CLIPZyme: Reaction-Conditioned Virtual Screening of Enzymes

Peter G. Mikhael [*1] **Itamar Chinn** [*1] **Regina Barzilay** [1]

Abstract

Computational screening of naturally occurring proteins has the potential to identify efficient catalysts among the hundreds of millions of sequences that remain uncharacterized. Current experimental methods remain time, cost and labor intensive, limiting the number of enzymes they can reasonably screen. In this work, we propose a computational framework for *in silico* enzyme screening. Through a contrastive objective, we train CLIPZyme to encode and align representations of enzyme structures and reaction pairs. With no standard computational baseline, we compare CLIPZyme to existing EC (enzyme commission) predictors applied to virtual enzyme screening and show improved performance in scenarios where limited information on the reaction is available (BEDROC_{85} of 44.69%). Additionally, we evaluate combining EC predictors with CLIPZyme and show its generalization capacity on both unseen reactions and protein clusters.

1. Introduction

Biosynthesis is the method of choice for the production of small molecules due to the cost effectiveness, scalability and sustainability of enzymes (Bornscheuer et al., 2012; Hossack et al., 2023). To find enzymes that can catalyze reactions of interest, practitioners often begin by identifying naturally occurring enzymes to repurpose or optimize. Only 0.23% of UniProt is well studied and efficient enzymes likely lie among the hundreds of millions of sequences that are yet to be explored (Ribeiro et al., 2023). As a result, the ability to computationally identify naturally occurring enzymes for any reaction can provide high quality starting points for enzyme optimization and has the potential to unlock a tremendous number of biosynthesis applications that may otherwise be inaccessible.

In this work, we propose CLIPZyme, a novel method to address the task of virtual enzyme screening by framing it as a retrieval task. Specifically, given a chemical reaction of interest, the aim is to obtain a list of enzyme sequences ranked according to their predicted catalytic activity. In order to identify reaction-enzyme pairs, methods must contend with several unique challenges. First, in some cases, small changes in enzyme structures can lead to a large impact on its activity. Yet in other cases, multiple enzymes with completely different structural domains catalyze the same exact reaction (Ribeiro et al., 2023). Similar principles hold for changes to the molecular structures of the reactants (substrates). This makes the task particularly challenging as methods must capture both extremes. Second, the efficacy of an enzyme is intricately linked to its interaction with the reaction's transition states (Martí et al., 2004; Liu et al., 2021), which are difficult to model. Finally, in addressing the challenge of screening extensive datasets of uncharacterized enzymes, the scalability of computational methods becomes a critical factor.

CLIPZyme is a contrastive learning method for virtual enzyme screening. Originally developed to align between image-caption pairs, CLIP-style training has been successfully extended to model the binding of drugs and peptides to their target protein (Singh et al., 2023; Palepu et al.). Unlike binding, however, the need to achieve transition state stabilization makes enzymatic catalysis a more nuanced process (in fact, very strong binding may inhibit an enzyme). Therefore, in order to represent the transition state, we develop a novel encoding scheme that first models the molecular structures of both substrates and products then simulates a pseudo-transition state using the bond changes of the reaction. To leverage the 3D organization of evolutionarily conserved enzyme domains, we encode AlphaFold-predicted structures (Jumper et al., 2021; Varadi et al., 2022). Since enzyme embeddings can be precomputed efficiently, screening large sets of proteins sequences for a new query reaction is computationally feasible.

Since no standard method currently exists for virtual enzyme screening, we utilize enzyme commission (EC) number prediction as a baseline. Specifically, the EC number

[*]Equal contribution [1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, U.S.A. Correspondence to: Peter G. Mikhael <pgmikhael@csail.mit.edu>, Itamar Chinn <itamarc@csail.mit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

is an expert-defined classification system that categorizes enzymes according to the reactions they catalyze. Each EC number is a four-level code where each level provides progressively finer detail on the catalyzed reaction. For this reason, if a novel reaction is associated with an EC class, EC predictors can be used to identify candidate enzymes matching that EC class.

We establish a screening set of 260,197 enzymes curated from BRENDA, EnzymeMap and CLEAN (Chang et al., 2021; Heid et al., 2023; Yu et al., 2023b). In our evaluation, we adopt the BEDROC metric, as is standard for virtual screening, and set its parameter $\alpha = 85$. This places the most importance on the first $\sim$10,000 ranked enzymes, which constitutes a reasonable experimental screening capacity. We compare CLIPZyme to CLEAN, a state-of-the-art EC prediction model, on the virtual screening task and showcase its superior performance. While CLIPZyme can perform virtual screening without any expert annotations of reactions, methods like CLEAN cannot. We show that even when given some knowledge of a novel reaction's EC class, CLIPZyme is still superior to EC prediction for virtual screening (BEDROC_{85} of 44.69% compared to 25.86%). Additionally, we show that combining CLIPZyme with EC prediction consistently achieves improved results. We also demonstrate that our reaction encoding outperforms alternative encoding schemes. Finally, we test our method on both unannotated reactions in EnzymeMap and a dataset of more challenging reactions involving terpene synthases (Samusevich et al., 2024).

2. Related Work

Reaction representation learning Methods to encode chemical reactions have been developed for a range of different computational tasks. This includes language models operating on reaction SMILES strings (Weininger, 1988; Schwaller et al., 2021) and graph-based methods operating on the individual molecular structures of a reaction or on the condensed graph representations (Jin et al., 2017; Fujita, 1986; Hoonakker et al., 2011). These have shown strong performance on tasks like reaction rate prediction and forward synthesis (Madzhidov et al., 2014; Heid & Green, 2021), but fail to take advantage of the data to effectively learn transition state representations. Models developed explicitly for transition state prediction are trained on simulations of very small molecules and are not scalable to enzymatic reactions (Duan et al., 2023; van Gerwen et al., 2023). In contrast to existing approaches that deterministically featurize bond changes, our method learns the features of these transition states directly from the data.

Catalysis of novel reactions Successful design of enzymes most often begins with finding natural proteins that can subsequently be repurposed or optimized (Seelig & Szostak, 2007; Sarai et al., 2024). One option is to use EC prediction to filter enzyme screening sets. However, EC numbers are predefined by experts and provide a relatively coarse characterization of enzymes. As a result, one EC can capture many different reactions, while none may be able to capture a completely novel reaction. Therefore, filtering large libraries of enzymes by EC may yield impractically large sets of enzymes or none at all. Lastly, state-of-the-art EC predictors still show limited success and mixed results (top F1 scores of 0.5-0.6) (Ayres et al.; Yu et al., 2023b; Ryu et al., 2019; Sanderson et al., 2023; Zhang et al., 2024). In this work, we move away from human-crafted enzyme classes and instead operate directly on molecular and protein structures.

Alternatively, the rational design of a new enzyme or active site requires a thorough understanding of the underlying mechanism (Röthlisberger et al., 2008; Jiang et al., 2008; Yeh et al., 2023; Feehan et al., 2021; Weitzner et al., 2019). While methods for protein sequence and structure generation have shown promise in creating custom folds and strong binders (Watson et al., 2023; Ingraham et al., 2019; Dauparas et al., 2022), unnatural enzymes still suffer from low activity relative to naturally occurring ones (Hossack et al., 2023). Instead, we focus on identifying natural protein leads that can be optimized further either computationally or experimentally (Seelig & Szostak, 2007; Bornscheuer et al., 2012; Sarai et al., 2024).

