ENZYFUSE: CROSS-VIEW INTERACTION OVER SPATIAL NEIGHBORHOODS FOR CATALYTIC RESIDUE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting catalytic residues supports mechanistic annotation and enzyme engineering. When annotated homologs are scarce, prediction can draw on evidence from the protein's sequence, evolutionary and structural views. The challenge is to learn from both view-specific evidence and its organization across spatially neighboring residues. We introduce EnzyFuse, a task-specific multi-view model built on pretrained sequence and alignment encoders and equivariant geometric message passing. Directly supervised view branches provide residue predictions, while an interaction branch combines their representations over spatial neighborhoods. Catalytic partner retrieval reuses the residue annotations as structured auxiliary supervision, identifying catalytic partners near an annotated anchor during training. A learned gate combines branch logits during training; inference uses a validation-selected average of the available logits. We evaluate EnzyFuse under joint sequence–structure similarity control, with post-split audits and controlled ablations to assess generalization and component contributions. Ablations and inference-time view masking indicate that evolutionary and structural information both contribute, and that the gain over a three-branch late-fusion model arises when spatial interaction is combined with catalytic partner retrieval. Under this protocol, a single EnzyFuse model outperforms all evaluated baselines, including a multi-model ensemble, in protein-averaged area under the precision–recall curve. This advantage persists on the subset with no detected neighbor among catalytic-labeled training proteins.

1 INTRODUCTION

Catalytic residue prediction identifies amino acids involved in chemical transformation and transition-state stabilization, supporting mechanistic studies, functional annotation and enzyme design (Pan et al., 2024; Rieger et al., 2026). Experimental characterization through mutagenesis and biochemical assays is costly and difficult to scale, motivating computational prediction. Annotation-transfer methods use sequence alignments or catalytic geometric templates to map known residues onto a query, but require a reliably matching annotated reference (Rieger et al., 2026; Hackett et al., 2026). When such references are scarce, sequence context, evolutionary patterns in multiple sequence alignments (MSAs), and three-dimensional spatial arrangements still provide complementary views of the protein. In particular, a lack of annotated homologs does not preclude constructing an MSA from unannotated relatives. We therefore ask **how learning from sequence, evolutionary and structural views can improve catalytic residue prediction in low-homology settings**.

Deep learning methods have advanced this task using different combinations of evidence. Squidly combines protein language models with function-guided contrastive learning using sequence alone at prediction time (Rieger et al., 2026). SCREEN incorporates sequence, evolutionary and structural features (Pan et al., 2024), whereas EasIFA combines sequence and structural representations with reaction information (Wang et al., 2024). However, differences in their inputs, architectures and supervision make cross-method comparisons insufficient to isolate the contributions of additional views and how they are learned. Catalytic residues are typically conserved, and their relative positions within an active site are constrained by the catalytic mechanism (Bartlett et al., 2002; Hackett et al., 2026). Predictive evidence therefore concerns both residue properties and how those prop-

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.

erties are organized in spatial neighborhoods. Learning task-specific information from each view and modeling local relations across views are consequently complementary goals. Assessing their generalization also requires controlling sequence and structural similarity between training and test proteins. Sequence-based separation alone can retain structural neighbors, making it difficult to distinguish generalizable structural information from benefits confined to proteins resembling training examples.

We introduce EnzyFuse, a task-specific multi-view model that combines view-level predictions with structure-guided cross-view interaction. It builds on pretrained protein language models (Lin et al., 2023; Rao et al., 2021b), equivariant message passing and direct branch supervision. The interaction branch processes the view representations over spatial neighborhoods, relating the sequence and evolutionary features of neighboring residues to their geometric arrangement. Catalytic partner retrieval further supervises these local relations by identifying other catalytic residues near an annotated anchor. This auxiliary task restructures the existing catalytic annotations into neighborhood-level targets; it introduces no additional experimental labels. All branches are trained jointly with a learned gate, whereas a validation-selected logit average is used at inference. Neither the gate nor catalytic annotations are needed for prediction. The methodological focus is this coupling of view-level prediction, spatial interaction and local catalytic supervision, rather than new encoders or a general-purpose fusion rule.

Building on prior efforts to control both sequence and structural similarity (Rieger et al., 2026), we evaluate EnzyFuse through model comparisons and controlled ablations of views and design choices. Post-split audits and similarity-stratified analyses assess proximity to catalytic-labeled training proteins, not to every protein used for training. A nearest-neighbor transfer control tests how much sequence-only splits reward annotation transfer. On the jointly controlled test set, EnzyFuse improves protein-averaged AUPRC over the evaluated baselines, including on enzymes with no detected neighbor among catalytic-labeled training proteins. Squidly retains higher global-threshold F1 and MCC. View-removal and design comparisons quantify the contributions of the input views, spatial interaction and partner supervision.

Our contributions are as follows:

- We develop a catalytic-residue predictor that couples directly supervised protein-view branches with structure-guided cross-view interaction over spatial residue neighborhoods.
- We formulate catalytic partner retrieval as structured auxiliary supervision derived from the existing catalytic labels, improving prediction without an additional inference task.
- We evaluate the model through joint sequence–structure splits, annotation-transfer controls and audits against catalytic-labeled training proteins. Model comparisons and ablations establish ranking gains and quantify view and component contributions within this framework.

2 Related Work

Catalytic residue prediction. Computational prediction of catalytic residues follows two lines. Annotation transfer maps known catalytic residues onto a query through sequence homology or geometric templates of active sites, and depends on a matching annotated reference at prediction time (Porter et al., 2004; Ribeiro et al., 2018; Hackett et al., 2026). Earlier learned predictors combined residue conservation with sequence and structural features in classical classifiers (Gutteridge et al., 2003; Petrova & Wu, 2006; Capra & Singh, 2007; Zhang et al., 2008). Deep learning methods, which also predict enzyme function from sequence or structure (Gligorijević et al., 2021; Yu et al., 2023), now train residue-level prediction models on enzyme annotations, and they draw on different combinations of evidence. Squidly relies on the sequence alone through a protein language model (Rieger et al., 2026), and AEGAN combines local structural neighborhoods with evolutionary profiles (Shen et al., 2023). SCREEN brings together language-model, evolutionary-profile and structure features (Pan et al., 2024), and EasIFA adds reaction information to sequence and structure representations (Wang et al., 2024). These methods show that both strong sequence representations (Rives et al., 2021; Elnaggar et al., 2021; Lin et al., 2023) and additional evolutionary, structural and reaction information can support catalytic residue prediction.

Combining protein views. The sequence, evolutionary and structural views describe the same protein and are related: coevolution in a multiple sequence alignment reflects residue contacts in the folded structure (Morcos et al., 2011; Rao et al., 2021a;b). Different views nonetheless make different information explicit, such as family-level conservation in the alignment and spatial proximity in the structure. For catalytic residues, the relevant evidence often spans views, since catalytic residues are typically conserved and lie together in the active-site environment (Bartlett et al., 2002). Work on multimodal learning (Baltrušaitis et al., 2018) points to two goals in combining views. Each input should receive an effective learning signal, since jointly trained networks can underuse some of their inputs (Wang et al., 2020; Wu et al., 2022; Peng et al., 2022; Huang et al., 2022), and relations across inputs that single-input models cannot learn should still be modeled (Du et al., 2023). EnzyFuse adopts these established ideas for catalytic residue prediction, combining direct view-level supervision with structure-guided interaction and a label-derived catalytic partner task. Its contribution concerns this task-specific design and its controlled evaluation.

Homology-controlled evaluation. Data preparation for catalytic residue prediction reduces redundancy and separates training from test data mainly by sequence identity, for example by reducing redundancy within the training data or by clustering training and evaluation data together (Pan et al., 2024; Wang et al., 2024). Sequence-based preparation leaves structural similarity unconstrained, and even sequence-level relatives can remain: in legacy test sets reused by SCREEN (Pan et al., 2024), up to 85.4% of test sequences exceed 90% identity to the training data (Rieger et al., 2026). Squidly incorporates both sequence and structural similarity into test-set construction (Rieger et al., 2026). Similarity between training and test data inflates performance estimates across machine-learning-based science (Kapoor & Narayanan, 2023), and similarity-aware splits are increasingly used in other protein tasks (Dallago et al., 2021; Bushuiev et al., 2024; Buttenschoen et al., 2024). To assess multi-view generalization, we use a joint sequence–structure split (Section 5.1) for controlled comparisons of the input views and fusion designs within EnzyFuse.

3 PRELIMINARIES

3.1 PROBLEM FORMULATION

A protein p is a chain of L residues with amino-acid sequence $\mathbf{s}_p = (s_1, \ldots, s_L)$. Two further views of the chain may be available: a multiple sequence alignment (MSA) $\mathcal{A}_p$ of its homologs, and a predicted structure (Jumper et al., 2021; Varadi et al., 2024) given by the C_α coordinates $\mathbf{X}_p = [\mathbf{x}_1, \ldots, \mathbf{x}_L] \in \mathbb{R}^{3\times L}$. Following prior work (Wang et al., 2024), we cast the task as per-residue prediction: a predictor outputs a probability $\hat{y}_i \in [0, 1]$ that residue i is catalytic. Because the MSA and the structure can be missing for a query, we write

$$m_p^a \in \{0, 1\}, \qquad a \in \mathcal{B} = \{\mathrm{seq}, \mathrm{msa}, \mathrm{str}\}, \qquad m_p^{\mathrm{seq}} \equiv 1, \tag{1}$$

to indicate whether view a is available for protein p, and require the predictor to remain defined when $m_p^a = 0$. Residue-level quantities are indexed by i, and the protein index is omitted where no ambiguity arises.

