Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates

Zhenqiao Song[1] Yunlong Zhao[2] Wenxian Shi[3] Wengong Jin[4] Yang Yang[5] Lei Li[1]

Abstract

Enzymes are genetically encoded biocatalysts capable of accelerating chemical reactions. How can we automatically design functional enzymes? In this paper, we propose EnzyGen, an approach to learn a unified model to design enzymes across all functional families. Our key idea is to generate an enzyme's amino acid sequence and their three-dimensional (3D) coordinates based on functionally important sites and substrates corresponding to a desired catalytic function. These sites are automatically mined from enzyme databases. EnzyGen consists of a novel interleaving network of attention and neighborhood equivariant layers, which captures both long-range correlation in an entire protein sequence and local influence from nearest amino acids in 3D space. To learn the generative model, we devise a joint training objective, including a sequence generation loss, a position prediction loss and an enzyme-substrate interaction loss. We further construct EnzyBench, a dataset with 3157 enzyme families, covering all available enzymes within the protein data bank (PDB). Experimental results show that our EnzyGen consistently achieves the best performance across all 323 testing families, surpassing the best baseline by 10.79% in terms of substrate binding affinity. These findings demonstrate EnzyGen's superior capability in designing well-folded and effective enzymes binding to specific substrates with high affinities. The code, model and dataset are released at https://github.com/LeiLiLab/EnzyGen.

[1]Language Technologies Institute, Carnegie Mellon University. [2]Department of Chemistry, Massachusetts Institute of Technology [3]Department of EECS, Massachusetts Institute of Technology. [4]Broad Institute of MIT and Harvard. [5]Department of Chemistry and Biochemistry, University of California Santa Barbara. Correspondence to: Lei Li <leili@cs.cmu.edu>, Yang Yang <yang@chem.ucsb.edu>.

1. Introduction

Enzymes are biological catalysts to accelerate challenging chemical reactions underlying a range of biological processes. They enjoy widespread applications in the production of pharmaceuticals (Wu et al., 2012), specialty chemicals (Carbonell et al., 2018) and biofuels (Liao et al., 2016). In enzymatic reactions, the substrate is a small molecule being converted by the enzyme catalyst. By binding with and acting on specific substrates, enzymes allow for dramatically accelerated rates in the transformation of their substrates (Bar-Even et al., 2011). Designing enzymes that can bind to specific substrates is a critical yet challenging problem.

Recently, deep learning methods are promising in protein design (Huang et al., 2016; Pearce & Zhang, 2021). Existing deep learning approaches for functional protein design fall into three categories: (a) generative models for protein sequence design guided by fitness landscape (Brookes et al., 2019; Rives et al., 2021; Ren et al., 2022; Song & Li, 2023; Wang et al., 2023); (b) structure to sequence generation based on a targeted protein backbone (Ingraham et al., 2019; Dauparas et al., 2022; Zheng et al., 2023); (c) co-design of a backbone structure and a protein sequence that encodes it (Anishchenko et al., 2021; Wang et al., 2022; Shi et al., 2022; Yeh et al., 2023). However, these methods face many limitations on enzyme design. First, fitness-guided approaches are limited by the lack of fitness data for the vast majority of enzyme families. Second, the structures of many enzymes remain unknown (Binz et al., 2005; Fischman & Ofran, 2018). Third, prior approaches do not model substrates in enzyme design process. Finally, there is no unified model applicable to all enzyme families.

In this paper, we aim to learn a unified generative model to design functional enzymes across thousands of enzyme families. The key design idea lies in the notion that the biological function of an enzyme is enabled by a subset of residues (amino acids), known as functionally important sites (Bickel et al., 2002; Chakrabarti & Lanczycki, 2007; Wang et al., 2022), and that an ideal enzyme should be able to bind its substrate(s) in enzymatic reactions. Therefore, we formulate the enzyme design problem as jointly generating the enzyme sequence and backbone structure given auto-

matically mined functionally important sites and substrates related to specific enzyme functions.

To this end, we propose EnzyGen, a unified model to co-design enzyme sequence and backbone structure. EnzyGen comprises an enzyme modeling module and a substrate representation module. We devise neighborhood attentive equivariant layers (NAELs) to jointly model enzyme amino acid sequence and residue coordinates. Each NAEL consists of a global attention sub-layer and a neighborhood equivariant sub-layer. The global attention sub-layer captures correlations among all residues within an entire enzyme sequence, while the neighborhood equivariant sub-layer updates residue representations and coordinates based on nearest neighbors in the 3D space. This architectural design facilitates information exchange with varying levels of granularity, promoting a comprehensive interaction among different residues. The substrate representation module also uses neighborhood equivariant layers to model a given substrate. Furthermore, We add category tags that correspond to BRENDA enzyme families to guide the design of enzymes with specific functions. To train EnzyGen, we design a joint training objective, which includes an amino acid type prediction loss, a coordinate reconstruction loss and an enzyme-substrate interaction loss.

Our contributions are listed as follows:

- We propose EnzyGen to jointly design enzyme sequence and backbone structure. EnzyGen is the first unified enzyme design model across thousands of enzyme families.
- We create EnzyBench, a comprehensive benchmark for enzyme design, including $101,974$ PDB entries across $3,157$ enzyme families, and a set of three metrics to evaluate enzyme quality before wet-lab experiments. EnzyBench includes a test set with 323 families for various function validation.
- We conduct experiments to evaluate EnzyGen and prior functional protein design methods. These results show that enzymes designed by EnzyGen lead to the best average enzyme-substrate ESP score (Kroll et al., 2023) of 0.65, substrate binding affinity of -9.44, and AlphaFold2 pLDDT (Jumper et al., 2021) of 87.45 across 323 families (BRENDA enzyme classification fourth-level categories). EnzyGen outperforms the best baseline by 10.79% in terms of substrate binding affinity.

2. Related Work

Methods for Functional Protein Design. Functional protein design has been studied with a wide variety of methods. Some works focus on designing functional proteins guided by fitness landscape, including searching (Brookes & List-garten, 2018; Brookes et al., 2019; Kumar & Levine, 2020; Das et al., 2021; Hoffman et al., 2022; Melnyk et al., 2021; Anishchenko et al., 2021; Ren et al., 2022) or directly generating (Jain et al., 2022; Song & Li, 2023) protein sequences through deep generative models. Another category of methods concentrates on designing protein sequences capable of folding into backbone structures to fulfill specific functions (Wang et al., 2018; Strokach et al., 2020; Dauparas et al., 2022; Hsu et al., 2022; Sumida et al., 2023). Owing to the availability of large-scale data, some models are trained on extensive protein sequences from diverse families (Ferruz et al., 2022) or incorporate additional control tags for enhanced modeling (Madani et al., 2023). On the path from sequence to function, the role of structure emerges as a crucial intermediary. Consequently, some studies focus on the design of functional structures (Trippe et al., 2022; Watson et al., 2023). Recognizing the inherently complex interplay between protein sequence and structure, another set of works relies on the simultaneous generation of a new backbone structure and an amino acid sequence that folds into it (Anishchenko et al., 2021; Wang et al., 2022; Shi et al., 2022; Yeh et al., 2023; Song et al., 2023). Compared to the design of other functional proteins, functional enzyme design is largely underdeveloped. Some of the above methods have been applied to specific enzyme families, such as generating satisfactory enzyme sequences using deep generative models (Detlefsen et al., 2022; Lin et al., 2022a; Giessel et al., 2022). However, there is currently no unified generative model that is applicable to all enzyme families.

Functionally Important Site Discovery in Enzymes. As our method conditions enzyme design on automatically mined functionally important sites, crucial in determining their structures or functions (Tristem, 2000), it is necessary to revisit research on identifying functionally important sites in enzymes. Several databases, such as PROSITE (Hulo et al., 2006), Gene Ontology (Ashburner et al., 2000) and InterPro (Hunter et al., 2009) identify and annotate protein functional sites based on sequentially conserved residues and information extracted from experimental studies and literature searches. Some efforts have also been undertaken to derive functional insights using 3D structural information (Fetrow & Skolnick, 1998; Fetrow et al., 2001; Ivanisenko et al., 2005). Other investigators have predicted functional sites using multiple sequence alignments (Armon et al., 2001; Panchenko et al., 2004; Bray et al., 2009; Hosseini & Ilie, 2022).

3. Proposed Method: EnzyGen

An enzyme consists of a chain of amino acids (also called residues) connected by peptide bonds, which folds into a proper 3D structure. Let $\mathcal{A}$ be the set of 20 common amino acids. We denote the sequence of a N-residue enzyme by
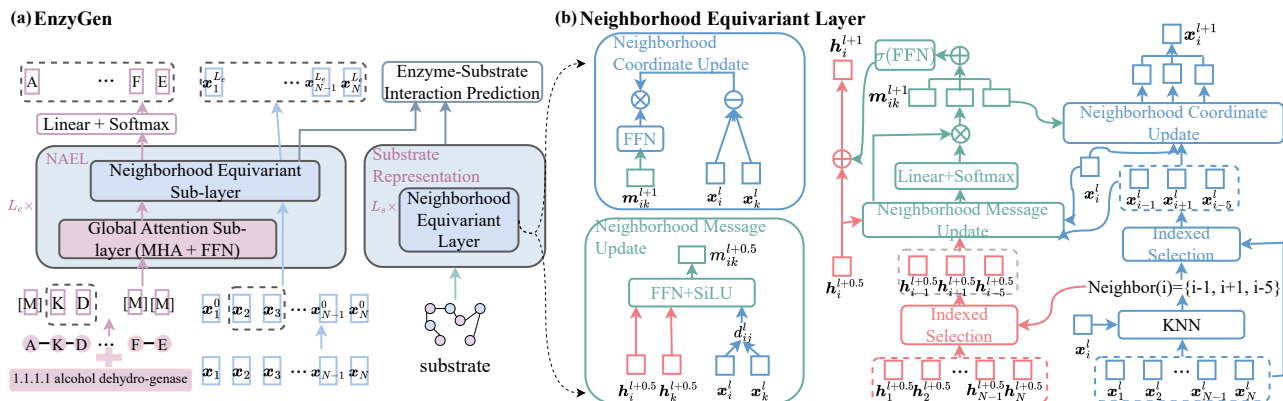
Figure 1. (a) EnzyGen architecture, consisting of an enzyme modeling module (left) and a substrate representation module (right). The enzyme modeling module aims to generate the enzyme sequence and backbone structure, and the substrate representation module targets at predicting if an enzyme can bind to a substrate. The dashed box in enzyme input denotes functionally important sites, while other sites need to be generated. [M] denotes mask token. "1.1.1.1" denotes the fourth-level enzyme class in the BRENDA enzyme classification (EC) tree. (b) Neighborhood equivariant layer: neighborhood message update (in green), neighborhood coordinate update (in blue) and neighborhood node feature update (in red). Indexed selection is choosing $\boldsymbol{x}_j$ (or $\boldsymbol{h}_j$) where j^{th} residue is in the K-nearest neighbors of ith residue.

$\boldsymbol{s} = \{s_1, s_2, ..., s_N\} \in \mathcal{A}^N$ and their C_α coordinates by $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]^T \in \mathbb{R}^{N \times 3}$. For residue type $s_i \in \mathcal{A}$, where $i \in \{1, 2, ..., N\}$, we denote its one-hot encoding as $\boldsymbol{s}_i = \text{onehot}(s_i)$ and the functionally important site index set as $\mathcal{M}$. We denote a substrate as $\mathcal{V} = \{v_1, ..., v_m\}$, where m is the number of its atoms, and $v_j = (\boldsymbol{h}_j, \boldsymbol{x}_j)$ denotes an atom where $j \in \{1, 2, ..., m\}$. Here, $\boldsymbol{h}_j$ represents pre-computed chemical features, and $\boldsymbol{x}_j \in \mathbb{R}^3$ is the corresponding coordinate. Each substrate is paired with a binary label y indicating whether the enzyme can bind with it. Inspired by multilingual machine translation models in natural language processing, which employ language tags to guide model generation in specific target languages (Johnson et al., 2017; Liu et al., 2020), we similarly use enzyme family tags to facilitate enzyme design with desired functions. To provide a systematical and professional enzyme family classification, we leverage the enzyme classification (EC) tree in BRENDA (Schomburg et al., 2004) as illustrated in Appendix Figure 5.

The problem studied in this paper can be formulated as follows: given a desired enzyme family c in EC tree, a substrate $\mathcal{V}$, a functionally important site index set $\mathcal{M}$, its residue set $\boldsymbol{s}_\mathcal{M}$ and the corresponding 3D coordinates $\boldsymbol{x}_\mathcal{M}$, generate an enzyme sequence $\boldsymbol{s}$ and C_α coordinates $\boldsymbol{x}$ of N residues satisfying the substrate binding constraint y. Essentially, we aim to learn a generative model with probability $P(\boldsymbol{s}, \boldsymbol{x}, y | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c, \mathcal{V})$.

3.1. Overall Model Architecture

We propose a model named EnzyGen, as illustrated in Figure 1 (a), to simultaneously generate an enzyme sequence and its 3D backbone structure, constrained by the enzyme's

small-molecule substrate. EnzyGen is a deep neural network consisting of an enzyme modeling module and a substrate representation module. The enzyme modeling module is composed of L_e stacked neighborhood attentive equivariant layers (NAELs). Each NAEL is composed of a global attention sub-layer using Transformer (Vaswani et al., 2017) and a neighborhood equivariant sub-layer (Figure 1 (b)) to incorporate information from the nearby residues based on C_α coordinates. The substrate representation module consists of L_s stacked neighborhood equivariant layers to pass messages inside the substrate $\mathcal{V}$, which will then provide binding constraint on the design of enzyme.

