FUSED GROMOV-WASSERSTEIN CONTRASTIVE LEARNING FOR EFFECTIVE ENZYME-REACTION SCREENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Enzymes are crucial catalysts that enable a wide range of biochemical reactions. Efficiently identifying specific enzymes from vast protein libraries is essential for advancing biocatalysis. Traditional computational methods for enzyme screening and retrieval are time-consuming and resource-intensive. Recently, deep learning approaches have shown promise. However, these methods focus solely on the interaction between enzymes and reactions, overlooking the inherent relational structure within each domain. To address these limitations, we introduce FGW-CLIP, a novel contrastive learning framework based on optimizing the fused Gromov-Wasserstein distance. FGW-CLIP incorporates multiple alignments, including inter-domain alignment between reactions and enzymes and intra-domain alignment within enzymes and reactions. By introducing a tailored regularization term, our method minimizes the Gromov-Wasserstein distance between enzyme and reaction spaces, which enhances information integration across these domains. Extensive evaluations demonstrate the superiority of FGW-CLIP in challenging enzyme-reaction tasks. On the widely-used EnzymeMap benchmark, FGW-CLIP achieves state-of-the-art performance in enzyme virtual screening, as measured by BEDROC and EF metrics. Moreover, FGW-CLIP consistently outperforms across all three splits of ReactZyme, the largest enzyme-reaction benchmark, demonstrating robust generalization to novel enzymes and reactions. These results position FGW-CLIP as a promising framework for enzyme discovery in complex biochemical settings, with strong adaptability across diverse screening scenarios.

1 INTRODUCTION

Enzymes play a vital role in various biological processes, such as biosynthesis and metabolism. They act as catalysts, speeding up chemical reactions in living organisms. However, in the vast protein sequence databases, such as UniProt (uni, 2025), only 0.23% of proteins are well studied (Ribeiro et al., 2023). Many potentially valuable enzymes may remain hidden among billions of unexplored sequences.

Traditional computational methods, including those based on sequence similarity (Altschul et al., 1990; Desai et al., 2011; Altschul et al., 1997), homology (Krogh et al., 1994; Steinegger et al., 2019), and structure (Roy et al., 2012; Zhang et al., 2017), often require substantial manual effort and computational resources. These methods also suffer from protein annotation errors. In recent years, machine learning methods have shown great promise for enzyme screening and retrieval. CLEAN (Yu et al., 2023) improved the assignment of EC numbers to enzymes. CLIPZyme (Mikhael et al., 2024) employs contrastive learning to align enzyme and reaction representations for in-silico enzyme screening. ReactZyme (Hua et al., 2024) introduces the largest enzyme-reaction benchmark to date. CARE (Yang et al., 2024) integrates textual descriptions of enzymes and reactions to enhance performance. While these methods emphasize enzyme–reaction interactions, they largely overlook the relational structure inherent in each domain.

In this work, we propose FGW-CLIP, a novel contrastive learning framework based on the optimization of fused Gromov-Wasserstein distance. This framework incorporates multiple alignment

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

strategies, including cross-domain representation alignment between reactions and enzymes, as well as intra-domain alignment within enzyme and reaction representations. By introducing tailored regularization terms, FGW-CLIP minimizes the Gromov-Wasserstein distance between the enzyme and reaction spaces during training, thereby promoting more effective information interaction across both domains.

We provide theoretical insights into FGW-CLIP from the perspective of optimizing the fused Gromov-Wasserstein distance. Furthermore, we offer empirical support by validating our approach on the widely-used benchmark EnzymeMap and ReactZyme for enzyme virtual screening. FGW-CLIP outperforms existing baselines, achieving state-of-the-art performance on both benchmarks.

Our key contributions are as follows:

- We introduce FGW-CLIP, a contrastive learning framework that, to the best of our knowledge, is the first to incorporate fused Gromov-Wasserstein (FGW) distance into enzyme–reaction representation learning. Unlike prior approaches that focus only on inter-domain alignment, FGW-CLIP jointly models both inter-domain and intra-domain alignment, effectively capturing the interplay between reaction and enzyme domains as well as their internal structure.
- We show that the FGW-CLIP objective can be interpreted as an FGW-type problem within the inverse optimal transport (IOT) framework, extending the IOT view of contrastive learning to jointly account for cross-domain alignment and intra-domain structure.
- FGW-CLIP achieves consistent improvements over strong baselines on EnzymeMap and ReactZyme. Ablation studies further show that GW-based structural alignment consistently improves retrieval accuracy.

2 Related Work

Contrastive learning. CLIP (Radford et al., 2021) learns aligned representations through cross-modal contrastive learning, inspiring subsequent approaches to multimodal retrieval. X-MoRe (Eom et al., 2023) leverages cross-modal retrieval at inference time to improve the zero-shot classification of CLIP. Inverse optimal transport infers interaction costs from observed matchings (Li et al., 2019) and provides a perspective for understanding contrastive learning (Shi et al., 2023). These connections motivate our integration of cross-domain contrastive learning with structural alignment.

Fused Gromov–Wasserstein distance. Gromov–Wasserstein (GW) compares distributions through their intra-domain relational structures (Mémoli, 2011; Scetbon et al., 2022), allowing objects from different spaces to be matched through their relationships to other objects. Fused Gromov–Wasserstein (FGW) combines this structural comparison with feature-level matching (Titouan et al., 2019; Ma et al., 2024), enabling alignment across heterogeneous domains. We adopt this perspective to jointly model reaction–enzyme correspondence and intra-domain relationships.

Enzyme screening. Enzyme screening ranks candidate enzymes for a query reaction, supporting the identification of potential catalysts. CLIPZyme (Mikhael et al., 2024) aligns enzyme and reaction representations through contrastive learning for virtual screening. ReactZyme (Hua et al., 2024) provides a benchmark for bidirectional enzyme–reaction retrieval under multiple data splits, while CARE (Yang et al., 2024) incorporates textual descriptions for enzyme classification and retrieval. Related methods, such as CLEAN (Yu et al., 2023), focus on EC number prediction. Our approach combines cross-domain matching with intra-domain supervision and GW-based structural alignment. An extended discussion is provided in Appendix A.

3 Preliminaries

We briefly introduce Optimal Transport (OT), Gromov–Wasserstein (GW) distance, and their fused variant FGW to explain the mathematical foundations of our method.

Optimal Transport compares probability distributions by minimizing the cost of transporting mass between them. The Wasserstein distance, a classical OT metric, quantifies this cost when the distributions reside in a shared metric space. Given two discrete probability vectors $p \in \mathbb{R}_+^n$ and $q \in \mathbb{R}_+^m$ and a non-negative cost matrix $M \in \mathbb{R}^{n\times m}$, OT seeks a coupling Γ that minimizes the transportation cost $\langle \Gamma, M\rangle$. Here, $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product, and the feasible set is $\Pi(p, q) = \{\Gamma \in \mathbb{R}_+^{n\times m} \mid \Gamma \mathbf{1}_m = p,\ \Gamma^\top \mathbf{1}_n = q\}$, where $\mathbf{1}_m$ and $\mathbf{1}_n$ are all-ones vectors. In practice, p and q are often chosen to be uniform.

Gromov–Wasserstein distance extends this framework to structured data in different metric spaces by aligning intra-domain relationships. Let (X_1, C^1, p) and (X_2, C^2, q) be two metric measure spaces, where $C^1 \in \mathbb{R}^{n\times n}$ and $C^2 \in \mathbb{R}^{m\times m}$ contain pairwise distances within each domain. GW evaluates how well these distances agree under a soft alignment Γ, using a discrepancy function $\mathcal{L}$, typically the squared loss $\mathcal{L}(a, b) = (a - b)^2$. By comparing relational structures rather than requiring directly comparable pointwise features, GW is suitable for matching heterogeneous objects such as reactions and enzymes. Intuitively, GW encourages consistency between within-reaction and within-enzyme relationships under a cross-domain alignment. This relational perspective complements contrastive learning, which directly aligns paired reaction and enzyme representations in a shared embedding space.

Fused Gromov–Wasserstein (FGW) combines feature-level correspondence with the structural alignment of GW. It integrates the ground cost M and intra-domain structures (C^1, C^2) through a fusion parameter $\alpha \in [0, 1]$:

$$\begin{aligned}\mathrm{FGW}(p, q, M, C^1, C^2) = \min_{\Gamma\in\Pi(p,q)} &(1-\alpha)\langle \Gamma, M\rangle \\ &+ \alpha \sum_{i,j,k,l} \mathcal{L}(C^1_{i,k}, C^2_{j,l})\Gamma_{i,j}\Gamma_{k,l}.\end{aligned} \tag{1}$$

The parameter α balances feature-level and structural alignment. This combination motivates our use of FGW for enzyme–reaction screening, where matching reactions with enzymes should account for both cross-domain correspondence and relational structures within each domain.

4 METHOD

4.1 OVERVIEW OF FGW-CLIP

Given large protein libraries, enzyme screening aims to identify enzymes that can catalyze specific chemical reactions. We formulate this task as a dense retrieval problem. Trained encoders generate representations for both reactions and enzymes. Reactions are then used as queries to retrieve enzymes, which are ranked based on their similarity to the query reactions. Enzyme–reaction retrieval is a bidirectional task, where reactions can be used to retrieve enzymes and enzymes can also be used to retrieve reactions. The top-ranked enzymes or reactions with the highest similarity scores are considered the most likely candidates for catalysis or functional annotation.

We propose FGW-CLIP from the perspective of optimizing the fused Gromov-Wasserstein distance. As shown in Figure 1, we use the pretrained 3D molecular model Uni-Mol (Zhou et al., 2023) to encode the molecules involved in a reaction, and obtain the reaction representation via a readout function. For enzymes, we adopt the pretrained protein language model ESM2 (Lin et al., 2023) to extract sequence embeddings. The ESM2 backbone is kept frozen during FGW-CLIP training, and a linear projection head is added on top to map the embeddings into a shared representation space. FGW-CLIP performs contrastive learning between reactions and enzymes, where each reaction–enzyme pair is labeled as positive if the enzyme can catalyze the reaction, and negative otherwise. In addition, we incorporate the intrinsic connections within enzymes and reactions by leveraging the Enzyme Commission (EC) number. Each reaction or enzyme is associated with one or more EC numbers, which we store as a list per data point. Taking reactions as an example, if a reaction's EC number appears in another sample's EC list within the batch, they are treated as a positive pair; otherwise, as a negative pair. To better align the structural relationships between and within the reaction and enzyme spaces, we introduce a novel regularization term that minimizes the GW distance between them. Furthermore, we integrate an auxiliary EC number prediction loss into the FGW-CLIP framework, enabling effective functional classification.