3. Method

We formulate enzyme screening as a retrieval task, where we have access to a predefined list of proteins and are asked to order them according to their ability to catalyze a specific chemical reaction. The representation of a protein P is denoted by $p \in \mathbb{R}^d$ and the query reaction R by $r \in \mathbb{R}^d$. We aim to learn a scoring function $s(r, p)$ such that a higher score corresponds to a higher likelihood that P catalyzes R. We jointly learn a reaction encoder, f_{rxn}, and a protein encoder, f_p, to compute r and p (Figure 1). We adopt a contrastive learning objective (Sohn, 2016; Radford et al., 2021) to maximize the cosine similarity between the embeddings of biochemical reactions and their associated enzymes (Equations (1) and (2)). We treat all enzymes in a training batch that are not annotated to catalyze a reaction as negative samples. Implementations details are provided in Appendices C and D.

$$s_{ij} = s(r_i, p_j) = \frac{r_i}{\|r_i\|} \cdot \frac{p_j}{\|p_j\|} \tag{1}$$

$$\mathcal{L}_{ij} = -\frac{1}{2N}\left(\log \frac{e^{s_{ij}/\tau}}{\sum_i e^{s_{ij}/\tau}} + \log \frac{e^{s_{ij}/\tau}}{\sum_j e^{s_{ij}/\tau}}\right) \tag{2}$$

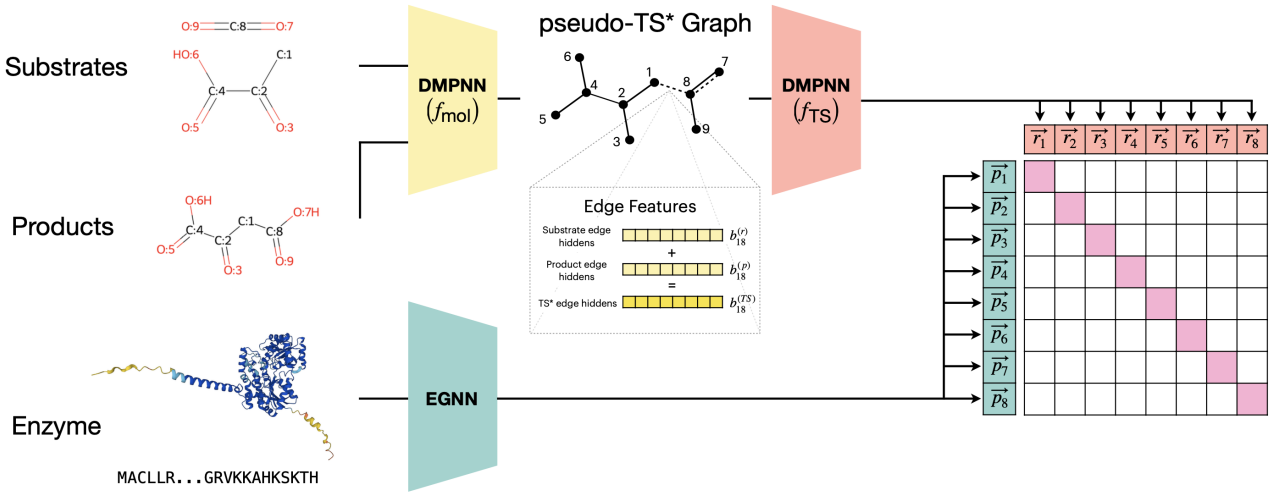

Figure 1. Overview of our approach. We encode atom-mapped chemical reactions using a DMPNN. We combine the substrate and product graphs by adding the hidden embeddings of their corresponding bonds to obtain an intermediate graph representing a pseudo transition state. A second DMPNN computes an embedding for the entire reaction. Enzymes are encoded with an EGNN using their predicted crystal structure and ESM-2 sequence embeddings. The reaction and enzyme representations are aligned with a CLIP objective.

3.1. Chemical Reaction Representation

To obtain a functionally meaningful representation of the reaction, we leverage the key insight that the active sites of enzymes have evolved to stabilize the transition state(s) of their corresponding reactions (Casadevall et al., 2023). As a result, there is a geometric complementarity between the 3D shape of the protein active site and the molecular structure of the transition state. This complementarity determines to a large extent the catalytic activity of enzymes (Martí et al., 2004; Liu et al., 2021). While we do not have access to ground truth or predicted transition states, we use the atom-mapping available in the dataset to learn a superposition of the reactant and product molecular graphs and obtain the reaction embedding.

Specifically, reactants and products are constructed as 2D graphs, where each molecular graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ has atom (node) features $v_i \in \mathcal{V}$ and bond (edge) features $e_{ij} \in \mathcal{E}$. A directed message-passing neural network (DMPNN) (Yang et al., 2019), f_{mol}, is used to separately encode the graph of the reactants G_x and that of the products G_y. This results in learned atom and bond features $a_i, b_{ij} \in \mathbb{R}^d$. To simulate the transition state, we construct a pseudo-transition state graph, $G_{TS} = (\mathcal{V}_{TS}, \mathcal{E}_{TS})$, by adding the bond features for edges connecting the same pairs of nodes in the reactants and the products. Edges between atom pairs that are not connected have bond features set to zero. We use the original atom features v_i as the nodes of $\mathcal{G}_{TS}$ to preserve the atom identities.

$$a_i, b_{ij} = f_{\text{mol}}(\mathcal{G}_x, \mathcal{G}_y) \tag{3}$$

$$v_i^{(TS)} := v_i^{(x)} \quad (= v_i^{(y)}) \tag{4}$$

$$e_{ij}^{(TS)} := b_{ij}^{(x)} + b_{ij}^{(y)} \tag{5}$$

We jointly train a second DMPNN, f_{TS} to encode G_{TS} and obtain the reaction embedding r by aggregating the learned node features.

$$a'_i, b'_{ij} = f_{\text{TS}}(\mathcal{G}_{TS}) \tag{6}$$

$$r = \sum_i a'_i \tag{7}$$

3.2. Protein Representation

Enzyme representation plays a pivotal role in modeling their function and interaction with substrates. To this end, we leverage advancements in both protein language models and graph neural networks.

Each protein is represented as a 3D graph $\mathcal{G}_p = (\mathcal{V}, \mathcal{E})$, with residue (node) features $h_i \in \mathcal{V}$ and bond (edge) features $e_{ij} \in \mathcal{E}$. Additionally each node i has coordinates $c_i \in \mathbb{R}^3$. The node features of $\mathcal{G}_p$ are initialized using embeddings from the ESM-2 model with 650 million parameters (`esm2_t33_650M_UR50D`) (Lin et al., 2022), which has demonstrated success in capturing many relevant protein features for a range of downstream tasks. The ESM model produces a feature vector for each residue denoted as $h \in \mathbb{R}^{1280}$.

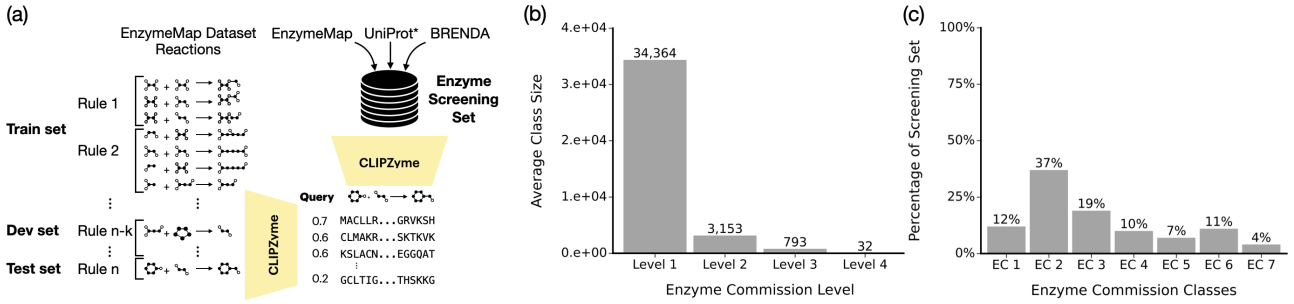

Figure 2. Overview of dataset construction and statistics. **(a)** Reaction-enzyme pairs are obtained from the EnzymeMap dataset (Heid et al., 2023) and split based on their reaction rules. At test time, a reaction is queried and enzymes are ranked from a screening set composed of sequences from EnzymeMap, UniProt*, and BRENDA. **(b)** Averge number of sequences in each EC class when considering different levels of the EC hierarchy. **(c)** Distribution of sequences in the screening set according to their first EC level.
*The UniProt dataset is obtained from Yu et al. (2023b).