Suppose θ are EnzyGen parameters. We can formulate the joint probability as:

$$
\begin{aligned}
P(\boldsymbol{s}, \boldsymbol{x}, y | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c, \mathcal{V}; \theta) &= P(\boldsymbol{s}, \boldsymbol{x} | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c; \theta) \\
&\quad \cdot P(y | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c, \mathcal{V}; \theta) \\
P(\boldsymbol{s}, \boldsymbol{x} | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c; \theta) &= P(\boldsymbol{s} | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c; \theta) \\
&\quad \cdot P(\boldsymbol{x} | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c; \theta) \\
P(\boldsymbol{s} | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c; \theta) &= \Pi_{i=1 \& i \notin \mathcal{M}}^N P(\boldsymbol{s}_i | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c; \theta) \\
P(\boldsymbol{s}_i | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c; \theta) &= \text{Softmax}(W_\mathcal{A} \cdot \boldsymbol{h}_i^{L_e}) \\
P(\boldsymbol{x} | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c; \theta) &= \Pi_{i=1 \& i \notin \mathcal{M}}^N P(\boldsymbol{x}_i | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M}, c; \theta) \\
\boldsymbol{x}_i | \boldsymbol{s}_\mathcal{M}, \boldsymbol{x}_\mathcal{M} &\sim \mathcal{N}(\boldsymbol{x}_i^{L_e}; \lambda I)
\end{aligned}
\tag{1}
$$

where $W_\mathcal{A}$ is the embedding matrix for 20 common amino acids, $\boldsymbol{h}_i^{L_e}$ is the output embedding for i^{th} residue at the last layer of enzyme modeling module and $\boldsymbol{x}_i^{L_e}$ is the corresponding output coordinate. $\mathcal{N}(\boldsymbol{x}_i^{L_e}; \lambda I)$ is the Gaussian distribution with mean $\boldsymbol{x}_i^{L_e}$ and covariance matrix λI (I is the identity matrix). λ is a hyperparameter. To find the optimal θ, we maximize the conditional log likelihood (i.e.,

minimizing the negative log likelihood):

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\theta) = \arg\min_{\theta} -\log P(\boldsymbol{s}, \boldsymbol{x}, y | \boldsymbol{s}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{M}}, c, \mathcal{V}; \theta)$$

$$= \arg\min_{\theta} \Big\{ -\sum_{i=1 \& i \notin \mathcal{M}}^{N} \log P(\boldsymbol{s}_i | \boldsymbol{s}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{M}}, c; \theta)$$

$$- \sum_{i=1 \& i \notin \mathcal{M}}^{N} \log P(\boldsymbol{x}_i | \boldsymbol{s}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{M}}, c; \theta)$$

$$- \log P(y | \boldsymbol{s}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{M}}, c, \mathcal{V}; \theta) \Big\} \tag{2}$$

For simplicity, we omit the number of enzyme samples in the dataset. The second log likelihood function can be further simplified as:

$$\log P(\boldsymbol{x}_i | \boldsymbol{s}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{M}}, c; \theta) = -\frac{\lambda}{2} ||\boldsymbol{x}_i - \boldsymbol{x}_i^{L_e}||_2^2 + \text{const} \tag{3}$$

Therefore, the overall training objective is:

$$\mathcal{L}(\theta) = -\sum_{i=1 \& i \notin \mathcal{M}}^{N} \log P(\boldsymbol{s}_i | \boldsymbol{s}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{M}}, c; \theta)$$

$$+ \frac{\lambda}{2} \sum_{i=1 \& i \notin \mathcal{M}}^{N} ||\boldsymbol{x}_i - \boldsymbol{x}_i^{L_e}||_2^2 - \log P(y | \boldsymbol{s}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{M}}, c, \mathcal{V}; \theta) \tag{4}$$

3.2. NAEL Global Attention Sub-Layer

This sub-layer computes global contextual embeddings for all enzyme residues, which does not consider the closeness of residues in 3D space. By allowing every residue to attend to all other residues across the whole sequence, we facilitate information flow through the entire enzyme sequence.

We adopt the Transformer layer (Vaswani et al., 2017) to compute global contextual embeddings. Specifically, each transformer layer is composed of one multi-head self-attention sub-layer (MHA) and one fully connected feed-forward network (FFN). A residue connection and a layer normalization are employed after each of the two sub-layers. The calculation of the global sequence attention can be formulated as follows:

$$\boldsymbol{h}_i^{l+0.5} = \text{LayerNorm}\left(\text{FFN}(\tilde{\boldsymbol{h}}_i^{l+0.5}) + \tilde{\boldsymbol{h}}_i^{l+0.5} \right),$$

$$\tilde{\boldsymbol{h}}_i^{l+0.5} = \text{LayerNorm}\left(\text{MHA}(\boldsymbol{h}_i^l, \boldsymbol{H}_e^l) + \boldsymbol{h}_i^l \right) \tag{5}$$

where $\boldsymbol{h}_i^l$ is the ith residue input representation at l^{th} layer and $\boldsymbol{H}_e^l = [\boldsymbol{h}_1^l, \boldsymbol{h}_2^l, ..., \boldsymbol{h}_N^l]^T$. The input residue embeddings for the first layer are either taken from an embedding lookup table for functionally important sites, or initialized with a special [mask] token embedding for other residues. The residue embedding will then be enhanced by the enzyme family embedding $\text{Emb}(c)$ corresponding to the desired enzyme family c:

$$\boldsymbol{h}_i^0 = \hat{\boldsymbol{h}}_i^0 + \text{Emb}(c)$$

$$\hat{\boldsymbol{h}}_i^0 = \begin{cases} W_A^T \boldsymbol{s}_i, & i \in \mathcal{M}, \\ \text{Emb}([\text{mask}]), & \text{otherwise} \end{cases} \tag{6}$$

Similar to token-level language tags employed in multilingual models (Chi et al., 2020; Song et al., 2021), we also utilize the enzyme family tags at the residue level.

3.3. NAEL Neighborhood Equivariant Sub-Layer

Properly modeling the interactions of a given residue and its nearest neighboring residues in the 3D space can lead to improved residue representations. We intend to model such impact with a carefully designed subnetwork while keeping the equivariance under 3D translation and rotation. To this end, we propose the neighborhood equivariant sub-layer. This sub-layer includes three components: neighborhood message update, neighborhood coordinate update and neighborhood node feature update (Figure 1 (b)). Updating residue representations and coordinates in 3D space with only nearest neighbors enables more efficient and economic message passing compared to prior approaches which compute messages on the complete pairwise residue graph.

Neighborhood message update. We first compute distances between residues using C_α coordinates, and select K nearest residues (Figure 1 (b) green region). We compute the messages between i^{th} residue and its K-nearest neighbors (denoted as Neighbor(i)) as follows:

$$\boldsymbol{m}_{ik}^{l+0.5} = \text{SiLU}(\text{FFN}([(\boldsymbol{h}_i^{l+0.5}; \boldsymbol{h}_k^{l+0.5}; ||\boldsymbol{x}_i^l - \boldsymbol{x}_k^l||_2]))$$

$$w_{ik}^{l+0.5} = \frac{\exp(W_a^l \boldsymbol{m}_{ik}^{l+0.5} + b_a^l)}{\sum_{k' \in \text{Neighbor(i)}} \exp(W_a^l \boldsymbol{m}_{ik'}^{l+0.5} + b_a^l)} \tag{7}$$

$$\boldsymbol{m}_{ik}^{l+1} = w_{ik}^{l+0.5} * \boldsymbol{m}_{ik}^{l+0.5}$$

where FFN is a two-layer fully connected feed-forward network with SiLU activation function after its first layer. [;] is concatenation operator and $||\boldsymbol{x}_i^l - \boldsymbol{x}_k^l||_2$ is the Euclidean distance between i^{th} and k^{th} residue coordinates at l^{th} layer. W_* and b_* are trainable parameters.

Neighborhood coordinate update. We update the C_α coordinate of ith residue as a K-nearest neighbor vector field in a radial direction (Figure 1 (b) blue region). The C_α coordinate $\boldsymbol{x}_i^l$ at l^{th} layer of ith residue is updated with the weighted sum of all relative differences $(\boldsymbol{x}_i^l - \boldsymbol{x}_k^l)_{\forall k \in \text{Neighbor(i)}}$:

$$\boldsymbol{x}_i^{l+1} = \boldsymbol{x}_i^l + \sum_{k \in \text{Neighbor(i)}} (\boldsymbol{x}_i^l - \boldsymbol{x}_k^l) \cdot \text{FFN}(\boldsymbol{m}_{ik}^{l+1}) \tag{8}$$

The input residue C_α coordinates are either the given coordinates for functionally important sites, or randomly initialized as 3D points on the spherical surface centered at its left residue, considering the Euclidean distances between neighboring C_α pairs are almost the same (around $r = 3.75$Å):

$$\boldsymbol{x}_i^0 = \begin{cases} \boldsymbol{x}_i, & i \in \mathcal{M}, \\ \boldsymbol{x}_{i-1}^0 + r \cdot [\sin \omega_1 \cos \omega_2, \sin \omega_1 \sin \omega_2, \cos \omega_1]^T, & i \notin \mathcal{M} \end{cases} \tag{9}$$
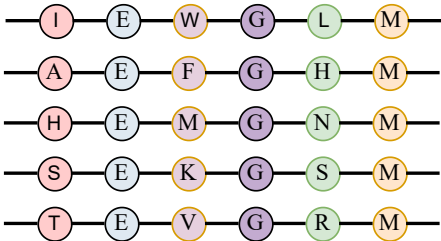
Figure 2. Discovering functionally important sites. Each row is a protein sequence in a same enzyme family (BRENDA fourth-level enzyme classification tree category). We use ClustalW2 to perform multiple sequence alignment and select common residuals above the identity threshold τ. According to the aligned sequences, E, G and M are common in all the sequences therefore these are selected as important sites. In experiment, $\tau = 30\%$.

where $\omega_1 \sim \text{Uniform}(0, \pi)$ is the angle to Z-axis and $\omega_2 \sim \text{Uniform}(0, 2\pi)$ is the angle to X-axis in polar coordinate system.

Neighborhood node feature update. We update the ith residue feature by gathering information from its K-nearest neighbors using a gating mechanism (Figure 1 (b) red region):

$$\boldsymbol{c}_i^{l+1} = \sum_{k \in \text{Neighbor}(i)} \boldsymbol{m}_{ik}^{l+1}$$
$$\boldsymbol{h}_i^{l+1} = \boldsymbol{h}_i^{l+0.5} + \sigma(\text{FFN}(\boldsymbol{c}_i^{l+1})) \odot \boldsymbol{c}_i^{l+1} \qquad (10)$$

where FFN is a two-layer fully connected feed-forward network with ReLU activation function after its first layer. σ denotes the sigmoid activation function.

We stack L_e layers of NAEL to model enzymes. The output embedding $\boldsymbol{h}_i^{L_e}$ and coordinate $\boldsymbol{x}_i^{L_e}$ for the ith residue are both from the last layer. Then the output probability of amino acid type for i^{th} residue is calculated as:

$$P(s_i = a | \boldsymbol{s}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{M}}, c) = \frac{\exp(h_{i,a}^o)}{\sum_{a'=1}^{20} \exp(h_{i,a'}^o)}, \quad \boldsymbol{h}_i^o = W_{\mathcal{A}} \cdot \boldsymbol{h}_i^{L_e} \qquad (11)$$

3.4. Substrate Representation Module

To facilitate effective and efficient message passing inside the substrate, we stack L_s neighborhood equivariant layers to learn the substrate representations. With the substrate atom embedding $\boldsymbol{h}_j^l$ and the corresponding coordinate $\boldsymbol{x}_j^l$ at the lth layer, we perform the substrate message passing as follows:

$$\boldsymbol{m}_{jk}^{l+1} = \phi_m(\boldsymbol{h}_j^l, \boldsymbol{h}_k^l, ||\boldsymbol{x}_j^l - \boldsymbol{x}_k^l||_2)$$
$$\boldsymbol{h}_j^{l+1} = \boldsymbol{h}_j^l + \phi_n(\boldsymbol{h}_j^l, \sum_{k \in \text{Neighbor}(j)} \boldsymbol{m}_{jk}^{l+1}) \qquad (12)$$

where ϕ_m and ϕ_n respectively denotes neighborhood message update and neighborhood node feature update. Following Guan et al. (2022), we initialize the node feature as

Dataset	Training	Validation	Test
Third-Level Category	256	30	30
Fourth-Level Category	3157	428	323
Size	98974	1500	1500
Average Seq. Length	327	331	328

Table 1. EnzyBench statistics. The first two rows are third and fourth-level categories in BRENDA enzyme classification tree. Size indicates the number of protein entries in PDB under the categories.

five chemical features $\hat{\boldsymbol{h}}_j^0 \in \mathbb{R}^5$ and $\boldsymbol{h}_j^0 = W_s \cdot \hat{\boldsymbol{h}}_j^0$ where $W_s \in \mathbb{R}^{d \times 5}$ is a mapping matrix and d is the representation dimensionality. Since we have the real 3D structure of the substrate, we do not need to update its atom coordinates.

Then based on the learned enzyme and substrate representations, we can predict whether the enzyme can bind with the given substrate as follows:

$$P(y | \boldsymbol{s}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{M}}, c, \mathcal{V}; \theta) = \text{Softmax}(W_b \cdot [f(\boldsymbol{H}_e^{L_e}); f(\boldsymbol{H}_s^{L_s})]) \qquad (13)$$

where $W_b \in \mathbb{R}^{2 \times 2d}$ is a mapping matrix and f denotes sum pooling operation. $\boldsymbol{H}_e^{L_e} = [\boldsymbol{h}_1^{L_e}, ..., \boldsymbol{h}_N^{L_e}]^T$ and $\boldsymbol{H}_s^{L_s} = [\boldsymbol{h}_1^{L_s}, ..., \boldsymbol{h}_m^{L_s}]^T$ denote the output embeddings of enzyme and substrate at the last layer of enzyme modeling module and substrate representation module, respectively.

3.5. Functionally Important Enzyme Site Discovery

Functional sites are often evolutionarily conserved in a given enzyme family (Tristem, 2000; Bickel et al., 2002). Following previous methods (Armon et al., 2001; Panchenko et al., 2004; Chakrabarti & Lanczycki, 2007; Bray et al., 2009), we leverage multiple sequence alignment (MSA) to mine evolutionarily conserved patterns, which are known as functionally or structurally important sites. Designing enzymes conditioning on these conserved patterns can facilitate the generation of functional enzymes. We employ the ClustalW2 method (Anderson et al., 2011) to perform multiple sequence alignment for all proteins in each enzyme family corresponding to BRENDA fourth-level EC tree category. We select the residues that occur at the same site in more than τ sequences. These residues are designated as functionally important sites. In practice, we set τ to be 30%. Figure 2 illustrate the procedure to select functionally important sites.

4. Experiments

4.1. EnzyBench Construction

We aim to evaluate the effectiveness of our EnzyGen across a diverse range of enzymes. To achieve this, we gather all enzymes from BRENDA (Schomburg et al., 2002), yielding

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4
PROTSEED	0.54	0.67	0.24	0.39	0.57	**0.83**	0.52	0.29	0.75	0.58	0.45	0.77	0.88	0.81	0.78
RFDiffusion+IF	0.45	0.94	**0.54**	0.39	0.47	0.43	0.48	**0.39**	0.52	0.46	0.53	0.50	0.51	0.60	0.55
ESM2+EGNN	0.58	0.90	0.35	0.35	0.63	0.79	0.53	0.32	0.80	0.59	0.51	0.76	0.88	0.88	0.77
EnzyGen	**0.64**	**0.98**	0.38	**0.42**	**0.72**	0.80	**0.61**	0.38	**0.86**	**0.66**	**0.53**	0.76	**0.92**	**0.93**	**0.80**

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	0.69	0.70	**0.90**	0.84	0.48	0.29	0.69	0.31	0.10	0.50	0.57	0.37	0.84	0.83	0.42	0.59
RFDiffusion+IF	0.53	0.33	0.61	0.62	**0.49**	**0.62**	0.45	**0.47**	**0.44**	0.55	0.63	**0.59**	0.59	0.84	0.45	0.53
ESM2+EGNN	0.70	0.71	0.78	0.82	0.43	0.22	0.56	0.35	0.11	0.61	0.73	0.37	**0.81**	0.89	0.54	0.61
EnzyGen	**0.79**	**0.76**	0.62	**0.88**	0.47	0.26	**0.73**	0.40	0.14	**0.66**	**0.78**	0.40	0.80	**0.93**	**0.57**	**0.65**

Table 2. ESP scores ($\uparrow$) of generated enzymes for 30 testing BRENDA EC categories. ESP evaluates the enzyme-substrate interaction function. Each number is the average ESP score of 50 testing cases for each category. IF denotes the inverse folding model ProteinMPNN. A ESP score of 0.6 or higher indicates a positive substrate binding. The last column (Avg) is the average across 30 categories. Notice that our EnzyGen outperforms all previous methods in ESP by a big margin.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4
PROTSEED	-6.61	-3.93	-4.27	-10.22	-6.71	-8.84	-9.58	-7.43	-10.01	-7.44	-5.46	-8.07	-9.68	-10.24	-11.68
RFDiffusion+IF	-7.11	-4.43	-4.70	**-10.74**	**-7.21**	-9.61	**-10.04**	-7.93	-10.64	-7.84	-6.19	-8.55	-10.60	-10.44	**-12.18**
ESM2+EGNN	-6.66	-4.47	-4.81	-10.73	-7.02	-9.57	-9.98	-8.61	-10.90	-7.95	-6.43	-8.79	-10.23	-10.75	-11.31
EnzyGen	**-8.44**	**-4.58**	**-5.10**	-10.34	-6.95	**-10.05**	-9.89	**-9.65**	**-11.91**	**-9.98**	**-8.05**	**-10.50**	**-11.65**	**-12.51**	-11.24

