzyme. Curr. Opin. Struct. Biol. 69, 41–49. https://doi.org/10.1016/j.sbi.2021.03.001.

Costa-Silva, T., Carvalho, A., Souza, C., De Castro, H., Bachmann, L., Said, S., Oliveira, W., 2021. Immobilized enzyme-driven value enhancement of lignocellulosic-based agricultural byproducts: application in aroma synthesis. J. Clean. Prod. 284, 124728 https://doi.org/10.1016/j.jclepro.2020.124728.

Cozzetto, D., Buchan, D.W., Bryson, K., Jones, D.T., 2013. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. BMC Bioinf. 14 (3), 1–11. https://doi.org/10.1186/1471-2105-14-S3-S1.

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. 14 (3), 313–320. https://doi.org/10.1038/s41557-021-00833-9.

Crean, R.M., Gardner, J.M., Kamerlin, S.C., 2020. Harnessing conformational plasticity to generate designer enzymes. J. Am. Chem. Soc. 142 (26), 11324–11342. https://doi.org/10.1021/jacs.0c04924.

Cui, Y., Wang, Y., Tian, W., Bu, Y., Li, T., Cui, X., Zhu, T., Li, R., Wu, B., 2021. Development of a versatile and efficient C-N lyase platform for asymmetric hydroamination via computational enzyme redesign. Nat. Catal. 4 (5), 364–373. https://doi.org/10.1038/s41929-021-00604-2.

Cui, Y., Sun, J., Wu, B., 2022. Computational enzyme redesign: large jumps in function. Trends Chem. 1216, 265–273. https://doi.org/10.1007/978-1-4939-1486-9_14.

Curado-Carballada, C., Feixas, F., Iglesias-Fernández, J., Osuna, S., 2019. Hidden conformations in *aspergillus Niger* monoamine oxidase are key for catalytic efficiency. Angew. Chem. Int. Ed. 58 (10), 3097–3101. https://doi.org/10.1002/anie.201812532.

Dagan, S., Hagai, T., Gavrilov, Y., Kapon, R., Levy, Y., Reich, Z., 2013. Stabilization of a protein conferred by an increase in folded state entropy. Proc. Natl. Acad. Sci. USA 110 (26), 10628–10633. https://doi.org/10.1073/pnas.1302284110.

Damnjanović, J., Nakano, H., Iwasaki, Y., 2014. Deletion of a dynamic surface loop improves stability and changes kinetic behavior of phosphatidylinositol-synthesizing *Streptomyces* phospholipase D. Biotechnol. Bioeng. 111 (4), 674–682. https://doi.org/10.1002/bit.25149.

De Ferrari, L., Mitchell, J.B., 2014. From sequence to enzyme mechanism using multi-label machine learning. BMC Bioinf. 15 (1), 1–13. https://doi.org/10.1186/1471-2105-15-150.

Dinmukhamed, T., Huang, Z., Liu, Y., Lv, X., Li, J., Du, G., Liu, L., 2021. Current advances in design and engineering strategies of industrial enzymes. Syst. Microbiol. Biomanuf. 1 (1), 15–23. https://doi.org/10.1007/s43393-020-00005-9.

Doersch, C., 2016. Tutorial on variational autoencoders arXiv preprint. https://arxiv.org/abs/1606.05908.

Dubourg-Felonneau, G., Shams, S., Akiva, E., Lee, L., 2021. Protein organization with manifold exploration and spectral clustering. bioRxiv. https://doi.org/10.1101/2021.12.08.471858.

Falda, M., Toppo, S., Pescarolo, A., Lavezzo, E., Di Camillo, B., Facchinetti, A., Cilia, E., Velasco, R., Fontana, P., 2012. Argot2: a large scale function prediction tool relying on semantic similarity of weighted gene ontology terms. BMC Bioinf. 13 (4), 1–9. https://doi.org/10.1186/1471-2105-13-S4-S14.

Feehan, R., Franklin, M.W., Slusky, J.S., 2021. Machine learning differentiates enzymatic and non-enzymatic metals in proteins. Nat. Commun. 12 (1), 1–11. https://doi.org/10.1038/s41467-021-24070-3.

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.

Flynn, B.T., Kozak, S.M., Lawton, M.R., Alcaine, S.D., 2021. Lactose oxidase: an enzymatic approach to inhibit *listeria monocytogenes* in milk. J. Dairy Sci. 104 (10), 10594–10608. https://doi.org/10.3168/jds.2021-20450.

Folkman, L., Stantic, B., Sattar, A., Zhou, Y., 2016. EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. J. Mol. Biol. 428 (6), 1394–1405. https://doi.org/10.1016/j.jmb.2016.01.012.

Foroozandeh Shahraki, M., Farhadyar, K., Kavousi, K., Azarabad, M.H., Boroomand, A., Ariaeenejad, S., Hosseini Salekdeh, G., 2021. A generalized machine-learning aided method for targeted identification of industrial enzymes from metagenome: A xylanase temperature dependence case study. Biotechnol. Bioeng. 118 (2), 759–769. https://doi.org/10.1002/bit.27608.

