CHEMISTRY-INFORMED DUAL ENCODERS FOR ENZYME-REACTION DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Enzyme discovery often begins by uncovering promiscuous activities in known enzymes or characterizing the reactions catalyzed by newly identified sequences. Understanding an enzyme's substrate and reaction scope requires connecting active-site features to reaction chemistry, a relationship that existing prediction methods often overlook. We propose Chemistry-Informed Reaction–Catalyst Embeddings (*CIRCE*), a contrastive dual encoder for bidirectional enzyme–reaction retrieval. *CIRCE* integrates function-aware protein representations with reaction representations that encode reactants, products, stoichiometry, cofactors, and additional chemical annotations. Incorporating chemical knowledge into both its representations and training objective enables *CIRCE* to learn features relevant to enzyme reactivity and promiscuity, with the aim of improving generalization to new reactions and accelerating enzyme discovery. Across enzyme screening and bidirectional retrieval benchmarks, *CIRCE* achieves state-of-the-art performance and robust generalization. These results establish its utility for enzyme discovery and provide a foundation for investigating the active-site determinants of catalytic promiscuity.

1 INTRODUCTION

Enzymes catalyze diverse chemical transformations with remarkable selectivity, sustaining metabolism across all forms of life and enabling advances in metabolic engineering, pharmaceutical synthesis, and sustainable chemical manufacturing (Nielsen & Keasling, 2016; Bornscheuer et al., 2012; Sheldon & Woodley, 2018). Enzyme discovery regards both identifying sequences for target reactions and uncovering additional activities in known enzymes. Such catalytic promiscuity expands the range of accessible transformations and provides a starting point for optimization through protein engineering. Yet only a small fraction of sequences in large protein databases, such as UniProt (Consortium, 2023), have experimentally annotated reactions, and characterizing their broader catalytic scope requires costly screening. Computational methods, linking enzymes to compatible reactions, can therefore accelerate discovery by prioritizing candidates for experimental validation.

Predicting enzyme–reaction compatibility requires understanding how active-site residues and their structural environment support substrate recognition and catalysis. Furthermore, enzyme–reaction relationships are inherently many-to-many: an enzyme may catalyze multiple reactions, and a reaction may be catalyzed by multiple enzymes. Auxiliary information can help capture these relationships and support the discovery of promiscuous activities and enzymes that catalyze novel reactions.

Existing computational approaches infer enzyme function and reaction compatibility through sequence similarity (Stephen, 1990; Desai et al., 2011; Rahman et al., 2014), profile-based homology searches (Krogh et al., 1994; Steinegger et al., 2019), or structural comparisons (Roy et al., 2012; Zhang et al., 2017). These approaches depend on the coverage and quality of reference annotations, while structure-based workflows can impose high computational costs. Recent deep learning methods address these limitations by learning from large datasets using biological supervision, such as Enzyme Commission (EC) labels (Yu et al., 2023a), or CLIP-like contrastive objectives (Mikhael et al., 2024; Hua et al., 2024; Zhou et al., 2025; Rocks et al., 2026). Some also incorporate auxiliary information, including enzyme textual descriptions (Zhang et al., 2026). We hypothesize that integrating information about functionally relevant residues with explicit reaction chemistry can better organize the joint enzyme–reaction latents to enhance predictions of enzyme reactivity and promiscuity.

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.

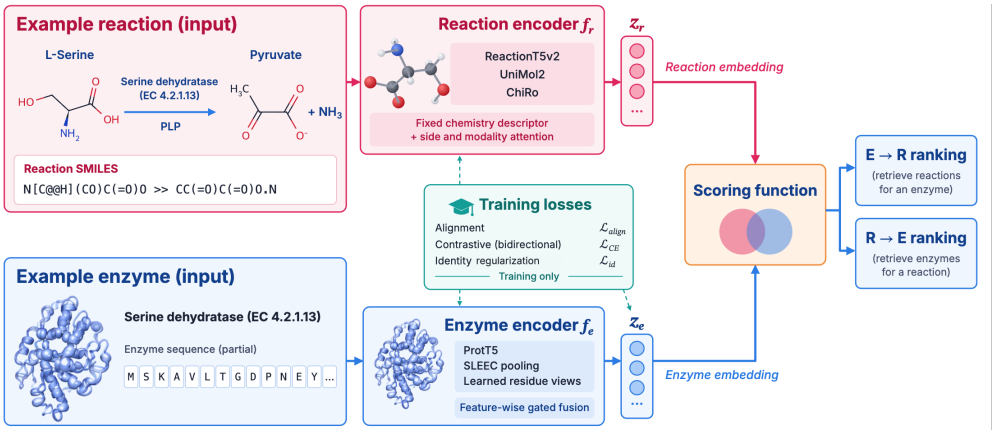


Figure 1: *CIRCE* **pipeline overview.** The dual-encoder approach consists of a ProT5-based enzyme encoder that combines global and SLEEC-guided pooling with K learned residue views through gated fusion. Then, the reaction encoder combines molecular features extracted by ReactionT5v2, UniMol2D, and ChIRo with a fixed chemistry descriptor. Each modality then has some residual correcting features. The training phase consists of a first alignment step, where the two main encoders align their representations, followed by residual refinement, where the residual heads are tuned for prediction correction. Finally, the final embeddings are scored according to the function in Eq. 8.

We introduce *CIRCE* (Chemistry-Informed Reaction–Catalyst Embeddings), a contrastive dual-encoder framework that learns a shared representation space for bidirectional retrieval of enzymes and reactions they may catalyze. The enzyme encoder integrates functional-residue prediction priors (Gong et al., 2026) with complementary learned residue representations, while the reaction encoder captures reactants, products, stoichiometry, cofactors, and additional chemical descriptors. A many-to-many contrastive objective accommodates multiple compatible partners for both enzymes and reactions, and a residual refinement stage uses known associations to improve cross-modal alignment. Together, these components link enzyme functional features to reaction chemistry, supporting catalyst discovery and exploration of enzyme reaction scope while providing a foundation for investigating the active-site determinants of reactivity and promiscuity.

We evaluate *CIRCE* on EnzymeMap (Heid et al., 2023) and ReactZyme (Hua et al., 2024) for virtual enzyme screening and bidirectional retrieval. On EnzymeMap, *CIRCE* improves full-library screening BEDROC_{85} over the released CLIPZyme checkpoint under the same evaluation protocol. On ReactZyme, it outperforms the ESM2/MAT-2D and EnzGFM baselines in both retrieval directions across all three splits. Appendix F details the controlled ablation protocols and inference-sensitivity results. Our primary contributions are:

- ◇ We introduce *CIRCE*, a dual-encoder framework integrating functional-residue priors and explicit reaction chemistry to support enzyme discovery and exploration of catalytic scope.
- ◇ We develop a many-to-many contrastive objective that captures multiple compatible enzymes and reactions, together with a residual refinement stage that uses known associations to improve cross-modal alignment.
- ◇ We demonstrate state-of-the-art performance on enzyme screening and bidirectional retrieval on EnzymeMap and ReactZyme, with consistent gains across benchmarks.

2 Related Work

Functional-site Prediction. Protein function is often localized to specific regions of the protein, known as functional sites. Several models try to identify such residues in the protein sequence to exploit their functional importance. EasIFA5 (Wang et al., 2024) combines protein-language-model (pLM) embeddings, three-dimensional structure, and reaction information through multimodal cross-attention to predict catalytic residues, while Squidly (Rieger et al., 2025) performs sequence-only catalytic-residue prediction using per-residue pLM embeddings and biology-informed contrastive

learning. Finally, SLEEC (Gong et al., 2026) jointly learns EC-oriented protein representations and sparse functional-residue scores through semi-supervised learning. Integrating functional-site predictions inferred from the SLEEC model, CIRCE employs a site-guided pooling mechanism that prioritizes catalytically relevant residues during enzyme–reaction retrieval. This focuses the learned protein representations on key sequence regions essential for catalytic activity.

Enzyme–Reaction Retrieval. Conventional methods for enzyme–reaction compatibility often rely on proxy tasks like EC classification (Dalkiran et al., 2018; Yu et al., 2023b) or substrate binding prediction (Zeng et al., 2022), ignoring broader contextual dynamics. Framing compatibility as a cross-modal retrieval task offers a flexible alternative by mapping enzymes and reactions into a unified embedding space. CLIPZyme (Mikhael et al., 2024) and ReactZyme (Hua et al., 2024) established this contrastive retrieval paradigm and its benchmarks, which subsequent works extended through EC supervision (CREEP (Yang et al., 2024)) or large-scale pretraining (Horizyn-1 (Rocks et al., 2026)). More recent architectures incorporate richer auxiliary signals: FGW-CLIP (Zhou et al., 2025) includes a structural geometry regularization, TIGER (Zhang et al., 2026) exploits text descriptions, and EnzymeCAGE (Liu et al., 2026) incorporates catalytic-pocket information. Despite these advances, disparate benchmarks prevent direct comparisons across frameworks. To bridge this gap, CIRCE combines molecular-set chemical features with functional-site-guided residue pooling within a dual-encoder architecture. Furthermore, we account for enzyme promiscuity via a multi-positive contrastive alignment objective, complemented by a residual refinement step over the training association graph.

Protein and Reaction Representation. Protein language models like ProtT5 (Elnaggar et al., 2022) extract contextual embeddings without MSAs, bridging sequence-only systems (e.g., Horizyn-1 (Rocks et al., 2026)) and structure-aware encoders (Mikhael et al., 2024; Liu et al., 2026). Differing from structure-dependent methods, CIRCE operates strictly sequence-only at inference using frozen ProtT5 embeddings. To capture catalytic context without 3D structures, it complements a global sequence summary with a fixed SLEEC-guided functional-site pool and learned query pools.

Reaction featurization spans traditional SMILES (Zhong et al., 2022) and fingerprints (Rogers & Hahn, 2010) to neural models capturing reaction context (ReactionT5v2 (Sagawa & Kojima, 2025)), 3D geometry (Uni-Mol2 (Ji et al., 2024)), and chirality (ChIRo (Adams et al., 2022)). CIRCE constructs a hybrid view fusing ReactionT5v2 with permutation-invariant Uni-Mol2 and ChIRo molecular pools alongside explicit chemical descriptors. Role-agnostic set summarization is restricted to explicit descriptors, while neural attention pools preserve reactant and product side identities.

3 METHOD

We first formalize reaction deorphanization and enzyme screening as a bidirectional retrieval task in 3.1. CIRCE framework is outlined in Figure 1. The trained encoders combine functional-residue pooling with multi-view learned embeddings for the enzyme representation, while adopting chemical descriptors for the reaction one (Sections 3.2–3.3). Representation alignment and residual refinement objectives guide the learning process of the joint embedding space (Sections 3.4–3.4).
At inference, candidate ranking is defined by a cosine similarity-based score function (Section 3.5).

3.1 PROBLEM FORMULATION

Let $\mathcal{R}$ and $\mathcal{E}$ denote the sets of reactions and enzymes under consideration, respectively. Their experimentally validated associations form a bipartite graph $\mathcal{D} \subseteq \mathcal{R} \times \mathcal{E}$, with binary adjacency $Y_{ij} = \mathbb{1}[(r_i, e_j) \in \mathcal{D}]$. This graph is many-to-many due to enzyme promiscuity: a reaction can pair with multiple enzymes, and an enzyme can catalyze multiple reactions. We aim to learn a scoring function $S(r, e)$ that measures enzyme–reaction compatibility, relying exclusively on the training graph $\mathcal{D}$. Let f_R and f_E denote the complete reaction and enzyme retrieval models, respectively; they are learned to be mapped in the same dimensional shared space. Their final ℓ_2-normalized outputs $\mathbf{z}_r = f_R(r)$ and $\mathbf{z}_e = f_E(e)$ define the compatibility score $S(r, e) = \mathbf{z}_r^\top \mathbf{z}_e \in [-1, 1]$. This score natively supports bidirectional retrieval. For an enzyme query e, $E \to R$ retrieval ranks candidate reactions $r \in \mathcal{C}_R(e) \subseteq \mathcal{R}$ by $S(r, e)$. Conversely, for a reaction query r, $R \to E$ retrieval ranks

candidate enzymes $e \in \mathcal{C}_E(r) \subseteq \mathcal{E}$. The relevant candidates comprise the observed neighbors of each query in the evaluation association graph.

3.2 Enzyme Representation

In enzymes, catalytic function depends on both global sequence context and specific catalytic residues. Given an enzyme sequence e, we extract residue embeddings $\mathbf{H}_e \in \mathbb{R}^{L\times 1024}$ using a frozen ProtT5 backbone (Elnaggar et al., 2022), where $L \leq 1022$ denotes the number of retained residue tokens applying truncation if needed. A first sequence representation is extracted via mean pooling across residues, yielding a global sequence embedding $\mathbf{x}_e^P \in \mathbb{R}^{1024}$. To isolate catalytic context without requiring 3D structures, we enrich this representation using a fixed SLEEC-guided (Gong et al., 2026) pooling $\mathbf{x}_e^S \in \mathbb{R}^{1024}$ alongside K learned residue representations.

SLEEC-Guided Pooling. Let $\mathcal{V}_e$ index the L valid residues of enzyme e, each residue is represented by embedding $\mathbf{h}_i \in \mathbb{R}^{1024}$ ($i \in \{1,\dots,L\}$). We train a SLEEC model (Gong et al., 2026) to assign each residue embedding $\mathbf{h}_i$ a functional logit ℓ_i that indicates its catalytic relevance. Details of the SLEEC training procedure are provided in Appendix A. The frozen pre-trained SLEEC model then assigns each residue embedding $\mathbf{h}_i$ a functional logit ℓ_i, which is centered and clipped as $\tilde{\ell}_i = \mathrm{clip}(\ell_i - L^{-1}\sum_{j\in\mathcal{V}_e}\ell_j, -10, 10)$. We then obtain the SLEEC-guided vector $\mathbf{x}_e^S \in \mathbb{R}^{1024}$, that aggregates residue features using normalized functional weights:

$$a_i^{\mathrm{S}} = (1-\eta)\frac{\exp(\tilde{\ell}_i)}{\sum_{j\in\mathcal{V}_e}\exp(\tilde{\ell}_j)} + \frac{\eta}{L}, \qquad \mathbf{x}_e^S = \sum_{i\in\mathcal{V}_e} a_i^{\mathrm{S}}\mathbf{h}_i, \tag{1}$$

where $\eta = 0.05$ ensures numerical stability and a non-zero minimal contribution from every valid residue, balancing active-site focus with surrounding sequence context.

Learned Auxiliary Views. To complement the representations $\mathbf{x}_e^P$ and $\mathbf{x}_e^S$, we introduce K learned auxiliary views that can adaptively identify residue-level patterns useful for enzyme–reaction retrieval. Each view is parameterized by a learnable query vector $\mathbf{q}_k \in \mathbb{R}^{256}$, $k = 1,\dots,K$. A linear residue adapter maps each residue embedding $\mathbf{h}_i$ to a normalized key $\mathbf{k}_i \in \mathbb{R}^{256}$, used to determine its relevance to the query, and a value $\mathbf{v}_i \in \mathbb{R}^{256}$, which carries the residue information to be aggregated. Given the normalized query, $\hat{\mathbf{q}}_k = \mathbf{q}_k/\|\mathbf{q}_k\|_2$, we compute

$$p_{ki} = \frac{\exp\left(\gamma\hat{\mathbf{q}}_k^\top\mathbf{k}_i\right)}{\sum_{j\in\mathcal{V}_e}\exp\left(\gamma\hat{\mathbf{q}}_k^\top\mathbf{k}_j\right)}, \qquad \mathbf{x}_e^k = \sum_{i\in\mathcal{V}_e}\left[(1-\eta)p_{ki} + \frac{\eta}{L}\right]\mathbf{v}_i, \tag{2}$$

where $\mathcal{V}_e$ denotes the set of valid residues and $L = |\mathcal{V}_e|$. The attention weights p_{ki} allow each query to emphasize a potentially different retrieval-relevant residue pattern, while the weighted by η prevents the representation from depending exclusively on a small subset of residues. The learned logit scale $\gamma = 10\cdot\sigma(\rho_\gamma)$ controls the concentration of the attention distribution, with larger values producing more localized pooling. We use $K = 4$ by default, providing multiple adaptive summaries of the enzyme alongside the fixed pooling views (see Appendix F.2 for the K ablation).

Feature-Wise Gated Fusion. Each feature view $\mathbf{x}_e^v$ for $v \in \mathcal{I} = \{P, S, 1, \dots, K\}$ is projected into a 256-dimensional space via an independent MLP and normalized: $\mathbf{y}_e^v = \mathrm{norm}(F_v(\mathbf{x}_e^v))$, where $\mathrm{norm}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2$. A gating network maps their concatenation to view weights $\omega_{e,vd}$ normalized coordinate-wise across views v. The fused representation is:

$$\bar{y}_{e,d} = \sum_{v\in\mathcal{I}}\omega_{e,vd}y_{e,d}^v, \qquad d = 1,\dots,256. \tag{3}$$

Two MLPs project the global $\mathbf{x}_e^P$ and the fused representation $\bar{\mathbf{y}}_e$ into normalized 512-dimensional vectors $\mathbf{g}_e$ and $\mathbf{f}_e$. We combine them as $\mathbf{b}_e = \mathrm{norm}(\mathbf{g}_e + s\mathbf{f}_e)$, where the learned scalar s controls how strongly the fused features contribute to the base embedding. More details in Appendix C.1.

3.3 Reaction Encoder

The reaction encoder fuses frozen ReactionT5v2 sequence features (Sagawa & Kojima, 2025), 3D and chiral molecular representations from Uni-Mol2 and ChIRo (Ji et al., 2024; Adams et al., 2022),

and explicit chemical descriptors derived from reaction SMILES. The explicit descriptor concatenates Morgan fingerprints (Rogers & Hahn, 2010), aggregated physicochemical properties, participant-set statistics, and train-fitted cofactor indicators. We provide a more in-depth description of the set-chemistry composition and the implementation of each module in Appendices C.4 and C.2.

Given a reaction r, its representation in the SMILES format natively encodes the reactant and product sets, denoted with the underscript $-$ and $+$ respectively. We exploit this reaction structure when extracting its molecular features for each backbone $m \in \{\text{Uni-Mol2}, \text{ChIRo}\}$, where independent reactant and product attention pools extract features $\mathbf{p}_-^m$ and $\mathbf{p}_+^m$. Their combination vector, $\mathbf{c}_r^m = [\mathbf{p}_-^m; \mathbf{p}_+^m; \mathbf{p}_+^m - \mathbf{p}_-^m; |\mathbf{p}_+^m - \mathbf{p}_-^m|]$. The vector retains both information about reaction components alongside their signed and absolute differences, encoding the molecular change in the reaction. The modality views are mapped via independent MLPs into 512-dimensional tokens and aggregated via modality attention pooling to yield the final ℓ_2-normalized reaction representation $\mathbf{b}_r \in \mathbb{R}^{512}$.