To encode the protein graphs, we utilize an $E(n)$-Equivariant Graph Neural Network (EGNN) with coordinate updates (Satorras et al., 2021). This network is particularly suited for our purpose as it preserves translation, rotation and reflection equivariant graph features but is computationally inexpensive. Alternative methods preserve additional symmetries that are relevant to proteins such as SE(3) equivariance but are much more computationally expensive. We follow the implementation outlined in Satorras et al. (2021) except that the relative distances between nodes are encoded using a sinusoidal function (Appendix C), as is common in protein structure modeling (Aykent & Xia, 2022; Atz et al., 2022; Vaswani et al., 2017).

4. Experimental Setup

4.1. EnzymeMap Dataset

Our method is developed on the EnzymeMap dataset (Heid et al., 2023), which includes biochemical reactions linked with associated UniProt IDs and their respective EC numbers. Each reaction is atom-mapped, allowing every atom in the product to be traced back to a corresponding atom in the reactants. To acquire the corresponding protein sequences, we select reactions linked to UniProt or SwissProt IDs and retrieve their sequences from their respective databases (uni, 2023). Additionally, we retrieve the predicted enzyme structures from the AlphaFold Protein Structure Database (Jumper et al., 2021; Varadi et al., 2022). We filter samples to include protein sequences up to 650 amino acids in length only. EnzymeMap provides a reaction rule for each reaction, which captures the bio-transformation in a reaction and can be applied to recreate the products of a reaction from its substrates (Ni et al., 2021). With the goal of extending our model to unfamiliar chemical reactions, we divide our dataset into training, development, and testing groups based on these reaction rules (Figure 2). This results in a total of 46,356 enzyme-driven reactions, encompassing 16,776 distinct chemical reactions, 12,749 enzymes, across 2,841 EC numbers and 394 reaction rules. We note that ECs can share chemical rules. For instance, the test set covers 319 unique ECs, 88 of which are included in the training set.

EnzymeMap includes additional reactions that are associated with an EC number but lack an annotated protein sequence. We identify 7,967 of these unannotated reactions involving 1,101 EC numbers, distinct from our training data in terms of reaction rules. This subset serves as an additional validation set, allowing us to evaluate how our method ranks enzymes in relation to the EC number for each reaction. More information on how the data was processed can be found in Appendix B.

4.2. Terpene Synthase Dataset

Terpenoids are a large and diverse family of biomolecules with wide applications to medicine and consumer goods. The reactions generating these natural compounds involve particularly complex chemical transformations that are typically catalyzed by a class of enzymes called terpene synthases (Samusevich et al., 2024). This enzyme class is noteworthy for utilizing a relatively small number of substrates ($\sim$11) but is capable of generating thousands of distinct products. This presents a significant challenge with substantial implications. To further evaluate our method's performance on reactions known to involve challenging chemistry, we use a dataset of terpenoid reactions made available by recent work in detecting novel terpene synthases (Samusevich et al., 2024). We exclude reactions that are themselves or their enzyme included in our training set, obtaining 110 unique reactions and 99 enzymes.

4.3. Enzyme Screening Set

To construct our screening set of enzymes, we include sequences annotated in the EnzymeMap dataset (Heid et al., 2023), Brenda release `2022_2` (Chang et al., 2021), and those used in developing CLEAN (UniProt release `2022_01`) (Yu et al., 2023b). We filter our set to those of sequence length < 650 with available AlphaFold predicted structures (Jumper et al., 2021; Varadi et al., 2022) and obtain a final list of 260,197 sequences.

4.4. Baselines

4.4.1. Ranking Enzymes via EC Prediction

The ultimate goal of enzyme screening is to identify candidate proteins from large protein databases, including the hundreds of millions of unannotated sequences. Since no standard computational procedure for enzyme screening has emerged, a reasonable approach is to first assign an EC number to the query reaction and then select all enzymes that share that EC class. To identify the EC classes of the enzymes in the screening set, one can use an EC predictor. On the other hand, assigning the full EC number of a reaction is not always straightforward or possible. For this reason, we consider baselines where between 1 to 4 levels of a query reaction's EC number are assignable (e.g., 1 level: 1.x.x.x to 4 levels: 1.2.3.4). We evaluate EC prediction and CLIPZyme on ranking the enzymes screening set for each reaction in the EnzymeMap test set.

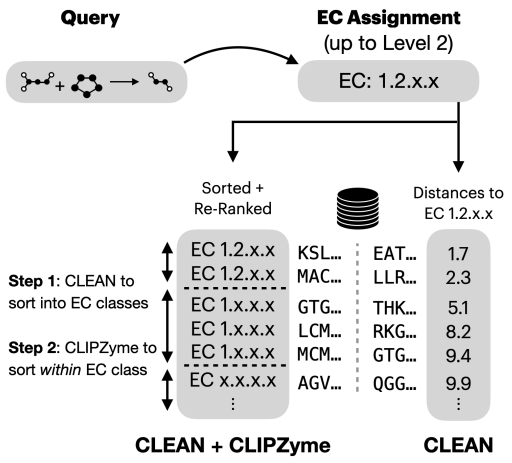

Figure 3. Approaches for adapting EC prediction to virtual enzyme screening. We first assign a reaction an EC up to some level of specificity (here, level 2). To obtain rankings based on CLEAN, we use each sequence's distance to the EC class. To combine CLEAN and CLIPZyme, enzymes are first sorted according to their predicted EC class. Then they are ranked within each class using CLIPZyme.

We use CLEAN, a state-of-the-art EC predictor, to obtain a ranked list of enzymes for each EC (Yu et al., 2023b). CLEAN computes a single representation for every EC in its dataset as the mean embedding of sequences in that class and uses these as test-time anchors. The predicted EC class of a new sequence is then determined by the Euclidean distance to each EC anchor. Accordingly, given a reaction's assigned EC number, we rank our screening set enzymes by their distances to the reaction's EC anchor (Figure 3). If a reaction's EC class does not exist in the CLEAN dataset, we broaden the search to one level higher. As an example, for a reaction with EC 1.2.3.4, if this EC is not in the CLEAN dataset, we rank enzymes according to their distances to the mean representation of EC 1.2.3 (and so on). For consistency with previous work, we use the CLEAN model trained on a split where none of the test enzymes share more than 50% sequence identity with those in the training set (Yu et al., 2023b).

We hypothesize that combining CLEAN to obtain EC predictions and CLIPZyme to rank them presents an opportunity for improved performance. Specifically, we predict the EC numbers for all of the enzyme sequences in our screening set using CLEAN. Given the reaction's assigned EC number, we first filter our screening set to those enzymes with the same exact predicted EC and rank this list using CLIPZyme (Figure 3). We then filter all remaining enzymes to those that belong to one EC level above and again rank that list using CLIPZyme. As an example, given an input reaction with assigned EC of 1.2.3.4, we identify all enzymes predicted for that EC and rank them with CLIPZyme. We then rank all remaining enzymes with predicted EC 1.2.3. This process is repeated until all enzymes are ranked.

4.4.2. Reaction Representation

We explore three alternative methods for encoding the reaction and compare against these in our results. The first uses the condensed graph reaction (CGR) representation (Hoonakker et al., 2011) by overlaying the reactants and products and concatenating the edge features. A DMPNN encodes the CGR to obtain a hidden representation of the reaction. The second approach is to use the full reaction SMILES (Weininger, 1988) as an input to a language model and obtain a final representation of the reaction. We follow the tokenization scheme for SMILES introduced by the Molecular Transformer (Schwaller et al., 2019) and train a transformer model as our encoder (Vaswani et al., 2017). We also consider the Weisfeiler-Lehman Difference Network (WLDN) architecture and implement it as described in Jin et al. (2017). We train all models until convergence, using the same data splits and hyper-parameters (Appendix D).

