A STEREOCHEMISTRY-AWARE VECTOR–SCALAR FRAMEWORK FOR ENZYME–SUBSTRATE SPECIFICITY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Enzyme specificity determines which substrates are selectively recognized and catalyzed, but existing computational methods often provide limited modeling of pair-specific three-dimensional residue–substrate interactions. Here, we develop StereoVSE, a stereochemistry-aware Vector–Scalar framework for enzyme–substrate specificity prediction and candidate residue prioritization. StereoVSE integrates protein sequence representations, molecular features, and enzyme–substrate structural configurations through a Vector–Scalar graph neural network and Bidirectional Residue–Atom Interaction Alignment, enabling substrate-conditioned residue–atom interaction learning. Experiments on halogenase and glycosyltransferase benchmarks demonstrate improved specificity prediction performance over existing methods. Furthermore, StereoVSE generates substrate-conditioned residue–atom attribution signals that complement sequence conservation analysis and enable candidate residue prioritization.

1 INTRODUCTION

Enzymes are efficient and selective biological catalysts with broad applications in pharmaceutical synthesis, green chemistry, metabolic engineering, and industrial biomanufacturing Bornscheuer et al. (2012); Sheldon & Woodley (2018); Wu et al. (2021); Bell et al. (2021); Chen & Arnold (2020); Leveson-Gower et al. (2019). Advances in directed evolution, protein engineering, and machine-learning-guided design have substantially expanded the accessible scope of biocatalytic transformations Zeymer & Hilvert (2018); Yang et al. (2019); Probst et al. (2022); Landwehr et al. (2025); Mazurenko et al. (2020); Romero et al. (2013). Nevertheless, identifying enzymes that can catalyze a desired substrate still requires extensive experimental screening.

Traditional computational approaches to enzyme–substrate specificity prediction have relied on sequence homology and conserved motifs, molecular or reaction similarity, structure-based docking, and family-specific statistical or machine-learning models Hermann et al. (2007); Carbonell et al. (2017); Hadadi et al. (2019); Mou et al. (2021). Nevertheless, many models encode enzymes and substrates independently and assess their compatibility only after global pooling, feature concatenation, or embedding alignment Öztürk et al. (2018); Feng et al. (2018); Karimi et al. (2019). In recent years, self-supervised molecular encoders, pretrained protein language models, and equivariant geometric networks have enabled task-relevant representations to be learned directly from large-scale molecular, sequence, and structural data Hu et al. (2020); Rong et al. (2020); Rives et al. (2021); Elnaggar et al. (2021); Satorras et al. (2021); Lin et al. (2023).

Although structure-aware methods partially address this issue, directional geometric information may still be compressed during subsequent pooling. Meanwhile, residue–atom attention is often used as an intermediate weighting signal rather than retained as a persistent interaction representation Schütt et al. (2021); Jing et al. (2021). Explicit atom–residue interaction modeling can provide finer-grained pair-specific informationLi et al. (2020); Karimi et al. (2020); Jain & Wallace (2019); Serrano & Smith (2019). However, many existing approaches either use contact information only as auxiliary features or compress interaction patterns into global representations, limiting their ability to preserve substrate-conditioned residue–atom associations. Recent enzyme–substrate models have progressively evolved from family-wide sequence–molecule representation learning to contrastive

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.

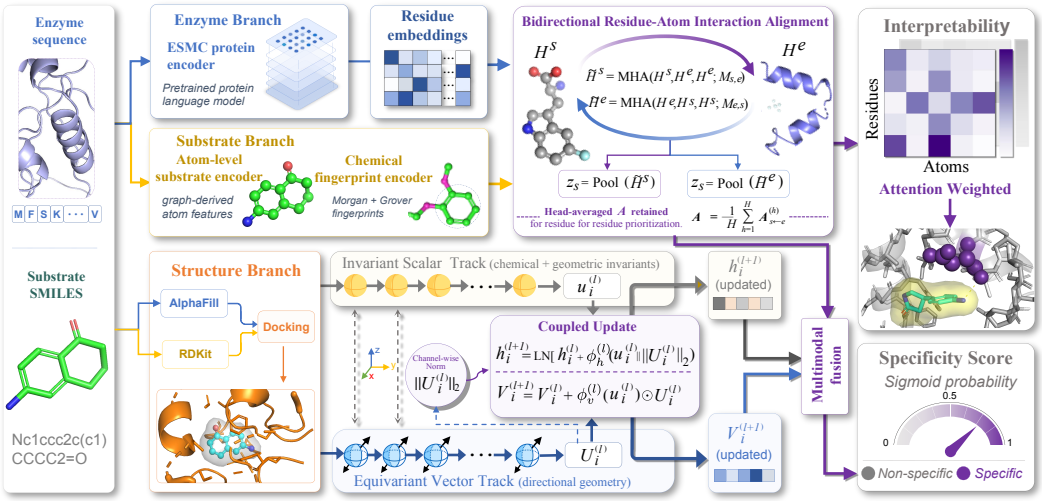


Figure 1: Overview of the StereoVSE framework. StereoVSE integrates enzyme sequence representations, substrate molecular features, and docking-derived three-dimensional structures. The fused multimodal representation is used for enzyme–substrate specificity prediction, and the learned interaction signals provide residue–atom interaction attribution signals for downstream residue prioritization.

enzyme–reaction alignment and structure-aware residue–atom interaction modeling Goldman et al. (2022); Kroll et al. (2023); Mikhael et al. (2024); Cui et al. (2025). These limitations motivate a unified framework that preserves residue-level protein semantics, atom-level substrate chemistry, directional three-dimensional geometry, and explicit residue–atom interactions throughout representation learning. Importantly, rather than treating attention weights alone as explanations, we evaluate BRIA-derived interaction signals through residue perturbation analysis and substrate-conditioned prediction changes.

To address these limitations, we formulate enzyme–substrate specificity prediction as a substrate-conditioned residue–atom interaction learning problem and propose **StereoVSE**, a stereochemistry-aware Vector–Scalar framework designed to preserve these interactions during multimodal representation learning. Here, stereochemistry-aware refers to modeling the three-dimensional stereochemical environment formed by enzyme residues and substrate atoms, including their spatial arrangement and directional interactions, rather than relying solely on sequence information or molecular connectivity. StereoVSE jointly models enzyme sequence semantics, substrate chemical structures, and pair-specific binding geometries while explicitly maintaining residue–atom correspondence through interaction-aware representation learning. The sequence branch uses the pretrained protein language model ESMC to generate contextualized residue representations Candido et al. (2026), while the substrate branch combines atom-level features with Morgan fingerprints Rogers & Hahn (2010) and GROVER-derived molecular representations Rong et al. (2020). The structural branch uses the precomputed enzyme–substrate configurations associated with the benchmark pairs Cui et al. (2025). Each configuration is converted into a local pocket–substrate graph and processed by a **Vector–Scalar Graph Neural Network**, which couples invariant chemical features with directional geometric states throughout message passing. We further introduce **Bidirectional Residue–Atom Interaction Alignment** (**BRIA**), which performs enzyme-conditioned atom updating and substrate-conditioned residue updating. Finally, the aligned enzyme and substrate representations are integrated with global chemical and docking-derived structural features for specificity prediction, while the learned interaction matrices provide residue-level signals for identifying candidate specificity-determining residues.

Our main contributions are summarized as follows:

- We propose **StereoVSE**, a unified framework that formulates enzyme–substrate specificity prediction through substrate-conditioned residue–atom interaction learning, integrating protein semantics, substrate chemistry, and binding geometry.

- We develop a vector–scalar geometric encoder that preserves directional binding information through equivariant residue–atom neighborhood aggregation.
- We introduce **BRIA** to explicitly maintain sample-specific substrate atom–enzyme residue alignments and derive substrate-conditioned attribution signals for candidate residue prioritization.

2 Related Work

Enzyme–substrate specificity prediction. Existing methods have progressively extended enzyme–substrate modeling from family-specific activity prediction to large-scale enzyme screening. Family-wide models jointly encode enzyme sequences and substrate structures Goldman et al. (2022), while ESP learns general protein and molecular representations for predicting small-molecule substrates across enzyme families Kroll et al. (2023). CLIPZyme formulates enzyme identification as reaction-conditioned retrieval through contrastive representation alignment Mikhael et al. (2024). More recently, EZSpecificity incorporates three-dimensional enzyme structures and substrate features using an equivariant graph architecture and an interaction module Cui et al. (2025). Although these methods improve compatibility prediction, most focus primarily on global enzyme–substrate representations, leaving residue–atom alignment patterns and residue-level candidate prioritization less explored.

Biomolecular representation and geometric learning. Protein language models such as ESM provide contextualized residue embeddings that capture evolutionary and functional information Lin et al. (2023). For substrates, Morgan fingerprints describe local chemical substructures, whereas GROVER learns complementary atom-level and global molecular representations through self-supervised graph pretraining Rogers & Hahn (2010). Meanwhile, geometric networks such as EGNN, PaiNN, and ViSNet demonstrate the benefits of symmetry-aware three-dimensional modeling and coupled scalar–vector message passing Satorras et al. (2021); Schütt et al. (2021); Wang et al. (2024); Jing et al. (2021); Gasteiger et al. (2020); Fuchs et al. (2020). Building on these advances, StereoVSE combines pretrained sequence and chemical representations with a Vector–Scalar geometric encoder operating on a docking-derived local pocket–substrate graph. It further introduces Bidirectional Residue–Atom Interaction Alignment (BRIA) to construct pair-specific interactions between substrate atoms and enzyme residues, supporting both specificity prediction and candidate specificity-determining residue identification.

3 Methods

3.1 Overview of StereoVSE

To address existing limitations, we combine multimodal representation learning, bidirectional residue–atom interaction alignment, and geometry-aware structural encoding. The multimodal module captures complementary enzyme and substrate features, while BRIA explicitly models residue–atom interactions and provides interpretable alignment. Meanwhile, the Vector–Scalar encoder integrates chemical semantics with three-dimensional geometric information through invariant and equivariant representations. Finally, these representations are fused for enzyme–substrate specificity prediction, enabling accurate and interpretable modeling of biochemical interactions. The model workflow is shown in Figure 1.

3.2 Multimodal Representation Learning

For an enzyme containing L_e residues, a pretrained protein language model produces contextualized residue embeddings. These embeddings are projected to a shared hidden dimension d:

$$\mathbf{H}^e = \mathrm{Proj}_e\left(\mathrm{ESMC}(S_e)\right) \in \mathbb{R}^{L_e \times d}. \tag{1}$$

For a substrate containing L_s atoms, an atom-level molecular encoder produces

$$\mathbf{H}^s = \mathrm{Proj}_s\left(\mathrm{MolEnc}(M_s)\right) \in \mathbb{R}^{L_s \times d}. \tag{2}$$

Table 1: Four-fold pair-wise validation results on both benchmarks are reported as mean ± standard deviation. Best and second-best means are shown in bold and underlined; Up Ratio denotes the relative improvement over the strongest baseline.