3.4 Training objectives

Anchor-Balanced Multi-Positive Encoder Alignment. The alignment loss for the reaction and enzyme encoders uses a bidirectional, decoupled multi-positive objective designed to accommodate the many-to-many structure of enzyme–reaction associations. An association batch first determines the distinct enzyme and reaction endpoints under consideration. For each anchor $a \in \{\mathcal{A}_{B,X}\}$ with $X \in \{R, E\}$, which may be either a reaction or an enzyme, $P_B(a)$ contains all opposite-modality endpoints in the batch B that have a recorded training association with a. The remaining endpoints, with no association with a in the training data, form $N_B(a)$. The objective is:

$$\mathcal{L}_{\text{align}} = \sum_{X \in \{R,E\}} \sum_{a \in \mathcal{A}_{B,X}} \sum_{p \in P_B(a)} \frac{\log\Big(1 + \sum_{n \in N_B(a)} \exp[s_{an} - s_{ap}]\Big)}{2|\mathcal{A}_{B,X}||P_B(a)|}. \tag{4}$$

where $\mathcal{A}_{B,X}$ is the set of type-X anchors for which $P_B(a)$ and $N_B(a)$ are nonempty, and $s_{ax} = \mathbf{b}_a^\top \mathbf{b}_x / \tau_1$ is the temperature-scaled cosine similarity between the base encoder embeddings, with $\tau_1 = 0.2$. Each term encourages a positive associated point $p \in P_B(a)$ to receive a higher similarity than every negative endpoint $n \in N_B(a)$ in the same batch. Because each positive is compared independently against the shared negative set, recorded partners do not compete with one another, allowing multiple known reactions or enzymes associated with the same anchor to remain close in the joint representation space. Averaging over $P_B(a)$ further prevents anchors with more recorded partners from dominating the objective. Importantly, elements of $N_B(a)$ represent unobserved training pairs rather than confirmed biological incompatibilities; thus, incompleteness in the association data may still introduce false negatives.

Residual Refinement. After the initial encoder alignment, we freeze both encoder backbones and learn a lightweight, modality-specific refinement of the shared embedding space using supervision from the complete training association graph. Specifically, we attach separate residual heads h_E and h_R to the enzyme and reaction encoders, respectively, each producing a 512-dimensional residual embedding. For an endpoint a of modality $X \in \{R, E\}$, the refined embedding is $\mathbf{u}_a = \text{norm}(\mathbf{b}_a + \kappa h_X(\mathbf{b}_a))$, where $\mathbf{b}_a$ is the frozen, normalized modality-specific embedding and $h_X(\mathbf{b}_a)$ is its learned modality-specific residual update. The coefficient $\kappa = 0.2$ limits the magnitude of the update, while the final normalization keeps the refined representation on the unit hypersphere, so that dot products correspond to cosine similarities. Let $\mathcal{T}_X$ denote the complete set of type-X training endpoints and $\bar{X}$ the modality opposite to X. For each anchor a, $P_{\mathcal{T}}(a) \subseteq \mathcal{T}_{\bar{X}}$ contains all its recorded training partners, whereas $N_{\mathcal{T}}(a) = \mathcal{T}_{\bar{X}} \setminus P_{\mathcal{T}}(a)$ contains the remaining, unobserved endpoints.We denote by $\mathcal{A}_{\mathcal{T},X} \subseteq \mathcal{T}_X$ the type-X anchors for which $P_{\mathcal{T}}(a)$ and $N_{\mathcal{T}}(a)$ are nonempty. The refined similarity $\bar{s}_{ax} = \mathbf{u}_a^\top \mathbf{u}_x / \tau_2$ is the temperature-scaled cosine similarity between the refined embeddings, with $\tau_2 = 0.2$.

$$\mathcal{L}_{\text{CE}} = -\frac{1}{2} \sum_{X \in \{R,E\}} \frac{1}{|\mathcal{A}_{\mathcal{T},X}|} \sum_{a \in \mathcal{A}_{\mathcal{T},X}} \frac{1}{|P_{\mathcal{T}}(a)|} \sum_{p \in P_{\mathcal{T}}(a)} \log \frac{\exp(\bar{s}_{ap})}{\sum_{x \in \mathcal{T}_{\bar{X}}} \exp(\bar{s}_{ax})}. \tag{5}$$

where $\bar{s}_{ax} = \mathbf{u}_a^\top \mathbf{u}_x / \tau_2$ is the temperature-scaled cosine similarity between the refined embeddings, with $\tau_2 = 0.2$. For each anchor, this objective increases the scores of all its recorded training

associates relative to the complete set of opposite-modality training endpoints. In contrast to the batch-local alignment objective, which uses $P_B(a)$ and $N_B(a)$, in this case the normalization is computed over all $x \in \mathcal{T}_{\bar{X}}$. This full-graph supervision allows the residual heads to refine the global retrieval geometry. To prevent the full-graph refinement from excessively distorting the representations learned during encoder alignment, we introduce an identity-preservation penalty $\mathcal{L}_{\text{id}}$:

$$\mathcal{L}_{\text{id}} = \frac{1}{2} \sum_{X \in \{R,E\}} \frac{1}{|\mathcal{A}_{\mathcal{T},X}|} \sum_{a \in \mathcal{A}_{\mathcal{T},X}} \left(1 - \mathbf{u}_a^\top \mathbf{b}_a\right). \tag{6}$$

Because both $\mathbf{u}_a$ and $\mathbf{b}_a$ are normalized, $1 - \mathbf{u}_a^\top \mathbf{b}_a$ is their cosine distance. The penalty therefore preserve the geometry learned during the encoder alignment. The complete objective is

$$\mathcal{L}_{\text{ref}} = \mathcal{L}_{\text{CE}} + \psi \mathcal{L}_{\text{id}}, \tag{7}$$

where $\psi = 10$ controls the trade-off between adapting the embeddings to the complete training association structure and preserving the encoder aligned representation space.

3.5 Scoring Function

At inference, we set the enzyme fusion multiplier to $\alpha_{\text{inf}} = 2$ and halve the training-time residual coefficient, such that $\kappa_{\text{inf}} = \kappa/2 = 0.1$. We define the inference-time enzyme base embedding as $\tilde{\mathbf{b}}_e = \text{norm}(\mathbf{g}_e + \alpha_{\text{inf}} s\mathbf{f}_e)$, retain the reaction base embedding $\mathbf{b}_r$ unchanged, and define the refined embeddings as $\mathbf{z}_e = \text{norm}(\tilde{\mathbf{b}}_e + \kappa_{\text{inf}} h_E(\tilde{\mathbf{b}}_e))$ and $\mathbf{z}_r = \text{norm}(\mathbf{b}_r + \kappa_{\text{inf}} h_R(\mathbf{b}_r))$. The final compatibility score is the cosine similarity between the refined embeddings:

$$S(r, e) = \mathbf{z}_r^\top \mathbf{z}_e. \tag{8}$$

Here, $\tilde{\mathbf{b}}_e$ differs from the training-time enzyme base embedding $\mathbf{b}_e = \text{norm}(\mathbf{g}_e + s\mathbf{f}_e)$ only through the fusion multiplier $\alpha_{\text{inf}} = 3$. Since $\mathbf{z}_r$ and $\mathbf{z}_e$ are unit-normalized, $S(r, e)$ is their cosine similarity and is used unchanged in both retrieval directions. All model components are trained separately for each benchmark split (Appendix D). We report the sensitivity to α_{inf} and κ_{inf} in Appendix F.3.

4 Functional Organized Embeddings

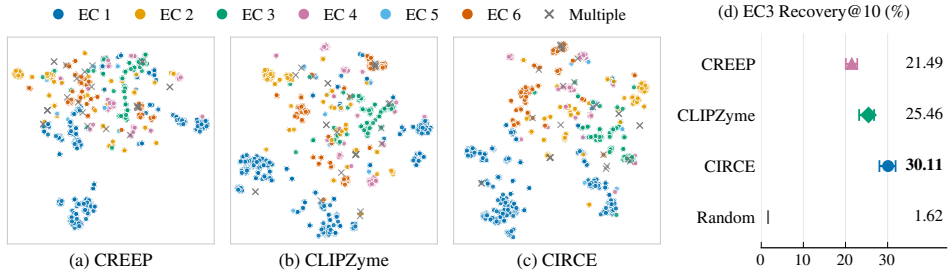


Figure 2: **Functional embedding organization.** *(a–c)* Independent t-SNE projections of EnzymeMap test-set enzyme representations, colored by first-level EC class; gray crosses denote enzymes with multiple EC annotations. Compared with competing baselines, `CIRCE` produces a clearer organization of the embedding space according to enzyme function. *(d)* Enzyme Recovery@10 for 1,067 EnzymeMap reaction queries, evaluated using third-level EC annotations. Gray indicates the expected performance under random retrieval. The higher recovery achieved by `CIRCE` indicates stronger functional alignment between reaction and enzyme representations.

The application of such an alignment method to the joint embedding space allows shaping the space to reflect functional relationships between enzymes and reactions. In this section, we evaluate the functional organization of the joint embedding space, comparing `CIRCE` to CREEP and CLIPZyme, evaluating the representation of EnzymeMap samples. To enable a fair comparison, all methods use the same enzymes and the same queries and candidate exclusions for functional recovery.
We focus our analysis on two aspects: *(a)* whether enzymes are grouped by function in the embedding space, and *(b)* whether reactions can be connected to functionally compatible enzymes.

Enzyme Organization. The aim of this experiment is to verify whether the joint embedding space is actually improving in the organization by enzyme function. Following FGW-CLIP (Zhou et al., 2025), we adopt the test data from EnzymeMap, containing $1,357$ enzymes associated with $1,067$ reactions. For each model, we extract the final ℓ_2-normalized output embeddings ($\mathbf{b}_e \in \mathbb{R}^{512}$) and apply t-SNE in a strictly unsupervised manner using cosine distance, a perplexity of 30, and identical random seed initializations. We in particular consider a subset of Enzyme Commission (EC) labeled 720 enzymes, annotated *post-hoc*, focusing on the first level of the EC hierarchy. `CIRCE` enzyme embeddings form local groups enriched for shared catalytic classes. On this cohort, as also shown in Figure 2(a–c), `CIRCE` also achieves higher class-balanced EC1 silhouette and EC3 neighbor agreement than CREEP and CLIPZyme in the original embedding space. Complete hyperparameter configurations, sensitivity analyses, and quantitative results are provided in Appendix G.2 and G.3.

Functional Recovery. In this setting, we aim to evaluate the relationship between reactions and enzyme representation space. We employ the $1,067$ test reactions from EnzymeMap and the 131,409 EC3-annotated accessions in the full 261,907-accession screening bank For each reaction, we get the associated enzymes and their EC3 labels, the first three EC levels. We then remove these, alongside enzymes sharing their sequences, and members of their detected sequence components (at least 50% identity and 80% coverage of both sequences). This filtering allows us to test whether functional recovery persists after removing known partners and close sequence alternatives. Each model ranks the remaining EC3-annotated candidates using its unchanged reaction–enzyme score. We define Recovery@10 as the percentage of top-ten candidates sharing at least one EC3 label with the query's associated enzymes. We first average within each EC3 class and then across classes, giving each function equal weight. Queries without a compatible alternative receive zero. We compare with expected random retrieval and the same models listed above. `CIRCE` achieves 30.11% Recovery@10, compared with 25.46% for CLIPZyme (Figure 2d). This result supports stronger cross-modal functional organization beyond recorded partner identities; the EC3 agreement showcases that the retrieval procedure, given a reaction, identifies functionally compatible candidates. More details on the uncertainty estimation and results are in Appendix G.1.

5 Experiments

We evaluate `CIRCE` on two complementary tasks: recovering enzymes from a large screening library and retrieving enzyme–reaction associations. EnzymeMap (Heid et al., 2023) measures early recovery under a reaction-rule split, whereas ReactZyme (Hua et al., 2024) evaluates retrieval under reaction-similarity, enzyme-similarity, and temporal splits. We also add a case study of D-fructose-to-D-tagatose isomerization, on both a curated 144-entry panel and two larger candidate banks, adopting the model trained on the benchmark ReactZyme experiments. More details in Appendix H.

5.1 Enzyme Screening

Datasets. We evaluate enzyme screening performance on the EnzymeMap benchmark (Heid et al., 2023), which comprises $46,356$ enzyme-catalyzed reactions spanning $16,776$ unique chemical reactions, $12,749$ enzymes across $2,841$ EC numbers, and 394 reaction rules. To assess generalization to unseen reactions without data leakage, we adopt the reaction-rule split protocol introduced in CLIPZyme (Mikhael et al., 2024). The dataset is partitioned by reaction rule IDs using an 80/10/10 ratio, resulting in $34,427$ training, $7,287$ validation, and $4,642$ test samples.

The screening task evaluates a model's ability to rank a candidate enzyme library for each query reaction. According to the established evaluation protocol (Mikhael et al., 2024; Zhou et al., 2025), we query $1,521$ unseen test reactions against $261,907$ candidate enzyme identifiers, built integrating enzymes from Brenda release 2022_2 (Chang et al., 2021) and UniProt release 2022_1 (Consortium, 2023) whose sequences are shorter than 650 residues. Furthermore, we evaluate a training-ID-excluded setting that removes training enzyme identifiers from the candidate pool, retaining $1,337$ queries with positive targets among $252,113$ candidates. This exclusion operates strictly on exact identifier matches and does not enforce sequence- or homology-based disjointness.

Baselines and Metrics. We compare our `CIRCE` to the released CLIPZyme (Mikhael et al., 2024) architecture trained on the EnzymeMap dataset. We test the models on the same setting described

Table 1: Reported performance on EnzymeMap screening. Published results (†) are taken from Zhou et al. (2025); *CIRCE* uses the validation-frozen dictionary-free reference recipe and exceeds the displayed published values in both candidate settings.

	Full library				Training-ID-excluded			
Method	$BEDROC_{85}$	$BEDROC_{20}$	EF_5	EF_{10}	$BEDROC_{85}$	$BEDROC_{20}$	EF_5	EF_{10}
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	39.13	58.86	13.40	7.81
FGW-CLIP†	48.66	66.69	14.91	8.18	45.14	61.43	13.57	7.61
CREEP	42.82	58.76	13.01	7.09	37.27	53.55	11.97	6.62
CIRCE	**53.05**	**71.32**	**16.03**	**8.62**	**49.64**	**68.70**	**15.48**	**8.45**

above. We also include as a baseline the FGW-CLIP results reported by Zhou et al. (2025), which could not have been reproduced due to the lack of access to the original implementation.

We report for each model the BEDROC (Boltzmann-Enhanced Discrimination of ROC) (Truchon & Bayly, 2007) and the enrichment factor (EF) metrics, which measure early retrieval performance and the concentration of true positives among top-ranked results, respectively. Higher values indicate better performance for both metrics. We compute the query-averaged $BEDROC_{85}$, $BEDROC_{20}$, and enrichment factors (EF) at the top 5% and 10% of the predictions (EF5 and EF10), according to the established evaluation protocol (Mikhael et al., 2024; Zhou et al., 2025).

Results. *CIRCE* shows a consistent improvement over CLIPZyme in both candidate configurations (Table 1). In the full library setting, $BEDROC_{85}$ increases from 44.69% to 53.05%, while EF_5 rises from 14.09 to 16.03. The improvements persist in the more challenging training-ID-excluded setting, where $BEDROC_{85}$ increases from 39.13% to 49.64%, and EF_5 increases from 13.40 to 15.48. Thus, the gains remain substantial after removing training enzyme identifiers from the candidate library. This establishes improved screening among identifiers absent from training; identical or homologous sequences may remain under other identifiers, so this setting alone does not establish generalization to unseen enzyme sequences.

5.2 Enzyme–Reaction retrieval

Dataset. We evaluate retrieval performance on the ReactZyme dataset (Hua et al., 2024), a benchmark composed from SwissProt (Boeckmann et al., 2003) and Rhea (Bansal et al., 2022) curated enzyme–reaction associations. The published dataset statistics are $178,463$ enzyme–reaction associations, $178,327$ unique enzymes, and $7,726$ unique reactions. We use its three evaluation protocols: Enzyme-Sim, Reaction-Sim, and Time. Time separates associations at the 31 December 2010 cutoff; Enzyme-Sim requires at least 60% sequence difference between training and test enzymes; Reaction-Sim holds out reactions with no overlap between training and test sets. Within each released training partition, we reserve 10% of association rows for validation using seed 42. The materialized protocols used in our experiments contain $178,459$ association rows each. Their training/validation/test counts are $149,554/16,618/12,287$ for Time, $152,748/16,972/8,739$ for Enzyme-Sim, and $147,393/16,377/14,689$ for Reaction-Sim (Table 6). See also Appendix E.

Baselines and Metrics. The main comparison follows Zhang et al. (2026): it includes ReactZyme encoder combinations, fingerprint, GNN, Bi-RNN, CLIPZyme (Mikhael et al., 2024), and both TIGER variants, for which we report the paper results due to the lack of publicly available implementations. We also retrained in the same setting CREEP (Yang et al., 2024), including this as an additional baseline. We retain the published rows at their reported precision and add *CIRCE* from its stored predictions. The ReactZyme encoder superscripts are defined in Table 2. Extended tables in the appendix include FGW-CLIP and further molecular encoder combinations from Zhou et al. (2025). For further comparison on the baselines, see Appendix E.2. We evaluate the performanceusing the

Table 2: Bidirectional retrieval performance on the Time, Enzyme-Sim, and Reaction-Sim splits of ReactZyme, following the evaluation protocol in Zhang et al. (2026). Published baseline results are reported at their original precision. Superscripts denote the reaction/enzyme encoder pair: [i] UniMol-3D/ESM, [ii] MAT-2D/ESM, and [iii] UniMol-3D/SaProt. E→R denotes reaction retrieval given an enzyme, whereas R→E denotes enzyme retrieval given a reaction.

