
EPIC: Epitope–Paratope Interleaved Contact Learning for Antibody–Antigen Interface Prediction

Yijingxiu Lu^{*,†}, Yin Li^{*}, Wuyu Yang^{*}, Xinyu Chen, Tianyuan Wang[†]

Clickmab Biotechnology

solanoon0113@gmail.com, wangtianyuan@clickmab.com

^{*}Equal contribution. [†]Corresponding authors.

Abstract

Predicting antibody–antigen interfaces directly from sequence is challenging because residue–residue contacts are sparse, the interface depends on the specific antibody–antigen pair, and predictions must remain consistent between residue–pair contacts and residue-level binding sites. We present EPIC, a geometry-aware multi-task framework for predicting antibody–antigen contacts, paratopes, and epitopes from sequence inputs. EPIC combines pretrained protein language model representations with sequence-derived geometry obtained from predicted monomer structures, without requiring experimentally determined structures at inference time. Its geometry-aware interleaved interaction block integrates dense sequence self-attention with geometry-aware self-attention over local distance, direction, and relative-orientation features, and uses contact-gated cross-attention to enable information exchange conditioned on the specific antibody–antigen pair. A low-rank pair decoder models residue–pair interactions, while pairwise contact supervision and antibody- and antigen-side marginal supervision jointly constrain a shared representation of the interface. Task-specific readouts produce contact, paratope, and epitope predictions. We evaluate EPIC on SAbDab2, with additional evaluation on AsEP to assess out-of-distribution generalization. EPIC achieves the best contact AUROC and recall, matches a dedicated sequence-only predictor in paratope AUROC, and trails methods using experimental structures in epitope ranking. These results show that interleaving pretrained sequence representations with predicted geometric priors enables a unified framework for jointly predicting epitopes, paratopes, and residue–residue contacts from antibody–antigen sequences.

1 Introduction

Accurate characterization of antibody–antigen interfaces provides structural hypotheses for molecular recognition, binding-site localization, and residue-level constraints for complex modeling. However, predicting these interfaces remains challenging because antibody–antigen interactions are sparse and partner-specific, experimentally resolved complex structures remain limited, and binding can involve conformational rearrangements that are difficult to capture computationally [McCoy et al., 2024, Yin and Pierce, 2024]. These limitations are particularly relevant in practical settings, where sequence information is readily available but reliable complex structures are often unavailable. Biologically, the epitope and paratope form the two complementary sides of the same binding interface, while residue–residue contacts specify their direct interactions; epitope, paratope, and contact prediction are therefore intrinsically coupled aspects of antibody–antigen interface prediction.

Recent antibody-aware approaches have incorporated bipartite graphs, multimodal interaction modeling, and cross-attention to capture partner-specific interactions [Liu et al., 2024, Wang et al., 2024, Ahmed et al., 2026]. Despite this progress, many existing methods rely on experimentally resolved or

otherwise explicit structural information, which limits their applicability when only antibody and antigen sequences are available. In addition, predictions at the epitope, paratope, and residue-pair levels are often modeled with partially independent objectives or readouts, even though they describe the same physical interface. Such decoupling can lead to inconsistent predictions across tasks and may cause optimization or model selection to favor one interface resolution at the expense of another.

The central challenge is therefore to predict antibody–antigen interfaces from sequence while preserving consistency between pairwise contacts and residue-level binding sites. To this end, we introduce **EPIC**, an epitope–paratope interleaved contact learning framework that integrates pretrained sequence representations with sequence-derived intra-chain geometric information and jointly models epitope, paratope, and residue-pair contacts. Our main contributions are as follows:

- We propose EPIC, a sequence-only framework that jointly predicts epitopes, paratopes, and residue-pair contacts as complementary views of the antibody–antigen interface. Its *Geometry-Aware Interleaved Interaction Block* integrates PLM representations, sequence-derived local geometry, and cross-chain attention to model partner-specific interactions.
- We design a *Symmetric Low-Rank Pair Decoder* that supports contact prediction and residue-level readouts through a shared pairwise representation. Joint supervision promotes strong agreement between contact-derived and residue-level predictions ($\rho \approx 0.97$ – 0.99 ; Appendix D), supporting consistent interface prediction across tasks.
- Experiments on SAbDab2 demonstrate the effectiveness of EPIC’s unified architecture, with leading contact AUROC and recall and paratope AUROC comparable to a dedicated sequence-only predictor. Additional evaluation on AsEP characterizes its out-of-distribution generalization across sequence-identity groups.

2 Related Work

Current computational interface prediction methods can be broadly divided into three categories according to their input information: *structure-based*, *sequence–structure*, and *sequence-only* methods. Among *structure-based* approaches, ImaPEp represents candidate paratope–epitope patches as two-dimensional images derived from antibody–antigen complex structures and uses a convolutional neural network to predict their binding compatibility [Li et al., 2024]. *Sequence–structure* methods combine residue identity or learned sequence representations with geometric information: DiscoTope 3.0 integrates antigen sequence features with inverse-folding representations and surface features derived from solved or predicted structures for B-cell epitope prediction [Høie et al., 2024]; EpiFormer encodes antibody and antigen geometric graphs with interleaved bidirectional cross-attention for antibody-aware epitope prediction [Ahmed et al., 2026]; and MIPE combines sequence and structural modalities through contrastive learning while estimating an antibody–antigen interaction matrix to jointly predict paratopes and epitopes [Wang et al., 2024]. *Sequence-only* methods do not use experimentally determined structures as model inputs during training or inference: SPPIDER-seq uses ESM-2 [Lin et al., 2023] embeddings and cross-attention for partner-aware protein–protein interface-site prediction [Porollo et al., 2026]; ParaDeep combines chain-aware bidirectional LSTMs and one-dimensional convolutions for residue-level paratope prediction directly from antibody sequences [Udomwong et al., 2025]; and Paraplume leverages protein-language-model embeddings for scalable, antigen-agnostic paratope prediction [Athènes et al., 2025].