We use three types of residue annotation from UniProtKB/Swiss-Prot (The UniProt Consortium, 2025): catalytic residues (`ACT_SITE`), binding sites (`BINDING`), and other sites of interest (`SITE`). We denote by $y_{i,c} \in \{0, 1\}$ the label of residue i in class $c \in \mathcal{C} = \{\mathrm{cat}, \mathrm{bind}, \mathrm{oth}\}$. This paper targets the catalytic class, which is also the class predicted by the methods we compare against (Pan et al., 2024; Rieger et al., 2026). Its labels define the catalytic set of each protein,

$$\mathcal{S}_p = \{\, i : y_{i,\mathrm{cat}} = 1 \,\} \subseteq \{1, \ldots, L\}, \qquad n_p = |\mathcal{S}_p|. \tag{2}$$

EnzyFuse predicts all three classes as independent binary outputs $\hat{y}_{i,c}$, with $\hat{y}_i = \hat{y}_{i,\mathrm{cat}}$. The binding and other classes serve only as additional training signal; their predictions are reported in Appendix H.

3.2 EQUIVARIANCE

The structure view is given in an arbitrary coordinate frame, whereas the catalytic role of a residue does not depend on that frame. The per-residue prediction should therefore be unchanged under any

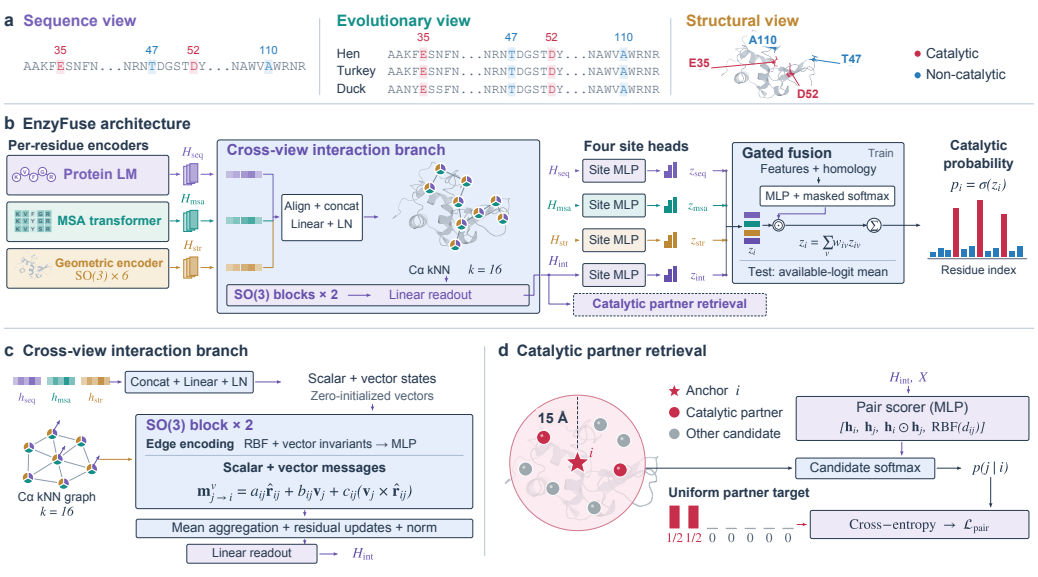


Figure 1: Overview of EnzyFuse. (a) Illustration of sequence, evolutionary and structural views, with catalytic residues in red and selected non-catalytic residues in blue. These annotations illustrate the views and are not inputs to the predictor. (b) Three view encoders and an interaction branch feed four separate site heads. The gate is used during training; inference averages available-branch logits. (c) Two $\mathrm{SO}(3)$-equivariant blocks integrate the views on a C_α graph with $k = 16$. (d) Training-only partner retrieval scores other residues within 15 Å of a catalytic anchor, with a uniform target over its catalytic partners.

rotation and translation of the chain, $\mathbf{x}_i \mapsto \mathbf{R}\,\mathbf{x}_i + \mathbf{t}$ for all i, with $\mathbf{R} \in \mathrm{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$. We build structural features from relative positions $\mathbf{x}_j - \mathbf{x}_i$, which removes translations. A per-residue feature $\mathbf{u}_i \in \mathbb{R}^{d_u}$ is *invariant* if rotation leaves it unchanged, and a feature $\mathbf{V}_i \in \mathbb{R}^{3\times d_v}$, whose columns are vectors in $\mathbb{R}^3$, is *equivariant* if it rotates with the structure:

$$\mathbf{u}_i\big(\mathbf{R}\mathbf{X}_p + \mathbf{t}\big) = \mathbf{u}_i\big(\mathbf{X}_p\big), \qquad \mathbf{V}_i\big(\mathbf{R}\mathbf{X}_p + \mathbf{t}\big) = \mathbf{R}\,\mathbf{V}_i\big(\mathbf{X}_p\big), \tag{3}$$

with $\mathbf{t}$ added to every column. Equivariant features propagate directional information through message passing (Thomas et al., 2018; Fuchs et al., 2020; Satorras et al., 2021; Schütt et al., 2021; Jing et al., 2021); our predictions are read out from invariant features. Reflection invariance is not required, because proteins are chiral.

4 Method

EnzyFuse encodes three protein views and combines four directly supervised branch predictions (Figure 1a,b). The interaction branch integrates the view representations over spatial neighborhoods (Figure 1c), while catalytic partner retrieval supervises local relations during training (Figure 1d). All branches are trained jointly, allowing interaction losses to update the view representations. A learned gate combines logits during training; inference averages the available branch logits.

4.1 View-specific encoding

The three encoders in Figure 1b produce residue representations of dimension $d = 512$. The sequence branch reads the final-layer representation of a frozen ESM-2 650M model (Lin et al., 2023) and projects it to $\mathbf{h}_i^{\mathrm{seq}} \in \mathbb{R}^d$ by a linear layer with layer normalization. The MSA branch reads the query-row representation of a frozen MSA Transformer (Rao et al., 2021b), together with an indicator of whether an alignment exists, and maps it to $\mathbf{h}_i^{\mathrm{msa}} \in \mathbb{R}^d$ by a two-layer perceptron. Both encoders are used without fine-tuning, so the two branches learn only how to read these representations for the task.

The structure branch operates on a k-nearest-neighbor graph $\mathcal{G}$ (Ingraham et al., 2019) over the C_α coordinates, with $k = 16$. Each node starts from an embedding of its residue type and its predicted local confidence (pLDDT), which forms the invariant feature $\mathbf{u}_i \in \mathbb{R}^{d_u}$ with $d_u = 128$. The equivariant feature $\mathbf{V}_i \in \mathbb{R}^{3\times d_v}$, $d_v = 16$, starts at zero. Six message-passing blocks (Gilmer et al., 2017) update both features. In each block, an edge network maps the invariant features $\mathbf{u}_i$ and $\mathbf{u}_j$ of the two endpoints, together with rotation-invariant edge features, to an edge embedding. The edge features are a radial expansion of the distance d_{ij}, the projections of $\mathbf{V}_i$ and $\mathbf{V}_j$ onto the unit edge direction $\hat{\mathbf{r}}_{ij}$, the cosines between corresponding channels of $\mathbf{V}_i$ and $\mathbf{V}_j$, and their norms. The edge embedding yields an invariant message $\mathbf{g}_{ij}$ and an equivariant message that combines the three directions available to an edge,

$$\Delta\mathbf{V}_i = \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)}\Big(\hat{\mathbf{r}}_{ij}\boldsymbol{\alpha}_{ij}^{\top} + \mathbf{V}_j\,\mathrm{diag}(\boldsymbol{\beta}_{ij}) + \big(\mathbf{V}_j\times\hat{\mathbf{r}}_{ij}\big)\,\mathrm{diag}(\boldsymbol{\gamma}_{ij})\Big), \tag{4}$$

where $\boldsymbol{\alpha}_{ij}, \boldsymbol{\beta}_{ij}, \boldsymbol{\gamma}_{ij} \in \mathbb{R}^{d_v}$ are predicted from the edge embedding; the cross product acts on each column of $\mathbf{V}_j$. Residual updates mix vector channels without bias and pass their invariant summaries to the scalar features (Appendix B). Each term in equation 4 is rotation-equivariant; the cross product permits chirality-sensitive features. The readout produces $\mathbf{h}_i^{\mathrm{str}} \in \mathbb{R}^d$ from invariant features only. Residues without coordinates receive zero representations and are marked unavailable.

4.2 Structure-guided cross-view interaction

Catalytic residues are typically conserved and lie together in the active-site environment (Bartlett et al., 2002). The view branches predict from their own representations. The interaction branch instead processes sequence, evolutionary and geometric information jointly over the spatial neighborhood of each residue, and is designed to capture how conservation and chemical environment are arranged around it (Figure 1c). On each node of the structure graph, it concatenates $\mathbf{h}_i^{\mathrm{seq}}$, $\mathbf{h}_i^{\mathrm{msa}}$ and $\mathbf{h}_i^{\mathrm{str}}$ and projects them to an invariant feature, while its equivariant feature starts at zero. Two message-passing blocks of the form in Section 4.1 then update both features on the same graph, and $\mathbf{h}_i^{\mathrm{int}} \in \mathbb{R}^d$ is read out from the invariant feature. Because messages combine the invariant features of neighboring nodes, the sequence and evolutionary representation of each neighbor enters the update of residue i. The branch is available exactly where the structure branch is, and its representation is zero elsewhere. When the alignment is missing, the MSA branch receives a zero input with its indicator set to zero, so $\mathbf{h}_i^{\mathrm{msa}}$ takes a default representation shared by all residues, which the interaction branch still reads.