	Halogenase			Glycosyltransferase		
Model	**AUC**	**AUPR**	**MCC**	**AUC**	**AUPR**	**MCC**
CPI (*Bioinformatics*, 2022)	0.7960 ± 0.0022	0.5694 ± 0.0020	0.4665 ± 0.0132	0.9105 ± 0.0087	0.6633 ± 0.0230	0.6002 ± 0.0431
ESP (*Nat. Commun.*, 2023)	0.8836 ± 0.0085	0.6885 ± 0.0112	0.5423 ± 0.0125	0.9034 ± 0.0116	$\underline{0.7093 \pm 0.0041}$	0.5501 ± 0.0410
CLIPZyme (*ICML*, 2024)	0.8499 ± 0.0132	0.6407 ± 0.0235	0.4991 ± 0.0133	$\underline{0.9149 \pm 0.0100}$	0.7089 ± 0.0286	0.6081 ± 0.0287
CATNIP (*Nature*, 2025)	0.9012 ± 0.0058	0.7409 ± 0.0036	0.5527 ± 0.0155	0.9092 ± 0.0119	0.7072 ± 0.0241	0.5340 ± 0.0257
EZSpecificity (*Nature*, 2025)	$\underline{0.9592 \pm 0.0025}$	$\underline{0.8664 \pm 0.0125}$	$\underline{0.7374 \pm 0.0239}$	0.9147 ± 0.0055	0.6805 ± 0.0113	$\underline{0.6198 \pm 0.0121}$
RAMMESI (*arXiv*, 2026)	0.8898 ± 0.0233	0.6753 ± 0.0480	0.5600 ± 0.0414	0.8967 ± 0.0058	0.6157 ± 0.0214	0.5515 ± 0.0369
StereoVSE	$\mathbf{0.9728 \pm 0.0015}$	$\mathbf{0.9112 \pm 0.0030}$	$\mathbf{0.7666 \pm 0.0141}$	$\mathbf{0.9364 \pm 0.0018}$	$\mathbf{0.7338 \pm 0.0156}$	$\mathbf{0.6813 \pm 0.0137}$
Up Ratio	↑ 1.42%	↑ 5.17%	↑ 3.96%	↑ 2.35%	↑ 3.45%	↑ 9.92%

Here, $\mathbf{H}^s$ preserves atom-level information required by the residue–atom interaction module. To complement these local atom features with global chemical context, we combine their pooled representation with Morgan fingerprints and pretrained GROVER features:

$$\mathbf{z}_{\text{chem}} = \phi_{\text{chem}}\left(\text{Pool}(\mathbf{H}^s) \parallel \text{Proj}_M(\mathbf{m}_{\text{Morgan}}) \parallel \text{Proj}_G(\mathbf{m}_{\text{GROVER}})\right). \tag{3}$$

The residue embeddings encode the sequence context of each enzyme position, whereas the substrate representations capture local atomic environments, molecular substructures, and global chemical semantics.

3.3 Three-Dimensional Pocket–Substrate Graph

The structural branch operates on the retained three-dimensional configuration associated with each benchmark pair. The pair-specific enzyme–substrate configurations adopted in this study were precomputed. Diverse enzyme–substrate binding poses generated by AutoDock-GPU were filtered based on distance metrics and affinity scoring systems to select the poses most favorable for catalysis Cui et al. (2025).

For an enzyme–substrate pair (e, s), let

$$\mathcal{C}_{e,s} = (\mathcal{V}_{\text{rec}}, \mathcal{V}_{\text{lig}}, \mathbf{X}_{\text{rec}}, \mathbf{X}_{\text{lig}}) \tag{4}$$

denote the retained configuration, where $\mathcal{V}_{\text{rec}}$ and $\mathcal{V}_{\text{lig}}$ are the receptor and substrate atoms, and $\mathbf{X}_{\text{rec}}$ and $\mathbf{X}_{\text{lig}}$ are their corresponding Cartesian coordinates.

To focus geometric learning on the local binding environment, a receptor atom is retained if its minimum Euclidean distance to any substrate atom does not exceed 10 Å. We use a 10 Å pocket radius as a fixed preprocessing hyperparameter to balance local-environment coverage and graph size. This radius retains both direct-contact atoms and the surrounding second-shell environment, while excluding distal protein regions:

$$\mathcal{V}_{\text{pocket}} = \left\{ i \in \mathcal{V}_{\text{rec}} \;\middle|\; \min_{j \in \mathcal{V}_{\text{lig}}} \left\| \mathbf{x}_i^{\text{rec}} - \mathbf{x}_j^{\text{lig}} \right\|_2 \leq 10 \text{ Å} \right\}. \tag{5}$$

The node set of the local graph is

$$\mathcal{V}_{e,s} = \mathcal{V}_{\text{pocket}} \cup \mathcal{V}_{\text{lig}}, \qquad \mathbf{X}_{e,s} = [\mathbf{X}_{\text{pocket}}; \mathbf{X}_{\text{lig}}]. \tag{6}$$

Edges are constructed dynamically using a k-nearest-neighbor rule in the retained coordinate frame:

$$(i, j) \in \mathcal{E}_{e,s} \quad \Longleftrightarrow \quad j \in \text{kNN}(i). \tag{7}$$

For each edge, we define

$$\mathbf{r}_{ij} = \mathbf{x}_j - \mathbf{x}_i, \qquad d_{ij} = \|\mathbf{r}_{ij}\|_2, \qquad \widehat{\mathbf{r}}_{ij} = \frac{\mathbf{r}_{ij}}{d_{ij} + \epsilon}. \tag{8}$$

Its scalar edge representation is

$$\mathbf{e}_{ij} = \phi_{\text{edge}}\left(\text{RBF}(d_{ij}) \parallel \mathbf{a}_i \parallel \mathbf{a}_j \parallel \text{Emb}(\tau_i) \parallel \text{Emb}(\tau_j)\right), \tag{9}$$

where $\mathbf{a}_i$ and $\mathbf{a}_j$ denote atom attributes, and $\tau_i, \tau_j \in \{\text{pocket}, \text{substrate}\}$ distinguish the two node types. The resulting structural graph is

$$G_{e,s}^{3D} = \left(\mathcal{V}_{e,s}, \mathcal{E}_{e,s}, \mathbf{X}_{e,s}, \{\mathbf{e}_{ij}\}_{(i,j)\in\mathcal{E}_{e,s}}\right). \tag{10}$$

This representation preserves both local chemical identities and the relative geometry between pocket and substrate atoms. Additional details on coordinate extraction and graph serialization are provided in Appendix F.2.

3.4 Vector–Scalar Geometric Encoding

Conventional molecular graph encoders frequently represent geometry only through pairwise distances, which removes explicit directional information. To model the orientation-dependent organization of the binding environment, our structural encoder maintains two coupled states for each node i: an invariant scalar state $\mathbf{h}_i^{(l)} \in \mathbb{R}^d$ and an equivariant vector state $\mathbf{V}_i^{(l)} \in \mathbb{R}^{d_v\times 3}$.

The initial scalar state is obtained from atom attributes, while the vector state is initialized through local directional aggregation:

$$\mathbf{h}_i^{(0)} = \phi_{\text{atom}}(\mathbf{a}_i), \qquad \mathbf{V}_i^{(0)} = \sum_{j\in\mathcal{N}(i)} \phi_{\text{init}}(\mathbf{e}_{ij}) \odot \widehat{\mathbf{r}}_{ij}. \tag{11}$$

The multiplication in the second equation is broadcast over the d_v vector channels.

At layer l, rotation-invariant descriptors are extracted from the vector states:

$$\mathcal{I}_{ij}^{(l)} = \left[\left\|\mathbf{V}_i^{(l)}\right\|_2 \,\|\, \left\|\mathbf{V}_j^{(l)}\right\|_2 \,\|\, \left\langle \mathbf{V}_i^{(l)}, \mathbf{V}_j^{(l)} \right\rangle\right]. \tag{12}$$

The norm and inner product are evaluated channel-wise. The scalar edge message then combines node states, chemical edge information, and the invariant geometric descriptors:

$$\mathbf{m}_{ij}^{(l)} = \phi_m^{(l)}\left(\mathbf{h}_i^{(l)} \,\|\, \mathbf{h}_j^{(l)} \,\|\, \mathbf{e}_{ij} \,\|\, \mathcal{I}_{ij}^{(l)}\right). \tag{13}$$

A learned attention coefficient determines the relative importance of each neighbor:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(\phi_\alpha^{(l)}(\mathbf{m}_{ij}^{(l)})\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\phi_\alpha^{(l)}(\mathbf{m}_{ik}^{(l)})\right)}. \tag{14}$$

Two separate gates regulate directional propagation from the neighboring vector state and from the relative interatomic direction:

$$\mathbf{g}_{ij}^{v} = \sigma\left(\phi_{g_v}^{(l)}(\mathbf{m}_{ij}^{(l)})\right), \qquad \mathbf{g}_{ij}^{r} = \sigma\left(\phi_{g_r}^{(l)}(\mathbf{m}_{ij}^{(l)})\right). \tag{15}$$

The scalar and vector messages are aggregated as

$$\begin{aligned} \mathbf{u}_i^{(l)} &= \sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l)} \phi_s^{(l)}\left(\mathbf{m}_{ij}^{(l)}\right), \\ \mathbf{U}_i^{(l)} &= \sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l)} \left(\mathbf{g}_{ij}^{v} \odot \mathbf{V}_j^{(l)} + \mathbf{g}_{ij}^{r} \odot \widehat{\mathbf{r}}_{ij}\right). \end{aligned} \tag{16}$$

The two streams are subsequently coupled:

$$\mathbf{h}_i^{(l+1)} = \text{LN}\left[\mathbf{h}_i^{(l)} + \phi_h^{(l)}\left(\mathbf{u}_i^{(l)} \,\|\, \left\|\mathbf{U}_i^{(l)}\right\|_2\right)\right], \quad \mathbf{V}_i^{(l+1)} = \mathbf{V}_i^{(l)} + \phi_v^{(l)}\left(\mathbf{u}_i^{(l)}\right) \odot \mathbf{U}_i^{(l)}. \tag{17}$$

The vector norm injects invariant geometric evidence into the scalar stream, whereas the scalar stream controls the propagation of directional features.

Under a global rigid transformation $\mathbf{x}_i' = \mathbf{R}\mathbf{x}_i + \mathbf{t}$ with $\mathbf{R} \in \text{SO}(3)$, the scalar and vector states satisfy

$$\mathbf{h}_i'^{(l)} = \mathbf{h}_i^{(l)}, \qquad \mathbf{V}_i'^{(l)} = \mathbf{V}_i^{(l)}\mathbf{R}^\top. \tag{18}$$

Thus, the scalar representations and final prediction remain invariant to global rotations and translations, while the vector states rotate equivariantly with the coordinates. After L layers, attention pooling over the final scalar states produces the structural representation

$$\mathbf{z}_{\text{3D}} = \text{AttnPool}\left(\left\{\mathbf{h}_i^{(L)}\right\}_{i\in\mathcal{V}_{e,s}}\right). \tag{19}$$

Further implementation details are provided in Appendix F.3.

3.5 Bidirectional Residue–Atom Interaction Alignment

Independent enzyme and substrate encoders do not explicitly model which residues interact with particular substrate atoms. We therefore introduce Bidirectional Residue–Atom Interaction Alignment (**BRIA**), which performs cross-attention in both directions:

$$\begin{aligned}\widetilde{\mathbf{H}}^s &= \mathrm{MHA}\left(\mathbf{H}^s, \mathbf{H}^e, \mathbf{H}^e; \mathbf{M}_{s,e}\right), \\ \widetilde{\mathbf{H}}^e &= \mathrm{MHA}\left(\mathbf{H}^e, \mathbf{H}^s, \mathbf{H}^s; \mathbf{M}_{e,s}\right),\end{aligned} \tag{20}$$

where $\mathbf{M}_{s,e}$ and $\mathbf{M}_{e,s}$ mask padded atoms and residues. The first direction produces enzyme-conditioned atom features, whereas the second produces substrate-conditioned residue features. Their pooled representations are

$$\mathbf{z}_s = \mathrm{Pool}\left(\widetilde{\mathbf{H}}^s\right), \qquad \mathbf{z}_e = \mathrm{Pool}\left(\widetilde{\mathbf{H}}^e\right). \tag{21}$$

In addition to interaction-aware fusion, BRIA provides residue-level prioritization. For attention head h, let $\mathbf{A}^{(h)}_{s\leftarrow e} \in \mathbb{R}^{L_s \times L_e}$ denote the atom-to-residue attention matrix. We retain its head-averaged form

$$\mathbf{A} = \frac{1}{H}\sum_{h=1}^{H} \mathbf{A}^{(h)}_{s\leftarrow e} \in \mathbb{R}^{L_s \times L_e}, \tag{22}$$

which captures pair-specific associations between individual substrate atoms and enzyme residues. This matrix is used for downstream ranking of candidate specificity-determining residues. Detailed multi-head projections and residue-scoring procedures are provided in Appendix F.4 and Appendix F.6.