3 Methodology

3.1 Geometry-Aware Residue-Pair Contact Backbone

Figure 1 illustrates the EPIC framework, which takes antibody and antigen sequences as input. A frozen ESM-2 model extracts residue-level embeddings, while OpenDDE [Aureka AI OpenDDE project, 2026] predicts the monomer structures. The N, C α , and C backbone coordinates of these predicted structures provide the geometric information for intra-chain encoding.

The antibody and antigen representations are then processed in parallel by separate intra-chain encoders with unshared parameters. Each encoder comprises two branches: a *dense sequence branch*, which applies multi-head self-attention with RoPE across the full chain, and a *local geometry branch*,

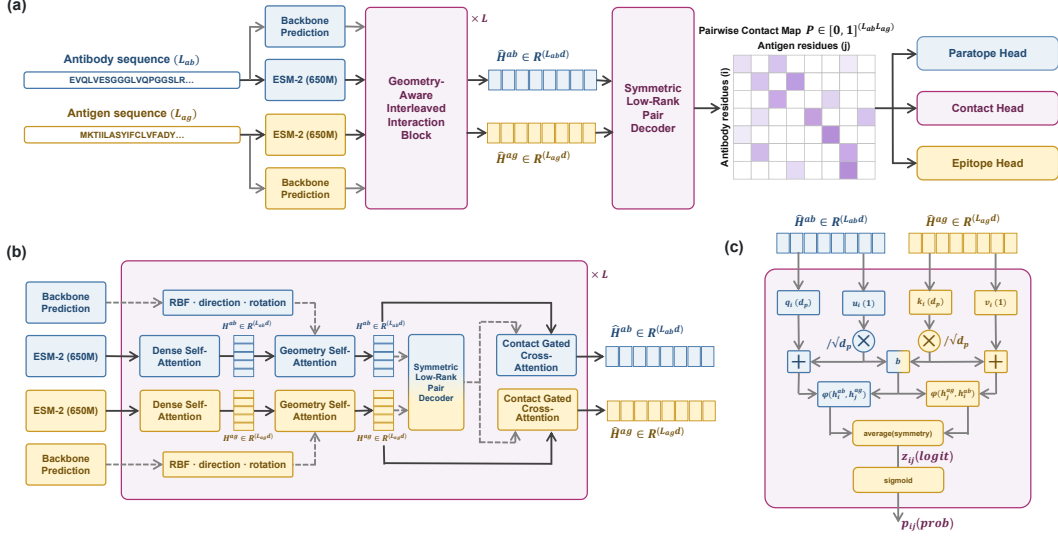


Figure 1: **Model overview of EPIC.** (a) EPIC combines PLM representations with geometry derived from sequence-predicted monomer backbones to jointly predict residue-pair contacts, paratopes, and epitopes. (b) The **Geometry-Aware Interleaved Interaction Block** fuses dense content self-attention with geometry-aware self-attention over distance, direction, and relative-rotation features, followed by contact-gated bidirectional cross-attention for partner-specific information exchange. (c) The **Symmetric Low-Rank Pair Decoder** computes directed antibody-antigen interaction scores and symmetrically aggregates them into residue-pair contact logits.

which derives geometric biases from the predicted backbone coordinates. Specifically, for each residue i , we define a right-handed local coordinate frame (R_i, o_i) using its N, C α , and C atoms, with the origin at C α . For each ordered intra-chain residue pair (i, j) , we construct geometric features g_{ij} that are invariant to global SE(3) transformations and encode inter-residue distance, local direction, relative orientation, and sequence separation:

$$g_{ij} = [\text{RBF}_{16}(\|o_j - o_i\|_2), R_i^\top \text{unit}(o_j - o_i), \text{vec}(R_i^\top R_j), \delta_{ij}^{\text{sgn}}, \delta_{ij}^{\text{abs}}]. \quad (1)$$

The encoding RBF_{16} is a 16-dimensional radial-basis encoding of inter-residue distance, and vec flattens the relative rotation matrix. We set $\text{unit}(0) = 0$, $\delta_{ij}^{\text{sgn}} = \text{clip}((j - i)/32, -1, 1)$, and $\delta_{ij}^{\text{abs}} = \text{clip}(|j - i|/32, 0, 1)$.

The geometry branch maps g_{ij} to an additive bias for each attention head and computes attention over a local graph containing self-loops and up to 32 nearest same-chain neighbors within 16Å. The geometric biases modulate attention within this neighborhood to encode local spatial relationships. For chain $c \in \text{Ab, Ag}$, the dense sequence branch and geometry branch produce H_{dense}^c and H_{geom}^c , respectively, which are fused through a residual gate:

$$H_c = H_{\text{dense}}^c + \tanh(\gamma_{\text{geo}}^c) (H_{\text{geom}}^c - X_c), \quad c \in \{\text{Ab, Ag}\} \quad (2)$$

In the fusion equation, $X_c \in \mathbb{R}^{L_c \times d}$ denotes the input residue representations supplied to both branches, L_c is the length of chain c , and γ_{geo}^c is a learnable gate parameter controlling the geometric contribution. The resulting antibody and antigen representations, H_{Ab} and H_{Ag} , are passed to the *Symmetric Low-Rank Pair Decoder*. For antibody residue i and antigen residue j , the decoder computes an intermediate contact logit z_{ij} :

$$\begin{aligned} \phi(x, y) &= \frac{q(x)^\top k(y)}{\sqrt{d_p}} + u(x) + v(y) + b, \\ z_{ij} &= \frac{1}{2} \left[\phi(h_i^{\text{Ab}}, h_j^{\text{Ag}}) + \phi(h_j^{\text{Ag}}, h_i^{\text{Ab}}) \right] \end{aligned} \quad (3)$$

The functions q and k are learned projections into a d_p -dimensional pair space ($d_p = 128$), u and v are learned scalar-valued residue projections, and b is a scalar bias. The directed score $\phi(x, y)$ is averaged with $\phi(y, x)$ to make the pair score invariant to exchanging its two arguments.

