EPIFORMER: LEARNING ANTIGEN-ANTIBODY INTERACTIONS FOR EPITOPE PREDICTION VIA GEOMETRIC DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Antibodies neutralize foreign antigens by binding to specific surface regions called epitopes. Computational epitope prediction is critical for understanding immune recognition and guiding antibody engineering. However, existing methods face three fundamental challenges: antibody-aware models encode each chain independently and combine them only at a late stage, failing to capture co-dependent structural features that define binding interfaces, whereas severe class imbalance and scarcity of known antibody-antigen complexes render standard training objectives ineffective. We propose *EpiFormer*, a general encoder-decoder framework that addresses these challenges jointly. Our key design principle is interleaved cross-attention within GNN encoding layers, enabling bidirectional antigen-antibody information flow throughout representation learning rather than only at the output. This early-fusion principle is backbone-agnostic, providing consistent gains across GNN architectures from simple GCNs to equivariant models. We further show that sparsity-aware objectives are effective when paired with early-fusion architectures for the epitope prediction task. *EpiFormer* improves over the previous best method by over 40% in F1 score on standard benchmarks, demonstrating generalizability and cross-dataset transferability. Notably, two behaviors of *EpiFormer* emerge from end-to-end training alone, where the learned cross-attention gates favor antigen-to-antibody information flow, consistent with the asymmetric roles of the two chains at the binding interface, and the model's preference for geometric over evolutionary features aligns with the established finding that epitope residues are not evolutionarily conserved.

1 INTRODUCTION

Antibodies are Y-shaped proteins that recognize and neutralize foreign substances by binding to specific surface regions on antigens called epitopes. Accurate identification of epitopes is critical for therapeutic antibody design, vaccine development, and understanding immune recognition (Norman et al., 2020; Joubbi et al., 2024). While traditional experimental approaches for epitope mapping are time-consuming and expensive, computational methods offer the potential for rapid screening of candidate therapeutics (Hummer et al., 2022).

Computational epitope prediction methods can be categorized as *antibody-agnostic* or *antibody-aware*. Antibody-agnostic methods, including DiscoTope3 (Høie et al., 2024), EpiGraph (Choi & Kim, 2024), and GraphBepi (Zeng et al., 2023), predict binding sites as intrinsic properties of the antigen without considering the specific antibody. However, epitopes are fundamentally antibody-specific, as different antibodies targeting the same antigen bind to different surface regions (Liu et al., 2024). Antibody-aware methods such as PECAN (Pittala & Bailey-Kellogg, 2020), MIPE (Wang et al., 2024b), and WALLE (Liu et al., 2024) address this by conditioning on the antibody structure.

However, antibody-aware epitope prediction faces three fundamental challenges. First, existing methods encode antigen and antibody independently and combine them only at a late stage (*late fusion*), failing to capture co-dependent structural features that define binding interfaces. Antibody CDR loops adapt their conformation to the epitope surface, and epitope residues similarly adjust to accommodate the paratope. Encoding each chain without knowledge of its counterpart misses

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.

these interaction-specific geometric signatures. Second, epitope residues constitute fewer than 5% of antigen surface residues, creating severe class imbalance that renders standard objectives ineffective. Third, the scarcity of known antibody-antigen complexes limits the data available for training.

We propose *EpiFormer*, a general encoder-decoder framework that addresses these challenges jointly. Our key design principle is *interleaved bidirectional cross-attention* within GNN encoding layers, enabling antigen-antibody information flow throughout representation learning rather than only at the output. This early-fusion principle is backbone-agnostic, providing consistent gains across GNN architectures from simple GCNs to equivariant models. We further develop sparsity-aware training objectives, including Dice loss, per-graph count regularization, bipartite edge prediction, and distance supervision, that are effective when paired with early-fusion architectures. Within this framework, we propose *EGNN-R*, a multi-relational E(3)-equivariant GNN where each edge relation type has its own learned transformation, providing additional gains.

Our main contributions are:

1. A **backbone-agnostic early-fusion framework** with interleaved bidirectional cross-attention within GNN encoding layers, allowing the antigen and antibody representations to inform each other at every layer rather than only in the decoder.
2. An **architecture-loss co-design principle**, where sparsity-aware objectives are developed to effectively address class imbalance and data scarcity for the epitope prediction task. These objectives improve early-fusion models while degrading late-fusion ones, and the two therefore have to be co-designed.
3. We show that compact geometric features outperform high-dimensional protein language model embeddings for epitope prediction, and that the learned cross-attention gates reveal a consistent antigen-antibody asymmetry that aligns with the distinct roles of the two chains.

2 Related Work

Antibody-agnostic methods predict epitopes without considering which antibody is binding. Early epitope prediction methods relied on sequence features such as hydrophilicity and accessibility (Jespersen et al., 2017). Structure-based approaches, such as DiscoTope3 (Høie et al., 2024), EpiGraph (Choi & Kim, 2024), and GraphBepi (Zeng et al., 2023), improved upon these by incorporating 3D information. **Antibody-aware methods** condition predictions on the binding counterpart. PECAN (Pittala & Bailey-Kellogg, 2020) encodes antibody and antigen with separate graph convolutions and combines them via bilinear attention. MIPE (Wang et al., 2024b) uses multi-modal contrastive learning to align sequence and structure representations, applying multi-head attention for late-stage fusion. WALLE (Liu et al., 2024) formulates epitope prediction as bipartite link prediction between antibody and antigen residue graphs. EpiScan (Wang et al., 2024a) incorporates CDR masking into sequence embeddings. All of these methods share a *late fusion* architecture in which antigen and antibody are encoded independently before cross-chain attention is applied.

E(3)-equivariant neural networks preserve geometric properties under rotations and translations, making them well-suited for molecular modeling. EGNN (Satorras et al., 2021) achieves equivariance by updating coordinates along displacement vectors scaled by learned invariant functions. Equiformer (Liao & Smidt, 2022) and EquiformerV2 (Liao et al., 2023) extend this with SE(3)/SO(2) equivariant attention using spherical harmonics. GearNet (Zhang et al., 2022) introduced multi-relational protein graphs with seven edge types encoding sequential and spatial relationships, but uses a shared message function across relations. Surface-based methods, including MaSIF (Gainza et al., 2020) and AtomSurf (Mallet et al., 2023), operate on molecular surface representations with geodesic or spectral convolutions.

Several methods from adjacent domains share architectural similarities with *EpiFormer*. **CheapNet** (Lim et al., 2025) predicts protein-ligand binding affinity using geometry-informed graph networks (GIGN) for each entity, followed by cross-attention. This is similar to our use of separate encoders with cross-attention, but CheapNet applies cross-attention only at the final layer (late fusion), whereas *EpiFormer* interleaves it at every encoder block. **EquiPocket** (Zhang et al., 2023) uses E(3)-equivariant message passing for ligand binding site prediction but is antibody-agnostic and operates at atom-level rather than the residue-level. **DiffDock** (Corso et al., 2022) employs

SE(3)-equivariant score networks for molecular docking, demonstrating the value of equivariant architectures for protein-ligand interactions. **Boltz-1/2** (Wohlwend et al., 2025; Passaro et al., 2025) use Pairformer architectures with triangle attention for structure prediction; Boltz-2 adds cross-attention between binding counterparts, but applies it to pairwise representations rather than node embeddings.

EpiFormer differs from these methods by addressing their limitations for epitope prediction. Like CheapNet, we use separate encoders with cross-attention, but apply attention at every layer rather than only at the output. Like GearNet, we use multi-relational graphs, but learn separate transformations per relation type. Similar to EGNN and EquiPocket, we maintain E(3)-equivariance, but extend it to the multi-relational setting. Recently, ATProt (Gao et al., 2024) applies cross-attention between antibody and antigen chains, but only once after all encoding layers are complete (late fusion). DiffDock-PP (Ketata et al., 2023) uses SE(3)-equivariant tensor product convolutions for rigid-body docking but does not perform residue-level binding site prediction. To our knowledge, no prior method interleaves cross-attention within structural GNN encoding layers for protein-protein interaction prediction. While prior antibody-aware methods (PECAN, MIPE, ATProt) encode chains independently before late fusion, *EpiFormer* enables cross-chain information flow throughout the encoding process. Antibody-antigen, however, remains one of the interface classes on which co-folding is least reliable, with AlphaFold3 reaching acceptable accuracy on 38% of complexes (Hitawala & Gray, 2025; Unsal et al., 2025), which leaves scope for a task-specific predictor (Appendix H.1).

3 METHODS

3.1 PRELIMINARIES

Graph construction. We build two independent residue graphs $\mathcal{G}_{\text{ag}} = (\mathcal{V}_{\text{ag}}, \mathcal{E}_{\text{ag}}, \mathcal{R})$ and $\mathcal{G}_{\text{ab}} = (\mathcal{V}_{\text{ab}}, \mathcal{E}_{\text{ab}}, \mathcal{R})$ from the antigen and antibody chains, which AsEP releases as chains extracted from bound complexes. The relative position of the two chains does not reach the model, since message passing runs on intra-chain edges only and the decoder receives two embedding matrices with no coordinates. Inter-chain distances enter only as auxiliary targets discarded at inference. Appendices A.3 and E state and test this setting. Vertex $v_i \in \mathcal{V}$ represents residue i, centered on C_α at coordinate $\mathbf{x}_i \in \mathbb{R}^3$ ($|\mathcal{V}_{\text{ag}}| = n$, $|\mathcal{V}_{\text{ab}}| = m$). Each node carries a geometric feature vector $\mathbf{h}_i \in \mathbb{R}^{d_h}$ (encoding residue type, backbone geometry, physicochemical properties) and a coordinate matrix $\mathbf{X}_i \in \mathbb{R}^{3\times 4}$ for four backbone atoms $\xi = \{\text{N}, \text{C}_\alpha, \text{C}_\beta, \text{O}\}$. Each edge $e_{i,j}$ carries a feature vector $\mathbf{f}_{i,j} \in \mathbb{R}^{d_f}$ (distances, angles) and a relation tuple $\mathbf{r}_{i,j} \subseteq \mathcal{R}$. The relation set $\mathcal{R} = \{\rho_1, \rho_2, \rho_3, \rho_4\}$ captures sequential bonds (ρ_1), short-range coupling (ρ_2), K-nearest neighbors (ρ_3), and medium-range contacts within 8 Å (ρ_4). We denote a residue graph as $\mathcal{G} = (\mathbf{H}, \mathbf{X}, \mathbf{F}, \mathbf{R})$; see Appendix B.1 for details.

Problem Formulation. We formulate epitope prediction as binary node classification. A residue $v \in \mathcal{V}_{\text{ag}}$ is labeled as an epitope if it is within 4.5Å of any residue in $\mathcal{V}_{\text{ab}}$, and non-epitope otherwise. The classifier $f : v_{\text{ag}} \to \{0, 1\}$ predicts $\hat{y}_{\text{ag}} = f(v_{\text{ag}}; \mathcal{G}_{\text{ag}}, \mathcal{G}_{\text{ab}})$.

Equivariance. Since molecular properties are unchanged under rigid body transformations, geometric GNNs incorporate E(3)-equivariance as an inductive bias (Jiao et al., 2023). For coordinates $\mathbf{X} \in \mathbb{R}^{3\times m}$ and scalar features $\mathbf{h} \in \mathbb{R}^d$, an E(3)-equivariant function satisfies $f(g \cdot \mathbf{X}, \mathbf{h}) = g \cdot f(\mathbf{X}, \mathbf{h})$ for all $g \in \text{E}(3)$, where group actions are translations ($g \cdot \mathbf{X} = \mathbf{X} + \mathbf{b}$) or rotations/reflections ($g \cdot \mathbf{X} = \mathbf{O}\mathbf{X}$, $\mathbf{O} \in \text{O}(3)$). E(3)-invariant functions instead satisfy $f(g \cdot \mathbf{X}, \mathbf{h}) = f(\mathbf{X}, \mathbf{h})$.

3.2 EPIFORMER

In this section, we present the architecture of *EpiFormer*, an encoder-decoder framework for antibody-antigen binding-site prediction. The model receives two disjoint multi-relational residue graphs, $\mathcal{G}_{\text{ag}}$ and $\mathcal{G}_{\text{ab}}$, processes them with independent E(3)-equivariant encoders that produce residue-level embeddings, and passes these embeddings to a cross-attention decoder that reconstructs the bipartite adjacency matrix $\hat{\mathcal{E}}_{\text{bg}} \in \{0, 1\}^{n\times m}$. The workflow is in Figure 1 and the algorithm in Appendix 1.

Encoder. *EpiFormer* contains two parallel encoders, one dedicated to the antigen chain and the other to the antibody chain (Figure 1a). Each node encodes Cartesian coordinates $\mathbf{x}_i \in \mathbb{R}^3$ and geometric

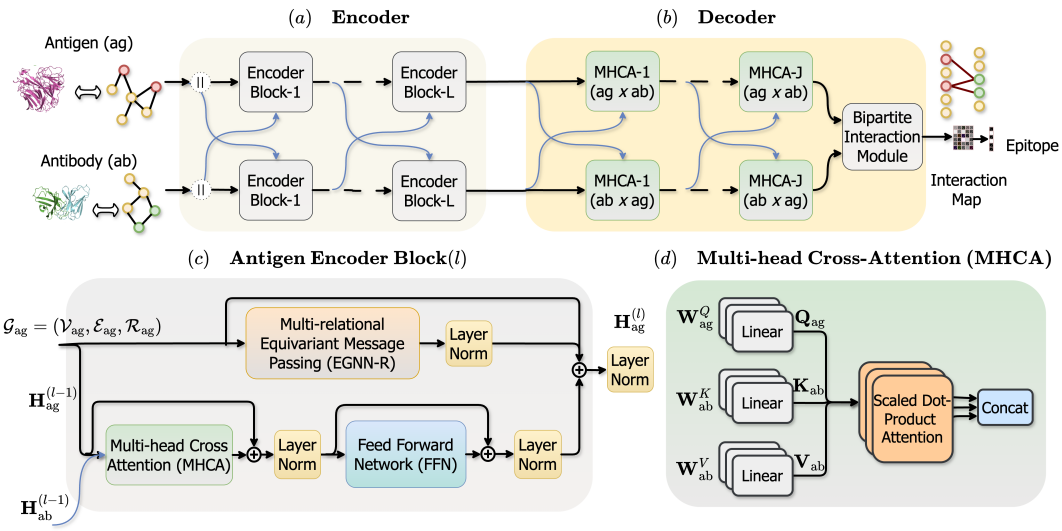


Figure 1: Overview of *EpiFormer*. (a) Parallel EGNN-R encoders with interleaved cross-attention. (b) Bidirectional cross-attention decoder producing the interaction map. (c) Encoder block schematic ("⊕" = addition). (d) Multi-head cross-attention between antigen and antibody residues.

descriptors $\mathbf{h}_i^{\text{geo}} \in \mathbb{R}^{d_{\text{geo}}}$, which are projected to the working width d_h:

$$\mathbf{h}_i^0 = \mathbf{W}_{\text{geo}}\,\mathbf{h}_i^{\text{geo}} \in \mathbb{R}^{d_h}. \tag{1}$$

The vector $\mathbf{h}_i^0$ serves as the initial node state for the first *EpiFormer* encoder block. The schematic of an *EpiFormer* block is shown in Figure 1 **(c)**. Let $\mathbf{H}_{\text{ag}}^{\ell} \in \mathbb{R}^{n\times d_h}$ and $\mathbf{H}_{\text{ab}}^{\ell} \in \mathbb{R}^{m\times d_h}$ be the current embeddings, which are passed in parallel to their EGNN-R and MHCA layers.

EGNN-R layer: We develop a relation-aware variant of EGNN (Satorras et al., 2021) to propagate structural and geometric information within each chain. Unlike shared-weight approaches, EGNN-R uses per-relation message functions because sequential edges (covalent constraints) and spatial edges (non-covalent contacts) represent fundamentally different physics. Let $\mathbf{h}_i^{\ell} \in \mathbb{R}^{d_h}$, $\mathbf{x}_i^{\ell} \in \mathbb{R}^3$ be the feature and coordinate of residue i at layer ℓ, $d_{ij} = \|\mathbf{x}_i^{\ell} - \mathbf{x}_j^{\ell}\|_2^2$ the squared distance, and $\boldsymbol{\delta}_{ij} = \mathbf{x}_i^{\ell} - \mathbf{x}_j^{\ell}$ the displacement:

$$m_{ij}^{\rho} = \phi_m^{\rho}\big(\mathbf{h}_i^{\ell}, \mathbf{h}_j^{\ell}, \gamma(d_{ij}), \mathbf{f}_{ij}\big), \qquad s_{ij}^{\rho} = \phi_x^{\rho}\big(m_{ij}^{\rho}\big), \tag{2}$$

$$\mathbf{h}_i^{(\ell+1)} = \mathbf{h}_i^{\ell} + \phi_h\Big(\mathbf{h}_i^{\ell}, \textstyle\sum_{j\in\mathcal{N}(i)}\sum_{\rho\in\mathbf{r}_{ij}} m_{ij}^{\rho}\Big), \quad \mathbf{x}_i^{(\ell+1)} = \mathbf{x}_i^{\ell} + \sum_{j\in\mathcal{N}(i)}\sum_{\rho\in\mathbf{r}_{ij}} \frac{\boldsymbol{\delta}_{ij}}{\sqrt{d_{ij}+\varepsilon}}\, s_{ij}^{\rho}. \tag{3}$$

Here, $\gamma(\cdot)$ is a 16-term radial basis function, $\mathbf{f}_{ij}$ the edge attribute vector, and each $\phi_{\{m,x\}}^{\rho}$ a two-layer MLP with relation-specific parameters ($\varepsilon = 10^{-8}$). We have four message MLPs $\phi_m^{\rho} : \mathbb{R}^{2d_h+d_f+16} \to \mathbb{R}^{d_q}$, four coordinate MLPs $\phi_x^{\rho} : \mathbb{R}^{d_q} \to \mathbb{R}^3$, and a shared node update MLP $\phi_h : \mathbb{R}^{d_h+d_q} \to \mathbb{R}^{d_h}$. With residual connections and layer normalization:

$$\mathbf{H}_{\text{ag}}^{\text{intra}} = \{\, W_{\text{ag}}^{\ell}\mathbf{h}_i^{\ell} \mid v_i \in \mathcal{V}_{\text{ag}}\},\qquad \mathbf{H}_{\text{ab}}^{\text{intra}} = \{\, W_{\text{ab}}^{\ell}\mathbf{h}_j^{\ell} \mid v_j \in \mathcal{V}_{\text{ab}}\}, \tag{4}$$

where W^{ℓ} are per-layer trainable parameters. The layer is $\mathrm{E}(3)$-equivariant by construction since only the displacement $\boldsymbol{\delta}_{ij}$ enters the coordinate update, while cross-attention operates on invariant features (proof in Appendix B.2).