Table 1. Enzyme virtual screening performance compared to using EC prediction alone and together with CLIPZyme. For a given reaction EC level, enzymes are ranked according to their Euclidean distance to EC class anchors when using CLEAN (Yu et al., 2023b). Alternatively, CLEAN is first used to place enzymes into successively broader EC levels matching that of the reaction, and CLIPZyme is used to reorder the enzymes within each level. BEDROC: Boltzmann-enhanced discrimination of receiver operating characteristic; EF: enrichment factor.

EC Level Assumed Available	Method	$BEDROC_{85}$ (%)	$BEDROC_{20}$ (%)	$EF_{0.05}$	$EF_{0.1}$
-	CLIPZyme	44.69	62.98	14.09	8.06
Level 1 (x.-.-.-)	CLEAN	0.96	6.53	1.22	1.72
	CLIPZyme + CLEAN	57.03	78.50	17.84	9.56
Level 2 (x.x.-.-)	CLEAN	4.86	14.10	3.23	2.49
	CLIPZyme + CLEAN	75.57	90.20	19.40	9.84
Level 3 (x.x.x.-)	CLEAN	25.86	36.75	8.03	4.81
	CLIPZyme + CLEAN	82.69	93.23	19.43	9.84
Level 4 (x.x.x.x)	CLEAN	89.74	93.42	18.97	9.60
	CLIPZyme + CLEAN	89.57	95.24	19.43	9.84

4.4.3. Protein Representation

We focus on achieving a balance between efficiency and the ability to process extensive enzyme datasets. To this end, we explore both sequence-based and structure-based approaches, acknowledging the critical influence of structure on enzymatic activity despite its inherent computational demands. We train ESM-2 (Lin et al., 2022) as a sequence-based baseline for protein encoding. We also encode the structure with an EGNN (Satorras et al., 2021) and compare initializing node embeddings from either the MSA-transformer (Rao et al., 2021) or ESM-2, to identify the best method in terms of both performance and speed.

4.5. Evaluation Setup

We aim to simulate the scenario where an enzyme is desired to catalyze a novel reaction, and it exists in nature but is not annotated. We compare different approaches to encoding the reactions and their enzymes, and compare our method to an alternative approach using EC prediction.

As our main aim is to generalize to novel chemical transformations, our test set consists of reactions with reaction rules that are unseen during training, queried against all 260,197 sequences. However, this means our screening set does include proteins used in training the model. Therefore, we also evaluate model performance when excluding proteins used in training. Additionally, we use MMSeqs2 (Steinegger & Söding, 2017) and Foldseek (van Kempen et al., 2023) to exclude proteins based on their similarity to the training set proteins in terms of sequence identity and protein fold, respectively. If the exclusion of a protein results in a test reaction having no actives in the screening set, we exclude the entire reaction.

Throughout our evaluations, we take the BEDROC score as our primary metric (Truchon & Bayly, 2007). We focus on the case $\alpha = 85$, where the top 3.5% of predictions contribute to 95% of the score, and as suggested in Truchon & Bayly (2007), we also calculate the BEDROC score for $\alpha = 20$. We also report the enrichment factor (EF) when taking the top 5% and 10% of predictions. Depending on the experimental context, the capacity to screen a diverse set of enzyme sequences can be further limited, so we also provide in Appendix A.5 results focusing on the top $\sim$1,000 predictions with $\alpha = 748.9$ and 173.3 and EF values for the top 0.5% and 1%. This estimates the fraction of catalyzing enzymes found in our top predictions relative to random selection.

5. Results

We present here an overview of our key results. In Table 1, we compare CLIPZyme's performance to that of EC prediction with CLEAN and show the benefit of combining methods. CLIPZyme shows improved performance in all comparisons. Table 2 shows the impact of different protein and reaction representations and highlights the superior performance of our novel reaction encoding. In Table 3, we show that CLIPZyme's performance extends to a challenging dataset of terpene synthase reactions and unannotated reactions. Lastly, we show in Table 4 that CLIPZyme's performance drops when screening enzymes that significantly differ from those it was trained on, but still maintains useful predictive value. Additional analysis, including alternative baselines, is provided in Appendix A.

Table 2. Performance of various protein and reaction encoding schemes on virtual screening for reactions in the EnzymeMap test set. The symbol ❄ denotes models where the weights are kept unchanged during training.

PROTEIN ENCODER	REACTION ENCODER	$BEDROC_{85}$(%)	$BEDROC_{20}$(%)	$EF_{0.05}$	$EF_{0.1}$
ESM❄	OURS (SECTION 3.1)	17.84	29.39	6.61	4.17
ESM	OURS (SECTION 3.1)	36.91	53.04	11.93	6.84
MSA-TRANSFORMER❄ + EGNN	OURS (SECTION 3.1)	28.76	46.53	10.34	6.67
ESM❄ + EGNN	CGR (HOONAKKER ET AL., 2011)	38.91	57.58	13.16	7.73
ESM❄ + EGNN	REACTION SMILES	29.94	46.01	10.34	6.32
ESM❄ + EGNN	WLDN (JIN ET AL., 2017)	29.84	46.70	10.71	6.41
ESM + EGNN	OURS (SECTION 3.1)	33.68	50.08	11.04	6.77
ESM❄ + EGNN	OURS (SECTION 3.1)	44.69	62.98	14.09	8.06

5.1. Enzyme Screening Evaluation on EnzymeMap

CLIPZyme effectively ranks the screening set against reactions in the EnzymeMap test set with an average $BEDROC_{85}$ of 44.69% and an enrichment factor of 14.09 when choosing the top 5% (Table 1). We compare CLIPZyme's performance alone (first row) to ranking using EC prediction with CLEAN. Since it is not always possible to assign all 4 levels of an EC to a chemical reaction, we examine scenarios where different EC levels are assumed to be known for query reactions in the test set.

For example, with only the first EC level known, using EC prediction alone obtains a $BEDROC_{85}$ score of 0.96% (Table 1). This improves to 25.86% when we are able to specify a reaction up to the third EC level. With four EC levels known, the CLEAN method becomes more effective than CLIPZyme alone. However, being able to assign all four EC levels for a reaction may not be always feasible in real-world applications.

Combining the CLEAN method with CLIPZyme achieves improved performance regardless of how many EC levels we assume to be known for reactions. Here, CLEAN is first used to predict the EC classes of enzymes in the screening set. Enzymes within the predicted EC class are re-ranked using CLIPZyme (Figure 3). Even basic knowledge of the first EC level of a chemical reaction enhances CLIPZyme's performance from a $BEDROC_{85}$ of 44.69% to 57.03%. With the first two levels assumed to be known, performance also improves to 75.57%. We observe that the same patterns hold when focusing only on the top ∼1,000 hits (Appendix A.5).

We note that EC classification may be insufficient for categorizing chemical reactions that do not fit in existing EC classes. As a result, any EC prediction method is not applicable in that setting, while CLIPZyme is as it operates directly on the reaction.

5.2. Impact of Reaction and Protein Representation

We explore a number of different encoding methods for both reaction and protein representations and find that the model is highly sensitive to changes in both (Table 2). Using the molecular structures of the reaction obtains better performance than language-based methods operating over the reaction SMILES, with the former achieving a $BEDROC_{85}$ of 44.69% compared to 29.94%. This suggests that structural representations may capture chemical transformations that correspond to enzyme activity more explicitly than language based ones. The patterns observed in structures may be more difficult for language models to capture without additional features or data. Employing a more expressive model also improves performance when compared to using WLDN as the reaction encoder. While all reaction representation methods include the full reaction, they differ in how the bond changes are encoded. Methods that explicitly delineate chemical transformations between substrates and products appear to obtain generally better performance.