The intermediate contact probabilities provide additive biases for bidirectional cross-attention. For attention head a , the antigen-to-antibody update is

$$\begin{aligned}\alpha_{ij}^{(a)} &= \operatorname{softmax}_{j:M_{ij}=1} \left(\frac{(q_i^{(a)})^\top k_j^{(a)}}{\sqrt{d_h}} + \lambda_a \sigma(z_{ij}) \right), \\ \hat{h}_i^{\text{Ab},(a)} &= \sum_{j:M_{ij}=1} \alpha_{ij}^{(a)} v_j^{(a)}, \\ q_i^{(a)} &= W_q^{(a)} h_i^{\text{Ab}}, \quad k_j^{(a)} = W_k^{(a)} h_j^{\text{Ag}}, \quad v_j^{(a)} = W_v^{(a)} h_j^{\text{Ag}}.\end{aligned}\tag{4}$$

The attention-head dimension is d_h , $W_q^{(a)}$, $W_k^{(a)}$, and $W_v^{(a)}$ are learned projection matrices, and λ_a controls the contact bias for head a . The candidate mask M_{ij} equals one for valid antibody-antigen residue pairs and zero otherwise, excluding padding and invalid positions. Softmax and value aggregation are performed over valid antigen residues for each antibody query. The reverse update exchanges the antibody and antigen roles and normalizes over valid antibody residues.

These modules form the Geometry-Aware Interleaved Interaction Block. We stack $L = 2$ blocks and apply the Symmetric Low-Rank Pair Decoder to the final residue representations to obtain contact logits for all three tasks. Below, h_i^{Ab} , h_j^{Ag} , and z_{ij} denote the final-block representations and logits.

The final residue-level prediction probabilities are determined by combining residue representations with contact evidence aggregated along the corresponding direction in the contact map. For antibody residue i , the paratope prediction probability is obtained from its residue representation and the contact evidence from the corresponding antigen residues; for antigen residue j , the epitope prediction probability is obtained analogously from its representation and the contact evidence from the corresponding antibody residues:

$$\begin{aligned}p_i^{\text{para}} &= \sigma(w_p^\top \text{LN}(h_i^{\text{Ab}}) + \text{LME}_{j:M_{ij}=1}(z_{ij})), \\ p_j^{\text{epi}} &= \sigma(w_e^\top \text{LN}(h_j^{\text{Ag}}) + \text{LME}_{i:M_{ij}=1}(z_{ij})), \\ p_{ij} &= \sigma(z_{ij}).\end{aligned}\tag{5}$$

The learned readout weights are $w_p, w_e \in \mathbb{R}^d$, and $\text{LME}_{j:M_{ij}=1}(z_{ij})$ and $\text{LME}_{i:M_{ij}=1}(z_{ij})$ denote contact-logit evidence aggregated over antigen residues and antibody residues, respectively. For a candidate index set S , $\text{LME}_{r \in S}(z_r) = \log \sum_{r \in S} \exp(z_r) - \log |S|$, a log-mean-exp that combines the contact evidence over the candidate partner residues while normalizing for their number. The quantities p_i^{para} , p_j^{epi} , and p_{ij} denote the predicted probabilities for the paratope, epitope, and residue-pair contact, respectively.

To address class imbalance in epitope prediction, we combine a class-balanced binary cross-entropy (CB-BCE) loss with two auxiliary terms: a Dice loss and a conditional-proportion loss. The Dice loss encourages overlap between the predicted epitope probabilities and the ground-truth labels, whereas the conditional-proportion loss encourages the predicted epitope fraction within each complex to match its ground-truth fraction. Both auxiliary terms operate on the epitope probabilities p_j^{epi} in Eq. (5) and are defined in Appendix A.2, Eq. (11). The paratope classification loss uses the same CB-BCE formulation, applied to paratope probabilities and labels. For contact prediction, $\mathcal{L}_c$ is a masked focal BCE loss with focusing parameter $\gamma = 1$; the masked BCE and masked focal BCE formulations are provided in Appendix A.2, Eqs. (9) and (10). We additionally use an auxiliary five-class distance loss, $\mathcal{L}_{\text{dist}}$, with bin boundaries at 4, 8, 16, and 32 Å. For each classification loss, we first average over examples within each observed class of each complex, then average equally over the observed classes, and finally average equally over complexes. The complete training objective is given in Appendix A.2, Eq. (8).

3.2 Experimental settings

Data and splits. We evaluate EPIC and baselines on a 3,663-complex SAbDab2 structural cohort [Capel et al., 2026]. We use the official SAbDab2 *ab-ag* split and evaluate all methods on the same test

split using the same label definitions, validity masks, and metric aggregation protocol. Additionally, we evaluate our model on AsEP [Liu et al., 2024] to assess out-of-distribution generalization across sequence-identity groups (Appendix F). For each antibody–antigen residue pair, we set $Y_{ij} = 1$ when the minimum heavy-atom distance between residues i and j is at most 4.5 Å. Applying this rule over the complete Cartesian candidate grid yields 190,509,766 candidate pairs, including 143,756 positive contacts. The corresponding residue-level labels of paratopes and epitopes are directly derived from the contact matrix, details of model training and evaluation are provided in the Appendix:

$$y_i^{\text{para}} = \bigvee_j Y_{ij}, \quad y_j^{\text{epi}} = \bigvee_i Y_{ij}. \quad (6)$$

4 Results

Table 1 summarizes EPIC’s performance across the three interface prediction tasks. Using sequence-derived inputs, EPIC achieves the highest contact AUROC and recall among the evaluated methods, exceeding ImaPEp by 21.6 and 3.5 percentage points, respectively. For paratope prediction, EPIC achieves AUROC of 0.9540 versus 0.9514 for Paraplume, with similar recall (0.7464 versus 0.7485). Epitope prediction remains more challenging: EPIC achieves lower AUROC and AUPRC than the strongest structure-informed baselines. The relatively high epitope recall should be interpreted in the context of the evaluation protocol: recall is computed globally at a selected operating point, whereas AUROC and AUPRC are evaluated per complex. At matched recall, EPIC’s epitope precision is approximately 1.4× the positive prevalence (Appendix B), indicating limited enrichment.