MHCA layer with feed-forward network: In parallel to geometric message passing, each encoder block applies bidirectional multi-head cross-attention (MHCA) (Vaswani et al., 2017) to enable

inter-chain communication. The MHCA mechanism shown in Figure 1 **(d)** produces cross-chain context representations $\widetilde{\mathbf{H}}_{\text{ag}}$ and $\widetilde{\mathbf{H}}_{\text{ab}}$. A learnable scalar gate α balances intra-chain geometry with cross-chain context:

$$\mathbf{H}_c^{(\ell+1)} = \mathbf{H}_c^{\ell} + \mathbf{H}_c^{\text{intra}} + \alpha_c \ \text{FFN}(\widetilde{\mathbf{H}}_c), \quad c \in \{\text{ag}, \text{ab}\}, \tag{5}$$

where $\alpha_{\text{ag}}, \alpha_{\text{ab}} \in \mathbb{R}^+$ are learnable parameters, $\widetilde{\mathbf{H}} = \text{MHCA}(\mathbf{H})$, and FFN is a two-layer Feed Forward Network. The MHCA is detailed in Appendix B.3.

Decoder. The decoder refines the residue embeddings produced by the encoder and performs bipartite interaction prediction. It consists of J identical layers, each containing bidirectional MHCA with FFN and residual connections, followed by a bipartite interaction head. The final embeddings $\mathbf{H}_{\text{ag}}^J$, $\mathbf{H}_{\text{ab}}^J$ are projected into queries and keys of width d_k in both directions:

$$\mathbf{S}_{\text{ag}\to\text{ab}} = \frac{(\mathbf{H}_{\text{ag}}^J \mathbf{W}_Q^{\text{out}})(\mathbf{H}_{\text{ab}}^J \mathbf{W}_K^{\text{out}})^\top}{\sqrt{d_k}}, \qquad \mathbf{S}_{\text{ab}\to\text{ag}} = \frac{(\mathbf{H}_{\text{ab}}^J \mathbf{W}_Q^{\prime\,\text{out}})(\mathbf{H}_{\text{ag}}^J \mathbf{W}_K^{\prime\,\text{out}})^\top}{\sqrt{d_k}}. \tag{6}$$

The two score maps are fused via a learnable mixing vector $\mathbf{w} \in \mathbb{R}^2$ and bias $b \in \mathbb{R}$ to produce logits $\mathbf{Z} = \mathbf{w}^\top [\, \mathbf{S}_{\text{ag}\to\text{ab}} \ (\mathbf{S}_{\text{ab}\to\text{ag}})^\top \,] + b$, and the interaction probabilities are $\hat{\mathcal{E}}_{\text{bg}} = \sigma(\mathbf{Z}) \in \mathbb{R}^{n\times m}$.

3.3 Joint Objective

EpiFormer is trained with a joint objective that combines the primary epitope node classification loss with auxiliary terms for bipartite edge reconstruction and inter-chain geometric classification. The overall training objective is a weighted sum of these components:

$$\mathcal{L} = \lambda_{\text{node}} \mathcal{L}_{\text{node}} + \lambda_{\text{edge}} \mathcal{L}_{\text{edge}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}}. \tag{7}$$

Node Classification Loss ($\mathcal{L}_{\text{node}}$). The node classification loss supervises epitope nodes only and combines three complementary objectives to handle class imbalance and enforce structural priors:

$$\mathcal{L}_{\text{node}} = \beta_{\text{BCE}} \mathcal{L}_{\text{BCE}}^{\text{epi}} + \beta_{\text{Dice}} \mathcal{L}_{\text{Dice}}^{\text{epi}} + \beta_{\text{sparsity}} \mathcal{L}_{\text{sparsity}}^{\text{epi}}, \tag{8}$$

where $\beta_{\{\cdot\}}$ weight the different terms. The probability that node v_{ag} is an epitope is derived from the bipartite interaction matrix by aggregating across the antibody dimension:

$$(\hat{y}_{\text{ag}})_i = \frac{1}{m} \sum_{j=1}^{m} (\hat{\mathcal{E}}_{\text{bg}})_{ij}, \tag{9}$$

where $m = |\mathcal{V}_{\text{ab}}|$ is the number of antibody residues.

Class-Reweighted Binary Cross-Entropy: The primary classification loss applies positive class reweighting ($\pi_{\text{epi}} > 1$) to address the severe class imbalance in epitope prediction:

$$\mathcal{L}_{\text{BCE}}^{\text{epi}} = -\frac{1}{n} \sum_{i=1}^{n} \big[\pi_{\text{epi}} \, (y_{\text{ag}})_i \log(\hat{y}_{\text{ag}})_i + (1 - (y_{\text{ag}})_i) \log(1 - (\hat{y}_{\text{ag}})_i) \big], \tag{10}$$

Dice Loss: The Dice loss treats epitope prediction as a segmentation problem (Sudre et al., 2017), directly measuring the overlap between predicted and true epitope regions:

$$\mathcal{L}_{\text{Dice}}^{\text{epi}} = 1 - \frac{2 \sum_i (\hat{y}_{\text{ag}})_i \, (y_{\text{ag}})_i + \epsilon}{\sum_i (\hat{y}_{\text{ag}})_i + \sum_i (y_{\text{ag}})_i + \epsilon}, \tag{11}$$

where $\epsilon > 0$ is a smoothing constant.

Sparsity Regularization: The sparsity term enforces cardinality matching between predicted and true epitope counts: $\mathcal{L}_{\text{sparsity}}^{\text{epi}} = \|\hat{y}_{\text{ag}} - y_{\text{ag}}\|_1$.

Edge Prediction Loss ($\mathcal{L}_{\text{edge}}$). This loss applies positive-class-reweighted binary cross-entropy over all antigen-antibody residue pairs:

$$\mathcal{L}_{\text{edge}} = -\frac{1}{nm} \sum_{i,j} \big[\pi_{\text{edge}} \, E_{ij} \log \hat{E}_{ij} + (1 - E_{ij}) \log(1 - \hat{E}_{ij}) \big], \tag{12}$$

where $E_{ij} = (\mathcal{E}_{\text{bg}})_{ij} \in \{0, 1\}$ are ground-truth contacts (1 if residues $v_{\text{ag}}, v_{\text{ab}}$ are within 4.5Å), $\hat{E}_{ij} = (\hat{\mathcal{E}}_{\text{bg}})_{ij}$ are predicted probabilities, and π_{edge} compensates for the extreme sparsity of positives.

Auxiliary Distance Classification Loss ($\mathcal{L}_{\text{geo}}$). The auxiliary geometric term classifies inter-chain distances into discrete bins ($\{0\text{–}4, 4\text{–}8, 8\text{–}16, 16\text{–}32\}$ Å), helping the model learn distance-aware representations. The loss applies class-balanced cross-entropy over near-contact antigen-antibody residue pairs (details in Appendix B.4).

4 EXPERIMENTS

4.1 SETTINGS

Dataset. We use the AsEP dataset (Liu et al., 2024), the largest benchmark of antibody-antigen complexes for epitope prediction, retaining 1,721 complexes after preprocessing (details in Appendix A.2). Each residue is represented by a geometric feature vector encoding amino-acid type, secondary structure, solvent accessibility, and local spatial geometry (Appendix B.1).

We adopt two pre-defined splitting strategies from AsEP (Liu et al., 2024): the **epitope-ratio split** stratifies complexes by the fraction of epitope residues, balancing task difficulty across splits; the **epitope-group split** clusters complexes by epitope identity, completely excluding test epitopes from training to evaluate generalization to novel binding sites. We also report a stricter **family-based split** that assigns whole antigen and antibody families to one partition (Appendix A.5).

Table 1: Epitope-ratio split results (mean ± std over 3 seeds). Epitope-group results are in Table 6. The AlphaFold3 row runs without MSAs (Appendix C.3).

Method	AUC↑	AUPRC↑	F1↑	MCC↑	Prec.↑	Rec.↑
Epitope and paratope prediction						
EpiGraph (Choi & Kim, 2024)	.790±.004	.222±.011	.064±.032	.089±.035	.339±.052	.035±.018
EpiScan (Wang et al., 2024a)	.669±.024	.135±.019	.203±.052	.124±.043	.148±.072	.327±.041
MIPE (Wang et al., 2024b)	.827±.035	.409±.024	.337±.071	.356±.061	**.637±.146**	.229±.047
WALLE (Liu et al., 2024)	.808±.005	.206±.011	.203±.002	.208±.003	.114±.002	**.926±.006**
DiscoTope3 (Høie et al., 2024)	.821±.003	.278±.010	.251±.002	.247±.003	.149±.001	.797±.006
GraphBepi (Zeng et al., 2023)	.783±.021	.232±.014	.079±.005	.118±.024	.430±.031	.044±.076
PECAN (Pittala & Bailey-Kellogg, 2020)	.740±.022	.156±.019	.229±.041	.182±.027	.150±.052	.488±.039
Protein binding-site prediction						
EquiPocket (Zhang et al., 2023)	.761±.017	.203±.027	.202±.010	.183±.033	.116±.023	.799±.029
AtomSurf (Mallet et al., 2023)	.606±.022	.094±.048	.139±.004	.062±.007	.076±.042	.817±.072
ESMBind (Schreiber, 2023)	.768±.037	.192±.006	.208±.042	.192±.018	.120±.052	.798±.026
Protein interaction and docking						
ATProt (Gao et al., 2024)	.798±.001	.236±.003	.252±.006	.237±.002	.152±.005	.736±.027
DiffDock-PP (Ketata et al., 2023)	.735±.037	.148±.024	.192±.019	.169±.032	.109±.012	.812±.036
DiffDock (Corso et al., 2022)	.677±.011	.123±.035	.154±.016	.089±.005	.143±.027	.168±.004
Structure and affinity prediction						
Boltz-1 (Wohlwend et al., 2025)	.771±.021	.242±.032	.257±.018	.223±.003	.155±.064	.750±.017
Boltz-2 (Passaro et al., 2025)	.798±.014	.316±.038	.255±.016	.232±.019	.152±.011	.815±.007
AlphaFold3 (no MSA) (Abramson et al., 2024)	0.562	0.077	0.153	0.071	0.097	0.361
CheapNet (Lim et al., 2025)	.719±.022	.194±.029	.184±.009	.149±.017	.105±.006	.765±.017
GearBind (Cai et al., 2024)	.799±.012	.215±.031	.244±.027	.236±.002	.144±.040	.787±.037
Molecular property prediction						
EquiformerV2 (Liao et al., 2023)	.827±.011	.264±.031	.260±.016	.262±.022	.154±.013	.823±.022
***EpiFormer* (ours)**	**.924±.003**	**.493±.012**	**.482±.011**	**.464±.009**	.363±.014	.720±.012

Baseline Methods. We compare against 19 baselines spanning five categories. The first group is dedicated epitope and paratope predictors (EpiGraph, EpiScan, MIPE, WALLE, DiscoTope3, GraphBepi, PECAN). Because labeled antibody-antigen complexes are scarce, we also adapt models from related structural-biology tasks: general protein binding-site prediction (EquiPocket, AtomSurf, ESMBind), protein interaction and docking (ATProt, DiffDock-PP, DiffDock), structure and affinity prediction (Boltz-1, Boltz-2, AlphaFold3, CheapNet, GearBind), and molecular property prediction (EquiformerV2). All baselines were retrained on AsEP with their published configurations; training details are in Appendix C.

4.2 Main Results

EpiFormer leads on all primary metrics across both evaluation splits. On the epitope-ratio split (Table 1), it achieves 0.924 AUC and 0.482 F1, surpassing the next-best method MIPE (0.337 F1) by a wide margin. On the harder epitope-group split (Table 6), *EpiFormer* reaches 0.826 AUC and 0.305 F1, maintaining a clear margin over all baselines.

Several patterns emerge from these results. Antigen-only methods such as EpiGraph, GraphBepi, and DiscoTope3 reach reasonable AUC but low F1, because they cannot condition on the specific antibody and therefore predict generic binding-prone surface rather than the true epitope. Among antibody-aware methods, MIPE is the strongest baseline. Its edge comes largely from protein language model embeddings (ESM and AbLang), though its high variance ($\pm$0.071 F1) points to sensitivity to initialization. Methods from adjacent domains, such as Boltz-1/2, GearBind, and EquiformerV2, are competitive on AUC but have no mechanism for the severe class imbalance of epitope prediction, which keeps their F1 low. The docking model DiffDock falls behind much simpler baselines, and ATProt, which fuses the two chains only after encoding, trails *EpiFormer* by a wide margin.

These differences are clearest in the balance between precision and recall. Several baselines reach high recall only by labeling most residues as positive; WALLE, for instance, recovers almost every epitope residue (0.926 recall) at the cost of very low precision (0.114). The Dice loss and sparsity regularizer in *EpiFormer* instead produce balanced predictions (0.363 precision, 0.720 recall), and this balance is what drives its higher F1 and MCC.

4.3 Analysis

Backbone-Agnostic Framework. The gains come more from the interleaved cross-attention framework rather than from the choice of GNN backbone. Replacing EGNN-R with a plain GCN (Kipf & Welling, 2016), while keeping the decoder, cross-attention, and losses fixed, still reaches 0.447 F1 and already surpasses all 19 baselines (Table 2). GAT (Veličković et al., 2017) and RGCN Schlichtkrull et al. (2018) behave similarly, and every backbone we tried falls within a narrow range. The encoder choice does add a smaller but consistent increment on top, with equivariance contributing +0.018 F1 (EGNN over GAT), multi-relational encoding +0.012, and learnable coordinate updates +0.017.

Table 2: GNN backbone ablation on the epitope-ratio split. All variants share the same cross-attention decoder and losses; only the encoder differs. Eq.=equivariant message passing, MR=multi-relational edges, Coord=learned coordinate updates.

GNN	Eq.	MR	Coord	AUC	AUPRC	F1	MCC
GCN	No	No	No	.905	.397	.447	.424
GIN	No	No	No	.890	.349	.408	.369
GAT	No	No	No	.908	.399	.452	.424
RGCN	No	Yes	No	.910	.405	.454	.422
EGNN	Yes	No	Yes	.912	.421	.470	.455
EGNN-R (frozen)	Part.	Yes	No	.908	.412	.465	.441
EGNN-R (ours)	**Yes**	**Yes**	**Yes**	**.924**	**.493**	**.482**	**.464**

Architecture-Loss Co-Design. We note that the sparsity-aware losses help only when the architecture already fuses the two chains early. We added *EpiFormer*'s auxiliary losses (Dice, count regularizer, and edge prediction) to six baselines and tuned the loss weights for each by grid search (Table 3). The losses raise F1 for three of the six, WALLE (+0.129), EquiformerV2 (+0.041), and DiscoTope3 (+0.031), and lower it for the remaining three, most sharply for CheapNet (−0.104). Neither factor alone is responsible, since *EpiFormer* with a plain BCE loss reaches 0.336 F1, our objective in the best late-fusion baseline reaches 0.332, and the two together reach 0.482. A loss that was the real driver would help all architectures. Therefore, fusion timing and supervision have to be co-designed (Appendix F.4).

Geometric Features Outperform PLMs. To probe our feature choice, we swapped the geometric features for protein language model (PLM) embeddings in *EpiFormer* and, conversely, removed PLM embeddings from four baselines that rely on them. *EpiFormer* is slightly worse with PLM features than with geometric ones (0.889 vs. 0.924 AUC), whereas the baselines drop sharply once their PLM features are removed, by 0.133 AUC for DiscoTope3, 0.122 for WALLE, and 0.088 for MIPE. This contrast suggests that geometric features are sufficient for *EpiFormer*, while these baselines lean heavily on the PLM signal. A plausible reason is that epitope residues are not evolutionarily

Table 3: Architecture and loss co-design: baselines augmented with *EpiFormer*'s auxiliary losses. Tier 1 (antigen-only) receives Dice and count regularization; Tier 2 (antibody-aware) additionally receives edge prediction; Tier 3 is *EpiFormer* itself, which uses the full joint objective.

Method	Tier	Orig. F1	F1	MCC	Δ F1
DiscoTope3	1	0.251	0.282	0.263	+0.031
EquiformerV2	1	0.260	**0.301**	**0.293**	+0.041
WALLE	2	0.203	**0.332**	0.280	**+0.129**
PECAN	2	0.229	0.218	0.157	−0.011
CheapNet	2	0.184	0.080	0.154	−0.104
MIPE	2	0.337	0.261	0.230	−0.076
EpiFormer	**3**	0.336	**0.482**	**0.464**	**+0.146**

conserved (Ponomarenko & Bourne, 2007), so the conservation signal that PLMs encode is of limited use here. Epitope prediction is better described as a relational task defined by spatial proximity to a specific antibody (Sela-Culang et al., 2015), which the geometric features capture directly. The full analysis is in Appendix F.7.

Cross-Dataset Generalization. To test out-of-distribution robustness, we evaluated *EpiFormer* on three external benchmarks with no overlap with AsEP, SAbDab (Dunbar et al., 2014), CoV-AbDab (Raybould et al., 2021), and ANABAG (Grandguillaume et al., 2025) (Table 7). We compare against MIPE, which is the second-best method and the strongest antibody-aware baseline. Under zero-shot transfer, *EpiFormer* achieves a mean AUC of 0.786, outperforming MIPE (0.688) despite substantial distribution shift (Figure 5). Leave-one-dataset-out (LODO) fine-tuning further improves the AUC to 0.890, with particularly strong gains on CoV-AbDab. Full results are in Appendix D.2.

Robustness and Generalization. To test what *EpiFormer* relies on before a complex is solved, we perturbed the structural input three ways. An arbitrary rigid transform on one chain moves all output probabilities by less than 10^{-5}, and randomizing the shared reference frame during training improves both splits. Substituting a conformation the antigen adopts against a different antibody costs 0.010 AUC, and 2 Å of correlated backbone error costs 0.007. The official splits also share antigens and antibody lineages (median test-to-train antigen identity 1.00 and 0.95). *EpiFormer* reaches 0.866 AUC and 0.407 F1 by training on a stricter family-based split, whereas an identical late-fusion model reaches 0.712 and 0.196, a gain of 0.211 F1 where the antigen carries no usable prior. Full results are in Appendices E and E.4.