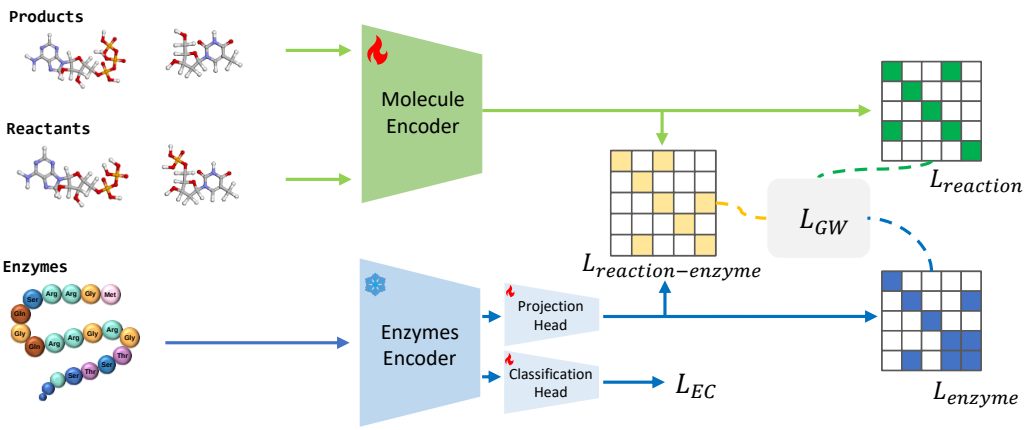

Figure 1: Overview of FGW-CLIP Framework. Reactants and products, along with their structures, are input into the Molecule Encoder, while enzyme sequences are input into the Enzymes Encoder. Representations of enzymes and reactions are aligned using $L_{\text{reaction-enzyme}}$. Internal alignments within reactions and enzymes are achieved via L_{reaction} and L_{enzyme}, respectively, utilizing EC numbers. Inspired by the goal of minimizing the Gromov-Wasserstein (GW) distance, we introduce a regularization term L_{GW}. During training, the Enzyme Encoder is frozen, and a projection head is appended to its output. For clarity, only one classification head is shown in the figure.

In the following sections, we elaborate on the core components of FGW-CLIP. Section 4.2 introduces the molecular encoder Uni-Mol and enzyme encoder ESM2. Section 4.3 details the training strategy of FGW-CLIP, focusing on contrastive learning between reactions and enzymes. Section 4.4 provides a theoretical insight into FGW-CLIP from the perspective of FGW distance optimization.

4.2 Pretraining Backbone of Reaction and Enzyme

For reaction representation, we employ Uni-Mol (Zhou et al., 2023) to encode the molecules involved in reactions. Uni-Mol is pretrained on large-scale molecular datasets with 3D conformations, which helps integrate 3D structural information. We obtain the reaction embedding using a sum-based readout function.

For enzyme representation, we adopt ESM2 (Lin et al., 2023), a protein language model pretrained on millions of protein sequences. In our framework, we leverage ESM2 to encode enzyme sequences by averaging all residue embeddings to obtain a single protein-level representation. ESM2 is kept frozen during training.

4.3 Training Strategy of FGW-CLIP

4.3.1 Contrastive Learning Objectives

We adopt contrastive learning to align reaction and enzyme representations based on experimentally validated catalytic relationships provided in the data. Each training sample consists of a reaction, its corresponding enzyme, and an associated EC number. Since the relationship between reactions and enzymes is not strictly one-to-one, we construct a ground truth label matrix from the data. If a reaction can be catalyzed by a specific enzyme, the corresponding entry in the matrix is set to 1; otherwise, it is 0.

Let R_i denote the index set of enzymes that can catalyze reaction i, and let E_i denote the index set of reactions catalyzed by enzyme i. The inter-domain contrastive loss is defined using the InfoNCE (Oord et al., 2018) formulation:

$$L_{\text{reaction-enzyme}} = \frac{1}{2}\sum_{i=1}^{N}\left[-\sum_{j\in R_i}\log\frac{e^{(\text{sim}(r_i,e_j)/\tau)}}{\sum_{k=1}^{N}e^{(\text{sim}(r_i,e_k)/\tau)}} - \sum_{j\in E_i}\log\frac{e^{(\text{sim}(e_i,r_j)/\tau)}}{\sum_{k=1}^{N}e^{(\text{sim}(e_i,r_k)/\tau)}}\right] \quad (2)$$

Here, r_i and e_i denote the learned representations of reaction i and enzyme i, respectively. $\text{sim}(\cdot,\cdot)$ represents the similarity function, and τ is the temperature parameter. We use cosine similarity as $\text{sim}(\cdot,\cdot)$ throughout our experiments.

In addition to modeling the catalytic relationship between reactions and enzymes, we also incorporate the internal relationships within enzymes and reactions. Let I_i denote the set of samples that share the same EC number with item i. The internal contrastive loss is given as:

$$L_{\text{internal}} = -\sum_{i=1}^{N}\sum_{j\in I_i}\log\frac{e^{(\text{sim}(x_i,x_j)/\tau)}}{\sum_{k=1}^{N}e^{(\text{sim}(x_i,x_k)/\tau)}} \quad (3)$$

Here, x_i and x_j refer to either reactions or enzymes embeddings, while x_k includes all possible embeddings in the same batch.

4.3.2 EC Prediction

The Enzyme Commission (EC) number is a standardized numerical classification scheme for enzymes based on the chemical reactions they catalyze. Each EC number consists of four levels (e.g., EC 1.1.1.1), where each level adds increasing specificity to the enzyme's function. In our framework, we use enzyme representations to predict EC classes across all four hierarchical levels. For each level, a separate classification head is employed, and we use cross-entropy loss for the predictions. The total EC classification loss is the sum of the losses from all four levels:

$$L_{\text{EC}} = L_{\text{level 1}} + L_{\text{level 2}} + L_{\text{level 3}} + L_{\text{level 4}} \quad (4)$$

This multi-level classification framework captures the hierarchical nature of enzyme functions and enhances the biological relevance of the learned representations.

4.3.3 Regularization Loss for GW Distance Optimization

To enhance the alignment between the reaction and the enzyme spaces, we introduce an additional regularization term motivated by the goal of minimizing the Gromov-Wasserstein (GW) distance.

This regularization is essential to maintain the internal structure of both spaces while aligning them effectively. The complete loss function is given as:

$$L_{\text{GW}} = -\sum_{i,j,i',j'=1}^{N}\Gamma_d^{\psi_1}(i,i')\Gamma_d^{\psi_2}(j',j)\Gamma^{\theta}(i,j)\Gamma^{\theta}(i',j') \quad (5)$$

Here, ψ_1 and ψ_2 denote the reaction and enzyme spaces, respectively. The terms $\Gamma_d^{\psi_1}(i,i')$ and $\Gamma_d^{\psi_2}(j,j')$ represent the pairwise structural relationships within the reaction and enzyme spaces. These matrices are computed based on intra-domain similarity. They are detached during training to block gradient flow from propagating and thus avoid interference with the optimization of other terms. $\Gamma^{\theta}(i,j)$ models the soft alignment between reaction i and enzyme j, computed from the similarity between their representations. The corresponding formulations are:

$$\Gamma_d^{\psi_1}(i,i') = \frac{e^{(\text{sim}_d(r_i,r_{i'})/\tau)}}{\sum_{k=1}^{N}e^{(\text{sim}_d(r_i,r_k)/\tau)}}, \quad \Gamma^{\theta}(i,j) = \frac{e^{(\text{sim}(r_i,e_j)/\tau)}}{\sum_{j'=1}^{N}e^{(\text{sim}(r_i,e_{j'})/\tau)}} \quad (6)$$

$\Gamma_d^{\psi_2}(j,j')$ is defined analogously using enzyme representations, and sim_d denotes sim with gradients detached. This objective leverages the internal structural relationships within both domains to guide inter-domain alignment. A full derivation of this regularization term within the FGW-CLIP framework is provided in Appendix D, and its computational complexity is discussed in Appendix E.2.

4.4 FGW-CLIP: Enhancing CLIP by Optimizing Gromov-Wasserstein Distance

By integrating the training objectives in Section 4.3, we can derive the overall training objective for FGW-CLIP, denoted as L_{FGW}, as follows:

$$L_{\text{FGW}} = (1-\alpha)(L_{\text{reaction-enzyme}} + L_{\text{reaction}} + L_{\text{enzyme}}) + 2\alpha L_{\text{GW}} + \lambda L_{\text{EC}} \tag{7}$$

The loss function involves multiple terms. Based on the approach outlined in Shi et al. (2023); Zhou et al. (2024), we establish a connection between L_{FGW} and the fused Gromov-Wasserstein distance optimization problem under a specific constraint through Proposition 4.1.