Table 1: Performance on SAbDab2 across contact, epitope, and paratope prediction. All methods are evaluated using the same split and protocol. Structure denotes experimentally determined structures; “–” indicates unsupported tasks.

Input modality	Method	Contact			Epitope			Paratope		
		AUROC	AUPRC	Recall	AUROC	AUPRC	Recall	AUROC	AUPRC	Recall
Structure-only	ImaPEp	0.7374	0.0957	0.3481	–	–	–	–	–	–
Structure + sequence	DiscoTope-3.0	–	–	–	0.6862	0.2577	0.5346	–	–	–
	EpiFormer	–	–	–	0.8008	0.2775	0.4785	–	–	–
	MIPE	–	–	–	0.8074	0.2804	0.5523	0.8811	0.6892	0.7954
Sequence-only	SPPIDER-seq	–	–	–	0.6718	0.1454	0.5339	0.9486	0.6511	0.7313
	ParaDeep	–	–	–	–	–	–	0.9089	0.5833	0.7349
	Paraplume	–	–	–	–	–	–	0.9514	0.7319	0.7485
	EPIC	0.9531	0.0761	0.3832	0.6572	0.2327	0.6097	0.9540	0.6962	0.7464

Given AsEP’s overlap with SAbDab2 training data, we assess OOD performance on 412 complexes below 0.5 sequence identity to the training set. Epitope AP is 0.186, versus 0.316 at identity ≥ 0.9 , indicating weaker generalization to less similar complexes (Appendix F). Scores are not directly comparable with SAbDab2 because AsEP evaluates only antibody CDRs and antigen surface residues.

5 Conclusion

We presented EPIC, a unified framework that jointly predicts antibody–antigen contacts, paratopes, and epitopes from sequence-derived representations and geometry. Its Geometry-Aware Interleaved Interaction Block integrates PLM embeddings with local intra-chain geometry, while a Symmetric Low-Rank Pair Decoder models cross-chain interactions to guide bidirectional information exchange and support residue-level predictions. Joint supervision connects the three tasks through a shared interface representation. On SAbDab2, EPIC achieves the highest contact AUROC and recall among the evaluated methods and paratope AUROC comparable to a dedicated sequence-only predictor, although its epitope AUROC and AUPRC remain below those of the strongest structure-informed baselines. Several limitations remain. Sparse contact labels pose challenges for learning and calibration, and performance decreases on low-identity AsEP complexes (Appendix F), highlighting the need for better out-of-distribution generalization. Predicted monomer structures may also fail to capture binding-induced conformational changes and flexible interfaces. The sensitivity of EPIC to backbone prediction quality has not yet been systematically evaluated. Future work will focus on learning under sparse supervision, generalization across more diverse binding partners and structural distributions, and modeling structural uncertainty and conformational variability.

References

- Mansoor Ahmed, Huirong Chai, Haoxin Wang, Hemanth Venkateswara, and Murray Patterson. EpiFormer: Learning antigen–antibody interactions for epitope prediction via geometric deep learning. *arXiv preprint arXiv:2606.04154*, 2026. URL <https://arxiv.org/abs/2606.04154>.
- Gabriel Athènes, Adam Woolfe, Thierry Mora, and Aleksandra M. Walczak. Paraplume: A fast and accurate paratope prediction method provides insights into repertoire-scale binding dynamics. *arXiv preprint arXiv:2510.05626*, 2025. URL <https://arxiv.org/abs/2510.05626>.
- Aureka AI OpenDDE project. Folding, reasoning, and scaling with open-source drug discovery engine, 2026.
- Henriette L. Capel, Odysseas Vavourakis, Benjamin H. Williams, Christopher R. Taylor, and Charlotte M. Deane. SAbDab2: The structural antibody database in the age of machine learning. *bioRxiv*, 2026. doi: 10.64898/2026.06.16.732554. URL <https://www.biorxiv.org/content/10.64898/2026.06.16.732554v1>.
- 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.
- Dong Li, Fabrizio Pucci, and Marianne Rooman. Prediction of paratope–epitope pairs using convolutional neural networks. *International Journal of Molecular Sciences*, 25(10):5434, 2024.
- 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.
- Chu’nan Liu, Lilian Denzler, Yihong Chen, Andrew Martin, and Brooks Paige. AsEP: Benchmarking deep learning methods for antibody-specific epitope prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0373.
- Katherine Maia McCoy, Margaret E Ackerman, and Gevorg Grigoryan. A comparison of antibody–antigen complex sequence-to-structure prediction methods and their systematic biases. *Protein Science*, 33(9):e5127, 2024.
- Aleksey Porollo, Om Jadhav, Aaron Alvarez, and Jichao Chen. SPPIDER-seq: Sequence-based partner-aware predictor of protein–protein interaction sites. *Bioinformatics*, 42(7):btac388, 2026. doi: 10.1093/bioinformatics/btac388.
- Piyachat Udomwong, Thanathat Pamonsupornwichit, Kanchanok Kodchakorn, and Chatchai Tayapiwatana. Paradeep: sequence-based deep learning for residue-level paratope prediction using chain-aware bilstm-cnn models. *Frontiers in Bioinformatics*, 5:1684042, 2025.
- Zhiwei Wang, Yongkang Wang, and Wen Zhang. Improving paratope and epitope prediction by multi-modal contrastive learning and interaction informativeness estimation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6053–6061, 2024. doi: 10.24963/ijcai.2024/669.
- Rui Yin and Brian G Pierce. Evaluation of alphafold antibody–antigen modeling with implications for improving predictive accuracy. *Protein Science*, 33(1):e4865, 2024.