	Time				Enzyme-Sim				Reaction-Sim			
	E→R		R→E		E→R		R→E		E→R		R→E	
Method	**Hit@1**	**MRR**	**Hit@1**	**MRR**	**Hit@1**	**MRR**	**Hit@1**	**MRR**	**Hit@1**	**MRR**	**Hit@1**	**MRR**
ReactZyme[i]	0.291	0.410	0.168	0.140	0.727	0.811	0.409	0.293	0.091	0.201	0.135	0.134
ReactZyme[ii]	0.325	0.218	0.218	0.179	0.599	0.728	0.362	0.259	0.109	0.199	0.093	0.096
ReactZyme[iii]	0.092	0.159	0.056	0.054	0.600	0.723	0.348	0.256	0.094	0.194	0.114	0.104
Fingerprint	0.236	0.298	0.144	0.117	0.579	0.639	0.255	0.204	0.094	0.194	0.114	0.104
GNN[i]	0.359	0.495	0.205	0.163	0.711	0.802	0.393	0.284	0.110	0.201	0.124	0.113
GNN[iii]	0.251	0.345	0.133	0.112	0.633	0.746	0.366	0.263	0.096	0.197	0.092	0.105
Bi-RNN[i]	0.354	0.494	0.254	0.211	0.811	0.875	0.509	0.387	0.109	0.197	0.124	0.121
Bi-RNN[ii]	0.391	0.530	0.265	0.227	0.815	0.886	0.589	0.456	0.118	0.240	0.171	0.170
CLIPZyme[i]	0.263	0.394	0.133	0.131	0.755	0.855	0.357	0.283	0.131	0.194	0.130	0.125
CLIPZyme[ii]	0.304	0.436	0.176	0.168	0.549	0.697	0.334	0.204	0.124	0.220	0.146	0.152
TIGER (ESM2Text)[†]	0.581	0.690	0.454	0.366	**0.931**	0.956	0.792	0.592	0.416	0.518	0.430	0.319
TIGER (ProtT3)[†]	0.583	0.683	0.454	0.372	0.908	0.940	0.784	0.579	0.386	0.472	0.428	0.337
CREEP	0.605	0.704	0.465	0.383	0.923	0.952	0.755	0.578	0.318	0.429	0.368	0.303
CIRCE	**0.668**	**0.764**	**0.564**	**0.456**	*0.930*	**0.960**	**0.842**	**0.636**	**0.425**	**0.528**	**0.477**	**0.377**

benchmark implementation's all-positive mean reciprocal rank (MRR):

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{|P(q)|} \sum_{p \in P(q)} \frac{1}{\mathrm{rank}_q(p)}, \tag{9}$$

where Q is the test query set, $P(q)$ its known positive candidates, and $\mathrm{rank}_q(p)$ the one-based rank of p. Following the layout of Zhou et al. (2025), we also report Top-k (Hit@k), Top-k-N (precision@k), and the query-averaged mean rank of positives, for $k \in \{1, 5, 10, 20\}$. Metric definitions and published-source checks are given in Appendix E.3.

Results. Table 2 compares Hit@1 and MRR in both retrieval directions across all three ReactZyme splits. On the Time split, *CIRCE* achieves E→R and R→E MRR values of 0.764 and 0.456, respectively, compared with 0.690 and 0.366 for TIGER (ESM2Text). On Enzyme-Sim, *CIRCE* obtains E→R/R→E MRR values of 0.960/0.636, while on Reaction-Sim it reaches 0.528/0.377. Thus, *CIRCE* achieves the highest MRR among the displayed methods in all six evaluation settings. For Hit@1, *CIRCE* achieves overall the highest value, with the only exception of Enzyme-Sim E→R, where its score of 0.930 is marginally below the 0.931 obtained by TIGER (ESM2Text). Detailed Hit@k, precision@k, and mean-rank results are reported in Tables 7, 8, and 9 in the appendix E.3.

6 Conclusions

We introduced *CIRCE*, a dual encoder that combines chemical descriptors with global, functional-site-guided, and learned residue-level views. Multi-positive alignment and full-graph refinement enable efficient bidirectional retrieval through independently computed enzyme and reaction embeddings. Using a consistent training recipe fitted separately to each benchmark partition, *CIRCE* achieves state-of-the-art performance for enzyme screening on EnzymeMap and enzyme–reaction bidirectional retrieval on ReactZyme. Embedding-space analyses further show that *CIRCE* organizes enzymes more clearly by function and retrieves functionally compatible candidates beyond known pairings and close sequence homologs. Together, these results suggest that incorporating biochemical knowledge into both the representations and the learning objective produces a more functionally meaningful shared space for enzyme–reaction retrieval.

ETHICS STATEMENT

The experiments reported in this work are computational evaluations of enzyme–reaction retrieval using existing protein and reaction resources. The intended application is to prioritize candidate enzymes for research in biocatalysis. Retrieval scores indicate learned compatibility, not verified catalytic activity, substrate specificity, or experimental safety; any downstream application requires independent experimental validation and appropriate biosafety assessment. Incomplete annotations and uneven coverage of enzyme families and reaction classes may bias rankings toward well-studied biology and chemistry. Enzyme-retrieval tools also have potential dual-use implications, so downstream use should consider risks associated with the target chemistry and biological system. Reuse or redistribution of source data and pretrained models should respect their applicable licenses and attribution requirements. To support transparent interpretation, we document data overlap, preprocessing controls, and the use of benchmark tests during architecture development in Appendix B.

AI USE STATEMENT

Generative AI tools assisted with drafting and revising manuscript text, organizing sections, searching and summarizing related literature, checking and formatting bibliography entries, proposing the title and method name, and adapting the LaTeX template. The authors remain responsible for the accuracy and originality of the final manuscript, including all AI-assisted text, claims, citations, and artifacts.

REPRODUCIBILITY STATEMENT

The architecture and alignment objective are described in Section 3, and the candidate-ranking protocol and evaluation metrics in Section 5.1. Appendices B and B document the data splits, candidate pools, train-fitted preprocessing, and overlap audits. Training settings, random seeds, checkpoint-selection criteria, baseline adaptations, and sensitivity protocols are provided in Appendices D, E, and F.2; the upstream functional-site scorer and its selection procedure are described in Appendix A. These details support reproduction of the reported configuration. The model-selection limitations discussed in Appendix B.

REFERENCES

Keir Adams, Lagnajit Pattanaik, and Connor W. Coley. Learning 3d representations of molecular chirality with invariance to bond rotations. In *International Conference on Learning Representations*, 2022.

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.

Uwe T Bornscheuer, GW Huisman, RJ Kazlauskas, S Lutz, JC Moore, and K Robins. Engineering the third wave of biocatalysis. *Nature*, 485(7397):185–194, 2012.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735.

Antje Chang, Lisa Jeske, Sandra Ulbrich, Julia Hofmann, Julia Koblitz, Ida Schomburg, Meina Neumann-Schaal, Dieter Jahn, and Dietmar Schomburg. Brenda, the elixir core data resource in 2021: new developments and updates. *Nucleic acids research*, 49(D1):D498–D508, 2021.

Jiajun Chen, Dawei Ni, Yingying Zhu, Wei Xu, Tarek A. A. Moussa, Wenli Zhang, and Wanmeng Mu. Discovery of a thermostable tagatose 4-epimerase powered by structure- and sequence-based protein clustering. *Journal of Agricultural and Food Chemistry*, 72(33):18585–18593, 2024. doi: 10.1021/acs.jafc.4c03793.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 01 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac1052. URL https://doi.org/10.1093/nar/gkac1052.

Alperen Dalkiran, Ahmet Sureyya Rifaioglu, Maria Jesus Martin, Rengul Cetin-Atalay, Volkan Atalay, and Tunca Doğan. Ecpred: a tool for the prediction of the enzymatic functions of protein sequences based on the ec nomenclature. *BMC bioinformatics*, 19(1):334, 2018.

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.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Chengyue Gong, Daiheng Zhang, Jeffrey Ouyang-Zhang, Qiang Liu, Adam Klivans, and Daniel Diaz. Enzyme classification via semi-supervised functional residue learning. *bioRxiv*, pp. 2026–02, 2026.

Esther Heid, Daniel Probst, William H. Green, and Georg K. H. Madsen. EnzymeMap: curation, validation and data-driven prediction of enzymatic reactions. *Chemical Science*, 14(48):14229–14242, 2023. doi: 10.1039/D3SC02048G.

Chenqing Hua, Bozitao Zhong, Sitao Luan, Liang Hong, Guy Wolf, Doina Precup, and Shuangjia Zheng. Reactzyme: A benchmark for enzyme-reaction prediction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 26415–26442. Curran Associates, Inc., 2024. doi: 10.52202/079017-0832.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. Uni-Mol2: Exploring molecular pretraining model at scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 46956–46978, 2024.

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.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

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. doi: 10.1038/s41929-026-01478-y.

Peter Mikhael, Itamar Chinn, and Regina Barzilay. CLIPZyme: Reaction-conditioned virtual screening of enzymes. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/mikhael24a.html.

Jens Nielsen and Jay D Keasling. Engineering cellular metabolism. *Cell*, 164(6):1185–1197, 2016.

Syed Asad Rahman, Sergio Martinez Cuesta, Nicholas Furnham, Gemma L Holliday, and Janet M Thornton. Ec-blast: a tool to automatically search and compare enzyme reactions. *Nature methods*, 11(2):171–174, 2014.

William JF Rieger, Mikael Boden, Frances Arnold, and Ariane Mora. Squidly: enzyme catalytic residue prediction harnessing a biology-informed contrastive learning framework. *bioRxiv*, pp. 2025–06, 2025.

Jason W. Rocks, Dat P. Truong, Dmitrij Rappoport, Samuel Maddrell-Mander, Daniel A. Martin-Alarcon, Toni M. Lee, Steven Crossan, and Joshua E. Goldford. Dual-encoder contrastive learning accelerates enzyme discovery. *Proceedings of the National Academy of Sciences*, 123(12):e2520070123, 2026. doi: 10.1073/pnas.2520070123.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 05 2010. ISSN 1549-9596. doi: 10.1021/ci100050t.

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.

Tatsuya Sagawa and Ryosuke Kojima. Reactiont5: a pre-trained transformer model for accurate chemical reaction prediction with limited data. *Journal of Cheminformatics*, 17(1):126, 2025. doi: 10.1186/s13321-025-01075-4. URL https://doi.org/10.1186/s13321-025-01075-4.

Roger A Sheldon and John M Woodley. Role of biocatalysis in sustainable chemistry. *Chemical reviews*, 118(2):801–838, 2018.

Kyung-Chul Shin, Tae-Eui Lee, Min-Ju Seo, Dae Wook Kim, Lin-Woo Kang, and Deok-Kun Oh. Development of tagaturonate 3-epimerase into tagatose 4-epimerase with a biocatalytic route from fructose to tagatose. *ACS Catalysis*, 10(20):12212–12222, 2020. doi: 10.1021/acscatal.0c02922.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35:1026–1028, 2017. doi: 10.1038/nbt.3988.

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):473, 2019.

F Altschul Stephen. Basic local alignment search tool. *J. mol. Biol.*, 215:403–410, 1990.

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.

Xiaorui Wang, Xiaodan Yin, Dejun Jiang, Huifeng Zhao, Zhenxing Wu, Odin Zhang, Jike Wang, Yuquan Li, Yafeng Deng, Huanxiang Liu, et al. Multi-modal deep learning enables efficient and accurate annotation of enzymatic active sites. *Nature communications*, 15(1):7348, 2024.

Yuxin Wang, Zijian Tan, Hongli Wei, Nan Zhang, Lifei Wang, Yifan Yang, Weidong Liu, and Leilei Zhu. Protein engineering of tagatose 4-epimerase for d-tagatose production. *Journal of agricultural and food chemistry*, 73(20):12353–12363, 2025.

Wenhao Xia, Shimeng Liu, Lihui Huang, Hao Wu, Tao Bai, Shanshan Wang, Fei Hao, Xiao Wang, and Wen Wang. Reshaping the binding pocket of D-tagaturonate epimerase UxaE to improve the epimerization activity of C4-OH for enabling green synthesis of D-tagatose. *Molecular Catalysis*, 566:114439, 2024. doi: 10.1016/j.mcat.2024.114439.

Jason Yang, Ariane Mora, Shengchao Liu, Bruce J Wittmann, Anima Anandkumar, Frances H 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, 2023a.

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, 2023b.

Wei Zeng, Xiuqin Li, Yunyun Yang, Jian Min, Jian-Wen Huang, Weidong Liu, Du Niu, Xuechun Yang, Xu Han, Lilan Zhang, et al. Substrate-binding mode of a thermophilic pet hydrolase and engineering the enzyme to enhance the hydrolytic efficacy. *Acs Catalysis*, 12(5):3033–3040, 2022.

Chengxin Zhang, Lydia 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.

Yuhang Zhang, Keyan Ding, Peilin Chen, Han Liu, Can Lin, Ruixi Chen, Shiqi Wang, and Qi Song. Tiger: Text-informed generalized enzyme-reaction retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35530–35548, 2026.

Zipeng Zhong, Jie Song, Zunlei Feng, Tiantao Liu, Lingxiang Jia, Shaolun Yao, Min Wu, Tingjun Hou, and Mingli Song. Root-aligned smiles: a tight representation for chemical reaction prediction. *Chemical Science*, 13(31):9023–9034, 2022.

Gengmo Zhou, Feng Yu, Wenda Wang, Zhifeng Gao, Guolin Ke, Zhewei Wei, and Zhen Wang. Multi-alignment contrastive learning for enzyme–reaction retrieval. *arXiv preprint arXiv:2512.08508*, 2025. doi: 10.48550/arXiv.2512.08508.

A SLEEC RETRAINING AND CHECKPOINT SELECTION

The functional-site prior used by `CIRCE` is obtained from a separately trained SLEEC Stage-1 residue classifier (Gong et al., 2026). This upstream training stage is independent of the reaction–enzyme retrieval objective. After checkpoint selection, the SLEEC classifier is frozen and only its residue-level logits are supplied to `CIRCE` as a soft attention prior.

Residue Annotations and Conservation Labels. We construct the supervised dataset from a local M-CSA export containing reference catalytic-residue annotations and their corresponding UniProt (Consortium, 2023) sequences. We assign positive labels to annotated catalytic positions and treat the remaining retained positions as negatives. Ambiguous amino acids are mapped to `X`. Sequences longer than 1,022 residues are deterministically reduced to 1,022 positions by retaining the N-terminal quarter, a centered middle half, and the C-terminal quarter; catalytic-residue indices are remapped accordingly.

The resulting dataset contains 983 proteins, 410,639 residue labels, and 4,776 annotated functional residues. Proteins are sorted, shuffled, and divided approximately equally at the protein level, producing 492 training proteins and 491 validation proteins. This local snapshot is larger than the 820-protein, 3,716-site dataset reported in the original SLEEC study (Gong et al., 2026). Our retraining therefore uses a larger M-CSA-derived snapshot and does not reproduce the original data curation. The training partition contains $201,195$ residues with $2,371$positives; validation contains $209,444$ residues with $2,405$ positives.

To complement these annotations, we generate a multiple-sequence alignment for each of the 983 query proteins using Ligns/MMseqs2 against either UniRef50 or UniRef90. The search uses three iterations, an E-value threshold of 0.1, sensitivity 7.5, at most 10,000 hits per query, and a maximum sequence identity of 0.95. The retained UniRef90 run uses monomer mode without alignment expansion. For query position i, we calculate the Shannon entropy

$$E_i = -\sum_{a \in \mathcal{A}} p_i(a) \log p_i(a), \tag{10}$$

where $\mathcal{A}$ is the set of 20 canonical amino acids and $p_i(a)$ is the empirical frequency of amino acid a in alignment column i. Gaps and non-canonical residues are excluded from the frequency calculation. For a protein of retained length L, the $\lceil 0.1L \rceil$ lowest-entropy positions are assigned pseudo-label $\widetilde{y}_i = 1$, and all remaining positions receive $\widetilde{y}_i = 0$. This produces 41,522 pseudo-positive positions among 410,639 positions. To prevent the pseudo-label term from being dominated by negatives, we retain all pseudo-positive positions and sample an equal number of pseudo-negatives with, resulting in an 83,044-residue auxiliary pool.

Classifier and Optimization. We compare frozen residue representations from ProtT5 (Elnaggar et al., 2022), ESM2-650M (Lin et al., 2023), and ESMC-6B (Candido et al., 2026). The corresponding residue dimensions are 1,024, 1,280, and 2,560. Representations are precomputed, and the protein language models are not updated during SLEEC training.

For each backbone b, the functional-residue logit is predicted by a two-layer multilayer perceptron:

$$u_i^{(b)} = \mathbf{w}_2^\top \operatorname{ReLU}\left(\mathbf{W}_1 \mathbf{H}_{e,i}^{(b)} + \mathbf{b}_1\right) + b_2, \tag{11}$$

where the hidden dimension is 256. We do not use layer normalization or dropout. The resulting heads contain 262,657, 328,193, and 655,873 trainable parameters for ProtT5, ESM2-650M, and ESMC-6B, respectively.

We train this classifier using a supervised M-CSA batch and a balanced pseudo-label batch at each optimization step. Let $u_i^{(s)}$ and $y_i^{(s)}$ denote a supervised logit and label, and let $u_j^{(p)}$ and $\widetilde{y}_j$ denote their pseudo-label counterparts. With confidence $c_j = \max\{\sigma(u_j^{(p)}), 1 - \sigma(u_j^{(p)})\}$, we minimize

$$\mathcal{L}_{\text{SLEEC}} = \frac{1}{B_s} \sum_{i=1}^{B_s} \operatorname{BCE}(u_i^{(s)}, y_i^{(s)}) + \frac{1}{B_p} \sum_{j=1}^{B_p} \mathbf{1}[c_j > 0.9] \operatorname{BCE}(u_j^{(p)}, \widetilde{y}_j). \tag{12}$$

Table 3: SLEEC Stage-1 development-set results. $F_1@0.5$ uses the default probability threshold, while $\tau^\star$, $P^\star$, $R^\star$, and $F_1^\star$ are selected on the threshold grid. The dagger marks the checkpoint used by `CIRCE`.

Backbone	MSA database	Selected step	$F_1@0.5$	$\tau^\star$	$P^\star$	$R^\star$	$F_1^\star$
ESM2-650M	UniRef50	28,500	0.4764	0.340	0.5437	0.4936	0.5174
ESM2-650M	UniRef90	38,000	0.4899	0.360	0.5474	0.4877	0.5158
ProtT5	UniRef50	38,500	0.5060	0.340	0.5508	0.5094	0.5293
ProtT5†	UniRef90	30,000	0.5006	0.329	0.5591	0.5052	0.5308
ESMC-6B	UniRef50	12,500	0.5540	0.360	0.5826	0.5721	0.5773
ESMC-6B	UniRef90	10,500	0.5548	0.380	0.6165	0.5435	**0.5777**

The confidence mask is determined by the classifier prediction rather than by the MSA entropy. Moreover, the masked pseudo-label loss is averaged over the entire pseudo-label batch, so its effective contribution varies with the fraction of examples satisfying the confidence criterion. We use unweighted binary cross-entropy for the supervised term despite the residue-level class imbalance.

All classifiers are trained for $50{,}000$ optimizer steps using AdamW with learning rate 10^{-5} and weight decay 10^{-2}. The supervised and pseudo-label batch sizes are both $8{,}192$ residues. Training uses four GPUs with data parallelism, and no learning-rate scheduler, warm-up, early stopping, gradient clipping, data augmentation, mixed-precision training, or positive-class weighting. The two data loaders are shuffled at construction, after which their batch sequences are cycled throughout optimization.

Checkpoint Selection. We evaluate on the M-CSA validation partition every 500 steps. In addition to reporting performance at the default probability threshold 0.5, we sweep the threshold set

$$\mathcal{T} = \{0.25, 0.28, 0.30, 0.32, 0.329, 0.34, 0.36, 0.371, 0.38, 0.40, 0.42, 0.45, 0.48, 0.50, 0.55\}. \tag{13}$$

For each run, the saved checkpoint maximizes residue-level validation F1 over both training step and $\tau \in \mathcal{T}$. Training nevertheless continues to $50{,}000$ steps. Table 3 reports the selected checkpoint from each complete run; an earlier $10{,}000$-step ESM2–UniRef90 pilot without threshold sweeping is omitted.