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	-8.00	-6.01	-7.20	-9.16	-9.53	-8.79	-8.67	-5.19	-5.44	-8.57	-10.11	-9.37	-9.74	-4.38	-9.11	-7.94
RFDiffusion+IF	-8.50	-6.51	-7.70	-11.65	-10.08	-9.29	-9.03	-5.54	-5.94	-9.07	-11.12	-9.87	-11.24	-4.88	-9.68	-8.57
ESM2+EGNN	-8.47	-5.81	-7.20	-11.45	-10.14	-8.91	-9.25	-5.59	-5.42	-8.23	-10.69	-11.15	**-11.34**	-5.01	-9.76	-8.52
EnzyGen	**-8.86**	**-7.01**	**-8.89**	**-11.90**	**-10.50**	**-10.60**	**-10.49**	**-6.37**	**-6.84**	**-9.23**	**-13.10**	**-11.35**	-11.03	**-5.51**	**-10.64**	**-9.44**

Table 3. Enzyme-substrate binding affinity ($\downarrow$) (kcal/mol) for 30 testing BRENDA EC categories, evaluated by a docking software Gnina. IF denotes the inverse folding model ProteinMPNN. The last column (Avg) is the average across 30 categories. Notice that our EnzyGen outperforms all previous methods in binding affinity by a big margin.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4
PROTSEED	77.10	81.80	71.19	74.24	78.67	77.40	74.54	75.18	77.11	74.79	75.55	75.90	81.05	74.05	76.50
RFDiffusion+IF	82.47	89.63	81.12	89.32	82.04	82.49	85.14	85.61	81.13	86.25	81.60	**87.51**	86.75	85.91	88.30
ESM2+EGNN	90.67	93.11	90.93	90.30	87.67	79.40	84.78	84.80	84.56	**90.21**	87.47	83.52	**88.92**	85.59	90.09
EnzyGen	**91.86**	**94.40**	**93.02**	**92.70**	**91.99**	**83.47**	**87.71**	**92.81**	**87.02**	89.69	**89.20**	85.19	87.55	**87.64**	**91.81**

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	78.00	76.29	77.89	75.75	78.76	73.56	82.40	76.70	75.90	75.16	74.62	83.46	76.36	78.87	83.31	77.07
RFDiffusion+IF	81.25	80.01	81.59	81.22	**92.04**	**89.72**	77.20	84.05	85.47	71.35	**82.87**	**84.49**	81.31	79.02	76.11	83.43
ESM2+EGNN	81.80	87.27	**87.05**	85.50	72.23	71.31	82.62	83.48	88.69	84.96	73.34	80.77	**87.72**	89.70	85.48	85.13
EnzyGen	**83.75**	**89.79**	85.40	**89.68**	74.44	77.14	**89.11**	**86.70**	**89.80**	**85.98**	76.31	84.32	85.71	**91.88**	**87.55**	**87.45**

Table 4. pLDDT ($\uparrow$) predicted by AlphaFold2. IF denotes the inverse folding model ProteinMPNN. The last column (Avg) is the average across 30 categories. Notice that our EnzyGen outperforms all previous methods in pLDDT. It shows EnzyGen's enzyme can fold more stably.

a total of $101,974$ PDB entries. These entries are classified into $3,157$ fourth-level enzyme categories based on the EC Tree (Appendix Figure 5). It is noteworthy that there exist a total of $8,422$ fourth-level categories; however, our focus lies specifically on those with experimentally confirmed structures. This focus results in our dataset comprising $3,157$ fourth-level categories. We then conduct MSAs for each fourth-level category to identify functionally important sites, using an MSA threshold (τ) of 30%. Given the relatively small size of fourth-level categories, we aggregate data belonging to the same third-level category in the EC Tree, resulting in 256 third-level categories. Then we select 30 third-level categories for validation and testing, respectively including 428 and 323 fourth-level categories.

For each of the 30 third-level categories, we we randomly split 100 PDB entries with 50 for validation and 50 for testing, while the remaining entries are utilized for training. To prevent data leakage, we cluster the PDB entries with a sequence identity threshold of 50%. We ensure that no PDB entries in the validation or test sets belong to the same cluster as those in the training set. We further collect $23,711$ experimentally confirmed <enzyme, substrate> pairs from Kroll et al. (2023), ensuring enzymes without a corresponding substrate have a randomly sampled negative one during the training process. We also make sure that all entries in the test set have a corresponding substrate to ensure function evaluation. Detailed data statistics are provided in Table 1 and Appendix C.
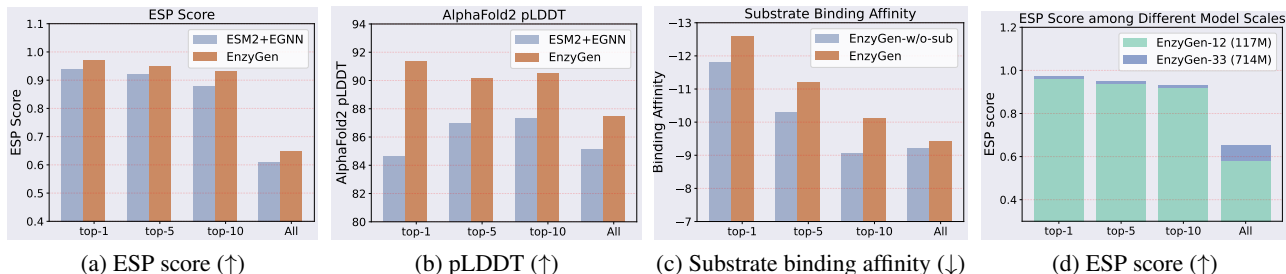
Figure 3. (a) Ablation study comparing EnzyGen against ESM2+EGNN on ESP score. (b) Ablation study comparing EnzyGen against ESM2+EGNN on AlphaFold2 pLDDT. (c) Ablation study comparing EnzyGen against the model removing enzyme-substrate interaction constraint (EnzyGen-w/o-sub). (d) Ablation study on different model scales.

4.2. Experimental Setup

Implementation Details. We use 33 NAELs to model enzymes and 3 neighborhood equivariant layers for substrate representation module with node feature dimensionality set to 1280. As suggested by Ying et al. (2018), the number of graph layers typically ranges from 2 to 6. To prevent over-fitting and facilitate efficiency, we utilize 3 neighborhood equivariant sub-layers for enzyme modeling, comprising a modified NAEL with every 11 global attention sub-layers followed by one neighborhood equivariant sub-layer. The parameters of the global attention sub-layer are initialized with the released 650M ESM-2 (Lin et al., 2022b) parameters. The total parameter count is 714M. The hyperparameters $\lambda/2$ and K are set to 1.0 and 30, respectively. The model undergoes training for $1,000,000$ steps using 8 NVIDIA RTX A6000 GPU cards, among which only the sequence generation and position prediction losses are adopted for the first $200,000$ steps. The batch size and learning rate are set to 8192 tokens and 3e-4 respectively. Sequences are decoded using the greedy decoding strategy. We incorporate four-level enzyme tags in the EC tree. For example, an enzyme in the category 1.1.1.1 should have four tags: 1, 1.1, 1.1.1, and 1.1.1.1.

Baseline Models. We compare the proposed Enzy-Gen against the following representative baselines: (1) **ESM2+EGNN** employs ESM2 (Lin et al., 2022b) as the sequence encoder and EGNN (Satorras et al., 2021) as the backbone structure predictor. The node features in EGNN are initialized with the output from ESM2. Both ESM2 and EGNN share the same layer configurations as our EnzyGen. (2) **PROTSEED** (Shi et al., 2022) co-designs protein sequences and backbone structures based on secondary structure and binary contact maps. (3) **RFDiffusion+ProteinMPNN** first applies RFDiffusion (Watson et al., 2023) to design an enzyme structure based on the given functionally important sites and then uses Protein-MPNN (Dauparas et al., 2022) to design a sequence based on the generated structure. To ensure a fair and reliable comparison, we train ESM2+EGNN, PROTSEED and Pro-

teinMPNN on our EnzyBench using their official codes and implementations. Since RFDiffusion does not provide a training script, we encountered challenges in training their code and reproducing their results. Consequently, we directly apply their released model to design the structures.

Function Evaluation. We evaluate the enzyme-substrate interaction function by the following metrics: (1) Following previous work (Li et al., 2013; Vidal-Limon et al., 2022), we use **enzyme-substrate binding affinity**, calculated by Gnina (McNutt et al., 2021) to quantify the strength of interactions between the designed enzymes and their substrates. A lower binding affinity indicates stronger binding. (2) We also use **ESP score**, developed by Kroll et al. (2023) to assess the ability of the designed enzymes to bind their corresponding substrates. In their paper, ESP model predicts enzyme-substrate interaction with 91% accuracy across multiple benchmarks. The scale of ESP score is 0-1, and a higher ESP score indicates a stronger enzyme-substrate interaction. Additionally, to evaluate if the designed enzymes are well-folded, we compute (3) **pLDDT** using AlphaFold2.

4.3. Main Results

The ESP scores, substrate binding affinities, and pLDDT results are shown in Tables 2, 3 and 4. For clarity, we average the performance of fourth-level categories under the same third-level one, with detailed results for 323 families in Appendix D.1. The results for top-1, top-5 and top-10 candidate are also provided in Appendix D.2.

EnzyGen excels in designing enzymes that bind their respective substrates with high affinities. Table 2 and 3 highlight the ability of EnzyGen in achieving the highest average ESP score and substrate binding affinity across 323 families. Notably, EnzyGen obtains an average ESP score of 0.65, surpassing the suggested enzyme-substrate interaction threshold of 0.6 from the ESP evaluator developer. In comparison, PROTSEED achieves 0.59, RFDiffusion+ProteinMPNN achieves 0.53, and ESM2+EGNN achieves 0.61 of average ESP score. EnzyGen's ESP score is significantly higher

than all previous baselines. This is further verified by the binding affinity scores. The enzymes generated by our Enzy-Gen exhibit stronger binding affinity to their corresponding substrates than all other methods across all 30 third-level testing EC categories. EnzyGen gains an average improvement of -0.87 in binding affinity over the previous best method (RFDiffusion+ProteinMPNN), which is significant. These results demonstrate that enzymes designed by our EnzyGen are able to bind their corresponding substrates with high affinities.

EnzyGen is able to design well-folded enzymes. According to the findings of Guo et al. (2022) and Binder et al. (2022), high pLDDT scores (e.g., > 80) indicate high confidence of the residue structure. EnzyGen attains an average pLDDT score of 87.45 across 323 families. It outperforms all three baselines (PROTSEED, RFDiffusion+ProteinMPNN, ESM2+EGNN). This demonstrates the capability of our model to design enzymes with stable folding, validating the value of incorporating structural information into the enzyme design process.

5. Analysis: Diving Deep into EnzyGen

5.1. Ablation Study: Do Interleaving Network and Substrate Constraint Help?

They both contribute to designed enzymes with better substrate binding. In comparison to directly concatenating ESM2 and EGNN, Figure 3 (a) illustrates that EnzyGen consistently enhances the average ESP score across 323 families when considering different candidate sets. In Figure 3 (b), EnzyGen achieves higher AlphaFold2 pLDDT than ESM2+EGNN across various candidate sets. These results affirm that the interleaving network within our proposed NAEL facilitates information exchange at different granularities, thereby aiding the design of well-folded enzymes with better substrate binding functions. We further study whether the proposed enzyme-substrate interaction constraint is useful. As shown in Figure 3 (c), incorporating the enzyme-substrate interaction constraint into the training process improves substrate binding affinities, boosting enzyme design with stronger substrate binding.

5.2. Ablation Study: Does Model Scale Help?

EnzyGen is scalable. To assess the scalability of our Enzy-Gen, we compare our EnzyGen (714M) with a randomly initialized 12-layer model (117M) trained on the same dataset. Figure 3 (d) presents the average ESP score across 323 categories for different candidate set. The results indicate that EnzyGen attains a higher enzyme-substrate interaction score as the model scales up, and this difference becomes more obvious when more candidates are considered. This confirms the scalability advantage of our EnzyGen.

5.3. Does Further Finetuning Bring Additional Benefits?

EnzyGen shows improvement with further finetuning. Drawing inspiration from the strategy of finetuning pre-trained multilingual models for specific languages (Liu et al., 2020; Conneau et al., 2020), we conduct further finetuning on each third-level category. Each category undergoes an additional finetuning of 30 epochs, leading to EnzyGen-finetune. As presented in Table 5, 15 categories exhibit improvements on ESP scores after finetuning. Simultaneously, the performance of 8 categories shows degradation, and it is noted that all of them have training sizes smaller than 1000, except for 2.3.1. These findings demonstrate that EnzyGen can indeed benefit from the further finetuning technique, with larger training data sizes yielding more substantial improvements.

5.4. Zero-shot Generalization Capability

To assess the generalization capabilities of our EnzyGen, we respectively select two unseen substrates (magnesium(2+), zinc(2+)) and two unseen fourth-level categories (2.7.13.3, 6.3.2.4) from Swiss-Prot, which are not included in our En-zyBench benchmark. For these enzyme cases, we utilize their structures provided by the AlphaFold Database. The ESP scores of the top-32 candidates for each new fourth-level category or substrate are presented in Figure 4 (a). Remarkably, the average ESP score is 0.83 across two fourth-level categories and 0.89 across two substrates, with more than 85% of the cases scoring over 0.6. These observations confirm that functionally important sites are crucial to design functional enzymes. Even not trained on these categories, **our EnzyGen exhibits the capability to design enzymes with substrate binding ability.**

5.5. Does EnzyGen Learn the Family Relationship?

Families with closer functions have closer embeddings. We cluster the 3,157 fourth-level family category embeddings in Figure 4 (b). It shows that enzyme families from the same superfamily are clustered together, displaying closer tag representations in the embedding space. This observation confirms that our family tag learns useful function information which can provide guidance for the design of desirable enzymes.