Table 3. Performance of CLIPZyme on additional biochemical reactions. The terpene synthase dataset is obtained from Samusevich et al. (2024) and includes reactions considered to involve more complex biotransformations. The unannotated subset of EnzymeMap consists of reactions in the dataset that are not assigned a UniProt or SwissProt identifier. In this case, virtual screening is evaluated as the ability to highly rank proteins with the correct EC class.

DATASET	$BEDROC_{85}$ (%)	$BEDROC_{20}$ (%)	$EF_{0.05}$	$EF_{0.1}$
TERPENE SYNTHASES	72.46	85.89	18.29	9.42
UNANNOTATED ENZYMEMAP	42.94	61.39	13.92	7.73

We find a similar sensitivity to enzyme encoding. We compare using ESM embeddings alone and using ESM embeddings together as node features for EGNN. We find that using an EGNN to capture the structural components of the enzyme improves performance compared to training a sequence-based model alone (44.69% compared to

Table 4. Performance when excluding sequences from the screening set with various levels of similarity to training set enzymes.

EXCLUSION CRITERIA	$BEDROC_{85}$ (%)	$BEDROC_{20}$ (%)	$EF_{0.05}$	$EF_{0.1}$
EXACT MATCH	39.13	58.86	13.40	7.81
MMSEQS 30% SIMILARITY	35.32	54.86	12.43	7.30
FOLDSEEK 30% SIMILARITY	21.44	35.39	7.93	4.93

36.91%), which indicates that enzyme structure is important for achieving good performance on this task. We also explore initializing the EGNN node features with embeddings from the pre-trained MSA-Transformer (Rao et al., 2021). These embeddings do not appear to improve performance, although they capture evolutionary information of the sequence. This, however, may be due to differences in quality of representations learned by ESM and MSA-Transformer in which ESM-2 was trained on much larger set of sequences. Lastly, we also evaluate the effect of fine-tuning ESM and find that it leads to worse performance compared to fixing the ESM weights and using its learned representations. We hypothesize that this may be due to overfitting of ESM on a relatively small dataset of proteins with $\sim$12K unique proteins.

5.3. Evaluation on Reaction-Specific Datasets

We extend our evaluation to two additional datasets to further assess CLIPZyme's utility in practical applications in Table 3. The first dataset encompasses reactions catalyzed by terpene synthases. We evaluated CLIPZyme using the same screening set and observed robust performance, evidenced by a $BEDROC_{85}$ score of 72.45%. Due to the small and uniform substrate pool, the model might be preferentially ranking terpene synthases as a whole, rather than effectively distinguishing between specific reactions.

Additionally, we present an evaluation using unannotated reactions from EnzymeMap. For the sake of evaluation, we assume the true enzymes in the screening set for a given reaction are those with EC classes matching that of the reaction. Under this setup, CLIPZyme achieves a $BEDROC_{85}$ of 42.94%, which aligns closely with the results from the annotated subset of EnzymeMap. Because the metrics are calculated relative to the EC classes of each protein, this result suggests that the CLIPZyme rankings correspond with the proteins' EC numbers.

5.4. Generalization to Novel Proteins

Our primary focus has been on evaluating the generalization of CLIPZyme on reactions unseen during training. However, given the ultimate goal of screening a wide array of both annotated and unannotated enzymes, it's crucial to understand the model's efficacy in ranking proteins dissimilar to those in the training set.

To do so, we exclude proteins that are similar to our training set according to three similarity metrics. We first exclude training set enzymes. Second, we apply MMSeqs2 (Steinegger & Söding, 2017) to remove enzymes with 30% or greater sequence similarity. Lastly, we exclude enzymes with 30% fold similarity as determined by Foldseek (van Kempen et al., 2023). By measuring performance on these three screening subsets, we demonstrate CLIPZyme's generalizability across both reactions and enzymes.

Each exclusion criteria led to a reduction in performance. For example, CLIPZyme's performance decreases by approximately 5 percentage points on both BEDROC metrics when excluding training set enzymes Table 4. The most marked impact was observed with Foldseek-based filtering, showing a 23.25 point decrease in $BEDROC_{85}$ scores. This aligns with our previous findings that protein structural features play a critical role in effective screening. Despite this, the model still demonstrated a notable ability to rank enzymes effectively as the top-ranked candidates consistently showed enrichment for active enzymes.

6. Conclusion

We present here the task of virtual enzyme screening and a contrastive method, CLIPZyme, to address it. We show that our method can preferentially rank catalytically active enzymes against reactions across multiple datasets. Without a standard baseline, we examine enzyme screening through EC prediction and highlight CLIPZyme's competitive ability. We furthermore show that combining EC prediction with CLIPZyme achieves significantly improved performance. Lastly, we evaluate CLIPZyme's capacity to generalize by evaluating it on additional challenging reaction datasets and on unseen protein clusters. In practical scenarios, where millions or even hundreds of millions of enzymes need screening, we foresee the necessity of methods like CLIPZyme with even higher sensitivity for effective enzyme design at scale.

Among its limitations, the current approach does not model the physical interactions between reactants and enzymes, and it is unable to capture the mechanisms that give rise to the observed reaction. Moreover, the available data covers a

relatively small chemical space and includes a restricted set of reactions and enzyme sequences (e.g., EC class 7 is completely unrepresented). We also note that our approach of random negative sampling may give rise to false negatives due to the promiscuity of many enzymes and the method may benefit from alternative sampling techniques. Directions for future work include modeling the 3D interactions characterizing biochemical reactions (e.g., through docking) and leveraging transition state sampling through quantum chemical simulations.

Software and Data

We downloaded EnzymeMap from https://github.com/hesther/enzymemap, and the Terpene Synthase data from https://zenodo.org/records/10359046. We make our code and data available at https://github.com/pgmikhael/clipzyme.

Acknowledgements

We are grateful to Jeremy Wohlwend, Anush Chiappino-Pepe, Tomáš Pluskal, Raman Samusevich, Esther Heid, Hannes Stark, and Bowen Jing for their insightful discussions. This work is supported by the Jameel Clinic for AI and Health at MIT, the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard, Novo Nordisk A/S, the Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS) consortium, the DTRA Discovery of Medical Countermeasures Against New and Emerging (DOMANE) Threats program, the DARPA Accelerated Molecular Discovery program, and the NSF Expeditions grant (award 1918839: Collaborative Research: Understanding the World Through Code).

Impact Statement

The ability to identify, repurpose, or create enzymes for novel reactions remains a grand challenge that has profound societal impact through applications ranging from therapeutic manufacturing to the biodegradation of plastics. By enriching the pool of candidate enzymes through virtual screening, experimental approaches to optimize enzymes gain a greater chance of achieving efficient catalysis. As with all computational methods that facilitate the generation of small molecules, successful virtual enzyme screening may also be exploited for the production of hazardous chemicals.

References

Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.

Atz, K., Isert, C., Böcker, M. N., Jiménez-Luna, J., and Schneider, G. δ-quantum machine-learning for medicinal chemistry. *Physical Chemistry Chemical Physics*, 24(18): 10775–10783, 2022.

Aykent, S. and Xia, T. Gbpnet: Universal geometric representation learning on protein structures. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4–14, 2022.

Ayres, G., Munsamy, G., Heinzinger, M., Ferruz, N., Yang, K., and Lorenz, P. Hifi-nn annotates the microbial dark matter with enzyme commission numbers.

Bornscheuer, U. T., Huisman, G., Kazlauskas, R., Lutz, S., Moore, J., and Robins, K. Engineering the third wave of biocatalysis. *Nature*, 485(7397):185–194, 2012.

Casadevall, G., Duran, C., and Osuna, S. Alphafold2 and deep learning for elucidating enzyme conformational flexibility and its application for design. *JACS Au*, 3(6):1554–1562, 2023.

Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D., and Schomburg, D. Brenda, the elixir core data resource in 2021: new developments and updates. *Nucleic acids research*, 49(D1):D498–D508, 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.

Duan, C., Du, Y., Jia, H., and Kulik, H. J. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. *arXiv preprint arXiv:2304.06174*, 2023.

Falcon, W. and The PyTorch Lightning team. PyTorch Lightning, March 2019. URL https://github.com/Lightning-AI/lightning.

Fan, H., Wang, Z., Yang, Y., and Kankanhalli, M. Continuous-discrete convolution for geometry-sequence modeling in proteins. In *The Eleventh International Conference on Learning Representations*, 2022.

Feehan, R., Montezano, D., and Slusky, J. S. Machine learning for enzyme engineering, selection and design. *Protein Engineering, Design and Selection*, 34:gzab019, 2021.

Fey, M. and Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Fujita, S. Description of organic reactions based on imaginary transition structures. 1. introduction of new concepts. *Journal of Chemical Information and Computer Sciences*, 26(4):205–212, 1986.

Heid, E. and Green, W. H. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *Journal of Chemical Information and Modeling*, 62(9):2101–2110, 2021.

Heid, E., Probst, D., Green, W. H., and Madsen, G. K. Enzymemap: Curation, validation and data-driven prediction of enzymatic reactions. 2023.

Hoonakker, F., Lachiche, N., Varnek, A., and Wagner, A. Condensed graph of reaction: considering a chemical reaction as one single pseudo molecule. *Int. J. Artif. Intell. Tools*, 20(2):253–270, 2011.

Hossack, E. J., Hardy, F. J., and Green, A. P. Building enzymes through design and evolution. *ACS Catalysis*, 13(19):12436–12444, 2023.

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

Jamasb, A. R., Morehead, A., Joshi, C. K., Zhang, Z., Didi, K., Mathis, S. V., Harris, C., Tang, J., Cheng, J., Liò, P., et al. Evaluating representation learning on the protein structure universe. In *The twelfth international conference on learning representations*, 2024.

Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas III, C. F., et al. De novo computational design of retro-aldol enzymes. *science*, 319(5868): 1387–1391, 2008.

Jin, W., Coley, C., Barzilay, R., and Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in neural information processing systems*, 30, 2017.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., and Dror, R. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2020.

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.

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*, 2022.

Liu, M., Nazemi, A., Taylor, M. G., Nandy, A., Duan, C., Steeves, A. H., and Kulik, H. J. Large-scale screening reveals that geometric structure matters more than electronic structure in the bioinspired catalyst design of formate dehydrogenase mimics. *ACS Catalysis*, 12(1): 383–396, 2021.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Madzhidov, T., Polishchuk, P., Nugmanov, R., Bodrov, A., Lin, A., Baskin, I., Varnek, A., and Antipin, I. Structure-reactivity relationships in terms of the condensed graphs of reactions. *Russian Journal of Organic Chemistry*, 50: 459–463, 2014.

Martí, S., Roca, M., Andrés, J., Moliner, V., Silla, E., Tuñón, I., and Bertrán, J. Theoretical insights in enzyme catalysis. *Chemical Society Reviews*, 33(2):98–107, 2004.

Mirdita, M., Von Den Driesch, L., Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research*, 45(D1):D170–D176, 2017.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Ni, Z., Stine, A. E., Tyo, K. E., and Broadbelt, L. J. Curating a comprehensive set of enzymatic reaction rules for efficient novel biosynthetic pathway design. *Metabolic Engineering*, 65:79–87, 2021.

Palepu, K., Ponnapati, M., Bhat, S., Tysinger, E., Stan, T., Brixi, G., Koseki, S., and Chatterjee, P. Design of peptide-based protein degraders via contrastive deep learning. biorxiv 2022. *Google Scholar*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P., Sercu, T., and Rives, A. Msa transformer. *bioRxiv*, 2021. doi: 10.1101/2021.02.12.430858. URL https://www.biorxiv.org/content/10.1101/2021.02.12.430858v1.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173–175, 2012.

Ribeiro, A. J., Riziotis, I. G., Borkakoti, N., and Thornton, J. M. Enzyme function and evolution through the lens of bioinformatics. *Biochemical Journal*, 480(22):1845–1863, 2023.

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

Ryu, J. Y., Kim, H. U., and Lee, S. Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proceedings of the National Academy of Sciences*, 116(28):13996–14001, 2019.

Samusevich, R., Hebra, T., Bushuiev, R., Bushuiev, A., Chatpatanasiri, R., Kulhánek, J., Čalounová, T., Perković, M., Engst, M., Tajovská, A., Sivic, J., and Pluskal, T. Discovery and characterization of terpene synthases powered by machine learning. *bioRxiv*, 2024. doi: 10.1101/2024.01.29.577750. URL https://www.biorxiv.org/content/early/2024/01/31/2024.01.29.577750.

Sanderson, T., Bileschi, M. L., Belanger, D., and Colwell, L. J. Proteinfer, deep neural networks for protein functional inference. *Elife*, 12:e80942, 2023.

Sarai, N. S., Fulton, T. J., O'Meara, R. L., Johnston, K. E., Brinkmann-Chen, S., Maar, R. R., Tecklenburg, R. E., Roberts, J. M., Reddel, J. C. T., Katsoulis, D. E., and Arnold, F. H. Directed evolution of enzymatic silicon-carbon bond cleavage in siloxanes. *Science*, 383(6681):438–443, 2024. doi: 10.1126/science.adi5554. URL https://www.science.org/doi/abs/10.1126/science.adi5554.

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.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H., and Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166, 2021.

Seelig, B. and Szostak, J. W. Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature*, 448(7155):828–831, 2007.

Singh, R., Sledzieski, S., Bryson, B., Cowen, L., and Berger, B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences*, 120(24): e2220778120, 2023.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

Truchon, J.-F. and Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of chemical information and modeling*, 47(2):488–508, 2007.

van Gerwen, P., Briling, K. R., Bunne, C., Somnath, V. R., Laplaza, R., Krause, A., and Corminboeuf, C. Equireact: An equivariant neural network for chemical reactions. *arXiv preprint arXiv:2312.08307*, 2023.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L., Söding, J., and Steinegger, M. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, pp. 1–4, 2023.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence

space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

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.

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*, 620(7976): 1089–1100, 2023.

Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Weitzner, B. D., Kipnis, Y., Daniel, A. G., Hilvert, D., and Baker, D. A computational method for design of connected catalytic networks in proteins. *Protein Science*, 28 (12):2036–2041, 2019.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

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.

Yu, H., Deng, H., He, J., Keasling, J. D., and Luo, X. Unikp: a unified framework for the prediction of enzyme kinetic parameters. *Nature Communications*, 14(1):8211, 2023a.

Yu, T., Cui, H., Li, J. C., Luo, Y., Jiang, G., and Zhao, H. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023b.

Zhang, Z., Lu, J., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Protir: Iterative refinement between retrievers and predictors for protein function annotation. *arXiv preprint arXiv:2402.07955*, 2024.

A. Additional Analysis

A.1. Performance of Alternative Architectures and Models

Since our work relies on EGNN (Satorras et al., 2021) to obtain protein representations, we show here the performance of alternative encoders. We consider GVP (Jing et al., 2020), GearNet (Jamasb et al., 2024), and CDConv (Fan et al., 2022) on this task (trained in the exactly the same way). We utilize the same implementation of these models per `github.com/DeepGraphLearning/ESM-GearNet` and observe that these protein encoders perform slightly worse than the EGNN on this task.