A Experimental Setup and Evaluation Protocol

A.1 Data Splitting, Structural Deduplication, and Label Definition

The processed SAbDab2 subset used in this study contains 3,663 antibody–antigen complex instances derived from 2,548 distinct PDB entries. We adopt the *ab-ag* split provided by SAbDab2, comprising 2,887 training complexes, 321 validation complexes, and 455 held-out test complexes. The split is constructed by first clustering antibodies and antigens separately according to sequence similarity. On the antibody side, clustering is performed based on CDRH3, CDRH123, and CDRL123. On

the antigen side, protein antigens are clustered using a 40% sequence-identity threshold, whereas peptide, DNA, and RNA antigens use an 80% threshold. Structural instances sharing any underlying cluster are subsequently grouped into connected components, and all instances within the same connected component are assigned to the same data partition. The training set is used to optimize model parameters, the validation set is used for training-time monitoring and checkpoint selection, and the held-out test set is used exclusively for final performance evaluation and does not participate in model selection.

For antibody residue i and antigen residue j , we define $Y_{ij} = 1$ if the minimum distance between any non-hydrogen atom of residue i and any non-hydrogen atom of residue j in the native complex structure is no greater than 4.5 Å, and $Y_{ij} = 0$ otherwise. The model computes contact logits over the full $L_{\text{Ab}} \times L_{\text{Ag}}$ Cartesian candidate grid, with padded positions excluded by a candidate mask. Residue-level labels are derived from the same contact matrix:

$$y_i^{\text{para}} = \bigvee_j Y_{ij}, \quad y_j^{\text{epi}} = \bigvee_i Y_{ij}. \quad (7)$$

Table 2: Dataset statistics for three prediction tasks on SAbDab2. For paratope and epitope prediction, positive and negative counts refer to residues; for contact prediction, they refer to antibody–antigen residue pairs.

	Split	# Samples	Positive	Negative	Positive Ratio (%)
Paratope	Train	2,887	50,300	532,564	8.6298
	Valid	321	6,138	56,558	9.7901
	Test	455	8,670	81,268	9.6400
Epitope	Train	2,887	47,815	735,935	6.1008
	Valid	321	5,817	78,097	6.9321
	Test	455	7,968	89,284	8.1931
Contact	Train	2,887	110,155	156,061,683	0.0705
	Valid	321	13,815	15,850,701	0.0871
	Test	455	19,786	18,453,626	0.1071

A.2 Training Configuration and Model Selection

The model uses frozen ESM-2 embeddings, a residue dimension of 128, two intra-chain encoder layers, and eight attention heads. Training uses bf16 precision, 0.1 dropout, the AdamW optimizer, a learning rate of 3×10^{-4} , a weight decay of 10^{-4} , and a batch size of 1. Experiments use random seeds 17, 29, and 43, train for at most 20 epochs, and apply early stopping when the validation metric does not improve for 5 consecutive epochs.

The joint loss is

$$\mathcal{L} = \mathcal{L}_c + 0.1\mathcal{L}_{\text{dist}} + 0.1\mathcal{L}_p + 0.5\mathcal{L}_e, \quad \mathcal{L}_e = \mathcal{L}_{\text{CB-BCE}}^e + \mathcal{L}_{\text{Dice}}^e + 0.2\mathcal{L}_{\text{cond}}^e. \quad (8)$$

Binary losses use residue masks m_i^{Ab} , m_j^{Ag} and the candidate-pair mask M_{ij} . A pair is valid when $M_{ij} = 1$; padding, invalid residues, and excluded pairs do not contribute to contact loss, contact aggregation, or evaluation. Residue-level losses use only valid residue positions. Within each complex, losses are averaged separately over the observed positive and negative classes, then equally across those classes. The resulting losses are averaged equally across complexes.

Specifically, $\mathcal{L}_p$ and $\mathcal{L}_{\text{CB-BCE}}^e$ use masked BCE and masked class-balanced BCE, respectively, while $\mathcal{L}_c$ uses masked focal BCE with $\gamma = 1$. Given logits s , binary labels y , and the corresponding mask m , the masked binary loss is computed only at positions where $m = 1$:

$$\ell_{\text{BCE}}(s, y) = -y \log \sigma(s) - (1 - y) \log(1 - \sigma(s)), \quad (9)$$

$$\ell_{\text{focal}}(s, y) = (1 - p_t)^\gamma \ell_{\text{BCE}}(s, y), \quad p_t = y\sigma(s) + (1 - y)(1 - \sigma(s)). \quad (10)$$

The auxiliary distance loss $\mathcal{L}_{\text{dist}}$ is a five-class inter-chain classification loss with boundaries at 4, 8, 16, and 32 Å. The paratope classification term uses the same class-averaging procedure as $\mathcal{L}_{\text{CB-BCE}}^e$, applied to paratope predictions and labels.

To mitigate class imbalance in epitope prediction and further constrain the overall proportion of epitope predictions within each complex, the epitope loss augments the masked class-balanced BCE with a Dice auxiliary loss and a conditional-proportion auxiliary loss:

$$\begin{aligned} \mathcal{L}_{\text{Dice}}^e &= \frac{1}{B} \sum_b \left[1 - \frac{2 \sum_r m_{br}^{\text{Ag}} \pi_{br} y_{br} + \varepsilon}{\sum_r m_{br}^{\text{Ag}} \pi_{br} + \sum_r m_{br}^{\text{Ag}} y_{br} + \varepsilon} \right], \\ \mathcal{L}_{\text{cond}}^e &= \frac{1}{B} \sum_b \left| \frac{\sum_r m_{br}^{\text{Ag}} \pi_{br}}{\sum_r m_{br}^{\text{Ag}}} - \frac{\sum_r m_{br}^{\text{Ag}} y_{br}}{\sum_r m_{br}^{\text{Ag}}} \right|. \end{aligned} \quad (11)$$

The mask m_{br}^{Ag} identifies valid antigen residues, and π_{br} and y_{br} are the predicted epitope probability and label for residue r in complex b . Dice loss measures prediction–label overlap; conditional-proportion loss penalizes the difference between predicted and true epitope fractions within each complex. We use $\varepsilon = 10^{-6}$ for numerical stability, and B denotes batch size.