4.3 Branch-wise prediction and fusion

Each branch $b \in \mathcal{K} = \{\mathrm{seq}, \mathrm{msa}, \mathrm{str}, \mathrm{int}\}$ has its own site head, a two-layer perceptron with GELU activation that produces logits $\mathbf{z}_i^b \in \mathbb{R}^{|\mathcal{C}|}$ for the three classes as independent binary outputs. The heads do not share parameters, so every branch forms a residue prediction of its own (Figure 1b). The figure shows the catalytic logits; each head also produces binding and other-site logits.

We separate the training-time combination from the inference rule. During training, a per-residue gate parametrizes the combined logits and receives gradients through their loss:

$$\mathbf{z}_i = \sum_{b\in\mathcal{K}} w_i^b\,\mathbf{z}_i^b, \qquad \mathbf{w}_i = \mathrm{softmax}\Big(\mathrm{MLP}\big([s_{\mathrm{seq}}, \mathbf{h}_i^{\mathrm{seq}}, \mathbf{h}_i^{\mathrm{msa}}, \mathbf{h}_i^{\mathrm{str}}, \mathbf{h}_i^{\mathrm{int}}]\big) + \log \boldsymbol{m}_i\Big), \tag{5}$$

where $\boldsymbol{m}_i$ collects the availability of the four branches at residue i: the sequence and MSA branches follow the protein-level indicators of equation 1, the structure branch additionally requires coordinates for residue i, and the interaction branch follows the structure branch. The gate is a two-layer perceptron with SiLU activation and 64 hidden units. Its input is not masked, but an unavailable branch receives zero weight at its output, and s_{seq} is the maximum sequence identity of the protein to the training set.[1] Each branch also receives direct supervision, so its prediction is not trained solely through the gate. At inference, the gate is not used, and the available logits are averaged,

$$\mathbf{z}_i = \frac{\sum_{b\in\mathcal{K}} m_i^b\,\mathbf{z}_i^b}{\sum_{b\in\mathcal{K}} m_i^b}, \tag{6}$$

[1] s_{seq} is computed with MMseqs2 over hits covering at least half of the query, excluding hits from the protein's own connected component; it is zero when no hit is found.

because this rule gives higher validation AUPRC than gated inference (0.7227 versus 0.7139; Appendix B). This empirical choice does not require the learned weights to generalize as a routing rule. Section 5.4 separately tests the gate's role during training. Inference requires neither the gate nor s_{seq}, and reduces to the sequence branch when both other views are unavailable. The catalytic prediction is $\hat{y}_i = \sigma(z_{i,\text{cat}})$.

4.4 Training objective

Residue classification. The masked binary cross-entropy $\mathcal{L}_{\text{res}}$ averages over labeled residue–class pairs in a minibatch $\mathcal{M}$ (Appendix B). Positives receive weight $\omega_c = \min(\sqrt{N_c^- / N_c^+}, 100)$, based on training-set class counts; $\mu_{i,c} = 0$ masks unknown labels. Catalytic labels are masked for enzymes with an EC number but no catalytic annotation. We supervise the combined logits and each available branch directly (Lee et al., 2015; Szegedy et al., 2015), independently of its fusion weight.

Ranking within a protein. The evaluation ranks the residues of each protein, whereas residue classification treats them independently. We therefore add a listwise term (Cao et al., 2007; Bruch et al., 2019) on the combined catalytic logits, averaged over the set $\mathcal{P} \subseteq \mathcal{M}$ of proteins with at least one labeled catalytic residue,

$$\mathcal{L}_{\text{rank}} = -\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{1}{n_p} \sum_{i \in \mathcal{S}_p} \log \frac{\exp(z_{i,\text{cat}})}{\sum_{j:\, \mu_{j,\text{cat}}=1} \exp(z_{j,\text{cat}})}, \tag{7}$$

where the normalization runs over the residues with a catalytic label. The term treats these residues as one candidate list and the catalytic set as a uniform target over it.

Catalytic partner retrieval. Whereas the ranking objective prioritizes catalytic residues within each protein, partner retrieval conditions on an annotated catalytic residue and scores the candidates in its spatial neighborhood (Figure 1d). For each catalytic residue i with coordinates, the candidates $\mathcal{R}_i$ are the other residues with coordinates within 15 Å of it. This radius-based set excludes the anchor and differs from the k-nearest-neighbor graph used for message passing. A scoring network f rates each candidate from $\mathbf{h}_i^{\text{int}}$, $\mathbf{h}_j^{\text{int}}$, their elementwise product and a radial expansion of d_{ij}. A softmax over $\mathcal{R}_i$ defines the candidate probability $p(j \mid i)$, which is trained against a uniform distribution over the catalytic candidates:

$$\mathcal{L}_{\text{pair}} = -\frac{1}{|\mathcal{P}'|} \sum_{p \in \mathcal{P}'} \frac{1}{|\mathcal{S}'_p|} \sum_{i \in \mathcal{S}'_p} \frac{1}{|\mathcal{R}_i \cap \mathcal{S}_p|} \sum_{j \in \mathcal{R}_i \cap \mathcal{S}_p} \log \frac{\exp f(i,j)}{\sum_{l \in \mathcal{R}_i} \exp f(i,l)}, \tag{8}$$

where $\mathcal{S}'_p$ is the set of catalytic residues of p with at least one catalytic candidate and $\mathcal{P}' \subseteq \mathcal{M}$ the proteins with a structure for which it is non-empty. This objective derives structured targets from the same catalytic labels used for residue classification; it provides no independent annotation source. The scoring network is used only during training.

Full objective. With $\lambda_{\text{aux}} = 0.5$ and $\lambda_{\text{rank}} = \lambda_{\text{pair}} = 0.05$, the training objective is

$$\mathcal{L} = \mathcal{L}_{\text{res}}(\mathbf{z}) + \lambda_{\text{aux}} \sum_{b \in \mathcal{K}} \mathcal{L}_{\text{res}}(\mathbf{z}^b) + \lambda_{\text{rank}}\, \mathcal{L}_{\text{rank}} + \lambda_{\text{pair}}\, \mathcal{L}_{\text{pair}}. \tag{9}$$

During training, the structure of each protein is also dropped with probability 0.3, so that the model learns to predict without it. Such a protein does not contribute to the structure, interaction or partner-retrieval losses at that step. Implementation details and hyperparameter settings are provided in Appendix B.

5 Experiments

5.1 Experimental setup

Data and homology-controlled splits. We use UniProtKB/Swiss-Prot release 2026_02. In a control split built by sequence identity alone (40%), 93.2% of test enzymes with predicted structures

Table 1: Comparison with existing methods on the catalytic test subset (6,804 proteins). Metrics are computed per protein and averaged over proteins. Retrained models report mean$_{\pm\text{std}}$ over three runs; released-weight and transfer baselines are evaluated once. F1 and MCC use a single threshold per method selected on the validation set. Best in **bold**, second underlined. *Five ensemble members per run, except for one run with four.

Method	AUPRC	AUROC	F1	MCC
AEGAN	$0.2352_{\pm 0.0126}$	$0.8868_{\pm 0.0042}$	$0.0878_{\pm 0.0190}$	$0.1376_{\pm 0.0139}$
EasIFA	$0.3647_{\pm 0.0132}$	$0.7655_{\pm 0.0093}$	$0.2833_{\pm 0.0124}$	$0.2922_{\pm 0.0130}$
SCREEN	$0.6722_{\pm 0.0051}$	$0.9600_{\pm 0.0041}$	$0.5266_{\pm 0.0085}$	$0.5394_{\pm 0.0080}$
Squidly*	$\underline{0.7012_{\pm 0.0108}}$	$\underline{0.9739_{\pm 0.0026}}$	$\mathbf{0.5779_{\pm 0.0109}}$	$\mathbf{0.5950_{\pm 0.0102}}$
SCREEN (released)	0.6271	0.9697	0.4498	0.4803
Homology transfer (MMseqs2)	0.2989	0.6496	0.3001	0.3011
Structure transfer (Foldseek)	0.3194	0.6619	0.3282	0.3267
EnzyMM	0.1162	0.6368	0.1433	0.1640
EnzyFuse (ours)	$\mathbf{0.7568_{\pm 0.0035}}$	$\mathbf{0.9903_{\pm 0.0007}}$	$\underline{0.5521_{\pm 0.0043}}$	$\underline{0.5726_{\pm 0.0067}}$

share a Foldseek cluster with a training protein (TM-score threshold 0.5; Zhang & Skolnick, 2004). We therefore construct a similarity graph in which proteins are linked if they share at least 30% sequence identity (MMseqs2; Steinegger & Söding, 2017) or a TM-score of at least 0.5 (Foldseek; Van Kempen et al., 2024). Each connected component is assigned as a whole to the training, validation or test set, so no graph edge crosses the split. Construction details are given in Appendix A; independent similarity audits are reported in Section 5.5. The training set contains 224,427 proteins with at least one annotation in any of the three classes of Section 3.1, and the validation and test sets contain 16,467 and 16,195 such proteins. We evaluate catalytic residue prediction, our primary task, on the subsets of 7,200 validation and 6,804 test proteins with catalytic annotations. Proteins without a predicted structure, 8.0% of these test proteins, are retained. The validation set is used for model selection and the test set for performance evaluation.