Proposition 4.1. *Given encoder f_{ψ_1} for data field X_1 and encoder f_{ψ_2} for data field X_2, x_{ψ_1} represents the ℓ_2 normalized embeddings of X_1 from f_{ψ_1}, while x_{ψ_2} represents the ℓ_2 normalized embeddings of X_2 from f_{ψ_2}. Γ^{f_1} represents the label on X_1, Γ^{f_2} represents the label on X_2, Γ^{cor} represents the label on the pairs (X_1, X_2). FGW-CLIP can be interpreted as optimizing the following constrained FGW-type objective:*

$$\begin{aligned}
&\min_{\theta,\psi_1,\psi_2} \Big\{ (1-\alpha) KL(\Gamma^{cor} || \Gamma^{\theta}) + \alpha GW(\Gamma_d^{\psi_1}, \Gamma_d^{\psi_2}, \Gamma^{\theta}) \\
&\quad + \lambda_1 KL(\Gamma^{f_1} || \Gamma^{\psi_1}) + \lambda_2 KL(\Gamma^{f_2} || \Gamma^{\psi_2}) \\
&+ \lambda_{ce} CE(y_{\psi_2}, f_{\psi_2}(X_2)) \Big\} \\
&\textit{subject to} \quad \Gamma^{\theta} = \arg\min_{\Gamma \in U(a^{cor})} \left(\langle C^{\theta}, \Gamma \rangle - \tau H(\Gamma) \right), \\
&\qquad\qquad \Gamma^{\psi_1} = \arg\min_{\Gamma \in U(a^{\psi_1})} \left(\langle C^{\psi_1}, \Gamma \rangle - \tau H(\Gamma) \right), \\
&\qquad\qquad \Gamma^{\psi_2} = \arg\min_{\Gamma \in U(a^{\psi_2})} \left(\langle C^{\psi_2}, \Gamma \rangle - \tau H(\Gamma) \right),
\end{aligned} \tag{8}$$

where $KL(X||Y) = \sum_{ij} x_{ij} \log \frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij}$ represents the Kullback-Leibler divergence, and $H(\Gamma) = -\sum_{i,j} \Gamma_{ij}(\log(\Gamma_{ij}) - 1)$ represents entropic regularization. $\Gamma^{\theta}, \Gamma^{\psi_1}, \Gamma^{\psi_2} \in R_+^{N \times N}$, $C^{\theta}, C^{\psi_1}, C^{\psi_2} \in R_+^{N \times N}$ are cost matrix and $C^{\theta}(i,j) = c - x_{\psi_1,i} x_{\psi_2,j}^T$, $C^{\psi_1}(i,j) = c - x_{\psi_1,i} x_{\psi_1,j}^T$, $C^{\psi_2}(i,j) = c - x_{\psi_2,i} x_{\psi_2,j}^T$. $a^{\psi_1}, a^{\psi_2}, a^{cor}$ represent the label vector of dataset X_1, X_2 and pair dataset (X_1, X_2). $\Gamma_d^{\psi_1}, \Gamma_d^{\psi_2}$ represent the values of $\Gamma^{\psi_1}, \Gamma^{\psi_2}$ respectively, with the gradients detached, $GW(\Gamma_d^{\psi_1}, \Gamma_d^{\psi_2}, \Gamma^{\theta}) = \sum_{i,j=1}^{N} \sum_{i',j'=1}^{N} |\Gamma_d^{\psi_1}(i,i') - \Gamma_d^{\psi_2}(j',j)|^2 \Gamma^{\theta}(i,j) \Gamma^{\theta}(i',j')$ is the Gromov-Wasserstein distance. CE is the cross-entropy loss of data field X_2, which is added to facilitate a specific classification task as a regularization term.

The proof is provided in Appendix D. In L_{FGW}, we utilize Γ^{ψ_1} and Γ^{ψ_2} to learn the structural information of two data domains X_1 and X_2, respectively. Through the optimization of the Gromov-Wasserstein distance, structural alignment at the domain level is achieved. We consider this overall structural alignment information as a supplement and enhancement to the existing label alignment information between the two domains X_1 and X_2. By optimizing this fused Gromov-Wasserstein distance, we can better extend the generalization capability of the CLIP model and alleviate the issue of insufficient effective labels between domains X_1 and X_2.

As illustrated in Eq. 7, the training of FGW-CLIP involves the joint optimization of multiple objective functions. The choice of weighting coefficients (α and λ) affects model performance. We find that smaller values of α reduce the impact of the GW loss, leading to weaker alignment, while larger α values overemphasize the GW optimization, interfering with intra-domain alignment. We also analyze the impact of different λ values on EC prediction loss. Since the EC loss tends to have a relatively large magnitude, assigning it a high weight can dominate the overall objective and suppress the contributions of other loss terms, leading to performance degradation. Therefore, we use a smaller weight of 0.1 to maintain a balanced scale between the EC loss and the other components.

5 EXPERIMENT

5.1 ENZYME SCREENING

5.1.1 EXPERIMENTAL SETUP.

We evaluate enzyme screening on EnzymeMap (Heid et al., 2023), following the reaction-rule-based training, validation, and test splits used by CLIPZyme (Mikhael et al., 2024). Test reactions serve as queries against the same screening database of 261,907 enzyme sequences. We compare FGW-CLIP with CLIPZyme and its variants under the same evaluation protocol. Dataset statistics, screening database construction, and baseline details are provided in Appendix B.

We report BEDROC (Truchon & Bayly, 2007) at $\alpha = 85$ and 20, and enrichment factors (EF) at the top 5% and 10%, following CLIPZyme. Both metrics assess early retrieval performance, with higher values indicating better results. Formal definitions are provided in Appendix C.

5.1.2 RESULTS

Table 1: Enzyme virtual screening performance on EnzymeMap. The higher the BEDROC and EF, the better.

Method	$BEDROC_{85}(\%)$	$BEDROC_{20}(\%)$	$EF_{0.05}$	$EF_{0.1}$
CLIPZyme (ESM)	36.91	53.04	11.93	6.84
CLIPZyme (CGR)	38.91	57.58	13.16	7.73
CLIPZyme	44.69	62.98	14.09	8.06
FGW-CLIP	**48.66**	**66.69**	**14.91**	**8.18**

Table 1 shows the performance comparison between FGW-CLIP and the current SOTA baseline CLIPZyme on EnzymeMap. The best results for each metric are shown in bold. We also include variants of CLIPZyme using different combinations of protein and reaction encoders. ESM denotes using the ESM2 protein language model for enzyme representation. CGR (Hoonakker et al., 2011) is a method for obtaining reaction representations based on graph structures. These results are consistent with those reported in the original CLIPZyme paper. As shown in the table, FGW-CLIP achieves the best performance across all four metrics for BEDROC and EF. Notably, it improves both BEDROC scores by approximately 4 percentage points. This is achieved with a frozen sequence-only enzyme encoder, whereas CLIPZyme uses predicted structures. Relative to the CLIP-only baseline with the same encoders (Table 4), the full objective still adds 2.7 points in $BEDROC_{85}$. This demonstrates the advantage of FGW-CLIP in optimizing the fused GW distance through contrastive learning, which jointly captures inter-domain alignment between enzymes and reactions as well as intra-domain structural relationships.

Table 2: Performance on EnzymeMap. Exclude enzymes that appeared in the training set from the screening set.

Method	$BEDROC_{85}(\%)$	$BEDROC_{20}(\%)$	$EF_{0.05}$	$EF_{0.1}$
CLIPZyme	39.13	58.86	13.40	**7.81**
FGW-CLIP	**45.14**	**61.43**	**13.57**	7.61

We further evaluate the generalization ability of FGW-CLIP by testing on unseen enzymes. Specifically, we exclude all enzymes in the screening set that also appeared in the training set. Table 2 presents the results, showing that FGW-CLIP outperforms CLIPZyme on 3 out of the 4 evaluation metrics, and performs comparably on $EF_{0.1}$. Notably, FGW-CLIP achieves a substantial improvement in both BEDROC scores, demonstrating its ability to identify catalytically relevant enzymes even without prior exposure. These results highlight FGW-CLIP's stronger capacity to capture generalizable features of enzyme–reaction interactions.

Table 3: Retrieval Performance on ReactZyme of the time-based split. The best results are **bolded**. Except for the mean rank, the higher the metric, the better. The baselines with * indicate that they use FANN (Puny et al., 2021) to enhance residue-level representations.

(a) Given the enzyme, retrieve positive reactions.

Time/enzyme-reaction	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Data(Ground-truth)	1.0000	1.0000	1.0000	1.0000	1.0000	0.2002	0.1001	0.0500	1.0004	0.9998
MAT-2D + ESM	0.3246	0.6044	0.7079	0.7972	0.3246	0.1209	0.0708	0.0399	40.4756	0.4549
MAT-2D + SaProt	0.2073	0.4020	0.5004	0.6120	0.2073	0.0804	0.0499	0.0306	75.3546	0.2898
UniMol-2D + ESM	0.2827	0.5210	0.6508	0.7612	0.2827	0.1041	0.0651	0.0380	53.4261	0.4011
UniMol-2D + SaProt	0.1957	0.3855	0.4380	0.6021	0.1957	0.0771	0.0438	0.0301	79.8460	0.2788
UniMol-2D + ESM*	0.2948	0.5866	0.6912	0.7831	0.2948	0.1173	0.0691	0.0391	45.0611	0.4289
UniMol-2D + SaProt*	0.2512	0.4329	0.6474	0.6879	0.2512	0.0866	0.0647	0.0344	63.1455	0.3176
MAT-3D + ESM	0.2858	0.4955	0.6548	0.7405	0.2858	0.0991	0.0655	0.0371	60.3628	0.4041
MAT-3D + SaProt	0.1210	0.2265	0.3108	0.4015	0.1210	0.0453	0.0311	0.0201	150.0301	0.1862
UniMol-3D + ESM	0.2905	0.5365	0.6586	0.7639	0.2905	0.1074	0.0659	0.0382	46.0553	0.4104
UniMol-3D + SaProt	0.0916	0.2134	0.2923	0.3882	0.0916	0.0426	0.0292	0.0194	168.8244	0.1591
UniMol-3D + ESM*	0.3588	0.6545	0.7815	0.8126	0.3588	0.1309	0.0781	0.0406	32.7443	0.4952
UniMol-3D + SaProt*	0.2508	0.4075	0.5448	0.6421	0.2508	0.0815	0.0546	0.0321	59.8345	0.3453
CLIPZyme (MAT-2D + ESM)	0.3041	0.5993	0.6943	0.7840	0.3041	0.1201	0.0695	0.0392	42.3645	0.4355
CLIPZyme (UniMol-3D + ESM)	0.2631	0.4534	0.6444	0.7516	0.2631	0.0907	0.0645	0.0376	45.3637	0.3940
FGW-CLIP	**0.3782**	**0.7143**	**0.8119**	**0.8839**	**0.3782**	**0.1429**	**0.0812**	**0.0442**	**24.1628**	**0.5229**

(b) Given the reaction, retrieve positive enzymes.

