CoBALT: Coupling Antigen–Antibody Binding-Site Prediction and Affinity Ranking through a Contact-Energy Bottleneck

Anonymous authors
Paper under double-blind review

Abstract

Antigen–antibody (Ag–Ab) recognition involves binding-interface organization and collective intermolecular energetics, captured from complementary perspectives by binding-site prediction (BSP) and binding-affinity ranking (BAR). Although intrinsically connected, the two tasks operate at different resolutions: BSP localizes residue-level interfaces, whereas BAR evaluates complex-level affinity. Unifying these perspectives requires reconciling their distinct signals under scarce joint annotations. We introduce **CoBALT**, a unified framework that treats BSP and BAR as complementary readouts of a shared Ag–Ab interaction landscape. From partner-conditioned residue representations, CoBALT constructs a contact–energy bottleneck that factorizes pairwise interactions into a contact field localizing binding interfaces and an energy-like field encoding affinity-relevant local contributions. To learn from disjoint annotations, we develop an alternating multitask scheme that trains the shared bottleneck on separate BSP pairs and BAR candidate sets, enabling cross-task transfer without dual-annotated pairs. In five-fold cross-validation, CoBALT achieves epitope and paratope AUPRCs of 65.60% and 88.58%, respectively, while improving BAR Spearman correlation from 0.65 to 0.70 and Precision@1 from 50.77% to 56.99% over the strongest baselines. These gains largely persist under antigen-homology-aware holdouts, demonstrating that the contact–energy bottleneck effectively connects binding-site detection with affinity ranking.

1 Introduction

Antigen–antibody (Ag–Ab) recognition is mediated by a localized molecular interface (Kim et al., 2023), where fine-grained interactions collectively determine the binding affinity of the complex (Kim et al., 2023; Miller et al., 2023). This coupling between the binding interface and complex-level affinity is reflected in two closely related computational tasks. Binding-site prediction (BSP) identifies the epitope and paratope residues that form the interface for a given Ag–Ab pair (Akbar et al., 2021; Chinery et al., 2023), while binding-affinity ranking (BAR) compares candidate antibodies by their affinities for a target antigen (Guest et al., 2021; Hie et al., 2024). Despite their shared biological basis, BSP and BAR have largely developed along separate methodological lines. BSP methods predict epitopes from antigen features alone or incorporate antibody information to model partner-conditioned interfaces (Zeng et al., 2023a), whereas affinity models use sequence- or structure-based representations to estimate and compare binding strengths (Vecchietti et al., 2025). Consequently, existing methods generally do not jointly exploit binding-site and affinity supervision. AbRank takes an initial step toward bridging BSP and BAR by transferring an encoder pretrained on binding-site prediction to affinity ranking (Liu et al., 2025). However, this sequential transfer does not jointly optimize the two tasks or enable mutual supervision between them.

Nevertheless, joint modeling of the two tasks is complicated by differences in supervisory granularity and limited overlap in annotations. Binding-site annotations identify where binding occurs but do not quantify how individual interface residues contribute to affinity (Nada et al., 2024). Conversely, affinity rankings capture differences in binding strength without resolving the residue-level interactions underlying those differences (Keen et al., 2025). Effectively combining these complementary signals therefore requires preserving their distinct semantics while capturing the connections be-

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.

tween them. Furthermore, given the scarcity of Ag–Ab pairs annotated for both tasks, it remains important to make effective use of datasets in which only one form of annotation is available (Dunbar et al., 2014; Zhou et al., 2025).

Thus, we introduce **CoBALT**, a **Co**ntact–Energy **B**ottleneck framework for **A**g–Ab **L**earning across **T**asks, to unify supervision from binding-site predictions and affinity rankings without requiring joint annotations. At its core, CoBALT transforms partner-conditioned residue representations into a shared interaction landscape comprising a contact-propensity field and an energy-like field. This bottleneck provides a common basis for the two tasks: BSP localizes binding interfaces through predicted contacts, while BAR derives complex-level affinity from the associated interaction energies. CoBALT alternates between disjoint BSP pairs and BAR candidate sets, allowing binding-site supervision to ground the interaction landscape in physical interfaces and affinity supervision to shape it according to differences in binding strength. Our main contributions are as follows:

- We introduce CoBALT, the first unified framework for coupling residue-level binding-site prediction with complex-level affinity ranking under disjoint supervision, overcoming the reliance on scarce Ag–Ab pairs annotated for both tasks.
- We propose a novel contact–energy bottleneck that factorizes residue-pair interactions into contact-propensity and energy-like fields, enabling heterogeneous supervision to jointly shape a shared Ag–Ab interaction landscape through alternating optimization.
- Experiments under random five-fold cross-validation and antigen-homology-aware holdouts demonstrate consistent improvements over strong baselines, while extensive ablations and interaction-field analyses further validate the benefits of cross-task supervision.

2 Related Work

Binding-site prediction. Existing BSP approaches can be broadly divided into partner-independent and partner-conditioned methods. Partner-independent approaches predict epitope propensity solely from antigen features. GraphBepi and EpiGraph, for example, integrate sequence and structural features through graph-based modeling (Zeng et al., 2023b; Choi & Kim, 2024), whereas DiscoTope-3.0 combines inverse-folding representations with positive–unlabeled learning (Høie et al., 2024). In contrast, partner-conditioned approaches incorporate the cognate antibody to model antibody-specific recognition. PECAN, EpiScan, and EpiFormer capture such contextual dependencies using graph- and attention-based architectures (Pittala & Bailey-Kellogg, 2020; Wang et al., 2024a; Ahmed et al., 2026). More direct approaches explicitly represent cross-partner interactions at the residue-pair level. MIPE combines multimodal contrastive learning with interaction-matrix estimation (Wang et al., 2024b), while WALLE casts antibody-specific epitope prediction as bipartite link prediction (Liu et al., 2024). These studies underscore the importance of partner conditioning and explicit residue-pair interaction modeling for antibody-specific BSP.

Binding-affinity prediction and ranking. Computational methods leverage sequence and structural information to estimate Ag–Ab binding strength and prioritize antibody candidates. Sequence-based methods often build on pretrained representations such as AntiBERTy (Ruffolo et al., 2021). MINT captures cross-chain dependencies through attention (Ullanat et al., 2026), while MVSF-AB augments language-model representations with residue-level physicochemical features (Li et al., 2025). Structural information can be incorporated at different levels: SaProt encodes amino acid identity and local structure without requiring an assembled Ag–Ab complex (Su et al., 2024), whereas ANTIPASTI derives normal-mode correlation maps from complex structures (Michalewicz et al., 2024). Although predicted affinity scores can rank candidates, regression objectives do not directly optimize relative ordering. Ranking-based methods instead learn pairwise or listwise preferences: AbRank uses a margin-filtered pairwise objective (Liu et al., 2025), while AbLWR combines positive–unlabeled pretraining with context-aware listwise ranking for candidates targeting identical or homologous antigens (Xu et al., 2026).

Integrating binding-site and affinity learning. Prior Ag–Ab studies have linked binding-site and affinity learning primarily through knowledge transfer. For example, AbRank initializes its affinity model with a pretrained WALLE encoder but does not retain binding-site supervision during affinity training (Liu et al., 2025). Joint learning has also been explored in compound–protein systems:

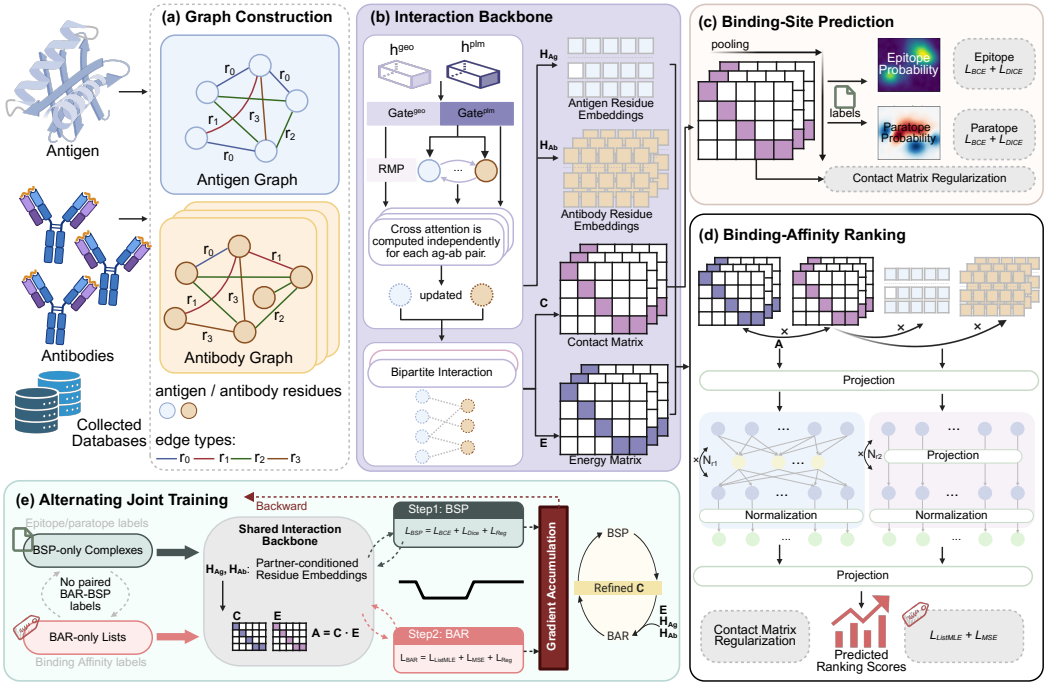


Figure 1: **Overview of the proposed CoBALT framework.** The architecture comprises the following key components: (a) Graph Construction: Antigen and antibody residues are represented as heterogeneous molecular graphs with geometric features and representations derived from protein language models (PLMs). (b) Interaction Backbone: A shared interaction backbone generates partner-conditioned antigen and antibody residue embeddings, $\mathbf{H}_{\mathrm{Ag}}$ and $\mathbf{H}_{\mathrm{Ab}}$, together with two pairwise interaction fields: a contact-propensity field C and an energy-like field E. (c-d) Task-Specific Decoders: (c) The binding-site prediction (BSP) decoder derives epitope and paratope predictions from contact-field marginals. (d) The binding-affinity ranking (BAR) decoder combines partner-conditioned residue embeddings with the contact–energy interaction fields for contextual ranking. (e) Alternating Joint Training: BSP- and BAR-only datasets are optimized through alternating joint training without requiring paired sample-level annotations.