5.6. Designing Novel Enzymes with EnzyGen

EnzyGen is able to design novel enzymes that bind to specific substrates. Figure 4 (c) and (d) showcase two designed enzymes based on PDB 1KAG (catalyzing the specific phosphorylation of the 3-hydroxyl group of shikimic acid) and 5L2P (hydrolyzing various p-nitrophenyl phosphates, aromatic esters and p-nitrophenyl fatty acids) respectively. Both of the two designed enzymes achieve pLDDT
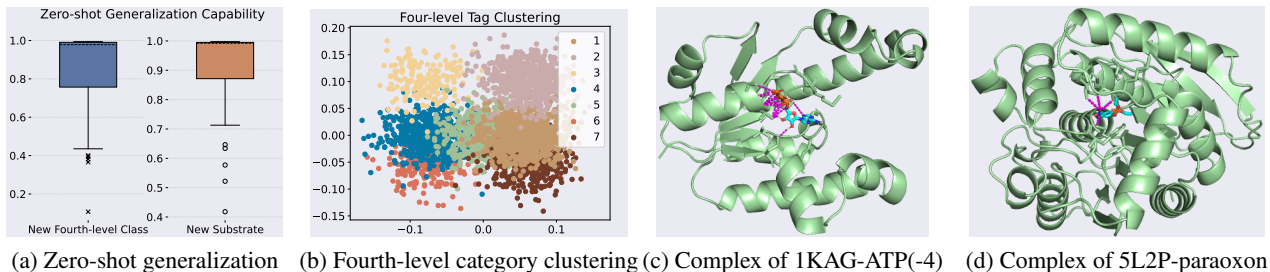
(a) Zero-shot generalization (b) Fourth-level category clustering (c) Complex of 1KAG-ATP(-4) (d) Complex of 5L2P-paraoxon

Figure 4. (a) ESP scores of designed enzymes from new fourth-level classes or with new substrates. Dash line denotes median. (b) Fourth-level category embedding clustering. Case Study: (c) Complex of designed 1KAG (2.7.1.71, catalyzing the specific phosphorylation of the 3-hydroxyl group of shikimic acid) and substrate ATP(-4), with pLDDT=90.39, Uniprot blastp recovery rate = 58.5%, (d) Complex of designed 5L2P (3.1.1.2, hydrolyzing various p-nitrophenyl phosphates, aromatic esters and p-nitrophenyl fatty acids) and substrate paraoxon, with pLDDT=89.44, Uniprot blastp recovery rate = 49.4%. Both cases show polar contacts (hydrogen bonds) depicted in purple.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4	
EnzyGen-finetune	**0.68**	0.95	0.38	0.37	0.68	**0.82**	0.57	0.38	0.86	**0.69**	**0.54**	**0.80**	**0.93**	**0.94**	**0.82**	
EnzyGen	0.64	**0.98**	0.38	**0.42**	**0.72**	0.80	**0.61**	0.38	0.86	0.66	0.53	0.76	0.92	0.93	0.80	

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
EnzyGen-finetune	0.79	**0.77**	**0.86**	0.88	**0.48**	0.24	0.63	0.40	**0.20**	0.68	**0.79**	0.37	**0.89**	0.93	0.56	**0.66**
EnzyGen	0.79	0.76	0.62	0.88	0.47	**0.26**	**0.73**	0.40	0.14	0.66	0.78	**0.40**	0.80	0.93	**0.57**	0.65

Table 5. ESP score (↑) of EnzyGen and EnzyGen-finetune. Avg denotes average. Finetuning leads to improvement in ESP scores.

approaching or exceeding 90 and substrate binding affinities below -10. Additionally, the Gnina docked complexes show polar contacts (hydrogen bonds) between the enzymes and substrates. This affirms the enzyme-substrate interaction function. We carry out blastp search of designed enzymes in Uniprot, yielding an amino acid identity of **58.5%** and **49.4%**, respectively, to the most similar enzyme. Together, these results show that our model is capable of designing novel enzymes with high substrate binding affinities. Additional cases are detailed in Appendix D.3, each exhibiting a pLDDT higher than 80, a binding affinity below -10, and a Uniprot blastp identity rate lower than 65%. Enzyme sequences for all analyzed cases are also available in Appendix Table 19.

6. Discussion

Enzymes, as genetically encoded biocatalysts, play a crucial role in accelerating chemical reactions and find extensive applications across various fields. Designing enzymes that can specifically bind to target substrates has always been a complex challenge. Our EnzyGen demonstrates a remarkable ability to design well-folded and effective enzymes that bind to specific substrates across diverse enzyme categories. Despite its impressive performance, certain limitations remain, which we will address in this section.

First, we collected enzyme data from the PDB, covering 3,157 fourth-level enzyme classes. Although there are 8,422 fourth-level classes in total, our dataset includes only

a subset of these categories. Future work could focus on expanding this dataset by incorporating data from larger databases, such as Swiss-Prot and the AlphaFold structure database, to encompass a broader range of enzyme classes.

Second, we address the substrate-binding constraint using an enzyme-substrate binding prediction loss. While this provides a basic constraint, it is relatively weak concerning the enzyme's catalytic functions. Future research could focus on designing enzymes that bind to a specific 3D substrate structure, potentially leading to the formation of a more effective enzyme-substrate complex.

7. Conclusion

This paper introduces EnzyGen, a unified generative model for designing functional enzymes across diverse families. EnzyGen simultaneously generates enzyme sequence and backbone structure guided by automatically identified functionally important sites and a given substrate. We employ a set of three losses to train EnzyGen. To comprehensively evaluate EnzyGen, we construct EnzyBench, a benchmark for enzyme design, including all available enzymes within PDB across 3157 enzyme families. Experimental results demonstrate EnzyGen's capability to design well-folded enzymes with strong enzyme-substrate interaction functions.

Acknowledgements

Impact Statement

This paper contributes to the effective and efficient design of functional enzymes, thereby advancing the fields of functional protein design and AI for molecule design. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Anderson, C. L., Strope, C. L., and Moriyama, E. N. Suitemsa: visual tools for multiple sequence alignment comparison and molecular sequence simulation. *BMC bioinformatics*, 12(1):1–14, 2011.

Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021.

Armon, A., Graur, D., and Ben-Tal, N. Consurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of molecular biology*, 307(1):447–463, 2001.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D. S., and Milo, R. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50(21):4402–4410, 2011.

Bickel, P. J., Kechris, K. J., Spector, P. C., Wedemayer, G. J., and Glazer, A. N. Finding important sites in protein sequences. *Proceedings of the National Academy of Sciences*, 99(23):14764–14771, 2002.

Binder, J. L., Berendzen, J., Stevens, A. O., He, Y., Wang, J., Dokholyan, N. V., and Oprea, T. I. Alphafold illuminates half of the dark human proteins. *Current Opinion in Structural Biology*, 74:102372, 2022.

Binz, H. K., Amstutz, P., and Plückthun, A. Engineering novel binding proteins from nonimmunoglobulin domains. *Nature biotechnology*, 23(10):1257–1268, 2005.

Bray, T., Chan, P., Bougouffa, S., Greaves, R., Doig, A. J., and Warwicker, J. Sitesidentify: a protein functional site prediction tool. *BMC bioinformatics*, 10(1):1–12, 2009.

Brookes, D., Park, H., and Listgarten, J. Conditioning by adaptive sampling for robust design. In *International conference on machine learning*, pp. 773–782. PMLR, 2019.

Brookes, D. H. and Listgarten, J. Design by adaptive sampling. *arXiv preprint arXiv:1810.03714*, 2018.

Carbonell, P., Wong, J., Swainston, N., Takano, E., Turner, N. J., Scrutton, N. S., Kell, D. B., Breitling, R., and Faulon, J.-L. Selenzyme: enzyme selection tool for pathway design. *Bioinformatics*, 34(12):2153–2154, 2018.

Chakrabarti, S. and Lanczycki, C. J. Analysis and prediction of functionally important sites in proteins. *Protein Science*, 16(1):4–13, 2007.

Chi, Z., Dong, L., Wei, F., Wang, W., Mao, X.-L., and Huang, H. Cross-lingual natural language generation via pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7570–7577, 2020.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobelt, H., Dos Santos, C., Chen, P.-Y., et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering*, 5(6):613–623, 2021.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Detlefsen, N. S., Hauberg, S., and Boomsma, W. Learning meaningful representations of protein sequences. *Nature communications*, 13(1):1914, 2022.

Ferruz, N., Schmidt, S., and Höcker, B. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Fetrow, J. S. and Skolnick, J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and t1ribonucleases. *Journal of molecular biology*, 281(5):949–968, 1998.

Fetrow, J. S., Siew, N., Di Gennaro, J. A., Martinez-Yamout, M., Dyson, H. J., and Skolnick, J. Genomic-scale comparison of sequence-and structure-based methods of function prediction: Does structure provide additional insight? *Protein science*, 10(5):1005–1014, 2001.

Fischman, S. and Ofran, Y. Computational design of antibodies. *Current opinion in structural biology*, 51:156–162, 2018.

Giessel, A., Dousis, A., Ravichandran, K., Smith, K., Sur, S., McFadyen, I., Zheng, W., and Licht, S. Therapeutic enzyme engineering using a generative neural network. *Scientific Reports*, 12(1):1536, 2022.

Guan, J., Qian, W. W., Peng, X., Su, Y., Peng, J., and Ma, J. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*, 2022.

Guo, H.-B., Perminov, A., Bekele, S., Kedziora, G., Farajollahi, S., Varaljay, V., Hinkle, K., Molinero, V., Meister, K., Hung, C., et al. Alphafold2 models indicate that protein sequence determines both structure and dynamics. *Scientific Reports*, 12(1):10696, 2022.

Hoffman, S. C., Chenthamarakshan, V., Wadhawan, K., Chen, P.-Y., and Das, P. Optimizing molecules using efficient queries from property evaluations. *Nature Machine Intelligence*, 4(1):21–31, 2022.

Hosseini, S. and Ilie, L. Pithia: Protein interaction site prediction using multiple sequence alignments and attention. *International Journal of Molecular Sciences*, 23 (21):12814, 2022.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*, pp. 8946–8970. PMLR, 2022.

Huang, P.-S., Boyken, S. E., and Baker, D. The coming of age of de novo protein design. *Nature*, 537(7620): 320–327, 2016.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P. S., Pagni, M., and Sigrist, C. J. The prosite database. *Nucleic acids research*, 34 (suppl_1):D227–D230, 2006.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L.,

Duquenne, L., et al. Interpro: the integrative protein signature database. *Nucleic acids research*, 37(suppl_1): D211–D215, 2009.

Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

Ivanisenko, V. A., Pintus, S. S., Grigorovich, D. A., and Kolchanov, N. A. Pdbsite: a database of the 3d structure of protein functional sites. *Nucleic Acids Research*, 33 (suppl_1):D183–D187, 2005.

Jain, M., Bengio, E., Hernandez-Garcia, A., Rector-Brooks, J., Dossou, B. F., Ekbote, C. A., Fu, J., Zhang, T., Kilgour, M., Zhang, D., et al. Biological sequence design with gflownets. In *International Conference on Machine Learning*, pp. 9786–9801. PMLR, 2022.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kroll, A., Ranjan, S., Engqvist, M. K., and Lercher, M. J. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature Communications*, 14(1):2787, 2023.

Kumar, A. and Levine, S. Model inversion networks for model-based optimization. *Advances in Neural Information Processing Systems*, 33:5126–5137, 2020.

Li, D., Lyons, J. A., Pye, V. E., Vogeley, L., Aragão, D., Kenyon, C. P., Shah, S. T., Doherty, C., Aherne, M., and Caffrey, M. Crystal structure of the integral membrane diacylglycerol kinase. *Nature*, 497(7450):521–524, 2013.

Liao, J. C., Mi, L., Pontrelli, S., and Luo, S. Fuelling the future: microbial engineering for the production of sustainable biofuels. *Nature Reviews Microbiology*, 14 (5):288–304, 2016.

Lin, E., Lin, C.-H., and Lane, H.-Y. De novo peptide and protein design using generative adversarial networks: an update. *Journal of Chemical Information and Modeling*, 62(4):761–774, 2022a.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M.,

Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022b.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742, 2020.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos Jr, J. L., Xiong, C., Sun, Z. Z., Socher, R., et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp. 1–8, 2023.

McNutt, A. T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., Sunseri, J., and Koes, D. R. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):1–20, 2021.

Melnyk, I., Das, P., Chenthamarakshan, V., and Lozano, A. Benchmarking deep generative models for diverse antibody sequence design. *arXiv preprint arXiv:2111.06801*, 2021.

Panchenko, A. R., Kondrashov, F., and Bryant, S. Prediction of functional sites by analysis of sequence and structure conservation. *Protein science*, 13(4):884–892, 2004.

Pearce, R. and Zhang, Y. Deep learning techniques have significantly impacted protein structure prediction and protein design. *Current opinion in structural biology*, 68: 194–207, 2021.

Ren, Z., Li, J., Ding, F., Zhou, Y., Ma, J., and Peng, J. Proximal exploration for model-guided protein sequence design. *bioRxiv*, 2022.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Satorras, V. G., Hoogeboom, E., and Welling, M. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Schomburg, I., Chang, A., and Schomburg, D. Brenda, enzyme data and metabolic information. *Nucleic acids research*, 30(1):47–49, 2002.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. Brenda, the enzyme database: updates and major new developments. *Nucleic acids research*, 32(suppl_1):D431–D433, 2004.

Shi, C., Wang, C., Lu, J., Zhong, B., and Tang, J. Protein sequence and structure co-design with equivariant translation. In *The Eleventh International Conference on Learning Representations*, 2022.

Song, Z. and Li, L. Importance weighted expectation-maximization for protein sequence design. In *International Conference on Machine Learning*, pp. 32349–32364. PMLR, 2023.

Song, Z., Zhou, H., Qian, L., Xu, J., Cheng, S., Wang, M., and Li, L. switch-glat: Multilingual parallel machine translation via code-switch decoder. In *International Conference on Learning Representations*, 2021.

Song, Z., Zhao, Y., Shi, W., Yang, Y., and Li, L. Functional geometry guided protein sequence and backbone structure co-design. *arXiv preprint arXiv:2310.04343*, 2023.

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., and Kim, P. M. Fast and flexible protein design using deep graph neural networks. *Cell systems*, 11(4):402–411, 2020.

Sumida, K. H., Núñez Franco, R., Kalvet, I., Pellock, S. J., Wicky, B. I., Milles, L. F., Dauparas, J., Wang, J., Kipnis, Y., Jameson, N., et al. Improving protein expression, stability, and function with proteinmpnn. *bioRxiv*, pp. 2023–10, 2023.

Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., and Jaakkola, T. S. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *The Eleventh International Conference on Learning Representations*, 2022.

Tristem, M. Molecular evolution—a phylogenetic approach. *Heredity*, 84(1):131–131, 2000.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vidal-Limon, A., Aguilar-Toala, J. E., and Liceaga, A. M. Integration of molecular docking analysis and molecular dynamics simulations for studying food proteins and bioactive peptides. *Journal of Agricultural and Food Chemistry*, 70(4):934–943, 2022.

Wang, J., Cao, H., Zhang, J. Z., and Qi, Y. Computational protein design with deep learning neural networks. *Scientific reports*, 8(1):1–9, 2018.

Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J. L., Castro, K. M., Ragotte, R., Saragovi, A., Milles, L. F., Baek, M., et al. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022.

Wang, Y., Tang, H., Huang, L., Pan, L., Yang, L., Yang, H., Mu, F., and Yang, M. Self-play reinforcement learning guides protein engineering. *Nature Machine Intelligence*, 5(8):845–860, 2023.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, pp. 1–3, 2023.

Wu, M.-C., Law, B., Wilkinson, B., and Micklefield, J. Bioengineering natural product biosynthetic pathways for therapeutic applications. *Current opinion in biotechnology*, 23(6):931–940, 2012.

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., et al. De novo design of luciferases using deep learning. *Nature*, 614(7949):774–780, 2023.

Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W., and Leskovec, J. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31, 2018.

Zheng, Z., Deng, Y., Xue, D., Zhou, Y., Ye, F., and Gu, Q. Structure-informed language models are protein designers. In *International conference on machine learning*, 2023.