Time/reaction-enzyme	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Data(Ground-truth)	1.0000	1.0000	1.0000	1.0000	1.0000	0.4718	0.2895	0.1677	2.8324	0.7497
MAT-2D + ESM	0.2175	0.3815	0.4924	0.6033	0.2175	0.1570	0.1206	0.0871	165.3066	0.1789
MAT-2D + SaProt	0.1260	0.2153	0.2921	0.3778	0.1260	0.0886	0.0716	0.0546	281.2419	0.0981
UniMol-2D + ESM	0.1435	0.2299	0.3554	0.4367	0.1435	0.0946	0.0871	0.0631	270.9385	0.1233
UniMol-2D + SaProt	0.0912	0.1494	0.2252	0.3488	0.0912	0.0615	0.0552	0.0504	536.5624	0.0805
UniMol-2D + ESM*	0.1486	0.2294	0.3529	0.4865	0.1486	0.0944	0.0865	0.0703	254.1982	0.1257
UniMol-2D + SaProt*	0.0988	0.1587	0.2273	0.3536	0.0988	0.0653	0.0557	0.0511	504.2854	0.0934
MAT-3D + ESM	0.2281	0.4240	0.5502	0.5879	0.2281	0.1703	0.1393	0.0852	152.1328	0.1931
MAT-3D + SaProt	0.1037	0.1800	0.2603	0.3671	0.1037	0.0723	0.0659	0.0532	411.5762	0.1056
UniMol-3D + ESM	0.1678	0.3155	0.3960	0.5011	0.1678	0.1267	0.1002	0.0748	177.4881	0.1400
UniMol-3D + SaProt	0.0558	0.0979	0.1359	0.1918	0.0558	0.0393	0.0344	0.0278	700.9714	0.0538
UniMol-3D + ESM*	0.2045	0.3792	0.4475	0.5168	0.2045	0.1523	0.1133	0.0749	167.5862	0.1628
UniMol-3D + SaProt*	0.1331	0.2044	0.3365	0.4119	0.1331	0.0821	0.0852	0.0597	322.5755	0.1122
CLIPZyme (MAT-2D + ESM)	0.1757	0.3447	0.4555	0.5343	0.1757	0.1312	0.1101	0.0756	173.3521	0.1678
CLIPZyme (UniMol-3D + ESM)	0.1331	0.2993	0.3554	0.4567	0.1331	0.1033	0.0949	0.0740	186.4576	0.1313
FGW-CLIP	**0.3967**	**0.6317**	**0.7172**	**0.8075**	**0.3967**	**0.2614**	**0.1866**	**0.1235**	**104.6430**	**0.3392**

5.2 Enzyme-Reaction Retrieval

5.2.1 Experimental setup.

We evaluate bidirectional enzyme–reaction retrieval on ReactZyme (Hua et al., 2024) using its time, enzyme similarity, and reaction similarity splits. Baselines include CLIPZyme and the reaction and enzyme encoder combinations reported in ReactZyme. We follow the ReactZyme evaluation protocol. Dataset and baseline details are provided in Appendix B.

We report Top-k Accuracy, Top-k Accuracy-N, Mean Rank, and Mean Reciprocal Rank (MRR). The first two measure whether any true label appears among the top-k results and the proportion of true labels within these results, respectively. Mean Rank and MRR average ranks and reciprocal ranks over all true labels. Lower Mean Rank and higher values of the other metrics indicate better performance. Definitions are provided in Appendix C.

5.2.2 Results

We evaluate enzyme–reaction retrieval on the three standard ReactZyme splits. Results for the time-based split are reported in Table 3, while the enzyme-similarity and reaction-similarity results are provided in Appendix F.

On the time-based split, FGW-CLIP achieves the best performance across all reported metrics in both retrieval directions. It also consistently outperforms the baselines on the enzyme-similarity

split. On the more challenging reaction-similarity split, FGW-CLIP achieves the best results on all metrics except Mean Rank.

The reaction-similarity test set contains only 386 unique reactions, all unseen during training (Hua et al., 2024). Because Mean Rank averages the ranks of all positives, a few poorly represented test reactions, whose positives are ranked low in both directions, can substantially inflate it, while the Top-k and MRR metrics remain strong.

When compared with CLIPZyme using the same encoder backbone, namely UniMol-3D for reactions and ESM for enzymes, FGW-CLIP consistently achieves better Top-k and MRR performance across all three splits. These results demonstrate that the improvements are not attributable solely to the choice of encoders.

5.3 Ablation Study

Table 4: Ablation study on EnzymeMap.

Method	$\text{BEDROC}_{85}(\%)$	$\text{BEDROC}_{20}(\%)$	$\text{EF}_{0.05}$	$\text{EF}_{0.1}$
CLIP only	45.94	61.11	13.28	7.41
FGW-CLIP w/o GW	45.83	64.17	14.37	8.12
FGW-CLIP	**48.66**	**66.69**	**14.91**	**8.18**

To evaluate the contribution of GW regularization, we compare FGW-CLIP with the cross-domain contrastive baseline (CLIP only) and FGW-CLIP w/o GW in Table 4.

Compared with CLIP only, FGW-CLIP w/o GW improves BEDROC_{20} and both enrichment-factor metrics while maintaining comparable BEDROC_{85} performance. These results suggest that intra-domain contrastive learning and the auxiliary loss improve retrieval beyond cross-domain reaction–enzyme alignment alone.

Adding GW regularization further improves BEDROC_{85} from 45.83 to 48.66 and BEDROC_{20} from 64.17 to 66.69, with additional gains in both EF metrics. This comparison supports the contribution of GW-based structural alignment to enzyme screening performance. These gains are consistent across different early retrieval metrics.

The two comparisons serve different purposes. CLIP only provides a reference for the combined benefit of the additional training objectives, whereas FGW-CLIP w/o GW isolates the effect of GW regularization within the framework.

Additional ablations on the GW weight and intra-domain structure construction are provided in Appendix E.3.

6 Conclusion

In this work, we proposed FGW-CLIP, a novel contrastive learning framework that enhances enzyme screening by optimizing the fused Gromov-Wasserstein (FGW) distance. Our approach jointly aligns the reaction and enzyme spaces while preserving the structural relationships within each domain. FGW-CLIP integrates three contrastive objectives—reaction–enzyme, enzyme–enzyme, and reaction–reaction alignment—along with a Gromov-Wasserstein regularization and an auxiliary EC prediction loss to better capture biochemical relationships. We conducted extensive experiments on EnzymeMap and ReactZyme, where FGW-CLIP consistently achieved state-of-the-art performance across key metrics such as BEDROC and MRR. Notably, with the same encoder backbone, FGW-CLIP outperformed other baselines on all ReactZyme splits, confirming its robustness and generalization.

In future work, we plan to improve the reaction representation by incorporating reaction directionality, addressing current limitations. Additionally, we will explore the integration of enzyme 3D information and functional-level optimization, and extend FGW-CLIP to broader biochemical applications such as drug discovery and metabolic pathway prediction.

AI USE STATEMENT

We used generative AI tools to improve manuscript readability, polish language, check grammar, and assist with information searches. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

Parit Bansal, Anne Morgat, Kristian B Axelsen, Venkatesh Muthukrishnan, Elisabeth Coudert, Lucila Aimo, Nevila Hyka-Nouspikel, Elisabeth Gasteiger, Arnaud Kerhornou, Teresa Batista Neto, et al. Rhea, the reaction knowledgebase in 2022. *Nucleic acids research*, 50(D1):D693–D700, 2022.

Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31 (1):365–370, 2003.

Antje Chang, Lisa Jeske, Sandra Ulbrich, Julia Hofmann, and Dietmar Schomburg. Brenda, the elixir core data resource in 2021: new developments and updates. *Nucleic Acids Research*, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Dhwani K Desai, Soumyadeep Nandi, Prashant K Srivastava, and Andrew M Lynn. Modenza: accurate identification of metabolic enzymes using function specific profile hmms with optimised discrimination threshold and modified emission probabilities. *Advances in bioinformatics*, 2011 (1):743782, 2011.

Seongha Eom, Namgyu Ho, Jaehoon Oh, and Se-Young Yun. Cross-modal retrieval meets inference: Improving zero-shot classification with cross-modal retrieval. *arXiv preprint arXiv:2308.15273*, 2023.

Esther Heid, Daniel Probst, William H. Green, and Georg K. H. Madsen. Enzymemap: curation, validation and data-driven prediction of enzymatic reactions. *Chem. Sci.*, 14:14229–14242, 2023. doi: 10.1039/D3SC02048G.

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

Chenqing Hua, Bozitao Zhong, Sitao Luan, Liang Hong, Guy Wolf, Doina Precup, and Shuangjia Zheng. Reactzyme: A benchmark for enzyme-reaction prediction. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL `https://openreview.net/forum?id=xepxnDQoGq`.

Anders Krogh, Michael Brown, I Saira Mian, Kimmen Sjölander, and David Haussler. Hidden markov models in computational biology: Applications to protein modeling. *Journal of molecular biology*, 235(5):1501–1531, 1994.

Alexander Kroll, Sahasra Ranjan, Martin KM Engqvist, and Martin J Lercher. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature communications*, 14(1):2787, 2023.

Ruilin Li, Xiaojing Ye, Haomin Zhou, and Hongyuan Zha. Learning to match via inverse optimal transport. *Journal of machine learning research*, 20(80):1–37, 2019.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Yong Liu, Chenqing Hua, Menglong Xu, Tao Zeng, Jiahua Rao, Zhongyue Zhang, Ruibo Wu, Jing-Ke Weng, Connor W Coley, and Shuangjia Zheng. A geometric foundation model for enzyme retrieval with evolutionary insights. *Nature Catalysis*, 9(2):148–160, 2026.

Xinyu Ma, Xu Chu, Yasha Wang, Yang Lin, Junfeng Zhao, Liantao Ma, and Wenwu Zhu. Fused gromov-wasserstein graph mixup for graph-level classifications. *Advances in Neural Information Processing Systems*, 36, 2024.

Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. Molecule attention transformer. *arXiv preprint arXiv:2002.08264*, 2020.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Peter G Mikhael, Itamar Chinn, and Regina Barzilay. Clipzyme: Reaction-conditioned virtual screening of enzymes. *arXiv preprint arXiv:2402.06748*, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. *arXiv preprint arXiv:2110.03336*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Antonio JM Ribeiro, Ioannis G Riziotis, Neera Borkakoti, and Janet M Thornton. Enzyme function and evolution through the lens of bioinformatics. *Biochemical Journal*, 480(22):1845–1863, 2023.

Ambrish Roy, Jianyi Yang, and Yang Zhang. Cofactor: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research*, 40(W1):W471–W477, 2012.

Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. 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.

Theo Sanderson, Maxwell L Bileschi, David Belanger, and Lucy J Colwell. Proteinfer, deep neural networks for protein functional inference. *Elife*, 12:e80942, 2023.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time gromov wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pp. 19347–19365. PMLR, 2022.

Liangliang Shi, Gu Zhang, Haoyu Zhen, Jintao Fan, and Junchi Yan. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In *International conference on machine learning*, pp. 31408–31421. PMLR, 2023.

Martin Steinegger, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J Haunsberger, and Johannes Söding. Hh-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics*, 20:1–15, 2019.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: protein language modeling with structure-aware vocabulary. *bioRxiv*, pp. 2023–10, 2023.

Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pp. 6275–6284. PMLR, 2019.

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

Yixuan Wei, Yue Cao, Zheng Zhang, Houwen Peng, Zhuliang Yao, Zhenda Xie, Han Hu, and Baining Guo. iclip: Bridging image classification and contrastive language-image pre-training for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2776–2786, 2023.

Jason Yang, Ariane Mora, Shengchao Liu, Bruce Wittmann, Animashree Anandkumar, Frances Arnold, and Yisong Yue. Care: a benchmark suite for the classification and retrieval of enzymes. *Advances in Neural Information Processing Systems*, 37:3094–3121, 2024.

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023.

Chengxin Zhang, Peter L Freddolino, and Yang Zhang. Cofactor: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic acids research*, 45(W1):W291–W299, 2017.

Yu Zhang, Qi Zhang, Zixuan Gong, Yiwei Shi, Yepeng Liu, Duoqian Miao, Yang Liu, Ke Liu, Kun Yi, Wei Fan, et al. Mlip: Efficient multi-perspective language-image pretraining with exhaustive data utilization. *arXiv preprint arXiv:2406.01460*, 2024.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

Gengmo Zhou, Zhen Wang, Feng Yu, Guolin Ke, Zhewei Wei, and Zhifeng Gao. S-molsearch: 3d semi-supervised contrastive learning for bioactive molecule search. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=wJAF8TGVUG.

A Extended Related Work

A.1 Contrastive Learning

Contrastive learning has found extensive applications in vision and multimodal representation learning. CLIP (Contrastive Language-Image Pretraining) (Radford et al., 2021) advances this paradigm by aligning image and text representations, enabling strong performance in tasks such as image classification, text generation, and human-computer interaction. MLIP (Zhang et al., 2024) improves CLIP by integrating spatial and frequency-domain information for more comprehensive multimodal understanding. iCLIP (Wei et al., 2023) further bridges contrastive learning and image classification by optimizing CLIP for both visual tasks and vision-language pairings. X-MoRe (Eom et al., 2023) improves the zero-shot classification of CLIP without additional training, by retrieving relevant captions from an external image-text corpus and ensembling the image- and text-modal predictions. Beyond these applications, inverse optimal transport infers an interaction cost from observed matchings (Li et al., 2019), a view under which contrastive learning can also be understood (Shi et al., 2023) and which motivates our framework.

A.2 Fused Gromov-Wasserstein Distance

The Gromov-Wasserstein (GW) Distance is a metric from optimal transport theory that measures the similarity between two metric spaces by considering the structures of the spaces rather than their

individual points. Mémoli (2011) proves that $GW^{1/2}$ defines a distance on the space of metric measure spaces quotiented by measure-preserving isometries, and Scetbon et al. (2022) further reduce its computational cost to linear time using low-rank couplings and costs. Since exact optimal transport computation is costly, entropic regularization with Sinkhorn iterations (Cuturi, 2013) provides a widely used approximation strategy for optimal transport problems. Fused Gromov-Wasserstein (FGW) distance (Titouan et al., 2019; Ma et al., 2024) extends the GW metric to calculate transportation distance between two unregistered probability distributions on different product metric spaces, such as combining graph signals and structures, making it suitable for attributed graphs.

A.3 Enzyme Screening

Enzyme virtual screening and recognition accelerate the discovery of new enzymes and drug candidates by accurately identifying functions and efficiently screening potential inhibitors from large libraries. CLIPZyme (Mikhael et al., 2024) is a contrastive learning framework that effectively encodes and aligns representations of enzymes and their corresponding reaction pairs for in-silico enzyme screening. ReactZyme (Hua et al., 2024) introduces a large-scale benchmark for enzyme-reaction prediction, enabling improved enzyme retrieval and function annotation. CARE (Yang et al., 2024) incorporates textual descriptions of enzymes and reactions, and provides a benchmark suite for enzyme classification and retrieval. CLEAN (Yu et al., 2023) improves EC number assignment by accurately annotating understudied enzymes and correcting mislabeled entries. In addition, traditional sequence-similarity-based tools and machine learning models, such as BLASTp (Altschul et al., 1990), DeepEC (Ryu et al., 2019) and ProteInfer (Sanderson et al., 2023), can also be used for EC number prediction. More recently, EnzymeCAGE (Liu et al., 2026) models enzyme–reaction catalytic specificity by combining pocket-level enzyme structures with evolutionary information for enzyme retrieval.

B Dataset Details and Baselines

B.1 Enzyme Screening

EnzymeMap Based on the original EnzymeMap dataset (Heid et al., 2023), this dataset consists of biochemical reactions linked to UniProt IDs and EC numbers. It contains 46,356 enzyme-driven reactions, including 16,776 unique chemical reactions, 12,749 unique enzymes, 2,841 EC numbers, and 394 reaction rules. Following the setup in CLIPZyme (Mikhael et al., 2024), we split the dataset into training, validation, and test sets based on reaction rule IDs with a ratio of 0.8/0.1/0.1, resulting in 34,427, 7,287, and 4,642 entries, respectively.

Enzyme Screening Set This dataset integrated the EnzymeMap dataset, Brenda release 2022_2 (Chang et al., 2020), and UniProt release 2022_01 (Yu et al., 2023), and filtered out the sequences that are longer than 650 amino acids. The final set includes 261,907 protein sequences. Note that Mikhael et al. (2024) report 260,197 sequences, whereas the screening set released with their code contains 261,907 sequences. We use the released set. During evaluation, reactions from the EnzymeMap test set are used as queries to screen potential matching enzymes from this pool.

Baselines In this task, we use the SOTA method CLIPZyme (Mikhael et al., 2024) and several of its variants as baselines. CLIPZyme is a contrastive learning approach for enzyme screening, and we follow its experimental setup.

B.2 Enzyme-Reaction Retrieval

ReactZyme ReactZyme compiled from the SwissProt and Rhea databases (Boeckmann et al., 2003; Bansal et al., 2022), provides a comprehensive resource for enzyme-reaction prediction. The dataset comprises 178,463 enzyme-reaction pairs, including 178,327 unique enzymes and 7,726 unique reactions. Compared to existing datasets such as ESP (Kroll et al., 2023) and EnzymeMap, ReactZyme contains significantly more enzyme-reaction pairs and captures substrate, product, and reaction-level information, though it lacks atom-mapping data.

Data Split ReactZyme provides explicit train-test splits for each setting. Additionally, 10% of the training set is randomly sampled as a validation set.

- Time Split: Based on annotation date, with pairs before 2010-12-31 for training (166,175 pairs) and after for testing (12,287 pairs), ensuring a 93%/7% split.
- Enzyme Similarity Split: Ensures at least 60% sequence difference between training and test enzymes, resulting in 169,724 training pairs and 8,739 test pairs (95%/5% split).
- Reaction Similarity Split: Ensures no overlap in reactions between training and test sets, with 163,771 training pairs and 14,692 test pairs (91%/9% split).

Baselines For this task, baseline models include Molecule Attention Transformer-2D (MAT-2D) (Maziarka et al., 2020) and UniMol-2D (Zhou et al., 2023) for 2D molecular graphs, and MAT-3D and UniMol-3D for 3D molecular conformations as reaction representations. For enzyme representations, ESM (Lin et al., 2023) and the structure-aware protein language model SaProt (Su et al., 2023) are utilized. Additionally, an equivariant graph neural network FANN (Puny et al., 2021) is employed to enhance residue-level representations. For CLIPZyme (Mikhael et al., 2024), we adopt the implementation provided in the ReactZyme paper (Hua et al., 2024). This follows the evaluation protocol defined by ReactZyme to ensure consistency.

EC annotation preprocessing ReactZyme contains incomplete EC annotations, including missing entries denoted by "-" and non-numeric tokens such as "n11". We therefore adopt a max-based completion strategy at each level of the EC hierarchy. Entries containing an "n"-prefixed token are assigned a reserved category indexed one above the largest observed valid category, while missing entries denoted by "-" are assigned the next reserved category. These reserved indices do not overlap with the observed EC categories and allow every sample to be represented by a complete four-level EC label. We apply this in all ReactZyme experiments.

C Evaluation metrics

Enzyme Screening We give the formal definitions of the two screening metrics described in the main text. BEDROC emphasizes early enrichment among the top-ranked candidates:

$$\text{BEDROC} = \frac{\sum_{i=1}^{n} e^{-\alpha r_i/N}}{\frac{R_\alpha(1-e^{-\alpha})}{e^{\alpha/N}-1}} \times \frac{R_\alpha \sinh(\alpha/2)}{\cosh(\alpha/2) - \cosh\left(\alpha/2 - \alpha R_\alpha\right)} + \frac{1}{1-e^{\alpha(1-R_\alpha)}}, \tag{9}$$

where n is the number of active compounds, N the total number of compounds, $R_\alpha = n/N$, and r_i the rank of the i-th active compound. Here α denotes the early-recognition parameter of BEDROC, which is unrelated to the fusion weight α in Eq. 7. EF is the fold increase of active compounds within the top χ fraction relative to their overall proportion:

$$\text{EF} = \frac{\sum_{i=1}^{n} \delta_i}{\chi n}, \quad \delta_i = \begin{cases} 1, & r_i \le \chi N \\ 0, & r_i > \chi N \end{cases} \tag{10}$$

where $\chi \in [0, 1]$ is the fraction of the ranked list considered.