MONN jointly predicts interactions and affinities using explicit pairwise interaction labels (Li et al., 2020), while DeepRelations derives affinity predictions from residue–atom relations under structural regularization and partial contact supervision (Karimi et al., 2021). However, to our knowledge, jointly learning BSP and BAR from disjoint supervision has yet to be explored in Ag–Ab modeling.

3 Methodology

Figure 1 provides an overview of CoBALT. In this section, we first formulate the BSP and BAR tasks and introduce the relevant notation, then describe the shared interaction backbone and task-specific decoders, and finally present the training strategy under disjoint supervision.

3.1 Preliminaries

Problem Definition. Let $\mathcal{A}$ denote the amino acid alphabet. Given an antigen, its amino acid sequence is denoted by $\mathbf{S}_{\mathrm{Ag}} = (b_1, \ldots, b_{L_{\mathrm{Ag}}})$, where $b_i \in \mathcal{A}$ is the amino acid type of residue i, and L_{Ag} is the number of antigen residues. For an antibody, we retain its six complementarity-determining regions (CDRs) and concatenate them into $\mathbf{S}_{\mathrm{Ab}} = (a_1, \ldots, a_{L_{\mathrm{Ab}}})$, where $a_i \in \mathcal{A}$ is the amino acid type of CDR residue i, and L_{Ab} denotes the total number of CDR residues. For each partner $p \in \{\mathrm{Ag}, \mathrm{Ab}\}$, its structure is represented as a residue-level point cloud $\mathbf{X}_p = \{\mathbf{x}_1^p, \ldots, \mathbf{x}_{L_p}^p\}$,

where $\mathbf{x}_i^p \in \mathbb{R}^{C\times 3}$ contains the coordinates of C atoms in residue i. An Ag–Ab pair is then defined as $\mathcal{X} \triangleq (\mathcal{X}_{\mathrm{Ag}}, \mathcal{X}_{\mathrm{Ab}})$, where $\mathcal{X}_p \triangleq (\mathbf{S}_p, \mathbf{X}_p)$.

Building on these notations, we formulate BSP and BAR as Ag–Ab interaction modeling at two complementary granularities. Specifically, BSP predicts residue-level epitope and paratope probabilities: $(\hat{\mathbf{y}}_{\mathrm{epi}}, \hat{\mathbf{y}}_{\mathrm{para}}) = f_{\mathrm{BSP}}(\mathcal{X})$, where $\hat{\mathbf{y}}_{\mathrm{epi}} \in [0,1]^{L_{\mathrm{Ag}}}$ and $\hat{\mathbf{y}}_{\mathrm{para}} \in [0,1]^{L_{\mathrm{Ab}}}$, with corresponding binary labels $\mathbf{y}_{\mathrm{epi}} \in \{0,1\}^{L_{\mathrm{Ag}}}$ and $\mathbf{y}_{\mathrm{para}} \in \{0,1\}^{L_{\mathrm{Ab}}}$. For BAR, we consider a ranking instance consisting of a fixed antigen and K candidate antibodies: $\mathcal{T} \triangleq \left\{ \left(\mathcal{X}_{\mathrm{Ag}}, \mathcal{X}_{\mathrm{Ab}}^{(k)} \right) \right\}_{k=1}^{K}$. The model assigns each candidate pair a binding score $\hat{r}_k = f_{\mathrm{BAR}} \left(\mathcal{X}_{\mathrm{Ag}}, \mathcal{X}_{\mathrm{Ab}}^{(k)}; \left\{ \mathcal{X}_{\mathrm{Ab}}^{(l)} \right\}_{l\neq k} \right)$, yielding $\hat{\mathbf{r}} = [\hat{r}_1, \ldots, \hat{r}_K]^\top$. The objective is to ensure that the ordering induced by $\hat{\mathbf{r}}$ is consistent with that induced by the ground-truth binding strengths $\mathbf{r} = [r_1, \ldots, r_K]^\top$. Overall, CoBALT jointly parameterizes the two task-specific mappings, $f_{\mathrm{CoBALT}} \triangleq (f_{\mathrm{BSP}}, f_{\mathrm{BAR}})$, which characterize Ag–Ab interactions within a unified framework.

Graph Construction We represent the antigen and antibody in $\mathcal{X}$ as separate residue-level molecular graphs $\mathcal{G}_{\mathrm{Ag}}$ and $\mathcal{G}_{\mathrm{Ab}}$, respectively. Each residue is represented as a node located at its C_α coordinate, with a node feature that combines geometric descriptors and a pretrained molecular embedding. Four types of intramolecular relations are introduced to capture residue connectivity (Zhang et al., 2022), and each edge is associated with a 100-dimensional feature characterizing the spatial and sequential relationship between the two incident residues (Wu et al., 2025). We represent each featurized residue graph as $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{R}; \mathbf{H}, \mathbf{F})$, where $\mathcal{N}$ and $\mathcal{E}$ denote the node and edge sets, $\mathcal{R}$ denotes the set of relation types, and $\mathbf{H}$ and $\mathbf{F}$ collect the node and edge features, respectively. Detailed definitions of these features and relations are provided in Appendix A.3.

3.2 Interaction Backbone

A central challenge in jointly modeling BSP and BAR is to learn shared interaction representations that support both residue-level interface identification and pair-level binding ranking. To this end, we design an interaction backbone that produces partner-conditioned residue representations and two complementary residue-pair fields: a contact field characterizing where interactions are likely to occur, and an energy-like field quantifying how individual residue pairs contribute to binding.

Partner-Conditioned Residue Representation. We first derive residue-level representations by integrating geometric features with PLM embeddings, and then condition these representations on the binding partner through intramolecular message passing and intermolecular cross-attention.

Specifically, for each residue i of molecular partner $p \in \{\mathrm{Ag}, \mathrm{Ab}\}$, we project its geometric feature and pretrained molecular embedding into a common latent space, yielding $\mathbf{h}_{i,p}^{\mathrm{geo}}$ and $\mathbf{h}_{i,p}^{\mathrm{plm}}$, respectively. The two representations are then combined through residue-specific gated fusion: $\mathbf{h}_{i,p} = \gamma_{i,p}^{\mathrm{geo}} \mathbf{h}_{i,p}^{\mathrm{geo}} + \gamma_{i,p}^{\mathrm{plm}} \mathbf{h}_{i,p}^{\mathrm{plm}}$. Stacking these residue embeddings yields the initial representation $\mathbf{H}_{(0)}^p$ for each molecular partner.

Subsequently, we apply L interaction blocks to jointly capture intramolecular structural context and intermolecular interactions. In each block ℓ, the incoming residue representations are layer-normalized and fed into two parallel branches (normalization is omitted from the equations for brevity). The relation-aware message-passing (RMP) branch (Satorras et al., 2021) captures the intramolecular structural context for each molecular partner $p \in \{\mathrm{Ag}, \mathrm{Ab}\}$:

$$\mathbf{M}_{\ell,\mathrm{intra}}^p = \mathrm{RMP}_\ell^p \left(\mathbf{H}_{\ell-1}^p, \mathcal{E}_p, \mathcal{R}_p, \mathbf{F}_p \right), \tag{1}$$

where $\mathcal{E}_p$, $\mathcal{R}_p$, and $\mathbf{F}_p$ denote its edge set, relation types, and edge features, respectively. In parallel, the bidirectional multi-head cross-attention (MHCA) branch (Vaswani et al., 2017) captures intermolecular context by enabling each molecular partner to aggregate information from its counterpart:

$$\mathbf{M}_{\ell,\mathrm{inter}}^p = \mathrm{MHCA}_\ell^{p\leftarrow q} \left(\mathbf{H}_{\ell-1}^p, \mathbf{H}_{\ell-1}^q \right), \qquad (p,q) \in \{(\mathrm{Ag},\mathrm{Ab}), (\mathrm{Ab},\mathrm{Ag})\}. \tag{2}$$

Finally, the normalized block input $\mathbf{H}_{\ell-1}^p$ is combined with the outputs of the two parallel branches $\mathbf{M}_{\ell,\mathrm{intra}}^p$ and $\mathbf{M}_{\ell,\mathrm{inter}}^p$ via additive residual aggregation to produce the updated partner-conditioned

residue representations $\mathbf{H}_\ell^p$. After L interaction blocks, the resulting representations $\mathbf{H}_{\mathrm{Ag}}$ and $\mathbf{H}_{\mathrm{Ab}}$ are subsequently used to construct the residue-pair interaction fields. Further computational details of the RMP and MHCA modules are presented in Appendices B.1 and B.2, respectively.

Residue-Pair Interaction Fields. Based on the partner-conditioned residue representations, we construct a contact field C and an energy-like field E over all Ag–Ab residue pairs, yielding two matrices of size $L_{\mathrm{Ag}} \times L_{\mathrm{Ab}}$. For each field type $t \in \{C, E\}$, we compute pairwise scores in both molecular directions:

$$S_{ij}^{t,1} = d^{-1/2} \left(W_{q,t}^{\mathrm{Ag}} \mathbf{h}_i^{\mathrm{Ag}}\right)^\top \left(W_{k,t}^{\mathrm{Ab}} \mathbf{h}_j^{\mathrm{Ab}}\right), \qquad S_{ij}^{t,2} = d^{-1/2} \left(W_{k,t}^{\mathrm{Ag}} \mathbf{h}_i^{\mathrm{Ag}}\right)^\top \left(W_{q,t}^{\mathrm{Ab}} \mathbf{h}_j^{\mathrm{Ab}}\right). \quad (3)$$

Here, $W_{q,t}^p$ and $W_{k,t}^p$ are learnable query and key projections for partner $p \in \{\mathrm{Ag}, \mathrm{Ab}\}$ and field t, respectively, and d is the projection dimension. The two directional scores are combined with field-specific learnable coefficients and a bias term as $S_{ij}^t = \pi_1^t S_{ij}^{t,1} + \pi_2^t S_{ij}^{t,2} + b_t$.

For the contact field, each pairwise logit is mapped to a score in $(0, 1)$ as $C_{ij} = \sigma\left(S_{ij}^C / \tau_C\right)$, where $\tau_C > 0$ controls the sharpness of the contact field. Meanwhile, the energy-like field transforms each pairwise logit into a bounded scalar $E_{ij} = B \tanh\left(S_{ij}^E / B\right)$, where $B > 0$ bounds the magnitude of the pairwise value.

3.3 Task-Specific Decoders

With these shared residue representations and interaction fields, we introduce two task-specific decoders, both of them operate on the same interaction backbone.