4.4 Ablations

We isolate the contribution of each component; full details are provided in Appendix F.

Joint Objective. The joint loss is essential. With BCE alone, *EpiFormer* achieves only 0.336 F1. Adding Dice and edge prediction yields marginal gains, whereas the sparsity regularizer is a more impactful term, improving F1 by +0.132 (Table 16). This indicates that threshold degeneration, where the model predicts all residues as positive, is a primary failure mode that the count regularizer directly corrects. We also note that adding InfoNCE on top degrades performance (−0.029 F1) due to conflicting gradients on boundary residues (Ji et al., 2024).

Interleaved Cross-Attention. Removing the encoder MHCA entirely (α=0.0) drops F1 by 0.042 and MCC by 0.051 (Table 20). Even when α is initialized at zero, which forces the model to discover cross-attention utility from scratch, it learns non-zero gate values (0.014 to 0.033 for the antibody stream), confirming that cross-modal information provides a useful learning signal. See Appendix F.6 for the full ablation.

Hyperparameter Sensitivity. F1 is largely insensitive to the exact loss weights, with most weights showing flat trends across both splits (Figure 11).

4.5 Qualitative Analysis

Interaction Map Visualization. Figure 2(a) shows the interaction maps for a representative complex (8DF5_3P). The ground-truth contact map has a sparse binding interface concentrated in a narrow

antigen region, and the distance map confirms that these contacts sit at close spatial proximity between epitope and paratope residues. The interaction matrix predicted by *EpiFormer* closely follows this pattern, with high scores aligned to the true epitope band and non-binding regions suppressed.

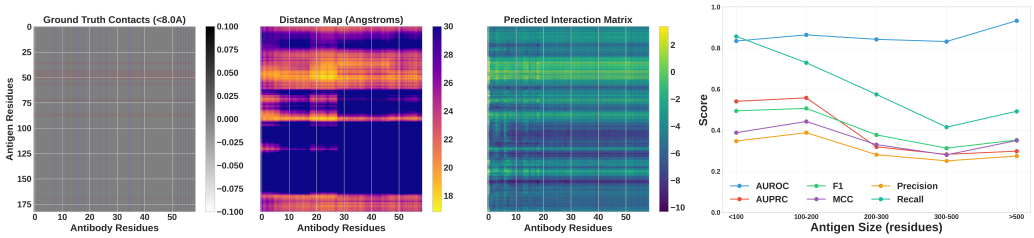


Figure 2: (a) Ground-truth contact map, pairwise distance map, and predicted interaction matrix for complex 8DF5_3P. (b) Performance across antigen size bins.

Performance by Antigen Size. Figure 2(b) stratifies performance by the antigen size. Ranking metrics (AUC, MCC, precision) remain stable across all size bins, but threshold-dependent metrics (AUPRC, F1, recall) decline for the largest antigens (>500 residues). This is expected, since as antigens grow the epitope makes up a smaller fraction of residues, which makes the classification task harder.

Asymmetry in Learned Cross-Attention. Figure 3 shows a consistent asymmetry in the learned gates. Across every encoder block and every initialization value, the antibody stream draws more on the antigen than the antigen draws on the antibody. This direction of information flow is consistent with the asymmetric roles of the two chains, since the antibody CDR loops are the variable part that conforms to a given epitope (Cagiada et al., 2025). The gates are not supervised, and one initialization is zero; hence the model finds this direction from the objective alone, which is consistent with the known roles of the two chains.

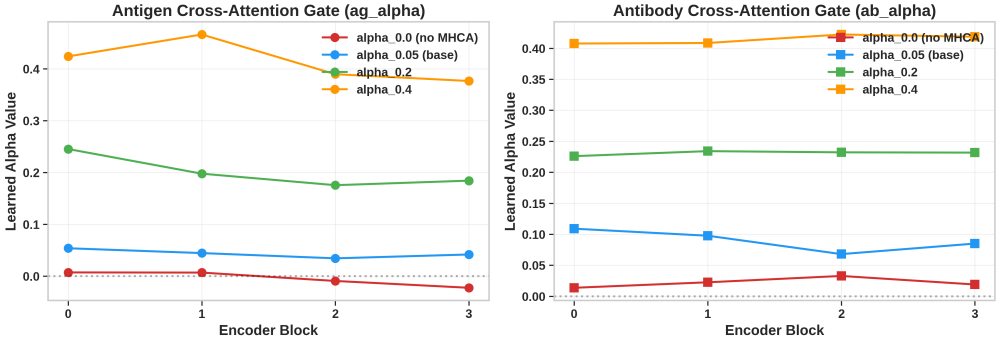


Figure 3: Learned cross-attention gate values (α) across encoder blocks for four initialization values ($\alpha_{\text{init}} \in \{0.0, 0.05, 0.2, 0.4\}$). (a) Antigen gate. (b) Antibody gate. The antibody gate consistently exceeds the antigen gate, reflecting the asymmetric roles of the two chains.

5 Conclusion

We presented *EpiFormer*, a general encoder-decoder framework for antibody-aware epitope prediction that addresses late fusion, class imbalance, and data scarcity jointly, and achieves state-of-the-art performance. Our key finding is that interleaved bidirectional cross-attention within GNN encoding layers provides consistent gains that are backbone-agnostic. We further showed that sparsity-aware training objectives must be co-designed with the fusion strategy for such task-specific applications.

Reproducibility Statement

We will make the code publicly available on GitHub and provide installation scripts with documentation to address any dependency issues. Furthermore, all the datasets used in this paper are publicly available, with the experimental details provided in the Appendix. We hope that this will support and accelerate future research and development.

Ethics Statement

This work adheres to the ICLR Code of Ethics. Our research focuses on computational methods for language and biological sequence design. We acknowledge that biological sequence design technologies could have dual-use implications. We have designed our experiments to focus on clearly beneficial applications and have not explored potentially harmful uses. All datasets used in this work are publicly available and have been previously published with appropriate ethical considerations. No human subjects were involved in this research. The computational methods developed here are general frameworks that require domain expertise and appropriate oversight for real-world applications. We encourage responsible use of these methods and recommend that practitioners consider ethical implications and potential risks when applying sequence design technologies in practice.

AI use statement

Generative AI tools were used for manuscript language editing and for assistance with code drafting and debugging. All reported numerical results were produced by the authors' computational pipelines and checked against saved outputs. The authors reviewed and revised all AI-assisted text and code and take responsibility for the final manuscript, analyses, and artifacts.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024. 6, 24, 27

Matteo Cagiada, Fabian C Spoendlin, King Ifashe, and Charlotte M Deane. Uncovering the flexibility of cdr loops in antibodies and tcrs through large-scale molecular dynamics. *bioRxiv*, pp. 2025–11, 2025. 9

Huiyu Cai, Zuobai Zhang, Mingkai Wang, Bozitao Zhong, Quanxiao Li, Yuxuan Zhong, Yanling Wu, Tianlei Ying, and Jian Tang. Pretrainable geometric graph neural network for antibody affinity maturation. *Nature communications*, 15(1):7785, 2024. 6, 25, 27

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020. 32

Sungjin Choi and Dongsup Kim. B cell epitope prediction by capturing spatial clustering property of the epitopes using graph attention network. *Scientific Reports*, 14(1):27496, 2024. 1, 2, 6, 27

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022. 2, 6, 24, 27

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42 (D1):D1140–D1146, 2014. 8, 26, 27

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. 22

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019. 16

Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, Davide Boscaini, Michael Bronstein, and Bruno Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020. 2

Ziqi Gao, Zijing Liu, Yu Li, and Jia Li. Towards stable representations for protein interface prediction. *Advances in Neural Information Processing Systems*, 37:73079–73097, 2024. 3, 6, 24, 27

Ilyas Grandguillaume, Fernando Luis Barroso da Silva, and Catherine Etchebest. Anabag: Annotated antibody–antigen data set with unique features for antibody engineering applications. *Journal of Chemical Information and Modeling*, 65(20):11454–11472, 2025. 8, 26

Fatima N Hitawala and Jeffrey J Gray. What does alphafold3 learn about antibody and nanobody docking, and what remains unsolved? In *MAbs*, volume 17, pp. 2545601. Taylor & Francis, 2025. 3, 25

Magnus Haraldson Høie, Frederik Steensgaard Gade, Julie Maria Johansen, Charlotte Würtzen, Ole Winther, Morten Nielsen, and Paolo Marcatili. Discotope-3.0: improved b-cell epitope prediction using inverse folding latent representations. *Frontiers in immunology*, 15:1322712, 2024. 1, 2, 6, 27, 31

Alissa Hummer, Brennan Abanades, and Charlotte Deane. Advances in computational structure-based antibody design. *Current Opinion in Structural Biology*, 74:102379, 2022. 1

Martin Closter Jespersen, Bjoern Peters, Morten Nielsen, and Paolo Marcatili. Bepipred-2.0: improving sequence-based b-cell epitope prediction using conformational epitopes. *Nucleic acids research*, 45(W1):W24–W29, 2017. 2

Cheng Ji, Zixuan Huang, Qingyun Sun, Hao Peng, Xingcheng Fu, Qian Li, and Jianxin Li. Regcl: rethinking message passing in graph contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8544–8552, 2024. 8, 32

Rui Jiao, Jiaqi Han, Wenbing Huang, Yu Rong, and Yang Liu. Energy-motivated equivariant pretraining for 3d molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8096–8104, 2023. 3

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020. 18, 19

Sara Joubbi, Alessio Micheli, Paolo Milazzo, Giuseppe Maccari, Giorgio Ciano, Dario Cardamone, and Duccio Medini. Antibody design using deep learning: from sequence and structure design to affinity maturation. *Briefings in Bioinformatics*, 25(4):bbae307, 2024. 1

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. 18

Mohamed Amine Ketata, Cedrik Laue, Richard Maser, and Gabriele Corso. Diffdock-pp: Rigid protein-protein docking with diffusion models. In *ICLR 2023 Workshop on Machine Learning for Drug Discovery*, 2023. 3, 6, 24, 27

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 7

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022. 2

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023. 2, 6, 25, 27

Hyukjun Lim, Sun Kim, and Sangseon Lee. Cheapnet: Cross-attention on hierarchical representations for efficient protein-ligand binding affinity prediction. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 6, 25, 27

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. 34

Chunan Liu, Lilian Denzler, Yihong Chen, Andrew Martin, and Brooks Paige. AsEP: Benchmarking deep learning methods for antibody-specific epitope prediction. *arXiv preprint arXiv:2407.18184*, 2024. 1, 2, 6, 27

Vincent Mallet, Souhaib Attaiki, Yangyang Miao, Bruno Correia, and Maks Ovsjanikov. Atomsurf: Surface representation for learning on protein structures. *arXiv preprint arXiv:2309.16519*, 2023. 2, 6, 24, 27

Richard Norman, Francesco Ambrosetti, Alexandre Bonvin, Lucy Colwell, Sebastian Kelm, Sandeep Kumar, and Konrad Krawczyk. Computational approaches to therapeutic antibody design: established methods and emerging trends. *Briefings in bioinformatics*, 21(5):1549–1567, 2020. 1

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025. 3, 6, 25, 27

Srivamshi Pittala and Chris Bailey-Kellogg. Learning context-aware structural representations to predict antigen and antibody binding interfaces. *Bioinformatics*, 36(13):3996–4003, 2020. 1, 2, 6, 27

Julia V Ponomarenko and Philip E Bourne. Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Structural Biology*, 7(64), 2007. 8, 34

Matthew IJ Raybould, Aleksandr Kovaltsuk, Claire Marks, and Charlotte M Deane. Cov-abdab: the coronavirus antibody database. *Bioinformatics*, 37(5):734–735, 2021. 8, 26, 27

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. 2, 4

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pp. 593–607. Springer, 2018. 7

Amelie Schreiber. ESMBind and QBind: LoRA, QLoRA, and ESM-2 for predicting binding sites and post translational modification. *BioRxiv*, pp. 2023–11, 2023. 6, 24, 27

Inbal Sela-Culang, Yanay Ofran, and Bjoern Peters. Antibody specific epitope prediction—emergence of a new paradigm. *Current opinion in virology*, 11:98–102, 2015. 8

Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539, 2011. 16

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *International Workshop on Deep Learning in Medical Image Analysis*, pp. 240–248. Springer, 2017. 5

Serbulent Unsal, Benjamin Holland, Inci Sardag, and Emel Timucin. Assessing the performance of af2 and af3-implementations on antibody-antigen complexes. *bioRxiv*, pp. 2025.07.25.666870, 2025. doi: 10.1101/2025.07.25.666870. Version 1; v2 is retitled "Confidence Scoring for AI-Predicted Antibody-Antigen Complexes: AntiConf as a Precision-driven Metric". 3, 25

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 7

Chuan Wang, Jiangyuan Wang, Wenjun Song, Guanzheng Luo, and Taijiao Jiang. EpiScan: accurate high-throughput mapping of antibody-specific epitopes using sequence information. *NPJ Systems Biology and Applications*, 10(1):101, 2024a. 2, 6, 27

Zhiwei Wang, Yongkang Wang, and Wen Zhang. Improving paratope and epitope prediction by multi-modal contrastive learning and interaction informativeness estimation. *arXiv preprint arXiv:2405.20668*, 2024b. 1, 2, 6, 27

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pp. 2024–11, 2025. 3, 6, 24, 27

Lirong Wu, Haitao Lin, Yufei Huang, Zhangyang Gao, Cheng Tan, Yunfan Liu, Tailin Wu, and Stan Z Li. Relation-aware equivariant graph networks for epitope-unknown antibody design and specificity optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 895–904, 2025. 18, 19, 20, 25, 31

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. 35

Yuansong Zeng, Zhuoyi Wei, Qianmu Yuan, Sheng Chen, Weijiang Yu, Yutong Lu, Jianzhao Gao, and Yuedong Yang. Identifying b-cell epitopes using alphafold2 predicted structures and pretrained language model. *Bioinformatics*, 39(4):btad187, 2023. 1, 2, 6, 27

Yang Zhang, Zhewei Wei, Ye Yuan, Chongxuan Li, and Wenbing Huang. Equipocket: an e (3)-equivariant geometric graph neural network for ligand binding site prediction. *arXiv preprint arXiv:2302.12177*, 2023. 2, 6, 24, 27

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022. 2, 19

APPENDIX

A Data and experimental protocol **16**

A.1 The AsEP dataset . . . 16
A.2 Preprocessing . . . 16
A.3 The structural input setting . . . 16
A.4 Train and test similarity under the official splits . . . 17
A.5 Construction of a family-based split . . . 17
A.6 External datasets and distribution shift . . . 17

B Model and implementation details **17**

B.1 Residue graph construction . . . 18
B.2 E(3)-equivariance of the EGNN-R layer . . . 19
B.3 Multi-head cross-attention with feed-forward network . . . 21
B.4 The auxiliary distance classification loss . . . 21
B.5 Training setup and hyperparameters . . . 21

C Baselines **23**

C.1 Baseline descriptions . . . 24
C.2 Architectural and loss component comparison . . . 25
C.3 Converting co-folding and docking predictions into epitope labels . . . 25

D Extended results **26**

D.1 Epitope-group split . . . 26
D.2 Cross-dataset evaluation . . . 26
D.3 Per-complex metric distributions . . . 26
D.4 Stratification by X-ray resolution . . . 26
D.5 Stratification by antigen size . . . 28

E Robustness of the structural input **28**

E.1 Randomizing the local coordinate frames . . . 29
E.2 Transfer to a donor antigen conformation . . . 30
E.3 Injected structural error . . . 30
E.4 Generalization under the family-based split . . . 31

F Ablations **31**

F.1 Encoder backbone . . . 31
F.2 RAAD and GearNet residue graphs . . . 31
F.3 The joint objective . . . 32
F.4 Separating the architecture from the objective . . . 33
F.5 Pooling strategy . . . 33

F.6 Cross-attention gate initialization 34
F.7 Protein language model features 34
F.8 Loss hyperparameter sensitivity 35

G Interpretability analyses **35**
G.1 Representation quality 35
G.2 Encoder and decoder cross-attention maps 36
G.3 Coordinate refinement 36

H Discussion **37**
H.1 Practical scope relative to co-folding and docking 37

A Data and experimental protocol

In this section, we provide the dataset, the input setting, the similarity between the official splits, and a new stricter split.

A.1 The AsEP dataset

Figure 4 summarizes the dataset. Antigen surface graphs average 288 residues (range 20–1500+), antibody CDR graphs average 61 residues, and the mean epitope size is 19 residues with a mean of 44 bipartite edges per complex.

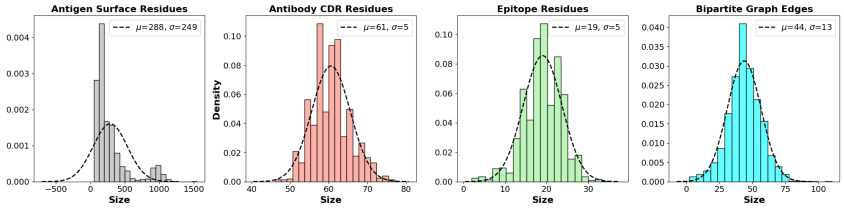


Figure 4: Size distributions of antigen surface residues, antibody CDR residues, epitope residues, and bipartite edges in the AsEP dataset.

A.2 Preprocessing

For each complex, we first separated the paired antigen and antibody chains into individual structure files. We then performed sequence-structure alignment using Clustal Omega (Sievers et al., 2011) to establish correspondence between SEQRES (complete sequence) and ATOMSEQ (resolved atoms) records. This alignment generated binary masks that enable reliable mapping of sequences to structural residues while preserving the native crystallographic ordering.

For antibody chains, we applied the alignment masks to reindex heavy (H) and light (L) chains by removing insertion codes to enforce consecutive 1-based residue numbering required for graph construction. Antigen chains underwent similar processing to maintain parity between sequences and structures. This step ensures that each residue in the protein sequence corresponds exactly to its structural counterpart during the graph representation. Then, we applied solvent-accessibility filters to retain only antigen surface residues, using the original AsEP seqres2surf masks to define the node set for antigen residue graphs. The binary epitope labels were projected onto the surface ATOMSEQ via alignment masks, while paratope labels were preserved for antibody residue nodes. This surface filtering step prevents non-surface residues from confounding epitope supervision while maintaining all necessary information for cross-chain interaction modeling.