Table 3 shows that ESMC-6B–UniRef90 has the highest selected F1 (0.5777). For retrieval, `CIRCE` uses ProtT5–UniRef90 at step 30,000 (0.5308) to share the encoder's existing residue features and avoid another language-model pass. The frozen scorer supplies continuous logits; its evaluation threshold does not binarize downstream attention. Neither M-CSA labels nor the residue-classification loss enter retrieval-model optimization.

This comparison is conditional on the upstream training procedure. The MSA-derived auxiliary pool is generated for all 983 M-CSA query proteins, including proteins assigned to the supervised validation partition. Consequently, although the ground-truth M-CSA labels of the validation proteins are excluded from the supervised training branch, those sequences can contribute conservation-derived pseudo-labels to the auxiliary branch. Furthermore, both checkpoint and threshold selection use the same validation partition. We therefore interpret Table 3 as a development-set comparison for selecting an upstream functional-residue scorer, rather than as an independent estimate of unseen-protein generalization.

B Data construction and leakage controls

We construct the training and evaluation sets independently for the released Time, Enzyme-Sim, and Reaction-Sim ReactZyme protocols (Hua et al., 2024). For each protocol, we shuffle the indices of the released training association rows with seed 42, assign $\lfloor 0.9N \rfloor$ rows to training, and use the remainder for internal validation. Rows are written in their original order after the index split, and each reaction table is restricted to the reaction identifiers needed by its association subset. This row-level split provides an internal validation set for checkpoint selection. Because it partitions associations, enzymes and reactions may appear in different pairs on both sides; validation therefore does not reproduce the distribution shift imposed by the benchmark test split. Table 4 reports the resulting

exact-identity overlaps. Exact pair intersections are zero for every train–validation, train–test, and validation–test comparison, but sequences or reactions can recur in different pairs. Reaction-Sim has no exact train–test reaction overlap, whereas the other protocols retain substantial reaction reuse.

For evaluation, each subset uses the unique protein identifiers in its association table as enzyme candidates. At test time, E→R ranks every test enzyme against the unique test reactions, and R→E ranks every test reaction against the test enzyme candidates. The counts are reported in Table 6.

Feature Preprocessing. The explicit set-chemistry schema is fitted separately for each protocol using only its training reactions. In particular, the means and standard deviations of the 64 aggregated descriptor features and the cofactor vocabulary are estimated on training data. The resulting schema is then frozen and applied unchanged to validation and test reactions. Cofactors absent from the training vocabulary remain unmarked in the fixed 37-slot vector, so held-out reactions cannot introduce new coordinates.

Table 4: Exact entity overlap. "Seq." denotes identical protein sequences and "Rxn." identical reaction SMILES. Identifier counts are equal, and all exact pair intersections are zero.

	Train–val.		Train–test		Val.–test	
Protocol	Seq.	Rxn.	Seq.	Rxn.	Seq.	Rxn.
Time	11	1,711	32	2,535	3	1,061
Enzyme-Sim	18	2,124	3	1,563	0	1,104
Reaction-Sim	22	2,085	10	0	2	0

ProtT5 (Elnaggar et al., 2022), SLEEC (Gong et al., 2026), ReactionT5v2 Sagawa & Kojima (2025), UniMol2 (Ji et al., 2024), and ChIRo (Adams et al., 2022) features are computed from one sequence or reaction at a time with frozen encoders and adopted as cache files for all experiments.

Case 1 screening banks. For the exploratory D-fructose-to-D-tagatose screen, we additionally assembled two candidate banks independent of the workbook's 144-entry panel. The CLIPZyme screening catalogue (Mikhael et al., 2024) contained 222,985 distinct protein sequences; exclusion of 131,935 exact matches to the V4 training proteins left 91,050. We selected reviewed Swiss-Prot (Consortium, 2023) entries annotated with an EC 5 isomerase number, retained entries whose sequences matched the local feature cache, excluded exact V4 training-sequence matches, and collapsed identical sequences to obtain 5,176 candidates. Since neither filtered bank contained a resolved sequence from the workbook's top-25 list, we added all 23 distinct resolved sequences before scoring, giving final banks of 91,073 and 5,199 candidates, respectively. The Excel list has 25 rows but only 24 sequence-resolved rows; two share a sequence. Candidate construction used no model scores or workbook rank-based filtering of the background proteins.

This exclusion guarantees only exact protein-sequence disjointness from V4 training. It does not eliminate related homologues. Moreover, the two Case 1 query molecules co-occur in seven molecule sets in the V4 training table, although none of those strings exactly equals the two-molecule query. Hence the screen cannot establish reaction novelty or a homology-disjoint transfer result. The reviewed EC 5 restriction uses prior functional knowledge, and the workbook's ranked list provides a retrospective priority comparison rather than measured activity labels for the full candidate banks.

C Implementation details of *CIRCE*

We provide the architecture details of the encoders and residual heads introduced in Section 3 and illustrated in Figure 1. We write $\mathrm{n}(\mathbf{x}) = \mathbf{x}/\max(\|\mathbf{x}\|_2, \epsilon)$ for numerical normalization. All attention distributions exclude padded positions, and independently labelled networks have separate parameters.

C.1 Enzyme representation

For an enzyme sequence e with L retained residues, a frozen ProtT5 model provides contextual residue vectors $\mathbf{H}_e = [\mathbf{h}_1, \ldots, \mathbf{h}_L]^\top \in \mathbb{R}^{L\times 1024}$ (Elnaggar et al., 2022). Here $\mathcal{V}_e = \{1, \ldots, L\}$

indexes valid residues; e is suppressed on per-residue vectors and attention weights. We construct $K+2$ protein views: one global view, one view defined by a frozen SLEEC functional-residue scorer (Gong et al., 2026), and K learned residue views, with $K = 4$ by default. The scorer is a pretrained resource; it is not fitted to each retrieval target in phase 1.

Global and SLEEC-Guided Pooling. The global view is the mean of the retained residue embeddings,

$$\mathbf{x}_e^P = \frac{1}{L}\sum_{i=1}^{L}\mathbf{h}_i. \tag{14}$$

The frozen scorer computes a scalar logit ℓ_i for each residue using a $1024 \to 256 \to 1$ MLP with a ReLU hidden activation. We center these logits within the protein, clip them, and form a smoothed attention distribution:

$$\bar{\ell} = L^{-1}\sum_{i=1}^{L}\ell_i, \qquad \tilde{\ell}_i = \mathrm{clip}(\ell_i - \bar{\ell}, -10, 10), \tag{15}$$

$$p_i^{\mathrm{S}} = \frac{\exp(\tilde{\ell}_i)}{\sum_{j=1}^{L}\exp(\tilde{\ell}_j)}, \qquad a_i^{\mathrm{S}} = (1-\eta)p_i^{\mathrm{S}} + \frac{\eta}{L}, \quad \eta = 0.05, \tag{16}$$

$$\mathbf{x}_e^S = \sum_{i=1}^{L} a_i^{\mathrm{S}}\mathbf{h}_i. \tag{17}$$

The vector $\mathbf{x}_e^S \in \mathbb{R}^{1024}$ is a protein-level summary: it averages the residue embeddings $\mathbf{h}_i$ with nonnegative weights a_i^{S} that sum to one. The superscripts P and S identify the global ProtT5 and SLEEC-weighted features, respectively; superscripts $k \in \{1, \ldots, K\}$ identify the learned features. The threshold used for SLEEC diagnostics does not select or remove residues in this encoder.

Learned Auxiliary Views. We augment the global and SLEEC summaries with K attention-pooled views learned from enzyme–reaction associations. These views provide adaptive pooling of contextual residue features before feature-wise fusion.

A shared residue adapter maps each contextual residue to $d_h = 256$ dimensions:

$$\tilde{\mathbf{h}}_i = \mathrm{GELU}(\mathbf{W}_{\mathrm{res}}\,\mathrm{LN}(\mathbf{h}_i) + \mathbf{b}_{\mathrm{res}}), \qquad \mathbf{k}_i = \mathrm{n}(\mathbf{W}_{\mathrm{key}}\tilde{\mathbf{h}}_i), \qquad \mathbf{v}_i = \mathbf{W}_{\mathrm{val}}\tilde{\mathbf{h}}_i + \mathbf{b}_{\mathrm{val}}. \tag{18}$$

The unit-normalized key $\mathbf{k}_i$ determines residue i's similarity to a query; the value $\mathbf{v}_i$ contains the features aggregated into the view. The key projection $\mathbf{W}_{\mathrm{key}}$ is bias-free, while the value projection is affine. Queries $\mathbf{q}_k \in \mathbb{R}^{256}$ are orthogonally initialized model parameters shared across enzymes and optimized during phase 1. Normalizing each query gives $\hat{\mathbf{q}}_k = \mathrm{n}(\mathbf{q}_k)$. Each query then produces a pooled view $\mathbf{x}_e^k$:

$$\gamma = 10\,\mathrm{sigmoid}(\rho_\gamma), \qquad \gamma_{\mathrm{init}} = 4, \tag{19}$$

$$p_{ki} = \frac{\exp\big(\gamma\,\hat{\mathbf{q}}_k^\top\mathbf{k}_i\big)}{\sum_{j=1}^{L}\exp\big(\gamma\,\hat{\mathbf{q}}_k^\top\mathbf{k}_j\big)}, \qquad a_{ki} = (1-\eta)p_{ki} + \frac{\eta}{L}, \tag{20}$$

$$\mathbf{x}_e^k = \sum_{i=1}^{L} a_{ki}\mathbf{v}_i, \qquad k \in \{1, \ldots, K\}. \tag{21}$$

For each query, the softmax is taken over valid residues, so $p_{ki} \geq 0$ and $\sum_i p_{ki} = 1$. The learned scalar ρ_γ parameterizes a bounded scale $\gamma \in (0, 10)$: larger values concentrate attention on residues with higher query–key cosine similarity. The uniform mixture gives $a_{ki} \geq \eta/L$ and $\sum_i a_{ki} = 1$. With $\eta = 0.05$, the resulting view is a mixture of the attention-weighted mean (95%) and the uniform mean (5%) of the same value vectors. These weights quantify feature aggregation, not probabilities that residues are catalytic.

Thus, $\mathbf{q}_k$ defines a shared pooling function, while a_{ki} and $\mathbf{x}_e^k$ are enzyme-specific. The alignment objective updates the queries, adapter, and projections through the fused enzyme embedding (Appendix D). Views can pool distant residues and overlap, without assigned biochemical roles. Because

pooling and fusion use only enzyme features, the views and their fused representation can be cached once the encoder is frozen and reused across candidate reactions. SLEEC contributes a separate view, not a bias to the learned attention logits.

The number of such views is controlled by K. At $K = 0$, only global and SLEEC summaries enter the fusion module. With $K = 1$, all learned residue information passes through one summary; larger K can retain complementary patterns, but also increases query, projection, and gate parameters. Views may overlap, so a larger count does not guarantee additional useful information. The benchmark model uses $K = 4$. Appendix F.2 specifies the matched view-count study and its capacity changes; no universally optimal count or correspondence to four functional regions is assumed. The residue queries remain frozen throughout the refinement training phase.

Feature-Wise Gated Fusion. The extracted enzyme features are combined by feature-wise gated fusion. Each feature is independently projected into a common 256-dimensional space using a network of the form

$$F_{a\to b}(\mathbf{x}) = \mathbf{W}_2\,\mathrm{Drop}_{0.1}(\mathrm{GELU}(\mathbf{W}_1\,\mathrm{LN}(\mathbf{x}) + \mathbf{b}_1)) + \mathbf{b}_2, \tag{22}$$

whose hidden width is 512 and whose training dropout probability is 0.1. The input width is $a = 1024$ for views P, S and $a = 256$ for learned views. Define $\mathbf{y}_e^v = \mathrm{n}(F_v(\mathbf{x}_e^v))$ for $v \in \mathcal{I} = \{P, S, 1, \ldots, K\}$, and let $V = |\mathcal{I}| = K + 2$. A $256V \to 256 \to 256V$ gate MLP with GELU maps their concatenation to logits $\mathbf{G}_e \in \mathbb{R}^{V\times 256}$,. The gate softmax is taken over views *separately for each feature coordinate* $d = 1, \ldots, 256$:

$$\omega_{e,vd} = (1-\zeta)\frac{\exp(G_{e,vd})}{\sum_{u\in\mathcal{I}}\exp(G_{e,ud})} + \frac{\zeta}{K+2}, \qquad \zeta = 0.05, \tag{23}$$

$$\bar{y}_{e,d} = \sum_{v\in\mathcal{I}} \omega_{e,vd}\, y^v_{e,d}. \tag{24}$$

For each enzyme and coordinate, $\sum_v \omega_{e,vd} = 1$: each fused coordinate is a weighted average of that coordinate across the projected views. The floor is $\zeta/(K+2)$ per view and coordinate, hence $\zeta/6$ at the default $K = 4$.

Two additional networks of the form in Eq. equation 22 produce separately normalized global and fused outputs in $\mathbb{R}^{512}$:

$$\mathbf{g}_e = \mathrm{n}(F_g(\mathbf{x}_e^P)), \qquad \mathbf{f}_e = \mathrm{n}(F_f(\bar{\mathbf{y}}_e)), \tag{25}$$

$$s = 0.5\,\mathrm{sigmoid}(\rho_s), \qquad \mathbf{b}_e = \mathrm{n}(\mathbf{g}_e + s\mathbf{f}_e). \tag{26}$$

Here $\mathbf{g}_e$ is computed directly from the sequence mean, while $\mathbf{f}_e$ is computed from the combined views. Their separate projections both output 512-dimensional vectors, making the final addition well-defined. The scalar s is shared across proteins, is initialized to 0.1, and remains strictly between 0 and 0.5. It controls the strength of the entire fused branch; the weights $\omega_{e,vd}$ instead control the contributions of individual views within that branch. The direct global branch retains sequence context alongside the information supplied by fusion.

The pooling module uses the contextual content of ProtT5 vectors without adding positional encodings. Permuting already computed residue vectors and their masks together leaves this pooling operation unchanged, whereas changing the amino-acid sequence can change the contextual vectors. Neither a slot nor a coordinate of the final 512-dimensional vector is assigned a named catalytic function. Attention weights are learned aggregation weights, not validated active-site labels.

C.2 Reaction Encoder

We use four feature sources: a 768-dimensional ReactionT5v2 reaction embedding (Sagawa & Kojima, 2025), 768-dimensional Uni-Mol2 molecule embeddings (Ji et al., 2024), 256-dimensional ChIRo molecule embeddings (Adams et al., 2022), and a 617-dimensional fixed chemistry vector defined in Appendix C.4. Its four blocks summarize participant substructures, molecular properties, set statistics, and cofactor presence. Their feature extractors are frozen; molecular pooling, modality projections, and fusion are trainable.

Molecular Pooling. For one reaction (with index r suppressed), molecular modality $m \in \{\mathrm{UniMol2}, \mathrm{ChIRo}\}$, and side $s \in \{-, +\}$, let $\mathbf{x}^m_{s,j}$ be the embedding of molecule j. Independent linear attention scorers for the two sides produce

$$\pi^m_{s,j} = \frac{\exp((\mathbf{w}^m_s)^\top \mathbf{x}^m_{s,j} + b^m_s)}{\sum_t \exp((\mathbf{w}^m_s)^\top \mathbf{x}^m_{s,t} + b^m_s)}, \qquad \mathbf{p}^m_s = \sum_j \pi^m_{s,j} \mathbf{x}^m_{s,j}. \tag{27}$$

The configured `directional_delta` composition is

$$\mathbf{c}^m_r = [\mathbf{p}^m_-;\ \mathbf{p}^m_+;\ \mathbf{p}^m_+ - \mathbf{p}^m_-;\ |\mathbf{p}^m_+ - \mathbf{p}^m_-|]. \tag{28}$$

It has dimension 3072 for Uni-Mol2 and 1024 for ChIRo. Direction is represented by the ordered sides and signed difference when physical sides are supplied. The first two blocks preserve the pooled molecular context of each side; the last two expose the sign and magnitude of coordinate-wise contrasts. These contrasts are between learned summaries, rather than an explicit atom-mapped list of bond changes. Only Uni-Mol2 and ChIRo use this pooling: ReactionT5v2 and the fixed chemistry descriptor already provide one vector per reaction.

The interpretation of these pooled features depends on the reaction representation. EnzymeMap inputs retain their physical reactant and product sides. ReactZyme's released participant-set inputs use the established self-reaction adapter, placing the same complete set on both sides. Their difference on these self-reactions must not be interpreted as an observed chemical transformation direction.

Modality Fusion. For $m \in \{\mathrm{T5}, \mathrm{UniMol2}, \mathrm{ChIRo}, \mathrm{chem}\}$, let $\mathbf{c}^m_r$ denote the corresponding reaction input. Each independent modality MLP has widths $d_m \to 1024 \to 512 \to 512$, with ReLU after the first two linear layers. A shared layer-normalization module then gives the modality token

$$\mathbf{t}^m_r = \mathrm{LN}(\mathrm{MLP}_m(\mathbf{c}^m_r)) \in \mathbb{R}^{512}. \tag{29}$$

An attention MLP of widths $512 \to 512 \to 1$, with a ReLU hidden activation, scores the tokens. If $\mathcal{M}_r$ is the set of available modalities,

$$\alpha^m_r = \frac{\exp(A(\mathbf{t}^m_r))}{\sum_{n \in \mathcal{M}_r} \exp(A(\mathbf{t}^n_r))}, \quad \mathbf{t}_r = \sum_{m \in \mathcal{M}_r} \alpha^m_r \mathbf{t}^m_r, \quad \mathbf{b}_r = \mathrm{n}(F_{\mathrm{out}}(\mathbf{t}_r)). \tag{30}$$

The output MLP F_{out} has widths $512 \to 1024 \to 512$ and a ReLU hidden activation. Unavailable modalities are masked out of attention. Molecule attention therefore combines molecules within one modality and side, whereas modality attention combines the four reaction-level tokens. Reaction modality dropout and projection dropout are zero in the reported recipe. The resulting $\mathbf{b}_r$ and $\mathbf{b}_e$ inhabit a shared 512-dimensional retrieval space.

C.3 Residual refinement heads

The enzyme and reaction heads, h_E and h_R, have the same architecture and independent parameters. Each takes a frozen, normalized 512-dimensional encoder embedding. Layer normalization is followed by a linear projection to 1,024 hidden units, a GELU activation, and a linear projection back to 512 dimensions. Both linear layers include biases; the heads use no dropout. The output layer's weights and biases are initialized to zero, so each head initially predicts a zero correction. The refined embedding is

$$\mathbf{u}_a = \mathrm{n}(\mathbf{b}_a + \kappa h_X(\mathbf{b}_a)), \qquad X \in \{R, E\}, \quad \kappa = 0.2. \tag{31}$$

This initialization preserves the base embedding at the start of refinement. Only the residual-head parameters are updated in phase 2; the enzyme and reaction encoders remain frozen. At inference, the same heads are used with the fusion and residual coefficients specified in Eq. 8. Optimization settings are given in Appendix D.