Appendix

A. Proof: SE(3) Equivariance

A.1. Proof: NAEL Is SE(3) Equivariant

In this section we prove that our proposed NAEL is translation equivariant on x for any translation vector $t \in \mathbb{R}^3$ and rotation equivariant for any R from SO(3) group. More formally, we will prove the NAEL satisfies:

$$H^{l+1}, Rx^{l+1} + t = \text{NAEL}(H^l, Rx^l + t) \qquad (14)$$

When $l = 0$:
For message update, the distance between two residues is invariant as $d_{ij}^l = ||Rx_i^l + t - (Rx_j^l + t)||^2 = (x_i^l - x_j^l)^T R^T R(x_i^l - x_j^l) = ||x_i^l - x_j^l||^2$. h_i^l is the embedding of residue or [mask] token, and thus it is always invariant. Therefore, $h_i^{l+0.5}$ is also invariant. Since $h_i^{l+0.5}$, $h_j^{l+0.5}$ and d_{ij}^l are all invariant, and thus m_{ij}^{l+1} is also invariant to translation t and rotation R on x.

For coordinate update, updated x^{l+1} is equivariant to the translation t and rotation R on input x^l:

$$
\begin{aligned}
&(Rx_i^l + t) + \sum_{j \in \text{Neighbor(i)}} (Rx_i^l + t - (Rx_j^l + t)) \cdot \text{FFN}(m_{ij}^{l+1}) \\
&= R(x_i^l + \sum_{j \in \text{Neighbor(i)}} (x_i^l - x_j^l) \cdot \text{FFN}(m_{ij}^{l+1})) + t \\
&= Rx_i^{l+1} + t
\end{aligned}
\qquad (15)
$$

For residue update, m_{ij}^{l+1} and $h_i^{l+0.5}$ is invariant to translation t and rotation R on x, so h_i^{l+1} is also invariant to translation t and rotation R on x.

When $l >= 1$:
We have proved that when $l = 0$, h_i^1 is invariant to rotation and translation on x. Taking H^1 as the second layer input and following the above process, we can prove h_i^2 is also invariant. Repeating this process from $l = 1$ to $L - 1$, we can get the same conclusion.

Combining the above two scenarios together, we have $H^{l+1}, Rx^{l+1} + t = \text{NAEL}(H^l, Rx^l + t)$ for $l = 0$ to $L-1$. Therefore, our proposed NAEL is $SE(3)$-equivariant.

A.2. Proof: EnzyGen Is SE(3) Equivariant

We provide the proof of corollary 3.2 as follows:

Proof.

$$
\begin{aligned}
&\text{EnzyGen}(H^0, Rx^0 + t) \\
&= \text{NAEL}^{L-1} \circ \text{NAEL}^{L-2} \circ \cdots \circ \text{NAEL}^0(H^0, Rx^0 + t)) \\
&= \text{NAEL}^{L-1} \circ \text{NAEL}^{L-2} \circ \cdots \circ \text{NAEL}^1(H^1, Rx^1 + t)) \\
&= ... \\
&= \text{NAEL}^{L-1}(H^{L-1}, Rx^{L-1} + t)) = H^L, Rx^L + t
\end{aligned}
\qquad (16)
$$

B. Enzyme Classification Tree

We provide an illustration of enzyme classification (EC) tree in figure 5. These categories are applied by our EnzyGen to guide the enzyme design with specific functions.

C. Data Statistics

We provide the third-level category information in the validation and test sets in Table 6.

D. Additional Experimental Results

D.1. EnzyGen Performance on 323 Testing Fourth-Level Categories

The ESP scores, substrate binding affinities and AlphaFold2 pLDDT results for the 323 testing fourth-level categories are respectively provided in Table 7, 9 and 11.

D.2. Performance on Different Candidate Set

D.2.1. TOP-1 CANDIDATE RESULTS

ESP scores of top-1 candidate and their corresponding substrate binding affinities as well as pLDDT are respectively reported in Table 8, 10 and 12. The tables show that our EnzyGen achieves the best average scores across different categories on all metrics. Notably, the ESP scores of different models consistently surpass those of the top-5 candidate. This observation is reasonable, as candidates with higher probabilities are more likely to achieve higher function scores.

D.2.2. TOP-5 CANDIDATE RESULTS

ESP scores of top-5 candidate and their corresponding substrate binding affinities as well as pLDDT are respectively reported in Table 13, 14 and 15. The tables show that our EnzyGen achieves the best average scores across different categories on all metrics.

Third-Level Category	Function
1.1.1	With NAD+ or NADP+ as acceptor
1.11.1	Peroxidases
1.14.13	With NADH or NADPH as one donor, and incorporation of one atom of oxygen into the other donor
1.14.14	With reduced flavin or flavoprotein as one donor and incorporation of one atom of oxygen into the other donor
1.2.1	With NAD+ or NADP+ as acceptor
2.1.1	Methyltransferases
2.3.1	Transferring groups other than aminoacyl groups
2.4.1	Hexosyltransferases
2.4.2	Pentosyltransferases
2.5.1	Transferring alkyl or aryl groups, other than methyl groups (only sub-subclass identified to date)
2.6.1	Transaminases
2.7.1	Phosphotransferases with an alcohol group as acceptor
2.7.10	Protein-tyrosine kinases
2.7.11	Protein-serine/threonine kinases
2.7.4	Phosphotransferases with a phosphate group as acceptor
2.7.7	Nucleotidyltransferases
3.1.1	Carboxylic-ester hydrolases
3.1.3	Phosphoric-monoester hydrolases
3.1.4	Phosphoric-diester hydrolases
3.2.2	Hydrolysing N-glycosyl compounds
3.4.19	Omega peptidases
3.4.21	Serine endopeptidases
3.5.1	In linear amides
3.5.2	In cyclic amides
3.6.1	In phosphorus-containing anhydrides
3.6.4	Acting on acid anhydrides to facilitate cellular and subcellular movement
3.6.5	Acting on GTP to facilitate cellular and subcellular movement
4.1.1	Carboxy-lyases
4.2.1	Hydro-lyases
4.6.1	Phosphorus-oxygen lyases (only sub-subclass identified to date)

Table 6. Third-level category information for validation and test sets.

D.2.3. TOP-10 CANDIDATE RESULTS

ESP scores of top-10 candidate and their corresponding substrate binding affinities as well as pLDDT are respectively reported in Table 16, 17 and 18. Again, our model achieves the best average performance in all cases. It is worth noting that our EnzyGen achieves an average enzyme-substrate ESP score of 0.93, an average substrate binding affinity of -10.12 and an average pLDDT of 90.52, demonstrating EnzyGen's ability to design both well-folded and functional enzymes.

D.3. More Novel Cases

We provide more designed enzymes in Figure 6. All these cases achieve a pLDDT higher than 80, a blastp amino acid identity rate lower than 65% in Uniprot, and a substrate binding affinity lower than -10. After Gnina docking, the corresponding substrates demonstrate polar contacts (hydrogen bonds) with the enzymes (shown in purple), affirming the enzyme-substrate interaction functions. We also provide the enzyme sequences of all the analyzed cases in Table 19.
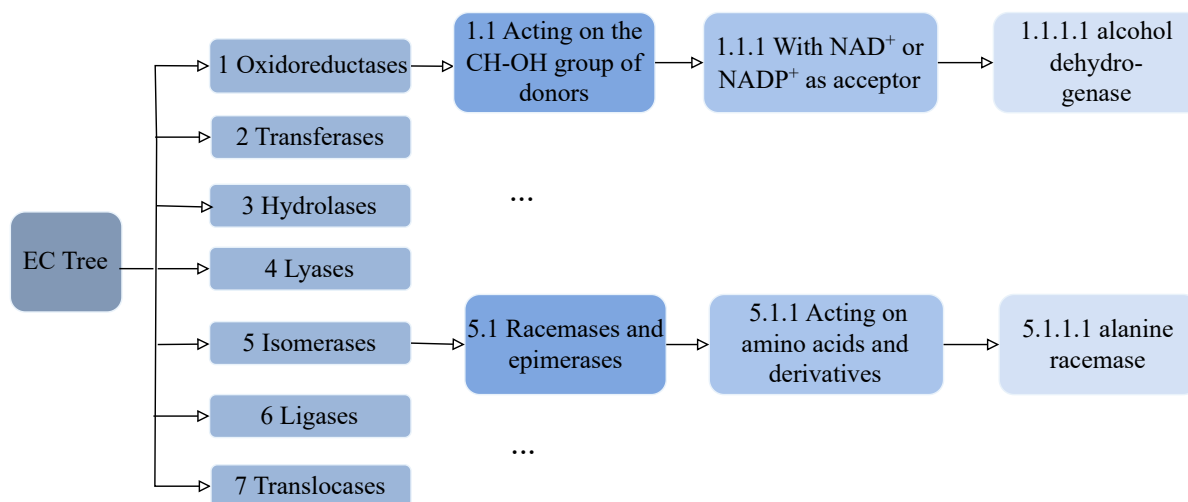
Figure 5. Enzyme Classification (EC) Tree in BRENDA.
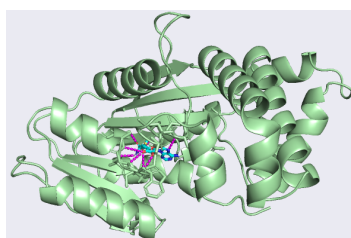
(a) Complex of designed 1L1E and S-adenosyl-L-methionine zwitterion. pLDDT=86.86, binding affinity=-11.02, Uniprot blastp recovery rate=56.5%
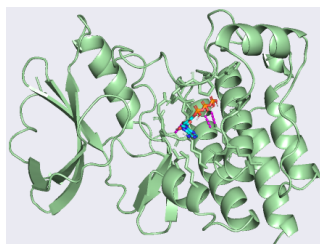
(b) Complex of designed 8DSO and ATP. pLDDT=80.95, binding affinity=-11.19, Uniprot blastp recovery rate=58.1%
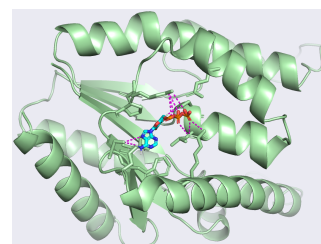
(c) Complex of designed 5TMP and ATP(4-). pLDDT=85.34, binding affinity=-11.76, Uniprot blastp recovery rate=59.8%
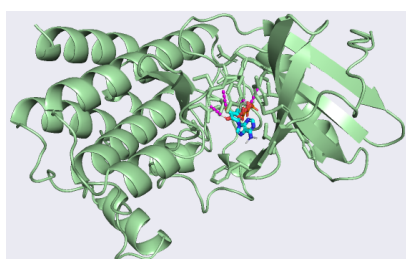
(d) Complex of designed 6W7O and ATP. pLDDT=80.84, binding affinity=-10.52, Uniprot blastp recovery rate=58.0%
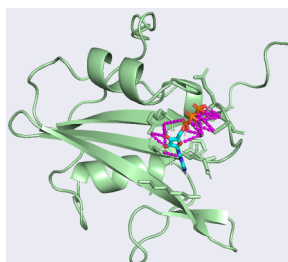
(e) Complex of designed 3UYO and ATP. pLDDT=81.69, binding affinity=-11.6, Uniprot blastp recovery rate=64.7%

(f) Complex of designed 4NI2 and ATP. pLDDT=81.32, binding affinity=-11.92, Uniprot blastp recovery rate=55%

Figure 6. Complexes of the EnzyGen designed enzymes and their corresponding substrates.

Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates

Protein Family	1.1.1.239	1.1.1.270	1.1.1.201	1.1.1.184	1.1.1.25	1.1.1.62	1.1.1.197	1.1.1.35	1.1.1.248	1.1.1.271	1.1.1.372	1.1.1.357	1.1.1.27	1.1.1.286	1.1.1.47	1.1.1.3
EnzyGen	0.62	0.15	0.09	0.16	0.92	0.38	0.39	0.55	0.89	0.62	0.86	0.71	0.98	0.14	0.99	0.39

Protein Family	1.1.1.56	1.1.1.105	1.1.1.77	1.1.1.34	1.1.1.399	1.1.1.42	1.1.1.49	1.1.1.44	1.1.1.40	1.11.1.24	1.11.1.12	1.11.1.15	1.11.1.9	1.11.1.11	1.11.1.5	1.11.1.6
EnzyGen	0.39	0.64	0.92	0.07	0.98	0.65	0.82	0.16	0.99	1.0	1.0	1.0	0.99	0.92	0.99	0.9

Protein Family	1.11.1.21	1.14.13.39	1.14.13.9	1.14.13.23	1.14.14.18	1.14.14.47	1.14.14.25	1.14.14.154	1.14.14.1	1.14.14.32	1.14.14.24	1.14.14.26	1.14.14.91	1.2.1.3	1.2.1.50	1.2.1.44
EnzyGen	0.99	0.99	0.4	0.19	0.6	0.27	0.24	0.03	0.56	0.36	0.99	0.22	0.04	0.44	0.66	0.48

Protein Family	1.2.1.12	1.2.1.59	1.2.1.13	1.2.1.11	1.2.1.10	1.2.1.5	1.2.1.47	1.2.1.36	1.2.1.9	1.2.1.26	1.2.1.88	1.2.1.84	1.2.1.89	1.2.1.90	2.1.1.309	2.1.1.37
EnzyGen	0.94	0.84	0.36	0.92	0.96	0.59	0.87	0.6	0.84	0.99	0.73	0.03	0.82	0.91	0.06	0.94

Protein Family	2.1.1.364	2.1.1.221	2.1.1.348	2.1.1.320	2.1.1.6	2.1.1.366	2.1.1.367	2.1.1.244	2.1.1.359	2.1.1.72	2.1.1.220	2.1.1.98	2.1.1.79	2.1.1.233		
EnzyGen	0.73	0.02	1.0	0.76	1.0	0.84	0.99	1.0	0.95	0.82	1.0	0.85	0.78	0.97	0.02	0.02

Protein Family	2.1.1.319	2.1.1.5	2.1.1.354	2.1.1.173	2.3.1.48	2.3.1.57	2.3.1.82	2.3.1.258	2.3.1.87	2.3.1.108	2.3.1.39	2.3.1.286	2.3.1.94	2.3.1.117	2.3.1.180	2.3.1.16
EnzyGen	0.49	0.72	0.82	0.76	0.75	0.34	0.97	0.35	0.07	0.2	0.43	0.31	0.67	0.53	0.98	0.79

Protein Family	2.3.1.97	2.3.1.9	2.3.1.47	2.3.1.26	2.3.1.50	2.3.1.37	2.4.1.186	2.4.1.135	2.4.1.38	2.4.1.255	2.4.1.143	2.4.1.264	2.4.1.182	2.4.1.345	2.4.1.115	2.4.1.15
EnzyGen	0.76	0.77	0.32	0.97	0.35	0.07	0.35	0.68	0.14	0.19	0.11	0.43	0.48	0.52	0.32	0.35

Protein Family	2.4.1.21	2.4.1.41	2.4.1.155	2.4.1.11	2.4.1.225	2.4.1.1	2.4.2.30	2.4.2.2	2.4.2.8	2.4.2.3	2.4.2.1	2.4.2.19	2.4.2.17	2.4.2.21	2.4.2.64	2.4.2.14
EnzyGen	0.98	0.96	0.29	0.38	0.56	0.84	0.89	0.49	0.93	0.99	0.98	0.93	0.92	0.69	0.65	0.15

Protein Family	2.4.2.36	2.5.1.18	2.5.1.31	2.5.1.68	2.5.1.55	2.5.1.16	2.5.1.1	2.5.1.15	2.5.1.79	2.5.1.87	2.5.1.29	2.5.1.59	2.5.1.126	2.5.1.47	2.5.1.46	2.5.1.54
EnzyGen	0.68	0.27	0.28	0.99	0.64	0.91	0.99	0.58	0.88	0.12	0.99	0.98	0.82	1.0	0.96	0.95