Fournière, M., Bedoux, G., Lebonvallet, N., Leschiera, R., Le Goff-Pain, C., Bourgougnon, N., Latire, T., 2021. Poly-and oligosaccharide ulva sp. fractions from enzyme-assisted extraction modulate the metabolism of extracellular matrix in human skin fibroblasts: potential in anti-aging dermo-cosmetic applications. Mar. Drugs 19 (3), 156. https://doi.org/10.3390/md19030156.

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., 2007. Improving catalytic function by ProSAR-driven enzyme evolution. Nat. Biotechnol. 25 (3), 338–344. https://doi.org/10.1038/nbt1286.

Fürst, M.J., Boonstra, M., Bandstra, S., Fraaije, M.W., 2019. Stabilization of cyclohexanone monooxygenase by computational and experimental library design. Biotechnol. Bioeng. 116 (9), 2167–2177. https://doi.org/10.1002/bit.27022.

Galmés, M.À., Nödling, A.R., He, K., Luk, L.Y., Świderek, K., Moliner, V., 2022. Computational design of an amidase by combining the best electrostatic features of two promiscuous hydrolases. Chem. Sci. 13 (17), 4779–4787. https://doi.org/10.1039/d2sc00778a.

Gardner, J.M., Biler, M., Risso, V.A., Sanchez-Ruiz, J.M., Kamerlin, S.C., 2020. Manipulating conformational dynamics to repurpose ancient proteins for modern catalytic functions. ACS Catal. 10 (9), 4863–4870. https://doi.org/10.1021/acscatal.0c00722.