Binding-Site Prediction. The BSP decoder aggregates the contact field C into soft residue-level marginals. For each antigen residue i, we summarize its interactions with all antibody residues by a log-mean-exp over the contact scores, normalized by the antibody length. The paratope-side score is defined symmetrically for each antibody residue j:

$$u_i^{\mathrm{Ag}} = \log \sum_{j=1}^{L_{\mathrm{Ab}}} \exp\left(C_{ij}\right) - \log L_{\mathrm{Ab}}, \qquad u_j^{\mathrm{Ab}} = \log \sum_{i=1}^{L_{\mathrm{Ag}}} \exp\left(C_{ij}\right) - \log L_{\mathrm{Ag}}. \quad (4)$$

The resulting scores are mapped to epitope ($\hat{y}_i^{\mathrm{epi}}$) and paratope ($\hat{y}_j^{\mathrm{para}}$) probabilities through a scaled and shifted sigmoid.

Binding-Affinity Ranking. The BAR decoder ranks candidate antibodies against a fixed antigen by disentangling where the antigen and antibody interact from how strongly each contacting residue pair contributes. The contact field C characterizes the former, while the energy-like field E captures the latter. The decoder first pools the residue embeddings using residue weights derived from C, yielding contact-weighted representations of the antigen ($\tilde{\mathbf{H}}_{\mathrm{Ag}}$) and antibody ($\tilde{\mathbf{H}}_{\mathrm{Ab}}$). It further extracts permutation-invariant statistics $\phi(\cdot)$ from C, E, and their interaction $A = C \odot E$. These features form a pair token specific to the k-th antibody candidate:

$$\mathbf{z}_k = \mathrm{MLP}\Big(\Big[\tilde{\mathbf{H}}_{\mathrm{Ag}}; \tilde{\mathbf{H}}_{\mathrm{Ab}}; \phi(C); \phi(E); \phi(A)\Big]\Big). \quad (5)$$

Each pair token is processed by two complementary modules. An absolute module, implemented as a feed-forward network (FFN), independently encodes each candidate, whereas a contextual module, composed of multi-head self-attention blocks (Pang et al., 2020), models dependencies among candidates:

$$\mathbf{a}_k = f_{\mathrm{abs}}(\mathbf{z}_k), \qquad \mathbf{g}_k = \left[f_{\mathrm{ctx}}(\mathbf{Z})\right]_k, \quad (6)$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_K]^\top$ stacks all candidate pair tokens. The two representations are concatenated and linearly projected to obtain the final ranking score: $\hat{r}_k = \mathbf{w}_{\mathrm{out}}^\top [\mathbf{a}_k; \mathbf{g}_k] + b_{\mathrm{out}}$.

3.4 Training Mechanism

Since few Ag–Ab pairs have both residue-level binding-site annotations and affinity measurements, we adopt an alternating joint training strategy to exploit both sources of supervision without requiring paired annotations. Let $\mathcal{D}_{\mathrm{BSP}}$ denote the BSP dataset with epitope and paratope labels, and

$\mathcal{D}_{\mathrm{BAR}}$ the BAR dataset with affinity measurements. Mini-batches drawn from separate data loaders are processed alternately, while gradients are propagated through the shared interaction backbone.

For BSP, we combine weighted binary cross-entropy (BCE) and Dice losses for epitope and paratope prediction:

$$\mathcal{L}_{\mathrm{BSP}} = \sum_{p\in\{\mathrm{epi},\mathrm{para}\}} \left(\lambda^{p}_{\mathrm{BCE}}\mathcal{L}^{p}_{\mathrm{BCE}} + \lambda^{p}_{\mathrm{Dice}}\mathcal{L}^{p}_{\mathrm{Dice}}\right). \tag{7}$$

The weighted BCE terms address the imbalance between interface and non-interface residues, while the Dice terms encourage overlap with the annotated binding sites. For BAR, the predicted scores are optimized using mean squared error (MSE) and ListMLE:

$$\mathcal{L}_{\mathrm{BAR}} = \lambda_{\mathrm{MSE}}\mathcal{L}_{\mathrm{MSE}} + \lambda_{\mathrm{ListMLE}}\mathcal{L}_{\mathrm{ListMLE}}. \tag{8}$$

The MSE term preserves supervision from measured affinity values, whereas ListMLE optimizes the ordering of antibodies within each antigen-specific ranking set.

Since the contact field C is learned without pairwise contact annotations, we introduce a weak contact-rate prior for each task $\nu \in \{\mathrm{BSP}, \mathrm{BAR}\}$:

$$\mathcal{R}_{\nu} = \frac{1}{|\mathcal{B}_{\nu}|}\sum_{b\in\mathcal{B}_{\nu}}\left(\operatorname*{mean}_{i,j} C^{(b)}_{ij} - \rho_{\nu}\right)^{2}, \tag{9}$$

where $\mathcal{B}_{\nu}$ denotes a task-specific mini-batch and ρ_{ν} is the corresponding target contact rate. This weak prior stabilizes the scale of C without imposing a hard biological constraint.

Combining task-specific supervision with the contact-rate prior, we train the BAR and BSP objectives using an alternating joint training schedule. At optimization step t, let $\mathcal{V}_t \subseteq \{\mathrm{BAR}, \mathrm{BSP}\}$ denote the set of tasks scheduled to provide a mini-batch at that step. The model parameters are updated as

$$\theta_{t+1} = \theta_t - \eta_t \sum_{\nu\in\mathcal{V}_t} \lambda_{\nu}\nabla_{\theta}\left(\mathcal{L}^{(t)}_{\nu} + \lambda^{\nu}_{\mathrm{reg}}\mathcal{R}^{(t)}_{\nu}\right), \tag{10}$$

where θ denotes all trainable model parameters and η_t is the learning rate.

4 Experiments

4.1 Experimental Setup

Datasets and Splits. For BSP, we use ANABAG (Grandguillaume et al., 2025), retaining 8,521 Ag–Ab pairs with residue-level epitope and paratope annotations. For BAR, we curate 116,072 unique pairs from the antibody fitness dataset (Chungyoun & Gray, 2025) and convert their equilibrium dissociation constants (K_D) into affinity scores using $-\log_{10}(K_D\,[\mathrm{M}])$. Following the sequence-identity thresholds adopted in prior studies (Liu et al., 2025; Lee et al., 2026), we evaluate CoBALT under two complementary protocols: (1) a standard *random split* with five-fold cross-validation and (2) *antigen-homology-aware splits* at sequence-identity thresholds of 75% and 80%. For each protocol, the data are partitioned into training, validation, and test sets. To prevent within- and cross-task data leakage, antigen-homology-aware split is coordinated across tasks at the antigen level, ensuring that antigens from the same family are assigned to the same partition. Additional details on data preprocessing and split construction are provided in Appendices A.1 and A.2, respectively.

Baselines and Evaluation Metrics. We compare CoBALT against a broad range of competitive baselines across both BSP and BAR. The BSP baselines include the partner-independent methods GraphBepi (Zeng et al., 2023b), DiscoTope-3.0 (Høie et al., 2024), and EpiGraph (Choi & Kim, 2024), as well as the partner-aware methods PECAN (Pittala & Bailey-Kellogg, 2020), WALLE (Liu et al., 2024), EpiScan (Wang et al., 2024a), and EpiFormer (Ahmed et al., 2026). The BAR baselines comprise three sequence-based methods—MINT (Ullanat et al., 2026), MVSF-AB (Li et al., 2025), and AntiBERTy (Ruffolo et al., 2021)—and two structure-based methods, AbRank (Liu et al., 2025) and SaProt (Su et al., 2024).

Table 1: **Task-specific performance under the random five-fold split.** Results are reported as mean ± standard deviation across folds, with percentage-based metrics expressed in %. The best and second-best results are shown in **bold** and underlined, respectively. Summary (w/t/l) counts CoBALT's significant wins, ties, and significant losses against all baselines, based on two-sided paired t-tests with corrected p-values ($\alpha = 0.05$), where ties indicate no significant difference.

Task	BSP					
Method	Epitope			Paratope		
	AUPRC ↑	AUROC ↑	F1 ↑	AUPRC ↑	AUROC ↑	F1 ↑
PECAN Pittala & Bailey-Kellogg (2020)	23.58 ± 0.80	80.14 ± 0.64	28.64 ± 1.23	69.28 ± 0.69	77.07 ± 0.45	65.27 ± 0.65
WALLE Liu et al. (2024)	40.61 ± 1.27	88.57 ± 0.37	45.06 ± 0.74	69.59 ± 0.39	79.12 ± 0.18	67.94 ± 0.41
EpiScan Wang et al. (2024a)	16.32 ± 1.16	72.41 ± 1.08	16.76 ± 1.26	$\underline{82.80 \pm 0.76}$	$\underline{88.75 \pm 0.55}$	$\underline{78.88 \pm 0.84}$
EpiFormer Ahmed et al. (2026)	$\underline{56.22 \pm 0.68}$	$\underline{92.06 \pm 0.23}$	$\underline{51.21 \pm 0.67}$	48.24 ± 0.33	53.97 ± 0.71	53.26 ± 3.93
GraphBepi Zeng et al. (2023b)	30.31 ± 0.48	84.17 ± 0.40	12.14 ± 1.67	80.40 ± 0.68	87.99 ± 0.29	74.51 ± 0.76
DiscoTope-3.0 Høie et al. (2024)	14.65 ± 0.31	72.79 ± 0.24	21.36 ± 0.50	58.57 ± 0.85	68.47 ± 0.50	55.10 ± 0.52
EpiGraph Choi & Kim (2024)	40.60 ± 3.42	88.99 ± 0.75	29.25 ± 8.36	79.18 ± 4.16	85.24 ± 3.25	73.44 ± 4.66
CoBALT (ours)	$\mathbf{65.60 \pm 1.24}$	$\mathbf{92.71 \pm 0.83}$	$\mathbf{57.00 \pm 1.75}$	$\mathbf{88.58 \pm 0.58}$	$\mathbf{91.32 \pm 0.38}$	$\mathbf{80.73 \pm 0.75}$
Summary (w/t/l)	*7/0/0*	*7/0/0*	*7/0/0*	*7/0/0*	*7/0/0*	*7/0/0*
Task	**BAR**					
Method	MAE ↓	Spearman ↑	Kendall ↑	P@1 ↑	Hit@3 ↑	PairAcc ↑
AntiBERTy Ruffolo et al. (2021)	0.76 ± 0.02	0.53 ± 0.01	0.45 ± 0.02	43.35 ± 1.48	64.05 ± 0.64	75.17 ± 0.60
MVSF-AB Li et al. (2025)	0.70 ± 0.01	0.63 ± 0.01	0.53 ± 0.02	45.05 ± 10.67	71.57 ± 0.72	80.41 ± 0.49
MINT Ullanat et al. (2026)	0.82 ± 0.01	0.52 ± 0.00	0.43 ± 0.01	32.94 ± 10.91	61.24 ± 0.26	75.05 ± 0.35
SaProt Su et al. (2024)	0.72 ± 0.03	0.60 ± 0.02	0.52 ± 0.02	47.23 ± 1.39	68.52 ± 1.06	78.85 ± 0.82
AbRank Liu et al. (2025)	$\mathbf{0.64 \pm 0.02}$	$\underline{0.65 \pm 0.01}$	$\underline{0.55 \pm 0.01}$	$\underline{50.77 \pm 0.69}$	$\underline{71.91 \pm 0.91}$	$\underline{81.18 \pm 0.62}$
CoBALT (ours)	$\underline{0.65 \pm 0.02}$	$\mathbf{0.70 \pm 0.01}$	$\mathbf{0.61 \pm 0.01}$	$\mathbf{56.99 \pm 1.13}$	$\mathbf{78.36 \pm 1.30}$	$\mathbf{84.42 \pm 0.36}$
Summary (w/t/l)	*4/1/0*	*5/0/0*	*5/0/0*	*5/0/0*	*5/0/0*	*5/0/0*