We excluded two complexes (5nj6_0P and 5ies_0P) from the AsEP dataset due to sequence alignment inconsistencies and unresolved residues, with the final dataset containing 1,721 complexes. The contact distribution between residues in the bipartite graph had a mean of 43.7 contacts with a standard deviation of 12.8. Additionally, the dataset includes 641 unique antigens and 973 epitope groups, highlighting the diversity and complexity of the antibody-antigen interactions captured in the AsEP dataset.

Finally, we used these preprocessed structures to generate HeteroData objects for the multi-relational graphs using PyTorch Geometric (Fey & Lenssen, 2019). Each residue node is attributed a geometric feature vector encoding amino acid type, backbone geometry, distances, and local coordinate frames (detailed in Section B.1).

A.3 The structural input setting

All the antigen-antibody pairs in the AsEP dataset have known complex structures, and the chains are extracted from those bound complexes. The task exists only because epitope labels can come from solved complexes alone. Hence, any supervised benchmark is restricted by construction to pairs whose complex is already known. The entire PDB yields only the 1,721 usable complexes in AsEP,

spanning 641 unique antigens after redundancy filtering. Set against antibody repertoires that contain billions of paired sequences, that is a very small fraction, and the vast majority of characterized antibodies, therapeutic candidates included, have no solved complex with their target.

The model itself is unaware that the chains come from a bound complex, and the relative position of the two chains does not reach it in the forward pass. More precisely, the message passing runs on intra-chain edges only, the cross-attention mask carries chain membership, and the decoder receives two embedding matrices with no coordinates (analogous to unbound structures). The inter-chain distances appear only as targets of an auxiliary loss that is discarded at inference (Appendix B.4).

Applying an arbitrary rigid transform to one chain moves all output probabilities by less than 10^{-5}, which confirms that the model is not using the docked pose. A weaker dependence on the shared reference frame does remain, and we measure it in Appendix E.1. All baseline methods use the same bound-extracted AsEP chains, which matches the comparison in Table 1. We further report the experiments under this setting in Appendix E.

A.4 Train and test similarity under the official splits

Neither official split controls for homology between its partitions. For each test complex of both splits, we measured the maximum antigen SEQRES identity, the maximum CDR identity and the maximum antigen TM-score against the training set, using MMseqs2 at its most sensitive setting and Foldseek in TM-align mode.

Table 4: Similarity of each test complex to the training set. Each entry is the median over test complexes, with the share above 90% in parentheses.

Median against training	**epitope-ratio** (n=170)	**epitope-group** (n=169)
Antigen sequence identity	1.00 (79%)	0.95 (53%)
Antibody CDR identity	0.94 (58%)	0.86 (37%)
Antigen structural TM-score	0.99 (82%)	0.95 (69%)

Antigens, antibody lineages and close structural homologs are therefore all shared, and not one of the 170 epitope-ratio test complexes falls below 70% on both the antigen and the antibody side. This is a property of AsEP rather than of our choice between its two splits, and all baseline methods in Table 1 share the same splits. The family-based split of Appendix A.5 addresses this.

A.5 Construction of a family-based split

To create a stricter split, we link two antigen-antibody complexes when their antigen SEQRES reach 30% identity over at least half the query, or when their CDRs reach 90% identity. In both cases, we assign each connected component whole to one partition. This yields 251 families and a 1205 / 172 / 344 split, in which no test complex exceeds 27.5% antigen identity or 89.6% CDR identity against anything in the training set.

The families are defined on sequence alone, which leaves a median test-to-train antigen TM-score of 0.70. The label statistics stay comparable, with a mean epitope ratio of 0.103 against 0.1 on epitope-ratio, where a shift in class balance does not distort F1 and MCC. The results are in Appendix E.4.

A.6 External datasets and distribution shift

B Model and implementation details

In this section, we provide the residue graph, the encoder layer, the cross-attention module, the auxiliary objective and the training setup needed to reimplement *EpiFormer*.

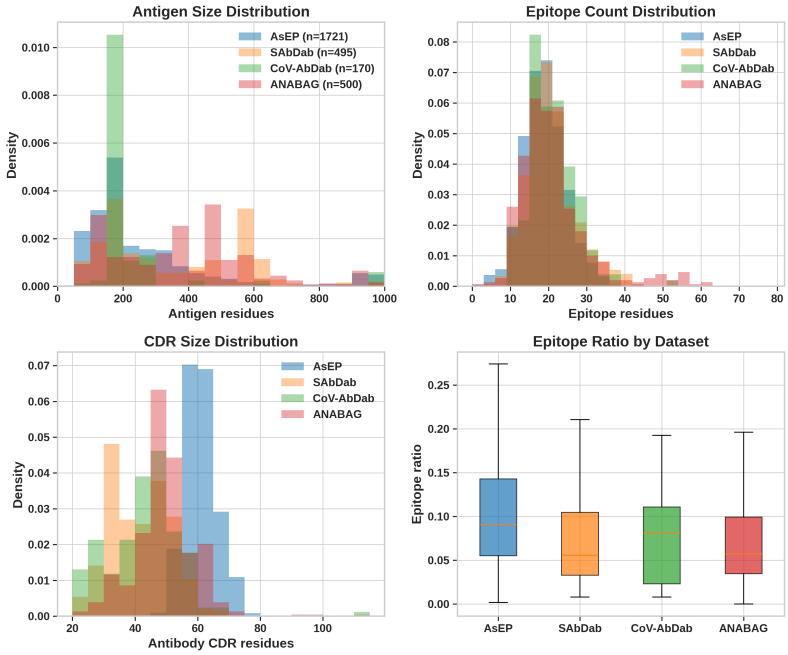


Figure 5: Distribution shift between AsEP and three external benchmarks across (a) antigen size, (b) epitope count, (c) CDR size, and (d) epitope ratio. These differences explain the zero-shot calibration gap recovered by LODO fine-tuning.

B.1 Residue Graph Construction

B.1.1 Node Features

Each residue node in our protein graph incorporates two complementary information sources that together provide a rich representation of both local structural properties and evolutionary context:

Local geometry & physicochemistry. Each residue $v_i \in \mathcal{V}$ is annotated with a 105-dimensional geometric and biochemical feature vector $\mathbf{h}_i^{\text{geo}} \in \mathbb{R}^{d_{\text{geo}}}$ that encodes the type, position, distance, direction, angle, and orientation of each residue. Such residue-level descriptors are widely employed in diverse protein-related studies in structural bioinformatics (Wu et al., 2025; Jing et al., 2020; Jumper et al., 2021). This vector is constructed as follows:

$$\mathbf{h}_i^{\text{geo}} = \left[E_{\text{type}}(v_i),\ E_{\text{pos}}(i),\ \sin(\eta_i),\ \cos(\eta_i),\ \text{RBF}(\|\mathbf{x}_{i,C_\alpha} - \mathbf{x}_{i,\xi}\|),\ Q_i^\top \frac{\mathbf{x}_{i,\xi} - \mathbf{x}_{i,C_\alpha}}{\|\mathbf{x}_{i,\xi} - \mathbf{x}_{i,C_\alpha}\|} \right], \quad (13)$$

where:

- E_{type}: Embedding for amino acid residue type (e.g., arginine, glycine).
- E_{pos}: Positional encoding of residue index in the sequence, enabling the model to distinguish between identical amino acids based on their sequence context. This positional information is crucial for understanding long-range dependencies and structural motifs, as amino acids at different sequence positions (N-terminus vs. C-terminus, loop regions vs. secondary structures) often play different functional roles even if they are the same amino acid type.
- η_i: Local backbone geometry encoded through six fundamental angles that determine how the protein chain folds at each residue v_i and are encoded by their sine and cosine (12 scalars). Bond angles (α_i, β_i, γ_i) describe the geometric constraints of covalent bonds, while dihedral angles (ψ_i, ϕ_i, ω_i) capture the rotational freedom that gives rise to secondary structures like helices and sheets.
- $\text{RBF}(\cdot)$: Radial basis function encoding distances between C_α and other backbone atoms ($\xi \in \{C_\beta, N, O\}$), with each distance represented by 16 Gaussian basis functions.

- $Q_i^\top \mathbf{u}_i$: Here, $Q_i \in \mathbb{R}^{3\times3}$ is the orthonormal rotation matrix defining the local coordinate system constructed from the C_α, C_β, and N atoms of residue i, and $\mathbf{u}_i = [\mathbf{u}_i^1, \mathbf{u}_i^2, \mathbf{u}_i^3] \in \mathbb{R}^{3\times3}$ contains the normalized direction vectors between these atoms (e.g., $\mathbf{u}_i^1 = \frac{\mathbf{x}_{i,C_\beta} - \mathbf{x}_{i,C_\alpha}}{\|\mathbf{x}_{i,C_\beta} - \mathbf{x}_{i,C_\alpha}\|}$). The matrix product $Q_i^\top \mathbf{u}_i$ transforms these direction vectors into the local coordinate frame and is flattened to yield a 9-dimensional feature vector. Note that the oxygen atom is stored in the coordinate matrix for other calculations (like the RBF distance features), but isn't used for the local coordinate frame construction.
- The coordinates are held in a 3×4 matrix, which is used in the calculation of node and edge features.

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{i,N} & \mathbf{x}_{i,C_\alpha} & \mathbf{x}_{i,C_\beta} & \mathbf{x}_{i,O} \end{bmatrix} \in \mathbb{R}^{3\times4}, \quad \text{where} \quad \mathbf{x}_{i,\xi} \in \mathbb{R}^3$$

B.1.2 Edge features

We compute a 100-dimensional edge feature vector $\mathbf{f}_{i,j} \in \mathbb{R}^{d_f}$ that describes the spatial and sequential relationship between two residues v_i and v_j. This vector integrates multiple complementary descriptors to provide a rich representation of inter-residue interactions (Jing et al., 2020) and is defined as follows:

$$\mathbf{f}_{i,j} = \left\{ E_{\text{type}}(e_{i,j}), E_{\text{pos}}(i-j), \text{RBF}(\|\mathbf{x}_{i,C_\alpha} - \mathbf{x}_{j,\xi}\|), Q_i^\top \frac{\mathbf{x}_{j,\xi} - \mathbf{x}_{i,C_\alpha}}{\|\mathbf{x}_{j,\xi} - \mathbf{x}_{i,C_\alpha}\|}, q\left(Q_i^\top Q_j\right) \mid \xi \right\}, \tag{14}$$

where $E_{\text{type}}(e_{i,j})$ is the one-hot encoding of relations $\mathbf{r}_{i,j}$ of length 4 between two residues, and the positional encoding $E_{\text{pos}}(i-j)$ encodes the relative sequential position sinusoidally to 16 scalars. The third and fourth terms are distance and direction encodings of four backbone atoms ξ in residue v_j in the local coordinate frame Q_i. These four inter-residue distances $\{d(C_\alpha, C_\beta), d(C_\alpha, N), d(C_\alpha, O), d(C_\alpha, C_\alpha)\}$ are each represented by 16 Gaussian basis functions. The last term $q\left(Q_i^\top Q_j\right)$ is the quaternion representation $q(\cdot)$ of $Q_i^\top Q_j$. By integrating sequence position, local geometry, and orientation, the model understands the residue identity from global pose and enables robust generalization across structures. These node and edge features are visualized in Figure 6(a).

B.1.3 Edge Relations

Since spatial proximity between residues alone cannot capture hydrogen bonding's directional specificity or electrostatic complementarity's charge-based selectivity, we use multi-relational edges to capture distinct interaction types (Zhang et al., 2022). By treating each relation separately, the model learns complex interaction patterns within the protein. Hence, to expand the contexts of these interactions, we divide the edges into four different types of relations $\mathcal{R} = \{\rho_1, \rho_2, \rho_3, \rho_4\}$, including *(i)* **sequential relations** ρ_1 and ρ_2 between two residues with relative sequential distance equal to 1 (peptide bond) and 2 (short-range torsion coupling); *(ii)* **spatial relations** between residues that are from the same component and spatially connected due to K-nearest neighbors (relation ρ_3 that captures local packing shell) or with a Euclidean distance less than 8Å (relation ρ_4) capturing medium-range contact between residues within the protein structure (Wu et al., 2025).

To illustrate the importance of edge relations, consider a discontinuous epitope that spans two antigen loops. The sequential edges (ρ_1, ρ_2) maintain the structural integrity of each loop, while the spatial edges (ρ_3, ρ_4) capture the three-dimensional proximity between residues from different loops, allowing the model to represent how distant sequence regions come together to form a cohesive binding interface. We provide a schematic of edge relations in Fig. 6 (b), where each edge $e_{i,j} \in \mathcal{E}$ is associated with a set of relations $\mathbf{r}_{i,j} \in \mathcal{R}$. Besides, two relations ρ_1 (with sequence distance equal to 1) can derive a relation ρ_2 (with sequence distance equal to 2), while an edge may connect two nodes (residues) due to both relations ρ_3 and ρ_4.

B.2 E(3)-equivariance of the EGNN-R layer

Theorem B.1. *Consider the EGNN-R layer in §3.2 with updates*

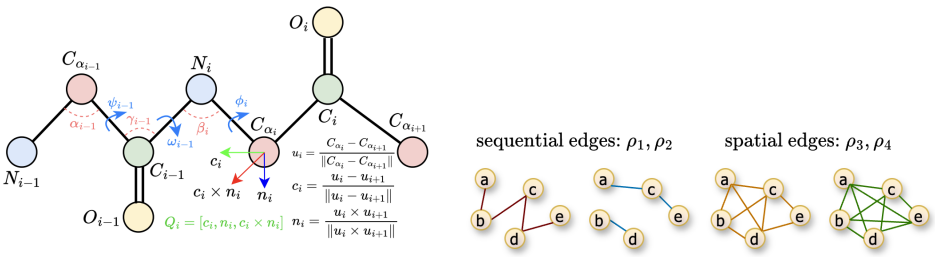


Figure 6: (a) Node and edge features encoding position, distance, direction, angle, and orientation (Figure credit: (Wu et al., 2025)). (b) Four edge relations (sequential ρ_1, ρ_2; spatial ρ_3, ρ_4). To avoid complexity, we visualize only some edges.

$$m_{ij}^{\rho} = \phi_m^{\rho}\big(\mathbf{h}_i^{\ell}, \mathbf{h}_j^{\ell}, \gamma(d_{ij}), \mathbf{f}_{ij}\big), \tag{15}$$

$$s_{ij}^{\rho} = \phi_x^{\rho}\big(m_{ij}^{\rho}\big), \tag{16}$$

$$\mathbf{h}_i^{(\ell+1)} = \mathbf{h}_i^{\ell} + \phi_h\Big(\mathbf{h}_i^{\ell}, \sum_{j\in\mathcal{N}(i)} \sum_{\rho\in\mathbf{r}_{ij}} m_{ij}^{\rho}\Big), \tag{17}$$

$$\mathbf{x}_i^{(\ell+1)} = \mathbf{x}_i^{\ell} + \sum_{j\in\mathcal{N}(i)} \sum_{\rho\in\mathbf{r}_{ij}} \frac{\boldsymbol{\delta}_{ij}}{\sqrt{d_{ij}+\varepsilon}}\, s_{ij}^{\rho}, \tag{18}$$

where $\boldsymbol{\delta}_{ij} = \mathbf{x}_i^{\ell} - \mathbf{x}_j^{\ell}$, $d_{ij} = \|\boldsymbol{\delta}_{ij}\|_2^2$, *and* $\varepsilon > 0$. *Assume: (i) node features* $\mathbf{h}_i^{\ell} \in \mathbb{R}^{d_h}$ *are scalar channels, (ii)* $\mathbf{h}_{ij}$ *and* $\mathbf{r}_{ij}$ *are categorical and independent of coordinates, (iii)* γ *is any scalar function of* d_{ij}*, (iv) each* $\phi_{\{m,x\}}^{\rho}$ *is an MLP from scalars to scalars. Let the* $E(3)$ *action be* $g = (R, t)$ *with* $R \in O(3)$ *and* $t \in \mathbb{R}^3$*, acting as* $\mathbf{x}_i^{\ell} \mapsto R\mathbf{x}_i^{\ell} + t$ *and* $\mathbf{h}_i^{\ell} \mapsto \mathbf{h}_i^{\ell}$*. Then the layer is* $E(3)$*-equivariant:*

$$\big\{\, \mathbf{x}_i^{\ell}, \mathbf{h}_i^{\ell} \,\big\}_{i=1}^{n} \mapsto \big\{\, R\mathbf{x}_i^{\ell} + t,\ \mathbf{h}_i^{\ell} \,\big\}_{i=1}^{n} \implies \big\{\, \mathbf{x}_i^{(\ell+1)}, \mathbf{h}_i^{(\ell+1)} \,\big\}_{i=1}^{n} \mapsto \big\{\, R\mathbf{x}_i^{(\ell+1)} + t,\ \mathbf{h}_i^{(\ell+1)} \,\big\}_{i=1}^{n}.$$

Consequently, any stack of such layers is $E(3)$*-equivariant by composition.*

Proof. Let $g = (R, t) \in E(3)$ act as stated. Edge data $\mathbf{f}_{ij}$ and $\mathbf{r}_{ij}$ are unchanged.

Invariants. Relative displacement and distance transform as

$$\boldsymbol{\delta}_{ij} \mapsto R\boldsymbol{\delta}_{ij}, \qquad d_{ij} = \|\boldsymbol{\delta}_{ij}\|^2 \mapsto \|R\boldsymbol{\delta}_{ij}\|^2 = d_{ij}. \tag{19}$$

Hence d_{ij}, $\gamma(d_{ij})$, and $(d_{ij}+\varepsilon)^{-1/2}$ are invariant scalars.

Scalar messages and coefficients. Each message $m_{ij}^{\rho} = \phi_m^{\rho}(\mathbf{h}_i^{\ell}, \mathbf{h}_j^{\ell}, \gamma(d_{ij}), \mathbf{f}_{ij})$ depends only on scalars that are invariant under g, so m_{ij}^{ρ} is invariant. Then $s_{ij}^{\rho} = \phi_x^{\rho}(m_{ij}^{\rho})$ is also invariant.

Feature update. The update

$$\mathbf{h}_i^{(\ell+1)} = \mathbf{h}_i^{\ell} + \phi_h\Big(\mathbf{h}_i^{\ell}, \sum_{j\in\mathcal{N}(i)} \sum_{\rho\in\mathbf{r}_{ij}} m_{ij}^{\rho}\Big) \tag{20}$$

uses only invariant scalars, so $\mathbf{h}_i^{(\ell+1)}$ is invariant. This matches the scalar action on features.