We also consider a baseline predicting enzyme-substrate turnover rates, namely UniKP (Yu et al., 2023a). The output of UniKP is defined for a single substrate-enzyme pair, while our dataset consists of chemical reactions. To obtain a single score per reaction-enzyme pair, we obtain the k_{cat} associated with each reactant involved in the reaction and take the average score. We note that UniKP is trained to predict enzyme activity constants (k_{cat}, k_m, k_{cat} / k_m) from enzyme and substrate pairs. In doing so, the model mainly sees positive pairs and implicitly assumes there is some enzymatic activity. This differs from our setting where we do not assume a given pair will necessarily react. Furthermore, the UniKP model doesn't consider the full reaction.

Table 5. Performance of alternative baselines. We show performance of alternative protein encoders (CDConv, GVP, GearNet) as well as a model for turnover rate prediction (UniKP).

MODEL	$BEDROC_{85}$ (%)	$BEDROC_{20}$ (%)	$EF_{0.05}$	$EF_{0.1}$
CDCONV	29.41	43.40	9.55	5.81
GVP	34.12	47.15	10.13	6.03
GEARNET	36.56	55.48	12.44	7.38
UNIKP	2.72	6.87	1.51	1.16

A.2. Sequence Clustering as a Baseline

We provide an additional baseline for virtual enzyme screening. We randomly select three reactions from the training set and note their EC numbers. We randomly choose one protein that has this EC number from the training set. We then cluster the sequences in the screening set against the randomly selected sequences. For simplicity, we use the same FoldSeek clusters as in the paper (0 sequence identity, and 30% coverage). Since we are not guaranteed that a protein in the training set will be found for a specific EC, we follow the logic used in the paper and try to find a protein that matches a coarser level of that EC (instead of considering the lack of an exact match as a total failure). We rank the sequences by their e-values (based on MMSeqs2 and sequence alone) and obtain the results in Table 6. Comparing the results in Table 1 and Table 6, we see that CLIPZyme yields improvement over clustering alone.

It is important to note that applying this strategy requires an alignment for (at least) every protein sampled for a reaction, which doesn't scale well with the size of the screening set. For instance, an evaluation with 1,000 test reactions yields alignments between 100s of millions of sequence pairs. Additionally, the results from clustering can also depend highly on the parameters of the similarity search and the randomly selected proteins.

Table 6. Enzyme virtual screening performance using Sequence Similarity and e-value ranking.

EC LEVEL	$BEDROC_{85}$	$BEDROC_{20}$	$EF_{0.05}$	$EF_{0.1}$
1	2.28	7.57	1.55	1.45
2	9.22	14.75	3.02	2.16
3	27.02	32.10	6.51	3.74
4	38.95	43.98	8.93	4.92

A.3. Enzyme Screening Within EC Classes

We also explore CLIPZyme's ability to discriminate between enzymes within the same EC class, where enzymes are more likely to share function and physical-chemical features. To do so, for each query reaction in the test set, we adjust the

screening set to include only those enzymes belonging to its EC class. The number of enzymes quickly diminishes when considering EC subclasses to the extent that the EC-based screening sets become too small for virtual screening (Figure 2b) – for example, the BEDROC metric is only valid only when $(\alpha \times \text{proportion of actives}) \ll 1$. For this reason, we consider only the top EC level in this analysis. We observe that it is more difficult to rank the correct enzymes higher when only considering sequences in the same EC class but that the top predictions are still enriched for the active enzymes (Table 7).

Table 7. Performance of CLIPZyme when limiting the screening set to enzymes belonging to the query reaction's top EC level.

BEDROC_{85} (%)	BEDROC_{20} (%)	$\text{EF}_{0.05}$	$\text{EF}_{0.1}$
36.25	51.61	11.30	6.83

A.4. Adapting CLEAN for Ranking Enzymes

We consider using both CLEAN EC predictions and computed distances to perform virtual screening similar to Section 4.4.1. Here we present an alternative reranking approach than that in the main body. We follow the exact same setup as reranking EC predictions using CLIPZyme but instead rerank using the distance to the EC anchors. For example, given a query reaction with EC 1.2.3.4, we first predict the EC numbers for all of the enzymes in the screening set using CLEAN. We then rank the enzymes with predicted EC of 1.2.3.4 by the distance from the anchor with EC 1.2.3.4 (computed as the mean embedding of all ECs in the CLEAN training set with EC 1.2.3.4). We then rank all remaining enzymes with predicted EC of 1.2.3.x by their distance to the anchor embeddings of EC 1.2.3.4 (this is the same anchor). This differs from the main body approach since CLEAN assigns EC numbers based on a varying threshold (i.e., max-separation) for each embedding. By first ordering by EC and then reranking within each EC we achieve different results than by ranking all at once by distance to the 1.2.3.4 anchor.

Table 8. Enzyme virtual screening performance when using CLEAN to first place enzymes into successively broader EC levels matching that of the reaction, then re-ranking them according to their Euclidean distance to the reaction's EC.

EC Level	BEDROC_{85} (%)	BEDROC_{20} (%)	$\text{EF}_{0.05}$	$\text{EF}_{0.1}$
Level 1	5.43	26.94	5.55	6.33
Level 2	35.56	71.10	18.95	9.72
Level 3	63.40	85.61	19.35	9.74
Level 4	92.65	96.16	19.48	9.80

A.5. Performance Assuming a Smaller Experimental Capacity

Table 9. Enzyme virtual screening performance compared to using EC prediction alone and together with CLIPZyme, considering a smaller experimental capacity.

EC Level Assumed Available	Method	$\text{BEDROC}_{748.9}$ (%)	$\text{BEDROC}_{173.3}$ (%)	$\text{EF}_{0.005}$	$\text{EF}_{0.01}$
-	CLIPZyme	24.66	37.00	79.09	47.67
Level 1 (x.-.-.-)	CLEAN	0.04	0.35	0.46	0.43
	CLIPZyme + CLEAN	28.39	46.12	100.12	62.75
Level 2 (x.x.-.-)	CLEAN	0.76	2.57	4.39	3.97
	CLIPZyme + CLEAN	41.57	65.03	146.56	84.48
Level 3 (x.x.x.-)	CLEAN	15.5	22.15	46.35	27.84
	CLIPZyme + CLEAN	49.71	73.47	165.97	92.63
Level 4 (x.x.x.x)	CLEAN	75.37	86.93	183.44	9.25
	CLIPZyme + CLEAN	70.95	84.66	180.11	95.07

The metrics used to evaluate the enzyme retrieval places the most emphasis on the top ∼10,000 proteins. The scale at which experimental screening is feasible depends on the similarity of the sequences to be tested. In cases where the enzymes retrieved are diverse, obtaining them as distinct sequences can be expensive, and a screening set of ∼10,000 proteins can be prohibitive. Accordingly, we include in Table 9 the performance when further restricting the screening set size to ∼1,000. In particular, we calculate the BEDROC score for values of α = 748.9 and α = 173.3. These can be interpreted as the metric where the top 0.4% of predictions contribute to 95% and 50% of the score, respectively.

B. Data Processing

B.1. Enzymemap

We obtain version 2 of the EnzymeMap dataset (Heid et al., 2023) and use only the reactions with assigned protein references from either SwissProt or UniProt. Our method requires that the same atoms appear on both sides of the reaction, so we exclude samples where this is not the case. We also filter reactions where the EC number is not fully specified, the sequence could not be retrieved from UniProt, or there wasn't a computable bond change. We restrict our data to proteins of sequence length no more than 650 (maintaining 90% of the sequences) and those with a predicted structure in the AlphaFold database. We remove duplicate reactions, where the same reaction and sequence appear for multiple organisms. We split reactions into train/development/test splits with a ratio of 0.8/0.1/0.1 based on the reaction rule IDs assigned in the dataset. The statistics for the final dataset are shown in Table 10.

Table 10. Statistics of the EnzymeMap dataset used to develop CLIPZyme after pre-processing.