Performance is evaluated using six metrics for each task. In BSP, AUPRC, AUROC, and F1 at a decision threshold of 0.5 are reported separately for residue-level epitope and paratope prediction. BAR evaluation considers both absolute affinity estimation and within-antigen ranking. Mean absolute error (MAE) measures prediction error on the unified affinity scale, while Spearman's and Kendall's correlations assess ranking consistency. Precision@1 (P@1) and Hit@3 evaluate the retrieval of top binders. Pairwise ordering accuracy (PairAcc) additionally measures accuracy among antibody pairs with ground-truth affinity differences greater than 0.5 (Liu et al., 2025).

4.2 Task-specific Results

We first evaluate CoBALT under the random five-fold split, as summarized in Table 1. On the BSP task, CoBALT achieves the best performance across all metrics, with statistically significant improvements over the baselines. For epitope prediction, CoBALT achieves the highest AUROC of 92.71%. Under the substantial class imbalance of this task, it also improves the more informative AUPRC by 9.38 percentage points, from EpiFormer's 56.22% to 65.60%, and F1 by 5.79 points, from 51.21% to 57.00%. CoBALT further improves paratope prediction despite the stronger baseline performance, raising AUPRC from EpiScan's 82.80% to 88.58% and achieving the highest AUROC and F1 scores of 91.32% and 80.73%, respectively.

For BAR, CoBALT remains competitive with the best-performing baseline, AbRank, in absolute affinity estimation, while showing clear advantages in within-antigen ranking. It improves Spearman correlation from 0.65 to 0.70 and Kendall correlation from 0.55 to 0.61. Top-binder retrieval also improves, with P@1 increasing from 50.77% to 56.99% and Hit@3 from 71.91% to 78.36%. In addition, PairAcc rises from 81.18% to 84.42%. These results show improved prioritization of stronger binders within the evaluated antigen-specific ranking chunks.

4.3 Antigen-Homology-Aware Splits

To evaluate generalization beyond closely related antigens, we assessed CoBALT using antigen-homology-aware splits with sequence-identity thresholds of 80% and 75%. Results for the 80% split are shown in Figure 2, while those for 75% split are provided in Appendix Figure A1.

Under the 80% split, CoBALT outperforms EpiFormer on epitope prediction, EpiScan on paratope prediction, and AbRank on BAR across all reported metrics. For epitope prediction, CoBALT achieves absolute gains of 33.5%, 9.8%, and 34.2% in AUPRC, AUROC, and F1, respectively. It also improves the three paratope metrics by 10.6%, 5.0%, and 6.8%. For BAR, CoBALT reduces MAE by 0.534 metric units and improves the ranking metrics by 4.9%–9.2%. This trend largely persists under the more stringent 75% split, with CoBALT outperforming the strongest baselines in 23 of the 24 task–metric settings. The only exception is BAR MAE at the 75% threshold, where CoBALT's error is 0.046 metric units higher than that of the strongest baseline. Overall, these results demonstrate robust generalization to antigen-cluster holdouts, with consistent improvements in BSP and BAR ranking performance.

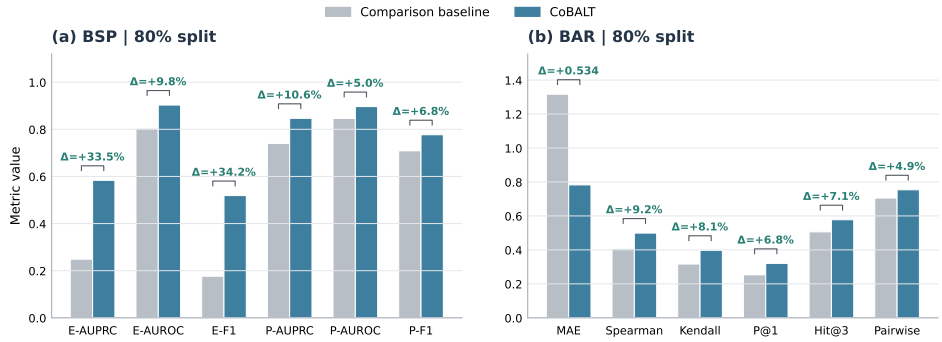


Figure 2: **Comparison with the strongest task-specific baselines on the 80% antigen-homology-aware split.** EpiFormer and EpiScan serve as the BSP baselines for epitope and paratope prediction, respectively, while AbRank serves as the BAR baseline. Gray and blue bars denote the baselines and CoBALT, respectively, with annotations indicating CoBALT's improvements.

4.4 Ablation Study and Analysis

We further conduct ablation studies to investigate the effects of joint training and the shared contact–energy bottleneck. As shown in Table 2, CoBALT-Single denotes separate models trained for BAR and BSP. *w/o BAR→C* blocks BAR gradients from the contact-propensity field C; *w/o C in BAR* and *w/o E* remove C or the energy-like field E, respectively; and *w/o interface in BAR* ranks candidates using only mean-pooled residue representations.

Compared with CoBALT-Single, joint training improves all BAR ranking metrics and substantially increases epitope AUPRC and F1 from 54.94% to 67.34% and from 46.90% to 59.25%, respectively, while consistently benefiting paratope prediction. These gains support our central premise that BSP and BAR can provide complementary supervision. Structure-based analysis also shows that the jointly learned interaction fields align better with structure-derived contacts than their single-task counterparts (Appendix Section C).

Blocking BAR gradients to C causes the largest ranking degradation, reducing Spearman from 0.70 to 0.54 and P@1 from 57.64% to 38.92%, while removing C, E, or the entire interface representation also impairs BAR. For BSP, removing the interface components generally impairs performance, while blocking BAR gradients to C yields mixed results, modestly improving some metrics but reducing epitope AUPRC and F1. Together, these results validate the contact–energy bottleneck as an effective bridge for cross-task supervision.

4.5 Case Study Analysis

To illustrate how CoBALT connects Ag–Ab binding interfaces with affinity predictions, we analyze two antigen panels: human immunodeficiency virus type 1 (HIV-1) envelope glycoprotein gp120 paired with three candidate antibodies, and hen egg-white lysozyme (HEWL) paired with nine. Both panels provide binding-site annotations and affinity measurements. The structural visualizations in Figure 3(a) and Appendix Figure A2(a), together with the residue-level prediction profiles in Appendix Figure A3, show that CoBALT identifies all annotated binding interfaces.

Table 2: **Comparison with single-task counterparts and component ablations on fold 1 of the random split.** Percentage-based metrics are reported in %, with PairAcc evaluated at a threshold of 0.50. The best and second-best results are shown in **bold** and underlined, respectively.

Model / Variant	BSP						BAR					
	Epitope			Paratope			MAE ↓	Spearman ↑	Kendall ↑	P@1 ↑	Hit@3 ↑	PairAcc ↑
	AUPRC ↑	AUROC ↑	F1 ↑	AUPRC ↑	AUROC ↑	F1 ↑						
CoBALT	**67.34**	93.69	**59.25**	**88.31**	91.19	80.41	0.68	**0.70**	**0.60**	**57.64**	**78.02**	**84.11**
CoBALT-Single	54.94	92.06	46.90	87.16	90.62	78.69	**0.67**	0.68	0.59	55.43	77.39	82.82
Component ablations												
w/o BAR→C	64.33	**94.21**	48.63	88.30	**91.37**	**80.72**	0.82	0.54	0.47	38.92	59.19	76.01
w/o C in BAR	58.59	92.91	42.52	84.43	88.44	77.31	0.68	0.65	0.56	50.15	72.42	81.56
w/o E	58.83	92.85	43.05	84.36	88.40	77.31	0.69	0.64	0.55	52.36	73.36	81.19
w/o interface in BAR	59.24	92.93	41.99	85.48	89.33	78.32	0.68	0.63	0.54	51.38	71.91	80.73

We further examine whether the learned interaction fields can distinguish antibodies that target similar regions yet differ in affinity (Appendix Figure A4). Despite exhibiting similar contact-field patterns, these antibodies show distinct energy-like signatures (Figure 3(c) and Appendix Figure A5(a)). Moreover, masking top-ranked contact regions causes larger changes in predicted affinity than masking random or low-ranked regions (Figure 3(b) and Appendix Figure A2(b)), indicating that the predictions rely preferentially on high-scoring contact-field entries. Together, these results demonstrate that CoBALT establishes a coherent link between binding-interface recognition and affinity discrimination through its learned interaction fields.