Protein Family	2.5.1.6	2.5.1.7	2.5.1.19	2.5.1.21	2.6.1.21	2.6.1.9	2.6.1.16	2.6.1.99	2.6.1.51	2.6.1.63	2.6.1.3	2.6.1.1	2.6.1.57	2.6.1.7	2.6.1.81	2.6.1.73
EnzyGen	0.98	0.98	0.98	0.95	0.05	0.23	0.32	0.06	0.7	0.99	0.99	0.41	0.31	1.0	0.29	0.94

Protein Family	2.6.1.83	2.6.1.62	2.6.1.85	2.6.1.82	2.7.1.107	2.7.1.154	2.7.1.26	2.7.1.71	2.7.1.21	2.7.1.24	2.7.1.113	2.7.1.153	2.7.1.23	2.7.1.35	2.7.1.33	2.7.1.68
EnzyGen	0.95	0.71	0.04	0.66	0.39	0.83	0.06	0.99	0.75	0.14	0.57	0.61	0.94	0.97	0.81	0.98

Protein Family	2.7.1.20	2.7.1.178	2.7.1.19	2.7.1.159	2.7.1.82	2.7.1.73	2.7.1.40	2.7.1.1	2.7.1.47	2.7.1.158	2.7.1.5	2.7.10.2	2.7.10.1	2.7.11.1	2.7.11.21	2.7.11.25
EnzyGen	0.68	0.57	0.34	0.09	0.78	0.13	0.99	0.98	0.89	0.99	0.98	0.88	0.98	0.92	0.63	1.0

Protein Family	2.7.11.22	2.7.11.30	2.7.11.24	2.7.11.26	2.7.11.2	2.7.11.16	2.7.4.6	2.7.4.3	2.7.4.8	2.7.4.14	2.7.4.9	2.7.4.25	2.7.4.10	2.7.4.22	2.7.4.24	2.7.4.1
EnzyGen	1.0	1.0	1.0	1.0	0.99	1.0	1.0	0.75	0.4	0.99	0.96	0.94	0.7	0.24	0.15	0.39

Protein Family	2.7.7.7	2.7.7.6	2.7.7.65	2.7.7.60	2.7.7.86	2.7.7.31	2.7.7.12	2.7.7.52	2.7.7.27	2.7.7.23	2.7.7.9	2.7.7.48	3.1.1.4	3.1.1.3	3.1.1.78	3.1.1.2
EnzyGen	0.91	0.29	0.42	0.8	0.38	0.98	0.98	0.41	0.88	0.15	0.57	0.38	0.16	0.19	0.51	0.84

Protein Family	3.1.1.23	3.1.1.72	3.1.1.47	3.1.1.34	3.1.1.8	3.1.1.7	3.1.1.5	3.1.3.32	3.1.3.46	3.1.3.3	3.1.3.104	3.1.3.16	3.1.3.108	3.1.3.7	3.1.3.11	3.1.3.36
EnzyGen	0.5	0.43	0.96	0.86	0.86	0.91	0.13	0.17	0.67	0.5	0.07	0.83	0.99	0.23	0.93	0.53

Protein Family	3.1.3.10	3.1.3.76	3.1.3.1	3.1.4.52	3.1.4.17	3.1.4.55	3.1.4.35	3.1.4.53	3.2.2.6	3.2.2.9	3.2.2.16	3.2.2.8	3.2.2.3	3.2.2.1	3.2.2.10	3.4.19.12
EnzyGen	0.25	0.21	0.09	0.23	0.95	0.03	0.98	0.99	0.41	0.82	0.2	0.55	0.51	0.87	0.77	0.44

Protein Family	3.4.19.5	3.4.19.14	3.4.19.13	3.4.21.53	3.4.21.92	3.4.21.107	3.4.21.22	3.5.1.98	3.5.1.26	3.5.1.97	3.5.1.124	3.5.1.9	3.5.1.44	3.5.1.1	3.5.1.2	3.5.1.25
EnzyGen	0.22	0.07	0.07	0.73	0.69	0.91	0.57	0.23	0.08	0.99	0.83	0.32	0.56	0.07	0.07	0.41

Protein Family	3.5.1.16	3.5.1.92	3.5.1.99	3.5.1.62	3.5.1.23	3.5.2.17	3.5.2.6	3.5.2.10	3.6.1.12	3.6.1.23	3.6.1.55	3.6.1.56	3.6.1.9	3.6.1.60	3.6.1.1	3.6.1.72
EnzyGen	0.13	0.11	0.22	0.44	0.27	0.15	0.13	0.63	0.97	0.11	0.63	0.45	0.88	0.93	0.9	0.54

Protein Family	3.6.1.15	3.6.1.58	3.6.1.8	3.6.1.22	3.6.1.27	3.6.1.29	3.6.1.5	3.6.4.6	3.6.4.12	3.6.4.13	3.6.4.10	3.6.5.2	3.6.5.1	3.6.5.5	4.1.1.23	4.1.1.85
EnzyGen	0.63	0.97	0.99	0.84	0.75	0.14	0.76	0.73	0.79	0.4	0.96	0.4	0.42	0.29	0.97	0.87

Protein Family	4.1.1.17	4.1.1.37	4.1.1.39	4.1.1.28	4.1.1.12	4.1.1.49	4.1.1.25	4.1.1.32	4.2.1.10	4.2.1.19	4.2.1.1	4.2.1.24	4.2.1.22	4.2.1.8	4.2.1.2	4.6.1.1
EnzyGen	0.02	0.57	0.88	0.02	0.33	0.99	0.04	0.45	0.81	0.85	1.0	0.64	0.01	0.81	0.96	0.65

| Protein Family	4.6.1.17	4.6.1.12	4.6.1.2													Avg
EnzyGen	0.49	0.4	0.52													0.65

Table 7. Enzyme-substrate ESP score (↑) for the 323 testing fourth-level categories. Note 0.6 or higher ESP score indicates a positive binding. Avg denotes average.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4	Avg
PROTSEED	0.98	0.99	0.48	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	–
RFDiffusion+IF	0.89	0.95	0.72	0.58	0.95	0.68	0.76	0.95	0.86	0.95	0.95	0.95	0.82	0.95	0.95	–
ESM2+EGNN	0.96	0.97	0.69	0.97	0.98	0.98	0.98	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98	–
EnzyGen	0.99	1.00	0.73	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	0.98	0.93	0.95	0.98	0.92	0.91	0.99	0.99	0.25	0.97	0.98	0.47	0.95	0.99	0.97	0.92
RFDiffusion+IF	0.98	0.97	0.93	0.94	0.92	0.94	0.97	0.96	0.60	0.97	0.97	0.94	0.92	0.98	0.97	0.94
ESM2+EGNN	0.98	0.97	0.93	0.94	0.92	0.94	0.97	0.96	0.60	0.97	0.97	0.94	0.92	0.98	0.97	0.94
EnzyGen	1.00	1.00	1.00	1.00	0.96	0.99	1.00	0.99	0.63	0.99	1.00	0.89	0.99	1.00	0.99	**0.97**

Table 8. Enzyme-substrate ESP score (↑) of the top-1 candidate. IF denotes the inverse folding model ProteinMPNN. Note 0.6 or higher ESP score indicates a positive binding. Avg denotes average.

Protein Family	1.1.1.239	1.1.1.270	1.1.1.201	1.1.1.184	1.1.1.25	1.1.1.62	1.1.1.197	1.1.1.35	1.1.1.248	1.1.1.271	1.1.1.372	1.1.1.357	1.1.1.27	1.1.1.286	1.1.1.47	1.1.1.3
EnzyGen	-4.29	-9.76	-11.74	-12.03	-3.29	-13.14	-3.32	-12.46	-4.75	-3.52	-3.29	-10.41	-8.47	-6.32	-12.56	-6.57

Protein Family	1.1.1.56	1.1.1.105	1.1.1.77	1.1.1.34	1.1.1.399	1.1.1.42	1.1.1.49	1.1.1.44	1.1.1.40	1.11.1.24	1.11.1.12	1.11.1.15	1.11.1.9	1.11.1.11	1.11.1.5	1.11.1.6
EnzyGen	-12.31	-13.01	-10.87	-9.23	-10.65	-9.37	-9.51	-9.01	-9.87	-5.03	-4.42	-4.48	-4.42	-4.38	-4.55	-5.44

Protein Family	1.11.1.21	1.14.13.39	1.14.13.9	1.14.13.23	1.14.14.18	1.14.14.47	1.14.14.25	1.14.14.154	1.14.14.1	1.14.14.32	1.14.14.24	1.14.14.26	1.14.14.91	1.2.1.3	1.2.1.50	1.2.1.44
EnzyGen	-6.64	-5.1	-6.24	-6.92	-12.34	-10.96	-8.63	-10.25	-7.32	-10.12	-8.76	-9.04	-9.84	-5.64	-5.66	-5.65

Protein Family	1.2.1.12	1.2.1.59	1.2.1.13	1.2.1.11	1.2.1.10	1.2.1.5	1.2.1.47	1.2.1.36	1.2.1.9	1.2.1.26	1.2.1.88	1.2.1.84	1.2.1.89	1.2.1.90	2.1.1.309	2.1.1.37
EnzyGen	-6.17	-14.72	-6.22	-10.11	-10.98	-7.9	-9.44	-6.31	-12.57	-13.56	-11.75	-7.13	-8.62	-10.26	-9.11	-9.84

Protein Family	2.1.1.364	2.1.1.221	2.1.1.348	2.1.1.320	2.1.1.6	2.1.1.366	2.1.1.367	2.1.1.244	2.1.1.359	2.1.1.72	2.1.1.220	2.1.1.45	2.1.1.98	2.1.1.79	2.1.1.8	2.1.1.233
EnzyGen	-9.04	-8.16	-10.96	-9.51	-11.43	-9.98	-8.9	-9.01	-9.24	-10.81	-9.64	-8.07	-7.94	-10.94	-11.95	

Protein Family	2.1.1.319	2.1.1.5	2.1.1.354	2.1.1.173	2.3.1.48	2.3.1.57	2.3.1.82	2.3.1.258	2.3.1.87	2.3.1.108	2.3.1.39	2.3.1.286	2.3.1.94	2.3.1.117	2.3.1.180	2.3.1.16
EnzyGen	-5.93	-7.97	-8.63	-8.9	-10.63	-6.52	-8.29	-9.47	-9.27	-9.41	-9.61	-8.76	-10.12	-10.59	-10.91	-9.98

Protein Family	2.3.1.97	2.3.1.9	2.3.1.47	2.3.1.26	2.3.1.50	2.3.1.37	2.4.1.186	2.4.1.135	2.4.1.38	2.4.1.255	2.4.1.143	2.4.1.264	2.4.1.182	2.4.1.345	2.4.1.115	2.4.1.15
EnzyGen	-7.56	-7.21	-8.34	-10.25	-9.61	-8.9	-8.61	-8.67	-10.83	-11.73	-5.25	-5.93	-6.35	-8.76	-9.44	-10.25

Protein Family	2.4.1.21	2.4.1.41	2.4.1.155	2.4.1.11	2.4.1.225	2.4.1.1	2.4.2.30	2.4.2.2	2.4.2.8	2.4.2.3	2.4.2.1	2.4.2.19	2.4.2.17	2.4.2.21	2.4.2.64	2.4.2.14
EnzyGen	-11.67	-10.1	-11.32	-11.53	-10.54	-11.84	-11.13	-13.41	-13.88	-9.98	-8.24	-10.61	-9.01	-9.44	-15.76	-13.06

Protein Family	2.4.2.36	2.5.1.18	2.5.1.31	2.5.1.68	2.5.1.55	2.5.1.16	2.5.1.1	2.5.1.15	2.5.1.79	2.5.1.87	2.5.1.29	2.5.1.59	2.5.1.126	2.5.1.47	2.5.1.46	2.5.1.54
EnzyGen	-10.64	-10.04	-11.41	-8.84	-8.79	-8.27	-8.21	-9.03	-9.98	-10.67	-12.95	-9.85	-9.87	-10.86	-11.23	-10.01

Protein Family	2.5.1.6	2.5.1.7	2.5.1.19	2.5.1.26	2.6.1.21	2.6.1.9	2.6.1.16	2.6.1.99	2.6.1.51	2.6.1.63	2.6.1.3	2.6.1.1	2.6.1.57	2.6.1.7	2.6.1.81	2.6.1.73
EnzyGen	-5.93	-5.82	-9.98	-7.08	-8.8	-8.98	-7.43	-7.02	-8.01	-7.81	-8.9	-8.02	-7.65	-8.9		

Protein Family	2.6.1.83	2.6.1.62	2.6.1.85	2.6.1.82	2.7.1.107	2.7.1.154	2.7.1.26	2.7.1.71	2.7.1.21	2.7.1.24	2.7.1.113	2.7.1.153	2.7.1.23	2.7.1.35	2.7.1.33	2.7.1.68
EnzyGen	-8.01	-8.87	-8.12	-9.23	-9.21	-10.64	-12.58	-9.09	-12.07	-12.07	-9.76	-9.23	-9.65	-10.5	-10.76	-10.87

Protein Family	2.7.1.20	2.7.1.178	2.7.1.19	2.7.1.159	2.7.1.82	2.7.1.73	2.7.1.40	2.7.1.1	2.7.1.47	2.7.1.158	2.7.1.5	2.7.10.2	2.7.10.1	2.7.11.1	2.7.11.21	2.7.11.25
EnzyGen	-8.28	-10.4	-12.43	-10.86	-11.56	-8.01	-10.87	-10.04	-10.85	-11.45	-11.86	-11.65	-9.44	-12.32	-13.13	-11.44

Protein Family	2.7.11.22	2.7.11.30	2.7.11.24	2.7.11.26	2.7.11.2	2.7.11.16	2.7.4.6	2.7.4.3	2.7.4.8	2.7.4.14	2.7.4.9	2.7.4.25	2.7.4.10	2.7.4.22	2.7.4.24	2.7.4.1
EnzyGen	-10.32	-12.53	-13.65	-11.02	-10.83	-12.36	-11.24	-8.76	-9.89	-13.56	-13.36	-10.32	-11.96	-8.99	-8.65	-13.56

Protein Family	2.7.7.7	2.7.7.6	2.7.7.65	2.7.7.60	2.7.7.86	2.7.7.31	2.7.7.12	2.7.7.52	2.7.7.27	2.7.7.23	2.7.7.9	2.7.7.48	3.1.1.4	3.1.1.3	3.1.1.78	3.1.1.2
EnzyGen	-9.02	-8.49	-8.8	-8.35	-8.02	-7.32	-7.02	-7.94	-6.53	-8.48	-7.03	-3.12	-7.79	-8.38	-5.88	-3.32

Protein Family	3.1.1.23	3.1.1.72	3.1.1.47	3.1.1.34	3.1.1.8	3.1.1.7	3.1.1.5	3.1.3.32	3.1.3.46	3.1.3.3	3.1.3.104	3.1.3.16	3.1.3.108	3.1.3.7	3.1.3.11	3.1.3.36
EnzyGen	-5.84	-5.58	-5.44	-5.23	-7.86	-7.33	-12.05	-10.07	-8.25	-6.45	-9.1	-9.36	-9.25	-9.04	-9.13	-9.07

Protein Family	3.1.3.10	3.1.3.76	3.1.3.1	3.1.4.52	3.1.4.17	3.1.4.55	3.1.4.35	3.1.4.53	3.2.2.6	3.2.2.9	3.2.2.16	3.2.2.8	3.2.2.3	3.2.2.1	3.2.2.10	3.4.19.12
EnzyGen	-8.12	-8.42	-7.95	-12.09	-11.35	-11.39	-12.14	-11.86	-10.5	-10.46	-10.56	-9.63	-5.2	-10.93	-10.76	-10.83