Coordinate update. The increment

$$\Delta\mathbf{x}_i = \sum_{j\in\mathcal{N}(i)} \sum_{\rho\in\mathbf{r}_{ij}} \frac{\boldsymbol{\delta}_{ij}}{\sqrt{d_{ij}+\varepsilon}}\, s_{ij}^{\rho} \tag{21}$$

is a sum of relative vectors scaled by invariant scalars. Under g each term becomes

$$\frac{\boldsymbol{\delta}_{ij}}{\sqrt{d_{ij}+\varepsilon}}\, s_{ij}^{\rho} \ \mapsto\ \frac{R\boldsymbol{\delta}_{ij}}{\sqrt{d_{ij}+\varepsilon}}\, s_{ij}^{\rho} = R\Big(\frac{\boldsymbol{\delta}_{ij}}{\sqrt{d_{ij}+\varepsilon}}\, s_{ij}^{\rho}\Big), \tag{22}$$

so $\Delta\mathbf{x}_i \mapsto R\,\Delta\mathbf{x}_i$. Therefore

$$\mathbf{x}_i^{(\ell+1)} = \mathbf{x}_i^{\ell} + \Delta\mathbf{x}_i \;\mapsto\; R\mathbf{x}_i^{\ell} + t + R\Delta\mathbf{x}_i = R\big(\mathbf{x}_i^{\ell} + \Delta\mathbf{x}_i\big) + t = R\mathbf{x}_i^{(\ell+1)} + t. \tag{23}$$

Composition. The composition of equivariant maps is equivariant. Hence, any stack of EGNN-R layers is $E(3)$-equivariant. □

B.3 Multi-head cross-attention with feed-forward network

The bidirectional multi-head cross-attention mechanism enables information exchange between antigen and antibody chains. Let n_{head} be the number of heads with per-head width $d_a = d_h/n_{\text{head}}$. For layer ℓ, independent linear projections produce queries, keys, and values:

$$\mathbf{Q}_{\text{ag}}^{\ell} = \mathbf{H}_{\text{ag}}^{(\ell-1)}\mathbf{W}_{\text{ag}}^{Q(\ell)}, \tag{24}$$

$$\mathbf{K}_{\text{ab}}^{\ell} = \mathbf{H}_{\text{ab}}^{(\ell-1)}\mathbf{W}_{\text{ab}}^{K(\ell)}, \tag{25}$$

$$\mathbf{V}_{\text{ab}}^{\ell} = \mathbf{H}_{\text{ab}}^{(\ell-1)}\mathbf{W}_{\text{ab}}^{V(\ell)}, \tag{26}$$

with analogous expressions for the reverse direction. After reshaping to n_{head} heads of width d_a, scaled dot-product attention computes the affinity matrices:

$$\mathbf{A}_{\text{ag}\leftarrow\text{ab}}^{\ell} = \text{softmax}\Big(\tfrac{1}{\sqrt{d_h}}\,\mathbf{Q}_{\text{ag}}^{\ell}{\mathbf{K}_{\text{ab}}^{\ell}}^{\top} + \mathbf{M}\Big), \tag{27}$$

where $\mathbf{M}$ is a batch mask (applied only in decoder) that assigns $-\infty$ to residue pairs from different complexes. The resulting context vectors are:

$$\widetilde{\mathbf{H}}_{\text{ag}}^{\ell} = \big[\mathbf{A}_{\text{ag}\leftarrow\text{ab}}^{\ell}\mathbf{V}_{\text{ab}}^{\ell}\big]\mathbf{W}_{O,\text{ag}}^{\ell}, \tag{28}$$

$$\widetilde{\mathbf{H}}_{\text{ab}}^{\ell} = \big[\mathbf{A}_{\text{ab}\leftarrow\text{ag}}^{\ell}\mathbf{V}_{\text{ag}}^{\ell}\big]\mathbf{W}_{O,\text{ab}}^{\ell}. \tag{29}$$

Each direction then applies a feed-forward network $\text{FFN}(\mathbf{x}) = \mathbf{W}_2\,\sigma(\mathbf{W}_1\mathbf{x}+\mathbf{b}_1)+\mathbf{b}_2$ with dropout, residual connections, and layer normalization.

B.4 The auxiliary distance classification loss

Let $\mathcal{M} = \{(i,j) : d_{ij} \le D_{\max}\}$ be the set of antigen-antibody residue pairs within the maximum distance cutoff, where d_{ij} is the Euclidean distance between residues i and j. The bins are defined by distances $\{d_0, d_1, d_2, d_3, d_4\} = \{0, 4, 8, 16, 32\}$ Å, creating $B = 4$ bins:

$$b(i,j) = \arg\max_{b\in\{1,\dots,4\}} \mathbf{1}[\,d_{b-1} \le d_{ij} < d_b\,]. \tag{30}$$

The network predicts per-pair distance logits $\mathbf{\Delta}_{ij} \in \mathbb{R}^5$, but only the first $B = 4$ components $\widehat{\mathbf{\Delta}}_{ij} \in \mathbb{R}^4$ are used for pairs in $\mathcal{M}$, ignoring the "far" class beyond $D_{\max} = 32$ Å. The class probabilities are:

$$p_{ijb} = \frac{\exp(\widehat{\Delta}_{ijb})}{\sum_{b'=1}^{4}\exp(\widehat{\Delta}_{ijb'})}. \tag{31}$$

The loss combines class balancing with distance weighting:

$$\mathcal{L}_{\text{geo}} = -\frac{1}{|\mathcal{M}|}\sum_{(i,j)\in\mathcal{M}} w_{ij}\sum_{b=1}^{4}\alpha_b\,\mathbf{1}[\,b(i,j)=b\,]\,\log p_{ijb}, \tag{32}$$

where $\alpha_b > 0$ are class-balance weights computed from empirical bin frequencies within $\mathcal{M}$ and $w_{ij} > 0$ are distance weights inversely proportional to d_{ij}, normalized to unit mean over $\mathcal{M}$.

B.5 Training setup and hyperparameters

The model is trained with an Adam optimizer and a ReduceLROnPlateau learning-rate schedule with decoupled weight decay. The learning rate is selected from the sweep-defined range and fixed

Algorithm 1: *EpiFormer*: High-Level Architecture

Input: Antigen graph $\mathcal{G}_{\text{ag}}$ and antibody graph $\mathcal{G}_{\text{ab}}$ with coordinates $\mathbf{X}$, geometric features $\mathbf{h}^{\text{geo}}$
Output: Bipartite interaction matrix $\hat{\mathcal{E}}_{\text{bg}} \in [0,1]^{n \times m}$

// **Feature Initialization**
foreach *chain* $\in \{ag, ab\}$ **do**
  Project geometric features to hidden dimension;
  $\mathbf{h}_i^0 \leftarrow \mathbf{W}_{\text{geo}}\, \mathbf{h}_i^{\text{geo}}$ for each residue i;
end
// **Encoder: Parallel Processing**
for *layer* $\ell = 1$ **to** L **do**
  // Intra-chain geometric message passing
  $(\mathbf{H}_{\text{ag}}^{\text{intra}}, \mathbf{X}^{\text{ag}}) \leftarrow \text{EGNN-R}(\mathcal{G}_{\text{ag}}, \mathbf{H}_{\text{ag}}^{(\ell-1)}, \mathbf{X}^{\text{ag}})$;
  $(\mathbf{H}_{\text{ab}}^{\text{intra}}, \mathbf{X}^{\text{ab}}) \leftarrow \text{EGNN-R}(\mathcal{G}_{\text{ab}}, \mathbf{H}_{\text{ab}}^{(\ell-1)}, \mathbf{X}^{\text{ab}})$;
  // Inter-chain cross-attention
  $\widetilde{\mathbf{H}}_{\text{ag}} \leftarrow \text{MHCA}(\mathbf{H}_{\text{ag}}^{\text{intra}}, \mathbf{H}_{\text{ab}}^{\text{intra}}, \mathbf{H}_{\text{ab}}^{\text{intra}})$;
  $\widetilde{\mathbf{H}}_{\text{ab}} \leftarrow \text{MHCA}(\mathbf{H}_{\text{ab}}^{\text{intra}}, \mathbf{H}_{\text{ag}}^{\text{intra}}, \mathbf{H}_{\text{ag}}^{\text{intra}})$;
  // Combine intra-chain and cross-chain information
  $\mathbf{H}_{\text{ag}}^{\ell} \leftarrow \mathbf{H}_{\text{ag}}^{(\ell-1)} + \mathbf{H}_{\text{ag}}^{\text{intra}} + \alpha_{\text{ag}}\, \text{FFN}(\widetilde{\mathbf{H}}_{\text{ag}})$;
  $\mathbf{H}_{\text{ab}}^{\ell} \leftarrow \mathbf{H}_{\text{ab}}^{(\ell-1)} + \mathbf{H}_{\text{ab}}^{\text{intra}} + \alpha_{\text{ab}}\, \text{FFN}(\widetilde{\mathbf{H}}_{\text{ab}})$;
end
// **Decoder: Cross-Attention Refinement**
Initialize decoder embeddings: $\mathbf{H}_{\text{ag}}^{\text{dec}} \leftarrow \mathbf{H}_{\text{ag}}^{L}, \mathbf{H}_{\text{ab}}^{\text{dec}} \leftarrow \mathbf{H}_{\text{ab}}^{L}$;
for *layer* $\ell = 1$ **to** L **do**
  // Inter-chain cross-attention
  $\widetilde{\mathbf{H}}_{\text{ag}}^{\text{dec}} \leftarrow \text{MHCA}(\mathbf{H}_{\text{ag}}^{\text{dec}}, \mathbf{H}_{\text{ab}}^{\text{dec}}, \mathbf{H}_{\text{ab}}^{\text{dec}})$;
  $\widetilde{\mathbf{H}}_{\text{ab}}^{\text{dec}} \leftarrow \text{MHCA}(\mathbf{H}_{\text{ab}}^{\text{dec}}, \mathbf{H}_{\text{ag}}^{\text{dec}}, \mathbf{H}_{\text{ag}}^{\text{dec}})$;
  // Combine intra-chain and cross-chain information
  $\mathbf{H}_{\text{ag}}^{\text{dec}(\ell)} \leftarrow \mathbf{H}_{\text{ag}}^{\text{dec}(\ell-1)} + \text{FFN}(\widetilde{\mathbf{H}}_{\text{ag}}^{\text{dec}})$;
  $\mathbf{H}_{\text{ab}}^{\text{dec}(\ell)} \leftarrow \mathbf{H}_{\text{ab}}^{\text{dec}(\ell-1)} + \text{FFN}(\widetilde{\mathbf{H}}_{\text{ab}}^{\text{dec}})$;
end
// **Bipartite Interaction Prediction**
Compute bidirectional attention scores:;
$\mathbf{S}_{\text{ag}\to\text{ab}} \leftarrow \frac{(\mathbf{H}_{\text{ag}}^{\text{dec}} \mathbf{W}_Q^{\text{out}})(\mathbf{H}_{\text{ab}}^{\text{dec}} \mathbf{W}_K^{\text{out}})^\top}{\sqrt{d_k}}$;
$\mathbf{S}_{\text{ab}\to\text{ag}} \leftarrow \frac{(\mathbf{H}_{\text{ab}}^{\text{dec}} \mathbf{W}_Q'^{\text{out}})(\mathbf{H}_{\text{ag}}^{\text{dec}} \mathbf{W}_K'^{\text{out}})^\top}{\sqrt{d_k}}$;
Fuse scores and apply sigmoid:;
$\mathbf{Z} \leftarrow \mathbf{w}^\top [\, \mathbf{S}_{\text{ag}\to\text{ab}}\, (\mathbf{S}_{\text{ab}\to\text{ag}})^\top \,] + b$;
$\hat{\mathcal{E}}_{\text{bg}} \leftarrow \sigma(\mathbf{Z})$;
// **Epitope Extraction**
Extract per-residue epitope probabilities via mean pooling:;
$(\hat{y}_{\text{ag}})_i = \frac{1}{m} \sum_{j=1}^{m} (\hat{\mathcal{E}}_{\text{bg}})_{ij}$;
Function MHCA(Q, K, V)**:**
  $Q_h \leftarrow QW_Q^h, K_h \leftarrow KW_K^h, V_h \leftarrow VW_V^h$; // Project per head h
  $\alpha_{ij}^h \leftarrow \text{Softmax}_j\left(\frac{Q_{h,i} \cdot K_{h,j}^\top}{\sqrt{d_h}}\right)$; // Attention scores
  $C_i^h \leftarrow \sum_j \alpha_{ij}^h V_{h,j}$; // Context vector
  Result: Concat$(C^1, \ldots, C^H)W_O$; // Combine heads
end
Function FFN(X)**:**
  $\hat{\mathcal{E}}_{\text{bg}} \leftarrow \text{SiLU}(XW_1 + b_1)W_2 + b_2$; // $W_1 \in \mathbb{R}^{d \times d_{ff}}$, $W_2 \in \mathbb{R}^{d_{ff} \times d}$
  Result: $\hat{\mathcal{E}}_{\text{bg}}$
end

at approximately $6.5e{-}5$ in the best configuration. A ReduceLROnPlateau scheduler monitors validation performance and decays the learning rate on stagnation, while an early stopping with patience of 10 epochs prevents overfitting and reduces variance in final selection. We used SiLU activation functions (Elfwing et al., 2018) throughout the model because they provide stable gradients via their smooth, non-monotonic curve, which are crucial for training deep graph networks. The hyperparameter tuning was performed via a Bayesian optimization sweep in Weights & Biases to minimize the validation loss, and the best hyperparameters were chosen within a predefined search

space using bounded uniform and log-uniform distributions. We ran the hyperparameter sweeping experiments separately for the epitope ratio and epitope group split settings.

- The model weight decay was sampled log-uniformly over $[1e{-}6,\ 1e{-}4]$ to prevent overfitting by penalizing large weights, with the best configuration using approximately $9.9e{-}5$.
- The model dropout was sampled log-uniformly over $[0.05, 0.5]$ to improve the generalizability of the model, and the best performing configuration for the epitope ratio model used a dropout of 0.132, while the epitope group model used a dropout of 0.053.
- The number of layers in the encoder module is treated as a hyperparameter and was chosen from the set $[3, 4, 5]$ while for the decoder, the number of layers was chosen from the $[2, 3, 4]$. We experimented with different encoder hidden dimensions and the best configuration of 128 was picked from $[64, 128, 256, 512]$ across different runs.
- We also experimented with different number of attention heads for the encoder and decoder MHCA (2,4,8,16) and picked the best model with 8 attention heads.
- A batch size of 8 was chosen from [4,8,16,32] across different runs.
- α_{ag} and α_{ab} are initialised to 0.05

For the loss coefficients, the best configuration for the epitope-ratio split uses $\lambda_{\text{edge}} = 1.0$, $\lambda_{\text{node}} = 0.4816$, $\lambda_{\text{geo}} = 0.0514$, $\beta_{\text{BCE}} = 9.3249$, $\beta_{\text{Dice}} = 2.2966$, $\beta_{\text{sparsity}} = 0.3068$, $\pi_{\text{epi}} = 15.2856$, $\pi_{\text{edge}} = 58.7077$, label smoothing $\epsilon = 0.1$, and a distance cutoff of 32 Å for $\mathcal{L}_{\text{geo}}$. The best run for the model trained on the epitope group split uses $\lambda_{\text{node}} = 0.143$, $\lambda_{\text{edge}} = 1.0$, $\lambda_{\text{geo}} = 0.158$, $\beta_{\text{BCE}} = 9.16$, $\beta_{\text{Dice}} = 1.83$, $\beta_{\text{sparsity}} = 0.64$, $\pi_{\text{epi}} = 53.18$, $\pi_{\text{edge}} = 44.11$, label smoothing $\epsilon = 0.1$, and a distance cutoff of 32 Å for $\mathcal{L}_{\text{geo}}$.

- The bipartite edge positive-class weight π_{edge} for the BCE-with-logits interaction loss was sampled log-uniformly over $[30,\ 150]$, accommodating variation in pairwise sparsity across complexes.
- The node objective weight λ_{node} was sampled uniformly over $[0.05,\ 0.5]$, exploring the trade-off between residue supervision and the other objectives.
- The binary cross-entropy multiplier within the node objective β_{BCE} was drawn uniformly over $[2,\ 10]$, spanning weak to strong emphasis on classification error.
- The Dice multiplier β_{Dice} was drawn uniformly over $[0.1,\ 3.0]$, reflecting its role as a secondary calibrator under class imbalance.
- The epitope positive-class weight π_{epi} was sampled log-uniformly over $[10,\ 60]$, covering roughly an order of magnitude in imbalance without biasing toward either extreme.
- The per-graph epitope count-regularizer weight β_{sparsity} was sampled uniformly over $[0.05,\ 1.0]$, enabling calibration of predicted positive counts at the complex level.
- The auxiliary distance-classification weight λ_{geo} was sampled uniformly over $[0.05,\ 0.3]$, with class balancing across distance bins and distance-aware pair weighting kept enabled and the maximum distance fixed at 32 Å for all trials.

The experiments were performed on an NVIDIA H100 GPU and it took around 35-60 minutes for a single hyperparameter sweeping experiment of around 50 epochs. To ensure full reproducibility of our experiments, we implement random seed management across all computational components, including NumPy (`numpy.random`), Python (`random`), PyTorch (`torch.manual_seed`), and CUDA operations (`torch.cuda.manual_seed_all`), while additionally controlling worker initialization in data loaders and disabling non-deterministic algorithms (`torch.backends.cudnn.deterministic=True`).

C BASELINES

In this section, we provide the baseline descriptions and the implementations under the similar experimental setup to *EpiFormer*.

C.1 Baseline descriptions

All baselines were evaluated on the 1,721-complex AsEP benchmark under identical settings: official splits (epitope-ratio and epitope-group), and the same six metrics (AUROC, AUPRC, F1, MCC, Precision, Recall). Below we describe each baseline's architecture and its adaptation to epitope prediction.

C.1.1 Protein Binding-Site Prediction Methods

EquiPocket (Zhang et al., 2023) operates at atom-level with 6D features, computing molecular surfaces via MSMS with 7D local geometric descriptors. The Surface-EGNN module performs E(3)-equivariant multi-channel convolution with DenseNet-style connections. We adapted to epitope prediction by adding residue aggregation (CA atom selection or mean pooling) to convert per-atom outputs to per-residue predictions.