Enzyme-Reaction Retrieval We give the formal definitions of the four retrieval metrics described in the main text, where N is the number of samples, M_i the number of true labels for the i-th sample, Rank_{ij} the rank of its j-th true label, $\mathbb{I}(\cdot)$ the indicator function, and k the cutoff threshold.

$$\text{Top-k} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\sum_{j=1}^{M_i} \mathbb{I}(\text{Rank}_{ij} \le k) > 0 \right) \tag{11}$$

$$\text{Top-k-N} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{j=1}^{M_i} \mathbb{I}(\text{Rank}_{ij} \le k)}{k} \tag{12}$$

$$\text{Mean Rank} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{M_i}\text{Rank}_{ij}}{M_i} \tag{13}$$

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}\frac{1}{\text{Rank}_{ij}} \tag{14}$$

D PROOF FOR FGW-CLIP

First, we establish a lemma for the loss in Eq. 2 induced by the inverse optimal transport problem.

Lemma D.1. *The loss in Equation 2 can be derived from the following optimization problem:*

$$\begin{aligned} &\min_{\theta} \quad KL(\hat{P}||P^{\theta}) \\ &\textit{subject to} \quad P^{\theta} = \arg\min_{P\in U(a)} \big(\langle C^{\theta}, P\rangle - \tau H(P)\big) \end{aligned} \tag{15}$$

where $C^{\theta} \in R^{N\times N}, C^{\theta}(i,j) = c - s_{ij}(\theta)$ and $\hat{P}(i,j) = \frac{f_{ij}}{N}$, where f_{ij} are label which equals to 1 when x_i and x_j are related else 0. I_i denotes the index associated with x_i. $U(\mathbf{a}) = \{\Gamma \in R_{+}^{N\times N} | \Gamma \mathbf{1}_N = \mathbf{a}\}$. Here $\mathbf{a}$ denotes a vector whose elements are the sums of labels, specifically defined as $\mathbf{a}(i) = \sum_{j=1}^{N} f_{ij}$.

We introduce the Lagrangian of equation 15 as follows:

$$L(P,d) = \langle C^{\theta}, P\rangle - \tau H(P) - \sum_{i=1}^{N} d_i\Big(\sum_{j=1}^{N} P_{ij} - a_i\Big) \tag{16}$$

The KKT conditions can be obtained as follows:

$$\frac{\partial L(P,d)}{\partial P_{ij}} = C_{ij}^{\theta} + \tau log P_{ij} - d_i = 0 \tag{17}$$

Given that $\sum_{j=1}^{N} P_{ij} = a_i$, we can derive the following expression:

$$P_{ij} = \frac{a_i e^{-C_{ij}^{\theta}/\tau}}{\sum_{k=1}^{N} e^{-C_{ik}^{\theta}/\tau}} \tag{18}$$

By solving the optimization problem 15 according to definition, we can obtain the following results:

$$L_{iot} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j\in I_i}\log\Big(\frac{e^{-C_{ij}^{\theta}/\tau}}{\sum_{k=1}^{N} e^{-C_{ik}^{\theta}/\tau}}\Big) + Constant \tag{19}$$

where the constant collects all terms independent of θ, including those involving a_i, since the row sums of P^{θ} are fixed to $\mathbf{a}$. To simplify the problem, we disregard the factor $1/N$ and the constant in L_{iot} in practical applications, resulting in the following expression:

$$L_{iot} = -\sum_{i=1}^{N}\sum_{j\in I_i}\log\Big(\frac{e^{-C_{ij}^{\theta}/\tau}}{\sum_{k=1}^{N} e^{-C_{ik}^{\theta}/\tau}}\Big) \tag{20}$$

D.1 PROOF FOR PROPOSITION 4.1

According to Lemma D.1, we can transform the original optimization problem into the following problem:

$$\min_{\theta,\psi_1,\psi_2}\Big\{-(1-\alpha)\sum_{i=1}^{N}\sum_{j\in I_i}\log\left(\frac{e^{-C_{ij}^{\theta}/\tau}}{\sum_{k=1}^{N}e^{-C_{ik}^{\theta}/\tau}}\right)+\alpha GW(\Gamma_d^{\psi_1},\Gamma_d^{\psi_2},\Gamma^{\theta})$$
$$-\lambda_1\sum_{i=1}^{N}\sum_{j\in I_i}\log\left(\frac{e^{-C_{ij}^{\psi_1}/\tau}}{\sum_{k=1}^{N}e^{-C_{ik}^{\psi_1}/\tau}}\right)-\lambda_2\sum_{i=1}^{N}\sum_{j\in I_i}\log\left(\frac{e^{-C_{ij}^{\psi_2}/\tau}}{\sum_{k=1}^{N}e^{-C_{ik}^{\psi_2}/\tau}}\right)\quad+\lambda_{ce}CE(y_{\psi_2},f_{\psi_2}(X_2))\Big\} \tag{21}$$

For the GW term, we can simplify it according to the definition as follows:

$$GW(\Gamma_d^{\psi_1},\Gamma_d^{\psi_2},\Gamma^{\theta})=\left(\Gamma_d^{\psi_1}\circ\Gamma_d^{\psi_1}a^{\psi_1}\right)^{\top}a^{\psi_1}+\left(\Gamma_d^{\psi_2}\circ\Gamma_d^{\psi_2}a^{\psi_2}\right)^{\top}a^{\psi_2}-2\,\mathrm{tr}\left(\left(\Gamma^{\theta}\right)^{\top}\Gamma_d^{\psi_1}\Gamma^{\theta}\Gamma_d^{\psi_2}\right) \tag{22}$$

which $\circ$ Hadamard product. Treating the marginals of Γ^{θ} as fixed, the first two terms are constant. Disregarding them, we can simplify the optimization objective as follows:

$$GW(\Gamma_d^{\psi_1},\Gamma_d^{\psi_2},\Gamma^{\theta})=-2tr\left((\Gamma^{\theta})^{\top}\Gamma_d^{\psi_1}\Gamma^{\theta}\Gamma_d^{\psi_2}\right) \tag{23}$$

Considering the symmetric positions of i and j, a classic technique is to transform the original optimization problem into the following form:

$$\min_{\theta,\psi_1,\psi_2}\Big\{-\frac{1-\alpha}{2}\Big(\sum_{i=1}^{N}\sum_{j\in I_i}\log\left(\frac{e^{-C_{ij}^{\theta}/\tau}}{\sum_{k=1}^{N}e^{-C_{ik}^{\theta}/\tau}}\right)\quad+\sum_{j=1}^{N}\sum_{i\in J_j}\log\left(\frac{e^{-C_{ij}^{\theta}/\tau}}{\sum_{k=1}^{N}e^{-C_{kj}^{\theta}/\tau}}\right)\Big)-2\alpha\,\mathrm{tr}\left((\Gamma^{\theta})^{\top}\Gamma_d^{\psi_1}\Gamma^{\theta}\Gamma_d^{\psi_2}\right)$$
$$-\lambda_1\sum_{i=1}^{N}\sum_{j\in I_i}\log\left(\frac{e^{-C_{ij}^{\psi_1}/\tau}}{\sum_{k=1}^{N}e^{-C_{ik}^{\psi_1}/\tau}}\right)-\lambda_2\sum_{i=1}^{N}\sum_{j\in I_i}\log\left(\frac{e^{-C_{ij}^{\psi_2}/\tau}}{\sum_{k=1}^{N}e^{-C_{ik}^{\psi_2}/\tau}}\right)\quad+\lambda_{ce}\,\mathrm{CE}\left(y_{\psi_2},f_{\psi_2}(X_2)\right)\Big\} \tag{24}$$

The above expression is consistent with the form of the overall loss obtained by FGW-CLIP. Given that $i\in I_i$, $j\in J_j$ and $C^{\theta}(i,j)=c-x_{\psi_1,i}x_{\psi_2,j}^{T}$, by reorganizing equation 24, it can be observed that:

$$\min_{\theta,\psi_1,\psi_2}\Big\{-\frac{1-\alpha}{2}\Big(\sum_{i=1}^{N}\sum_{j\in I_i}\frac{x_{\psi_1,i}x_{\psi_2,j}^{T}-c}{\tau}+\sum_{j=1}^{N}\sum_{i\in J_j}\frac{x_{\psi_1,i}x_{\psi_2,j}^{T}-c}{\tau}\Big)\quad+\alpha\,GW(\Gamma_d^{\psi_1},\Gamma_d^{\psi_2},\Gamma^{\theta})$$
$$+\frac{1-\alpha}{2}\Big(\sum_{i=1}^{N}\sum_{j\in I_i}\log\left(\sum_{k=1}^{N}e^{-C_{ik}^{\theta}/\tau}\right)\quad+\sum_{j=1}^{N}\sum_{i\in J_j}\log\left(\sum_{k=1}^{N}e^{-C_{kj}^{\theta}/\tau}\right)\Big)$$
$$-\lambda_1\sum_{i=1}^{N}\sum_{j\in I_i}\log\left(\frac{e^{-C_{ij}^{\psi_1}/\tau}}{\sum_{k=1}^{N}e^{-C_{ik}^{\psi_1}/\tau}}\right)-\lambda_2\sum_{i=1}^{N}\sum_{j\in I_i}\log\left(\frac{e^{-C_{ij}^{\psi_2}/\tau}}{\sum_{k=1}^{N}e^{-C_{ik}^{\psi_2}/\tau}}\right)\quad+\lambda_{ce}\,\mathrm{CE}\left(y_{\psi_2},f_{\psi_2}(X_2)\right)\Big\} \tag{25}$$