Evaluation protocol. For each protein in the catalytic test subset, we compute the average precision (AP) of the catalytic predictions over its residues, and report AUPRC, our primary metric, as the mean AP over proteins (Pan et al., 2024; Rieger et al., 2026), which suits the strong class imbalance of the task (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). AUROC is averaged over proteins in the same way. A protein for which a method returns no prediction receives an AP of zero, so the score reflects both prediction quality and coverage. F1 and MCC are computed per protein at a single method-specific threshold and averaged over proteins. For retrained models, we use three random seeds and select a threshold for each run by maximizing protein-averaged validation F1; the mean threshold across runs is then fixed for test evaluation. Transfer methods and the released-weight SCREEN model are evaluated once. Statistical comparisons are based on paired differences in per-protein AP, with AP first averaged across runs within each method. These differences are tested with a Wilcoxon signed-rank test (Wilcoxon, 1945), and a bootstrap 95% confidence interval is computed for their mean (Efron & Tibshirani, 1994). A difference is reported as significant when $p < 0.05$ and the interval excludes zero.

5.2 Comparison with existing methods

Under the protocol of Section 5.1, we compare EnzyFuse with deep learning models and annotation-transfer methods (Table 1). Baseline configurations are detailed in Appendix C. A single EnzyFuse model achieves the highest AUPRC and AUROC among the evaluated methods, including the Squidly ensemble. Its AUPRC of 0.757 exceeds that of this ensemble, the strongest baseline, by 0.056 (95% confidence interval $[0.049, 0.062]$) and that of retrained SCREEN by 0.085. The AUPRC improvement over each baseline is statistically significant (Appendix D). The advantage is not confined to a few enzymes: over most of the range of per-protein AUPRC, a larger fraction of enzymes reaches a given value with EnzyFuse than with any baseline, whereas near an AUPRC of 1 EnzyFuse and the Squidly ensemble coincide (Figure 2). However, the Squidly ensemble achieves higher F1 and MCC at the validation-selected global thresholds.

Table 2: Ablations, each retrained with three seeds. Δ is relative to EnzyFuse, with a bootstrap 95% CI (Wilcoxon $p \leq 2.1 \times 10^{-15}$); view rows give seed means only. †Also removes partner retrieval.

Variant	AUPRC	Δ	95% CI
EnzyFuse	$0.7568_{\pm 0.0035}$	—	—
Objective			
w/o partner retrieval	$0.7360_{\pm 0.0055}$	-0.0208	$[-0.0232, -0.0187]$
Interaction			
w/o spatial edges	$0.7233_{\pm 0.0045}$	-0.0335	$[-0.0365, -0.0305]$
w/o interaction branch†	$0.7254_{\pm 0.0048}$	-0.0314	$[-0.0376, -0.0274]$
Views			
w/o sequence	$0.6838_{\pm 0.0117}$	-0.0730	—
w/o MSA	$0.6792_{\pm 0.0099}$	-0.0776	—
w/o structure	$0.7184_{\pm 0.0032}$	-0.0384	—

To assess whether the AUPRC advantage also holds at the top of each protein's residue ranking, we examine top-k hit rates. These measure how often the k highest-scoring residues of a protein include at least one annotated catalytic residue, without applying a global score threshold. The top-ranked residue is annotated as catalytic in 70.5% of test enzymes for EnzyFuse, compared with 64.0% for SCREEN, the strongest baseline at this rank. Among the top three residues, the hit rates are 84.8% for EnzyFuse and 78.5% for the Squidly ensemble, the strongest baseline at $k = 3$ (Table 4). These results show that EnzyFuse more often places annotated catalytic residues among the few candidates prioritized for experimental follow-up.

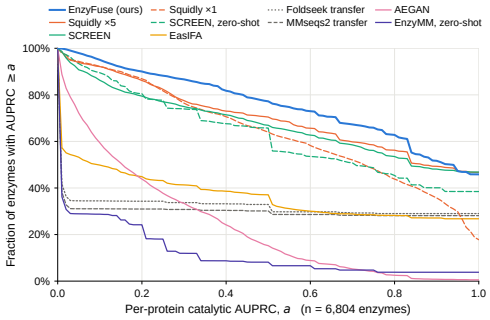


Figure 2: Fraction of the 6,804 test enzymes whose per-protein AUPRC is at least a.

Annotation-transfer methods obtain much lower AUPRC under this evaluation protocol. Homology transfer and EnzyMM return predictions for only 31.7% and 34.7% of test proteins, respectively. Limited prediction coverage therefore contributes to their low overall AUPRC, as proteins without predictions receive an AP of zero. Section 5.5 further examines performance as a function of similarity between test and training proteins.

5.3 Contributions of protein views

We next assess the incremental predictive value of each view within EnzyFuse. Removing a view and retraining lowers AUPRC (Table 2): by 0.078 for the MSA view, 0.073 for the sequence view and 0.038 for structural input. In particular, the evolutionary and structural inputs improve prediction even when pretrained sequence representations remain available. The no-structure variant also disables spatial interaction and partner retrieval, so its decrease measures the combined contribution of structural input and the components that use it.

Predictions from the jointly trained branch heads provide complementary descriptive evidence. On the 6,259 test enzymes with all three views, averaging the branch logits reaches an AUPRC of 0.76, compared with 0.71 for the strongest individual head, the interaction branch (Appendix E). These heads are not separately trained single-view baselines. Controlled masking at inference is reported in Appendix F; the next subsection examines the structure-dependent components separately.

5.4 Cross-view interaction and supervision

We examine the components that model local cross-view relations while retaining the input views and training protocol (Table 2). Removing the interaction branch also removes partner retrieval and lowers AUPRC by 0.031 (95% confidence interval $[-0.038, -0.027]$). Replacing its spatial message passing with a per-residue perceptron with a matched number of parameters lowers AUPRC

by 0.034, supporting the value of spatial message passing beyond parameter-matched per-residue processing. Removing partner retrieval alone lowers AUPRC by 0.021 ($[-0.023, -0.019]$), despite leaving the inference procedure unchanged. Thus, both spatial interaction and supervision of local catalytic relations contribute within the full framework. Appendix E reports comparisons for enzymes with no detected neighbor among catalytic-labeled training proteins, alongside secondary ablations. Appendix G examines how partner supervision changes the ranking of residues near catalytic sites.

The gate's training role is distinct from its use at inference. With equal-logit averaging at inference, replacing the training gate with equal weights lowers AUPRC from 0.7568 to 0.7418 (Table 8). This comparison supports retaining gated training under the current protocol, despite the validation preference for mean-logit inference. It does not establish reliable homology-dependent routing or isolate the contribution of the similarity input.

5.5 Generalization under homology control

We group test enzymes by h, the larger of their maximum sequence identity and maximum query-normalized TM-score against catalytic-labeled training proteins (Appendix A). This reference set excludes training proteins without catalytic labels; the audit therefore does not establish separation from all training inputs. Enzymes without a detected hit ($h = 0$) form the largest group, 3,322 of the 6,804 test enzymes (Figure 3). In this group, EnzyFuse exceeds the Squidly ensemble by 0.064 ($[0.053, 0.075]$) and SCREEN by 0.116, and it also leads for $0 < h < 0.3$ (+0.048 and +0.080) and $0.4 \leq h < 0.6$ (+0.086 and +0.090). Neither comparison is significant for $0.3 \leq h < 0.4$. Among the 243 enzymes with $h \geq 0.6$, SCREEN is higher by 0.036 ($[0.018, 0.054]$), while the difference from Squidly is not significant under our joint criterion. The gains therefore extend to low- and intermediate-similarity groups, but are not uniform across all similarity ranges.

Excluding test enzymes with a detected hit among catalytic-labeled training proteins at sequence identity of at least 0.3 or a query-normalized TM-score of at least 0.5 leaves a strict subset of 5,746 enzymes (84.5%). On this subset EnzyFuse reaches an AUPRC of 0.724 and remains ahead of every baseline, by 0.066 ($[0.059, 0.074]$) over the Squidly ensemble and by 0.102 over SCREEN (Appendix D). For comparison, a baseline that copies the annotation of the most similar training protein reaches an AUPRC of 0.845 on a split built by sequence identity alone, against 0.299 on our split (Appendix A). This separate control shows that high scores on a sequence-only split need not establish generalization beyond annotation transfer. Results with views masked at inference and by chain length are reported in Appendix F.

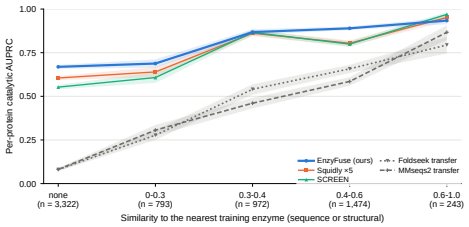


Figure 3: Per-protein AUPRC by similarity h to catalytic-labeled training proteins ("none": $h = 0$); group sizes below the axis. Bands: s.d. over three runs.