AtomSurf (Mallet et al., 2023) jointly encodes molecular surfaces using DiffusionNet (spectral convolution with 22D geometric features including HKS) and residue graphs using GCN with ProNet features, connected through bidirectional KNN-based message passing. We preserve the joint surface-graph encoding for antigen-only processing with mean-pooled surface features aggregated to the residue level.

ESMBind (Schreiber, 2023) adapts the ESM-2 protein language model (35M parameters by default) with LoRA (Low-Rank Adaptation) for parameter-efficient fine-tuning, training only $\tilde{0}.5\%$ of parameters. The model processes antigen amino acid sequences through ESM-2 with LoRA adapters (rank $r = 4$, scaling $\alpha = 8$, dropout 0.1) applied to attention layers, followed by a per-residue classification head for epitope prediction. We fine-tune ESMBind on the epitope sequences using class-weighted BCE loss to handle label imbalance, and support multiple ESM-2 variants (8M to 650M parameters).

C.1.2 Protein Interaction and Docking Methods

ATProt (Gao et al., 2024) uses SEGCN (Structure-Enhanced Graph Convolutional Network) for general protein-protein interaction site prediction. We ported the architecture from DGL to PyG, added cross-attention between antibody and antigen chains for antibody awareness, and trained from scratch on AsEP with the same evaluation protocol. ATProt applies cross-attention only once after all encoding layers (late fusion), in contrast to EpiFormer's interleaved design.

DiffDock-PP (Ketata et al., 2023) uses SE(3)-equivariant tensor product convolutions for rigid-body protein-protein docking. We adapted the architecture for residue-level epitope classification by replacing the docking score head with per-residue binary classification, adding cross-attention between antibody and antigen representations. Both antibody and antigen are encoded using the same TensorProductConvLayer backbone with ESM-2 embeddings.

DiffDock (Corso et al., 2022) generates ligand poses through diffusion over SE(3) using E(3)-equivariant score networks. We use inference-only mode with DiffDock-L to dock antibody CDR regions onto antigen structures, labeling antigen residues within contact distance of predicted CDR poses as epitopes, weighted by confidence scores.

C.1.3 Structure and Affinity Prediction Methods

Boltz-1 (Wohlwend et al., 2025) uses a Pairformer trunk with triangle multiplication ($z_{ij} += \sum_k a_{ik} \cdot b_{jk}$), triangle attention, and attention with pair bias for structure-aware representations. We adapt by projecting ESM-2 embeddings to single (384D) and pairwise (128D) representations, reducing to 8 layers, and adding an epitope prediction head on concatenated single and mean-pooled pairwise features.

AlphaFold3 (Abramson et al., 2024) uses a diffusion-based architecture with Pairformer trunk processing single and pairwise representations through adaptive LayerNorm, gated linear units, and

triangle operations. We use inference-only mode with pretrained weights [1], converting inter-chain contact predictions from predicted complexes to epitope labels using distance thresholds. Due to imprecise docking without evolutionary information, we used a relaxed 15Å contact threshold. AlphaFold3 without MSA achieved AUC=0.56 with the relaxed 15Å threshold, only marginally better than random, confirming that structure prediction alone without co-evolutionary signals cannot reliably identify epitopes.

CheapNet (Lim et al., 2025) uses Geometry-Informed Graph Neural Network blocks for local structure encoding with intra/inter-molecular edges and cross-attention between entity representations. We adapted from graph-level affinity regression to node-level classification by replacing ligand-protein atom pairs with antibody-antigen residue pairs, projecting 35D residue features from RAAD (Wu et al., 2025)/ESM-2, and adding a per-residue MLP classifier with BCE loss.

GearBind (Cai et al., 2024) is a pretrainable geometric GNN based on GearNet, pretrained on CATH using contrastive learning and fine-tuned on SKEMPI for $\Delta\Delta G_{\text{bind}}$ prediction. We extract encoder representations for antigen residues, replace the affinity regression head with a binary classifier, and initialize from CATH-pretrained weights.

Boltz-2 (Passaro et al., 2025) extends Boltz-1 with cross-attention between binding partners. We additionally process antibody embeddings (AntiBERTy, 512D) alongside antigen (ESM-2, 1280D), using 2-layer CrossPairAttention for antibody-aware epitope prediction with the same triangle-based Pairformer backbone.

C.1.4 Molecular Property Prediction Methods

EquiformerV2 (Liao et al., 2023) is an improved equivariant transformer using SO(2) convolutions for efficiency, processing 3D graphs with type-L features and attention in irreducible representation space. We adapt from energy/force regression to residue-level classification by constructing antigen graphs with CA coordinates, combining RAAD features with ESM-2 embeddings, and replacing the output head with per-residue binary classification.

C.2 Architectural and loss component comparison

Table 5 compares the architectural components and loss functions used by each baseline against *EpiFormer*.

C.3 Converting co-folding and docking predictions into epitope labels

AlphaFold3, Boltz-1, Boltz-2, and DiffDock predict a complex rather than a residue label. The conversion is similar for all four methods. We predict the complex, extract the inter-chain contacts, and map the contacting antigen residues onto the same AsEP-labeled set, so that all methods are scored against the same residues.

Our AlphaFold3 test does not represent the method at its strongest. We ran it without MSAs because the full pipeline costs around 60-80 minutes per complex over roughly 300 test complexes, and the relaxed 15 Å contact threshold was adopted **in AlphaFold3's favour**, since the standard 4.5 Å cutoff returned almost no contacts. Our experiments therefore measure co-folding without MSAs, and we label it *AlphaFold3 (no MSA)* in Table 1. Boltz-2 is a properly configured co-folding baseline in the same table and reaches 0.255 F1 against *EpiFormer*'s 0.482. Independent evaluations of AlphaFold3 with full MSAs report that it reaches acceptable accuracy (DockQ above 0.23) on 38% of antibody-antigen complexes and high accuracy on 11%. The widely quoted 60% requires roughly a thousand seeds per target, which is not a screening protocol (Hitawala & Gray, 2025). The current co-folding methods do not exceed 44% acceptable accuracy on a 200-complex antibody-antigen set (Unsal et al., 2025). Antibody-antigen is one of the interface classes on which co-folding remains unreliable, which motivates a task-specific predictor for this task.

[1] The AF3 model weights were obtained from DeepMind. We used AF3 without MSA due to computational resource constraints.

Table 5: Architectural and loss component comparison across baselines and EpiFormer.

Method	Architecture										Loss Components				
	Ab	St	PLM	Graph	Geom	M-rel	E(3)	X-Att	Inter	Gate	BCE	Dice	Contr	Count	Edge
Epitope and paratope prediction															
EpiGraph	×	✓	✓	✓	×	×	×	×	×	×	✓	×	×	×	×
EpiScan	✓	×	×	×	×	×	×	×	×	×	✓	✓	×	×	×
MIPE	✓	✓	✓	✓	✓	✓	×	✓	×	×	✓	×	✓	×	×
WALLE	✓	✓	✓	✓	×	×	×	×	×	×	✓	×	×	✓	✓
DiscoTope3	×	✓	✓	×	✓	×	×	×	×	×	✓	×	×	×	×
GraphBepi	×	✓	✓	✓	✓	×	×	×	×	×	✓	×	×	×	×
PECAN	✓	✓	×	✓	✓	×	×	✓	×	×	✓	×	×	×	×
Protein binding-site prediction															
EquiPocket	×	✓	×	✓	✓	×	✓	×	×	×	✓	✓	×	×	×
AtomSurf	×	✓	✓	✓	✓	×	×	×	×	×	✓	×	×	×	×
ESMBind	×	×	✓	×	×	×	×	×	×	×	✓	×	×	×	×
Protein interaction and docking															
ATProt	✓	✓	×	✓	✓	×	×	✓	×	×	✓	×	×	×	×
DiffDock-PP	✓	✓	✓	✓	✓	×	✓	✓	×	×	✓	×	×	×	×
DiffDock	×	✓	×	✓	✓	×	✓	×	×	×	×	×	×	×	×
Structure and affinity prediction															
Boltz-1	×	✓	✓	×	✓	×	×	×	×	×	✓	×	×	×	×
Boltz-2	✓	✓	✓	×	✓	×	×	✓	×	×	✓	×	×	×	×
AlphaFold3	✓	✓	×	×	✓	×	✓	×	×	×	×	×	×	×	×
CheapNet	✓	✓	×	✓	✓	×	×	✓	×	×	✓	×	×	×	×
GearBind	×	✓	×	✓	✓	✓	×	×	×	×	✓	×	×	×	×
Molecular property prediction															
EquiformerV2	×	✓	✓	✓	✓	×	✓	×	×	×	✓	×	×	×	×
EpiFormer	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓

Architecture columns: Ab = antibody-aware; St = uses 3D structure; PLM = protein language model features; Graph = graph neural network; Geom = explicit geometric or surface features; M-rel = multi-relational edges; E(3) = E(3)/SE(3)-equivariant; X-Att = cross-attention; Inter = interleaved (per-layer) cross-attention; Gate = gated cross-attention.

D Extended results

In this section, we provide the detailed results.

D.1 Epitope-group split

Table 6 reports results on the more challenging epitope-group split, where test epitopes are completely unseen during training. EpiFormer achieves 0.826 AUC and 0.305 F1, outperforming all baselines.

D.2 Cross-dataset evaluation

We evaluated EpiFormer on three external benchmarks with no overlap with AsEP training data, SAbDab (Dunbar et al., 2014) (494 diverse antibody-antigen complexes from the Structural Antibody Database, post-2023), CoV-AbDab (Raybould et al., 2021) (170 SARS-CoV-2 neutralizing antibody complexes), and ANABAG (Grandguillaume et al., 2025) (499 curated complexes). We compare EpiFormer against MIPE, which is the second-best method on AsEP (Table 1) and the strongest antibody-aware baseline. Table 7 reports both zero-shot (no retraining) and leave-one-dataset-out (LODO) fine-tuning settings.

D.3 Per-complex metric distributions

Figure 7 shows per-complex metric distributions on the epitope-ratio test set. EpiFormer achieves higher medians and tighter interquartile ranges than MIPE and WALLE across all six metrics, indicating more consistent predictions. WALLE exhibits high recall but low precision, showing its tendency to over-predict epitope residues.

D.4 Stratification by X-ray resolution

Table 8 stratifies performance by X-ray crystallographic resolution. EpiFormer maintains strong AUC and F1 across all resolution bins with no catastrophic degradation at low resolution (> 3.0 Å).

Table 6: Performance comparison on the AsEP epitope-group split. Best in **bold**, second-best underlined.

Method	AUC↑	AUPRC↑	F1↑	MCC↑	Prec.↑	Rec.↑
Epitope and paratope prediction						
EpiGraph (Choi & Kim, 2024)	0.779	0.194	0.056	0.096	0.401	0.030
EpiScan (Wang et al., 2024a)	0.443	0.080	0.127	0.048	0.089	0.221
MIPE (Wang et al., 2024b)	0.740	0.228	0.172	0.206	**0.495**	0.104
WALLE (Liu et al., 2024)	0.713	0.137	0.170	0.143	0.095	**0.830**
DiscoTope3 (Høie et al., 2024)	0.763	0.208	0.210	0.189	0.123	0.729
GraphBepi (Zeng et al., 2023)	0.781	0.220	0.144	0.161	0.386	0.088
PECAN (Pittala & Bailey-Kellogg, 2020)	0.729	0.166	0.254	0.195	0.169	0.513
Protein binding-site prediction						
EquiPocket (Zhang et al., 2023)	0.711	0.132	0.194	0.159	0.113	0.676
AtomSurf (Mallet et al., 2023)	0.699	0.144	0.202	0.151	0.127	0.500
ESMBind (Schreiber, 2023)	0.685	0.130	0.174	0.126	0.100	0.663
Protein interaction and docking						
ATProt (Gao et al., 2024)	0.710	0.155	0.200	0.160	0.120	0.600
DiffDock-PP (Ketata et al., 2023)	0.650	0.110	0.165	0.120	0.095	0.680
DiffDock (Corso et al., 2022)	0.443	0.095	0.146	0.081	0.131	0.164
Structure and affinity prediction						
Boltz-1 (Wohlwend et al., 2025)	0.702	0.174	0.229	0.169	0.146	0.536
Boltz-2 (Passaro et al., 2025)	0.702	0.168	0.220	0.156	0.140	0.518
AlphaFold3 (no MSA) (Abramson et al., 2024)	0.570	0.076	0.155	0.079	0.098	0.375
CheapNet (Lim et al., 2025)	0.641	0.095	0.152	0.100	0.084	0.776
GearBind (Cai et al., 2024)	0.734	0.154	0.206	0.169	0.123	0.627
Molecular property prediction						
EquiformerV2 (Liao et al., 2023)	0.729	0.160	0.212	0.169	0.131	0.561
EpiFormer **(ours)**	**0.826**	**0.290**	**0.305**	**0.290**	0.196	0.698

Table 7: Cross-dataset generalization on SAbDab (Dunbar et al., 2014), CoV-AbDab (Raybould et al., 2021), and ANABAG. Zero-shot: models trained on AsEP evaluated without retraining. LODO (leave-one-dataset-out): models fine-tuned on the two external datasets not used for testing. Per-dataset best in **bold**.

		Zero-shot		LODO	
Dataset	**Method**	**AUC**	**F1**	**AUC**	**F1**
SAbDab (494)	**EpiFormer**	**0.782**	**0.297**	**0.858**	**0.364**
	MIPE	0.614	0.140	0.746	0.190
CoV-AbDab (170)	**EpiFormer**	**0.819**	**0.358**	**0.890**	**0.359**
	MIPE	0.730	0.175	0.813	0.212
ANABAG (499)	**EpiFormer**	**0.758**	**0.301**	0.837	0.367
	MIPE	0.719	0.220	**0.772**	**0.217**

Figure 8 shows the full per-complex distributions; EpiFormer's advantage over MIPE and WALLE is consistent across all resolution bins and metrics.

Table 8: Performance stratified by X-ray resolution on the epitope-ratio test set (per-complex mean ± std).

Resolution	N	Method	AUC	F1	MCC	Rec.
< 2.0 Å	17	**EpiFormer**	**.908±.071**	**.622±.238**	**.553±.265**	.721±.269
	3	MIPE	.515±.212	.119±.205	.011±.203	.107±.185
	17	WALLE	.734±.204	.372±.146	.207±.209	**.964±.090**
2.0–3.0 Å	75	**EpiFormer**	**.898±.080**	**.484±.235**	**.434±.243**	.553±.291
	6	MIPE	.674±.221	.163±.398	.145±.406	.167±.408
	75	WALLE	.719±.170	.238±.109	.137±.119	**.915±.146**
> 3.0 Å	78	**EpiFormer**	**.916±.070**	**.464±.206**	**.437±.211**	.565±.271
	3	MIPE	.803±.128	.000±.000	.010±.011	.000±.000
	78	WALLE	.748±.172	.202±.096	.162±.121	**.932±.167**

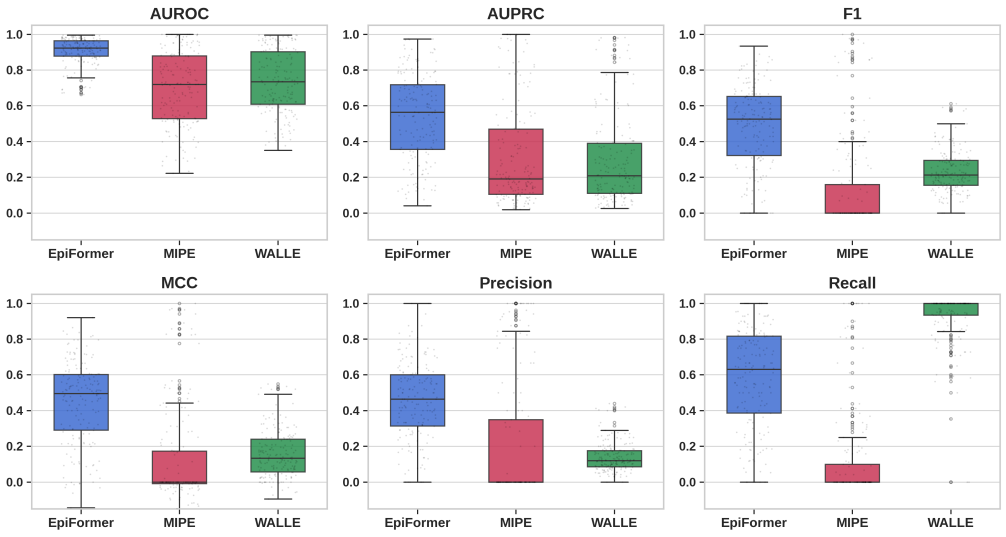


Figure 7: Per-complex metric distributions on the epitope-ratio test set (170 complexes) for EpiFormer, MIPE, and WALLE.

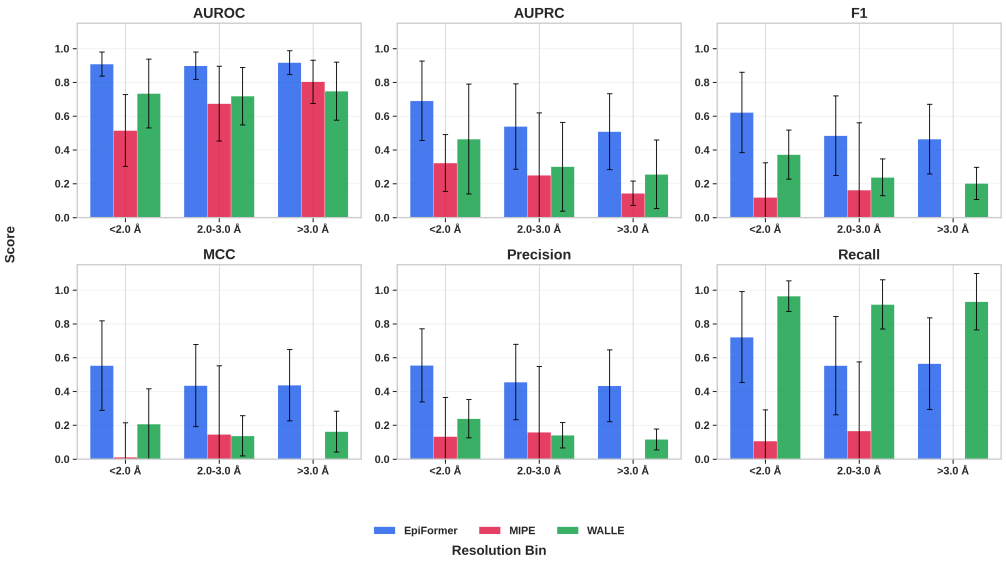


Figure 8: Per-complex metric distributions (AUROC, AUPRC, F1, MCC, Precision, Recall) stratified by X-ray resolution for EpiFormer, MIPE, and WALLE.