Gargiulo, S., Soumillion, P., 2021. Directed evolution for enzyme development in biocatalysis. Curr. Opin. Chem. Biol. 61, 107–113. https://doi.org/10.1016/j.cbpa.2020.11.006.
Gomez de Santos, P., Mateljak, I., Hoang, M.D., Fleishman, S.J., Hollmann, F., Alcalde, M., 2023. Repertoire of computationally designed peroxygenases for enantiodivergent C-H oxyfunctionalization reactions. J. Am. Chem. Soc. 145 (6), 3443–3453. https://doi.org/10.1021/acscatal.2c03006.
Greenhalgh, J.C., Fahlberg, S.A., Pfleger, B.F., Romero, P.A., 2021. Machine learning-guided acyl-ACP reductase engineering for improved in vivo fatty alcohol production. Nat. Commun. 12 (1), 1–10. https://doi.org/10.1038/s41467-021-25831-w.
Gumulya, Y., Baek, J.-M., Wun, S.-J., Thomson, R.E., Harris, K.L., Hunter, D.J., Behrendorff, J.B., Kulig, J., Zheng, S., Wu, X., 2018. Engineering highly functional thermostable proteins using ancestral sequence reconstruction. Nat. Catal. 1 (11), 878–888. https://doi.org/10.1038/s41929-018-0159-5.
Han, B., Hou, Y., Jiang, T., Lv, B., Zhao, L., Feng, X., Li, C., 2018. Computation-aided rational deletion of C-terminal region improved the stability, activity, and expression level of GH2 β-glucuronidase. J. Agric. Food Chem. 66 (43), 11380–11389. https://doi.org/10.1021/acs.jafc.8b03449.
Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., Bikard, D., 2021. Generating functional protein variants with variational autoencoders. PLoS Comput. Biol. 17 (2), e1008736 https://doi.org/10.1371/journal.pcbi.1008736.
Hochberg, G.K., Thornton, J.W., 2017. Reconstructing ancient proteins to understand the causes of structure and function. Annu. Rev. Biophys. 46, 247–269. https://doi.org/10.1146/annurev-biophys-070816-033631.
Hollmann, F., Fernandez-Lafuente, R., 2021. Grand challenges in biocatalysis. Front. Catal. 1, 633893 https://doi.org/10.3389/fctls.2021.633893.
Huang, L.-T., Gromiha, M.M., Ho, S.-Y., 2007. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics 23 (10), 1292–1293. https://doi.org/10.1093/bioinformatics/btm100.
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 (7897), 523–528. https://doi.org/10.1038/s41586-021-04383-5.
James, L.C., Tawfik, D.S., 2003. Conformational diversity and protein evolution-a 60-year-old hypothesis revisited. Trends Biochem. Sci. 28 (7), 361–368. https://doi.org/10.1016/S0968-0004(03)00135-X.
Jemth, P., Karlsson, E., Vögeli, B., Guzovsky, B., Andersson, E., Hultqvist, G., Dogan, J., Güntert, P., Riek, R., Chi, C.N., 2018. Structure and dynamics conspire in the evolution of affinity between intrinsically disordered proteins. Sci. Adv. 4 (10), eaau4130. https://doi.org/10.1126/sciadv.aau4130.
Jiang, L., Althoff, E.A., Clemente, F.R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J.L., Betker, J.L., Tanaka, F., Barbas III, C.F., 2008. *De novo* computational design of retro-aldol enzymes. Science 319 (5868), 1387–1391. https://doi.org/10.1126/science.1152692.
Jiang, Y., Ding, N., Shao, Q., Stull, S.L., Cheng, Z., Yang, Z.J., 2023. Substrate positioning dynamics involves a non-electrostatic component to mediate catalysis. J. Phys. Chem. Lett. 14 (50), 11480–11489. https://doi.org/10.1021/acs.jpclett.3c02444.
Jones, B.S., Ross, C.M., Foley, G., Pozhydaieva, N., Sharratt, J.W., Kress, N., Seibt, L.S., Thomson, R.E., Gumulya, Y., Hayes, M.A., 2024. Engineering biocatalysts for the C-H activation of fatty acids by ancestral sequence reconstruction. Angew. Chem. Int. Ed. 63 (18), e202314869 https://doi.org/10.1002/anie.202314869.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.
Kaczmarski, J.A., Mahawaththa, M.C., Feintuch, A., Clifton, B.E., Adams, L.A., Goldfarb, D., Otting, G., Jackson, C.J., 2020. Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme. Nat. Commun. 11 (1), 1–14. https://doi.org/10.1038/s41467-020-19695-9.
Khersonsky, O., Röthlisberger, D., Wollacott, A.M., Murphy, P., Dym, O., Albeck, S., Kiss, G., Houk, K., Baker, D., Tawfik, D.S., 2011. Optimization of the in-silico-designed Kemp eliminase KE70 by computational design and directed evolution. J. Mol. Biol. 407 (3), 391–412. https://doi.org/10.1016/j.jmb.2011.01.041.
Khersonsky, O., Kiss, G., Röthlisberger, D., Dym, O., Albeck, S., Houk, K.N., Baker, D., Tawfik, D.S., 2012. Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. Proc. Natl. Acad. Sci. USA 109 (26), 10358–10363. https://doi.org/10.1073/pnas.1121063109.
Khersonsky, O., Lipsh, R., Avizemer, Z., Ashani, Y., Goldsmith, M., Leader, H., Dym, O., Rogotner, S., Trudeau, D.L., Prilusky, J., 2018. Automated design of efficient and functionally diverse enzyme repertoires. Mol. Cell 72 (1), 178–186. https://doi.org/10.1016/j.molcel.2018.08.033.
Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D., Houk, K., 2013. Computational enzyme design. Angew. Chem. Int. Ed. 52 (22), 5700–5725. https://doi.org/10.1002/anie.201204077.
Koskinen, P., Törönen, P., Nokso-Koivisto, J., Holm, L., 2015. PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. Bioinformatics 31 (10), 1544–1552. https://doi.org/10.1093/bioinformatics/btu851.
Kouba, P., Kohout, P., Haddadi, F., Bushuiev, A., Samusevich, R., Sedlar, J., Damborsky, J., Pluskal, T., Sivic, J., Mazurenko, S., 2023. Machine learning-guided protein engineering. ACS Catal. 13 (21), 13863–13895. https://doi.org/10.1021/acscatal.3c02743.
Krivák, R., Hoksza, D., 2018. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J. Chemother. 10 (1), 1–12. https://doi.org/10.1186/s13321-018-0285-8.
Kumar, R., Prabhu, N.P., Yadaiah, M., Bhuyan, A.K., 2004. Protein stiffening and entropic stabilization in the subdenaturing limit of guanidine hydrochloride. Biophys. J. 87 (4), 2656–2662. https://doi.org/10.1529/biophysj.104.044701.