6 Conclusion

EnzyFuse combines established protein encoders with a task-specific design for learning from complementary views: directly supervised predictions, structure-guided interaction and label-derived catalytic partner retrieval. Under joint sequence–structure similarity control, it improves AUPRC over the evaluated baselines, including on enzymes with no detected neighbor among catalytic-labeled training proteins. Controlled ablations support the contribution of the additional views, spatial interaction and local catalytic supervision within this framework.

Limitations. Unannotated residues are treated as negatives, making evaluation sensitive to incomplete labels. Full-view preprocessing requires alignment search and predicted structures, although either view may be missing. Evaluation excludes proteins longer than 1022 residues, and performance declines on long chains (Appendix F). Squidly remains stronger on global-threshold F1 and MCC. The audit covers only catalytic-labeled training proteins; no detected neighbor in this subset does not establish separation from all training inputs or evolutionary independence.

AI USE STATEMENT

We used AI tools to check and revise grammar in the writing and to replace some word choices. This use does not affect any conclusion or result of this study.

ETHICS STATEMENT

This work uses only publicly available protein sequences, predicted structures and functional annotations, and involves no human subjects or personal data. Catalytic residue prediction could in principle inform enzyme design for any purpose; the model predicts residue roles and does not generate sequences or synthesis procedures.

REPRODUCIBILITY STATEMENT

The benchmark construction, including the homology-controlled splits, is described in Section 5.1 and Appendix A. Model and training hyperparameters are listed in Appendix B. Inference code, the trained checkpoints for all three seeds, configuration files and the feature-preprocessing pipeline are available at `https://anonymous.4open.science/r/EnzyFuse-1CEF/`. The test-set features and labels are released at `https://huggingface.co/datasets/enzfuse-anno/enzyfuse-data` and can be downloaded using a script in the repository. With these resources, a single command reproduces the per-seed test-set results reported in the paper.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2018.

Gail J Bartlett, Craig T Porter, Neera Borkakoti, and Janet M Thornton. Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*, 324(1):105–121, 2002.

Sebastian Bruch, Xuanhui Wang, Michael Bendersky, and Marc Najork. An analysis of the softmax cross entropy loss for learning-to-rank with binary relevance. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 75–78, 2019.

Anton Bushuiev, Roman Bushuiev, Petr Kouba, Anatolii Filkin, Marketa Gabrielova, Michal Gabriel, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, et al. Learning to design protein-protein interactions with enhanced generalization. In *International Conference on Learning Representations*, volume 2024, pp. 21010–21035, 2024.

Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 129–136, 2007. doi: 10.1145/1273496.1273513.

John A Capra and Mona Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, 2007.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240, 2006.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for SO(3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12200–12209, 2021.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pp. 8632–8656. PMLR, 2023.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2021.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3D roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pp. 1263–1272. Pmlr, 2017.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168, 2021.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Alex Gutteridge, Gail J Bartlett, and Janet M Thornton. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *Journal of Molecular Biology*, 330(4):719–734, 2003.

Raymund E Hackett, Ioannis G Riziotis, Martin Larralde, António JM Ribeiro, Georg Zeller, and Janet M Thornton. Investigating enzyme function by geometric matching of catalytic motifs. *Protein Science*, 35(9):e70744, 2026.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *International Conference on Machine Learning*, pp. 9226–9259. PMLR, 2022.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in Neural Information Processing Systems*, 32, 2019.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 2023.

Radoslav Krivák and David Hoksza. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10(1):39, 2018.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pp. 562–570. Pmlr, 2015.

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

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

Tong Pan, Yue Bi, Xiaoyu Wang, Ying Zhang, Geoffrey I Webb, Robin B Gasser, Lukasz Kurgan, and Jiangning Song. SCREEN: A graph-based contrastive learning tool to infer catalytic residues and assess enzyme mutations. *Genomics, Proteomics & Bioinformatics*, 22(6):qzae094, 2024.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8237. IEEE, 2022.

Natalia V Petrova and Cathy H Wu. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics*, 7(1):312, 2006.

Craig T Porter, Gail J Bartlett, and Janet M Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32(suppl_1):D129–D133, 2004.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2021a.

Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8844–8856, 2021b.

António J M Ribeiro, Gemma L Holliday, Nicholas Furnham, Jonathan D Tyzack, Katherine Ferris, and Janet M Thornton. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Research*, 46(D1):D618–D623, 2018.

William JF Rieger, Mikael Bodén, Frances Arnold, and Ariane Mora. Squidly harnesses enzyme functional hierarchy and contrastive learning to efficiently predict catalytic residues from sequence. *eLife*, 14:RP108186, 2026.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, pp. 9323–9332. PMLR, 2021.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems*, 30, 2017.

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pp. 9377–9388. PMLR, 2021.

Xiaowei Shen, Shiding Zhang, Jianyu Long, Changjing Chen, Meng Wang, Ziheng Cui, Biqiang Chen, and Tianwei Tan. A highly sensitive model based on graph neural networks for enzyme key catalytic residue prediction. *Journal of Chemical Information and Modeling*, 63(14):4277–4290, 2023.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

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

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Jérôme Tubiana, Dina Schneidman-Duhovny, and Haim J Wolfson. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nature Methods*, 19(6):730–739, 2022.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, 2024.

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375, 2024.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12692–12702. IEEE, 2020.

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.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

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

Tuo Zhang, Hua Zhang, Ke Chen, Shiyi Shen, Jishou Ruan, and Lukasz Kurgan. Accurate sequence-based prediction of catalytic residues. *Bioinformatics*, 24(20):2329–2338, 2008.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

A Data and split construction

Labels. Residue labels are derived from the `ACT_SITE`, `BINDING` and `SITE` features of UniProtKB/Swiss-Prot release 2026_02 (575,503 entries). In each class, a residue covered by a feature of that class is positive and every other residue is negative; a residue may be positive in several classes. For an enzyme that carries an EC number but no `ACT_SITE` feature, the catalytic labels are treated as unknown and masked in every loss term, whereas its binding and other-site labels are used. Proteins longer than 1022 residues, the context length of ESM-2, are removed.

Similarity graph and split. Two proteins are linked when MMseqs2 reports at least 30% sequence identity or Foldseek reports a TM-score of at least 0.5 between their AlphaFold DB models. The connected components of this graph are assigned as whole units to the training, validation and test sets, stratified by component size, first-level EC class and the presence of experimentally supported site annotations. In a control split built by clustering sequences alone (MMseqs2, 40% identity, coverage 0.8), 34,587 of the 37,106 test enzymes with a predicted structure (93.2%) share a Foldseek cluster (TM-score ≥ 0.5) with a training protein; under the same criterion, no test enzyme of our split does (54,346 test proteins and 24,160 test enzymes before the filters of Section 5.1).

Evaluation sets. After filtering, the training set contains 224,427 proteins with at least one annotation in any class, and the validation and test sets contain 16,467 and 16,195 such proteins. Catalytic residue prediction is evaluated on the 7,200 validation and 6,804 test proteins with at least one catalytic annotation, which carry 12,246 and 11,036 catalytic residues. Predicted structures are missing for 8.0% of these test proteins; they are kept and scored like all other proteins.

Residual similarity audit. The split prevents any edge of the similarity graph from crossing it, but does not bound the similarity of every cross-split pair. We therefore search each of the 6,804 catalytic test proteins against catalytic-labeled training proteins with MMseqs2 (`-s 7.5 -e 1e-3 --max-seqs 300`, hits with query coverage ≥ 0.5) and Foldseek (`-e 1e-3 --max-seqs 300 -c 0.5 --cov-mode 0 --alignment-type 2`). The nearest training sequence reaches 30% identity over at least half of the query for 12.2% of test proteins (832), and the nearest structure reaches a TM-score of 0.5 normalized by query length for 8.8% (597). The TM-score normalization matters: normalizing by alignment length instead gives 36.7% at the same threshold, because short partial alignments then count as matches. The 5,746 proteins (84.5%) that meet neither criterion form the strict subset of Section 5.5. Only 3,194 test proteins have a structural hit at all; proteins without a hit or without a structure are counted as below threshold. The similarity h of Section 5.5 is the larger of the maximum sequence identity and the maximum query-normalized TM-score, with zero for no hit; Table 6 lists the values behind Figure 3. Throughout these analyses, "no detected neighbor" means no detected neighbor among catalytic-labeled training proteins, not among all proteins used to train EnzyFuse.

Copying the nearest neighbor. To measure how much a split rewards annotation transfer, we copy the catalytic annotation of the most similar training protein through the alignment, without any learning, and score the result with the protocol of Section 5.1; proteins without a hit score zero. On a conventional split built by 30% sequence-identity clustering (9,029 test proteins), sequence copying answers 88.8% of queries and reaches an AUPRC of 0.845, and structure copying answers 91.8% and reaches 0.838. On our split, the two answer 31.7% and 41.3% of queries and reach 0.299 and 0.319. AUPRC on answered queries is high (0.91–0.95 on the conventional split), while the fraction of answered queries differs markedly between splits. As a control, placing the same number of positives at random positions yields an AUPRC of 0.007–0.009. The two splits have different test sets, so the comparison characterizes each split rather than isolating a causal effect of the split procedure.

B IMPLEMENTATION DETAILS

Figure notation. In Figure 1, H_b collects the residue representations $\mathbf{h}_i^b$ for branch b, and z_b denotes its catalytic logits. The displayed catalytic probability p_i is $\hat{y}_i = \sigma(z_{i,\mathrm{cat}})$. Panel (c) illustrates one vector channel of equation 4: a_{ij}, b_{ij} and c_{ij} correspond to entries of $\boldsymbol{\alpha}_{ij}$, $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\gamma}_{ij}$. Panel (d) uses $\mathbf{h}_i$ as shorthand for $\mathbf{h}_i^{\mathrm{int}}$ and X for the residue coordinates.