D.5 Stratification by antigen size

Table 9 stratifies performance by antigen surface residue count. EpiFormer's AUC improves with antigen size, while F1 declines due to increasing class imbalance (the mean epitope ratio drops from 13.9% in small antigens to 2.2% in large ones). MIPE collapses for medium and large antigens. Figure 9 shows the full per-complex distributions across all six metrics.

E Robustness of the structural input

The AsEP chains are extracted from solved complexes (Appendix A.3). In this section, we therefore test whether *EpiFormer* generalizes to inputs that, in reality, don't form complexes as in the bench-

Table 9: Performance stratified by antigen surface residue count on the epitope-ratio test set (per-complex mean ± std).

Size Bin	N	Epi. Ratio	Method	AUC	F1
< 200 (small)	90	13.9%	**EpiFormer**	**.901±.081**	**.568±.202**
			WALLE	.707±.180	.293±.114
			MIPE	.630±.241	.222±.396
200–500 (med)	62	6.6%	**EpiFormer**	**.905±.068**	**.412±.206**
			WALLE	.730±.156	.167±.078
			MIPE	.561±.145	.000±.000
> 500 (large)	18	2.2%	**EpiFormer**	**.953±.047**	**.356±.260**
			WALLE	.876±.136	.175±.090
			MIPE	.845±.109	.000±.000

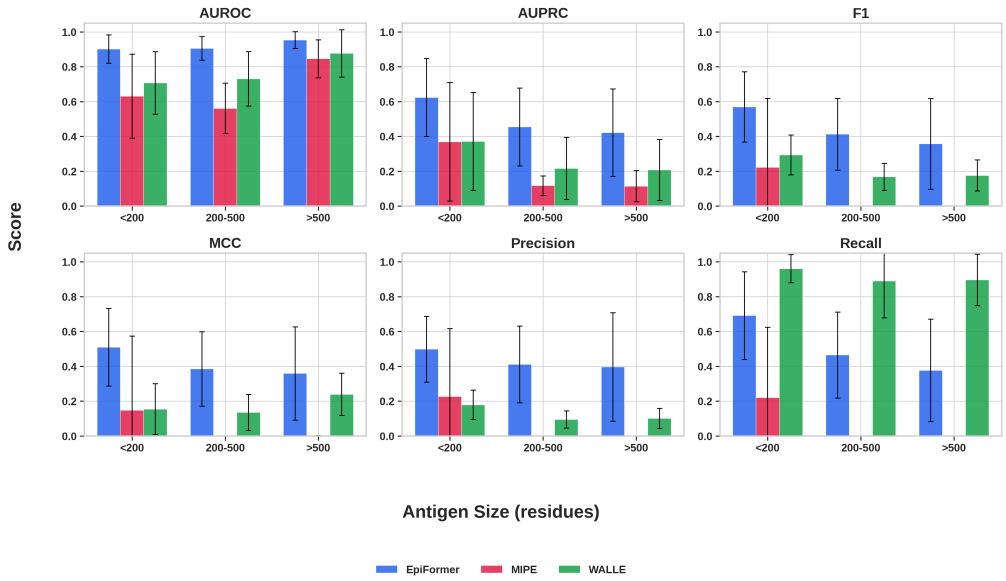


Figure 9: Per-complex metric distributions (AUROC, AUPRC, F1, MCC, Precision, Recall) stratified by antigen size for EpiFormer, MIPE, and WALLE.

Table 10: Inference latency scaling by antigen size (ms/sample, batch size 1).

Antigen Size	# Samples	WALLE (ms)	MIPE (ms)	EpiFormer (ms)
Small (<200 res.)	∼100	19.16	108.93	100.56
Medium (200–500 res.)	∼55	22.94	221.79	123.00
Large (>500 res.)	∼15	24.10	1087.26	161.62
Ratio (Large/Small)		**1.26×**	**9.98×**	**1.61×**

marks. We remove the shared reference frame, substitute a conformation that the antigen adopts against a different antibody, inject structural error, retrain under a split that removes antigen families and antibody lineages, and compare the model against an antigen-only lookup.

E.1 Randomizing the local coordinate frames

The two chains of an AsEP complex are deposited in a common reference frame, and a model could in principle exploit the relative orientation that this encodes. To quantify this, we retrained *EpiFormer* with the frame block randomized at each step.

Here, *global* applies one rotation to both chains, removing the global reference frame while preserving their relative orientation, whereas *per-chain* rotates each chain independently and destroys the orientation as well. Two patterns emerge from these results.

Table 11: Retraining with randomized local coordinate frames.

Frames	AUC (ratio)	F1 (ratio)	AUC (group)	F1 (group)
As given, and under any rigid transform	0.927	0.496	0.869	0.347
Global randomization	**0.937**	**0.505**	**0.890**	**0.384**
Per-chain randomization	0.821	0.323	0.722	0.219

- The dependence on a global frame is a removable nuisance rather than a source of signal, since training that dependence away improves all metrics on both splits.
- Providing rotation without translation is a harder scenario, since it provides a direction rather than a site. The model still has to predict which face of the antigen to attend to, which is why removing that orientation costs 0.116 and 0.168 AUROC (global minus per-chain) without collapsing the model.

E.2 Transfer to a donor antigen conformation

The practical scenario for this task is one in which a practitioner has an antigen structure, solved or predicted, together with an antibody of interest, and needs to know where on that antigen it binds without solving or docking the complex. AsEP contains antigens that appear in several complexes bound by different antibodies, which lets us construct that scenario from the benchmark itself.

For each test complex, we selected the donor crystal whose epitope overlaps least with the test epitope, substituted the antigen coordinates with the donor's, and kept the test complex's own labels, so that the test epitope lies in a conformation the antigen adopts when a different antibody binds it. All features and graphs are rebuilt from the donor coordinates, and we evaluate only on residues resolved in both structures.

Table 12: Antigen coordinates substituted with those of a donor complex bound by a different antibody, keeping the test complex's own labels.

Subset	n	Crystal control AUC	Donor conformation AUC
Epitope-ratio, all donors	53	0.943	0.933
Epitope-group, all donors	22	0.912	0.896
Epitope-ratio, disjoint epitopes	29	0.917	0.905

The median Cα RMSD between donor and native antigen is 0.73 Å, and 37 of the 53 donors share no epitope residue with the test complex at all. The performance is close to unchanged under the substitution, which indicates that the model uses the antigen surface.

E.3 Injected structural error

A predicted antigen carries errors that a crystal structure does not have. Hence, we injected correlated backbone error into the antigen. The displacement field is smoothed along the sequence so that whole loops move together, which is how structure-prediction error actually behaves, and each residue is moved rigidly so that bond geometry stays physical.

Table 13: Correlated backbone error injected into the antigen, on both official splits.

Target Cα RMSD	0.0 Å	1.0 Å	2.0 Å	3.0 Å	5.0 Å
Epitope-ratio AUC	0.927	0.925	0.920	0.911	0.888
Epitope-ratio F1	0.496	0.493	0.483	0.464	0.432
Epitope-group AUC	0.869	0.862	0.853	0.843	0.823
Epitope-group F1	0.347	0.342	0.334	0.325	0.302

At 2 Å (roughly the backbone accuracy of a good predicted structure), the cost is 0.007 and 0.016 AUROC. For a relatively poor model (5 Å error), *EpiFormer* reaches 0.432 F1 against 0.337 for the

best baseline evaluated on the crystal structure. This is consistent with the established work for this task, since DiscoTope-3.0 showed that a structure-based epitope predictor holds up when solved structures are replaced by predicted ones (Høie et al., 2024).

E.4 Generalization under the family-based split

We further retrained *EpiFormer* on the split of Appendix A.5, alongside an otherwise identical model with the encoder cross-attention switched off so that only the decoder fuses the two chains. This second model is the late fusion mechanism used in the current antibody-aware baselines.

Table 14: Early against late fusion under the family-based split.

Family-based split (n=344)	**AUROC**	**F1**	**MCC**
***EpiFormer*, early fusion**	**0.866**	**0.407**	**0.368**
Same model, late fusion only	0.712	0.196	0.135

Removing antigen families and antibody lineages costs *EpiFormer* 0.054 AUROC and 0.065 F1 (against 0.920 and 0.471), which is the expected cost of a harder split. The comparison between the two rows is more informative, since early fusion holds an advantage of 0.211 F1 and 0.232 MCC over late fusion, well beyond its margin on the official splits. Generalizing to an unseen antigen family is therefore a scenario where conditioning on the antibody early becomes important.

F Ablations

In this section, we ablate the encoder backbone, the residue graph, the joint objective, the pooling strategy, the cross-attention gate, the input features and the loss weights.

F.1 Encoder backbone

To assess the impact of geometric message passing, we replaced the EGNN-R layers with alternative GNN architectures while keeping all other components fixed (Table 2).

Even simple backbones like GCN (0.447 F1) outperform all other evaluated baselines, showing that the EpiFormer framework, namely its interleaved cross-attention and joint losses, is the primary performance driver. EGNN-R provides an additional +0.028 F1 over the best non-equivariant variant (RGCN, .454) through equivariant coordinate updates and multi-relational encoding. We include vanilla EGNN and EGNN-R with frozen coordinates to decompose the contributions, which are +0.018 F1 from equivariance (EGNN vs GAT, .470 vs .452), +0.007 from multi-relational edges (RGCN vs GCN, .454 vs .447), and +0.017 from coordinate updates (EGNN-R vs EGNN-R frozen, .482 vs .465).

F.2 RAAD and GearNet residue graphs

GearNet differs from our implementation on two axes: the residue graph and the encoder. Our claimed novelty is in the encoder, while we use the graph type proposed in RAAD (Wu et al., 2025). We rebuilt the dataset with the SequentialEdge, SpatialEdge, and KNNEdge constructions over seven relations with their 59-D edge features, and retrained *EpiFormer*.

Table 15: Residue graph construction with the encoder held fixed. Both rows use the *EpiFormer* architecture and objective.

Residue graph	**Relations**	**ratio F1**	**ratio MCC**	**group F1**	**group MCC**
GearNet, official construction	7	0.472	0.453	0.338	0.315
RAAD (ours)	**4**	0.471	0.454	**0.356**	**0.357**

The two graphs produce similar F1 and MCC on epitope-ratio, and on epitope-group they differ slightly. This is the expected outcome, because the graph type proposed in RAAD (Wu et al., 2025)

is not a contribution of this paper. It is a standard residue-level protein graph built from sequence adjacency, k-nearest neighbors, and a spatial cutoff, and GearNet builds the same object with a finer relation partition.

However, EGNN-R is our extension of the GearNet encoder, which learns a separate message function per relation type and performs equivariant coordinate updates inside each block. GearNet's encoder is relational but not equivariant, and it does not update coordinates, as in RGCN. From Table 2, EGNN-R gains 0.028 F1 and 0.042 MCC over the relational non-equivariant family that GearNet belongs to. *EpiFormer* uses the RAAD graph construction, as it has fewer edge relations and produces marginal gains with our encoder.

F.3 THE JOINT OBJECTIVE

We performed ablations to evaluate the contribution of each loss component (primary, auxiliary, and regularizers) to epitope prediction, as shown in Table 16.

Contrastive Learning Loss ($\mathcal{L}_{\text{InfoNCE}}$). We also performed contrastive learning with the SimCLR InfoNCE (Information Noise Contrastive Estimation) loss (Chen et al., 2020) to learn discriminative representations by contrasting positive and negative residue pairs within and across protein chains. The contrastive loss combines intra-chain and inter-chain objectives:

$$\mathcal{L}_{\text{contrastive}} = \lambda_{\text{intra}}\mathcal{L}_{\text{intra}} + \lambda_{\text{inter}}\mathcal{L}_{\text{inter}}, \tag{33}$$

where λ_{intra} and λ_{inter} balance the relative importance of within-chain and cross-chain contrastive learning.

INTRA-CHAIN CONTRASTIVE LOSS ($\mathcal{L}_{\text{INTRA}}$). The intra-chain loss encourages similar representations for residues with the same label (epitope/non-epitope or paratope/non-paratope) within each protein chain:

$$\mathcal{L}_{\text{intra}} = \mathcal{L}^{\text{ag}}_{\text{intra}} + \mathcal{L}^{\text{ab}}_{\text{intra}}. \tag{34}$$

For each chain (antigen or antibody), the loss is computed as:

$$\mathcal{L}^{\text{chain}}_{\text{intra}} = -\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \log \frac{\sum_{j\in\mathcal{P}_{i^+}} \exp(\mathbf{h}_i^T\mathbf{h}_j/\tau)}{\sum_{k\in\mathcal{N}_i} \exp(\mathbf{h}_i^T\mathbf{h}_k/\tau)}, \tag{35}$$

where $\mathcal{P} = \{i : y_i = 1\}$ is the set of positive (binding) residues, $\mathcal{P}_{i^+} = \{j \in \mathcal{P} : j \neq i\}$ are other positive residues sharing the same label as anchor i, $\mathcal{N}_i$ includes all negative residues for anchor i, $\mathbf{h}_i, \mathbf{h}_j$ are L_2-normalized residue embeddings, and τ is the temperature parameter controlling concentration.

INTER-CHAIN CONTRASTIVE LOSS ($\mathcal{L}_{\text{INTER}}$). The inter-chain loss promotes alignment between epitope and paratope representations across antigen-antibody pairs:

$$\mathcal{L}_{\text{inter}} = \mathcal{L}_{\text{ag}\to\text{ab}} + \mathcal{L}_{\text{ab}\to\text{ag}}. \tag{36}$$

The bidirectional formulation ensures symmetric learning:

$$\mathcal{L}_{\text{ag}\to\text{ab}} = -\frac{1}{|\mathcal{P}_{\text{ag}}|}\sum_{i\in\mathcal{P}_{\text{ag}}} \log \frac{\sum_{j\in\mathcal{P}_{\text{ab}}} \exp(\mathbf{h}_i^{\text{ag}T}\mathbf{h}_j^{\text{ab}}/\tau)}{\sum_{k\in\mathcal{N}_{\text{cross}}} \exp(\mathbf{h}_i^{\text{ag}T}\mathbf{h}_k/\tau)}, \tag{37}$$

where $\mathcal{P}_{\text{ag}}, \mathcal{P}_{\text{ab}}$ are epitope and paratope residue sets, $\mathcal{N}_{\text{cross}}$ includes negative residues from both chains, and the loss pulls epitope embeddings closer to paratope embeddings while pushing them away from non-binding residues. Our experiments show that contrastive learning didn't contribute to improving the classification performance. We attribute this to conflicting optimization objectives between BCE loss and standard InfoNCE loss, a phenomenon demonstrated in a recent work (Ji et al., 2024).

Table 16: Loss function ablation on the epitope-ratio split. Each row adds the listed term to the configuration above it; best in **bold**.

Loss Config	AUC	F1	MCC	Δ F1
$\mathcal{L}_{\text{BCE}}$	.916	.336	.366	−.160
$+\,\mathcal{L}_{\text{edge}}$	.916	.323	.356	−.173
$+\,\mathcal{L}_{\text{Dice}}$	.916	.329	.361	−.167
$+\,\mathcal{L}_{\text{sparsity}}$	.917	.461	.440	−.035
+ $\mathcal{L}_{\text{geo}}$ (full)	**.926**	**.496**	**.475**	—
$+\,\mathcal{L}_{\text{InfoNCE}}$	.920	.467	.443	−.029

Table 17: Architecture against objective on the epitope-ratio split. The second row is the best of six baselines retrained with our full objective.

Control	Architecture	Objective	F1
Architecture alone	*EpiFormer*, early fusion	plain BCE	0.336
Objective alone, best case	WALLE, late fusion	ours, tuned	0.332
Together	***EpiFormer*, early fusion**	**ours**	**0.482**

F.4 Separating the architecture from the objective

We note that neither the architecture nor the objective alone accounts for the gains.

1. **Architecture alone.** *EpiFormer* with a plain BCE loss reaches 0.336 F1, and already matches the strongest baseline in the field (MIPE at 0.337).
2. **Objective alone.** We retrained six baselines from scratch with our full objective and tuned their weights per baseline by grid search. The gains become apparent along the fusion type, as the models that are not late-fusion improve (WALLE +0.129, EquiformerV2 +0.041, DiscoTope3 +0.031), while the late-fusion baselines degrade (PECAN −0.011, MIPE −0.076, CheapNet −0.104).
3. *EpiFormer* with that same objective reaches 0.482 F1, a +0.146 gain over its own BCE-only version and larger than the gain for any baseline.

A loss function that was the real driver would be expected to help all architectures, whereas this one mainly does not help the late-fusion models, because bipartite edge prediction, sparsity regularization on the contacts, and inter-chain distance supervision need cross-chain representations. In a late-fusion model, there is nothing for those gradients to act on at the point where they are applied. Hence, the timing of fusion and the form of supervision are design choices.

F.5 Pooling strategy

Table 18: Pooling strategy comparison on the epitope-ratio split.

Pooling	AUC	F1	MCC
Max	0.917	0.473	0.448
Mean	**0.921**	**0.483**	**0.454**
Top-2	0.920	0.471	0.454
Top-3	0.923	0.481	0.458
Softmax Attn.	0.921	0.480	0.452

To verify that *EpiFormer*'s gains are not explained by pooling choice, we applied top-k pooling to antibody-aware baselines with explicit interaction matrices (Table 19). Top-k degrades all four baselines, most severely for MIPE (−0.331 F1), while *EpiFormer*'s F1 varies by only 0.011 across strategies.

Table 19: Top-k pooling applied to antibody-aware baselines. Original AUC/F1/MCC are the published configurations reported in Table 1; Δ F1 is computed relative to each method's original pooling on the same test split.