Lehmann, M., Kostrewa, D., Wyss, M., Brugger, R., D'Arcy, A., Pasamontes, L., van Loon, A.P., 2000. From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. Protein Eng. 13 (1), 49–57. https://doi.org/10.1093/protein/13.1.49.
Li, G., Fang, X., Su, F., Chen, Y., Xu, L., Yan, Y., 2018a. Enhancing the thermostability of *Rhizomucor miehei* lipase with a limited screening library by rational-design point mutations and disulfide bonds. Appl. Environ. Microbiol. 84 (2) https://doi.org/10.1128/AEM.02129-17 e02129–02117.
Li, R., Wijma, H.J., Song, L., Cui, Y., Otzen, M., Tian, Y.E., Du, J., Li, T., Niu, D., Chen, Y., 2018b. Computational redesign of enzymes for regio-and enantioselective hydroamination. Nat. Chem. Biol. 14 (7), 664–670. https://doi.org/10.1038/s41589-018-0053-0.
Li, D., Han, T., Xue, J., Xu, W., Xu, J., Wu, Q., 2021a. Engineering fatty acid photodecarboxylase to enable highly selective decarboxylation of trans fatty acids. Angew. Chem. Int. Ed. 60 (38), 20695–20699. https://doi.org/10.1002/anie.202107694.
Li, G., Qin, Y., Fontaine, N.T., Ng Fuk Chong, M., Maria-Solano, M.A., Feixas, F., Cadet, X.F., Pandjaitan, R., Garcia-Borràs, M., Cadet, F., Reetz, M.T., 2021b. Machine learning enables selection of epistatic enzyme mutants for stability against unfolding and detrimental aggregation. ChemBioChem 22 (5), 904–914. https://doi.org/10.1002/cbic.202000612.
Li, J., Qu, G., Shang, N., Chen, P., Men, Y., Liu, W., Mei, Z., Sun, Y., Sun, Z., 2021c. Near-perfect control of the regioselective glucosylation enabled by rational design of glycosyltransferases. Green Synth. Catal. 2 (1), 45–53. https://doi.org/10.1016/j.gresc.2021.01.005.
Li, F., Yuan, L., Lu, H., Li, G., Chen, Y., Engqvist, M.K., Kerkhoven, E.J., Nielsen, J., 2022. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. Nat. Catal. 5, 662–672. https://doi.org/10.1038/s41929-022-00798-z.
Liao, J., Warmuth, M.K., Govindarajan, S., Ness, J.E., Wang, R.P., Gustafsson, C., Minshull, J., 2007. Engineering proteinase K using machine learning and synthetic genes. BMC Biotechnol. 7 (16), 1–19. https://doi.org/10.1186/1472-6750-7-16.
Lin, H.-Y., Chen, X., Dong, J., Yang, J.-F., Xiao, H., Ye, Y., Li, L.-H., Zhan, C.-G., Yang, W.-C., Yang, G.-F., 2021. Rational redesign of enzyme via the combination of quantum mechanics/molecular mechanics, molecular dynamics, and structural biology study. J. Am. Chem. Soc. 143 (38), 15674–15687. https://doi.org/10.1021/jacs.1c06227.
Lin, C.L., Petersen, M.A., Mauch, A., Gottlieb, A., 2022. Towards lager beer aroma improvement via selective amino acid release by proteases during mashing. J. Inst. Brew. 128 (1), 15–21. https://doi.org/10.1002/jib.682.
Lipton, Z.C., Berkowitz, J., Elkan, C., 2015. A critical review of recurrent neural networks for sequence learning. arXiv preprint. https://doi.org/10.48550/arXiv.1506.00019.
Lobzaev, E., Herrera, M.A., Campopiano, D.J., Stracquadanio, G., 2022. Designing human Sphingosine-1-phosphate lyases using a temporal Dirichlet variational autoencoder. bioRxiv. https://doi.org/10.1101/2022.02.14.480330.
Löffler, P., Schmitz, S., Hupfeld, E., Sterner, R., Merkl, R., 2017. Rosetta: MSF: a modular framework for multi-state computational protein design. PLoS Comput. Biol. 13 (6), e1005600 https://doi.org/10.1371/journal.pcbi.1005600.
Lu, Y., Yeung, N., Sieracki, N., Marshall, N.M., 2009. Design of functional metalloproteins. Nature 460 (7257), 855–862. https://doi.org/10.1038/nature08304.
Lu, H., Diaz, D.J., Czarnecki, N.J., Zhu, C., Kim, W., Shroff, R., Acosta, D.J., Alexander, B. R., Cole, H.O., Zhang, Y., 2022. Machine learning-aided engineering of hydrolases for PET depolymerization. Nature 604 (7907), 662–667. https://doi.org/10.1038/s41586-022-04599-z.
Magliery, T.J., 2015. Protein stability: computation, sequence statistics, and new experimental methods. Curr. Opin. Struct. Biol. 33, 161–168. https://doi.org/10.1016/j.sbi.2015.09.002.
Maria-Solano, M.A., Serrano-Hervás, E., Romero-Rivera, A., Iglesias-Fernández, J., Osuna, S., 2018. Role of conformational dynamics in the evolution of novel enzyme function. Chem. Commun. 54 (50), 6622–6634. https://doi.org/10.1039/c8cc02426j.
Maria-Solano, M.A., Iglesias-Fernández, J., Osuna, S., 2019. Deciphering the allosterically driven conformational ensemble in tryptophan synthase evolution. J. Am. Chem. Soc. 141 (33), 13049–13056. https://doi.org/10.1021/jacs.9b03646.
Maria-Solano, M.Á., Kinateder, T., Iglesias-Fernández, J., Sterner, R., Osuna, S., 2021. *In silico* identification and experimental validation of distal activity-enhancing mutations in tryptophan synthase. ACS Catal. 11 (21), 13733–13743. https://doi.org/10.1021/acscatal.1c03950.
Meghwanshi, G.K., Kaur, N., Verma, S., Dabi, N.K., Vashishtha, A., Charan, P.D., Purohit, P., Bhandari, H.S., Bhojak, N., Kumar, R., 2020. Enzymes for pharmaceutical and therapeutic applications. Biotechnol. Appl. Biochem. 67 (4), 586–601. https://doi.org/10.1002/bab.1919.
Meng, Q., Capra, N., Palacio, C.M., Lanfranchi, E., Otzen, M., Van Schie, L.Z., Rozeboom, H.T.J., Thunnissen, A.-M.W., Wijma, H.J., Janssen, D.B., 2020. Robust ω-transaminases by computational stabilization of the subunit interface. ACS Catal. 10 (5), 2915–2928. https://doi.org/10.1021/acscatal.9b05223.
Mori, Y., Noda, S., Shirai, T., Kondo, A., 2021. Direct 1, 3-butadiene biosynthesis in *Escherichia coli* via a tailored ferulic acid decarboxylase mutant. Nat. Commun. 12 (1), 1–12. https://doi.org/10.1038/s41467-021-22504-6.