Encoders and branches. The sequence branch uses layer 33, the final layer, of ESM-2 650M. The MSA branch uses the query row of the final (twelfth) layer of ESM-MSA-1b; alignments are built by an MMseqs2 search against UniRef50 (Suzek et al., 2015), and each alignment is subsampled to at most 256 sequences, retaining the query and stratifying the other rows over eight bins of identity to it. Its input has 769 dimensions: 768 from the encoder and one indicator set to 1 for every residue of a protein with an alignment and to 0 otherwise. Encoder outputs are precomputed and stored in half precision. The structure branch embeds the residue type (21 classes, 32 dimensions) and pLDDT, maps them to 128 invariant dimensions, and uses 16 equivariant channels; the radial expansion uses 16 Gaussian basis functions over 0–12 Å (Schütt et al., 2017). The interaction branch projects the concatenated 3×512 representation to 128 dimensions with layer normalization (Ba et al., 2016) and reads out to 512 dimensions after two blocks. Each site head maps $512 \to 512 \to 3$ with GELU (Hendrycks & Gimpel, 2016) and dropout 0.1 (Srivastava et al., 2014). The gate maps its input to 64 hidden units with SiLU (Elfwing et al., 2018) and then to four weights, in the manner of a mixture of experts (Jacobs et al., 1991); its input also contains a slot for a structural similarity score, which is set to zero in all reported models.

Message-passing updates. The vector and scalar states are updated with residual connections:

$$\mathbf{V}_i \leftarrow \mathrm{VN}\big(\mathbf{V}_i + \mathbf{V}_i\mathbf{W}_{\mathrm{self}} + \Delta\mathbf{V}_i\mathbf{W}_{\mathrm{msg}}\big), \qquad \mathbf{u}_i \leftarrow \mathrm{LN}\Big(\mathbf{u}_i + \mathrm{MLP}\big([\mathbf{u}_i, \bar{\mathbf{g}}_i, \iota(\mathbf{V}_i)]\big)\Big). \quad (10)$$

Here $\mathbf{W}_{\mathrm{self}}, \mathbf{W}_{\mathrm{msg}} \in \mathbb{R}^{d_v \times d_v}$ mix channels without bias, VN is vector normalization (Deng et al., 2021), and $\bar{\mathbf{g}}_i$ averages invariant messages over $\mathcal{N}(i)$. The invariant summary $\iota(\mathbf{V}_i)$ contains updated vector norms and cosines between adjacent channels, passing directional information to the scalar state.

Masked classification loss. For a minibatch $\mathcal{M}$, the residue loss is

$$\mathcal{L}_{\mathrm{res}}(\mathbf{z}) = \frac{1}{\sum_{p,i,c} \mu_{i,c}} \sum_{p \in \mathcal{M}} \sum_{i=1}^{L_p} \sum_{c \in \mathcal{C}} \mu_{i,c} \Big[-\omega_c y_{i,c} \log \sigma(z_{i,c}) - (1 - y_{i,c}) \log\big(1 - \sigma(z_{i,c})\big) \Big]. \quad (11)$$

The denominator is bounded below by one. Branch-specific losses additionally mask residues where that branch is unavailable.

Partner retrieval. The scoring network maps $[\mathbf{h}_i^{\mathrm{int}}, \mathbf{h}_j^{\mathrm{int}}, \mathbf{h}_i^{\mathrm{int}} \odot \mathbf{h}_j^{\mathrm{int}}, \mathrm{RBF}_{16}(d_{ij})]$, with the radial expansion over 0–15 Å, to 128 hidden units with SiLU and then to a scalar. A protein contributes to $\mathcal{L}_{\mathrm{pair}}$ only if it has at least two catalytic residues with coordinates; the term is skipped for a minibatch without such proteins. The listwise term contributes zero for a minibatch without catalytic residues, and the denominators of the masked cross-entropy are bounded below by one.

Optimization and model selection. We use AdamW (Loshchilov & Hutter, 2019) with learning rate 3×10^{-4} and weight decay 10^{-4}, 500 warm-up steps (Goyal et al., 2017) followed by cosine decay (Loshchilov & Hutter, 2017) over 12 epochs, and gradient clipping at 5.0. A minibatch holds 128 proteins, concatenated without padding, and training runs in single precision. Each model is trained for 12 epochs, and the epoch with the highest catalytic AUPRC on the validation set, under equal weighting at inference, is kept. Training uses three random seeds (42, 43, 44); each run takes 4.6–7.1 hours on a single NVIDIA RTX A6000 or RTX 6000 Ada GPU.

Inference rule. On the validation set, equal weighting of the available branches gives a catalytic AUPRC of 0.7227 against 0.7139 for the learned gate. The paired difference (gate minus equal weighting) is -0.0088, with a bootstrap 95% confidence interval of $[-0.0112, -0.0065]$ and Wilcoxon $p = 5.1 \times 10^{-49}$. We therefore use equal weighting at inference. This comparison changes the inference rule, not the learned branch parameters. In contrast, Table 8 changes the training rule while retaining mean-logit inference: replacing the gate with equal weights lowers test AUPRC from 0.7568 to 0.7418. These comparisons address different roles of the gate. They do not establish a mechanism for the inference difference or isolate the effect of s_{seq} from that of the gate itself.

Threshold-dependent metrics. For F1 and MCC, candidate thresholds range from 0.02 to 0.96 in 48 steps. A zero denominator gives a score of zero, and proteins without a positive or without a negative residue are excluded from these two metrics. Bootstrap confidence intervals use 2,000 resamples of the test proteins.

C BASELINE REPRODUCTION

Retrained methods. SCREEN, Squidly, AEGAN and EasIFA are retrained with their released code using three random seeds on the same 224,427-protein training set as EnzyFuse. SCREEN truncates sequences to 1,000 residues. Methods that cannot process a test protein score zero on it. Squidly averages five members per run, except for one run with four trained members. The single-member comparison in Appendix D aggregates all 14 trained members. SCREEN is additionally evaluated once with its released weights, without retraining.

Departures from the published setups. AEGAN's paper describes a sampling ratio of one positive to ten negatives, whereas its released code samples four negatives per positive with replacement and removes duplicates, which gives an effective ratio of 3.91 in our runs; we score all residues for every method. EasIFA's data loader fixes a maximum length of 600 residues, and batches containing a longer chain fail its validity check; 27.4% of test enzymes exceed this limit. EasIFA's multi-class evaluation raises a division by zero at its first evaluation step; we patch this single expression.

Annotation transfer. Homology transfer searches each test protein against the training set with MMseqs2 (`-e 1e-3 -s 7.5 --max-seqs 50`) and copies the catalytic positions of the best hit that carries a catalytic annotation through the alignment; a protein without such a hit scores zero. It answers 31.7% of test proteins, with an AUPRC of 0.933 on the answered proteins and 0.299 overall. Requiring a query coverage of at least 0.5 halves the answer rate and lowers overall AUPRC by 0.137, so we report the setting without a coverage threshold. EnzyMM answers 34.7% of test proteins.

Structure transfer applies the same procedure with Foldseek (`-e 1e-3 --max-seqs 25`) against the AlphaFold DB models of catalytic training proteins, after removing models with mean pLDDT below 70 or a sequence that differs from Swiss-Prot. The first hit with a catalytic annotation donates its sites through the structural alignment. The same structure-quality filters retain 5,917 test queries; the method ultimately answers 2,807 test proteins. Mapped residues receive the hit's sequence-identity score and all others receive zero. Proteins with no usable structure, no hit or no mapped catalytic position count as unanswered.

D COMPLETE RESULTS AND STATISTICS

Table 3 compares EnzyFuse with each method protein by protein, and Table 4 reports top-k hit rates, which rank the residues within each enzyme and involve no threshold across enzymes; EnzyFuse has the highest hit rate at every k.

Table 3: Paired comparison of EnzyFuse with each method on per-protein catalytic AP over the $6{,}804$ test proteins. Δ is EnzyFuse minus the baseline; the interval is a bootstrap 95% confidence interval of the mean difference, and p is from a Wilcoxon signed-rank test. Wins, ties and losses count proteins.

Baseline	ΔAUPRC	95% CI	p	Wins / ties / losses
Squidly, single member	$+0.1176$	$[+0.1116, +0.1240]$	4.2×10^{-296}	4330 / 1104 / 1370
Squidly, 5-model ensemble	$+0.0556$	$[+0.0490, +0.0624]$	1.3×10^{-57}	2506 / 2598 / 1700
SCREEN, retrained	$+0.0846$	$[+0.0781, +0.0911]$	5.3×10^{-137}	2809 / 2554 / 1441
SCREEN, original weights	$+0.1297$	$[+0.1206, +0.1384]$	2.0×10^{-160}	3260 / 1856 / 1688
EasIFA, retrained	$+0.3921$	$[+0.3820, +0.4021]$	$< 10^{-300}$	4640 / 1440 / 724
AEGAN, retrained	$+0.5216$	$[+0.5135, +0.5294]$	$< 10^{-300}$	6384 / 33 / 387
MMseqs2 homology transfer	$+0.4579$	$[+0.4470, +0.4680]$	$< 10^{-300}$	4877 / 1619 / 308
Foldseek structure transfer	$+0.4374$	$[+0.4269, +0.4480]$	$< 10^{-300}$	4727 / 1552 / 525
EnzyMM	$+0.6406$	$[+0.6323, +0.6487]$	$< 10^{-300}$	6363 / 215 / 226

Table 4: Ranking within each enzyme. Top-k hit rate is the percentage of test enzymes with at least one catalytic residue among their k highest-scoring residues. No threshold across enzymes is involved. *Five ensemble members per run, except for one run with four.