	TRAINING SPLIT	DEVELOPMENT SPLIT	TEST SPLIT
NUMBER OF SAMPLES	34,427	7,287	4,642
NUMBER OF REACTIONS	12,629	2,669	1,554
NUMBER OF PROTEINS	9,794	1,964	1,407
NUMBER OF ECS	2,251	465	319

B.2. Protein Structures

We obtain all protein structures as CIF files from the AlphaFold Protein Structure Database (Jumper et al., 2021; Varadi et al., 2022). We parse these files using the BioPython MMCIFParser. We then construct graphs for use in the PyTorch Geometric library (Fey & Lenssen, 2019). First we filter out the atoms from the CIF file to only include the C_α atoms of the protein. Each graph node as a result represents a residue and the associated coordinates from the CIF file. The edges are determined using the k-nearest neighbors (kNN) method, creating a connected graph that reflects the chemical interactions within the protein. We use a distance of 10 angstroms as a cutoff for the edges.

B.3. MSA Embeddings

We explore using the hidden representations from the MSA Transformer (Rao et al., 2021) as node embeddings of the enzyme 3D structure. Rather than using HHblits (Remmert et al., 2012), we opt for MMSeqs2 (Steinegger & Söding, 2017) because of its speed and efficient search. We follow the pipeline employed by ColabFold (Mirdita et al., 2022) but use only the UniRef30 (`uniref30_2302`) database and do not use an expanded search (Suzek et al., 2015; Mirdita et al., 2017). We sample 128 sequences for each MSA using a greedy search (maximum similarity) to obtain the input for the MSA-Transformer. We keep only the hidden representations of the query enzyme sequence and discard those from the MSA search. For an enzyme of length n, this yields sequence embeddings $h \in \mathbb{R}^{n \times 768}$.

B.4. Computing Screening Set Enzyme Clusters

To exclude from our enzyme screening set those proteins that are similar to sequences used in our training dataset, we compute protein clusters using MMSeqs2 (Steinegger & Söding, 2017) and Foldseek (van Kempen et al., 2023). For MMSeqs2, we use the default parameters with `--min-seq-id`= 0.3 and `--similarity`= 0.8. For Foldseek, we use the default parameters with `--min-seq-id`= 0 and `--c`= 0.3.

C. Implementation Details

All models are developed in PyTorch v2.0.1 (Paszke et al., 2019) and trained using PyTorch Lightning v2.0.9 (Falcon & The PyTorch Lightning team, 2019).

f_{mol} and f_{TS} We implement our reaction encoder (3.1) as two DMPNNs (Yang et al., 2019). We use standard node and edge features (Table 11) to initialize the reactant and product graphs, with input node dimensions of 9 and input edge dimensions of 3. The first encoder, f_{mol} has 5 layers and a hidden dimension of 1,280. The node features for the second encoder, f_{TS} are unchanged, while edges are obtained from taking the sum of the hidden edge representations from f_{mol}. Hence the node dimensions are still 9, while the input edge features have dimensions 1,280. The model also consists of 5 layers and a hidden size of 1,280. We aggregate the graph as a sum over the node features.

Table 11. Chemical properties used as node and edge features in constructing molecular graphs.

ENTITY	FEATURES
ATOM (NODE) FEATURES	ATOMIC NUMBER, CHIRALITY, DEGREE, FORMAL CHARGE, NUMBER OF HYDROGENS, NUMBER OF RADICAL ELECTRONS, HYBRIDIZATION, AROMATICITY, BELONGING TO A RING
BOND (EDGE) FEATURES	BOND TYPE, STEREOCHEMISTRY, CONJUGATION

Condensed Graph Reaction We construct the condensed graph reaction as described in Heid & Green (2021). Specifically, the atom and edge features for the reactants and products are created as binary vectors for the properties detailed in Table 11. For node features $x_i^{(r)}, x_i^{(p)}$ and edge features $e_{ij}^{(r)}, e_{ij}^{(p)}$, we compute $x' = x_i^{(r)} - x_i^{(p)}$ and $e'_{ij} = e_{ij}^{(r)} - e_{ij}^{(p)}$. We do not use the atomic number in calculating x'. Concatenating these with our reactants' features, our final CGR graph consists of 225 atom and 26 edge features, $x_i^{CGR} = [x_i^{(r)} \parallel x'_i]$ and $e_{ij}^{CGR} = [e_{ij}^{(r)} \parallel e'_{ij}]$, respectively.

Reaction SMILES The reaction SMILES is first canonicalized then tokenized according to Schwaller et al. (2019) without atom-mapping. We create a vocabulary based on this tokenization scheme and use a tranformer architecture (Vaswani et al., 2017) as implemented by the Hugging Face library (we use the BertModel) (Wolf et al., 2020). The transformer is initialized with 4 layers, a hidden and intermediate size of 1,280, and 16 attention heads. An absolute positional encoding is used over a maximum sequence length of 1,000. We prepend the reaction with a `[CLS]` token and use its hidden representation as the reaction embedding.

WLDN We implement WLDN as originally described in Jin et al. (2017) and initialize it with 5 layers and a hidden dimension of 1,280. The difference graph is calculated as the difference between atom-mapped node embeddings of the substrate and product graph. We apply a separate 1-layer WLN to obtain the final graph-level representation.

EGNN Node features are initialized with residue-level embeddings from ESM-2 (the 650M parameter variant with 33 layers) (Lin et al., 2022). We use a hidden size of 1,280, 6 layers, and a message dimension of 24. Both features and coordinates are normalized and updated at each step. Neighborhood aggregation is done as an average, and protein-level features are taken as a sum over the final node embeddings. Repurposing the positional encodings used in Vaswani et al. (2017), pairwise distances are transformed with sinusoidal embeddings. For a given relative distance d_{ij} between nodes i and j, the encoding function $f : \mathbb{N} \to \mathbb{R}^d$ transforms this distance into a d-dimensional sinusoidal embedding. The encoding is defined as follows:

$$f(d_{ij})^{(k)} = \begin{cases} \sin\left(\frac{1}{\theta^{k/2}} \cdot d_{ij}\right), k < \frac{d}{2}, \\ \cos\left(\frac{1}{\theta^{\frac{k-\frac{d}{2}}{2}}} \cdot d_{ij}\right), k \geq \frac{d}{2}. \end{cases} \tag{8}$$

where k is the index of the dimension of the distance vector, θ is a hyperparameter that controls the frequency of the sinusoids, which in our case is set to 10,000. The resulting embedding for a particular relative distance d_{ij} is constructed by concatenating the sine-encoded and cosine-encoded vectors, thus interleaving sinusoidal functions along the dimensionality of the embedding space.

CLEAN We train CLEAN with the supervised contrastive ("Supcon-Hard") loss following the training protocol and parameters loss described in the project's repository (https://github.com/tttianhao/CLEAN). Specifically, we use the supervised contrastive loss and the data split in which none of the test enzymes share $> 50\%$ sequence identity with those in the training set. At inference, we use the same approach described in Yu et al. (2023b) to compute the EC anchors. We obtain the predicted distance between each enzyme in our screening set and each EC anchor. We extend this to parent classes of the ECs. For instance, the representation for EC 1.2.3.x is the mean embedding of all CLEAN proteins in that class. We also predict the EC numbers for all of the enzyme sequences in our screening set using the "max-separation" algorithm.

D. Training Details

All models are trained with a batch size of 64 with bfloat16 precision and trained until convergence (approximately 30 epochs). We use a learning rate of $1e^{-4}$ with a cosine learning rate schedule and 100 steps of linear warm-up. Warm-up starts with a learning rate of $1e^{-6}$, and the minimum learning rate after warm-up is set to $1e^{-5}$. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with a weight decay of 0.05 and $(\beta_1, \beta_2) = (0.9, 0.999)$. When training the ESM model, we initialize with the pretrained weights of `esm2_t33_650M_UR50D` and use a mean of the residue embeddings for the sequence representation. We train all models on 8 NVIDIA A6000 GPUs.