Method	Pooling	AUC	F1	MCC	Δ F1
WALLE	Original	0.808	0.203	0.208	—
WALLE	+ Top-2	0.809	0.159	0.145	−0.044
PECAN	Original	0.740	0.229	0.182	—
PECAN	+ Top-3 + edge loss	0.641	0.198	0.118	−0.031
MIPE	Original	0.827	0.337	0.356	—
MIPE	+ Top-2	0.651	0.006	0.022	−0.331
CheapNet	Original	0.719	0.184	0.149	—
CheapNet	+ Top-2 + edge loss	0.724	0.175	0.141	−0.009
EpiFormer	Mean pooling	0.924	0.482	0.464	—
EpiFormer	Top-2 pooling	0.920	0.471	0.454	−0.011
EpiFormer	Top-3 pooling	0.923	0.481	0.458	−0.001

F.6 Cross-attention gate initialization

Table 20 shows the effect of the cross-attention gate initialization α on performance. Removing encoder cross-attention entirely (α=0.0) causes a notable drop in F1. Even when initialized at zero, the model learns non-zero gating values, confirming that it discovers cross-attention utility without explicit supervision. The default initialization (α=0.05) achieves the best results. Figure 10 visualizes these trends across all four metrics.

Table 20: MHCA gate initialization ablation (epitope-ratio split).

Alpha Init	AUC	AUPRC	F1	MCC	Prec.	Δ F1
0.0 (no MHCA)	0.906	0.464	0.454	0.424	0.354	−0.042
0.05 (base)	**0.926**	**0.509**	**0.496**	**0.475**	**0.383**	—
0.2	0.918	0.485	0.469	0.449	0.351	−0.027
0.4	0.919	0.490	0.484	0.457	0.383	−0.012

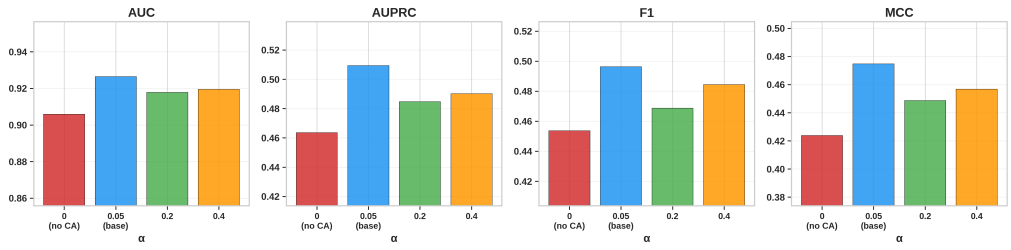


Figure 10: Test set performance across four interleaved-MHCA gate initialization values ($\alpha \in \{0.0, 0.05, 0.2, 0.4\}$).

F.7 Protein language model features

We compared our geometric feature representation against protein language model (PLM) embeddings across EpiFormer and four baselines. PLM embeddings $\mathbf{z}_i^{\text{plm}} \in \mathbb{R}^{d_c}$ are extracted from pre-trained models (ESM-2 (Lin et al., 2023) for antigens, AntiBERTy for antibodies) and projected via $\mathbf{h}_i^{\text{plm}} = \mathbf{W}_{\text{plm}}\mathbf{z}_i^{\text{plm}}$. Table 21 shows that EpiFormer achieves higher performance with geometric features than with PLMs, while all baselines degrade when PLMs are removed. This is consistent with the observation that epitope residues are not evolutionarily conserved (Ponomarenko & Bourne, 2007), making the conservation signals encoded by PLMs uninformative for binding prediction. Additionally, combining high-dimensional PLM features (1,280D) with lower-dimensional geometric features

(105D) creates a modality dominance problem (Wu et al., 2022) where the model overfits to the PLM pathway.

EpiFormer improves without PLMs, while other methods relying on PLMs for their primary representations degrade substantially. This suggests that the interleaved cross-attention captures the relational binding signal directly from geometric features, whereas other architectures depend on PLMs to compensate for weaker structural reasoning.

Table 21: PLM sensitivity analysis on the epitope-ratio test set. PLM embeddings are replaced with geometric features.

Method	Input Features	AUC	F1	MCC	△ AUC
EpiFormer	ESM2-650M + AntiBERTy	0.889	0.433	0.404	—
EpiFormer	**Geometric**	**0.924**	**0.482**	**0.464**	**+0.035**
EquiformerV2	ESM-2 (default)	0.815	0.241	0.243	—
EquiformerV2	Geometric	0.798	0.241	0.230	−0.017
DiscoTope3	ESM-IF1 (default)	0.817	0.249	0.243	—
DiscoTope3	Geometric	0.684	0.166	0.123	−0.133
WALLE	ESM-2 + AntiBERTy	0.809	0.202	0.206	—
WALLE	Geometric	0.687	0.145	0.109	−0.122
MIPE	ESM + AbLang	0.827	0.337	0.356	—
MIPE	Geometric	0.739	0.148	0.184	−0.088

F.8 Loss hyperparameter sensitivity

In order to verify that our result is not an artifact of hyperparameter selection, we swept each of the seven weights of the joint objective over the range given in Appendix B.5, holding the rest at their selected values (Figure 11). It can be seen that most weights show flat trends on both splits, or retain a broad plateau rather than a sharp optimum.

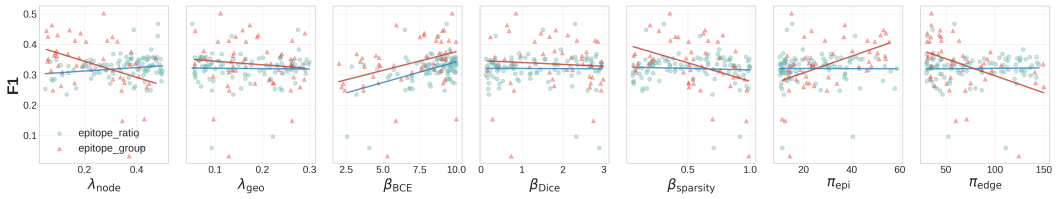


Figure 11: F1 sensitivity to seven loss hyperparameters across both splits. Most weights show flat trends, indicating robustness.

G Interpretability analyses

In this section, we examine what the trained model represents, from the separability of its residue embeddings to the asymmetry it learns between the two chains.

G.1 Representation quality

Figure 13 visualizes the learned antigen representations via t-SNE, showing that epitope residues form distinct clusters separated from non-epitope residues, with some overlap reflecting the inherent difficulty of the classification task. Figure 12(a) quantifies this separation, showing that silhouette scores increase progressively from the early encoder layers through the decoder, which confirms that each layer improves class separability. Figure 12(b) shows the learned gating weights for the default configuration (α_{init}=0.05); the antibody gate consistently exceeds the antigen gate across all encoder blocks, consistent with the learned asymmetry between antigen-antibody complexes.

We evaluated different strategies for mapping the bipartite interaction matrix $\hat{\mathcal{E}}_{\text{bg}}$ to per-residue epitope probabilities (Table 18). All pooling strategies converge to a narrow F1 range (0.471–0.483), with mean pooling achieving the best F1 while being the simplest. This suggests that

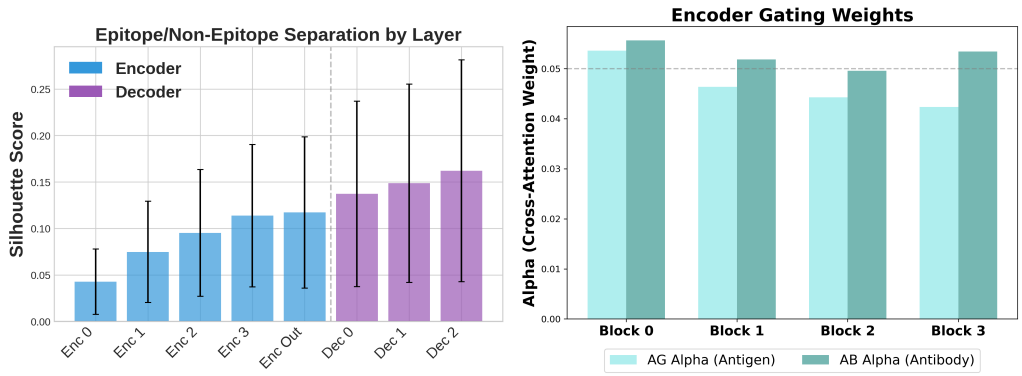


Figure 12: (a) Layer-wise silhouette scores showing progressive class separability from encoder to decoder. (b) Learned gating weights (α_{ag}, α_{ab}) controlling the cross-attention contribution per encoder block (default initialization α=0.05).

the interleaved encoder produces a well-calibrated interaction matrix largely independent of the downstream aggregation choice.

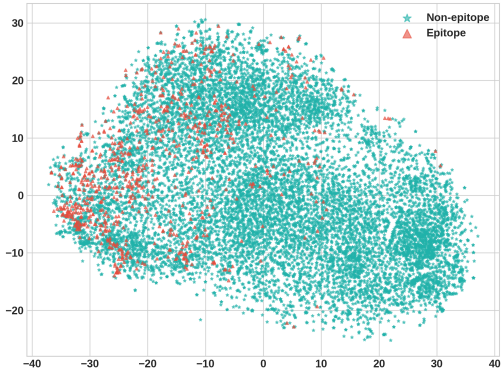


Figure 13: t-SNE visualization of antigen residue embeddings from the final encoder layer, colored by epitope (red) and non-epitope (teal) labels.

G.2 Encoder and Decoder Cross-Attention Maps

Figure 14 shows encoder cross-attention heatmaps for a representative complex (6xq0_1P) across the four α variants and all encoder blocks. When α=0.0 (no cross-attention), attention maps remain diffuse with no structure. With cross-attention enabled, attention progressively concentrates on ground-truth epitope and paratope residues across deeper encoder blocks. Figure 15 extends this analysis to the decoder, comparing decoder cross-attention maps and predicted interaction matrices against ground-truth contacts. All cross-attention variants produce focused decoder attention, but the α=0.05 configuration yields the sharpest predicted interaction matrix.

G.3 Coordinate Refinement

Figure 16 shows per-block coordinate displacement magnitude across the EGNN-R encoder blocks. Displacements decrease across blocks, confirming convergent geometric refinement. Epitope residues undergo consistently larger positional adjustments than non-epitope residues, and paratope residues show a similar pattern on the antibody side. The difference is statistically significant at Block 3 ($p = 1.80 \times 10^{-7}$, Wilcoxon rank-sum test).

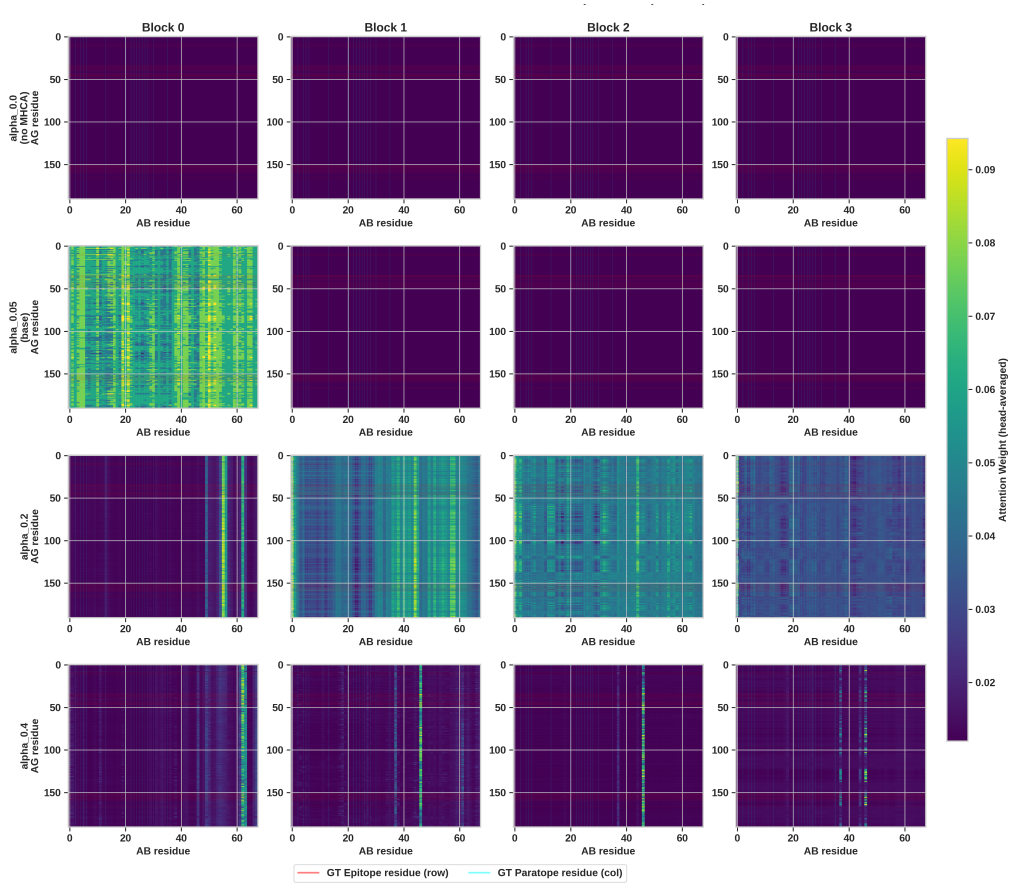


Figure 14: Encoder AG→AB cross-attention heatmaps for complex 6xq0_1P. Rows: $\alpha_{\text{init}} \in \{0.0, 0.05, 0.2, 0.4\}$; columns: encoder blocks 0–3. Ground-truth epitope (red) and paratope (cyan) residues are overlaid.

H Discussion

In this section, we set out the role of *EpiFormer* relative to full complex prediction, and its limitations.

H.1 Practical scope relative to co-folding and docking

If a co-folding model predicts an antibody-antigen complex accurately, the epitope can be derived directly from the interface. Co-folding also solves a different problem at a different cost. These methods predict a full complex with the epitope as a byproduct, whereas *EpiFormer* predicts the residue-level interface directly, conditioned on a given antibody. We note four main applications of *EpiFormer*.

- Co-folding remains unreliable on antibody-antigen interfaces specifically, as reported in Appendix C.3.
- Sparse residue-level classification benefits from an objective built for it, since a co-folding model optimizes structural accuracy and not per-residue labels.
- Many downstream uses need only the epitope map, including epitope binning, escape analysis and antigen engineering.
- Throughput matters at screening scale, where *EpiFormer* produces predictions at 118 ms and 0.174 GB per complex.

We therefore present *EpiFormer* as a complement to co-folding.

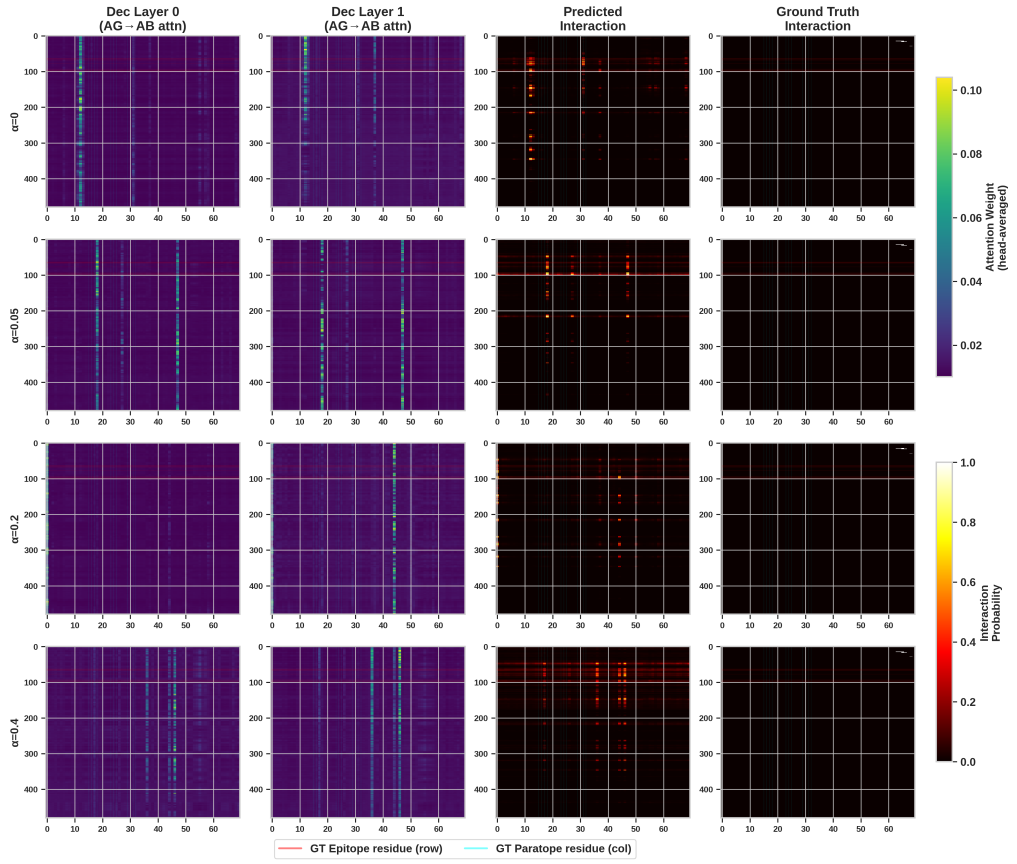


Figure 15: Decoder cross-attention and predicted interaction maps vs. ground-truth contacts for complex 6xq0_1P across four α_{init} variants. Columns: decoder layers 0–1, predicted interaction, ground truth.

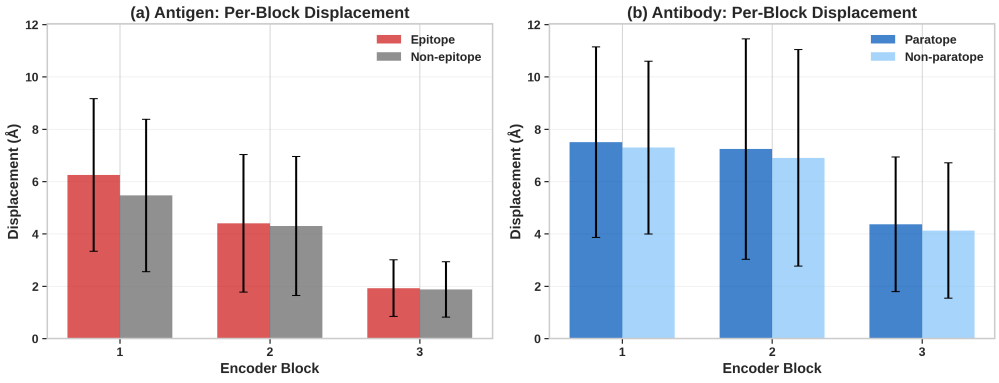


Figure 16: Per-block coordinate displacement magnitude across EGNN-R encoder blocks 1–3 (block 0 has no prior coordinates to compare against). (a) Antigen: epitope residues undergo larger displacements than non-epitope residues. (b) Antibody: paratope residues show a similar pattern. Displacements decrease across blocks, indicating convergent refinement.