C.4 Fixed chemistry descriptor

The fixed chemistry descriptor gives the reaction encoder an explicit summary of participant substructures, physicochemical properties, and recognized cofactors alongside its pretrained reaction and molecular embeddings. These quantities are available directly from participant SMILES; their construction does not require enzyme–reaction association labels. This supplies observable chemical

Table 5: Molecular properties used in the reaction chemistry descriptor. For each property, four reaction-level statistics are computed, producing a $16 \times 4 = 64$ dimensional descriptor.

#	Molecular property	RDKit calculation or definition
1	Exact molecular weight	`CalcExactMolWt`
2	Heavy atoms	Number of atoms with atomic number $Z > 1$
3	Heteroatoms	Number of atoms with $Z \notin \{0, 1, 6\}$
4	Formal charge	Net charge from `GetFormalCharge`
5	H-bond donors	`NumHDonors`
6	H-bond acceptors	`NumHAcceptors`
7	Polar surface area	Topological polar surface area from `CalcTPSA`
8	$\log P$	Crippen estimate from `MolLogP`
9	Rotatable bonds	`NumRotatableBonds`
10	Rings	`RingCount`
11	Aromatic fraction	Aromatic heavy atoms divided by $\max(N_{\text{heavy}}, 1)$
12	Stereocenters	Centers returned by `FindMolChiralCenters`, including unassigned centers
13	Wildcard atoms	Number of atoms with symbol `*`
14	Nitrogen atoms	Number of atoms with symbol `N`
15	Oxygen atoms	Number of atoms with symbol `O`
16	Phosphorus/sulfur atoms	Combined number of `P` and `S` atoms

cues whose contribution to retrieval is learned through the modality projection and fusion. The 617-dimensional chemistry vector consists of four blocks:

$$617 = 512 + 64 + 4 + 37. \tag{32}$$

The first block averages 512-bit radius-2 Morgan fingerprints of valid participant molecules. The second block applies sum, mean, maximum, and standard-deviation aggregation to 16 molecule descriptors, yielding 64 coordinates. Table 5 lists the properties in their exact order within the per-molecule vector $\mathbf{d}(m)$. Each is computed with RDKit from a valid participant molecule. The aggregate block concatenates all 16 sums, then all 16 means, maxima, and population standard deviations; these are summaries across participants, irrespective of reaction side.

The four set statistics are participant count, unique-participant fraction, valid-participant fraction, and wildcard-containing fraction among valid participants. The final block is a 37-position capacity for the recognized core-cofactor vocabulary fitted on training reactions; unused positions are zero. Only the 64 aggregated property coordinates are standardized. Fingerprints, set statistics, and cofactor indicators are not standardized in this construction.

More precisely, let $\mathcal{P}_r$ be the multiset of valid participant molecules, with $n_r = |\mathcal{P}_r| > 0$. For a molecule m, let $\boldsymbol{\phi}(m) \in \{0,1\}^{512}$ denote its radius-2 Morgan fingerprint and $\mathbf{d}(m) \in \mathbb{R}^{16}$ its property vector. Define

$$\bar{\boldsymbol{\phi}}_r = \frac{1}{n_r} \sum_{m \in \mathcal{P}_r} \boldsymbol{\phi}(m), \tag{33}$$

$$\mathbf{d}_r = \left[\sum_{m \in \mathcal{P}_r} \mathbf{d}(m);\ \text{mean}_{m \in \mathcal{P}_r}\, \mathbf{d}(m);\ \max_{m \in \mathcal{P}_r} \mathbf{d}(m);\ \text{std}_{m \in \mathcal{P}_r}\, \mathbf{d}(m) \right], \tag{34}$$

$$\mathbf{c}_r^{\text{chem}} = [\bar{\boldsymbol{\phi}}_r;\ (\mathbf{d}_r - \boldsymbol{\mu}_{\text{tr}}) \oslash \boldsymbol{\sigma}_{\text{tr}};\ \mathbf{s}_r^{\text{set}};\ \mathbf{q}_r^{\text{cof}}] \in \mathbb{R}^{617}. \tag{35}$$

Here $\boldsymbol{\mu}_{\text{tr}}, \boldsymbol{\sigma}_{\text{tr}} \in \mathbb{R}^{64}$ are the training means and stabilized standard deviations defined above, and $\oslash$ denotes elementwise division. All property aggregations and standardization are coordinate-wise, and std uses the population convention. The vector $\mathbf{s}_r^{\text{set}} \in \mathbb{R}^4$ contains the four set statistics above; $\mathbf{q}_r^{\text{cof}} \in \{0,1\}^{37}$ records presence of recognized core cofactors in the training vocabulary. An unseen cofactor does not create a new coordinate. Reactions with no valid participant have zero fingerprint and property blocks and an unavailable-chemistry mask; modality attention excludes this input.

The descriptor remains fixed during retrieval training. The fingerprint and property definitions, training-derived normalization, and cofactor vocabulary remain unchanged while the reaction encoder is trained. A learned chemistry-specific MLP maps $\mathbf{c}_r^{\text{chem}}$ to a 512-dimensional token, which participates in the same modality attention as the other three sources (Section C.2).

D TRAINING AND MODEL SELECTION

We train the same architecture independently on EnzymeMap and each ReactZyme splits. ProtT5, SLEEC, ReactionT5v2, UniMol2, and ChIRo remain frozen. Trainable residue-pooling, fusion, and reaction-projection components are initialized separately for each benchmark split. The main model uses $K = 4$ learned residue views.

Encoder Alignment. The alignment phase uses AdamW with learning rate 10^{-4}, weight decay 10^{-2}, association batch size 512, and inverse temperature $\beta = 1/\tau_1 = 5$. The benchmark tables use the end-of-epoch-20 checkpoints on ReactZyme and the end-of-epoch-15 checkpoint on EnzymeMap. Within each batch, endpoint identifiers are deduplicated, and all known training positives between the retained endpoints enter the alignment objective. The learned protein fusion scale s is initialized to 0.1 in all four benchmark fits.

Residual Refinement. The residual refinement phase freezes the aligned encoders and trains the two residual heads specified in Appendix C.3. We perform 100 full-graph updates using AdamW with learning rate 10^{-4}, weight decay 10^{-3}, temperature $\tau_2 = 0.2$, and gradient clipping at global norm 1. The objective combines uniform-positive cross-entropy with an identity penalty of weight 10. The residual training coefficient is $\kappa = 0.2$.

Model Selection and Inference. At inference, we use a protein fusion multiplier $\alpha_{\text{inf}} = 3$ and residual coefficient $\kappa_{\text{inf}} = 0.1$ in Eq. 8. These coefficients are applied to the learned checkpoint parameters, with the fusion scale multiplied by two exactly once. Retrieval uses the cosine similarity of the normalized embeddings. We use neither ensembles nor hubness correction. Appendix F.3 reports coefficient sensitivity for these same checkpoints and candidate pools.

The encoder checkpoints and the residual heads after 100 updates were selected on validation data during development. Before evaluating the inference coefficients on the test sets, we fix $(\alpha_{\text{inf}}, \kappa_{\text{inf}}) = (2, 0.1)$ on both ReactZyme validation directions in every split and all eight EnzymeMap validation metrics. This criterion does not establish that the selected pair maximizes every validation metric. Baselines retain their own validation-based selection rules (Appendix E.2). The broader study included test-set inspection during development, as disclosed in Appendix B. The matched ablation protocols in Appendix F retain these budgets and change only their explicitly stated experimental factor.

E EXPERIMENTAL PROTOCOLS AND SUPPLEMENTARY ANALYSES

E.1 SCREENING AND RETRIEVAL PROTOCOLS

EnzymeMap. The processed resource contains 46,356 associations, 16,776 reactions, 12,749 enzymes, 2,841 EC identifiers, and 394 reaction rules. The screening library combines EnzymeMap, BRENDA 2022_2, and UniProt 2022_01, retaining sequences of at most 650 residues. The 4,642 test associations yield 1,551 distinct reaction strings. Removing 30 strings present in training leaves 1,521 screening queries. Positives are the corresponding protein identifiers in the raw EnzymeMap records intersected with the candidate library, giving 4,544 positive query–candidate labels. Excluding training enzyme identifiers leaves 252,113 candidates and 1,337 queries with at least one remaining positive.

For n positives among N candidates, EF at fraction f is computed as

$$\text{EF}_f = \frac{\sum_{j=1}^{\lfloor fN \rfloor} y_{(j)}}{fn}, \tag{36}$$

where $y_{(j)}$ indicates whether the candidate at rank j is positive. We use $f = 0.05$ and 0.10 and average over eligible queries. BEDROC follows the released CLIPZyme notebook, including its zero-based rank and sorting conventions. The released CLIPZyme checkpoint was trained on the corresponding EnzymeMap partition and is reevaluated with the same query eligibility, candidate identifiers, and positive labels. FGW-CLIP is included through published results; it was not reproduced locally for this comparison.

ReactZyme. ReactZyme links SwissProt enzymes to Rhea reactions. The published dataset statistics are 178,463 associations, 178,327 enzymes, and 7,726 reaction identifiers. The materialized protocols used here contain 178,459 association rows each. Table 6 gives the resulting partitions and complete test catalogs. The temporal protocol uses the released 31 December 2010 cutoff. Validation is a row-level partition of the released training associations: it does not require disjoint enzyme or reaction identities. The construction and identity-overlap checks are detailed in Appendix B.

Table 6: ReactZyme partitions used by the local comparisons. Train/validation/test columns count association rows. The complete test reaction and enzyme catalogs, $|R_{\text{test}}|$ and $|E_{\text{test}}|$, exchange query and candidate roles between retrieval directions.

Split	Train	Val.	Test	$\lvert R_{\text{test}}\rvert$	$\lvert E_{\text{test}}\rvert$
Reaction-Sim	147,393	16,377	14,689	386	14,688
Enzyme-Sim	152,748	16,972	8,739	1,573	8,734
Time	149,554	16,618	12,287	2,634	12,277

E.2 Baseline adaptations and selection

CREEP. We adapt the protein–reaction variant released with CARE (Yang et al., 2024) to the exact training associations of each benchmark. ProtT5 and rxnfp are fine-tuned from their pretrained language-model weights with freshly initialized 256-dimensional linear projection heads. Protein pooling averages all 512 padded token positions; reaction pooling uses the CLS token. Training uses symmetric EBM-NCE on unnormalized projected vectors (temperature 0.1, one cyclic negative), without EC text, EC-derived positive pairs, or a generative loss. All fits use Adam with learning rate 10^{-5} and zero weight decay, and 40 epochs.

On ReactZyme, reaction strings concatenate the released participant set followed by », without assigning physical sides. Each split is trained independently with batch size 64 (the released default is 16). Checkpoints maximize mean bidirectional validation all-positive MRR.

On EnzymeMap, we train a separate model on the 34,427 official training associations, using the original directed reaction SMILES and a global batch size of 256 across four GPUs. Initialization uses the pretrained language-model backbones and fresh projection heads, without a ReactZyme-trained retrieval checkpoint. Validation is performed every three epochs and at epoch 40. The checkpoint with the highest full-library validation BEDROC_{85} is selected, retaining the earlier epoch in an exact tie; this selects epoch 39. The same selected checkpoint is evaluated on both test settings: 1,521 queries against 261,907 candidate identifiers and 1,337 queries against 252,113 identifiers after excluding training IDs. Test labels are used only for evaluation after checkpoint selection.

These adaptations use the same downstream associations, while retaining each method's pretraining data, auxiliary inputs, objective, and optimization budget. The exploratory participant-set adaptation of Horizyn is omitted from the quantitative tables because it used a different input adaptation and an unmatched optimization protocol. Horizyn remains part of the related work discussion. The local comparisons therefore describe performance relative to the included controls rather than an exhaustive or compute-matched comparison of retrieval architectures.

E.3 Retrieval metrics and published comparisons

For a query q, let $T_k(q)$ contain the k highest-ranked candidates. In addition to Eq. 9, we report

$$\text{Top}k = \frac{1}{|Q|}\sum_{q\in Q} \mathbb{1}[P(q)\cap T_k(q)\neq\emptyset], \tag{37}$$

$$\text{Top}k\text{-N} = \frac{1}{|Q|}\sum_{q\in Q}\frac{|P(q)\cap T_k(q)|}{k}, \tag{38}$$

$$\text{MeanRank} = \frac{1}{|Q|}\sum_{q\in Q}\frac{1}{|P(q)|}\sum_{p\in P(q)}\text{rank}_q(p). \tag{39}$$

We compute these metrics from the same score matrices as the main MRR results, keeping the trained model and evaluation data fixed.

F CONTROLLED ABLATIONS AND PARAMETER SENSITIVITY

F.1 REFERENCE CONFIGURATION AND EXPERIMENTAL CONTROLS

We evaluate the contributions of learned residue views and inference coefficients using the same model as in Tables 2 and 1. Its learned components are fitted independently on each benchmark's training associations. Each comparison retains the reference architecture, frozen feature extractors, feature preprocessing, positive-pair construction, optimizer, batch size, training budget, and evaluation protocol.

Retraining controls repeat both stages with the budgets in Appendix D; inference controls change only the stated coefficient at fixed fitted weights. These interventions answer different questions: removing a component at inference does not estimate the performance of a model trained without it. All comparisons retain the main-paper test graphs, complete candidate banks, and metrics. BEDROC is reported in percent and EF in fold enrichment.

We first remove the learned residue views and retrain the model on EnzymeMap. We then evaluate 17 inference settings on each of the four benchmark models and examine their effects on cross-modal functional recovery. All evaluated settings are reported, including those that reduce performance.

Table 10: Reference hyperparameters used in the controlled ablations. Only the parameter under study is varied in each comparison.

Parameter	**Role**	**Reference value**
K	Learned residue views	4
$\tau_1 = 1/\beta$	Alignment temperature	0.2 ($\beta = 5$)
s_0	Initial fusion scale	0.3
View width	Residue/query representation width	256
View dropout	Projection dropout	0.1
λ_{ent}	Entropy-penalty weight	0.01
λ_{div}	Diversity-penalty weight	0.01
η	Uniform attention mixture	0.05
ζ	Uniform gate mixture	0.05
$n_{\min}$	Entropy-floor support	4
m_{div}	Centered cosine margin	0.9
Batch size	Association rows per batch	1,024
Protein length	Maximum retained residues	1,022
Direction weights	R→E / E→R alignment weights	$0.5/0.5$
Training positives	Known in-batch positive edges	All known
Encoder budget	Retained training epoch	20 / 15*
Encoder optimizer	AdamW learning rate / weight decay	$10^{-4}/10^{-2}$
Encoder precision	Mixed-precision training	BF16
τ_2	Refinement temperature	0.2
κ	Training residual scale	0.2
ψ	Identity-penalty weight	10
Head width	Residual-MLP hidden size	1,024
Refinement budget	Full-graph optimization updates	100
Refinement optimizer	AdamW learning rate / weight decay	$10^{-4}/10^{-3}$
Refinement loss	Uniform-positive cross-entropy	`positive_ce`
α_{inf}	Inference fusion multiplier	2
κ_{inf}	Inference residual scale	0.1
Dictionary weight	Auxiliary dictionary-score weight	0

*ReactZyme / EnzymeMap.

Table 7: Detailed ReactZyme Time-split results, previously shown in the main comparison. Layout and baseline values follow Table 3 of Zhou et al. (2025); TIGER (ESM2Text) values use Tables 4–7 of Zhang et al. (2026), and ProtT3 uses their Table 1 (unreported entries: –). All rows above `CIRCE` are published values; `CIRCE` uses the validation-frozen dictionary-free reference scores. * denotes FANN-enhanced features, and ‡ flags a source inconsistency discussed above. Ground truth is an oracle, not a trained model. Higher is better except for mean rank.

(a) Given the enzyme, retrieve positive reactions.

Method	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Ground truth (oracle)	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	0.6550‡	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, EC Mode	0.3830	0.7090	0.8038	0.8823	0.3830	0.1418	0.0804	0.0442	23.1991	0.5222
FGW-CLIP, EC Max	0.3782	0.7143	0.8119	0.8839	0.3782	0.1429	0.0812	0.0442	24.1628	0.5229
TIGER (ESM2Text)	0.5810	0.8190	0.8740	0.9164	0.5810	0.1638	0.0874	0.0458	13.3309	0.6902
TIGER (ProtT3)	0.583	–	–	–	–	–	–	–	–	0.683
CIRCE	0.6682	0.8783	0.9195	0.9515	0.6682	0.1758	0.0920	0.0476	13.0522	0.7645

(b) Given the reaction, retrieve positive enzymes.

Method	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Ground truth (oracle)	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, EC Mode	0.3918	0.6230	0.7141	0.8060	0.3918	0.2582	0.1845	0.1233	106.2501	0.3394
FGW-CLIP, EC Max	0.3967	0.6317	0.7172	0.8075	0.3967	0.2614	0.1866	0.1235	104.6430	0.3392
TIGER (ESM2Text)	0.4536	0.6708	0.7676	0.8477	0.4536	0.2853	0.2016	0.1328	45.1359	0.3658
TIGER (ProtT3)	0.454	–	–	–	–	–	–	–	–	0.372
CIRCE	0.5642	0.7642	0.8299	0.8884	0.5642	0.3313	0.2245	0.1420	66.7177	0.4562

Table 8: ReactZyme Enzyme-Sim split. Layout and baseline values follow Table 7 of Zhou et al. (2025); TIGER (ESM2Text) uses Tables 8–11 of Zhang et al. (2026). ProtT3 uses their Table 1. *CIRCE* is computed from its dictionary-free reference scores. Conventions are as in Table 7.

(a) Given the enzyme, retrieve positive reactions.

Method	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Ground truth (oracle)	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, EC Mode	0.7875	0.9742	0.9891	0.9953	0.7875	0.1949	0.0990	0.0498	2.4126	0.8683
FGW-CLIP, EC Max	0.7642	0.9746	0.9875	0.9953	0.7642	0.1950	0.0988	0.0498	2.3837	0.8581
TIGER (ESM2Text)	0.9308	0.9850	0.9916	0.9962	0.9308	0.1970	0.0991	0.0498	1.5807	0.9561
TIGER (ProtT3)	0.908	–	–	–	–	–	–	–	–	0.940
CIRCE	0.9291	0.9928	0.9963	0.9977	0.9291	0.1986	0.0997	0.0499	1.7767	0.9592

(b) Given the reaction, retrieve positive enzymes.