Since x_{ψ_1},x_{ψ_2} are L_2 normalized, we have $x_{\psi_1,i}x_{\psi_2,j}^{T}=1-\frac{1}{2}\|x_{\psi_1,i}-x_{\psi_2,j}\|^2$. Disregarding constants, we can derive that:

$$\min_{\theta,\psi_1,\psi_2} \Big\{ \frac{1-\alpha}{2}\Big(\sum_{i=1}^{N}\sum_{j\in I_i} \frac{\|x_{\psi_1,i}-x_{\psi_2,j}\|^2}{2\tau} + \sum_{j=1}^{N}\sum_{i\in J_j} \frac{\|x_{\psi_1,i}-x_{\psi_2,j}\|^2}{2\tau} \Big) \quad + \alpha GW(\Gamma_d^{\psi_1}, \Gamma_d^{\psi_2}, \Gamma^\theta)$$
$$+ \frac{1-\alpha}{2}\Big(\sum_{i=1}^{N}\sum_{j\in I_i} \log\left(\sum_{k=1}^{N} e^{-C_{ik}^{\theta}/\tau} \right) \quad + \sum_{j=1}^{N}\sum_{i\in J_j} \log\left(\sum_{k=1}^{N} e^{-C_{kj}^{\theta}/\tau} \right) \Big)$$
$$- \lambda_1 \sum_{i=1}^{N}\sum_{j\in I_i} \log\left(\frac{e^{-C_{ij}^{\psi_1}/\tau}}{\sum_{k=1}^{N} e^{-C_{ik}^{\psi_1}/\tau}} \right) - \lambda_2 \sum_{i=1}^{N}\sum_{j\in I_i} \log\left(\frac{e^{-C_{ij}^{\psi_2}/\tau}}{\sum_{k=1}^{N} e^{-C_{ik}^{\psi_2}/\tau}} \right) \quad + \lambda_{ce}\,\mathrm{CE}\left(y_{\psi_2}, f_{\psi_2}(X_2)\right) \Big\} \tag{26}$$

From the equation, we can deduce that L_{FGW} optimizes a fused Gromov-Wasserstein-type objective under regularization conditions, in which the label-alignment term and the GW structural term share the same coupling Γ^θ.

E Experiment Details

For the training of FGW-CLIP on EnzymeMap, we use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$, along with a weight decay of 1×10^{-4}. The initial learning rate is 0.001, and we apply a polynomial decay scheduler with a warmup ratio of 0.06. The batch size is 32, distributed across 4 NVIDIA GeForce RTX 4090 24G GPUs. The dropout rate is set to 0.1. The temperature parameter τ used in all contrastive losses is 0.05. The weight λ for the EC prediction loss L_{EC} is set to 0.1. For the reaction branch, we adopt Uni-Mol as the molecular encoder, with the same hyperparameters as in its original paper. The molecular readout function is set to sum. For the enzyme branch, we use the pretrained ESM2 protein language model as the encoder. ESM2 is frozen during training, and we append a linear projection head to map its output to the shared representation space. The training epoch is 100, and the last checkpoint is selected.

For training on ReactZyme, the number of epochs is set to 50, and the best model checkpoint is selected based on validation loss. All experiments on ReactZyme are conducted using 4 NVIDIA A100-SXM4-80GB GPUs. The remaining parameters are the same as those used during EnzymeMap training.

E.1 Visualization

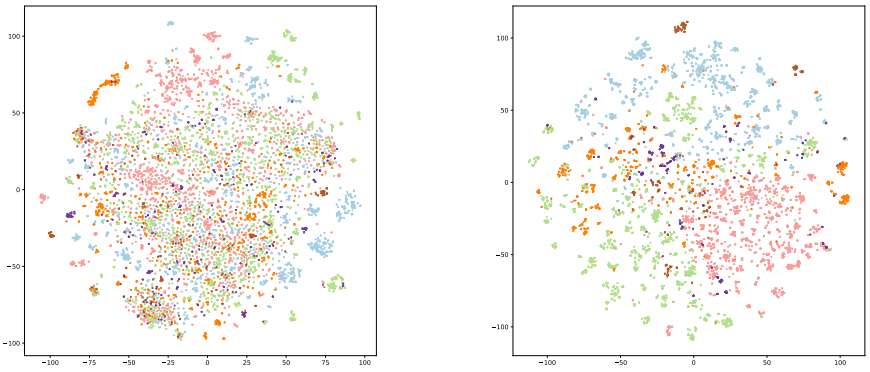

Figure 2: t-SNE visualization of enzyme representations. Left: pretrained checkpoint. Right: FGW-CLIP. Colors represent top-level EC numbers (EC 1 to EC 6).

To visually demonstrate the distinctions between embeddings learned by FGW-CLIP and those from pretrained ESM2 checkpoint, we present a comparative visualization in Figure 2. The enzymes

depicted are sourced from EnzymeMap, with distinct colors representing different top-level EC numbers ranging from EC 1 to EC 6. Upon comparison, the classification boundaries in Figure 2 generated by FGW-CLIP exhibit greater clarity, and the intra-class molecular distances appear more appropriately scaled. Notably, some clusters subdivide into multiple subclusters, potentially reflecting the inherent hierarchical structure within the molecular compositions.

E.2 Computational Complexity of GW Distance

The GW term in Eq. 5 is defined as a quadruple summation over index pairs (i, j, i', j'), which costs $\mathcal{O}(N^4)$ if evaluated naively, where N is the batch size. Equivalently, it can be written in the trace form derived in Appendix D and computed via matrix multiplications in $\mathcal{O}(N^3)$. Moreover, this computation is applied only within mini-batches, rather than over the entire dataset. As a result, the actual runtime is independent of the total dataset size.

To empirically evaluate this, we benchmarked our method with and without GW regularization on the EnzymeMap dataset. Results show:

- The per-step runtime increases by only 0.0037 seconds, a relative increase of 3.3%;
- GPU memory consumption remains virtually unchanged on a single GPU.

These findings indicate that the practical overhead of GW regularization is minimal in our mini-batch training scheme.

E.3 Additional Ablation Studies

Table 5: Ablation study of different GW optimization strategies and weight α on EnzymeMap.

Method	$\text{BEDROC}_{85}(\%)$	$\text{BEDROC}_{20}(\%)$	$\text{EF}_{0.05}$	$\text{EF}_{0.1}$
$\alpha = 0.05$	45.59	63.84	14.21	**8.24**
$\alpha = 0.3$	47.46	64.86	14.29	8.02
$\alpha = 0.5$	47.44	64.31	14.18	7.89
Label, $\alpha = 0.1$	47.08	65.28	14.64	8.08
No Detach, $\alpha = 0.1$	47.63	64.96	14.39	8.06
FGW-CLIP (Detach, $\alpha = 0.1$)	**48.66**	**66.69**	**14.91**	8.18

We further investigate the impact of different GW loss weights α and incorporation strategies within the FGW-CLIP framework, as summarized in Table 5.

The parameter α controls the trade-off between inter-domain alignment and intra-domain structure preservation. We find that $\alpha = 0.1$ yields the best overall performance. A smaller value ($\alpha = 0.05$) underweights the GW loss, leading to weaker alignment. In contrast, a larger value ($\alpha = 0.5$) overemphasizes GW optimization, which disrupts the intra-domain alignment and impairs the model's ability to capture meaningful intra-domain structures.

We also compare several GW integration strategies. The "Label" variant replaces the similarity matrices $\text{sim}_d(r_i, r_{i'})$ and $\text{sim}_d(e_j, e_{j'})$ in Eq. 6 with binary labels derived from internal contrastive learning (Eq. 3). The "No Detach" variant disables gradient blocking for these matrices, allowing gradients to propagate through them during training.

As shown in Table 5, the default "Detach" strategy, where gradients are not propagated through intra-domain similarity matrices, achieves the best performance. This detachment improves optimization stability by preventing interference between inter-domain alignment and intra-domain structure preservation. This strategy helps strike a meaningful balance, further highlighting the effectiveness of our design.

Table 6: Retrieval Performance on ReactZyme of the enzyme-similarity-based split. The best results are **bolded**. Except for the mean rank, the higher the metric, the better. The baselines marked with * indicate that they use FANN to enhance residue-level representations.

(a) Given the enzyme, retrieve positive reactions.

Enzyme/enzyme-reaction	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Data(Ground-truth)	1.0000	1.0000	1.0000	1.0000	1.0000	0.2001	0.1001	0.0500	1.0003	0.9999
MAT-2D + ESM	0.5987	0.8759	0.9328	0.9572	0.5987	0.1774	0.0939	0.0485	5.3021	0.7280
MAT-2D + SaProt	0.6691	0.8893	0.9358	0.9553	0.6691	0.1801	0.0942	0.0484	5.4356	0.7733
UniMol-2D + ESM	0.6077	0.8759	0.9338	0.9533	0.6077	0.1774	0.0940	0.0483	7.0311	0.7349
UniMol-2D + SaProt	0.5717	0.8357	0.8891	0.9474	0.5717	0.1693	0.0895	0.0480	15.2646	0.6912
UniMol-2D + ESM*	0.6256	0.8749	0.9348	0.9493	0.6256	0.1772	0.0941	0.0481	7.0024	0.7491
UniMol-2D + SaProt*	0.6038	0.8695	0.9338	0.9375	0.6038	0.1761	0.0940	0.0475	7.0746	0.7346
MAT-3D + ESM	0.4544	0.6408	0.8573	0.9118	0.4544	0.1300	0.0863	0.0462	30.8473	0.5093
MAT-3D + SaProt	0.5539	0.6904	0.8821	0.9296	0.5539	0.1400	0.0888	0.0471	15.3962	0.6735
UniMol-3D + ESM	0.7267	0.9062	0.9487	0.9632	0.7267	0.1835	0.0955	0.0488	4.5799	0.8112
UniMol-3D + SaProt	0.5998	0.8665	0.9229	0.9454	0.5998	0.1755	0.0929	0.0479	7.4701	0.7226
UniMol-3D + ESM*	0.7111	0.9017	0.9547	0.9592	0.7111	0.1826	0.0961	0.0486	4.8395	0.8023
UniMol-3D + SaProt*	0.6328	0.8853	0.9348	0.9513	0.6328	0.1793	0.0941	0.0482	6.9597	0.7457
CLIPZyme (MAT-2D + ESM)	0.5489	0.7768	0.9290	0.9460	0.5489	0.1554	0.0929	0.0473	8.3524	0.6971
CLIPZyme (UniMol-3D + ESM)	0.7547	0.9478	0.9679	0.9780	0.7547	0.1896	0.0968	0.0489	3.9820	0.8546
FGW-CLIP	**0.7642**	**0.9746**	**0.9875**	**0.9953**	**0.7642**	**0.1950**	**0.0988**	**0.0498**	**2.3837**	**0.8581**

(b) Given the reaction, retrieve positive enzymes.