3.6 Multimodal Fusion and Optimization

The interaction-aware enzyme and substrate representations are fused with the global chemical representation and the three-dimensional structural representation:

$$\mathbf{z}_{\mathrm{final}} = \phi_{\mathrm{fuse}}\left(\mathbf{z}_e \parallel \mathbf{z}_s \parallel \mathbf{z}_{\mathrm{chem}} \parallel \mathbf{z}_{\mathrm{3D}}\right). \tag{23}$$

The predicted enzyme–substrate compatibility probability is

$$\widehat{y} = \sigma\left(f_{\mathrm{pred}}\left(\mathbf{z}_{\mathrm{final}}\right)\right). \tag{24}$$

The model is optimized using class-weighted binary cross-entropy with ℓ_2 regularization:

$$\mathcal{L} = \mathcal{L}_{\mathrm{WBCE}}(y, \hat{y}) + \lambda_{\mathrm{reg}}\|\Theta\|_2^2.$$

4 Experiments

4.1 Datasets

We evaluate our method on two family-specific enzyme–substrate benchmark datasets: the halogenase dataset and the glycosyltransferase dataset Cui et al. (2025); Yang et al. (2018). Each sample is an enzyme–substrate pair (e, s) with a binary label $y_{e,s} \in \{0, 1\}$, where a positive label indicates that enzyme e catalyzes substrate s.

These two datasets represent complementary substrate-recognition scenarios. Halogenases are associated with catalyst-controlled substrate specificity in biocatalytic halogenation, whereas glycosyltransferases involve diverse donor–acceptor recognition patterns Agarwal et al. (2017); Dong et al. (2005); Lairson et al. (2008); Breton et al. (2006). The glycosyltransferase benchmark is substantially larger and more imbalanced than the halogenase benchmark, providing a challenging setting for evaluating positive-pair retrieval. Detailed dataset descriptions and statistics are provided in Appendix A.1.

4.2 Baselines and Evaluation Metrics

We compare StereoVSE with six representative baselines: CPI Yaseen et al. (2022), ESP Kroll et al. (2023), CLIPZyme Mikhael et al. (2024), CATNIP Paton et al. (2025), EZSpecificity Cui et al. (2025) and RAMMESI Liu et al. (2026). Detailed descriptions of these methods are provided in Appendix A.2.

We evaluate overall performance using AUC, AUPR, and Matthews correlation coefficient (MCC). AUC measures ranking performance, AUPR is more informative under class imbalance, and MCC provides a balanced assessment of binary classification quality Saito & Rehmsmeier (2015); Davis & Goadrich (2006). For the ablation study, we additionally report F1 score and accuracy (ACC) to provide a more comprehensive comparison of model variants. Higher values indicate better performance for all metrics.

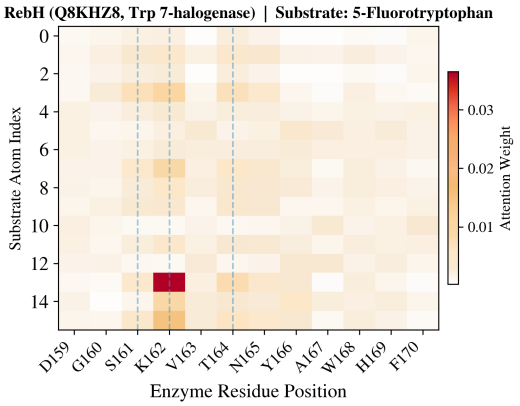


Figure 2: Each entry denotes a BRIA interaction weight between a substrate atom and a nearby enzyme residue.

4.3 Overall Performance Comparison

Under the random-split setting, four-fold cross-validation was performed on the halogenase dataset and the glycosyltransferase dataset at the enzyme–substrate pair level, and the mean performance across the four folds was reported. No identical enzyme–substrate pair was shared across the training, validation, and test subsets within a fold.

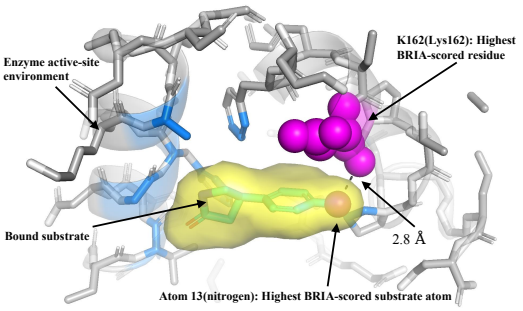


Figure 3: Representative RebH–5-fluorotryptophan three-dimensional configuration.

As shown in Table 1, Under the evaluated benchmark settings, StereoVSE achieves higher AUC, AUPR, and MCC than the compared baselines on both datasets. On the larger and more imbalanced glycosyltransferase dataset, the StereoVSE model achieves substantial improvements in the AUPR and MCC metrics. These results indicate that StereoVSE effectively distinguishes positive enzyme–substrate pairs under the class-imbalanced benchmark settings. Additional visualization analysis is available in Appendix C.

We further evaluate StereoVSE under three cold-start settings on the halogenase benchmark: enzyme-cold, substrate-cold, and enzyme–substrate pair-cold, which respectively hold out enzyme identities, substrate identities, and enzyme–substrate associations. StereoVSE consistently outperforms the evaluated baselines across all three settings (Table 2), showing that its representations remain informative under these distribution shifts. Full protocols are provided in Appendix E.

4.4 Ablation Study

We conduct ablation experiments on both the halogenase and glycosyltransferase datasets using the same data splits and training protocols as the main evaluation. Four variants are considered: (1) ESMC → ESM-2 replaces ESMC with ESM-2 Lin et al. (2023) as the pretrained protein encoder; (2) w/o 3D removes the structural branch constructed from the precomputed enzyme–substrate configurations; (3) w/o BRIA removes the bidirectional residue–atom interaction alignment module; and (4) w/o V–S removes the coupled vector–scalar geometric encoder.

Table 2: Cold-start evaluation on the halogenase benchmark under enzyme-cold, substrate-cold, and enzyme–substrate pair-cold settings. Best results are bolded.

	Enzyme-cold			Substrate-cold			Enzyme–substrate pair-cold		
Model	**AUC↑**	**AUPR↑**	**MCC↑**	**AUC↑**	**AUPR↑**	**MCC↑**	**AUC↑**	**AUPR↑**	**MCC↑**
CPI	0.688	0.408	0.254	0.683	0.432	0.291	0.676	0.392	0.268
ESP	0.726	0.459	0.318	0.810	0.543	0.338	0.741	0.471	0.352
CATNIP	0.709	0.443	0.201	0.814	0.514	0.172	0.762	0.486	0.241
CLIPZyme	0.909	0.705	0.606	0.907	0.719	0.604	0.895	0.676	0.578
RAMMESI	0.791	0.494	0.357	0.782	0.497	0.319	0.815	0.548	0.421
EZSpecificity	0.939	0.815	0.672	0.933	0.774	0.640	0.931	0.781	0.646
StereoVSE	**0.960**	**0.866**	**0.722**	**0.948**	**0.796**	**0.650**	**0.958**	**0.843**	**0.694**

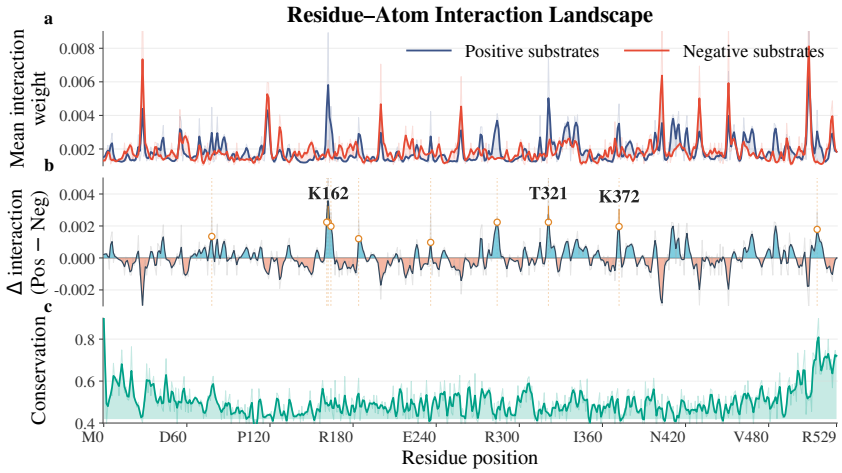


Figure 4: Sequence-wide residue–atom interaction landscape. (a) Mean residue interaction profiles for positive and negative substrates. (b) Differential interaction scores between the two groups, with representative candidate SDRs highlighted. (c) Sequence conservation profile of the same enzyme.

The ablation results on the halogenase benchmark are summarized in Table 3, while the corresponding results on the glycosyltransferase benchmark are provided in Appendix D.

All ablated variants underperform the complete model. Removing the pocket–substrate 3D branch causes the largest degradation, demonstrating the importance of local structural information. Removing BRIA decreases performance, suggesting that explicitly modeling substrate-conditioned residue–atom alignment provides complementary information beyond independent enzyme and substrate representations. Replacing ESMC with ESM-2 has a comparatively smaller effect.

Table 3: Ablation results on the halogenase benchmark.

Variant	**AUC**	**AUPR**	**F1**	**ACC**	**MCC**
ESMC→ESM2	0.9598	0.8743	0.7824	0.9134	0.7303
w/o 3D	0.7615	0.4975	0.4951	0.8085	0.3819
w/o BRIA	0.9550	0.8525	0.7819	0.9024	0.7216
w/o V-S	0.9164	0.7757	0.7317	0.8886	0.6615
StereoVSE	**0.9728**	**0.9112**	**0.8138**	**0.9238**	**0.7666**

4.5 Case Study

We evaluate whether StereoVSE provides residue-level interaction signals beyond enzyme–substrate specificity prediction using a halogenase dataset of 383 enzymes and 6,206 enzyme–substrate pairs. We first performed population-level analysis across nineteen enzymes to characterize overall interaction ranking behavior. Representative examples from individual enzymes are then presented to illustrate specific interaction patterns. For each enzyme, BRIA interaction weights are averaged over substrate atoms, and the difference between positive and negative substrate groups is used to rank candidate specificity-determining residues (SDRs). We interpret these scores as relative attribution

signals rather than physical interaction energies. We further compare the resulting residue rankings with sequence conservation Capra & Singh (2007) and evaluate selected candidates using in silico mutagenesis.

4.5.1 Interpretable Identification of Specificity-Determining Residues by BRIA

The BRIA module computes a unique residue–atom interaction matrix for every enzyme–substrate complex, which quantifies substrate-conditioned residue–atom alignment scores and enables us to rank candidate specificity-determining residues (SDRs). As visualized in Fig. 2, the BRIA interaction heatmap exhibits a prominent interaction hotspot linking residue K162 to Atom 13 of the substrate for the RebH–5-fluorotryptophan complex. As shown in Fig. 3, structural visualization confirms that Lys162 (K162) is 2.80 Å from the terminal nitrogen atom (Atom 13) of 5-fluorotryptophan. This spatial proximity is consistent with a potential residue–substrate association and supports the interpretation of the BRIA attribution signal.