Method	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Ground truth (oracle)	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, EC Mode	0.6940	0.8893	0.9319	0.9701	0.6940	0.4327	0.3176	0.2213	11.3119	0.5253
FGW-CLIP, EC Max	0.6978	0.8931	0.9389	0.9739	0.6978	0.4338	0.3200	0.2216	10.8511	0.5300
TIGER (ESM2Text)	0.7921	0.9408	0.9688	0.9809	0.7921	0.4606	0.3301	0.2238	6.8100	0.5921
TIGER (ProtT3)	0.784	–	–	–	–	–	–	–	–	0.579
CIRCE	0.8423	0.9612	0.9841	0.9936	0.8423	0.4634	0.3303	0.2243	6.4443	0.6361

Table 9: ReactZyme Reaction-Sim split. Layout, baseline values, and repeated method labels follow Table 8 of Zhou et al. (2025); TIGER (ESM2Text) uses Tables 12–15 of Zhang et al. (2026). ProtT3 uses their Table 1. `CIRCE` is computed from its dictionary-free reference scores. ‡ flags source inconsistencies (Appendix E.3); remaining conventions are as in Table 7.

(a) Given the enzyme, retrieve positive reactions.

Method	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Ground truth (oracle)	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, EC Mode	0.2486	0.5088	0.5716	0.6774	0.2486	0.1018	0.0572	0.0339	49.5410	0.3722
FGW-CLIP, EC Max	0.2561	0.5117	0.5799	0.6919	0.2561	0.1024	0.0580	0.0346	41.8125	0.3804
TIGER (ESM2Text)	0.4155	0.6416	0.6827	0.7540	0.4155	0.1283	0.0682	0.0377	23.2479	0.5180
TIGER (ProtT3)	0.386	–	–	–	–	–	–	–	–	0.472
CIRCE	0.4248	0.6417	0.6891	0.7649	0.4248	0.1283	0.0689	0.0382	28.9188	0.5284

(b) Given the reaction, retrieve positive enzymes.

Method	Top1	Top5	Top10	Top20	Top1-N	Top5-N	Top10-N	Top20-N	Mean Rank	MRR
Ground truth (oracle)	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, EC Mode	0.3886	0.5907	0.6736	0.7228	0.3886	0.2705	0.2091	0.1492	567.1841	0.3181
FGW-CLIP, EC Max	0.3782	0.5933	0.6658	0.7150	0.3782	0.2746	0.2163	0.1530	552.7454	0.3113
TIGER (ESM2Text)	0.4305	0.6113	0.6994	0.7616	0.4305	0.2854	0.2269	0.1680	219.7977	0.3185
TIGER (ProtT3)	0.428	–	–	–	–	–	–	–	–	0.337
CIRCE	0.4767	0.6606	0.7254	0.7902	0.4767	0.3010	0.2334	0.1706	279.6507	0.3769

F.2 Removing the learned residue views

We compare the main-paper $K = 4$ encoder with a freshly trained $K = 0$ control on EnzymeMap. Both models use 15 encoder epochs, batch size 512, $\beta = 5$, and 100 fresh refinement updates, with inference coefficients $(2, 0.1)$. All remaining settings follow Table 10.

At $K = 0$, the learned-residue adapter, key/value projections, queries, and learned-view projections are absent. The global and SLEEC views, reaction encoder, fusion mechanism, and residual heads remain; the retrieval dimension remains 512. Both attention penalties are zero because no learned attention distributions remain. The enzyme module has 2,763,521 trainable parameters at $K = 0$, compared with 4,739,330 at $K = 4$, excluding frozen resources, the reaction encoder, and residual heads. This is a module-removal control, not a parameter-matched test.

Table 11 reports the resulting comparison. The $K = 4$ model is higher on all eight screening endpoints. Removing the learned-view module reduces BEDROC_{85} by 1.79 percentage points in the full library and 2.55 points in the training-ID-excluded library. This provides evidence for the complete learned-view module at the stated training budget; it does not distinguish additional capacity, attention diversity, and regularization as separate causes.

Table 11: Matched $K = 0$ versus $K = 4$ retraining on EnzymeMap. Both models use 15 encoder epochs, batch size 512, 100 refinement updates, and $(\alpha_{\text{inf}}, \kappa_{\text{inf}}) = (2, 0.1)$, without a dictionary score. B85 and B20 are BEDROC percentages; EF5 and EF10 are fold enrichments. Differences are $K = 0$ minus the reference, in percentage points or fold-enrichment units. Shading identifies the reference.

Control	Full library				Training-ID-excluded			
	B85	B20	EF5	EF10	B85	B20	EF5	EF10
$K = 0$	51.26	69.88	15.55	8.55	47.10	66.57	14.90	8.33
$K = 4$ (reference)	53.05	71.32	16.03	8.62	49.64	68.70	15.48	8.45
$\Delta(K = 0)$	-1.79	-1.44	-0.48	-0.06	-2.55	-2.13	-0.58	-0.11

F.3 Inference-coefficient sensitivity

We isolate the inference-time effects of the enzyme fusion multiplier α_{inf} and residual coefficient κ_{inf} in Eq. 8. Before evaluating the alternatives, we reproduce every entry of the four reference score matrices and all 14 reference metrics exactly. The reference is $(\alpha_{\text{inf}}, \kappa_{\text{inf}}) = (2, 0.1)$, as used in the saved benchmark evaluations.

We vary $\alpha_{\text{inf}} \in \{0, 0.5, 1, 1.5, 2, 2.5, 3, 4\}$ while fixing $\kappa_{\text{inf}} = 0.1$, and $\kappa_{\text{inf}} \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ while fixing $\alpha_{\text{inf}} = 2$. Three additional pairs, $(1, 0)$, $(3, 0)$, and $(3, 0.2)$, provide the native phase-1 control and comparisons at fusion multiplier three. Deduplicating the shared reference gives 17 settings per model and 68 complete evaluations. Each ReactZyme evaluation retains both retrieval directions and their complete candidate pools. EnzymeMap uses all 1,521 queries against 261,907 accessions; excluding training identifiers leaves 1,337 queries with positives among 252,113 candidates. We retain all six ReactZyme MRR and eight EnzymeMap screening endpoints at every setting.

These settings distinguish the contributions of fusion and refinement. At $\alpha_{\text{inf}} = 0$, the fused enzyme branch is removed at inference, while the direct global branch and residual heads remain. This suppresses the entire fused representation, which includes the SLEEC-guided and learned residue views as well as a global view; it does not isolate any one of these signals and is not the retrained $K = 0$ ablation. At $\kappa_{\text{inf}} = 0$, both residual heads are bypassed. The pair $(1, 0)$ therefore recovers the native phase-1 embeddings, whereas $(2, 0)$ isolates removal of the residual correction at the reference fusion setting. The value $\kappa_{\text{inf}} = 0.2$ matches the residual scale used during training. Applying a different fusion multiplier also changes the enzyme embedding presented to the fixed enzyme residual head; these coefficients need not act independently.

Comparing $(2, 0)$ with $(2, 0.1)$ isolates the applied residual correction at fixed fusion. Comparing $(1, 0)$ with $(2, 0.1)$ changes both factors and cannot attribute their combined effect to refinement alone. All stage controls are included in the complete tables below.

Fusion Multiplier. Figures 3 and 4 show the complete one-coefficient-at-a-time curves, and Tables 12 and 13 report all 17 settings. At fixed $\kappa_{\text{inf}} = 0.1$, suppressing the fused branch reduces all 14 endpoints relative to the reference. For example, full-library BEDROC$_{85}$ decreases from 53.05% to 49.09%, and Time $E \to R$ MRR decreases from 0.7645 to 0.6561. Increasing the multiplier from two to three instead reduces five of six ReactZyme MRR values and all eight EnzymeMap metrics; Time $E \to R$ increases to 0.7697. The reference is nevertheless not the maximizer of every metric: multiplier 1.5 gives higher $R \to E$ MRR on all three ReactZyme splits and higher EnzymeMap BEDROC$_{85}$, whereas multiplier two gives higher EF5 in both screening libraries. The fused representation contributes useful retrieval information, but its preferred weight depends on the endpoint.

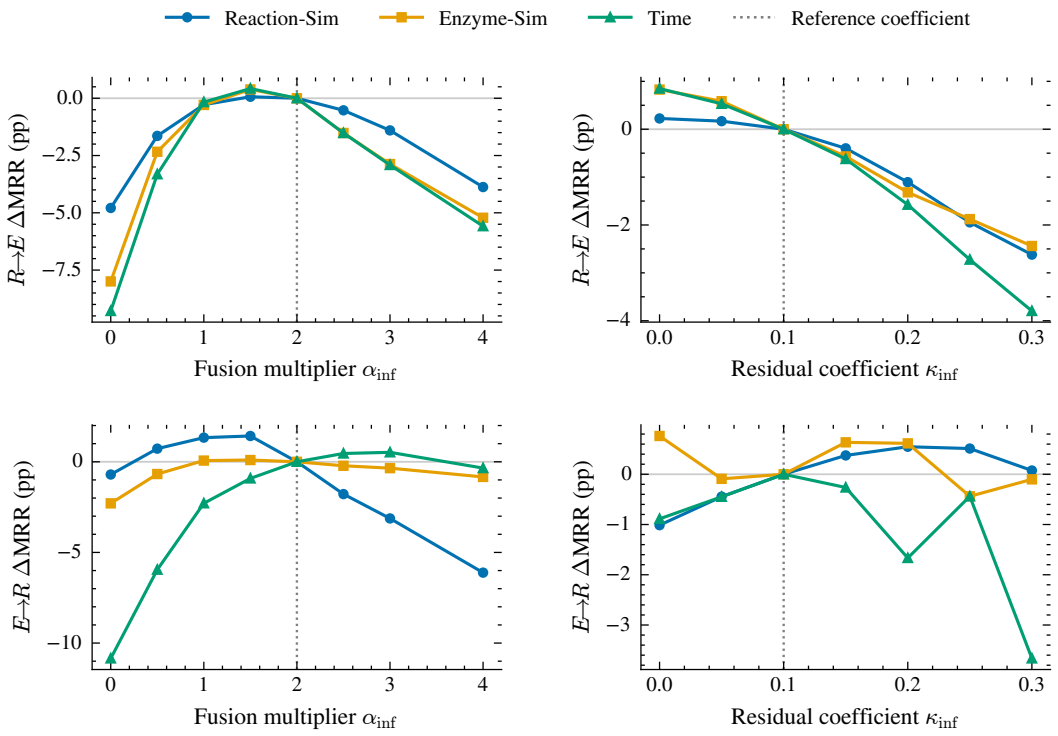


Figure 3: Dictionary-free ReactZyme coefficient sensitivity at fixed checkpoints. Curves show the change in test MRR, in percentage points, relative to $(\alpha_{\text{inf}}, \kappa_{\text{inf}}) = (2, 0.1)$. Left: fusion multiplier at fixed residual coefficient 0.1. Right: residual coefficient at fixed fusion multiplier two. Rows correspond to the two retrieval directions; colors and marker shapes identify splits. Dotted vertical lines mark the reference coefficient. Lines connect evaluated settings and are not fitted trends. These are single-fit sensitivity curves, without training-variance confidence intervals; absolute values appear in Table 12.

Residual Coefficient. At fusion multiplier two, moving from $\kappa_{\text{inf}} = 0$ to 0.1 increases full-library BEDROC$_{85}$ from 52.10% to 53.05% and training-ID-excluded BEDROC$_{85}$ from 48.81% to 49.64%. BEDROC$_{20}$ and EF5 also improve in both libraries, while full-library EF10 decreases slightly. The same change reduces $R \to E$ MRR on every ReactZyme split, improves $E \to R$ MRR on Reaction-Sim and Time, and reduces it on Enzyme-Sim. Halving the training coefficient from 0.2 to 0.1 improves four of six ReactZyme MRR endpoints and EF10 in both screening libraries, but reduces the other six screening endpoints. For example, full-library BEDROC$_{85}$ is 53.42% at coefficient 0.2, compared with 53.05% at 0.1. The shared inference coefficient therefore balances different retrieval metrics; its reduction does not improve every endpoint.

Table 12: All dictionary-free ReactZyme coefficient settings. Entries are query-averaged all-positive test MRR. The shaded row reproduces the primary benchmark scores; shading does not denote the column maxima. All rows use the same trained weights within each split.

		Reaction-Sim		Enzyme-Sim		Time	
α_{inf}	κ_{inf}	$R \to E$	$E \to R$	$R \to E$	$E \to R$	$R \to E$	$E \to R$
0	0.1	0.3290	0.5214	0.5562	0.9362	0.3633	0.6561
0.5	0.1	0.3605	0.5357	0.6128	0.9524	0.4230	0.7050
1	0	0.3772	0.5399	0.6364	0.9640	0.4599	0.7299
1	0.1	0.3741	0.5418	0.6332	0.9598	0.4545	0.7416
1.5	0.1	0.3776	0.5427	0.6400	0.9601	0.4604	0.7553
2	0	0.3791	0.5183	0.6444	0.9668	0.4646	0.7556
2	0.05	0.3786	0.5240	0.6420	0.9583	0.4615	0.7600
2	0.1	0.3769	0.5284	0.6361	0.9592	0.4562	0.7645
2	0.15	0.3729	0.5322	0.6304	0.9655	0.4499	0.7619
2	0.2	0.3658	0.5339	0.6229	0.9653	0.4404	0.7479
2	0.25	0.3574	0.5336	0.6174	0.9548	0.4289	0.7601
2	0.3	0.3507	0.5292	0.6117	0.9582	0.4182	0.7279
2.5	0.1	0.3716	0.5106	0.6210	0.9570	0.4410	0.7691
3	0	0.3689	0.4892	0.6210	0.9654	0.4407	0.7645
3	0.1	0.3629	0.4972	0.6074	0.9557	0.4270	0.7697
3	0.2	0.3491	0.5046	0.5965	0.9617	0.4107	0.7493
4	0.1	0.3381	0.4673	0.5840	0.9508	0.4003	0.7610

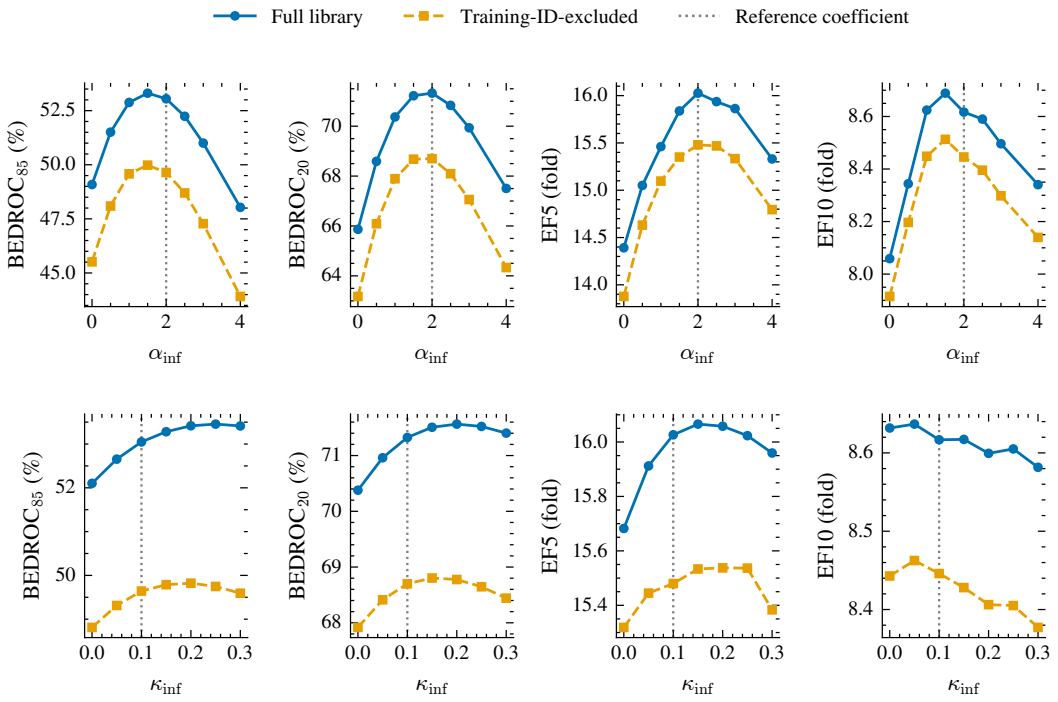


Figure 4: Dictionary-free EnzymeMap coefficient sensitivity on the complete screening libraries. Top: fusion multiplier, fixing $\kappa_{\text{inf}} = 0.1$. Bottom: residual coefficient, fixing $\alpha_{\text{inf}} = 2$. Blue circles denote the full library; orange squares denote the training-ID-excluded library. BEDROC is expressed as a percentage and EF as fold enrichment. Dotted vertical lines mark the reference coefficient. Each panel uses its own metric scale; Table 13 provides all numerical results.

Table 13: All dictionary-free EnzymeMap coefficient settings. B85 and B20 denote BEDROC_{85} and BEDROC_{20} in percent; EF5 and EF10 are fold enrichments. Query and candidate pools are fixed within each library. The shaded row is the unchanged primary benchmark reference.

		Full library				Training-ID-excluded			
α_{inf}	κ_{inf}	B85	B20	EF5	EF10	B85	B20	EF5	EF10
0	0.1	49.09	65.87	14.39	8.06	45.52	63.18	13.88	7.91
0.5	0.1	51.51	68.59	15.05	8.34	48.09	66.09	14.63	8.20
1	0	51.77	69.21	15.29	8.48	48.50	66.87	14.86	8.38
1	0.1	52.88	70.37	15.46	8.62	49.58	67.89	15.10	8.45
1.5	0.1	53.31	71.22	15.84	8.69	49.98	68.68	15.35	8.51
2	0	52.10	70.38	15.68	8.63	48.81	67.92	15.32	8.44
2	0.05	52.65	70.96	15.91	8.64	49.31	68.41	15.44	8.46
2	0.1	53.05	71.32	16.03	8.62	49.64	68.70	15.48	8.45
2	0.15	53.28	71.51	16.07	8.62	49.79	68.80	15.53	8.43
2	0.2	53.42	71.56	16.06	8.60	49.82	68.78	15.54	8.41
2	0.25	53.45	71.52	16.02	8.61	49.75	68.65	15.54	8.41
2	0.3	53.41	71.40	15.96	8.58	49.59	68.44	15.38	8.38
2.5	0.1	52.24	70.83	15.94	8.59	48.70	68.10	15.47	8.40
3	0	50.30	69.33	15.69	8.57	46.79	66.61	15.14	8.39
3	0.1	51.00	69.94	15.86	8.50	47.28	67.06	15.33	8.30
3	0.2	51.07	69.88	15.80	8.44	47.09	66.83	15.30	8.23
4	0.1	48.03	67.51	15.33	8.34	43.91	64.33	14.79	8.14

F.4 Refinement and cross-modal functional recovery

We apply the inference-stage controls in Appendix F.3 to the EC3 recovery diagnostic of Section 4. Within each benchmark, all columns of Table 14 use the same aligned encoder checkpoint and residual heads trained for 100 updates. The only changed parameters are the inference fusion multiplier α_{inf} and residual coefficient κ_{inf}. Their mappings to the implementation are given in Table 10.