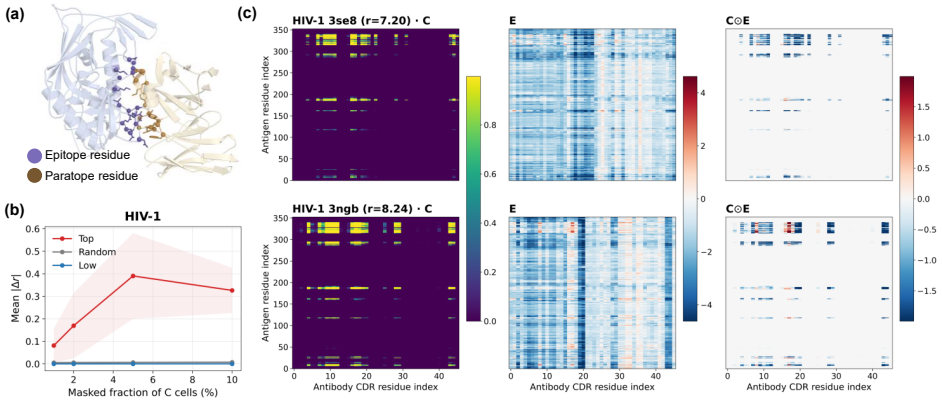


Figure 3: **Case study of the learned interaction fields for HIV-1 gp120.** (a) Structural visualization of the annotated epitope and paratope in the 3se9 complex. (b) Effects on predicted affinity of masking top-ranked, randomly selected, or low-ranked contact regions. (c) Contact fields, energy-like fields, and their element-wise products for the 3se8 and 3ngb complexes.

5 Conclusion

This work presents CoBALT, a unified framework for Ag–Ab binding-site prediction and affinity ranking under heterogeneous and disjoint supervision. Through a shared interaction backbone that learns structured contact and energy-like fields, CoBALT connects residue-level interface localization with complex-level affinity modeling, allowing binding-site annotations and affinity measurements to jointly inform Ag–Ab interaction learning without requiring paired labels. CoBALT consistently outperforms strong task-specific baselines under both random and antigen-homology-aware evaluations, indicating improved generalization to less homologous antigens. Our case studies further suggest that the learned interaction fields provide a coherent connection between binding-interface recognition and affinity discrimination. Future work will explore the use of larger and more diverse Ag–Ab datasets, investigate potential extensions to additional biophysical objectives, and assess the applicability of CoBALT to practical antibody screening and optimization.

AI USE STATEMENT

In this work, we used generative AI tools for grammar checking, language polishing, and formatting assistance. We did not use generative AI tools for literature review, experimental design, result validation, data analysis, or drafting the initial manuscript. All AI-assisted revisions were carefully reviewed by the authors to ensure technical accuracy. We take full responsibility for the final content of this work, including any material produced with the assistance of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. The study focuses on computational modeling of antibody–antigen interactions and does not involve human participants, personally identifiable information, or new animal experiments. All datasets were used in accordance with their applicable licenses and usage guidelines. We have taken care to minimize potential bias and avoid discriminatory outcomes throughout the research process. We encourage responsible use in accordance with relevant biosafety and ethical guidelines.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the CoBALT architecture, training objectives and procedures, data preparation and splits, baselines, and evaluation metrics in the paper and supplementary materials. We also document the hyperparameter settings, model configurations, and computational resources required to replicate the main experiments. All code and trained models are available in an anonymous public repository (https://anonymous.4open.science/r/CoBALT-341B.).

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.

Rahmad Akbar, Philippe A Robert, Milena Pavlović, Jeliazko R Jeliazkov, Igor Snapkov, Andrei Slabodkin, Cédric R Weber, Lonneke Scheffer, Enkelejda Miho, Ingrid Hobæk Haff, et al. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Reports*, 34(11), 2021.

Lewis Chinery, Newton Wahome, Iain Moal, and Charlotte M Deane. Paragraph—antibody paratope prediction using graph neural networks with minimal feature vectors. *Bioinformatics*, 39(1): btac732, 2023.

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.

Michael Chungyoun and Jeffrey Gray. Fitness landscape for antibodies 2: Benchmarking reveals that protein ai models cannot yet consistently predict developability properties. *bioRxiv*, 2025.

James Dunbar and Charlotte M Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016.

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.

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.

Johnathan D Guest, Thom Vreven, Jing Zhou, Iain Moal, Jeliazko R Jeliazkov, Jeffrey J Gray, Zhiping Weng, and Brian G Pierce. An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. *Structure*, 29(6):606–621, 2021.

Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature biotechnology*, 42(2):275–283, 2024.

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.

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.

Susan Jones and Janet M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 1996.

Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Explainable deep relational networks for predicting compound–protein affinities and contacts. *Journal of chemical information and modeling*, 61(1):46–66, 2021.

Meredith M Keen, Alasdair D Keith, and Eric A Ortlund. Epitope mapping via in vitro deep mutational scanning methods and its applications. *Journal of Biological Chemistry*, 301(1), 2025.

Jisun Kim, Matthew McFee, Qiao Fang, Osama Abdin, and Philip M Kim. Computational and artificial intelligence-based methods for antibody development. *Trends in pharmacological sciences*, 44(3):175–189, 2023.

Hyunho Lee, Karla Castro, Sean Renwick, Lucas Stalder, Wiona Glänzer, Rachita Kumar, NingNing Chen, Andreas Scheck, Alex Yermanos, Derek Mason, et al. Contrastive learning for antibody-antigen sequence-to-specificity prediction. *bioRxiv*, pp. 2026–02, 2026.

Minghui Li, Yao Shi, Shengqing Hu, Shengshan Hu, Peijin Guo, Wei Wan, Leo Yu Zhang, Shirui Pan, Jizhou Li, Lichao Sun, et al. Mvsf-ab: accurate antibody–antigen binding affinity prediction via multi-view sequence feature learning. *Bioinformatics*, 41(5):btae579, 2025.

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

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

Chu'nan Liu, Lilian Denzler, Yihong Chen, Andrew Martin, and Brooks Paige. Asep: Benchmarking deep learning methods for antibody-specific epitope prediction. *Advances in Neural Information Processing Systems*, 37:11700–11734, 2024.

Chunan Liu, Aurelien Pelissier, Yanjun Shao, Lilian Denzler, Andrew CR Martin, Brooks Paige, and María Rodríguez Martínez. Abrank: A benchmark dataset and metric-learning framework for antibody-antigen affinity ranking. *arXiv preprint arXiv:2506.17857*, 2025.

Kevin Michalewicz, Mauricio Barahona, and Barbara Bravi. Antipasti: Interpretable prediction of antibody binding affinity exploiting normal modes and deep learning. *Structure*, 32(12):2422–2434, 2024.

Nathaniel L Miller, Thomas Clark, Rahul Raman, and Ram Sasisekharan. Learned features of antibody-antigen binding affinity. *Frontiers in Molecular Biosciences*, 10:1112738, 2023.

Hossam Nada, Yongseok Choi, Sungdo Kim, Kwon Su Jeong, Nicholas A Meanwell, and Kyeong Lee. New insights into protein–protein interaction modulators in drug discovery and therapeutic advance. *Signal transduction and targeted therapy*, 9(1):341, 2024.

Liang Pang, Jun Xu, Qingyao Ai, Yanyan Lan, Xueqi Cheng, and Jirong Wen. Setrank: Learning a permutation-invariant ranking model for information retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 499–508, 2020.

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

Jeffrey A Ruffolo, Jeffrey J Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint arXiv:2112.07782*, 2021.

Jeffrey A Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature communications*, 14(1):2389, 2023.

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.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, volume 2024, pp. 6987–7009, 2024.

Varun Ullanat, Bowen Jing, Samuel Sledzieski, and Bonnie Berger. Learning the language of protein-protein interactions. *Nature Communications*, 17(1):1199, 2026.

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.

Luiz Felipe Vecchietti, Bryan Nathanael Wijaya, Azamat Armanuly, Begench Hangeldiyev, Hyunkyu Jung, Sooyeon Lee, Meeyoung Cha, and Ho Min Kim. Artificial intelligence-driven computational methods for antibody design and optimization. In *MAbs*, volume 17, pp. 2528902. Taylor & Francis, 2025.

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.

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.

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.

Fan Xu, Zhi-an Huang, Haohuai He, Yidong Song, Wei Liu, Dongxu Zhang, Yao Hu, and Kay Chen Tan. Ablwr: A context-aware listwise ranking framework for antibody-antigen binding affinity prediction via positive-unlabeled learning. *arXiv preprint arXiv:2604.11272*, 2026.

Xincheng Zeng, Ganggang Bai, Chuance Sun, and Buyong Ma. Recent progress in antibody epitope prediction. *Antibodies*, 12(3):52, 2023a.

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, 2023b.

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.

Yuwei Zhou, Wenwen Liu, Ziru Huang, Yushu Gou, Siqi Liu, Lixu Jiang, Yue Yang, and Jian Huang. A comprehensive antigen-antibody complex database unlocking insights into interaction interface. *Elife*, 14:RP104934, 2025.

APPENDIX

A DETAILS FOR BSP AND BAR DATASETS

A.1 DATASET COLLECTION AND PRE-PROCESSING

Binding-site prediction. For the BSP task, we begin with 8,583 Ag–Ab pairs collected from ANABAG, which provides experimentally determined complex structures and residue-level annotation tables (Grandguillaume et al., 2025). Following the ANABAG annotations, interface residues are identified based on the reduction in solvent-accessible surface area (SASA) upon complex formation (Jones & Thornton, 1996; Grandguillaume et al., 2025). Specifically, a residue is labeled as an epitope residue if it lies on the antigen, or as a paratope residue if it lies on the antibody, when $\mathrm{SASA}_{\mathrm{free}} - \mathrm{SASA}_{\mathrm{bound}} > 0$. To identify paratope residues located within CDRs, we number the antibody sequences with ANARCI using the Chothia scheme and determine the corresponding CDR boundaries (Dunbar & Deane, 2016).

We then exclude complexes for which (i) the required structure or annotation files are unavailable, (ii) the antibody sequence cannot be numbered by ANARCI, or (iii) the antigen contains no labeled epitope residues or the antibody CDRs contain no labeled paratope residues. After filtering, 8,521 complexes remain. Across the resulting dataset, antigens have a median length of 364 residues and contain a median of 24 epitope residues, whereas the concatenated antibody CDRs have a median total length of 52 residues and contain a median of 26 paratope residues.

Binding-affinity ranking. For the BAR task, we curate Ag–Ab pairs from FLAB (Chungyoun & Gray, 2025), an antibody fitness compilation containing more than four million antibody sequences from 32 studies spanning seven therapeutic properties. We retain records associated with binding measurements, excluding studies whose assay types or measurement units cannot be reliably determined. We further remove duplicate Ag–Ab records and entries for which antigen or antibody structures cannot be successfully predicted. After filtering, the resulting BAR dataset contains 116,072 unique Ag–Ab pairs, comprising 58,209 antibodies and 4,657 antigens.