Validation losses, calibration, and metrics use the same validity masks as training. Invalid positions are excluded from validation statistics and checkpoint selection.

For complexes without positive contacts, only the negative-class loss actually observed within the masked region is computed, without constructing additional positive samples. Checkpoints are selected on the validation set only; the test split is not involved in this process. Training a single model requires one RTX 4090, bf16 precision, and batch size one, for 21.33 aggregate GPU-hours; a checkpoint has 2,395,374 parameters, and inference is a single forward pass per complex.

A.3 Evaluation Metrics

We evaluate the model using AUROC, AUPRC, and Recall. AUROC and AUPRC are evaluated on a *per-complex* basis: the corresponding metric is first computed independently for each antibody–antigen complex and then averaged across complexes. This evaluation protocol assigns equal weight to each complex and prevents complexes containing more residues from disproportionately dominating the reported performance.

For task t , let $\mathcal{E}_t^{\text{AUROC}}$ denote the set of complexes for which AUROC is well defined. The reported AUROC is computed as

$$\text{AUROC}_t = \frac{1}{|\mathcal{E}_t^{\text{AUROC}}|} \sum_{n \in \mathcal{E}_t^{\text{AUROC}}} \text{AUROC} \left(y_n^{(t)}, s_n^{(t)} \right). \quad (12)$$

Similarly, let $\mathcal{E}_t^{\text{AP}}$ denote the set of complexes containing at least one positive label. The reported AUPRC corresponds to the mean per-complex average precision (AP):

$$\text{AUPRC}_t = \frac{1}{|\mathcal{E}_t^{\text{AP}}|} \sum_{n \in \mathcal{E}_t^{\text{AP}}} \text{AP} \left(y_n^{(t)}, s_n^{(t)} \right). \quad (13)$$

Here, $\text{AP}(\cdot)$ denotes average precision computed from the precision–recall curve.

In contrast to AUROC and AUPRC, Recall is computed globally by aggregating predictions over all residues in the evaluation set rather than averaging over individual complexes:

$$\text{Recall}_t = \frac{\text{TP}_t}{\text{TP}_t + \text{FN}_t}. \quad (14)$$

Here, TP_t and FN_t denote the total numbers of true-positive and false-negative residues, respectively, across the entire evaluation set. Recall is aggregated globally rather than per complex so that it follows the reporting convention of the SAbDab2 benchmark and of the original publications of the compared

methods, whereas AUROC and AUPRC are averaged per complex to prevent larger complexes from dominating. The same aggregation conventions are applied to every method. Per-complex AP and globally aggregated recall summarize different aspects of performance and should not be interpreted as a single precision–recall pair. Additional operating-point analyses explicitly report pooled precision and recall computed from the same predictions.

B Matched-Protocol Baseline Comparisons

All methods in Table 1 use the same 455 SAbDab2 *ab-ag* test complexes, labels, validity masks, and evaluation protocol. AUROC and AP are averaged per complex; recall is pooled across complexes. The supplementary comparisons use the same protocol; confidence-interval procedures are defined in Appendix I. EpiFormer was retrained in its published geometric-only configuration using experimental bound-complex backbones, three seeds, and validation-selected checkpoints. Paraplume’s released model was evaluated unchanged; 69 test complexes overlap its training PDBs, so results are also reported on the 386 non-overlapping complexes. EPIC uses OpenDDE-predicted monomer geometry. Supplementary EPIC results use independent retraining because the submitted checkpoint could not be restored exactly, and their absolute values differ from Table 1. MIPE was evaluated under the same main-table protocol but was not additionally retrained for this comparison.

The independently retrained EPIC achieves paratope AP of 0.671 versus 0.387 for EpiFormer. On the 386 complexes without overlap with Paraplume’s training data, EPIC’s paratope AP is lower by 0.034, with the confidence interval excluding zero. At epitope recall 0.6097, EPIC achieves precision of 0.117, compared with 0.282 for EpiFormer. EPIC’s precision is approximately $1.4\times$ the positive prevalence of 0.082, and EpiFormer’s precision is $2.4\times$ EPIC’s.

The epitope recall of 0.6097 in Table 1 describes a selected operating point, not ranking discrimination. At threshold 0.5, recall is 0.19 for the full retrained model, 0.25 without either epitope auxiliary loss, 0.26 with Dice loss only, and 0.30 with conditional-proportion loss only. Matched-recall precision remains similar across these variants, and none of the contrasts is statistically supported. These experiments do not establish an effect of the auxiliary losses on recall. The threshold is applied to the epitope probabilities in Eq. (5).

Table 3: Matched-protocol comparison on the 455 SAbDab2 test complexes. EPIC is the retrained model with OpenDDE geometry (3 seeds); differences are EPIC minus baseline with paired PDB-block bootstrap 95% intervals.

Baseline	Metric	EPIC	Baseline	Difference
EpiFormer	Epitope AP	0.214	0.372	−0.158 [−0.178, −0.139]
	Epitope AUROC	0.614	0.776	−0.162 [−0.181, −0.143]
	Epitope precision at recall 0.1	0.183	0.497	−
	Epitope precision at recall 0.6097	0.117	0.282	−
	Paratope AP	0.671	0.387	+0.284 [+0.263, +0.304]
Paraplume	Paratope AP, all 455	0.671	0.710	−0.039 [−0.063, −0.014]
	Paratope AP, 386 without training overlap	0.658	0.691	−0.034 [−0.062, −0.004]

C Component Ablations

We trained two independent sets of ablation models on the official SAbDab2 split, each with seeds 17, 29, and 43. Campaign A includes 36 models: the full model and 11 variants, each trained with three seeds and without the distance auxiliary loss. Campaign B includes 39 models: an independent implementation of the Section 3 equations, 11 variants, and a matched variant without the distance auxiliary loss, each trained with three seeds. All variants use a common initialization protocol and validation-based checkpoint selection.