Musil, M., Konegger, H., Hon, J., Bednar, D., Damborsky, J., 2018. Computational design of stable and soluble biocatalysts. ACS Catal. 9 (2), 1033–1054. https://doi.org/10.1021/acscatal.8b03613.

Nagi, A.D., Regan, L., 1997. An inverse correlation between loop length and stability in a four-helix-bundle protein. Fold. Des. 2 (1), 67–75. https://doi.org/10.1016/S1359-0278(97)00007-2.

Nanda, V., Koder, R.L., 2010. Designing artificial enzymes by intuition and computation. Nat. Chem. 2 (1), 15–24. https://doi.org/10.1038/nchem.473.

Norn, C., Wicky, B.I., Juergens, D., Liu, S., Kim, D., Koepnick, B., Anishchenko, I., Baker, D., Ovchinnikov, S., 2020. Protein sequence design by explicit energy landscape optimization. bioRxiv. https://doi.org/10.1101/2020.07.23.218917.

Obexer, R., Godina, A., Garrabou, X., Mittl, P.R., Baker, D., Griffiths, A.D., Hilvert, D., 2017. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. Nat. Chem. 9 (1), 50–56. https://doi.org/10.1038/nchem.2596.

O'Reilly, E., 2022. Building enzymes from scratch. Nat. Chem. 14 (3), 246–248. https://doi.org/10.1038/s41557-021-00884-y.

Osuna, S., 2021. The challenge of predicting distal active site mutations in computational enzyme design. Wiley Interdiscip. Rev. Comput. Mol. Sci. 11, e1502 https://doi.org/10.1002/wcms.1502.

Pan, X., Yang, J., Van, R., Epifanovsky, E., Ho, J., Huang, J., Pu, J., Mei, Y., Nam, K., Shao, Y., 2021. Machine-learning-assisted free energy simulation of solution-phase and enzyme reactions. J. Chem. Theory Comput. 17 (9), 5745–5758. https://doi.org/10.1021/acs.jctc.1c00565.

Park, S.Y., Hwang, I.S., Lee, H.J., Song, C.E., 2017. Biomimetic catalytic transformation of toxic alpha-oxoaldehydes to high-value chiral alpha-hydroxythioesters using artificial glyoxalase I. Nat. Commun. 8, 14877. https://doi.org/10.1038/ncomms14877.

Parthasarathy, S., Murthy, M.R.N., 2000. Protein thermal stability: insights from atomic displacement parameters (B values). Protein Eng. Des. Sel. 13 (1), 9–13. https://doi.org/10.1093/protein/13.1.9.

Pavelka, A., Chovancova, E., Damborsky, J., 2009. HotSpot wizard: a web server for identification of hot spots in protein engineering. Nucleic Acids Res. 37 (suppl_2), W376–W383. https://doi.org/10.1093/nar/gkp410.

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.

Punia, S., 2020. Barley starch modifications: physical, chemical and enzymatic-A review. Int. J. Biol. Macromol. 144, 578–585. https://doi.org/10.1016/j.ijbiomac.2019.12.088.

Ran, X., Jiang, Y., Shao, Q., Yang, Z.J., 2023. EnzyKR: a chirality-aware deep learning model for predicting the outcomes of the hydrolase-catalyzed kinetic resolution. Chem. Sci. 14 (43), 12073–12082. https://doi.org/10.1039/D3SC02752J.

Reich, S., Kress, N., Nestl, B.M., Hauer, B., 2014. Variations in the stability of NCR ene reductase by rational enzyme loop modulation. J. Struct. Biol. 185 (2), 228–233. https://doi.org/10.1016/j.jsb.2013.04.004.

Rentschler, E., Schuh, K., Krewinkel, M., Baur, C., Claassen, W., Meyer, S., Kuschel, B., Stressler, T., Fischer, L., 2015. Enzymatic production of lactulose and epilactose in milk. J. Dairy Sci. 98 (10), 6767–6775. https://doi.org/10.3168/jds.2015-9900.

Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Poviloniene, S., Laurynenas, A., Viknander, S., Abuajwa, W., 2021. Expanding functional protein sequence spaces using generative adversarial networks. Nat. Mach. Intell. 3 (4), 324–333. https://doi.org/10.3934/mbe.2023195.

Richter, F., Leaver-Fay, A., Khare, S.D., Bjelic, S., Baker, D., 2011. *De novo* enzyme design using Rosetta3. PLoS One 6 (5), e19230. https://doi.org/10.1371/journal.pone.0019230.

Richter, F., Blomberg, R., Khare, S.D., Kiss, G., Kuzin, A.P., Smith, A.J.T., Gallaher, J., Pianowski, Z., Helgeson, R.C., Grjasnow, A., Xiao, R., Seetharaman, J., Su, M., Vorobiev, S., Lew, S., Forouhar, F., Kornhaber, G.J., Hunt, J.F., Montelione, G.T., Tong, L., Houk, K.N., Hilvert, D., Baker, D., 2012. Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. J. Am. Chem. Soc. 134 (39), 16197–16206. https://doi.org/10.1021/ja3037367.

Risso, V.A., Martinez-Rodriguez, S., Candel, A.M., Krüger, D.M., Pantoja-Uceda, D., Ortega-Muñoz, M., Santoyo-Gonzalez, F., Gaucher, E.A., Kamerlin, S.C.L., Bruix, M., Gavira, J.A., Sanchez-Ruiz, J.M., 2017. *De novo* active sites for resurrected precambrian enzymes. Nat. Commun. 8, 16113. https://doi.org/10.1038/ncomms16113.

Romero-Rivera, A., Garcia-Borras, M., Osuna, S., 2017. Role of conformational dynamics in the evolution of retro-aldolase activity. ACS Catal. 7 (12), 8524–8532. https://doi.org/10.1021/acscatal.7b02954.