Affinity labels are standardized following the preprocessing protocol of FLAB. Measurements are converted to molar units where applicable and transformed as $r = -\log_{10}(K_D\,[\mathrm{M}])$, such that larger values of r indicate stronger binding.

A.2 DATA SPLITS

We evaluate CoBALT using two complementary data-splitting protocols. The random split assigns samples to the training, validation, and test sets without imposing any sequence-based constraints. In contrast, the antigen-homology-aware split holds out entire families of homologous antigens, providing a more stringent evaluation of generalization to dissimilar antigen targets.

Random Split. For BAR, the antibodies associated with each antigen are grouped into fixed-size ranking chunks of K. These chunks are partitioned into five folds with approximately equal numbers of Ag–Ab pairs. For BSP, Ag–Ab pairs sharing the same PDB identifier are kept in the same group and these groups are partitioned using the same rotating five-fold scheme.

Antigen-homology-aware splits. To mitigate homology leakage, we construct a unified catalog of antigen sequences from BAR and BSP, assigning a common sequence identifier to identical antigens shared between the two datasets. We independently cluster the combined catalog at sequence-identity thresholds of 75% and 80% using MMseqs2 (Steinegger & Söding, 2017). Consistent with previous antigen-clustering pipelines (Liu et al., 2024), we require each alignment to cover at least 80% of both sequences. Additionally, we require a minimum alignment length of 50 residues to exclude short local matches.

Clustering and data partitioning are performed jointly across BAR and BSP. All BAR and BSP pairs whose antigens belong to the same sequence cluster are assigned to the same partition. This constraint prevents homologous antigens from appearing in the training set of one task and the validation or test set of the other. The resulting groups for each sequence-identity threshold are

assigned in their entirety to the training, validation, and test sets, targeting approximate proportions of 70%, 10%, and 20%, respectively.

A.3 Graph construction

Structure Prediction. For BSP samples, residue-level graphs are constructed from experimentally determined Ag–Ab complex structures curated in ANABAG (Grandguillaume et al., 2025). By contrast, the BAR samples curated from FLAB (Chungyoun & Gray, 2025) are sequence-based and lack experimentally determined structures. We therefore predict the antigen and antibody structures independently using ESMFold (Lin et al., 2023) and IgFold (Ruffolo et al., 2023), respectively, and construct residue-level graphs from the predicted monomer structures. No structural docking or putative Ag–Ab complex construction is performed prior to interaction modeling.

Node features. Inspired by previous work (Ahmed et al., 2026; Wu et al., 2025), we represent each residue with a 105-dimensional feature vector comprising amino-acid identity, sequence positional encodings, backbone angular features, local distance encodings, and local-frame directional descriptors. We augment these features with residue-level PLM embeddings. Specifically, antigen residues are represented using 1,280-dimensional embeddings from the 650M-parameter ESM-2 model (Lin et al., 2023), whereas antibody residues are represented using 512-dimensional embeddings from the AntiBERTy encoder employed by IgFold (Ruffolo et al., 2021; 2023).

Edge relations. We adopt the four types of intramolecular relations defined in EpiFormer (Ahmed et al., 2026), comprising two sequence-based relations and two spatial relations. The sequence-based relations connect residue pairs separated by one or two positions along the same chain, respectively. Of the two spatial relations, one connects each residue to its 10 nearest neighbors based on C_α distances, whereas the other connects all residue pairs with an C_α distance below 8 Å (Wu et al., 2025). These relation types capture complementary aspects of sequential connectivity and spatial proximity and are not mutually exclusive. Thus, a residue pair may be associated with multiple relation types (Zhang et al., 2022). All relations are defined within each molecular partner separately.

Edge features. Following the edge-encoding scheme adopted in previous work (Ahmed et al., 2026; Jing et al., 2020), we represent each intramolecular edge with a 100-dimensional feature vector. This vector comprises a 4-dimensional relation-type encoding, a 16-dimensional sinusoidal encoding of the relative sequence separation, 64 radial basis function features encoding the inter-residue distance, 12 local-frame directional features, and a 4-dimensional relative orientation descriptor. Together, these features characterize the sequential and geometric relationships between connected residues while remaining independent of their absolute positions and orientations in the global coordinate frame.

B Implementation Details

B.1 Relation-Aware Message Passing

Relation-aware message passing (RMP) is applied separately to the antigen and antibody graphs to integrate residue representations, edge attributes, and geometric information. At interaction block ℓ, the RMP module for molecular partner $p \in \{\mathrm{Ag}, \mathrm{Ab}\}$ takes as input the node representations $\mathbf{H}^p_{\ell-1}$, residue coordinates $\mathbf{X}^p_{\ell-1}$, edge set $\mathcal{E}_p$, relation types $\mathcal{R}_p$, and edge features $\mathbf{F}_p$. For each relation $\xi \in \mathcal{R}_p$, let $\mathcal{E}_{p,\xi} \subseteq \mathcal{E}_p$ denote the subset of directed edges assigned to relation type ξ. Because a residue pair may be associated with multiple relations, these relation-specific subsets are not necessarily disjoint. For clarity, we omit the molecular-partner and layer indices below and use $\mathbf{h}_i$ and $\mathbf{x}_i$ to denote the input representation and coordinate of residue i, respectively. The coordinates are initialized from the C_α positions.

Relation-specific messages. Let $\mathcal{R} = \{0, 1, 2, 3\}$ denote the four intramolecular relation types defined in Appendix A.3. After omitting the molecular-partner index, $\mathcal{E}_\xi$ denotes the directed edge set associated with relation ξ. For an edge $(j, i) \in \mathcal{E}_\xi$ from source residue j to target residue i, we compute the relative coordinate vector and the corresponding Euclidean distance as

$$\boldsymbol{\delta}_{ji} = \mathbf{x}_j - \mathbf{x}_i, \qquad d_{ji} = \|\boldsymbol{\delta}_{ji}\|_2 . \tag{B.1}$$

We encode d_{ji} using 16 fixed Gaussian radial basis functions (RBFs) with centers uniformly spaced between 0 and 20 Å:

$$[\boldsymbol{\gamma}(d_{ji})]_k = \exp\left[-\frac{1}{2}\left(\frac{d_{ji}-\mu_k}{\sigma}\right)^2\right], \qquad \mu_k = \frac{20k}{15}\,\text{Å}, \quad k = 0,\ldots,15, \quad \sigma = 1\,\text{Å}. \tag{B.2}$$

Each relation type is associated with an independently parameterized message network. The message transmitted along edge (j, i) under relation ξ is computed as

$$\mathbf{m}_{ji}^{(\xi)} = \phi_m^{(\xi)}\left(\left[\mathbf{h}_j \,\|\, \mathbf{h}_i \,\|\, \mathbf{f}_{ji}^{(\xi)} \,\|\, \boldsymbol{\gamma}(d_{ji})\right]\right), \tag{B.3}$$

where $\mathbf{f}_{ji}^{(\xi)}$ is the feature vector associated with the typed edge (j, i, ξ), $\phi_m^{(\xi)}$ is the message network for relation ξ, and $\|$ denotes concatenation along the feature dimension. The use of relation-specific message networks allows different types of connectivity to be modeled independently.

Message aggregation and node update. For each target residue i, incoming messages are summed over its neighbors within each relation and subsequently across relation types:

$$\mathbf{a}_i = \sum_{\xi\in\mathcal{R}} \sum_{j\in\mathcal{N}_\xi^-(i)} \mathbf{m}_{ji}^{(\xi)}, \qquad \mathcal{N}_\xi^-(i) = \{j \mid (j,i) \in \mathcal{E}_\xi\}. \tag{B.4}$$

The aggregated message is then combined with the input residue representation through a node-update network shared across all relation types:

$$\Delta\mathbf{h}_i = \phi_h\left([\mathbf{h}_i \,\|\, \mathbf{a}_i]\right). \tag{B.5}$$

Stacking the feature increments $\Delta\mathbf{h}_i$ over all residues yields the intramolecular message $\mathbf{M}_{\ell,\text{intra}}^p$ used by interaction block ℓ.

Coordinate update. The same relation-specific edge messages are used to update the residue coordinates. For each relation ξ, a relation-specific network maps the edge message to a scalar weight:

$$w_{ji}^{(\xi)} = \phi_x^{(\xi)}\left(\mathbf{m}_{ji}^{(\xi)}\right). \tag{B.6}$$

The coordinate update is obtained by averaging the weighted relative coordinate vectors within each relation and then summing the resulting contributions across relations:

$$\mathbf{x}_i^+ = \mathbf{x}_i + \sum_{r\in\mathcal{R}} \frac{1}{\left|\mathcal{N}_r^-(i)\right| + \varepsilon} \sum_{j\in\mathcal{N}_r^-(i)} \boldsymbol{\delta}_{ji} w_{ji}^{(r)}, \qquad \varepsilon = 10^{-8}. \tag{B.7}$$

A relation with no incoming neighbors contributes zero to the coordinate update. Because the scalar weights are functions of rigid-motion-invariant features, while the coordinate increments are constructed from relative coordinate vectors, the update is invariant to global translations and equivariant to global rotations.

Both the node-feature and coordinate updates are computed using the coordinates available at the beginning of the current interaction block. The updated coordinates are not used to recompute messages within the same block. Instead, they are passed to the subsequent block, where they affect the distance encodings and relation-specific message computations.

B.2 Multi-Head Cross-Attention

Multi-head cross-attention (MHCA) enables residues of one molecular partner to aggregate information from the other. Let p denote the receiving partner and q its counterpart, where$(p, q) \in \{(\text{Ag}, \text{Ab}), (\text{Ab}, \text{Ag})\}$. At interaction block ℓ, the representations of the two partners are

$$\mathbf{H}_{\ell-1}^p \in \mathbb{R}^{L_p\times d}, \qquad \mathbf{H}_{\ell-1}^q \in \mathbb{R}^{L_q\times d}, \tag{B.8}$$

where L_p and L_q denote their respective numbers of residues and d is the hidden dimension. The batch dimension is omitted for clarity. In the $p \leftarrow q$ direction, residues of partner p provide the queries, whereas residues of partner q provide the keys and values.