The comparison $(\alpha_{\text{inf}}, \kappa_{\text{inf}}) = (1, 0) \to (1, 0.2)$ measures the effect of applying the trained heads at native fusion and their training residual coefficient. The comparison $(2, 0) \to (2, 0.1)$ measures their effect at the inference fusion and residual coefficients. The additional $(2, 0.2)$ control separates the choice of residual coefficient from the fusion multiplier. Comparing $(1, 0)$ directly with $(2, 0.1)$ changes both factors and cannot attribute the result solely to refinement.

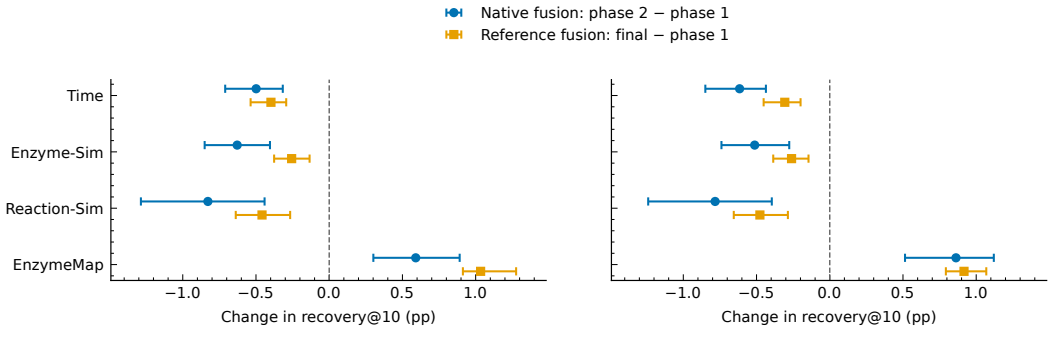


Figure 5: Changes in functional Recovery@10 attributable to applying the fixed residual heads at a fixed fusion multiplier. Left: partner exclusion; right: additional sequence-component exclusion. Blue compares $(1, 0.2)$ with $(1, 0)$; orange compares $(2, 0.1)$ with $(2, 0)$. Positive values indicate increased recovery. Whiskers are paired 95% reaction-family multiplier intervals as defined in Appendix G.1.

Table 14: EC3 class-macro functional Recovery@10 for fixed-checkpoint stage controls, in percent. Column labels are $(\alpha_{\mathrm{inf}}, \kappa_{\mathrm{inf}})$. The first column is native phase 1; the second applies full refinement at native fusion; the final column is the unchanged main-paper inference setting. Intermediate controls separate fusion and residual scaling.

Partition	Exclusion	$(1, 0)$	$(1, 0.2)$	$(2, 0)$	$(2, 0.2)$	$(2, 0.1)$
Time	Partners	31.83	31.33	32.10	31.20	31.70
Time	Components	29.97	29.35	30.13	29.39	29.82
Enzyme-Sim	Partners	29.36	28.73	29.60	28.77	29.34
Enzyme-Sim	Components	27.57	27.05	27.84	27.11	27.57
Reaction-Sim	Partners	24.85	24.02	24.78	23.89	24.32
Reaction-Sim	Components	24.17	23.39	24.15	23.25	23.67
EnzymeMap	Partners	31.36	31.95	30.66	31.29	31.69
EnzymeMap	Components	29.23	30.09	29.19	29.62	30.11

The scope of these comparisons is limited to the evaluated interventions. No matched retraining results establish the separate benefits of SLEEC, set chemistry, individual reaction modalities, the choice of alignment loss, or the individual regularizers. Temperature, additional view counts, and learned-view suppression are not reported as completed experiments. The single-seed studies do not establish training-seed robustness or prospective generalization.

G Organization of the learned embedding space

G.1 Functional recovery beyond recorded partners

To assess functional recovery beyond known associations, we apply the diagnostic in Section 4 to the models used in the main comparison, with no further training or selection. On ReactZyme, we form the known-partner graph from the union of the available training, validation, and test tables, matching exact benchmark reaction identifiers. A representational forward suffix is removed when present. On EnzymeMap, we use all 46,356 released association records and match the native directed reaction strings. We do not merge additional chemical aliases. Thus, "recorded partners" means partners in these available source datasets, not every association in the literature. Enzyme EC labels come from the native UniProt–Rhea table for ReactZyme and the released EC-to-UniProt mapping for EnzymeMap. An EC3 label is retained only when all three prefix fields are specified. For example, $1.1.1.-$ contributes 1.1.1, whereas $1.-.-.-$ does not contribute an EC3 label. Multiple labels are preserved, and annotations of identical full sequences are combined. We do not infer missing annotations by homology or model prediction. Reaction labels are the union over recorded partners, including those outside the retrieval candidate bank; the native reaction's EC field is not substituted for this partner-derived definition.

Candidate Exclusion. We start from the complete benchmark candidate sets: sequence identifiers on ReactZyme and 261,907 UniProt accessions on EnzymeMap. After requiring specified EC3 annotations, the latter contains 131,409 accessions. For each reaction, we remove recorded partner identifiers and all accessions with an identical full sequence to a recorded partner. Other accessions sharing a sequence remain separate candidate identifiers, as in the screening benchmark. Consequently, recovery at k does not necessarily represent k distinct protein sequences.

The stricter control additionally removes every candidate in a detected close-sequence component of any recorded partner. We construct the search universe from all annotated candidate sequences and available partner sequences across the four partitions, yielding 184,423 distinct full sequences. MMseqs2 searches this universe against itself with sensitivity 7.5 and a 10,000-hit cap. We retain alignments with identity at least 50%, coverage at least 80% of both sequences, and $E \leq 10^{-3}$, then take connected components of the undirected hit graph. These transitive, search-derived components are operational exclusions rather than an exhaustive guarantee of sequence dissimilarity.

After exclusion, a query remains eligible if it has at least one EC3 label and at least 50 annotated candidates. Eligibility is fixed across all compared models and all values of k within an exclusion policy. Queries with no remaining EC3-compatible candidate are retained with recovery zero. Table 15 gives the exact denominators. The observed class set may differ between exclusion policies, so their aggregate differences should not be attributed solely to removal of candidates without checking the query cohort.

Table 15: Functional-recovery cohorts. The eligible-bank range is after query-specific exclusions. Queries without a compatible alternative remain in the metric. All methods share the cohort within each row.

Partition	Exclusion	Queries	EC3 classes	Families	Eligible-bank range	No alternative
Time	Partners	2,601	219	1072	11,488–12,145	45
Time	Components	2,601	219	1072	11,480–12,145	49
Enzyme-Sim	Partners	1,559	200	727	8,390–8,667	38
Enzyme-Sim	Components	1,559	200	727	8,387–8,667	41
Reaction-Sim	Partners	381	105	278	12,132–14,380	32
Reaction-Sim	Components	381	105	278	12,132–14,380	32
EnzymeMap	Partners	1,073	68	53	131,367–131,408	1
EnzymeMap	Components	1,067	68	53	130,334–131,408	3

Recovery and Uncertainty. Let $P(r)$ contain a reaction's recorded partners and $\Lambda_3(e)$ contain an enzyme's specified EC3 labels. The reaction labels are $\Lambda_R(r) = \bigcup_{e \in P(r)} \Lambda_3(e)$. Let $T_k(r)$ contain the top-k candidates after the applicable exclusions. We define

$$F_k^{R \to E} = \frac{1}{|\mathcal{C}_3|} \sum_{c \in \mathcal{C}_3} \frac{1}{|\mathcal{Q}_c|} \sum_{r \in \mathcal{Q}_c} \frac{1}{k} \sum_{e \in T_k(r)} \mathbb{1}[\Lambda_R(r) \cap \Lambda_3(e) \neq \emptyset] \,, \tag{40}$$

where $\mathcal{Q}_c$ contains eligible reaction queries carrying class c and $\mathcal{C}_3$ contains the represented classes. This statistic averages query agreement within each EC3 class and then equally across classes. A multilabel query enters every class it carries, but its agreement uses any shared label. We report $100F_k^{R \to E}$ as a percentage. The ordinary query mean is retained separately.

For the random-retrieval reference, let n_r be the number of eligible candidates and m_r the number whose EC3 annotations overlap the query's labels. Uniform random retrieval without replacement has expected agreement m_r/n_r for every $k \leq n_r$. The attainable agreement is bounded by $\min(k, m_r)/k$. Both quantities are aggregated with the same class weights as Eq. 40; no random-ranking simulation is needed. They account for different annotation frequencies, multiple labels, and query-specific exclusions.

For an exact-score tie block of size b containing h compatible candidates, selecting t positions from that block contributes expected compatible count th/b. We include the entire boundary tie block, even when it extends past rank k. Candidate-index ordering is retained as a sensitivity analysis. No EC annotation is used to break ranking ties. This expectation addresses exact ties, not arbitrary near-tie perturbations. The final `CIRCE` score matrices reproduce the main benchmark matrices exactly on all four partitions before any candidate exclusion.

For uncertainty, we assign independent exponential weights of mean one to whole reaction families, sharing each family's weight across all of its queries and all methods. Within each draw, weights are normalized within each observed EC3 class before class averaging. Positive weights retain every represented class. We use 1,000 draws and report the 2.5th and 97.5th percentiles for individual methods and paired differences. ReactZyme families use participant-Tanimoto components at threshold 0.7; EnzymeMap families connect reactions sharing native rules.

Figure 6 extends the main EnzymeMap analysis to both exclusion policies and all four partitions, reporting every neighborhood size from 1 to 50.

`CIRCE` has the highest observed recovery@10 in all four partitions under both filters (Table 16). The paired intervals against CREEP on Time and against both comparators on Reaction-Sim nevertheless include zero (Table 17). On EnzymeMap, the margin over CLIPZyme is 4.15 percentage points under partner exclusion and 4.65 under component exclusion, with positive paired intervals. `CIRCE`

Table 16: Cross-modal functional Recovery@10 in percent. "Partners" excludes all locally recorded partners and exact-sequence aliases; "Components" additionally excludes detected partner sequence components. All entries use class-macro averaging and expected exact-tie agreement. CLIPZyme has no matched ReactZyme export in this comparison.

Partition	Exclusion	*CIRCE*	Horizyn	CREEP	CLIPZyme	Random
Time	Partners	31.70	30.79	31.14	–	1.37
Time	Components	29.82	28.37	29.64	–	1.37
Enzyme-Sim	Partners	29.34	25.76	26.90	–	1.41
Enzyme-Sim	Components	27.57	23.89	25.34	–	1.40
Reaction-Sim	Partners	24.32	23.60	23.02	–	1.13
Reaction-Sim	Components	23.67	22.94	22.35	–	1.11
EnzymeMap	Partners	31.69	26.99	23.16	27.55	1.67
EnzymeMap	Components	30.11	25.25	21.49	25.46	1.62

Table 17: Paired differences in functional recovery@10: *CIRCE* minus each comparator, in percentage points. Brackets give 95% reaction-family multiplier intervals, conditional on the fixed checkpoints and candidate banks.

Partition	Exclusion	Comparator	Difference	95% interval
Time	Partners	Horizyn	$+0.90$	$[+0.30, +1.60]$
Time	Partners	CREEP	$+0.56$	$[-0.12, +1.59]$
Time	Components	Horizyn	$+1.45$	$[+0.86, +2.17]$
Time	Components	CREEP	$+0.18$	$[-0.48, +1.18]$
Enzyme-Sim	Partners	Horizyn	$+3.59$	$[+2.91, +4.19]$
Enzyme-Sim	Partners	CREEP	$+2.44$	$[+1.54, +3.36]$
Enzyme-Sim	Components	Horizyn	$+3.68$	$[+2.95, +4.31]$
Enzyme-Sim	Components	CREEP	$+2.23$	$[+1.33, +3.17]$
Reaction-Sim	Partners	Horizyn	$+0.72$	$[-1.01, +2.28]$
Reaction-Sim	Partners	CREEP	$+1.30$	$[-0.33, +2.78]$
Reaction-Sim	Components	Horizyn	$+0.73$	$[-1.01, +2.36]$
Reaction-Sim	Components	CREEP	$+1.32$	$[-0.30, +2.82]$
EnzymeMap	Partners	Horizyn	$+4.70$	$[+3.61, +6.19]$
EnzymeMap	Partners	CREEP	$+8.54$	$[+7.19, +9.86]$
EnzymeMap	Partners	CLIPZyme	$+4.15$	$[+2.94, +6.23]$
EnzymeMap	Components	Horizyn	$+4.86$	$[+3.61, +6.04]$
EnzymeMap	Components	CREEP	$+8.62$	$[+7.17, +9.75]$
EnzymeMap	Components	CLIPZyme	$+4.65$	$[+3.04, +6.87]$

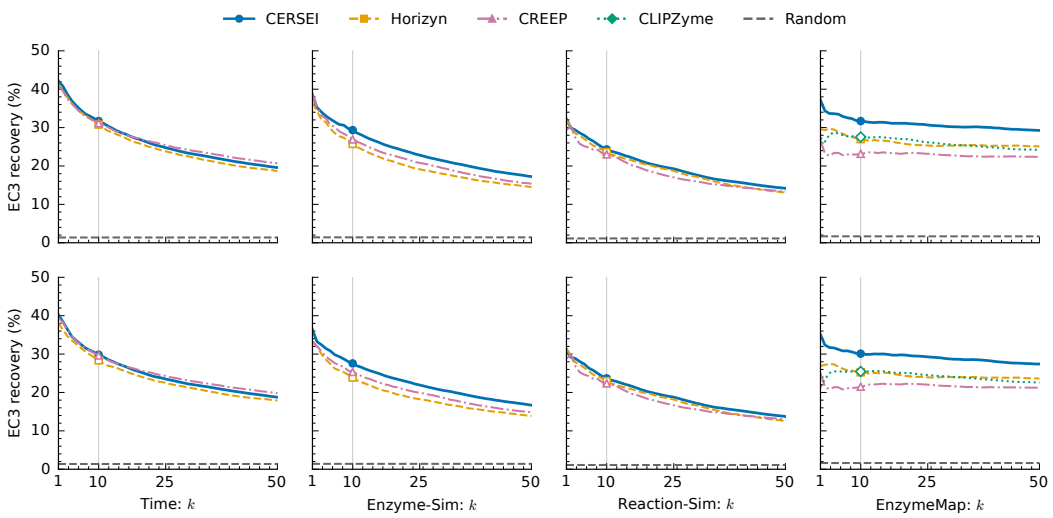


Figure 6: Cross-modal EC3 functional recovery for $k = 1$–50. Columns show Time, Enzyme-Sim, Reaction-Sim, and EnzymeMap. The top row excludes recorded partners and exact-sequence aliases; the bottom row additionally excludes their detected close-sequence components. All methods share eligible queries and candidates within each panel. Markers identify $k = 10$; gray curves give exact expected random retrieval. Values use Eq. 40, with expected agreement within exact-score ties. Neighborhood sizes are correlated, not independent replications.

exceeds all evaluated comparators throughout $k = 1$–50 on EnzymeMap under both policies and on Enzyme-Sim under component exclusion. This ordering is not universal: on Time, CREEP becomes higher from $k = 18$ under partner exclusion.

The stricter filter changes the candidate bank for only 28 of 381 Reaction-Sim queries, compared with 894 of 1,067 EnzymeMap queries; its additional evidence is therefore limited on Reaction-Sim. Appendix F.4 separates fusion and refinement effects, which are not uniformly positive.

G.2 Projection sensitivity and fidelity checks

To assess the stability of the enzyme maps in Section 4, we repeat the projections with fixed model embeddings and different initializations, and measure how well each projection preserves local neighborhoods.

The cohort contains all 720 EC-annotated sequences among the 1,357 unique enzymes associated with test reactions; 637 sequences lack a specified top-level EC label. No sequences are selected by prediction or projection appearance. The single-class counts are 285, 116, 85, 71, 16, and 119 for EC1 classes 1–6, respectively; 28 proteins carry more than one top-level class and are displayed as gray crosses. We preserve this imbalance rather than selecting a visually balanced subset. Colors are applied after the projection is fitted.

For CLIPZyme, we use the lexicographically first test-positive accession per sequence, since its accession-specific vectors can differ. The other methods encode sequences. We verify identical sequence identities and compare reconstructed cross-modal scores with each model's saved screening scores at those same representative accessions. Embedding dimensions are 256, 1,280, and 512 for CREEP, CLIPZyme, and `CIRCE`, respectively.

Projection Settings. We normalize the enzyme embeddings and independently fit each model's t-SNE projection with cosine distance, perplexity 30, early exaggeration 12, learning rate 200, and 1,500 iterations. The scikit-learn 1.7.2 Barnes–Hut implementation uses angle 0.5, minimum gradient norm 10^{-7}, and a 300-iteration no-progress threshold. For each seed, all models receive the same two-dimensional Gaussian initialization with standard deviation 10^{-4}. EC labels are not used in fitting.



Figure 7: All prespecified t-SNE initializations. Columns show CREEP, CLIPZyme, and *CIRCE*; rows use seeds 25092026, 25092027, and 25092028, respectively. The first row is the main-paper projection. Every panel uses the same 720 sequence identities and EC colors. These maps illustrate local organization and sensitivity of the two-dimensional layout; they are not independent model-training runs.

In addition to the main seed 25092026, seeds 25092027 and 25092028 were specified in advance and are retained in Figure 7. Each uses the settings above and an initialization shared across methods. Protein order is fixed by a hash of the sequence identifier and primary seed. The visualizations use equal aspect ratios without rescaling one coordinate relative to the other. Each square plotting window is centered on the coordinate bounds and contains every point; no coordinates are altered or cropped for display.

We quantify projection fidelity by comparing each enzyme's ten nearest neighbors under original-space cosine distance with its ten nearest neighbors under two-dimensional Euclidean distance. At the primary seed, the mean retained fractions are 71.00%, 70.93%, and 73.79% for CREEP, CLIPZyme, and *CIRCE*, respectively. Across the two additional seeds, they range from 70.38–71.15%, 71.38–71.68%, and 73.88–74.03%. The comparative claim in Figure 2d rests on the original-space retrieval measurement defined in Section 4, independently of these two-dimensional projections.

Table 18: Original-space enzyme organization on the matched map cohort. Neighbor agreement is in percent and uses 720 enzymes; silhouette uses the 692 single-EC1 enzymes. "Components" additionally excludes the query's 50%-identity component. Neighbor agreement and silhouette are class-macro averages. Higher is better.

Measurement	CREEP	CLIPZyme	CIRCE	Random
EC1 agreement@10	72.69	72.49	**74.27**	19.27
EC1 agreement@10, components	67.80	68.00	**69.54**	18.96
EC3 agreement@10	29.94	30.79	**33.18**	2.02
EC3 agreement@10, components	25.64	25.91	**28.71**	1.85
EC1 cosine silhouette	0.070	0.067	**0.100**	–

G.3 Quantitative enzyme organization on the visualization cohort

We quantify the functional organization illustrated in Figure 2a–c using the original embedding spaces. We measure local EC-label agreement and separation of annotated EC1 classes. These diagnostics evaluate different aspects of organization and need not favor the same model.