Figure 4 shows whether BRIA captures substrate-specific signals at the group level. We averaged the residue–atom interaction weights across catalyzable and non-catalyzable substrates and calculated the difference between the two groups for each residue (positive minus negative). This differential profile removes interactions shared by both groups and highlights residues receiving stronger substrate-conditioned attribution signals for catalyzable substrates. K162, T321, and K372 show prominent positive differential scores and were therefore identified as candidate specificity-determining residues. These positions do not consistently overlap with peaks in sequence conservation, indicating that BRIA captures substrate-dependent residue interaction patterns that are complementary to sequence conservation analysis. Figure 5 further visualizes the relationship between differential BRIA attribution and sequence conservation for the top-ranked candidate residues.

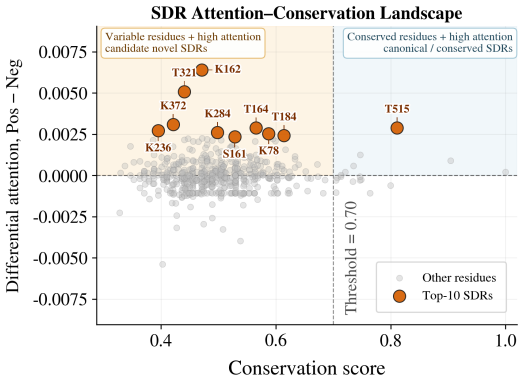


Figure 5: Differential interaction–conservation landscape. Orange points indicate the top ten residues ranked by differential interaction.

Finally, we assessed mutation sensitivity by re-encoding mutant sequences with ESMC while keeping the wild-type docking graph fixed. For ThaL, the T467A substitution caused substrate-dependent prediction changes, with substrate 232 showing the largest decrease. Detailed results are provided in Appendix E. These results suggest that BRIA-ranked residues are associated with substrate-specific prediction sensitivity.

5 Conclusion

In this work, we introduced StereoVSE, a stereochemistry-aware Vector–Scalar framework for enzyme–substrate specificity prediction with residue-level interaction analysis. StereoVSE integrates pretrained residue-level protein representations, atom-level and global molecular features, and docking-derived pocket–substrate geometries within a unified multimodal architecture. Its Vector–Scalar graph neural network preserves both invariant chemical information and directional three-dimensional geometry, while the proposed Bidirectional Residue–Atom Interaction Alignment module explicitly models substrate-dependent associations between enzyme residues and substrate atoms. Experimental results on two enzyme–substrate benchmarks show that StereoVSE achieves improved performance compared with the evaluated baselines. StereoVSE can serve as an accurate specificity predictor while providing computational attribution signals that may assist hypothesis generation for enzyme engineering. Future work will validate the identified residue–substrate interactions through biochemical experiments.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript, including improving grammar, clarity, and readability. We also used generative AI tools for retrieval and discovery purposes, such as identifying potentially relevant related work. We did not use generative AI tools to generate experimental results, perform data analysis, or make scientific conclusions.

All AI-assisted text was reviewed and revised by the authors. The authors verified the technical accuracy of the final text and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Vinayak Agarwal, Zachary D Miles, Jaclyn M Winter, Alessandra S Eustáquio, Abrahim A El Gamal, and Bradley S Moore. Enzymatic halogenation and dehalogenation reactions: pervasive and mechanistically diverse. *Chemical reviews*, 117(8):5619–5674, 2017.

Elizabeth L Bell, William Finnigan, Scott P France, Anthony P Green, Martin A Hayes, Lorna J Hepworth, Sarah L Lovelock, Haruka Niikura, Sílvia Osuna, Elvira Romero, et al. Biocatalysis. *Nature Reviews Methods Primers*, 1(1):46, 2021.

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. doi: 10.1038/nature11117.

Christelle Breton, Lenka Šnajdrová, Charlotte Jeanneau, Jaroslav Koča, and Anne Imberty. Structures and mechanisms of glycosyltransferases. *Glycobiology*, 16(2):29R–37R, 2006.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735. URL `https://www.biorxiv.org/content/10.64898/2026.06.03.729735v1`.

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

Pablo Carbonell, Jerry Wong, Neil Swainston, Eriko Takano, Nicholas J Turner, Nigel S Scrutton, Douglas B Kell, Rainer Breitling, and Jean-Loup Faulon. Selenzyme: enzyme selection tool for pathway design. *bioRxiv*, pp. 188979, 2017.

Kai Chen and Frances H Arnold. Engineering new catalytic activities in enzymes. *Nature Catalysis*, 3(3):203–213, 2020.

Haiyang Cui, Yufeng Su, Tanner J Dean, Tianhao Yu, Zhengyi Zhang, Jian Peng, Diwakar Shukla, and Huimin Zhao. Enzyme specificity prediction using cross-attention graph neural networks. *Nature*, 647(8090):639–647, 2025.

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.

Changjiang Dong, Silvana Flecks, Susanne Unversucht, Caroline Haupt, Karl-Heinz Van Pee, and James H Naismith. Tryptophan 7-halogenase (prna) structure suggests a mechanism for regioselective chlorination. *Science*, 309(5744):2216–2219, 2005.

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.

Qingyuan Feng, Evgenia Dueva, Artem Cherkasov, and Martin Ester. Padme: A deep learning-based framework for drug–target interaction prediction, 2018.

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.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1eWbxStPH.

Samuel Goldman, Ria Das, Kevin K Yang, and Connor W Coley. Machine learning modeling of family wide enzyme-substrate specificity screens. *PLoS computational biology*, 18(2):e1009853, 2022.

Noushin Hadadi, Homa MohammadiPeyhani, Ljubisa Miskovic, Marianne Seijo, and Vassily Hatzimanikatis. Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. *Proceedings of the National Academy of Sciences*, 116(15):7298–7307, 2019.

Johannes C Hermann, Ricardo Marti-Arbona, Alexander A Fedorov, Elena Fedorov, Steven C Almo, Brian K Shoichet, and Frank M Raushel. Structure-based activity prediction for an enzyme of unknown function. *Nature*, 448(7155):775–779, 2007.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlWWJSFDH.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=1YLJDvSx6J4.

Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019.

Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Explainable deep relational networks for predicting compound–protein affinities and contacts. *Journal of chemical information and modeling*, 61(1):46–66, 2020.

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

LL Lairson, Bernard Henrissat, GJ Davies, and SG Withers. Glycosyltransferases: structures, functions, and mechanisms. *Annu. Rev. Biochem.*, 77(1):521–555, 2008.

Grant M Landwehr, Jonathan W Bogart, Carol Magalhaes, Eric G Hammarlund, Ashty S Karim, and Michael C Jewett. Accelerated enzyme engineering by machine-learning guided cell-free expression. *Nature Communications*, 16(1):865, 2025. doi: 10.1038/s41467-024-55399-0.

Reuben B Leveson-Gower, Clemens Mayer, and Gerard Roelfes. The importance of catalytic promiscuity for enzyme design and evolution. *Nature Reviews Chemistry*, 3(12):687–705, 2019.

Shuya Li, Fangping Wan, Hantao Shu, Tao Jiang, Dan Zhao, and Jianyang Zeng. Monn: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell systems*, 10(4):308–322, 2020.

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.

Chen Liu, Bingxin Zhou, Xinyuan Wang, Ming Li, Guisheng Fan, and Liang Hong. Retrieval-augmented multimodal learning for enzyme-substrate interaction prediction under low-homology shift, 2026.

Stanislav Mazurenko, Zbynek Prokop, and Jiri Damborsky. Machine learning in enzyme engineering. *ACS catalysis*, 10(2):1210–1223, 2020.

Peter Mikhael, Itamar Chinn, and Regina Barzilay. CLIPZyme: Reaction-conditioned virtual screening of enzymes. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35647–35663. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/mikhael24a.html.

Zhongyu Mou, Jason Eakes, Connor J Cooper, Carmen M Foster, Robert F Standaert, Mircea Podar, Mitchel J Doktycz, and Jerry M Parks. Machine learning-based prediction of enzyme substrate scope: application to bacterial nitrilases. *Proteins: Structure, Function, and Bioinformatics*, 89 (3):336–347, 2021.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Alexandra E Paton, Daniil A Boiko, Jonathan C Perkins, Nicholas I Cemalovic, Thiago Reschützegger, Gabe Gomes, and Alison RH Narayan. Connecting chemical and protein sequence space to predict biocatalytic reactions. *Nature*, 646(8083):108–116, 2025.

Daniel Probst, Matteo Manica, Yves Gaetan Nana Teukam, Alessandro Castrogiovanni, Federico Paratore, and Teodoro Laino. Biocatalysed synthesis planning using data-driven learning. *Nature communications*, 13(1):964, 2022. doi: 10.1038/s41467-022-28536-w.

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.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Philip A Romero, Andreas Krause, and Frances H Arnold. Navigating the protein fitness landscape with gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3):E193–E201, 2013.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

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, 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.

Sofia Serrano and Noah A Smith. Is attention interpretable? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2931–2951, 2019.

Roger A Sheldon and John M Woodley. Role of biocatalysis in sustainable chemistry. *Chemical reviews*, 118(2):801–838, 2018. doi: 10.1021/acs.chemrev.7b00203.

Yusong Wang, Tong Wang, Shaoning Li, Xinheng He, Mingyu Li, Zun Wang, Nanning Zheng, Bin Shao, and Tie-Yan Liu. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 15(1):313, 2024.

Shuke Wu, Radka Snajdrova, Jeffrey C Moore, Kai Baldenius, and Uwe T Bornscheuer. Biocatalysis: enzymatic synthesis for industrial applications. *Angewandte Chemie International Edition*, 60(1):88–119, 2021. doi: 10.1002/anie.202006648.

Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019. doi: 10.1038/s41592-019-0496-6.

Min Yang, Charlie Fehl, Karen V Lees, Eng-Kiat Lim, Wendy A Offen, Gideon J Davies, Dianna J Bowles, Matthew G Davidson, Stephen J Roberts, and Benjamin G Davis. Functional and informatics analysis enables glycosyltransferase activity prediction. *Nature chemical biology*, 14(12): 1109–1117, 2018.

Adiba Yaseen, Imran Amin, Naeem Akhter, Asa Ben-Hur, and Fayyaz Minhas. Insights into performance evaluation of compound–protein interaction prediction methods. *Bioinformatics*, 38 (Supplement_2):ii75–ii81, 2022.

Cathleen Zeymer and Donald Hilvert. Directed evolution of protein catalysts. *Annual review of biochemistry*, 87(1):131–157, 2018. doi: 10.1146/annurev-biochem-062917-012034.

APPENDICES

A DATASETS AND BASELINES

A.1 DATASET

These two enzyme families provide complementary evaluation settings. However, both benchmarks are curated within specific enzyme families, and broader evaluation across diverse enzyme classes remains an important future direction. Halogenases catalyze the introduction of halogen atoms into organic substrates and are widely studied in biocatalytic functionalization. Their substrate profiles often reflect strong catalyst-dependent recognition and selectivity, making the dataset suitable for evaluating fine-grained differences among enzyme–substrate pairs Cui et al. (2025). Glycosyltransferases catalyze glycosyl-group transfer and involve diverse recognition patterns between enzymes, donors, and acceptors Yang et al. (2018). The corresponding benchmark is larger and contains a substantially lower fraction of positive pairs, making it particularly useful for evaluating model performance under sparse and imbalanced supervision.

HALOGENASE BENCHMARK

The halogenase dataset is derived from the HaloS collection and contains 383 enzymes, 974 substrates, and 6,206 enzyme–substrate pairs. Among these, 1,285 pairs are positive and 4,921 are negative, corresponding to a positive ratio of 20.71% and a positive-to-negative ratio of approximately 1:3.83. Compared with the glycosyltransferase benchmark, this dataset provides a relatively denser set of positive catalytic relationships.