Cross-attention computation. Let h denote the number of attention heads and let $d_h = d/h$ be the dimension of each head. For head $a \in \{1, \dots, h\}$, the query, key, and value matrices are computed as

$$\begin{aligned} \mathbf{Q}_{\ell,a}^{p\leftarrow q} &= \mathbf{H}_{\ell-1}^{p} \mathbf{W}_{Q,\ell,a}^{p\leftarrow q} + \mathbf{b}_{Q,\ell,a}^{p\leftarrow q}, \\ \mathbf{K}_{\ell,a}^{p\leftarrow q} &= \mathbf{H}_{\ell-1}^{q} \mathbf{W}_{K,\ell,a}^{p\leftarrow q} + \mathbf{b}_{K,\ell,a}^{p\leftarrow q}, \\ \mathbf{V}_{\ell,a}^{p\leftarrow q} &= \mathbf{H}_{\ell-1}^{q} \mathbf{W}_{V,\ell,a}^{p\leftarrow q} + \mathbf{b}_{V,\ell,a}^{p\leftarrow q}, \end{aligned} \tag{B.9}$$

where $\mathbf{Q}_{\ell,a}^{p\leftarrow q} \in \mathbb{R}^{L_p \times d_h}$, and $\mathbf{K}_{\ell,a}^{p\leftarrow q}, \mathbf{V}_{\ell,a}^{p\leftarrow q} \in \mathbb{R}^{L_q \times d_h}$. Each query residue is compared with all residues of the partner molecule using scaled dot-product attention:

$$\mathbf{S}_{\ell,a}^{p\leftarrow q} = \frac{\mathbf{Q}_{\ell,a}^{p\leftarrow q} \left(\mathbf{K}_{\ell,a}^{p\leftarrow q}\right)^{\top}}{\sqrt{d_h}}, \qquad \mathbf{A}_{\ell,a}^{p\leftarrow q} = \mathrm{softmax}_{\mathrm{row}}\left(\mathbf{S}_{\ell,a}^{p\leftarrow q}\right), \tag{B.10}$$

where $\mathbf{S}_{\ell,a}^{p\leftarrow q}, \mathbf{A}_{\ell,a}^{p\leftarrow q} \in \mathbb{R}^{L_p \times L_q}$. The row-wise softmax normalizes the attention weights over the L_q residues of partner q for each residue of partner p.

When padded batches are used, invalid partner residues are masked before the softmax operation and therefore receive zero attention weight. Dropout is subsequently applied to the attention weights, and the output of head a is

$$\mathbf{O}_{\ell,a}^{p\leftarrow q} = \mathrm{Dropout}\left(\mathbf{A}_{\ell,a}^{p\leftarrow q}\right) \mathbf{V}_{\ell,a}^{p\leftarrow q} \in \mathbb{R}^{L_p \times d_h}. \tag{B.11}$$

Multi-head output. The outputs of all attention heads are concatenated and transformed by an output projection:

$$\mathbf{M}_{\ell,\mathrm{inter}}^{p} = \left[\mathbf{O}_{\ell,1}^{p\leftarrow q} \parallel \cdots \parallel \mathbf{O}_{\ell,h}^{p\leftarrow q}\right] \mathbf{W}_{O,\ell}^{p\leftarrow q} + \mathbf{b}_{O,\ell}^{p\leftarrow q} \in \mathbb{R}^{L_p \times d}, \tag{B.12}$$

where $\parallel$ denotes concatenation along the feature dimension.

B.3 Baseline Models

BSP Baselines. The partner-independent BSP baselines predict residue-level binding propensity without explicitly modeling the corresponding molecular partner. To evaluate both sides of the interface, these methods are applied separately to the antigen and antibody. Partner-conditioned methods receive the antigen and antibody associated with the same sample. For all BSP methods, predictions on the antigen are compared with the epitope annotations, while predictions on the antibody CDRs are compared with the paratope annotations.

- **GraphBepi** (Zeng et al., 2023b) performs residue-level classification over protein graphs. We adapted it to BSP by training separate models for antigen epitope and antibody paratope prediction. Residues were represented using embeddings from the 3B-parameter ESM-2 model (`esm2_t36_3B_UR50D`), augmented with backbone dihedral-angle and secondary-structure features. The resulting graph representations were passed to an MLP to generate residue-level probabilities. Training used binary cross-entropy, with model selection based on validation AUPRC. The implementation followed the official GraphBepi codebase[1].
- **DiscoTope-3.0** (Høie et al., 2024) predicts B-cell epitopes by combining ESM-IF1 inverse-folding representations with sequence and structural descriptors. Residue-level propensity scores were obtained by averaging the outputs of the official ensemble of pretrained XGBoost classifiers. We used these classifiers directly, without additional fine-tuning on our BSP dataset. Code and pretrained models were obtained from the official repository[2].
- **EpiGraph** (Choi & Kim, 2024) employs graph attention to capture the spatial clustering of protein binding sites. Antigen residues and antibody CDR residues were organized into separate graphs, with edges drawn between residues whose Cα atoms were within 10 Å. Each node combined an ESM-IF1 structural embedding with an ESM-2 650M sequence

[1] `https://github.com/biomed-AI/GraphBepi`

[2] `https://github.com/Magnushhoie/DiscoTope-3.0`

embedding. Independent graph-attention predictors were used for the epitope and paratope tasks, and no cross-edges were introduced between the two molecular graphs. Both predictors were jointly optimized with residue-level binary cross-entropy. We implemented EpiGraph based on its official repository[3].

- **PECAN** (Pittala & Bailey-Kellogg, 2020) models residue-level binding sites through graph convolution and cross-protein attention. Following its original task formulation, we constructed two task-specific models: the epitope model treated the antigen as the primary protein and the antibody as its interaction partner, whereas the paratope model reversed these roles. Input features encoded amino-acid identity, BLOSUM62-based pseudo-PSSM information, and local amino-acid composition. After graph-convolutional encoding, the `attend_on_other` module computed pairwise interactions and aggregated representations from the partner protein. Experiments were implemented using the authors' released code[4].
- **WALLE** (Liu et al., 2024) formulates antibody-specific binding-site prediction as bipartite contact reconstruction. IgFold-derived antibody embeddings and ESM-2 650M antigen embeddings were encoded in two separate residue graphs constructed according to spatial proximity. Graph convolutional networks independently processed the two graphs, after which an inner-product decoder generated a dense Ag–Ab contact-probability matrix. Residue-level epitope and paratope scores were obtained by aggregating the predicted contacts. Optimization combined weighted contact-reconstruction loss with the edge-count regularization introduced in the original method. We followed the implementation distributed with the AsEP dataset[5].
- **EpiScan** (Wang et al., 2024a) jointly identifies antigen epitopes and antibody paratopes through a learned Ag–Ab interaction map. Antigen inputs included geometric, evolutionary, and local-composition features, while antibody CDR residues were represented by sequence-derived IgFold embeddings. Their pairwise interaction logits were subsequently refined by a convolution, followed by log-sum-exp pooling to obtain epitope and paratope predictions. Weighted binary cross-entropy was applied to both tasks during training. The model was implemented with reference to the official EpiScan repository[6].
- **EpiFormer** (Ahmed et al., 2026) provides an antibody-aware geometric architecture for epitope prediction. Its residue-level antigen and antibody graphs incorporated geometric features together with ESM-2 650M and AntiBERTy embeddings, respectively. EGNN-based message passing and bidirectional cross-attention were used to exchange information between the two graphs. A cross-attention decoder then constructed an Ag–Ab interaction matrix, from which residue-level epitope probabilities were pooled. We retained the official node-level training objective, including its classification, overlap, epitope-count, and auxiliary distance-prediction terms. The model was adapted from the authors' official implementation[7].

BAR Baselines. The BAR baselines assign a scalar binding score to each Ag–Ab pair. Candidate antibodies associated with the same antigen are subsequently ordered according to their predicted scores. Sequence-based baselines operate primarily on amino-acid representations, whereas structure-based baselines additionally incorporate structural information derived from the separately predicted antigen and antibody structures. Complex-based affinity predictors are excluded because they require experimentally determined or docked Ag–Ab complex structures. Such structures are unavailable for the BAR samples considered here, and generating approximate complex structures through molecular docking would introduce additional noise due to potential inaccuracies in the predicted binding poses while incurring substantial computational cost.

- **AntiBERTy** (Ruffolo et al., 2021) is an antibody-specific transformer pre-trained on 558M natural antibody sequences. Similar to our MINT setup, we utilized the `AntiBERTyRunner` library to extract embeddings for Ab-Ag pairs, which were then fed

[3] https://github.com/sj584/EpiGraph
[4] https://github.com/vamships/PECAN
[5] https://github.com/biochunan/AsEP-dataset
[6] https://github.com/gzBiomedical/EpiScan
[7] https://github.com/mansoor181/epiformer

into an MLP for binding affinity prediction optimized by MSE loss. The implementation was based on the official repository[8].

- **MVSF-AB**(Li et al., 2025) is a multi-view sequence fusion model designed for Ag–Ab binding affinity prediction. We used the released implementation to extract 768-dimensional ProteinBERT embeddings for antibody light-chain, heavy-chain, and antigen sequences. These embeddings were combined with physicochemical features and fed into the feature-fusion network for affinity prediction optimized by MSE loss. The implementation was based on the official repository[9].
- **MINT** (Ullanat et al., 2026) is PLM-based method designed to model protein interactions in a contextualized manner. We employed MINT as a feature extractor to obtain embeddings for Ab-Ag pairs, followed by a simple MLP to predict binding affinity. We used the pre-trained checkpoint *bernett_mlp.pth* available on Hugging Face[10] with the configuration file *esm2_t33_650M_UR50D.json*. The code was adapted from the official repository[11].
- **SaProt** (Su et al., 2024) is a structure-aware protein language model that represents proteins using a joint amino-acid and structural alphabet (AA+3Di). We used the 650M-parameter `SaProt_650M_AF2` checkpoint to extract mean-pooled final-layer embeddings independently for the antibody heavy chain, antibody light chain, and antigen. The three embeddings were concatenated and fed into a multilayer perceptron (MLP) for Ag–Ab binding-affinity prediction, with the model optimized using the MSE loss. The SaProt encoder was kept frozen during affinity-head training. Structure-aware AA+3Di sequences were generated from the corresponding antibody and antigen structures using Foldseek. Our implementation was based on the official SaProt repository[12].
- **AbRank** (Liu et al., 2025) is a metric-learning framework for Ab-Ag affinity ranking. To adapt it for our specific evaluation, we retrained the model on the datasets provided by the authors, strictly removing any test set leakage. We utilized the official implementation from GitHub[13] with the standard configuration file *train-abrank.yaml* for all experiments.