Protein Family	3.4.19.5	3.4.19.14	3.4.19.13	3.4.21.53	3.4.21.92	3.4.21.107	3.4.21.22	3.5.1.98	3.5.1.26	3.5.1.97	3.5.1.124	3.5.1.9	3.5.1.44	3.5.1.1	3.5.1.2	3.5.1.25
EnzyGen	-10.2	-9.25	-11.56	-11.95	-10.34	-8.32	-9.87	-5.24	-3.97	-4.24	-7.37	-4.79	-8.98	-6.58	-6.64	-9.46

Protein Family	3.5.1.16	3.5.1.92	3.5.1.99	3.5.1.62	3.5.1.23	3.5.2.17	3.5.2.6	3.5.2.10	3.6.1.12	3.6.1.23	3.6.1.55	3.6.1.56	3.6.1.9	3.6.1.60	3.6.1.1	3.6.1.72
EnzyGen	-9.01	-6.92	-4.23	-6.52	-5.24	-5.7	-6.63	-12.82	-8.7	-10.37	-11.7	-7.86	-4.62	-8.61	-9.44	-8.32

Protein Family	3.6.1.15	3.6.1.58	3.6.1.8	3.6.1.22	3.6.1.27	3.6.1.29	3.6.1.5	3.6.4.6	3.6.4.12	3.6.4.13	3.6.4.10	3.6.5.2	3.6.5.1	3.6.5.5	4.1.1.23	4.1.1.85
EnzyGen	-9.01	-9.87	-9.23	-9.65	-8.52	-8.25	-8.53	-11.12	-13.85	-13.82	-13.41	-11.35	-11.89	-12.56	-11.04	-10.94

Protein Family	4.1.1.17	4.1.1.37	4.1.1.39	4.1.1.28	4.1.1.12	4.1.1.49	4.1.1.25	4.1.1.32	4.2.1.10	4.2.1.19	4.2.1.1	4.2.1.24	4.2.1.22	4.2.1.8	4.2.1.2	4.6.1.1
EnzyGen	-10.17	-10.92	-11.53	-10.36	-10.63	-10.35	-6.32	-7.35	-5.41	-5.44	-5.53	-6.99	-8.65	-9.34	-4.23	-10.7

Protein Family	4.6.1.17	4.6.1.12	4.6.1.2													Avg
EnzyGen	-10.02	-11.76	-10.23													-9.44

Table 9. Substrate binding affinity (kcal/mol, ↓) for the 323 testing fourth-level categories. Avg denotes average.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4	Avg
PROTSEED	-8.37	-6.90	-7.08	-11.69	-6.61	-7.94	-12.26	-11.83	-12.93	-10.09	-8.30	-8.15	-10.09	-12.70	-10.01	–
RFDiffusion+IF	-6.54	-7.39	-7.09	-11.63	-11.89	-12.59	-12.39	-9.49	-14.11	-7.87	-8.81	-10.99	-14.19	-14.71	-11.53	–
ESM2+EGNN	-12.93	-10.03	-8.74	-10.39	-8.20	-14.38	-12.40	-11.41	-14.42	-12.00	-11.01	-8.74	-13.46	-13.32	-12.23	–
EnzyGen	-16.63	-7.20	-6.85	-13.03	-14.72	-12.43	-14.21	-11.84	-16.31	-15.36	-9.23	-13.31	-12.92	-15.01	-13.56	

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	-9.06	-9.79	-8.27	-9.02	-13.22	-9.77	-10.59	-6.70	-7.65	-10.32	-11.32	-13.50	-8.97	-5.74	-10.11	-9.63
RFDiffusion+IF	-9.56	-6.81	-10.50	-10.00	-12.11	-12.35	-12.05	-9.66	-7.22	-13.16	-14.26	-11.10	-9.47	-5.92	-9.84	-10.51
ESM2+EGNN	-11.30	-10.19	-10.85	-10.24	-13.65	-12.42	-12.80	-8.80	-9.61	-12.59	-15.68	-13.29	-10.36	-10.35	-14.61	-11.68
EnzyGen	-11.34	-13.10	-11.36	-12.66	-12.79	-15.25	-11.95	-11.15	-13.08	-12.24	-13.93	-14.17	-11.60	-9.34	-11.76	**-12.61**

Table 10. Substrate binding affinity (kcal/mol, ↓) for the top-1 candidate. IF denotes the inverse folding model ProteinMPNN. Avg denotes average.

Protein Family	1.1.1.239	1.1.1.270	1.1.1.201	1.1.1.184	1.1.1.25	1.1.1.62	1.1.1.197	1.1.1.35	1.1.1.248	1.1.1.271	1.1.1.372	1.1.1.357	1.1.1.27	1.1.1.286	1.1.1.47	1.1.1.3
EnzyGen	92.18	71.92	97.56	94.23	96.24	90.45	92.13	96.45	77.26	94.31	88.15	94.42	88.4	93.72	97.8	95.54

Protein Family	1.1.1.56	1.1.1.105	1.1.1.77	1.1.1.34	1.1.1.399	1.1.1.42	1.1.1.49	1.1.1.44	1.1.1.40	1.11.1.24	1.11.1.12	1.11.1.15	1.11.1.9	1.11.1.11	1.11.1.5	1.11.1.6
EnzyGen	86.04	96.22	98.32	96.49	96.1	92.63	96.29	96.9	96.35	94.71	95.63	78.96	69.85	96.01	96.04	95.77

Protein Family	1.11.1.21	1.14.13.39	1.14.13.9	1.14.13.23	1.14.14.18	1.14.14.47	1.14.14.25	1.14.14.154	1.14.14.1	1.14.14.32	1.14.14.24	1.14.14.26	1.14.14.91	1.2.1.3	1.2.1.50	1.2.1.44
EnzyGen	93.84	93.16	90.26	95.01	90.42	96.8	96.37	94.63	88.46	96.79	89.97	95.54	80.31	93.12	40.98	73.57

Protein Family	1.2.1.12	1.2.1.59	1.2.1.13	1.2.1.11	1.2.1.10	1.2.1.5	1.2.1.47	1.2.1.36	1.2.1.9	1.2.1.26	1.2.1.88	1.2.1.84	1.2.1.89	1.2.1.90	2.1.1.309	2.1.1.37
EnzyGen	97.54	84.78	88.9	96.02	89.11	98.09	95.01	95.72	97.27	97.94	92.61	29.82	90.8	95.95	67.58	58.33

Protein Family	2.1.1.364	2.1.1.221	2.1.1.348	2.1.1.320	2.1.1.6	2.1.1.366	2.1.1.367	2.1.1.244	2.1.1.359	2.1.1.72	2.1.1.220	2.1.1.45	2.1.1.98	2.1.1.79	2.1.1.8	2.1.1.233
EnzyGen	95.73	36.24	93.46	79.4	82.73	95.49	93.45	82.57	88.35	95.56	70.41	97.37	97.51	86.87	33.08	40.74

Protein Family	2.1.1.319	2.1.1.5	2.1.1.354	2.1.1.173	2.3.1.48	2.3.1.57	2.3.1.82	2.3.1.258	2.3.1.87	2.3.1.108	2.3.1.39	2.3.1.286	2.3.1.94	2.3.1.117	2.3.1.180	2.3.1.16
EnzyGen	94.27	74.91	96.81	86.12	85.3	91.39	94.01	95.19	92.16	94.57	83.54	80.12	66.06	89.17	90.06	90.08

Protein Family	2.3.1.97	2.3.1.9	2.3.1.47	2.3.1.26	2.3.1.50	2.3.1.37	2.4.1.186	2.4.1.135	2.4.1.44	2.4.1.255	2.4.1.143	2.4.1.264	2.4.1.182	2.4.1.345	2.4.1.115	2.4.1.15
EnzyGen	93.78	96.91	95.27	90.74	95.23	97.45	87.95	92.44	94.95	94.12	89.15	97.0	93.14	89.82	94.9	92.29

Protein Family	2.4.1.21	2.4.1.41	2.4.1.155	2.4.1.11	2.4.1.225	2.4.1.1	2.4.2.30	2.4.2.2	2.4.2.8	2.4.2.3	2.4.2.1	2.4.2.19	2.4.2.17	2.4.2.21	2.4.2.64	2.4.2.14
EnzyGen	87.5	88.72	87.26	95.22	90.59	94.47	85.37	57.66	86.97	94.28	89.57	94.67	87.78	97.23	84.38	56.9

Protein Family	2.4.2.36	2.5.1.18	2.5.1.31	2.5.1.68	2.5.1.55	2.5.1.16	2.5.1.1	2.5.1.15	2.5.1.79	2.5.1.87	2.5.1.29	2.5.1.59	2.5.1.126	2.5.1.47	2.5.1.46	2.5.1.54
EnzyGen	95.83	90.4	93.98	90.52	96.3	96.63	94.58	87.21	84.97	48.7	93.43	96.64	93.75	97.17	95.03	92.67

Protein Family	2.5.1.6	2.5.1.7	2.5.1.19	2.5.1.26	2.6.1.21	2.6.1.9	2.6.1.16	2.6.1.99	2.6.1.51	2.6.1.63	2.6.1.3	2.6.1.1	2.6.1.57	2.6.1.7	2.6.1.81	2.6.1.73
EnzyGen	85.44	95.85	97.85	94.61	98.39	95.83	92.46	33.65	87.92	97.82	92.37	91.69	63.4	97.38	92.31	94.62

Protein Family	2.6.1.83	2.6.1.62	2.6.1.85	2.6.1.82	2.7.1.107	2.7.1.154	2.7.1.26	2.7.1.71	2.7.1.21	2.7.1.24	2.7.1.113	2.7.1.153	2.7.1.23	2.7.1.35	2.7.1.33	2.7.1.68
EnzyGen	97.21	84.11	65.48	91.0	82.46	80.82	34.34	90.39	86.82	94.29	88.44	85.81	84.82	95.64	90.21	92.01

Protein Family	2.7.1.20	2.7.1.178	2.7.1.19	2.7.1.159	2.7.1.82	2.7.1.73	2.7.1.40	2.7.1.1	2.7.1.47	2.7.1.158	2.7.1.5	2.7.10.2	2.7.10.1	2.7.11.1	2.7.11.21	2.7.11.25
EnzyGen	95.98	84.15	89.49	37.84	82.38	38.89	95.48	89.56	96.63	94.46	96.0	88.84	85.6	87.22	93.53	88.0

Protein Family	2.7.11.22	2.7.11.30	2.7.11.24	2.7.11.26	2.7.11.2	2.7.11.16	2.7.4.6	2.7.4.3	2.7.4.8	2.7.4.14	2.7.4.9	2.7.4.25	2.7.4.10	2.7.4.22	2.7.4.24	2.7.4.1
EnzyGen	89.9	87.49	82.13	95.35	87.32	93.13	96.78	89.4	92.22	93.66	90.56	91.12	76.71	92.77	94.7	72.34

Protein Family	2.7.7.7	2.7.7.6	2.7.7.65	2.7.7.60	2.7.7.86	2.7.7.31	2.7.7.12	2.7.7.52	2.7.7.27	2.7.7.23	2.7.7.9	2.7.7.48	3.1.1.4	3.1.1.3	3.1.1.78	3.1.1.2
EnzyGen	89.35	70.58	59.94	95.05	63.21	95.83	92.19	31.92	93.16	31.28	88.78	97.82	94.83	88.53	54.0	89.44

Protein Family	3.1.1.23	3.1.1.72	3.1.1.47	3.1.1.34	3.1.1.8	3.1.1.7	3.1.1.5	3.1.3.32	3.1.3.46	3.1.3.3	3.1.3.104	3.1.3.16	3.1.3.108	3.1.3.7	3.1.3.11	3.1.3.36
EnzyGen	81.44	98.51	93.26	96.08	94.52	96.32	26.47	50.77	92.35	94.74	42.83	82.73	70.04	78.52	88.95	90.43

Protein Family	3.1.3.10	3.1.3.76	3.1.3.1	3.1.4.52	3.1.4.17	3.1.4.55	3.1.4.35	3.1.4.53	3.2.2.6	3.2.2.9	3.2.2.16	3.2.2.8	3.2.2.3	3.2.2.1	3.2.2.10	3.4.19.12
EnzyGen	88.66	92.35	97.53	82.9	88.1	38.32	96.68	92.31	77.62	85.88	33.32	95.85	94.01	66.52	57.02	65.74

Protein Family	3.4.19.5	3.4.19.14	3.4.19.13	3.4.21.53	3.4.21.92	3.4.21.107	3.4.21.22	3.5.1.98	3.5.1.26	3.5.1.97	3.5.1.124	3.5.1.9	3.5.1.44	3.5.1.1	3.5.1.2	3.5.1.25
EnzyGen	77.49	92.47	89.32	85.78	90.28	91.91	95.52	87.98	88.62	94.71	90.08	45.22	71.21	95.29	96.26	95.9

Protein Family	3.5.1.16	3.5.1.92	3.5.1.99	3.5.1.62	3.5.1.23	3.5.2.17	3.5.2.6	3.5.2.10	3.6.1.12	3.6.1.23	3.6.1.55	3.6.1.56	3.6.1.9	3.6.1.60	3.6.1.72
EnzyGen	48.52	95.56	97.24	88.51	32.35	32.7	90.82	97.73	66.37	91.28	87.33	84.84	93.72	85.65	97.47	42.75

Protein Family	3.6.1.15	3.6.1.58	3.6.1.8	3.6.1.22	3.6.1.27	3.6.1.29	3.6.1.5	3.6.4.6	3.6.4.12	3.6.4.13	3.6.4.10	3.6.5.2	3.6.5.1	3.6.5.5	4.1.1.23	4.1.1.85
EnzyGen	77.06	95.34	61.92	78.6	73.23	57.51	78.07	77.19	69.1	79.55	79.61	84.59	86.29	69.46	84.22	95.33

Protein Family	4.1.1.17	4.1.1.37	4.1.1.39	4.1.1.28	4.1.1.12	4.1.1.49	4.1.1.25	4.1.1.32	4.2.1.10	4.2.1.19	4.2.1.1	4.2.1.24	4.2.1.22	4.2.1.8	4.2.1.2	4.6.1.1
EnzyGen	77.3	78.75	96.5	95.62	91.49	93.31	95.01	82.3	92.92	88.36	91.39	96.7	93.19	97.28	92.94	86.06

Protein Family	4.6.1.17	4.6.1.12	4.6.1.2													Avg
EnzyGen	67.84	95.33	87.6													87.45

Table 11. AlphaFold2 pLDDT (↑) for the 323 testing fourth-level categories. Avg denotes average.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4	Avg
PROTSEED	86.28	82.47	73.36	86.48	91.38	64.90	91.83	74.33	89.81	74.43	72.72	79.61	78.34	70.53	92.71	–
RFDiffusion+IF	83.88	97.98	92.53	84.18	88.08	84.90	76.39	98.00	71.68	76.93	76.79	97.88	77.24	95.80	81.25	–
ESM2+EGNN	96.35	97.19	88.50	92.12	96.64	77.03	95.73	94.09	94.69	96.40	96.44	93.70	94.43	92.31	97.50	–
EnzyGen	96.49	97.81	95.67	98.20	98.60	95.38	95.16	94.18	94.88	96.80	98.49	84.15	92.73	96.03	95.67	–