Rosenthal, K., Lütz, S., 2018. Recent developments and challenges of biocatalytic processes in the pharmaceutical industry. Curr. Opin. Green Sustainable Chem. 11, 58–64. https://doi.org/10.1016/j.cogsc.2018.03.015.

Röthlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O., 2008. Kemp elimination catalysts by computational enzyme design. Nature 453 (7192), 190–195. https://doi.org/10.1038/nature06879.

Ruscio, J.Z., Kohn, J.E., Ball, K.A., Head-Gordon, T., 2009. The influence of protein dynamics on the success of computational enzyme design. J. Am. Chem. Soc. 131 (39), 14111–14115. https://doi.org/10.1021/ja905396s.

Russ, W.P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., 2020. An evolution-based model for designing chorismate mutase enzymes. Science 369 (6502), 440–445. https://doi.org/10.1126/science.aba3304.

Selberg, A.G., Gaucher, E.A., Liberles, D.A., 2021. Ancestral sequence reconstruction: from chemical paleogenetics to maximum likelihood algorithms and beyond. J. Mol. Evol. 89 (3), 157–164. https://doi.org/10.1007/s00239-021-09993-1.

Shao, Q., Jiang, Y., Yang, Z.J., 2022. EnzyHTP: a high-throughput computational platform for enzyme modeling. J. Chem. Inf. Model. 62 (3), 647–655. https://doi.org/10.1021/acs.jcim.1c01424.

Shao, Q., Jiang, Y., Yang, Z.J., 2023. EnzyHTP computational directed evolution with adaptive resource allocation. J. Chem. Inf. Model. 63 (17), 5650–5659. https://doi.org/10.1021/acs.jcim.3c00618.

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., 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. USA 112 (12), 3704–3709. https://doi.org/10.1073/pnas.1500545112.

Sijinamanoj, V., Muthukumar, T., Muthuraja, R., Rayappan, K., Karmegam, N., Saminathan, K., Govarthanan, M., Kathireswari, P., 2021. Ligninolytic valorization of agricultural residues by *aspergillus nomius* and *Trichoderma harzianum* isolated from gut and comb of *Odontotermes obesus* (Termitidae). Chemosphere 284, 131384. https://doi.org/10.1016/j.chemosphere.2021.131384.

Spence, M.A., Kaczmarski, J.A., Saunders, J.W., Jackson, C.J., 2021. Ancestral sequence reconstruction for protein engineers. Curr. Opin. Struct. Biol. 69, 131–141. https://doi.org/10.1016/j.sbi.2021.04.001.

Steipe, B., Schiller, B., Plückthun, A., Steinbacher, S., 1994. Sequence statistics reliably predict stabilizing mutations in a protein domain. J. Mol. Biol. 240 (3), 188–192. https://doi.org/10.1006/jmbi.1994.1434.

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., Robinson, G.E., 2015. Big data: astronomical or genomical? PLoS Biol. 13 (7), e1002195 https://doi.org/10.1371/journal.pbio.1002195.

St-Jacques, A.D., Rodriguez, J.M., Eason, M.G., Foster, S.M., Khan, S.T., Damry, A.M., Goto, N.K., Thompson, M.C., Chica, R.A., 2023. Computational remodeling of an enzyme conformational landscape for altered substrate selectivity. Nat. Commun. 14 (1), 6058. https://doi.org/10.1038/s41467-023-41762-0.

Stourac, J., Vavra, O., Kokkonen, P., Filipovic, J., Pinto, G., Brezovsky, J., Damborsky, J., Bednar, D., 2019. Caver web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. Nucleic Acids Res. 47 (W1), W414–W422. https://doi.org/10.1093/nar/gkz378.

Su, B.-M., Shao, Z.-H., Li, A.-P., Naeem, M., Lin, J., Ye, L.-D., Yu, H.-W., 2020. Rational design of dehydrogenase/reductases based on comparative structural analysis of prereaction-state and free-state simulations for efficient asymmetric reduction of bulky aryl ketones. ACS Catal. 10 (1), 864–876. https://doi.org/10.1021/acscatal.9b04778.

Sullivan, B.J., Nguyen, T., Durani, V., Mathur, D., Rojas, S., Thomas, M., Syu, T., Magliery, T.J., 2012. Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. J. Mol. Biol. 420 (4–5), 384–399. https://doi.org/10.1016/j.jmb.2012.04.025.

Teng, S., Srivastava, A.K., Wang, L., 2010. Sequence feature-based prediction of protein stability changes upon amino acid substitutions. BMC Genomics 11 (2), 1–8. https://doi.org/10.1186/1471-2164-11-S2-S5.

Teze, D., Zhao, J., Wiemann, M., Kazi, Z.G.A., Lupo, R., Zeuner, B., Vuillemin, M., Rønne, M.E., Carlström, G., Duus, J.Ø., Sanejouand, Y.-H., O'Donohue, M.J., Nordberg Karlsson, E., Fauré, R., Stålbrand, H., Svensson, B., 2021. Rational enzyme design without structural knowledge: a sequence-based approach for efficient generation of transglycosylases. Chemistry 27 (40), 10323–10334. https://doi.org/10.1002/chem.202100110.