We use the same 720 EC-annotated EnzymeMap sequences for all methods, as in Appendix G.2, with identical query and candidate identities. There are six represented EC1 classes and 68 EC3 classes. EC1 retains the first field of an EC number; EC3 retains its first three specified fields. Both levels are available for every sequence. We use the final normalized embeddings of the existing screening models, with dimensions 256, 1,280, and 512 for CREEP, CLIPZyme, and CIRCE, respectively.

Functional Neighbor Agreement. For enzyme e, let $\Lambda_\ell(e)$ contain its EC prefixes at depth $\ell \in \{1, 3\}$ and let $\mathcal{N}_k(e)$ contain its k nearest eligible enzymes by cosine similarity. We define

$$A_{\ell,k}(e) = \frac{1}{k} \sum_{e' \in \mathcal{N}_k(e)} \mathbb{1}[\Lambda_\ell(e) \cap \Lambda_\ell(e') \neq \emptyset], \qquad M_{\ell,k} = \frac{1}{|\mathcal{C}_\ell|} \sum_{c \in \mathcal{C}_\ell} \frac{1}{|\mathcal{Q}_c|} \sum_{e \in \mathcal{Q}_c} A_{\ell,k}(e), \tag{41}$$

where $\mathcal{C}_\ell$ is the set of represented classes and $\mathcal{Q}_c$ contains queries annotated with c. Averaging within classes and then equally across classes prevents frequent functions from dominating. Multiple annotations are retained: a query enters each of its class means, while agreement tests any shared label, rather than requiring a match to the particular averaging class.

We always exclude self matches. A second policy also excludes candidates in the query's detected 50%-identity sequence component, obtained from MMseqs2 matches with at least 80% coverage of both sequences and $E \leq 10^{-3}$. All queries remain eligible, with 719 candidates under self exclusion and 693–719 under component exclusion. Exact score ties receive expected agreement under uniform ordering of the complete tie block, including ties crossing the cutoff. The random reference is the expected compatible fraction in each eligible bank, with the same class averaging. We evaluate $k = 1$–50 and summarize $k = 10$.

Class Separation. The silhouette calculation requires a single class per enzyme. We therefore use the identical 692 sequences with one EC1 label for all methods, excluding the 28 multiclass sequences only from this calculation. For enzyme e, let a_e be its mean cosine distance to other members of its class and b_e its smallest mean distance to any other EC1 class. Its silhouette is

$$\sigma_e = \frac{b_e - a_e}{\max(a_e, b_e)}. \tag{42}$$

Positive values indicate smaller within-class distances than distances to the closest competing class. We report both the ordinary mean and the class-macro mean of σ_e. This measurement uses the annotated partition and does not fit clusters.

To estimate uncertainty in neighbor agreement and silhouette, we use 1,000 paired exponential multiplier draws over the existing 30%-identity sequence families, assigning one weight to all queries in a family and recomputing the weighted class means. There are 239 families in the full cohort; silhouette uses the families represented in its single-label subset. The same draws are used across methods. Candidate banks and represented classes remain fixed. The 2.5th and 97.5th percentiles

Table 19: Paired differences between CIRCE and each comparator, with 95% sequence-family multiplier intervals. Neighbor differences are percentage points; silhouette differences use the silhouette scale. Components denotes the additional sequence-component exclusion.

Measurement	Versus CREEP	Versus CLIPZyme
EC1 agreement@10	$+1.57$ $[-0.42, 3.91]$	$+1.78$ $[0.20, 3.39]$
EC1 agreement@10, components	$+1.74$ $[-0.44, 4.27]$	$+1.54$ $[-0.19, 3.33]$
EC3 agreement@10	$+3.24$ $[2.53, 4.19]$	$+2.39$ $[1.84, 3.15]$
EC3 agreement@10, components	$+3.07$ $[2.32, 4.07]$	$+2.80$ $[2.12, 3.61]$
EC1 macro silhouette	$+0.0296$ $[0.0139, 0.0458]$	$+0.0333$ $[0.0183, 0.0459]$

of paired differences form the reported 95% intervals. These describe conditional query-family sensitivity, not model-training or candidate-bank uncertainty. This analysis is exploratory: EC1 agreement@10 was designated the primary diagnostic before calculating these measurements, but the checkpoints and test data had already been examined in earlier experiments. Intervals are not adjusted for multiple comparisons.

Results. CIRCE has higher class-macro EC1 silhouette, 0.100, than CREEP (0.070) and CLIPZyme (0.067), with positive paired intervals (Tables 18 and 19). This supports greater average EC1 separation, although the modest absolute values indicate overlapping classes. The primary EC1 neighbor advantage is smaller: 74.27% versus 72.69% and 72.49%. Its paired interval against CREEP includes zero, so this comparison alone does not establish a clear broad-class neighborhood advantage.

At EC3, CIRCE achieves 33.18% agreement@10, versus 29.94% for CREEP and 30.79% for CLIPZyme. After component exclusion, the scores are 28.71%, 25.64%, and 25.91%, respectively. Both paired intervals remain positive, and CIRCE has higher class-macro EC3 agreement at every evaluated $k = 1$–50 under both exclusion policies. Class balancing matters: with ordinary query averaging, component-excluded EC3 agreement is 63.50% for CIRCE, 62.32% for CREEP, and 62.93% for CLIPZyme. Thus, the stronger advantage concerns equal weighting of functions rather than a large gain for the average enzyme. Ordinary mean EC1 silhouettes are 0.1290, 0.0927, and 0.0847, respectively.

These diagnostics support stronger class-balanced functional neighborhoods and greater average EC1 separation on this cohort. Sequence-component exclusion does not remove all homology or training exposure, and a shared EC3 label does not establish identical substrate specificity. These enzyme-to-enzyme measurements are distinct from the reaction-to-enzyme recovery experiment in Appendix G.1; they do not validate new enzyme–reaction associations. All claims above use the original spaces rather than distances in the t-SNE projections.

G.4 Functional resolution and limits across partitions

We extend the enzyme analysis to all four EC levels and benchmark partitions to examine how functional resolution affects the comparison. All measurements use the main-paper inference coefficients and fixed checkpoints. ReactZyme comparisons use split-specific CREEP and Horizyn refits (Yang et al., 2024; Rocks et al., 2026); EnzymeMap uses a Horizyn refit and the released CLIPZyme checkpoint (Mikhael et al., 2024). Their downstream training associations match the respective benchmark, but their pretrained inputs and optimization budgets differ. In particular, Horizyn is an additional representation comparator, not a budget-matched architectural control. Frozen ProtT5 means provide an input-representation reference.

We select at most 2,000 protein queries and 500 reaction queries per partition by an annotation-independent hash ordering (seed 23092026), retaining identical identities across methods and the full diagnostic candidate banks. In Time/Enzyme-Sim/Reaction-Sim order, these contain 12,277/8,734/14,688 protein sequences and 2,634/1,573/386 reactions. EnzymeMap uses the 1,357 test-positive sequences, of which 720 have EC annotations at every level, and all 1,521 test reactions. These protein banks differ from the full-accession screening library.

Table 20: Class-balanced enzyme agreement@10 after sequence-component exclusion, in percent. EC1–EC4 use their respective annotated pools. Dashes denote unevaluated model–dataset combinations, not zero performance. These are descriptive point estimates; small differences are not treated as statistically established improvements.

Partition	Level	CIRCE	Horizyn	CREEP	CLIPZyme	ProtT5
Time	EC1	86.35	84.73	85.14	–	70.55
	EC2	57.84	57.53	58.35	–	40.62
	EC3	51.61	51.27	51.28	–	33.68
	EC4	27.11	27.94	25.99	–	17.26
Enzyme-Sim	EC1	88.97	84.82	86.57	–	65.04
	EC2	66.51	61.45	63.12	–	41.59
	EC3	57.86	53.51	54.23	–	33.41
	EC4	38.07	38.04	37.02	–	20.47
Reaction-Sim	EC1	94.98	91.86	92.48	–	81.43
	EC2	75.55	71.43	71.68	–	60.71
	EC3	60.52	55.12	55.75	–	44.86
	EC4	34.50	33.27	32.53	–	24.13
EnzymeMap	EC1	69.54	67.04	–	68.00	56.86
	EC2	34.51	32.38	–	32.39	23.42
	EC3	28.71	27.53	–	25.91	19.38
	EC4	8.99	8.97	–	8.62	5.96

For each EC depth $\ell \in \{1, 2, 3, 4\}$, we retain specified native prefixes, preserving multiple annotations and excluding missing labels. Labels are not imputed. Eligible ReactZyme protein query counts decrease from 1,994/1,994/2,000 at EC1 to 1,746/1,891/1,609 at EC4. Protein neighborhoods exclude self matches and the query's detected MMseqs2 component (Steinegger & Söding, 2017) at 50% identity (80% bidirectional coverage, $E \leq 10^{-3}$). Reaction neighborhoods exclude self matches and participant-fingerprint Tanimoto similarity ≥ 0.5, using chirality-aware, radius-2, 2,048-bit Morgan fingerprints of the combined participants after removing atom maps. These filters restrict query–candidate similarity; they do not remove all training exposure or homology.

With label sets $\Lambda(q)$ and cosine-ranked eligible neighbors $\mathcal{N}_k(q)$, we compute

$$a_k(q) = \frac{1}{k} \sum_{x \in \mathcal{N}_k(q)} \mathbb{1}[\Lambda(q) \cap \Lambda(x) \neq \emptyset], \qquad A_k = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|\mathcal{Q}_c|} \sum_{q \in \mathcal{Q}_c} a_k(q), \tag{43}$$

where $\mathcal{Q}_c$ contains queries with class c and $\mathcal{C}$ is the represented class set. A multilabel query enters each of its class means, but agreement tests any shared label. All eligible queries retain at least 50 candidates; identities remain fixed for $k = 1$–50. Reaction ties preserve candidate order, and protein ties follow the label-independent top-k operator. These descriptive point estimates use a different tie convention from the matched visualization-cohort analysis in Appendix G.3.

EC Hierarchy. Table 20 retains all four depths and partitions. The strongest consistent enzyme-space differences occur at intermediate resolution: on Enzyme-Sim and Reaction-Sim, CIRCE exceeds both trained comparators at EC3 throughout $k = 1$–50. The advantage does not extend uniformly to EC4. Horizyn is higher on Time at every tested k, and Enzyme-Sim and EnzymeMap are nearly tied at $k = 10$. On EnzymeMap, Horizyn is higher at 40 of 50 EC4 neighborhood sizes. Finer labels also reduce the number of compatible candidates; agreement is neither chance-adjusted nor normalized by its attainable ceiling.

A coverage control retains only complete EC4 annotations and derives all shorter prefixes from them, fixing identities and rankings across depths. The query counts are 1,746/1,891/1,609/720 and the candidate counts are 10,830/8,264/11,989/720 in Time/Enzyme-Sim/Reaction-Sim/EnzymeMap order. With ordinary query averaging, EC4 agreement remains lower than Horizyn on Time (47.14% versus 49.01%) and Enzyme-Sim (60.54% versus 60.94%), and higher on Reaction-Sim (85.10% versus 83.70%) and EnzymeMap (22.39% versus 22.15%). This control changes the weighting from the table's class-macro metric.

Table 21: Reaction-neighborhood agreement@10 after excluding candidates with participant Tanimoto similarity ≥ 0.5, in percent. ReactZyme uses association-derived EC3 labels; EnzymeMap uses native rule IDs. Values should be compared within rows because the annotation targets and banks differ.

Partition	*CIRCE*	Horizyn	CREEP	CLIPZyme
Time	15.86	11.93	14.52	–
Enzyme-Sim	14.25	10.34	12.74	–
Reaction-Sim	8.28	5.88	8.44	–
EnzymeMap	19.26	17.36	–	16.94

Reaction Neighborhoods. Reaction labels are native rule IDs on EnzymeMap and unions of known partners' EC3 labels on ReactZyme. The latter measures association-derived coherence, not an independently annotated reaction mechanism. After annotation and fingerprint filtering, the four query counts are 494/496/381/500. Table 21 shows agreement at ten neighbors. The ordering depends on k: CLIPZyme is higher on EnzymeMap at $k = 1$–3, and CREEP is higher on Reaction-Sim at 46 of 50 neighborhood sizes and on Enzyme-Sim at $k = 45$–50. Thus, enzyme-space gains do not imply a uniform reaction-space advantage.

In a separate EnzymeMap check, single-rule reactions with native mapping quality ≥ 0.5 are matched to different-rule candidates within 0.02 participant Tanimoto units, allowing candidate reuse. The 427 queries produce 55,704 matched candidate pairs. Averaging same-rule minus matched different-rule cosine differences within queries and then rule classes gives 0.2787 for *CIRCE*, 0.1571 for Horizyn, and 0.2473 for CLIPZyme. These positive gaps show rule-sensitive organization within each model; their uncalibrated magnitudes do not rank retrieval quality.

We also compare these functional neighborhoods with retrieval of exact recorded partners. Using complete retrieval banks and class-macro positive coverage $|P(q) \cap T_k(q)|/|P(q)|$, with expected coverage under exact ties, EnzymeMap R→E coverage@10 is 14.62% for *CIRCE*, 16.56% for Horizyn, and 16.90% for CLIPZyme. The local Horizyn refit also has higher all-positive MRR in all six ReactZyme split–direction combinations than the analyzed *CIRCE* exports, under its different optimization budget. Better functional neighborhoods therefore do not establish better performance on every retrieval criterion.

These are exploratory fixed-checkpoint measurements. Correlated neighborhood sizes are not independent replications. For native rule labels, family reweighting can be degenerate when a class shares one component, so curve ordering is reported descriptively rather than as statistical significance. The evidence establishes neither universal embedding superiority nor catalytic activity for newly retrieved pairs.

H Case Studies

As a case study, we consider the conversion of D-fructose to D-tagatose, a reaction catalyzed by naturally occurring and engineered tagatose 4-epimerases (T4Eases) (Shin et al., 2020; Xia et al., 2024; Chen et al., 2024). We construct a database of candidate enzymes and query the D-fructose→D-tagatose transformation using *CIRCE* in R→E mode. Candidate enzymes are ranked by cosine similarity to the query reaction, providing a measure of their predicted compatibility with the transformation.

For structural analysis, we use the experimentally resolved wild-type TsT4Ease structure reported by Wang et al. (RCSB ID 9LHY). The study identifies E141, D174, H175, and H355 as catalytic residues and reports that D284 contributes to coordination of the catalytic metal ion (Wang et al., 2025). After identifying the highest-ranked candidate enzymes, we refold their sequences using ESMFold2-Fast and compare the candidates' active-site geometry with the experimentally resolved TsT4Ease structure for further structural analysis (Candido et al., 2026).

Datasets. We evaluate retrieval across a hand-curated set of T4Ease-related enzymes collected from the literature. This dataset contains 144 T4Ease variants and homologs drawn from studies, and none of the sequences are present in the model's train set. After sequence deduplication, 123 unique

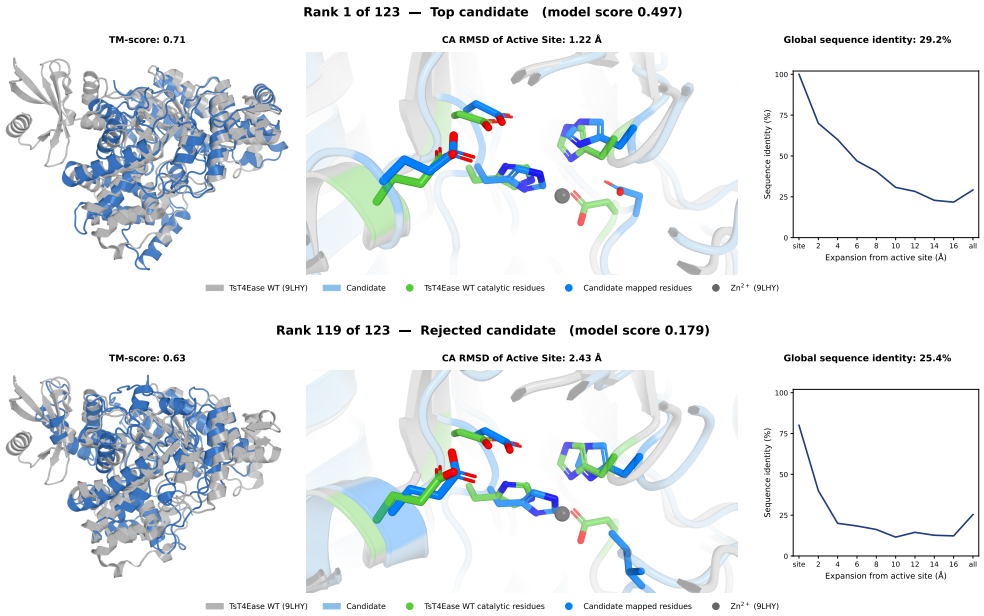


Figure 8: **Structural comparison of a top-ranked and a rejected candidate against TsT4Ease WT.** The top row shows the highest-ranked candidate from the curated literature set of 123 enzymes. All five annotated TsT4Ease active-site residues are conserved at the corresponding positions in this candidate, and sequence similarity remains high in the surrounding active-site region, reaching 60% within 4 Å of the active site. In contrast, the bottom row shows a low-ranked candidate rejected by the model. Although four of the five key catalytic residues are conserved, the candidate contains an isoleucine in place of the aspartic acid involved in coordinating the catalytic metal. Sequence similarity also decreases sharply away from the catalytic residues, falling to only 20% within 4 Å of the active site. Together, these examples suggest that `CIRCE` preferentially retrieves enzymes with greater conservation of the catalytic site and its surrounding local environment, despite receiving neither structural nor active-site information at inference time.

entries remained. This dataset enables a fine-grained evaluation of whether `CIRCE` can distinguish promising candidate homologs. Since many of the entries have not been experimentally validated for D-fructose 4-epimerization and are instead proposed as candidates based on sequence similarity to known T4Eases or related epimerases, we use this set to examine whether `CIRCE` can prioritize candidates that are more likely to catalyze the target reaction.

Across the 123 unique sequences in the dataset, the five catalytic residues of TsT4Ease WT (9LHY) are largely invariant: H355 is conserved in all 123 sequences, D174 in 122 (99.2%) and H175 in 121 (98.4%), while D284 is conserved in 99 (80.5%) and E141 in 96 (78.0%). Ninety sequences (73.2%) retain all five residues, and their absence is strongly associated with low rank. Specifically, 39 to the top 40 ranked candidates have all five active site residues intact compared to only 12 of the worst 40 candidates. As shown in 8, the scores generated by `CIRCE` correlate with local sequence similarity around the active site. Sequence identity to `9LHY` within a 2 Å shell of the catalytic residues correlates with retrieval score with a spearman of 0.65. Further, sequence identity to `9LHY` within a 4 Å and 6 Å shells of the catalytic residues correlates with retrieval score with a spearman of 0.49 and 0.47 respectively.