Table 4 reports $\Delta = \text{full} - \text{ablation}$ in per-complex AP; positive values indicate lower AP after removing a component. A contrast is marked supported (*) when all three seeds agree in sign and its simultaneous 95% confidence interval excludes zero. Intervals use paired resampling over PDB-entry blocks and account for the family of contrasts. Other contrasts are classified as inconclusive.

Table 4: Component ablations, Δ = full – ablation in per-complex AP (*: all seeds agree in sign and the simultaneous 95% interval excludes zero; brackets give that interval). Positive values mean the removed component helps. These campaigns are independent retrainings, so their absolute values are not comparable to Table 1.

Component removed	Task	Campaign A	Campaign B
Predicted-geometry branch	epitope	+0.031 [+0.011, +0.051]*	+0.025 [+0.008, +0.043]*
	paratope	+0.006 [−0.006, +0.018]	+0.018 [+0.005, +0.032]*
	contact	+0.0045 [+0.0006, +0.0085]*	+0.0031 [−0.0002, +0.0064]
Residue-level supervision (contact-only training)	paratope	+0.032 [+0.014, +0.049]*	+0.040 [+0.023, +0.058]*
	epitope	−0.000 [−0.019, +0.018]	+0.001 [−0.020, +0.023]
Contact supervision (residue-only training)	contact	+0.018 [+0.013, +0.023]*	+0.017 [+0.013, +0.022]*
	epitope	+0.010 [−0.007, +0.026]	+0.011 [−0.008, +0.029]
Contact gate (content-only cross-attention)	paratope	+0.021 [+0.008, +0.034]*	+0.007 [−0.005, +0.018]
	contact, epitope	inconclusive	
Cross-attention (removed entirely)	paratope	+0.026 [+0.011, +0.042]*	+0.016 [−0.001, +0.032]
Low-rank pair decoder (unary + bias)	contact	+0.001 [−0.003, +0.004]	−0.001 [−0.004, +0.003]
	paratope	−0.002 [−0.020, +0.015]	−0.017 [−0.030, −0.004]
Rank 128 → 1	all tasks	inconclusive	
Contact-marginal residue readout (unary readout)	all tasks	inconclusive	
Epitope Dice + proportion losses	epitope	+0.006 [−0.007, +0.019]	−0.000 [−0.019, +0.018]
Distance auxiliary	epitope	−	+0.011 [−0.006, +0.028]

The contact-only and residue-only variants also use different validation metrics for checkpoint selection, so these comparisons assess training regimes rather than isolated loss terms. The `no_geometry_branch` variant removes both predicted geometry and geometry-aware attention, measuring their joint contribution. Separating these contributions would require a variant that incorporates the same geometric features through a simpler mechanism, such as an additive pair bias; this variant was not evaluated.

Removing the predicted-geometry branch reduces per-complex epitope AP by approximately 0.03 in both campaigns, corresponding to an approximately 14% relative improvement from retaining the branch. This is the only architectural contribution supported in both campaigns. Removing residue-level supervision reduces paratope AP, and removing contact supervision reduces contact AP; the effect of contact supervision on epitope AP is not statistically supported. Neither campaign establishes an accuracy benefit from the low-rank pair decoder over a unary decoder, the contact-marginal residue readout, or the epitope auxiliary losses. Paratope improvements from the contact gate and cross-attention are supported in Campaign A but not Campaign B.

D Cross-Task Consistency

For each test complex, we compare residue-head predictions with contact-derived marginals computed using Eq. 5 (Table 5). Their rank correlations are $\rho \approx 0.97$ – 0.99 . Contact-derived epitope scores achieve AP comparable to the epitope head (0.217 versus 0.214). Removing contact evidence from the readouts leaves rank agreement similar, while removing contact supervision reduces paratope rank agreement to $\rho = +0.17$ and gives contact-derived paratope AP of 0.206 versus 0.670 for the residue head. These comparisons support the contribution of joint supervision to cross-task agreement.

Table 5: Cross-task consistency on the 455 test complexes, 3 seeds. The head–marginal ρ is the per-complex rank correlation between a residue head and the contact map’s own marginal, and the marginal-alone columns score residues using only that marginal. Results use the retrained models described in Appendix C; their absolute values differ from Table 1.

Condition	Epitope			Paratope		
	head–marginal ρ	AP, head	AP, marginal alone	head–marginal ρ	AP, head	AP, marginal alone
Full model	+0.97	0.214	0.217	+0.99	0.671	0.644
Contact evidence removed from readouts	+0.96	0.215	0.221	+0.97	0.677	0.641
No contact supervision (residue-only)	+0.75	0.204	0.188	+0.17	0.670	0.206
No residue supervision (contact-only)	+0.98	0.213	0.217	+0.98	0.631	0.642

E Training–Test Antigen Similarity

None of the 455 test complexes shares an antigen, antibody, or CDRH3 cluster with the training set. Every test antigen has local-alignment identity below 50% to every training antigen, normalized by test-antigen length (median 20%, 90th percentile 26%). The main-table results therefore evaluate antigens from clusters absent from training.

Test antigens with lower training-set similarity tend to be longer ($\rho = -0.28$) and have lower epitope prevalence and lower epitope and contact AP. After normalizing AP by prevalence, neither task shows a statistically supported association with sequence similarity (epitope Spearman $\rho = 0.04$, $p = 0.38$; contact $\rho = 0.005$, $p = 0.91$). A regression controlling for prevalence and antigen length also yields a similarity coefficient whose PDB-block bootstrap interval includes zero. The split contains no antigens closely related to training antigens and therefore cannot evaluate that regime. Generalization to unseen antibody families and conformationally flexible interfaces has not been separately evaluated. Independently predicted monomer backbones may not capture binding-induced conformational changes.