Enzyme/reaction-enzyme	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Data(Ground-truth)	1.0000	1.0000	1.0000	1.0000	1.0000	0.4833	0.3370	0.2263	3.2778	0.7321
MAT-2D + ESM	0.3624	0.6091	0.7225	0.7986	0.3624	0.2961	0.2444	0.1820	22.5053	0.2586
MAT-2D + SaProt	0.3999	0.6455	0.7583	0.8390	0.3999	0.3138	0.2565	0.1912	23.5890	0.2883
UniMol-2D + ESM	0.3435	0.5701	0.7007	0.7530	0.3435	0.2771	0.2370	0.1716	25.4892	0.2512
UniMol-2D + SaProt	0.3049	0.5273	0.6347	0.6872	0.3049	0.2563	0.2147	0.1566	30.5631	0.2245
UniMol-2D + ESM*	0.3584	0.5892	0.7338	0.7543	0.3584	0.2864	0.2482	0.1719	25.0362	0.2674
UniMol-2D + SaProt*	0.3534	0.5713	0.7051	0.7640	0.3534	0.2777	0.2385	0.1741	25.1678	0.2635
MAT-3D + ESM	0.3827	0.6373	0.7089	0.8048	0.3827	0.3098	0.2398	0.1834	26.4117	0.2841
MAT-3D + SaProt	0.3751	0.5493	0.6572	0.7394	0.3751	0.2670	0.2223	0.1685	24.5678	0.2763
UniMol-3D + ESM	0.4088	0.6892	0.7953	0.8666	0.4088	0.3350	0.2690	0.1975	24.2505	0.2930
UniMol-3D + SaProt	0.3477	0.5458	0.6980	0.7762	0.3477	0.2653	0.2361	0.1769	34.9487	0.2562
UniMol-3D + ESM*	0.3928	0.6612	0.7628	0.8324	0.3928	0.3214	0.2580	0.1897	23.8241	0.2837
UniMol-3D + SaProt*	0.3655	0.6161	0.7376	0.7552	0.3655	0.2995	0.2495	0.1721	22.8901	0.2633
CLIPZyme (MAT-2D + ESM)	0.3337	0.6077	0.6514	0.7687	0.3337	0.2844	0.2235	0.1811	30.4196	0.2038
CLIPZyme (UniMol-3D + ESM)	0.3570	0.6371	0.7552	0.8431	0.3570	0.2885	0.2577	0.1834	25.5786	0.2828
FGW-CLIP	**0.6978**	**0.8931**	**0.9389**	**0.9739**	**0.6978**	**0.4338**	**0.3200**	**0.2216**	**10.8511**	**0.5300**

Table 7: Retrieval Performance on ReactZyme of the reaction-similarity-based split. The best results are **bolded**. Except for the mean rank, the higher the metric, the better. The baselines marked with * indicate that they use FANN to enhance residue-level representations.

(a) Given the enzyme, retrieve positive reactions.

Reaction/enzyme-reaction	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Data(Ground-truth)	1.0000	1.0000	1.0000	1.0000	1.0000	0.2000	0.1000	0.0500	1.0000	1.0000
MAT-2D + ESM	0.0914	0.2968	0.4373	0.5908	0.0914	0.0596	0.0438	0.0296	39.9146	0.2005
MAT-2D + SaProt	0.0963	0.3018	0.4123	0.5070	0.0963	0.0606	0.0413	0.0254	72.0597	0.1936
UniMol-2D + ESM	0.0949	0.2694	0.4363	0.4232	0.0949	0.0541	0.0437	0.0212	65.2719	0.1865
UniMol-2D + SaProt	0.0944	0.2754	0.4143	0.4571	0.0944	0.0553	0.0415	0.0229	59.7940	0.1956
UniMol-2D + ESM*	0.0929	0.2610	0.4313	0.4271	0.0929	0.0524	0.0432	0.0214	72.7932	0.1810
UniMol-2D + SaProt*	0.0926	0.2699	0.4343	0.5309	0.0926	0.0542	0.0435	0.0266	89.8456	0.1857
MAT-3D + ESM	0.0930	0.2595	0.4203	0.4431	0.0930	0.0521	0.0421	0.0222	81.3234	0.1893
MAT-3D + SaProt	0.0915	0.2565	0.4293	0.5269	0.0915	0.0515	0.0430	0.0264	94.9242	0.1804
UniMol-3D + ESM	0.0912	0.2580	0.4213	0.4571	0.0912	0.0518	0.0422	0.0229	92.2778	0.1856
UniMol-3D + SaProt	0.1085	0.2699	0.4034	0.5429	0.1085	0.0542	0.0404	0.0272	42.3597	0.1988
UniMol-3D + ESM*	0.1104	0.3023	0.4573	0.5669	0.1104	0.0607	0.0458	0.0284	38.9685	0.2011
UniMol-3D + SaProt*	0.0962	0.2545	0.4024	0.5289	0.0962	0.0511	0.0403	0.0265	50.9663	0.1972
CLIPZyme (MAT-2D + ESM)	0.1235	0.3064	0.5719	0.6000	0.1235	0.0613	0.0572	0.0300	**35.6457**	0.2201
CLIPZyme (UniMol-3D + ESM)	0.1305	0.3420	0.5320	0.6220	0.1305	0.0684	0.0532	0.0311	48.4672	0.1937
FGW-CLIP	**0.2561**	**0.5117**	**0.5799**	**0.6919**	**0.2561**	**0.1024**	**0.0580**	**0.0346**	41.8125	**0.3804**

(b) Given the reaction, retrieve positive enzymes.

Reaction/reaction-enzyme	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Data(Ground-truth)	1.0000	1.0000	1.0000	1.0000	1.0000	0.5389	0.3870	0.2711	19.5272	0.6715
MAT-2D + ESM	0.1347	0.2000	0.2326	0.2753	0.1347	0.1083	0.0902	0.0749	529.4258	0.1341
MAT-2D + SaProt	0.0933	0.1627	0.2213	0.2561	0.0933	0.0881	0.0858	0.0697	504.8481	0.1076
UniMol-2D + ESM	0.0931	0.1321	0.1769	0.1863	0.0931	0.0715	0.0686	0.0507	550.0562	0.0946
UniMol-2D + SaProt	0.0910	0.1380	0.1818	0.2345	0.0910	0.0747	0.0705	0.0638	567.8300	0.0989
UniMol-2D + ESM*	0.1033	0.1502	0.2076	0.2547	0.1033	0.0813	0.0805	0.0693	590.4462	0.0928
UniMol-2D + SaProt*	0.0905	0.1339	0.1813	0.2407	0.0905	0.0725	0.0703	0.0655	549.8296	0.0961
MAT-3D + ESM	0.1269	0.1962	0.2251	0.2712	0.1269	0.1062	0.0873	0.0738	532.6187	0.1184
MAT-3D + SaProt	0.0909	0.1407	0.1849	0.2528	0.0909	0.0762	0.0717	0.0688	539.1481	0.1044
UniMol-3D + ESM	0.0924	0.1332	0.1790	0.2172	0.0924	0.0721	0.0694	0.0591	548.3340	0.0943
UniMol-3D + SaProt	0.0933	0.1703	0.2130	0.2613	0.0933	0.0922	0.0826	0.0711	**493.1189**	0.0962
UniMol-3D + ESM*	0.1244	0.2058	0.2440	0.2848	0.1244	0.1114	0.0946	0.0775	559.1225	0.1129
UniMol-3D + SaProt*	0.0917	0.1418	0.1847	0.2234	0.0917	0.0768	0.0716	0.0608	552.4546	0.1051
CLIPZyme (MAT-2D + ESM)	0.1457	0.2173	0.2456	0.2893	0.1457	0.1135	0.1001	0.0783	501.2071	0.1521
CLIPZyme (UniMol-3D + ESM)	0.1298	0.1993	0.2215	0.2544	0.1298	0.0866	0.0830	0.0741	526.4793	0.1245
FGW-CLIP	**0.3782**	**0.5933**	**0.6658**	**0.7150**	**0.3782**	**0.2746**	**0.2163**	**0.1530**	552.7454	**0.3113**

F THE RESULTS ON THE ENZYME-SIMILARITY-BASED SPLIT AND REACTION-SIMILARITY-BASED SPLIT OF REACTZYME

We provide detailed results for the enzyme and reaction retrieval experiments on the enzyme similarity-based and reaction similarity-based splits of ReactZyme in Tables 6 and 7, respectively. The experiments follow the same setup described for the time-based split in the main text, with FGW-CLIP compared against the baselines. The results for CLIPZyme in Table 3 of the main text and Tables 6 and 7 here are sourced from Tables 15, 16, and 17 of the ReactZyme paper.

For the enzyme similarity split (Table 6), FGW-CLIP demonstrates superior performance across both retrieval tasks. In the reaction similarity split (Table 7), FGW-CLIP achieves SOTA results on all metrics except Mean Rank. The weaker Mean Rank arises because the test set of this split contains only 386 unique reactions, all unseen during training (Hua et al., 2024). Mean Rank averages the ranks of all positives and is therefore dominated by a few poorly ranked ones: if a small number of these test reactions are poorly represented, their positives are ranked low in both retrieval directions. Top-k and MRR, which focus on the top of the ranking, are much less affected, showcasing the robustness of FGW-CLIP in retrieval tasks with no reaction overlap between training and test sets.

Experiments in Tables 6 and 7 validate the findings presented in the main text, further emphasizing FGW-CLIP's effectiveness and its consistent outperformance of baselines, including those using the same encoders.

G LIMITATION

Currently, the framework is limited to enzyme screening and retrieval, and the reaction representation does not explicitly model reaction directionality. In the future, it will be applied to more extensive enzyme prediction and design tasks.