Method	Top-1 (%)	Top-3 (%)	Top-5 (%)
AEGAN	15.3	32.3	41.9
EasIFA	33.1	42.6	46.0
SCREEN	64.0	75.8	80.2
Squidly*	63.5	78.5	85.5
SCREEN (released)	56.8	75.1	81.8
Homology transfer (MMseqs2)	30.6	31.0	31.0
Structure transfer (Foldseek)	33.5	34.4	34.4
EnzyMM	15.2	26.9	28.3
EnzyFuse (ours)	70.5	84.8	89.4

Table 5 repeats the comparison on the strict subset of Section 5.5, and Table 6 gives the comparison in each similarity group.

Table 5: Strict subset of $5{,}746$ test enzymes without a detected hit among catalytic-labeled training proteins at sequence identity ≥ 0.3 or query-normalized TM-score ≥ 0.5. EnzyFuse reaches 0.7239; Δ is EnzyFuse minus the baseline, with a bootstrap 95% CI and Wilcoxon p.

Baseline	AUPRC	Δ	95% CI	p
Squidly $\times 5$	0.6577	$+0.0662$	$[+0.0589, +0.0737]$	5.9×10^{-63}
Squidly $\times 1$	0.5970	$+0.1269$	$[+0.1198, +0.1339]$	1.1×10^{-258}
SCREEN	0.6217	$+0.1022$	$[+0.0953, +0.1096]$	2.8×10^{-152}
SCREEN, original weights	0.6029	$+0.1210$	$[+0.1109, +0.1309]$	3.1×10^{-115}
EasIFA	0.3250	$+0.3989$	$[+0.3884, +0.4098]$	$< 10^{-300}$
AEGAN	0.2418	$+0.4821$	$[+0.4736, +0.4908]$	$< 10^{-300}$
MMseqs2 transfer	0.1982	$+0.5257$	$[+0.5153, +0.5361]$	$< 10^{-300}$
Foldseek transfer	0.2329	$+0.4910$	$[+0.4791, +0.5020]$	$< 10^{-300}$
EnzyMM	0.1158	$+0.6081$	$[+0.5993, +0.6169]$	$< 10^{-300}$

Figure 4 compares EnzyFuse with the five-model Squidly ensemble enzyme by enzyme.

Table 6: Values behind Figure 3: per-protein AUPRC by similarity h to catalytic-labeled training proteins, with paired comparisons against the two strongest baselines.

h	n	EnzyFuse	Baseline	AUPRC	Δ	95% CI	p
none (h = 0)	3,322	0.6688	Squidly ×5	0.6050	+0.0639	[+0.0526, +0.0750]	3.3×10^{-36}
none (h = 0)	3,322	0.6688	SCREEN	0.5527	+0.1161	[+0.1062, +0.1265]	9.2×10^{-111}
(0, 0.3)	793	0.6879	Squidly ×5	0.6398	+0.0481	[+0.0322, +0.0653]	2.7×10^{-6}
(0, 0.3)	793	0.6879	SCREEN	0.6075	+0.0804	[+0.0652, +0.0964]	5.2×10^{-29}
[0.3, 0.4)	972	0.8680	Squidly ×5	0.8613	+0.0067	[−0.0034, +0.0168]	2.2×10^{-1}
[0.3, 0.4)	972	0.8680	SCREEN	0.8660	+0.0020	[−0.0087, +0.0123]	2.3×10^{-1}
[0.4, 0.6)	1,474	0.8896	Squidly ×5	0.8041	+0.0856	[+0.0714, +0.0998]	2.7×10^{-27}
[0.4, 0.6)	1,474	0.8896	SCREEN	0.7992	+0.0904	[+0.0740, +0.1064]	1.8×10^{-22}
[0.6, 1.0]	243	0.9340	Squidly ×5	0.9519	−0.0179	[−0.0406, +0.0042]	1.9×10^{-3}
[0.6, 1.0]	243	0.9340	SCREEN	0.9696	−0.0356	[−0.0542, −0.0181]	1.2×10^{-4}

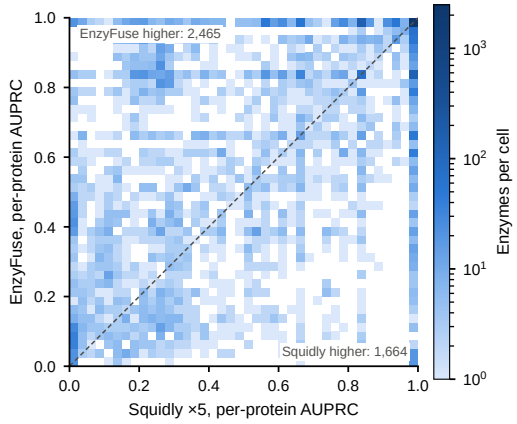


Figure 4: Per-protein AUPRC of EnzyFuse against the five-model Squidly ensemble. Color gives the number of enzymes per cell on a logarithmic scale; the corner labels count enzymes on which each method scores higher.

E Extended design analysis

Table 7 reports the available design comparisons on the 3,322 test proteins with no detected neighbor among catalytic-labeled training proteins ($h = 0$). Removing partner retrieval or the interaction branch lowers mean AUPRC by 0.0040 and 0.0087, respectively. These aggregate comparisons show the same direction as the full-test results, but do not establish statistical significance.

Table 7: Design comparisons on the 3,322 test proteins with no detected neighbor among catalytic-labeled training proteins ($h = 0$). Available aggregate AUPRC values are reported; paired confidence intervals are not available for these comparisons. †Also removes partner retrieval.

Variant	AUPRC	Δ	95% CI
EnzyFuse	0.6688	—	—
Objective			
w/o partner retrieval	0.6648	−0.0040	—
Interaction			
w/o interaction branch†	0.6601	−0.0087	—

Table 8 reports secondary ablations of the ranking objective, partner-retrieval radius, training-time fusion, MSA representation and training-data scope.

Table 8: Secondary ablations of EnzyFuse, each trained with three seeds and scored on all test proteins. Δ is relative to EnzyFuse, with a bootstrap 95% CI.

Variant	AUPRC	Δ	95% CI
EnzyFuse	$0.7568_{\pm 0.0035}$	—	—
Objective			
w/o listwise ranking term	$0.7068_{\pm 0.0092}$	-0.0500	$[-0.0564, -0.0438]$
partner radius 12 Å instead of 15 Å	$0.7373_{\pm 0.0057}$	-0.0195	$[-0.0246, -0.0145]$
Fusion			
equal weights instead of the gate in training	$0.7418_{\pm 0.0048}$	-0.0150	$[-0.0196, -0.0106]$
Input representation			
MSA as a profile (PSSM) instead of MSA Transformer	$0.6885_{\pm 0.0115}$	-0.0683	$[-0.0752, -0.0616]$
Training data			
catalytic proteins only	$0.7215_{\pm 0.0074}$	-0.0353	$[-0.0412, -0.0296]$

Figure 5 reports the AUPRC of each branch head of the final model.

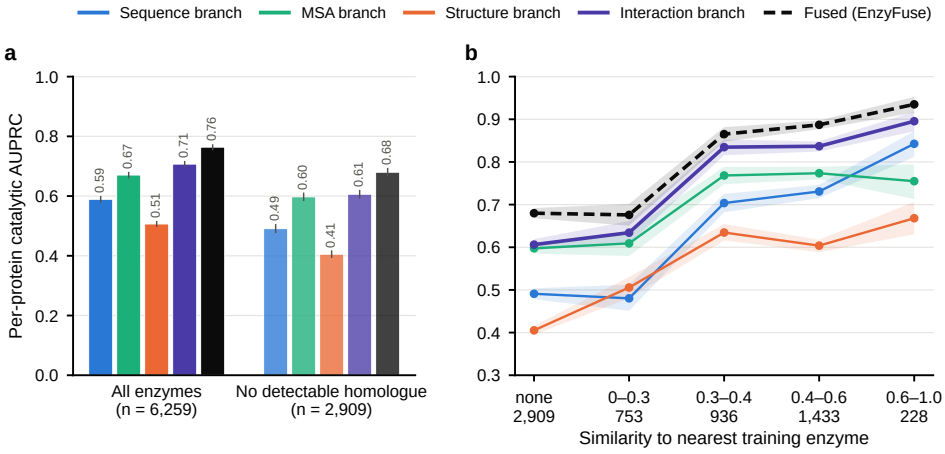


Figure 5: Per-protein catalytic AUPRC of each branch head and of the averaged prediction of EnzyFuse, on the 6,259 test enzymes with all three views. (a) All enzymes and the 2,909 with no detected neighbor among catalytic-labeled training proteins ($h = 0$). (b) By similarity to that reference set; bin sizes are given below the axis. These are heads of a jointly trained model, not separately trained single-view models.