Tingley, J.P., Low, K.E., Xing, X., Abbott, D.W., 2021. Combined whole cell wall analysis and streamlined *in silico* carbohydrate-active enzyme discovery to improve biocatalytic conversion of agricultural crop residues. Biotechnol. Biofuels 14 (1), 1–19. https://doi.org/10.1186/s13068-020-01869-8.

Tyka, M.D., Keedy, D.A., André, I., DiMaio, F., Song, Y., Richardson, D.C., Richardson, J. S., Baker, D., 2011. Alternate states of proteins revealed by detailed energy landscape mapping. J. Mol. Biol. 405 (2), 607–618. https://doi.org/10.1016/j.jmb.2010.11.008.

Tyka, M.D., Jung, K., Baker, D., 2012. Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers. J. Comput. Chem. 33 (31), 2483–2491. https://doi.org/10.1002/jcc.23069.

Unger, E.K., Keller, J.P., Altermatt, M., Liang, R., Matsui, A., Dong, C., Hon, O.J., Yao, Z., Sun, J., Banala, S., 2020. Directed evolution of a selective and sensitive serotonin sensor via machine learning. Cell 183 (7), 1986–2002.e1926. https://doi.org/10.1016/j.cell.2020.11.040.

Vaissier, V., Sharma, S.C., Schaettle, K., Zhang, T., Head-Gordon, T., 2018. Computational optimization of electric fields for improving catalysis of a designed Kemp eliminase. ACS Catal. 8 (1), 219–227. https://doi.org/10.1021/acscatal.7b03151.

van Donkelaar, L.H.G., Mostert, J., Zisopoulos, F.K., Boom, R.M., van der Goot, A.-J., 2016. The use of enzymes for beer brewing: thermodynamic comparison on resource use. Energy 115, 519–527. https://doi.org/10.1016/j.energy.2016.09.011.

Vasina, M., Velecký, J., Planas-Iglesias, J., Marques, S.M., Skarupova, J., Damborsky, J., Bednar, D., Mazurenko, S., Prokop, Z., 2022. Tools for computational design and high-throughput screening of therapeutic enzymes. Adv. Drug Deliv. Rev. 114143 https://doi.org/10.1016/j.addr.2022.114143.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Proces. Syst. 30, 5998–6008. In: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vázquez-Figueroa, E., Chaparro-Riggers, J., Bommarius, A.S., 2007. Development of a thermostable glucose dehydrogenase by a structure-guided consensus concept. ChemBioChem 8 (18), 2295–2301. https://doi.org/10.1002/cbic.200700500.

Vieille, C., Zeikus, J.G., 1996. Thermozymes: identifying molecular determinants of protein structural and functional stability. Trends Biotechnol. 14 (6), 183–190. https://doi.org/10.1016/0167-7799(96)10026-3.

Völler, J.-S., 2020. Ensemble-based enzyme design. Nat. Catal. 3 (10), 774. https://doi.org/10.1038/s41929-020-00529-2.

Wang, Y., Xu, M., Yang, T., Zhang, X., Rao, Z., 2020. Surface charge-based rational design of aspartase modifies the optimal pH for efficient β-aminobutyric acid production. Int. J. Biol. Macromol. 164, 4165–4172. https://doi.org/10.1016/j.ijbiomac.2020.08.229.

Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J.L., Castro, K.M., Ragotte, R., Saragovi, A., Milles, L.F., Baek, M., Anishchenko, I., Yang, W., Hicks, D.R., Expòsit, M., Schlichthaerle, T., Chun, J.-H., Dauparas, J., Bennett, N., Wicky, B.I.M., Muenks, A., DiMaio, F., Correia, B., Ovchinnikov, S., Baker, D., 2022. Scaffolding protein functional sites using deep learning. Science 377 (6604), 387–394. https://doi.org/10.1126/science.abn2100.

Warshel, A., Sharma, P.K., Kato, M., Xiang, Y., Liu, H., Olsson, M.H., 2006. Electrostatic basis for enzyme catalysis. Chem. Rev. 106 (8), 3210–3235. https://doi.org/10.1021/cr0503106.

Weitzner, B.D., Kipnis, Y., Daniel, A.G., Hilvert, D., Baker, D., 2019. A computational method for design of connected catalytic networks in proteins. Protein Sci. 28 (12), 2036–2041. https://doi.org/10.1002/pro.3757.

Wijma, H.J., Janssen, D.B., 2013. Computational design gains momentum in enzyme catalysis engineering. FEBS J. 280 (13), 2948–2960. https://doi.org/10.1111/febs.12324.

Wijma, H.J., Floor, R.J., Bjelic, S., Marrink, S.J., Baker, D., Janssen, D.B., 2015. Enantioselective enzymes by computational design and *in silico* screening. Angew. Chem. Int. Ed. 54 (12), 3726–3730. https://doi.org/10.1002/anie.201411415.

Wu, B., Wijma, H.J., Song, L., Rozeboom, H.J., Poloni, C., Tian, Y., Arif, M.I., Nuijens, T., Quaedflieg, P.J.L.M., Szymanski, W., Feringa, B.L., Janssen, D.B., 2016. Versatile peptide C-terminal functionalization via a computationally engineered peptide amidase. ACS Catal. 6 (8), 5405–5414. https://doi.org/10.1021/acscatal.6b01062.