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	92.04	77.49	79.17	80.26	76.65	96.63	74.94	79.85	66.33	80.00	82.03	86.16	83.35	85.18	79.94	80.97
RFDiffusion+IF	87.40	85.01	70.33	73.15	96.12	94.79	77.33	87.29	54.46	83.95	81.33	69.30	82.97	82.47	78.36	82.93
ESM2+EGNN	93.47	97.15	89.00	81.39	96.87	24.73	79.01	97.69	98.18	42.35	88.55	26.98	90.32	93.36	35.89	84.60
EnzyGen	97.82	98.51	91.55	41.93	97.35	85.52	97.30	98.55	91.58	77.43	96.18	95.08	74.63	97.86	68.76	**91.34**

Table 12. AlphaFold2 pLDDT (↑) for the top-1 candidate. IF denotes the inverse folding model ProteinMPNN. Avg denotes average.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4	Avg
PROTSEED	0.94	0.67	0.41	0.98	0.89	0.99	0.93	0.80	0.98	0.98	0.98	0.98	0.99	0.99	0.95	–
RFDiffusion+IF	0.77	0.95	0.54	0.52	0.85	0.62	0.64	0.63	0.77	0.75	0.86	0.75	0.56	0.93	0.95	–
ESM2+EGNN	0.94	0.97	0.65	0.97	0.97	0.97	0.96	0.96	0.97	0.96	0.97	0.97	0.98	0.98	0.97	–
EnzyGen	0.97	1.00	0.66	1.00	1.00	1.00	0.99	0.97	1.00	0.99	1.00	1.00	1.00	1.00	1.00	

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	0.98	0.78	0.91	0.97	0.92	0.81	0.98	0.85	0.20	0.50	0.97	0.39	0.91	0.99	0.96	0.85
RFDiffusion+IF	0.88	0.57	0.89	0.95	0.70	0.95	0.59	0.76	0.58	0.85	0.83	0.84	0.91	0.95	0.81	0.77
ESM2+EGNN	0.97	0.97	0.92	0.93	0.91	0.87	0.97	0.92	0.32	0.96	0.97	0.79	0.91	0.98	0.97	0.92
EnzyGen	1.00	1.00	1.00	1.00	0.94	0.88	1.00	0.98	0.34	0.99	1.00	0.78	0.99	1.00	0.99	**0.95**

Table 13. Enzyme-substrate ESP score (↑) for the top-5 candidate. IF denotes the inverse folding model ProteinMPNN. Note 0.6 or higher ESP score indicates a positive binding. Avg denotes average.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4	Avg
PROTSEED	-5.99	-5.93	-5.55	-11.17	-6.64	-9.61	-11.83	-8.80	-11.84	-8.57	-6.01	-8.72	-11.30	-12.38	-12.59	–
RFDiffusion+IF	-7.03	-6.43	-6.70	-13.24	-8.42	-12.59	-11.36	-10.01	-12.74	-9.80	-7.99	-11.90	-12.30	-12.82	-12.90	–
ESM2+EGNN	-6.85	-8.46	-7.82	-14.01	-6.97	-12.04	-12.65	-11.80	-15.14	-9.34	-5.87	-10.01	-12.65	-12.20	-11.43	–
EnzyGen	-14.11	-5.63	-6.71	-11.02	-10.05	-10.93	-12.66	-11.68	-13.94	-14.41	-9.14	-12.43	-12.70	-13.80	-12.47	

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	-10.11	-7.00	-9.20	-9.68	-13.02	-10.12	-10.76	-5.53	-6.00	-10.57	-13.73	-11.76	-9.35	-6.74	-11.77	-9.41
RFDiffusion+IF	-10.65	-9.15	-9.70	-10.11	-11.35	-13.10	-11.57	-6.88	-8.09	-11.07	-12.54	-11.45	-9.16	-6.64	-10.55	-10.27
ESM2+EGNN	-10.23	-9.34	-9.77	-9.75	-13.48	-11.00	-11.27	-6.19	-12.08	-8.94	-13.61	-13.25	-9.69	-6.56	-10.84	-10.44
EnzyGen	-9.76	-10.73	-9.97	-12.42	-12.08	-11.58	-11.02	-9.51	-11.09	-11.58	-13.83	-12.90	-11.41	-6.52	-10.64	**-11.22**

Table 14. Substrate binding affinity (kcal/mol, ↓) for the top-5 candidate. IF denotes the inverse folding model ProteinMPNN. Avg denotes average.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4	Avg
PROTSEED	81.07	81.80	74.37	84.17	79.28	72.95	69.86	78.31	80.06	79.16	76.61	79.83	80.33	79.20	84.10	–
RFDiffusion+IF	82.80	93.51	81.12	81.17	82.44	78.82	88.36	83.53	82.32	94.44	78.41	81.93	87.38	81.50	82.43	–
ESM2+EGNN	93.35	91.25	90.00	87.68	95.28	69.96	91.43	86.50	86.93	91.76	93.92	89.53	90.64	88.98	94.51	–
EnzyGen	92.85	92.68	90.25	98.04	98.29	82.15	93.39	93.34	95.11	75.91	98.22	76.31	91.46	95.70	94.52	–

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	80.55	77.08	77.89	79.35	78.42	78.24	78.00	77.22	75.51	83.91	75.16	68.56	82.58	80.92	82.22	78.56
RFDiffusion+IF	81.95	81.99	76.11	81.04	92.59	89.93	78.36	87.08	79.25	87.46	60.63	88.13	75.89	77.25	80.03	82.60
ESM2+EGNN	90.02	93.50	87.60	86.07	94.22	62.27	82.19	87.72	95.26	95.92	76.83	80.24	72.48	83.38	79.12	86.95
EnzyGen	96.80	98.04	89.22	81.27	96.82	71.04	96.64	98.12	80.57	80.78	95.23	80.05	87.88	93.38	90.87	**90.17**

Table 15. AlphaFold2 pLDDT (↑) for the top-5 candidate. IF denotes the inverse folding model ProteinMPNN. Avg denotes average.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4	Avg
PROTSEED	0.90	0.67	0.38	0.98	0.78	0.99	0.82	0.55	0.98	0.96	0.96	0.98	0.99	0.99	0.80	–
RFDiffusion+IF	0.68	0.95	0.54	0.47	0.69	0.54	0.55	0.48	0.71	0.66	0.77	0.67	0.51	0.83	0.89	–
ESM2+EGNN	0.89	0.96	0.59	0.94	0.93	0.96	0.91	0.88	0.95	0.95	0.94	0.95	0.96	0.97	0.95	–
EnzyGen	0.95	1.00	0.64	0.99	0.99	1.00	0.99	0.88	0.99	0.99	0.99	0.99	1.00	1.00	1.00	

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	0.97	0.70	0.90	0.95	0.89	0.68	0.97	0.73	0.14	0.50	0.91	0.37	0.86	0.99	0.94	0.81
RFDiffusion+IF	0.78	0.49	0.85	0.89	0.55	0.92	0.47	0.67	0.53	0.73	0.66	0.76	0.84	0.95	0.60	0.69
ESM2+EGNN	0.95	0.95	0.90	0.90	0.84	0.62	0.94	0.86	0.23	0.93	0.95	0.71	0.89	0.97	0.93	0.88
EnzyGen	0.99	0.99	0.99	1.00	0.91	0.69	1.00	0.94	0.27	0.98	1.00	0.74	0.99	1.00	0.99	**0.93**

Table 16. Enzyme-substrate ESP score (↑) for the top-10 candidate. IF denotes the inverse folding model ProteinMPNN. Note 0.6 or higher ESP score indicates a positive binding. Avg denotes average.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4	Avg
PROTSEED	-4.65	-4.93	-4.76	-11.09	-7.28	-9.46	-11.11	-9.21	-10.29	-7.88	-5.13	-8.38	-10.66	-11.37	-13.20	–
RFDiffusion+IF	-7.26	-5.43	-5.70	-11.15	-6.65	-11.45	-11.20	-9.33	-10.88	-8.84	-7.74	-10.02	-11.60	-11.66	-12.90	–
ESM2+EGNN	-5.38	-7.50	-6.75	-12.81	-8.16	-11.06	-10.54	-9.95	-13.16	-8.24	-5.53	-9.44	-11.88	-12.44	-10.84	–
EnzyGen	-10.49	-5.05	-5.82	-10.34	-8.18	-10.05	-10.38	-11.17	-11.91	-11.93	-8.84	-10.50	-11.81	-12.51	-11.50	–

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	-8.81	-7.01	-8.20	-8.71	-11.37	-9.50	-9.96	-5.21	-5.27	-13.57	-11.53	-10.37	-7.88	-5.24	-9.99	-8.60
RFDiffusion+IF	-9.60	-7.51	-8.70	-8.82	-10.89	-11.41	-10.07	-6.43	-7.72	-10.07	-12.27	-10.87	-8.08	-6.14	-10.46	-9.36
ESM2+EGNN	-8.83	-8.52	-8.92	-8.67	-9.73	-10.16	-10.60	-5.03	-8.67	-8.81	-10.89	-12.41	-8.53	-5.52	-9.45	-9.35
EnzyGen	-9.17	-9.07	-9.62	-12.13	-11.26	-10.72	-10.69	-8.30	-8.66	-11.15	-13.10	-11.35	-11.29	-5.99	-10.64	**-10.12**

Table 17. Substrate binding affinity (kcal/mol, ↓) for the top-10 candidate. IF denotes the inverse folding model ProteinMPNN. Avg denotes average.

Protein Family	1.1.1	1.11.1	1.14.13	1.14.14	1.2.1	2.1.1	2.3.1	2.4.1	2.4.2	2.5.1	2.6.1	2.7.1	2.7.10	2.7.11	2.7.4	Avg
PROTSEED	75.84	81.80	72.90	83.84	77.32	77.66	75.62	77.34	79.14	75.58	74.20	77.36	79.65	78.30	77.38	–
RFDiffusion+IF	75.57	91.36	81.12	87.85	79.33	81.03	85.67	86.62	77.50	88.98	79.52	86.97	86.75	81.85	87.24	–
ESM2+EGNN	93.22	91.68	90.10	86.29	94.90	78.73	90.94	87.93	87.07	91.15	93.99	87.45	90.51	88.13	94.98	–
EnzyGen	90.05	91.30	92.56	97.56	98.08	89.81	88.91	93.67	91.56	86.30	97.75	73.66	90.53	94.81	96.02	–

Protein Family	2.7.7	3.1.1	3.1.3	3.1.4	3.2.2	3.4.19	3.4.21	3.5.1	3.5.2	3.6.1	3.6.4	3.6.5	4.1.1	4.2.1	4.6.1	Avg
PROTSEED	76.90	76.29	77.89	79.41	77.11	72.06	79.91	73.83	79.13	75.16	68.74	83.46	76.33	78.37	85.71	77.47
RFDiffusion+IF	80.92	82.53	81.13	85.91	94.10	91.67	76.09	80.76	87.57	66.85	83.48	82.41	81.11	77.05	77.93	82.90
ESM2+EGNN	89.00	92.74	87.41	86.43	89.05	58.58	82.56	91.94	90.93	85.09	74.46	82.17	88.15	89.26	84.14	87.30
EnzyGen	96.47	97.62	90.15	85.19	95.71	71.34	96.32	97.54	81.33	81.01	92.42	86.13	89.63	94.89	87.32	**90.52**

Table 18. AlphaFold2 pLDDT (↑) for the top-10 candidate. IF denotes the inverse folding model ProteinMPNN. Avg denotes average.

PDB Source ID	Sequence
1KAG	MLPPIFLVGPMGAGKTSVGRELARRLGLEFLDSDREIEERVGLTVAWIFEELGEANFRRREEEVLRELFSLEPPVLATGGGAVTNRKNREFLKRHGTVVYLEVPV EELLRRLRALPLLAKLEEKFRALFERLVALYRAAADIVVRNGLLVNLLVLLVR
5L2P	MPLPPHLEYIIQMMLEKGGKGFTTMSEVEEIRDSLLQSASNTEPVEVDIDKIKETFKTSYGVKARQYFPIDNKAYPVVLYLHGGSWVIGSKFTHDKVCRAITVSC NCKVISVDYRLAPEYKFPAAVYECYDATKWIYENAKKINIDITKIAIAGNSAGGNLAAAVSLLSKEKNIKLKYQILLYPAVSFDLDTKSYKEFADGHFLDTDMIKFI GNNYLFNSKEDKLNPYASPINFADLSNLPPALIVTAEYDPLRDVGEAYANKLLQAGVDVTSIRYKGLVHAFASVLDEIGDTINIMGKLLKEYFK
1L1E	MNLEERFFVLFLDPTQTYNCAYFERREMAEQEAQIAKIDAALGALGPEPGMTLLDIGCGWGATTRRLIEKYDVTVVGLDLAKEQANHVRQLLANSDNLRVRAA MLEKWDFFEEPVDRVVSIGAFEHFGYQDRAALFQKTKEHLKPDGGLLLHHIVVPKRREKQEQLLSPQHPLAKFKKLIKKEIFPGGNLPSYPQLMQAAEKAGWE VTREESLQLDGAYWWDWWATAAAAHADEAIAIQSEFAYEYQHEALAAAAGDFELGYWDIQIFVLQK
8DSO	HHHNSKSESLLLLKKKGLLGFGGKVFYGKWRGQYVVAVKKIRSQSSDESQFIEEVQVMQQLRHENLVQLYGVCTEGKDIYIITEYMAAGLLLYYVKKKKFQM QQILTICKAICSALEYLEEKTFLHHDLLKRNLLLNSEGVAKISDFGLAKFISSSSSTKEPFPPVVLEQSKFSSKSDIWAFGVLMWEEYSYGKMPFEKFTRQETLEHV AQGLRLYPPDIASEPVYKIMKECWHEKADERPSFSILLSNLEELQEELA
5TMP	RGKFIVIEGLDGAGKSTNIDVVVEQLQRDGRREEWVFTREPGGTPLAEKLRELLLTPDDEPDDVDMDTEMVLMFEAARSQLVETVIKPALARGAWVLCDRHDL STYAYQGGGRGLPVAKLQALESFAINGLRPDLTLYLDVPVEIGLRRAKQRGKLERFEQERFQQRVAAVYAGRLERAQQDDSWQTIDATQPLSTVSDAIRTHLQRL QSEL
6W7O	HHLKLTFLFKKGELGGFFGGVFYGKWRGQYEVAVKKIRKQSSSESQFIDEVQVMQQLSHENLVQLYGCCTEGDDIYIITEYMANGSLLQYYLKKQHQFQQSQL LQIAKDIVSALEYLEELTLHHPGLSKRNLLLNSEGVAKITDFGLSRRVLDSDISIISSSFSLLPPIPEEELQQSKFTSKSDIWSFGVLMWEEYSYGKMPWQRFSQQE MEEHIRQQLRLYPPDLASPPVYTIMKECWHEKADERPSFSILLSNLVLLLE
3UYO	GSTKSTSEESHSWYHGPIPREDAERLLVSGIHGSFLVRESESTAGDHSLSLFYEGVKYHYRINTALDGKLYVSEGIKFDTLAELVQHYKTHADGLCYVLSEPCPPV GEEEEEEE
4NI2	QVIAQKHDNVSILFSDIVGFTSLASQCSAQDLVMTLNELFSRFDKLCAENDVYKIETIGDAYCCVSGLPRKEPDHSQQICEMALDDMMEVSSQVKTPHGKPINMRI GIHSGRVVAGVVGLMRRYYCLFGNDVNLANHMESGGVPGKINVPKETYELLKNDFSIEPRRGGSEECPENFPKKIVGYCLFVRRAAA

Table 19. Enzyme sequences for all analyzed cases.