GLYCOSYLTRANSFERASE BENCHMARK

The glycosyltransferase dataset contains 7,201 enzymes, 2,102 substrates, and 30,865 enzyme–substrate pairs. It includes 2,405 positive and 28,460 negative pairs, giving a positive ratio of 7.79% and a positive-to-negative ratio of approximately 1:11.83. Only 664 enzymes have at least one positive substrate in the benchmark. The limited positive supervision and strong class imbalance make positive catalytic-pair identification especially challenging and motivate the use of AUPR and MCC in addition to AUC.

As summarized in Table A1, the two benchmarks differ substantially in both scale and label distribution. The halogenase dataset offers a comparatively dense setting for assessing fine-grained specificity, whereas the glycosyltransferase dataset evaluates whether a model can recover a small number of positive catalytic pairs from a much larger set of negative combinations. Together, they provide complementary tests of predictive performance under different levels of class imbalance.

A.2 BASELINE METHODS

We compare StereoVSE with six representative enzyme–substrate and protein–compound prediction methods.

CPI CPI is a compound–protein interaction prediction framework that models compatibility between protein sequences and molecular structures Yaseen et al. (2022). We include it as a general protein–compound interaction baseline for evaluating whether enzyme-specific multimodal modeling provides additional benefits.

ESP ESP is a general enzyme–substrate prediction framework that combines enzyme sequence representations with substrate molecular features to predict whether an enzyme can act on a candidate small molecule Kroll et al. (2023). It serves as a representative sequence–chemical baseline for large-scale enzyme–substrate screening.

CATNIP CATNIP is a bidirectional enzyme–substrate recommendation framework Paton et al. (2025). It combines molecular descriptors, enzyme-sequence similarity, nearest-neighbor retrieval, and gradient-boosted models to rank compatible enzyme–substrate partners.

CLIPZyme CLIPZyme employs contrastive representation learning to align enzyme and reaction-related molecular representations for enzyme identification and compatibility modeling Mikhael et al. (2024). It provides a strong contrastive-learning baseline for evaluating multimodal enzyme–substrate representation alignment.

EZSpecificity EZSpecificity serves as a closely related structure-aware baseline for evaluating the contribution of StereoVSE's interaction-aware geometric representation Cui et al. (2025).

RAMMESI RAMMESI is a retrieval-augmented multimodal framework for enzyme–substrate interaction prediction under low-homology shifts Liu et al. (2026). It models enzyme–substrate compatibility through bidirectional cross-modal interaction and adaptive fusion, while using inference-time retrieval of similar enzymes to improve robustness and imbalance-aware loss functions to address sparse positive labels.

B Parameter Sensitivity Analysis

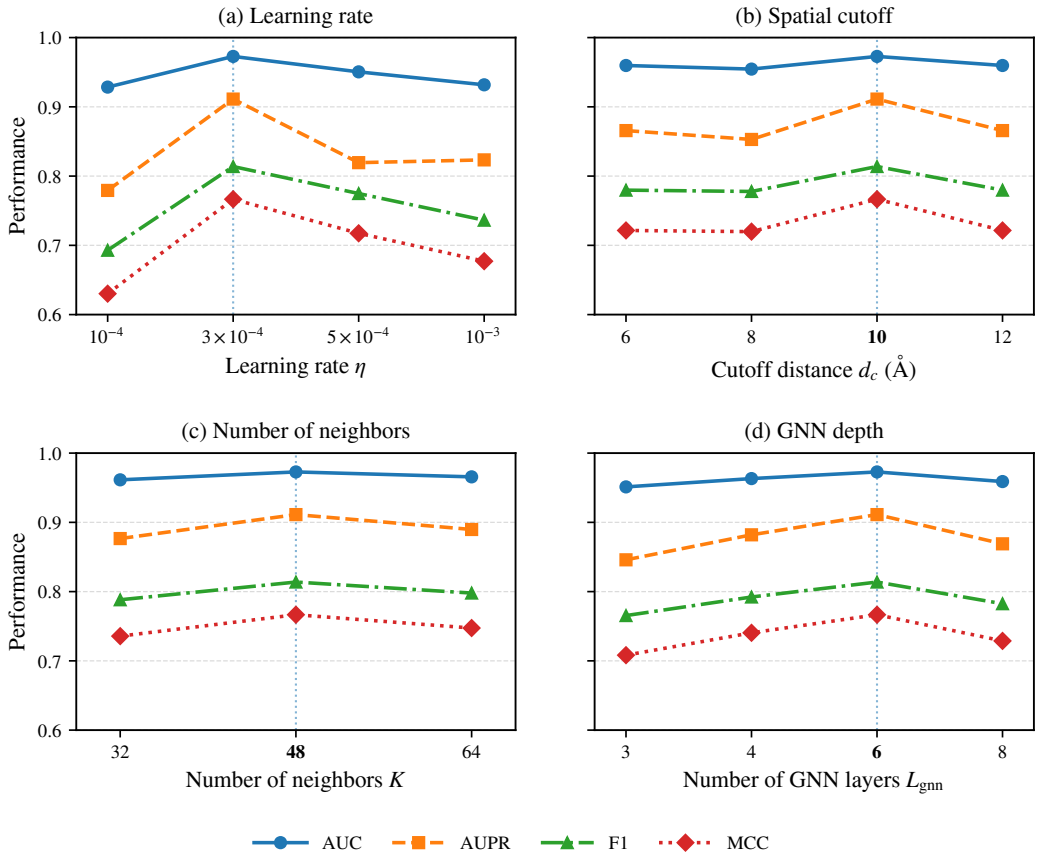


Figure A1: Hyperparameter sensitivity of StereoVSE with respect to (a) learning rate, (b) spatial cutoff distance, (c) number of neighboring nodes, and (d) GNN depth.

Figure A1 presents a one-at-a-time sensitivity analysis of four key hyperparameters, including the learning rate, spatial cutoff distance, number of neighboring nodes, and GNN depth. StereoVSE achieves the best overall performance with a learning rate of 3×10^{-4}, a cutoff distance of 10 Å, 48 neighboring nodes, and six GNN layers. Across the evaluated settings, the performance generally improves as each parameter approaches its selected value and decreases when the parameter becomes either too small or too large. In particular, AUPR, F1, and MCC exhibit greater variation than AUC, indicating that these metrics are more sensitive to hyperparameter selection under class

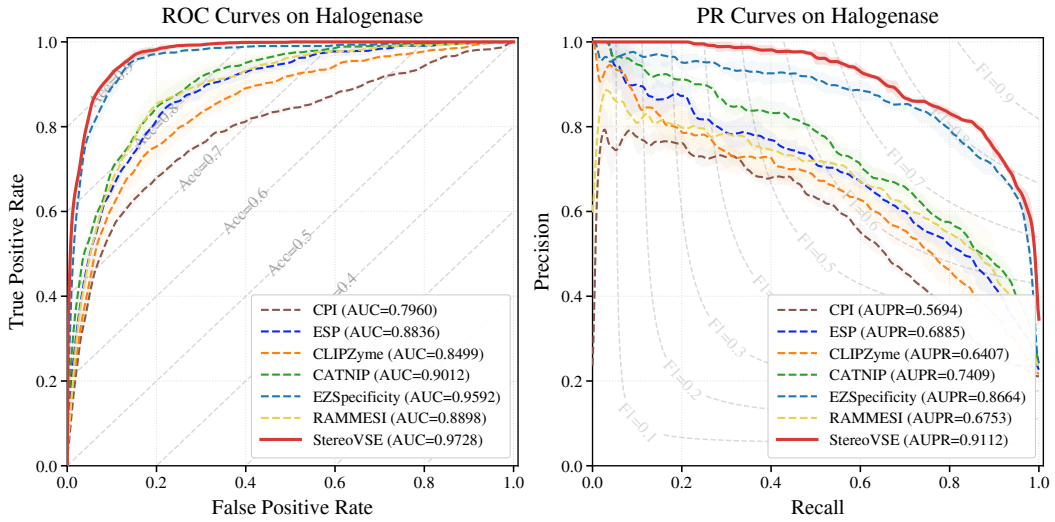


Figure A2: ROC and PR curves of different enzyme–substrate specificity prediction models on the halogenase dataset.

imbalance. The results suggest that the selected configuration provides an effective balance between structural neighborhood coverage, message-passing capacity, and optimization stability.

C Additional Performance Curves

Figure A2 presents the ROC and PR curves on the halogenase dataset, while Figure A3 presents the corresponding results on the glycosyltransferase dataset. In the ROC curve, StereoVSE remains closest to the upper-left corner across nearly all thresholds, indicating superior sensitivity-specificity trade-off. In the PR curve, StereoVSE maintains consistently higher precision over a wide range of recall values, demonstrating its stronger ability to recover positive catalytic pairs while controlling false positives.

Table A1: Statistics of the halogenase and glycosyltransferase benchmarks after preprocessing.

Metric	**Halogenase**	**Glycosyltransferase**
Enzymes	383	7,201
Substrates	974	2,102
Total pairs	6,206	30,865
Positive pairs	1,285 (20.71%)	2,405 (7.79%)
Negative pairs	4,921 (79.29%)	28,460 (92.21%)
Positive-to-negative ratio	1:3.83	1:11.83

D Ablation Study

As shown in Table A2, the ablation results demonstrate that all major components contribute to the overall performance of StereoVSE on the glycosyltransferase benchmark. Removing the pocket–substrate 3D branch leads to the most substantial degradation, highlighting the importance of pair-specific structural information. The performance drop caused by removing the Vector–Scalar module further demonstrates the value of preserving directional geometric features during message passing. Excluding BRIA also weakens the model, supporting the role of explicit residue–atom interaction modeling. Replacing ESMC with ESM-2 also leads to a performance drop, indicating that ESMC itself is also important to the model's effectiveness.

Table A2: Ablation results on the glycosyltransferase benchmark. The best results are shown in bold.

Ablation Module	AUC↑	AUPR↑	F1↑	ACC↑	MCC↑
ESMC → ESM-2	0.9157	0.6719	0.7007	0.8394	0.6206
w/o Pocket-Substrate 3D	0.8192	0.4542	0.5581	0.7187	0.4306
w/o BRIA	0.9053	0.6273	0.6885	0.8046	0.5924
w/o Vector-Scalar	0.8486	0.5052	0.6025	0.7461	0.5010
StereoVSE	**0.9364**	**0.7338**	**0.7500**	**0.8762**	**0.6813**

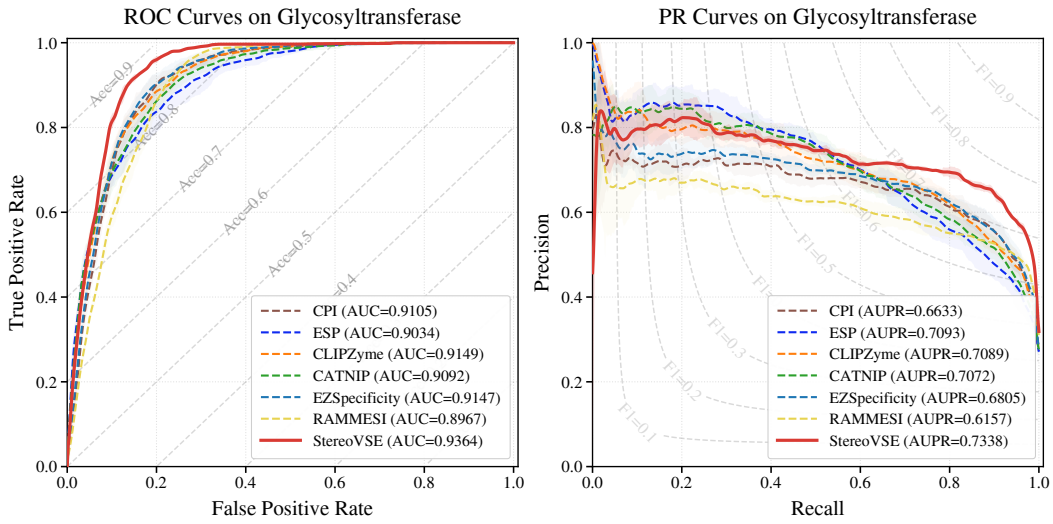


Figure A3: ROC and PR curves of different enzyme–substrate specificity prediction models on the glycosyltransferase dataset.