Figure 6 counts which branches recover each catalytic site. For each enzyme, a branch recovers a catalytic residue if the residue is among the n_p highest-scoring residues of that branch, where n_p is the number of annotated catalytic residues of the enzyme, in at least two of the three training runs. The single branches recover 40–64% of the 10,178 sites and the averaged prediction 69%; 313 sites are recovered only by the interaction branch. Because this diagnostic uses the true number of catalytic residues, it describes branch agreement rather than a deployable selection rule.

F Robustness and Scope

Masking views at inference. To separate the effect of a missing view from the difficulty of the proteins that lack it, we mask views at inference on the 6,259 test enzymes that have all three (Figure 7). Masking the structure lowers AUPRC from 0.764 to 0.740, masking the MSA lowers it to 0.664, and masking both, which leaves the sequence branch alone, lowers it to 0.589. On the 2,909 of these enzymes with no detected neighbor among catalytic-labeled training proteins ($h = 0$),

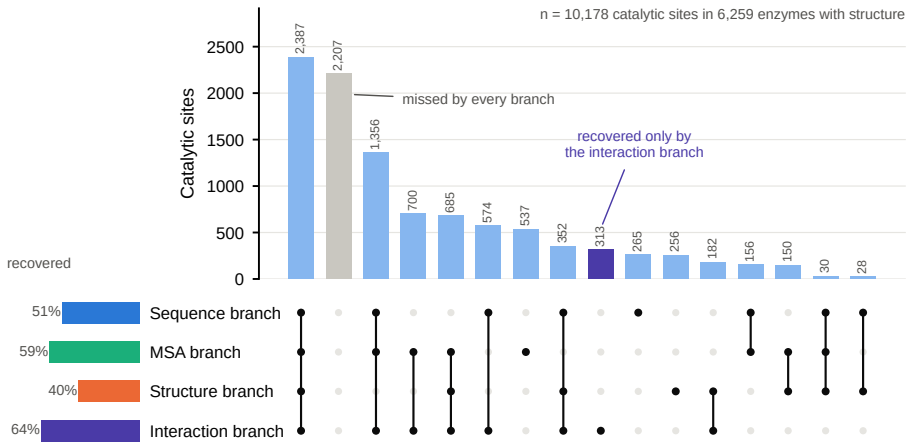


Figure 6: Catalytic sites recovered by each combination of branch heads, for the 10,178 catalytic sites of the 6,259 test enzymes with all three views. Left bars give the fraction of sites recovered by each branch.

the corresponding values are 0.680, 0.669, 0.551 and 0.491. With the structure masked, EnzyFuse remains above the five-model Squidly ensemble and SCREEN on both sets.

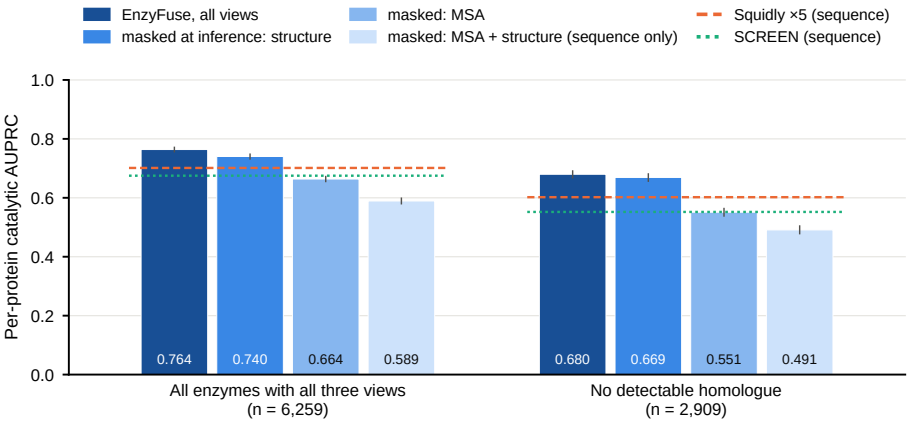


Figure 7: Catalytic AUPRC when views are masked at inference, on test enzymes with all three views and on the subset with no detected neighbor among catalytic-labeled training proteins ($h = 0$). Horizontal lines mark the Squidly ensemble and SCREEN on the same proteins. Error bars are standard deviations over three training runs.

Chain length. Figure 8 reports per-protein AUPRC as a function of chain length.

G CATALYTIC PARTNER RETRIEVAL

Figure 9 compares the final model with the same model trained without partner retrieval, on the 6,259 test enzymes with a structure and at least one catalytic residue with coordinates. Predicted probabilities are averaged over the three runs and converted to rank percentiles within each enzyme (one denotes the highest score); each distance bin is averaged within enzymes and then across enzymes, and all bins have 95% bootstrap confidence intervals that exclude zero. With partner retrieval, residues not annotated as catalytic within the 15 Å retrieval radius of a catalytic residue move

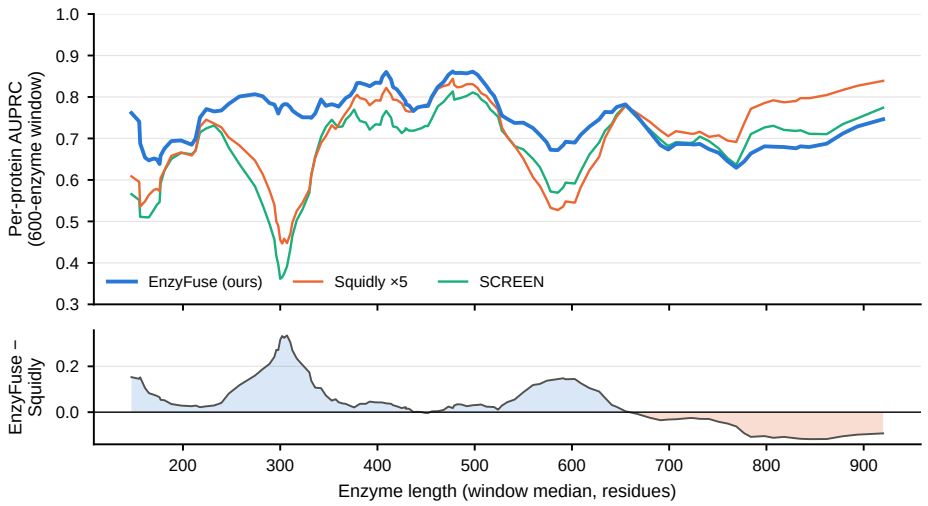


Figure 8: Per-protein catalytic AUPRC over windows of 600 enzymes ordered by chain length (top), and the difference between EnzyFuse and the five-model Squidly ensemble (bottom).

down in the within-protein ranking, most strongly at 6–8 Å, whereas residues farther than 20 Å and catalytic residues themselves move slightly up (Figure 9a). Among the five highest-ranked residues without a catalytic annotation in each enzyme, the fraction lying within 12 Å of a catalytic site decreases from 59.2% to 55.9% (Wilcoxon $p = 8.7 \times 10^{-42}$; Figure 9b). This pattern is consistent with improved local discrimination between annotated catalytic residues and their spatial neighbors.

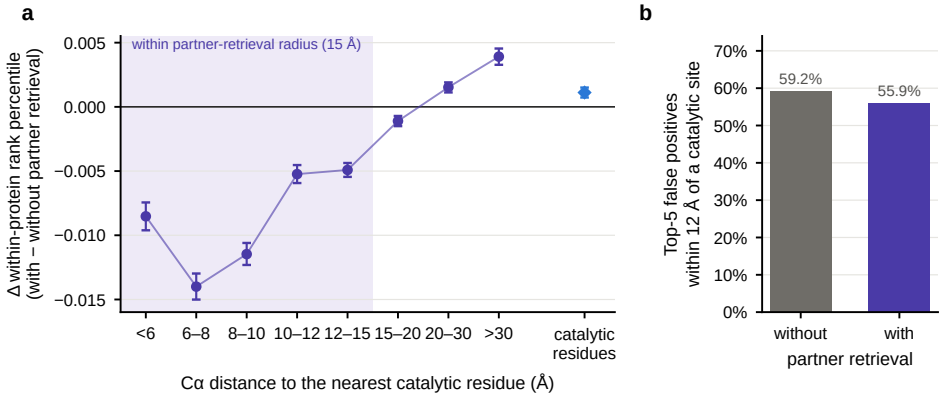


Figure 9: Effect of partner retrieval on the ranking of residues within each enzyme. (a) Change in the within-protein rank percentile of residues without catalytic annotations, grouped by C_α distance to the nearest catalytic residue, and of catalytic residues; the shaded region marks the 15 Å retrieval radius. (b) Fraction of the five highest-scoring residues without catalytic annotations that lie within 12 Å of a catalytic site (labeled top-5 false positives in the plot).

H BINDING AND OTHER FUNCTIONAL SITES

EnzyFuse predicts binding and other sites with the same heads, trained jointly with the catalytic class. Each class is evaluated on the test proteins with at least one positive and one negative residue of that class. Binding sites reach a per-protein AUPRC of 0.682 ± 0.003 on 12,254 proteins, against 0.041 for uniformly random scores, and other sites reach 0.314 ± 0.016 on 2,607 proteins, against 0.027. The random-score results average five repeats on the same class-specific cohorts. These auxiliary outputs are not compared with baselines here; dedicated predictors target ligand-binding

pockets (Krivák & Hoksza, 2018) or protein–protein binding sites (Tubiana et al., 2022), which differ from these residue annotations.