Xie, W.J., Asadi, M., Warshel, A., 2022. Enhancing computational enzyme design by a maximum entropy strategy. Proc. Natl. Acad. Sci. USA 119 (7), e2122355119. https://doi.org/10.1073/pnas.2122355119.

Xie, W.J., Liu, D., Wang, X., Zhang, A., Wei, Q., Nandi, A., Dong, S., Warshel, A., 2023. Enhancing luciferase activity and stability through generative modeling of natural enzyme sequences. Proc. Natl. Acad. Sci. 120 (48), e2312848120 https://doi.org/10.1073/pnas.2312848120.

Xu, E., Campanella, O.H., Ye, X., Jin, Z., Liu, D., BeMiller, J.N., 2020. Advances in conversion of natural biopolymers: A reactive extrusion (REX)-enzyme-combined strategy for starch/protein-based food processing. Trends Food Sci. Technol. 99, 167–180. https://doi.org/10.1016/j.tifs.2020.02.018.

Xu, W., Chen, Y., Li, D., Wang, Z., Xu, J., Wu, Q., 2022. Rational design of fatty acid photodecarboxylase enables the efficient decarboxylation of medium-and short-chain fatty acids for the production of gasoline bio-alkanes. Mol. Catal. 524, 112261 https://doi.org/10.1016/j.mcat.2022.112261.

Yang, Y., Niroula, A., Shen, B., Vihinen, M., 2016. PON-sol: prediction of effects of amino acid substitutions on protein solubility. Bioinformatics 32 (13), 2032–2034. https://doi.org/10.1093/bioinformatics/btw066.

Yang, M., Yang, S.-X., Liu, Z.-M., Li, N.-N., Li, L., Mou, H.-J., 2019. Rational design of alginate lyase from Microbulbifer sp. Q7 to improve thermal stability. Mar. Drugs 17 (6), 378. https://doi.org/10.3390/md17060378.

Yang, Y., Zeng, L., Vihinen, M., 2021. PON-Sol2: prediction of effects of variants on protein solubility. Int. J. Mol. Sci. 22 (15), 8027. https://doi.org/10.3390/ijms22158027.

Yang, Z.J., Shao, Q., Jiang, Y., Jurich, C., Ran, X., Juarez, R.J., Yan, B., Stull, S.L., Gollu, A., Ding, N., 2023. Mutexa: a computational ecosystem for intelligent protein engineering. J. Chem. Theory Comput. 19 (21), 7459–7477. https://doi.org/10.1021/acs.jctc.3c00602.

Yeh, A.H.-W., Norn, C., Kipnis, Y., Tischer, D., Pellock, S.J., Evans, D., Ma, P., Lee, G.R., Zhang, J.Z., Anishchenko, I., 2023. *De novo* design of luciferases using deep learning. Nature 614 (7949), 774–780. https://doi.org/10.1038/s41586-023-05696-3.

Yu, H., Yan, Y., Zhang, C., Dalby, P.A., 2017. Two strategies to engineer flexible loops for improved enzyme thermostability. Sci. Rep. 7, 1–15. https://doi.org/10.1038/srep41212.

Zamora, R.A., Ramirez-Sarmiento, C.A., Castro-Fernández, V., Villalobos, P., Maturana, P., Herrera-Morande, A., Komives, E.A., Guixé, V., 2020. Tuning of conformational dynamics through evolution-based design modulates the catalytic adaptability of an extremophilic kinase. ACS Catal. 10 (19), 10847–10857. https://doi.org/10.1021/acscatal.0c01300.

Zanghellini, A., Jiang, L., Wollacott, A.M., Cheng, G., Meiler, J., Althoff, E.A., Röthlisberger, D., Baker, D., 2006. New algorithms and an *in silico* benchmark for computational enzyme design. Protein Sci. 15 (12), 2785–2794. https://doi.org/10.1110/ps.062353106.

Zaugg, J., Gumulya, Y., Malde, A.K., Bodén, M., 2017. Learning epistatic interactions from sequence-activity data to predict enantioselectivity. J. Comput. Aided Mol. Des. 31 (12), 1085–1096. https://doi.org/10.1007/s10822-017-0090-x.

Zha, W., Zhang, F., Shao, J., Ma, X., Zhu, J., Sun, P., Wu, R., Zi, J., 2022. Rationally engineering santalene synthase to readjust the component ratio of sandalwood oil. Nat. Commun. 13 (1), 2508. https://doi.org/10.1038/s41467-022-30294-8.

Zhang, Y., Ma, C., Dischert, W., Soucaille, P., Zeng, A.P., 2019. Engineering of phosphoserine aminotransferase increases the conversion of L-homoserine to 4-hydroxy-2-ketobutyrate in a glycerol-independent pathway of 1, 3-propanediol production from glucose. Biotechnol. J. 14 (9), 1900003. https://doi.org/10.1002/biot.201900003.

Zhang, S., Zhang, J., Zhu, Y., 2020. ProdaMatch: A fast and accurate active site matching algorithm for *de novo* enzyme design. Comput. Chem. Eng. 140, 106921 https://doi.org/10.1016/j.compchemeng.2020.106921.