E Extended Experiments

E.1 Cold-start Evaluation

To evaluate model performance under distribution shifts beyond random pair-level partitioning, we conduct three complementary cold-start evaluations on the halogenase benchmark. Each sample consists of an enzyme–substrate pair with a binary label indicating whether the enzyme can catalyze the corresponding substrate. All cold-start protocols preserve the original class imbalance distribution.

Enzyme-cold split The enzyme-cold split tests the model's ability to make predictions for protein sequences that were entirely unseen during training. Specifically, the full set of 383 unique halogenase enzymes was partitioned into three disjoint subsets. Critically, all substrate molecules appearing in the test set also appear in the training set; the only factor that differs is the enzyme identity. Under this protocol, a model must infer whether a known substrate will be accepted by a novel enzyme based solely on the enzyme's sequence and structural features, without having observed any reaction paired with that enzyme during training. This setting closely mirrors practical enzyme discovery campaigns, where a newly identified or engineered enzyme is screened against a pre-existing compound library to determine its catalytic promiscuity and substrate scope.

Substrate-cold split The substrate-cold split addresses the orthogonal challenge of predicting whether a known enzyme will accept a chemically novel substrate molecule. Here, the 974 unique substrate molecules were partitioned into three disjoint sets. Enzymes, in contrast, are shared across all subsets, meaning the model has observed every enzyme during training but must generalize to

substrates with distinct chemical structures, scaffolds, or functional groups. This scenario emulates virtual screening or substrate scouting applications, where a well-characterized biocatalyst is tested against new chemical matter to identify novel transformations or expand its substrate portfolio.

Enzyme–substrate pair-cold split. The enzyme–substrate cold split evaluates prediction performance on enzyme–substrate associations that are not observed during training. Specifically, enzyme–substrate pairs are partitioned such that all test pairs are unseen, while enzymes and substrates may appear in different combinations across training and test sets. This setting assesses whether the model can infer novel enzyme–substrate compatibility beyond memorizing previously observed associations. For the substrate-cold split, $\mathcal{S}_{\text{train}} \cap \mathcal{S}_{\text{test}} = \emptyset$, while test enzymes are observed in training. For the enzyme–substrate pair-cold split, every test enzyme and substrate is individually observed during training, but their specific association is not: $\mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{test}} = \emptyset$, $\mathcal{E}_{\text{test}} \subseteq \mathcal{E}_{\text{train}}$, and $\mathcal{S}_{\text{test}} \subseteq \mathcal{S}_{\text{train}}$.

StereoVSE maintains competitive performance across the evaluated cold-start settings. The results indicate that its representations remain informative when enzyme identities, substrate identities, or specific enzyme–substrate associations are held out under the corresponding evaluation protocols.

E.2 Downstream Analysis

Assessing StereoVSE's ability to capture single-enzyme substrate preference

To further examine whether StereoVSE can characterize the substrate specificity of an individual enzyme, we performed an enzyme-fixed analysis using Q8KHZ8, a flavin-dependent tryptophan halogenase. In this analysis, the enzyme sequence and structural background remained unchanged, while the substrate was varied across 52 chemically distinct compounds. The resulting variation in model output therefore primarily reflects the substrate-dependent response learned for Q8KHZ8, rather than differences between enzyme sequences.

As summarized in Table A3, the Q8KHZ8 subset contained 12 positive and 40 negative enzyme–substrate annotations. StereoVSE produced a clearly differentiated specificity profile across these substrates. Ten of the 12 positive substrates received prediction scores above 0.5, whereas 39 of the 40 negative substrates received scores below 0.5. The mean prediction score was 0.742 for positive substrates but only 0.034 for negative substrates, indicating substantial separation between the two groups. Importantly, the model did not assign a uniform catalytic propensity to Q8KHZ8. Instead, its outputs varied from values close to one for several supported substrates to values approaching zero for substrates annotated as negative.

Table A4 presents representative substrates from the two extremes of the predicted specificity spectrum. The eight highest-scoring substrates were all annotated as positive and included chemically diverse aromatic scaffolds, such as diarylmethane-, coumarin-, diaryl ether-, naphthylamine-, anthranilate-, and sulfonamide-like structures. Thus, the learned specificity profile was not restricted to a single molecular scaffold. In contrast, the five lowest-scoring substrates were all annotated as negative and received scores below 2×10^{-37}.

These results suggest that StereoVSE captures a substrate-dependent functional profile for Q8KHZ8: the same enzyme representation produces markedly different predictions when paired with different substrate structures. This observation is consistent with the intended task of enzyme-specificity modeling, in which an enzyme may accept multiple chemically related or structurally diverse substrates while remaining selective against other compounds.

Family-Wide Complementarity Between BRIA and Sequence Conservation

To determine whether the complementarity observed for individual enzymes extends across the halogenase family, we compare differential BRIA scores with sequence conservation for 19 enzymes satisfying the downstream-analysis criteria. For each enzyme, residues are assigned both a differential interaction score and a conservation score. We assess their agreement using two complementary measures: the Spearman rank correlation over all residues and the overlap between the corresponding Top-20 residue rankings.

Table A3: Summary of the substrate-specificity profile predicted for Q8KHZ8.

Statistic	Value
Number of evaluated substrates	52
Positive benchmark annotations	12
Negative benchmark annotations	40
Positive substrates with score > 0.5	10/12
Negative substrates with score < 0.5	39/40
Mean score of positive substrates	0.742
Mean score of negative substrates	0.034

Table A4: Representative substrates from the predicted specificity spectrum of Q8KHZ8. The enzyme was fixed while the substrate was varied across 52 compounds. For compact presentation, only the eight highest-scoring and five lowest-scoring substrates are shown. The prediction score denotes the sigmoid output of StereoVSE.

Position	Substrate ID	Prediction score	Benchmark label	Substrate SMILES
1	279	1.0000	Positive	`CC(C)(C)C(c1ccc(N)cc1)c1ccc(N)cc1`
2	247	1.0000	Positive	`O=c1ccc2ccc(O)cc2o1`
3	203	1.0000	Positive	`Cc1cc(N)ccc1Oc1ccncc1`
4	99	0.9999	Positive	`Nc1ccc2ccccc2c1`
5	157	0.9917	Positive	`CC(=O)N(C)c1ccc(N)cc1`
6	290	0.9647	Positive	`O=C1CC(c2ccc(O)cc2)Oc2ccccc21`
7	88	0.9599	Positive	`Nc1ccccc1C(=O)O`
8	158	0.7397	Positive	`Cc1cc(N)ccc1S(=O)(=O)Nc1nccs1`
		· · · intermediate specificity scores are omitted for compact presentation · · ·		
48	436	1.85×10^{-37}	Negative	`Cc1[nH]c2ccccc2c1CCNCc1ccco1`
49	163	1.76×10^{-38}	Negative	`CC(=O)Nc1ccc2c(c1)CCN2`
50	502	3.11×10^{-40}	Negative	`O=C1OC(c2ccc(O)cc2)(c2ccc(O)cc2)c2ccccc21`
51	733	6.06×10^{-41}	Negative	`COc1ccc(C2CC(=O)c3c(O)cc(O)cc3O2)cc1O`
52	672	3.79×10^{-57}	Negative	`CC(C)=CCC1=C(O)C(=O)c2cc(O)c(C)c(O)c2C1=O`

Figure A4a shows the distribution of per-enzyme Spearman correlations. The correlations are concentrated near zero, with a mean of -0.035 and a median of -0.022, indicating weak rank-level correspondence between differential BRIA scores and sequence conservation.

Sequence conservation and differential BRIA scores capture distinct residue-level properties. Conservation reflects family-level evolutionary constraints, whereas differential BRIA scores are conditioned on substrate compatibility and characterize changes in the learned interaction profile.

Figure A4b compares the corresponding Top-20 residue rankings. The mean overlap is 2.632%, corresponding to approximately 0.53 shared residues per enzyme, while the median overlap is 5.000%. Together with the near-zero Spearman correlations, this low overlap shows that BRIA and sequence conservation prioritize largely distinct residue subsets.

The two rankings can therefore be examined jointly when selecting candidate residues for structural inspection, in silico perturbation, or experimental mutagenesis.

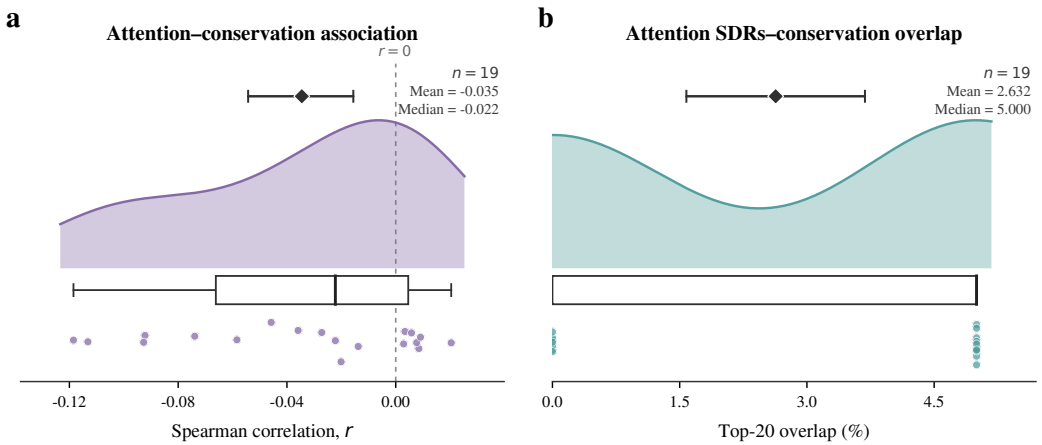


Figure A4: Family-wide comparison between BRIA-derived candidate SDR rankings and sequence conservation. (a) Distribution of per-enzyme Spearman rank correlations between differential BRIA scores and conservation scores across 19 halogenases. Values near zero indicate limited agreement between the two residue-ranking signals. (b) Distribution of the Top-20 overlap between residues ranked by differential interaction and those ranked by sequence conservation. The low overlap indicates that the two approaches generally prioritize different residue subsets.

In Silico Mutagenesis of Candidate SDRs

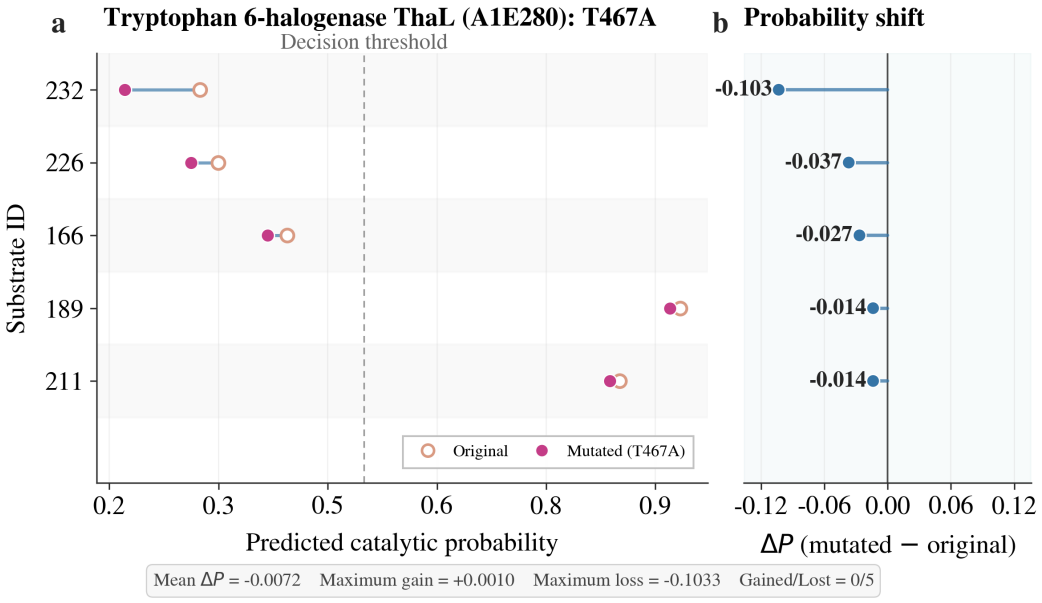


Figure A5: In silico mutagenesis shows substrate-dependent prediction sensitivity. (a) Original and mutated catalytic probabilities for five representative substrates following the T467A substitution in tryptophan 6-halogenase ThaL. The dashed line indicates the decision threshold. (b) Corresponding probability shifts, $\Delta P = P_{\text{mut}} - P_{\text{orig}}$, where negative values indicate reduced predicted catalytic compatibility after mutation.