F Out-of-Distribution Evaluation on AsEP

To complement the SAbDab2 training–test similarity analysis (Appendix E), we evaluate the frozen SAbDab2-trained EPIC checkpoints on 1,553 AsEP complexes [Liu et al., 2024]. Both datasets derive from SAbDab and contain substantial sequence overlap: median identity to the SAbDab2 training set is 0.97, and 412 complexes (27%) have identity below 0.5. Table 6 reports performance by sequence-identity group.

Epitope AP is 0.316 for complexes with sequence identity ≥ 0.9 and 0.186 for the 412 complexes below 0.5 identity. In the low-identity group, epitope and contact AP are $1.9\times$ and $2.9\times$ their respective prevalences, compared with $2.6\times$ and $20\times$ on the SAbDab2 test set. AsEP evaluates antibody CDRs and antigen surface residues, whereas EPIC is trained on full chains. These differences limit direct comparison with Table 1; comparisons between identity groups use the same AsEP evaluation setting.

Table 6: Out-of-distribution evaluation on AsEP. The frozen SAbDab2-trained EPIC is scored on all 1,553 AsEP complexes, stratified by sequence identity to the SAbDab2 training set; the three main strata are disjoint, and the < 0.3 row is a subset of the < 0.5 stratum. Each cell reports per-complex AP, with the positive prevalence of the stratum in parentheses. Absolute values are not comparable to Table 1 because AsEP restricts the antibody axis to CDR residues and the antigen axis to surface residues.

Identity to SAbDab2 train	n	Contact AP (prev.)	Epitope AP (prev.)	Paratope AP (prev.)
< 0.5	412	0.012 (0.004)	0.186 (0.099)	0.450 (0.322)
of which < 0.3	34	0.009 (0.004)	0.164 (0.096)	0.412 (0.294)
0.5–0.9	187	0.011 (0.003)	0.187 (0.082)	0.414 (0.301)
≥ 0.9	954	0.021 (0.004)	0.316 (0.108)	0.443 (0.320)

G Contact Ranking at Practical Operating Points

Positive contacts constitute 0.0705% of training pairs. EPIC achieves higher contact AUROC and recall than ImaPEp but lower AUPRC (Table 1). The 455 test complexes contain 19,786 positive pairs among 18,473,412 candidates, giving a pooled prevalence of 0.107%. Table 7 reports top-ranked contact precision and recall for the retrained models used in Appendix C. Complexes without positive pairs contribute zero precision, and $L_{\min}$ denotes the shorter chain length.

Pooled precision is 0.016 at 10% recall and 0.008 at 50% recall; no threshold achieves 5% pooled precision at non-trivial recall. Top-ranked contacts show approximately 18–28-fold enrichment above prevalence, but most remain false positives. These predictions may help prioritize candidate interface regions or provide ranking priors for docking and restraint weighting, but their precision is insufficient for high-confidence residue-pair restraints. Under this class imbalance, AUROC should be interpreted alongside precision-based metrics.

Table 7: Contact ranking at practical budgets per complex. Approximate enrichment is calculated from the displayed mean per-complex precision divided by the pooled prevalence ($19,786/18,473,412 \approx 0.107\%$), then rounded to the nearest integer. Values are averaged over the three seeds of the retrained ablation model and are not comparable to Table 1.

Ranking budget per complex	Precision	Recall	Enrichment	True contacts recovered
top 10	0.030	0.007	$\sim 28\times$	0.30
top 50	0.026	0.031	$\sim 24\times$	1.3
top $L_{\min}$	0.023	0.068	$\sim 21\times$	2.9
top 1% of pairs	0.019	0.138	$\sim 18\times$	5.9

H Qualitative Case Studies

Figure 2 compares the interface predictions of EPIC and ImaPEp on two complexes. For 6WAR, EPIC recovers more native contacts and paratope residues, with the same number of recovered epitope residues. For 8VC7, EPIC recovers more contacts and paratope residues but fewer epitope residues. These examples illustrate that improved contact recovery does not necessarily increase residue-level epitope recovery. The aggregate comparisons are reported in Table 1.

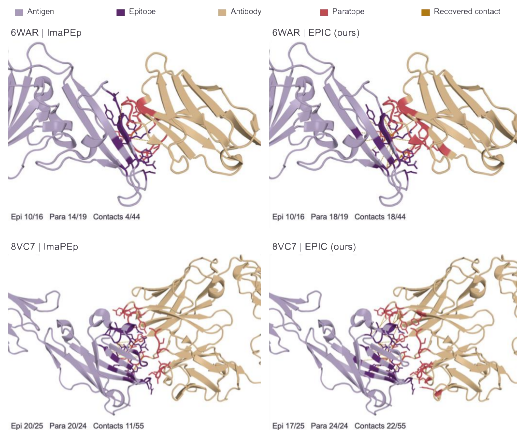


Figure 2: **Interface prediction examples for 6WAR and 8VC7.** ImaPEp predictions are shown on the left and EPIC predictions on the right. EPIC recovers more contacts and paratope residues in both examples, while epitope recovery is unchanged for 6WAR and lower for 8VC7.

I Statistical Uncertainty

Confidence intervals use a paired bootstrap with 10,000 replicates over 257 PDB-entry blocks, keeping complexes from the same PDB entry together. Per-complex scores are averaged over three seeds before resampling, and intervals for each family of contrasts are simultaneous across that family. Seed variability is reported separately as the standard deviation of mean per-complex scores across the three seeds.

For the retrained model whose per-complex AP is used in Appendix C and Appendix B, the metrics and their intervals are: contact AP 0.0219 [0.0191, 0.0248], epitope AP 0.214 [0.197, 0.233], and paratope AP 0.671 [0.649, 0.693]; the corresponding seed standard deviations are 0.001, 0.007, and 0.011, and the seed standard deviation of per-complex AUROC is at most 0.008 for every task.

The checkpoint used for Table 1 could not be restored exactly, so the ablation, ranking, and uncertainty analyses use independently retrained models. The reported intervals characterize these models and do not quantify uncertainty in the main-table values. All operating points, calibration maps, and thresholds were fitted on validation data and fixed before test evaluation; test results were not used for selection.