Finally, we evaluated whether the model predictions were sensitive to substitutions at BRIA-ranked candidate residues. Figure A5 presents an in silico T467A substitution in tryptophan 6-halogenase ThaL. This mutation replaces the polar, hydroxyl-containing threonine side chain with the smaller and nonpolar side chain of alanine, thereby perturbing the local residue representation while pre-

serving the remaining enzyme sequence. The resulting changes in predicted catalytic probability were then evaluated separately for each queried substrate.

The mutation produced heterogeneous responses across the five representative substrates shown in Figure A5. Substrate 232 exhibited the largest decrease, with $\Delta P = -0.1033$, indicating substantially greater sensitivity to the T467A substitution than the other displayed substrates. The next largest decreases were observed for substrate 226 ($\Delta P = -0.037$) and substrate 166 ($\Delta P = -0.027$), whereas substrates 189 and 211 showed only modest changes of approximately -0.014. Thus, the magnitude of the mutation effect varied by nearly one order of magnitude across the displayed substrates, despite applying the same amino acid substitution to the same enzyme. None of the five displayed enzyme–substrate pairs crossed the decision threshold after mutation. Substrates 232, 226, and 166 remained below the threshold, whereas substrates 189 and 211 remained above it. The T467A substitution therefore affected prediction confidence without changing the binary classifications of these pairs.

Across the complete evaluated substrate set, the mean probability shift was $\Delta P = -0.0072$, with a maximum gain of $+0.0010$ and a maximum loss of -0.1033. The variation in ΔP across substrates indicates that the model response to T467A is substrate dependent, with the strongest decrease observed for substrate 232.

These results connect BRIA-based residue ranking with sequence-perturbation sensitivity and identify ThaL T467A with substrate 232 as a candidate combination for subsequent structural analysis or experimental evaluation.

F Additional Methodological Details

This section expands the compact formulation presented in the Methods section of the main manuscript. It specifies the dataset and derived-graph notation, structural preprocessing, complete Vector–Scalar message passing, the multi-head implementation of BRIA, multimodal optimization, and downstream residue-analysis procedures.

F.1 Input and Derived-Graph Notation

The raw labeled dataset contains enzyme sequences, substrate SMILES strings, and binary compatibility labels:

$$\mathcal{D}_{\text{raw}} = \{(S_e, M_s, y_{e,s})\}_{n=1}^{N}, \qquad y_{e,s} \in \{0, 1\}. \tag{A1}$$

For each benchmark pair, a retained three-dimensional configuration is available:

$$\mathcal{C}_{e,s} = (\mathcal{V}_{\text{rec}}, \mathcal{V}_{\text{lig}}, \mathbf{X}_{\text{rec}}, \mathbf{X}_{\text{lig}}). \tag{A2}$$

The local structural graph is deterministically derived from this retained configuration:

$$G_{e,s}^{3D} = \Psi_{\text{graph}}(\mathcal{C}_{e,s}). \tag{A3}$$

Accordingly, the processed instances used by the model are

$$\mathcal{D}_{\text{proc}} = \left\{\left(S_e, M_s, G_{e,s}^{3D}, y_{e,s}\right)\right\}_{n=1}^{N}. \tag{A4}$$

Here, Ψ_{graph} includes pocket extraction, atom-feature construction, coordinate serialization, and dynamic graph connectivity. It does not include a new docking operation.

F.2 Structural-Complex Preprocessing

Pocket and Substrate Coordinate Extraction

Let $\mathcal{V}_{\text{rec}}$ and $\mathcal{V}_{\text{lig}}$ denote the receptor and substrate atoms in the retained configuration. A receptor atom is included in the local pocket if its minimum distance to any substrate atom does not exceed 10 Å:

$$\mathcal{V}_{\text{pocket}} = \left\{ i \in \mathcal{V}_{\text{rec}} \;\middle|\; \min_{j \in \mathcal{V}_{\text{lig}}} \left\| \mathbf{x}_i^{\text{rec}} - \mathbf{x}_j^{\text{lig}} \right\|_2 \leq 10\ \text{Å} \right\}. \tag{A5}$$

The selected receptor atoms and substrate atoms preserve the coordinate frame of the retained configuration. Their coordinates are concatenated as

$$\mathbf{X}_{e,s} = [\mathbf{X}_{\text{pocket}}; \mathbf{X}_{\text{lig}}]\,. \tag{A6}$$

Because the pocket and substrate coordinates originate from the same retained configuration, no additional coordinate alignment is required.

Graph Serialization and Dynamic Connectivity

Pocket and substrate atom features, node-type indicators, and Cartesian coordinates are serialized for model loading. The graph node set is

$$\mathcal{V}_{e,s} = \mathcal{V}_{\text{pocket}} \cup \mathcal{V}_{\text{lig}}. \tag{A7}$$

The edge set is constructed dynamically using a k-nearest-neighbor rule:

$$\mathcal{E}_{e,s} = \{(i,j) \mid j \in \text{kNN}(i)\}\,. \tag{A8}$$

For each directed edge (i,j),

$$\mathbf{r}_{ij} = \mathbf{x}_j - \mathbf{x}_i, \qquad d_{ij} = \|\mathbf{r}_{ij}\|_2\,, \qquad \widehat{\mathbf{r}}_{ij} = \frac{\mathbf{r}_{ij}}{d_{ij}+\epsilon}. \tag{A9}$$

The scalar edge representation is

$$\mathbf{e}_{ij} = \phi_{\text{edge}}\left(\text{RBF}(d_{ij}) \| \mathbf{a}_i \| \mathbf{a}_j \| \,\text{Emb}(\tau_i) \| \,\text{Emb}(\tau_j)\right), \tag{A10}$$

where $\mathbf{a}_i$ and $\mathbf{a}_j$ are atom attributes and τ_i, τ_j distinguish pocket and substrate nodes. The complete graph is

$$G^{3D}_{e,s} = \left(\mathcal{V}_{e,s}, \mathcal{E}_{e,s}, \mathbf{X}_{e,s}, \{\mathbf{e}_{ij}\}_{(i,j)\in\mathcal{E}_{e,s}}\right). \tag{A11}$$

F.3 Detailed Vector–Scalar Message Passing

For each node i, the encoder maintains an invariant scalar state $\mathbf{h}_i^{(l)} \in \mathbb{R}^d$ and an equivariant vector state $\mathbf{V}_i^{(l)} \in \mathbb{R}^{d_v \times 3}$.

Initial scalar features are obtained from atom attributes:

$$\mathbf{h}_i^{(0)} = \phi_{\text{atom}}\left(\mathbf{a}_i\right). \tag{A12}$$

The initial vector features are constructed through local directional aggregation:

$$\mathbf{V}_i^{(0)} = \sum_{j\in\mathcal{N}(i)} \phi_{\text{init}}\left(\mathbf{e}_{ij}\right) \odot \widehat{\mathbf{r}}_{ij}. \tag{A13}$$

Here, $\phi_{\text{init}}(\mathbf{e}_{ij}) \in \mathbb{R}^{d_v}$, and multiplication with $\widehat{\mathbf{r}}_{ij} \in \mathbb{R}^3$ is broadcast over vector channels.

At layer l, channel-wise invariant descriptors of the vector states are

$$\mathcal{I}_{ij}^{(l)} = \left[\left\|\mathbf{V}_i^{(l)}\right\|_2 \| \left\|\mathbf{V}_j^{(l)}\right\|_2 \| \left\langle \mathbf{V}_i^{(l)}, \mathbf{V}_j^{(l)} \right\rangle\right]. \tag{A14}$$

Each entry in $\|\mathbf{V}_i^{(l)}\|_2 \in \mathbb{R}^{d_v}$ is the norm of one vector channel, and $\langle \mathbf{V}_i^{(l)}, \mathbf{V}_j^{(l)} \rangle \in \mathbb{R}^{d_v}$ contains channel-wise inner products.

The scalar edge message is

$$\mathbf{m}_{ij}^{(l)} = \phi_m^{(l)}\left(\mathbf{h}_i^{(l)} \| \mathbf{h}_j^{(l)} \| \mathbf{e}_{ij} \| \mathcal{I}_{ij}^{(l)}\right). \tag{A15}$$

Neighboring messages are normalized by

$$\alpha_{ij}^{(l)} = \frac{\exp\left(\phi_\alpha^{(l)}\left(\mathbf{m}_{ij}^{(l)}\right)\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\phi_\alpha^{(l)}\left(\mathbf{m}_{ik}^{(l)}\right)\right)}. \tag{A16}$$

The vector-state and relative-direction gates are

$$\mathbf{g}_{ij}^{v} = \sigma\left(\phi_{g_v}^{(l)}\left(\mathbf{m}_{ij}^{(l)}\right)\right), \qquad \mathbf{g}_{ij}^{r} = \sigma\left(\phi_{g_r}^{(l)}\left(\mathbf{m}_{ij}^{(l)}\right)\right). \tag{A17}$$

The resulting directional edge message is

$$\mathbf{M}_{ij}^{(l)} = \mathbf{g}_{ij}^{v} \odot \mathbf{V}_{j}^{(l)} + \mathbf{g}_{ij}^{r} \odot \widehat{\mathbf{r}}_{ij}. \tag{A18}$$

Aggregated scalar and vector messages are

$$\begin{aligned} \mathbf{u}_{i}^{(l)} &= \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} \phi_{s}^{(l)}\left(\mathbf{m}_{ij}^{(l)}\right), \\ \mathbf{U}_{i}^{(l)} &= \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} \mathbf{M}_{ij}^{(l)}. \end{aligned} \tag{A19}$$

The scalar and vector streams are coupled through

$$\begin{aligned} \mathbf{h}_{i}^{(l+1)} &= \mathrm{LN}\left[\mathbf{h}_{i}^{(l)} + \phi_{h}^{(l)}\left(\mathbf{u}_{i}^{(l)} \| \left\|\mathbf{U}_{i}^{(l)}\right\|_{2}\right)\right], \\ \mathbf{V}_{i}^{(l+1)} &= \mathbf{V}_{i}^{(l)} + \phi_{\mathrm{vec}}^{(l)}\left(\mathbf{u}_{i}^{(l)}\right) \odot \mathbf{U}_{i}^{(l)}. \end{aligned} \tag{A20}$$

The vector norm transfers invariant geometric evidence into the scalar stream, whereas the scalar representation gates the update of directional features.

After L layers, attention pooling yields

$$\mathbf{z}_{3D} = \mathrm{AttnPool}\left(\left\{\mathbf{h}_{i}^{(L)}\right\}_{i \in \mathcal{V}_{e,s}}\right). \tag{A21}$$

Under a global rigid transformation

$$\mathbf{x}_{i}' = \mathbf{R}\mathbf{x}_{i} + \mathbf{t}, \qquad \mathbf{R} \in SO(3), \tag{A22}$$

the scalar and vector states satisfy

$$\mathbf{h}_{i}'^{(l)} = \mathbf{h}_{i}^{(l)}, \qquad \mathbf{V}_{i}'^{(l)} = \mathbf{V}_{i}^{(l)}\mathbf{R}^{\top}. \tag{A23}$$

This property follows because distances, vector norms, and channel-wise inner products are rotation invariant, whereas $\widehat{\mathbf{r}}_{ij}$ and the vector states rotate consistently. Hence, the pooled scalar representation and final prediction are invariant to global rotations and translations.

F.4 Multi-Head Implementation of BRIA

BRIA performs cross-attention between atom-level substrate features $\mathbf{H}^{s} \in \mathbb{R}^{L_s \times d}$ and residue-level enzyme features $\mathbf{H}^{e} \in \mathbb{R}^{L_e \times d}$.

For attention head h, substrate atoms serve as queries and enzyme residues serve as keys and values:

$$\mathbf{Q}_{s}^{(h)} = \mathbf{H}^{s}\mathbf{W}_{Q,s}^{(h)}, \qquad \mathbf{K}_{e}^{(h)} = \mathbf{H}^{e}\mathbf{W}_{K,e}^{(h)}, \qquad \mathbf{V}_{e}^{(h)} = \mathbf{H}^{e}\mathbf{W}_{V,e}^{(h)}. \tag{A24}$$

The atom-to-residue alignment matrix is

$$\mathbf{A}_{s \leftarrow e}^{(h)} = \mathrm{softmax}_{\mathrm{res}}\left(\frac{\mathbf{Q}_{s}^{(h)}\left(\mathbf{K}_{e}^{(h)}\right)^{\top}}{\sqrt{d_h}} + \mathbf{M}_{s,e}\right) \in \mathbb{R}^{L_s \times L_e}, \tag{A25}$$

where softmax is applied across the residue dimension and $\mathbf{M}_{s,e}$ is an additive padding mask. The enzyme-conditioned atom representation is

$$\mathbf{Z}_{s \leftarrow e}^{(h)} = \mathbf{A}_{s \leftarrow e}^{(h)}\mathbf{V}_{e}^{(h)}. \tag{A26}$$

For the reverse direction,

$$\mathbf{Q}_{e}^{(h)} = \mathbf{H}^{e}\mathbf{W}_{Q,e}^{(h)}, \qquad \mathbf{K}_{s}^{(h)} = \mathbf{H}^{s}\mathbf{W}_{K,s}^{(h)}, \qquad \mathbf{V}_{s}^{(h)} = \mathbf{H}^{s}\mathbf{W}_{V,s}^{(h)}. \tag{A27}$$

The residue-to-atom alignment matrix is

$$\mathbf{A}_{e\leftarrow s}^{(h)} = \mathrm{softmax}_{\mathrm{atom}}\left(\frac{\mathbf{Q}_e^{(h)}\left(\mathbf{K}_s^{(h)}\right)^\top}{\sqrt{d_h}} + \mathbf{M}_{e,s}\right) \in \mathbb{R}^{L_e\times L_s}, \tag{A28}$$

and the substrate-conditioned residue representation is

$$\mathbf{Z}_{e\leftarrow s}^{(h)} = \mathbf{A}_{e\leftarrow s}^{(h)}\mathbf{V}_s^{(h)}. \tag{A29}$$

The multi-head outputs are

$$\begin{aligned}\widetilde{\mathbf{H}}^s &= \mathrm{Concat}_{h=1}^{H}\left(\mathbf{Z}_{s\leftarrow e}^{(h)}\right)\mathbf{W}_{O,s},\\ \widetilde{\mathbf{H}}^e &= \mathrm{Concat}_{h=1}^{H}\left(\mathbf{Z}_{e\leftarrow s}^{(h)}\right)\mathbf{W}_{O,e}.\end{aligned} \tag{A30}$$

The interaction-aware representations used for fusion are

$$\mathbf{z}_s = \mathrm{Pool}\left(\widetilde{\mathbf{H}}^s\right), \qquad \mathbf{z}_e = \mathrm{Pool}\left(\widetilde{\mathbf{H}}^e\right). \tag{A31}$$

For residue-level prioritization, we retain the head-averaged atom-to-residue matrix

$$\mathbf{A} = \frac{1}{H}\sum_{h=1}^{H}\mathbf{A}_{s\leftarrow e}^{(h)} \in \mathbb{R}^{L_s\times L_e}. \tag{A32}$$

The matrix is retained before sequence-level pooling so that individual atom–residue associations remain accessible for downstream analysis.

F.5 Multimodal Fusion and Optimization

The interaction-aware enzyme and substrate representations are concatenated with the global chemical and structural representations:

$$\mathbf{z}_{\mathrm{final}} = \phi_{\mathrm{fuse}}\left(\mathbf{z}_e\|\mathbf{z}_s\|\mathbf{z}_{\mathrm{chem}}\|\mathbf{z}_{3D}\right). \tag{A33}$$

The enzyme–substrate compatibility probability is

$$\widehat{y} = \sigma\left(f_{\mathrm{pred}}\left(\mathbf{z}_{\mathrm{final}}\right)\right). \tag{A34}$$

To account for class imbalance, the prediction objective is weighted binary cross-entropy:

$$\mathcal{L}_{\mathrm{pred}} = -\frac{1}{N}\sum_{n=1}^{N} w_n\left[y_n\log\widehat{y}_n + (1-y_n)\log(1-\widehat{y}_n)\right]. \tag{A35}$$

Including parameter regularization, the complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{reg}}\|\Theta\|_2^2. \tag{A36}$$

F.6 BRIA-Guided Identification of Candidate SDRs

For enzyme e, let $\mathcal{P}_e^+$ and $\mathcal{P}_e^-$ denote the sets of compatible and incompatible substrates, respectively. For sample n, the interaction score of residue j is obtained by averaging the head-averaged BRIA matrix over valid substrate atoms:

$$a_j^{(n)} = \frac{1}{L_s^{(n)}}\sum_{i=1}^{L_s^{(n)}} A_{ij}^{(n)}. \tag{A37}$$

The positive- and negative-substrate interaction profiles are

$$\overline{a}_j^{+} = \frac{1}{|\mathcal{P}_e^+|}\sum_{n\in\mathcal{P}_e^+} a_j^{(n)}, \qquad \overline{a}_j^{-} = \frac{1}{|\mathcal{P}_e^-|}\sum_{n\in\mathcal{P}_e^-} a_j^{(n)}. \tag{A38}$$

The differential interaction score is

$$\Delta A_j = \overline{a}_j^{+} - \overline{a}_j^{-}. \tag{A39}$$

For within-enzyme ranking, the differential scores are standardized:

$$\mathrm{SDRScore}_j = \frac{\Delta A_j - \mu_{\Delta A}}{\sigma_{\Delta A} + \epsilon}, \tag{A40}$$

where $\mu_{\Delta A}$ and $\sigma_{\Delta A}$ are calculated across the residues of the same enzyme. The top-K positions are selected as candidate specificity-determining residues:

$$\mathcal{R}_{\mathrm{SDR}}^{K} = \mathrm{TopK}\left(\{\mathrm{SDRScore}_j\}_{j=1}^{L_e}\right). \tag{A41}$$

We use the resulting top-ranked positions as candidate specificity-determining residues for downstream analysis.

F.7 Comparison with Evolutionary Conservation

Let C_j denote the evolutionary conservation score of residue j. The rank association between interaction-derived and conservation-derived scores is quantified using Spearman correlation:

$$\rho = \mathrm{corr}_{\mathrm{Spearman}}\left(\Delta\mathbf{A}, \mathbf{C}\right), \tag{A42}$$

where $\Delta\mathbf{A} = [\Delta A_1, \ldots, \Delta A_{L_e}]$ and $\mathbf{C} = [C_1, \ldots, C_{L_e}]$.

Their top-K overlap is

$$\mathrm{Overlap@K} = \frac{\left|\mathcal{R}_{\mathrm{SDR}}^{K} \cap \mathcal{R}_{\mathrm{cons}}^{K}\right|}{K}. \tag{A43}$$

Low correlation or overlap indicates that BRIA-derived interaction scores and evolutionary conservation prioritize different residue subsets.

F.8 Sequence-Conditioned In Silico Mutagenesis

For candidate residue j and target amino acid r', the mutated sequence is

$$S_e^{j\to r'} = (r_1, \ldots, r_{j-1}, r', r_{j+1}, \ldots, r_{L_e}). \tag{A44}$$

The complete mutated sequence is re-encoded by the protein language model. The predictions for the original and perturbed sequences are

$$\begin{aligned} P_{\mathrm{orig}} &= \mathrm{StereoVSE}\left(S_e, M_s, G_{e,s}^{3D}\right), \\ P_{\mathrm{mut}} &= \mathrm{StereoVSE}\left(S_e^{j\to r'}, M_s, G_{e,s}^{3D}\right). \end{aligned} \tag{A45}$$

The mutation-associated prediction shift is

$$\Delta P = P_{\mathrm{mut}} - P_{\mathrm{orig}}. \tag{A46}$$

For a collection of L substrates, the mutation-response profile is

$$\mathbf{g}_{j\to r'} = [\Delta P_1, \Delta P_2, \ldots, \Delta P_L]. \tag{A47}$$

The wild-type docking-derived graph $G_{e,s}^{3D}$ is held fixed during this analysis. The mutated sequence is re-encoded, while receptor structure prediction, substrate docking, pocket extraction, and graph construction are not repeated. Accordingly, ΔP quantifies sequence-level prediction sensitivity under a fixed wild-type structural context and does not model mutation-induced structural rearrangement.

Table A5: Hardware and software configurations used in the computational experiments.

Component	Configuration
Operating system	Ubuntu 22.04.4 LTS
System architecture	Linux x86_64
GPU	4× NVIDIA H100 PCIe
CPU	2× Intel Xeon Gold 6267C at 2.60 GHz
CPU resources	48 physical cores, 96 threads
System memory	1.5 TB RAM
Storage	7.0 TB and 1.8 TB NVMe SSDs
Python	3.10.12
PyTorch	2.5.1+cu124
PyTorch Lightning	1.9.0
PyTorch Geometric	2.7.0
RDKit	2023.09.2
CUDA Toolkit	12.4
cuDNN	9.1.0
NVIDIA Driver	595.58.03
scikit-learn	1.6.1
NumPy	1.26.4
Pandas	2.3.3

G Computing Environment

All experiments were conducted on a Linux server running Ubuntu 22.04.4 LTS with the x86_64 architecture. The server was equipped with four NVIDIA H100 PCIe GPUs, two Intel Xeon Gold 6267C processors operating at 2.60 GHz, and 1.5 TB of system memory. The two processors provided 48 physical CPU cores and 96 hardware threads in total. The available local storage consisted of a 7.0 TB NVMe SSD for experimental data and model checkpoints, together with a 1.8 TB NVMe SSD for the system and user environment.

The proposed model was implemented in Python 3.10.12 using PyTorch 2.5.1. PyTorch Lightning 1.9.0 was used to organize the training and validation procedures, while PyTorch Geometric 2.7.0 was employed for graph construction and geometric message passing. Molecular preprocessing, conformer handling, and chemical feature extraction were performed using RDKit 2023.09.2. GPU acceleration was supported by CUDA 12.4 and cuDNN 9.1.0. Data preprocessing, dataset partitioning, and performance evaluation were conducted using scikit-learn 1.6.1, NumPy 1.26.4, and Pandas 2.3.3. All model variants and ablation experiments were executed under the same software environment to ensure a consistent and fair comparison.

Table A5 summarizes the hardware and software configurations used in our experiments.