B.4 Evaluation Metrics

In this section, we introduce the details of the metrics used in our two tasks. For the BSP tasks, epitope and paratope predictions are evaluated on the following metrics, separately:

- **Area under the precision–recall curve (AUPRC):** summarizes the precision–recall trade-off across all decision thresholds and is particularly informative under class imbalance.
- **Area under the receiver operating characteristic curve (AUROC):** summarizes the trade-off between the true-positive and false-positive rates across all decision thresholds.
- **F1 score:** measures the harmonic mean of precision and recall after binarizing the predicted probabilities at a fixed threshold of 0.5.

For the BAR task, we evaluate model performance from two complementary perspectives: the accuracy of absolute affinity predictions and the relative ordering of antibody candidates. MAE evaluates absolute prediction accuracy, whereas Spearman's correlation, Kendall's correlation, P@1, Hit@3, and PairAcc assess ranking quality. The ranking metrics are computed within each antigen-specific ranking chunk and then averaged across the corresponding valid chunks.

- **Mean absolute error (MAE):** measures the average absolute difference between the ground-truth and predicted affinity scores. Lower values indicate better performance.
- **Spearman's rank correlation:** measures the monotonic agreement between the ground-truth and predicted rankings. It ranges from -1 to 1, where 1 indicates perfect rank agreement, -1 indicates exactly reversed rankings, and 0 indicates no monotonic association.

[8] https://github.com/jeffreyruffolo/AntiBERTy
[9] https://github.com/TAI-Medical-Lab/MVSF-AB
[10] https://huggingface.co/varunullanat2012/mint
[11] https://github.com/VarunUllanat/mint
[12] https://github.com/westlake-repl/SaProt
[13] https://github.com/biochunan/AbRank-WALLE-Affinity

- **Kendall's rank correlation:** measures pairwise concordance between the ground-truth and predicted rankings. A value of 1 implies a perfect match, while -1 implies a perfectly reversed order.
- **P@1:** indicates whether the top-ranked prediction is a ground-truth top binder. It is evaluated only on ranking sets with $K > 1$.
- **Hit@3:** indicates whether a ground-truth top binder is included among the three highest-ranked predictions. It is evaluated only on ranking sets with $K > 3$.
- **PairAcc:** measures the proportion of correctly ordered candidate pairs. Following prior work (Liu et al., 2025), only pairs whose ground-truth affinity difference exceeds 0.5 are considered. Ranking sets containing no eligible pairs are excluded.

B.5 Training Details

All experiments were implemented in PyTorch and trained on a workstation equipped with four NVIDIA GeForce RTX 3090 GPUs. We used the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. The per-GPU batch sizes were 1 for BAR affinity training and 2 for BSP binding-site training, corresponding to effective global batch sizes of 4 and 8, respectively. Models were trained for a maximum of 50 epochs, with validation performed after every epoch. The learning rate was reduced by a factor of 0.5 after five consecutive validation epochs without improvement in the joint validation score, and early stopping was applied with a patience of 15 epochs. Gradients were clipped to a maximum norm of 1.0, and the checkpoint with the best joint validation score was selected for final evaluation. A dropout rate of 0.1 and random seed 42 were used throughout. Appendix Table A1 records the detailed hyperparameter configurations for model training.

Table A1: **Detailed Hyperparameter Configuration.** Architectural and training hyperparameters used for CoBALT.

Category	Hyperparameter	Value
Model	RMP Layers	4
	MHCA Attention Heads	4
	Dropout	0.1
	Affinity Blocks	6
	Attention Heads	8
	Chunk Size (K)	20
	$\rho_\nu, \nu \in BSP, BAR$	0.02
	Contact Temperature τ_C	1.0
	Energy Bound B	20.0
Optimizer	Type	AdamW
	Learning Rate	1×10^{-4}
	Weight Decay	1×10^{-4}
	Gradient Clipping	1.0
	Random Seed	42
BSP Objective	λ_{BSP}	1.0
	λ_{BCE}^{epi}	2.0
	λ_{BCE}^{para}	0.5
	λ_{DICE}^{epi}	0.5
	λ_{DICE}^{para}	0.1
	λ_{reg}^{BSP}	1×10^{-3}
BAR Objective	λ_{BAR}	2.0
	λ_{MSE}	1.0
	$\lambda_{ListMLE}$	1.0
	$L_{ListMLE}$ Tie Noise	1×10^{-4}
	λ_{reg}^{BAR}	1×10^{-4}
Hardware	GPUs	4 × NVIDIA GeForce RTX 3090
	Distributed Launcher	`torchrun`

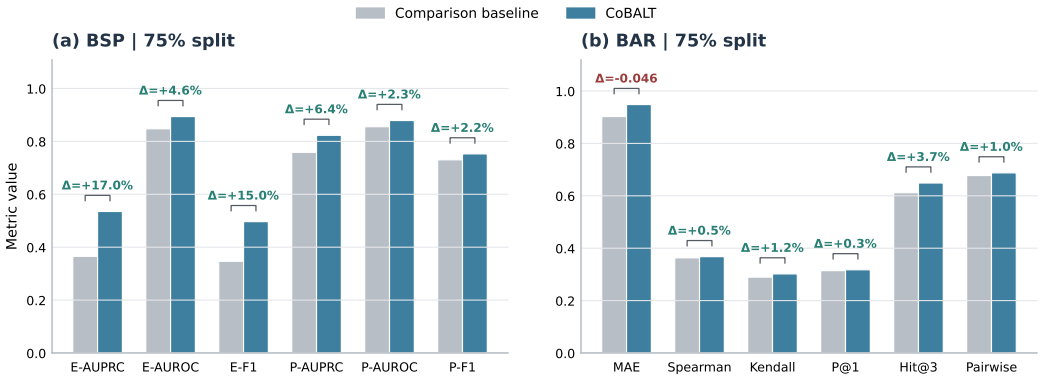


Figure A1: **Comparison with the strongest task-specific baselines on the 75% antigen-homology-aware split.** EpiFormer and EpiScan serve as the BSP baselines for epitope and paratope prediction, while AbRank serves as the BAR baseline. Gray and blue bars denote the baselines and CoBALT, respectively, with annotations indicating CoBALT's improvements.

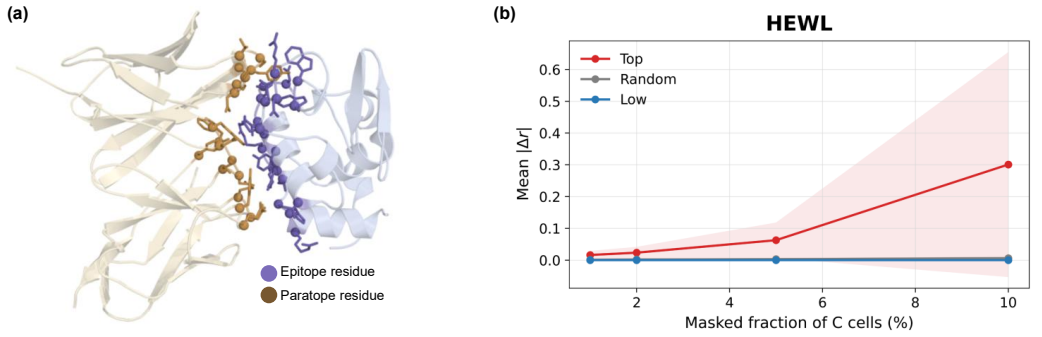


Figure A2: (a) Structural visualization of the annotated epitope and paratope in the 3a6b complex. (b) Contact-field perturbation analysis for HEWL, showing changes in predicted affinity induced by masking top-ranked, randomly selected, or low-ranked cells in the learned contact field C. Masking top-ranked cells produces larger prediction changes than masking random or low-ranked cells, indicating that the model's predictions are more sensitive to high-scoring entries of the learned contact field.

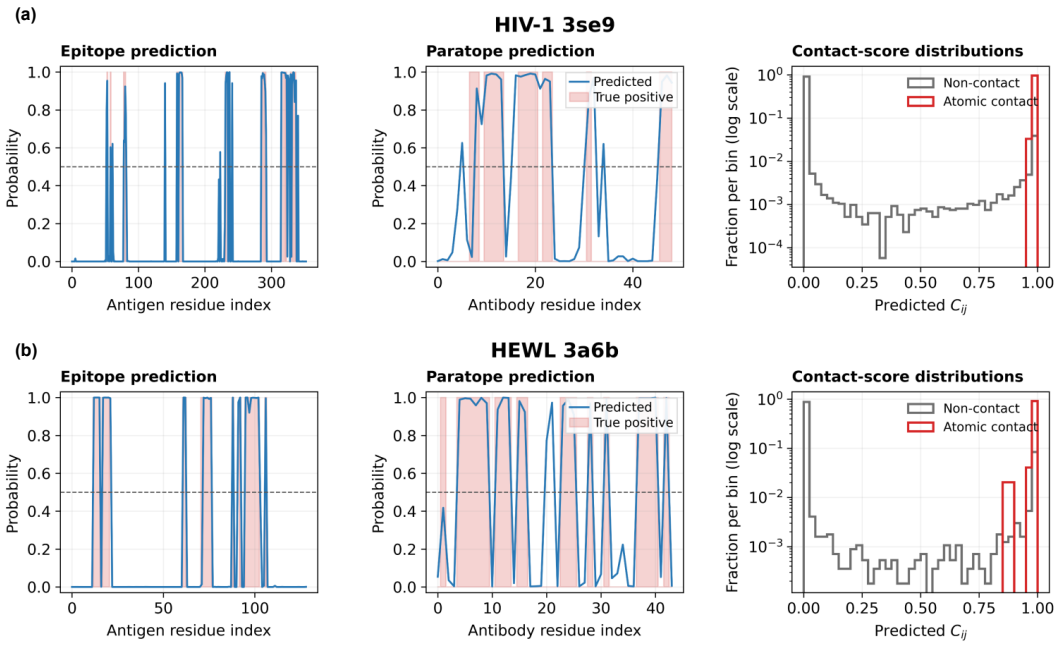


Figure A3: Residue-level binding-site predictions and contact-score distributions for representative (a) HIV-1 and (b) HEWL antigen groups. The first and second columns show the predicted epitope and paratope probabilities, respectively. Blue curves represent model predictions, shaded regions indicate experimentally annotated binding-site residues, and dashed horizontal lines denote the site-selection threshold. The third column compares the distributions of predicted contact scores for the ground truth contacting and non-contacting residue pairs. Overall, the predicted site probabilities are enriched at the annotated binding interfaces, and the scores effectively distinguish contacting from non-contacting residue pairs.

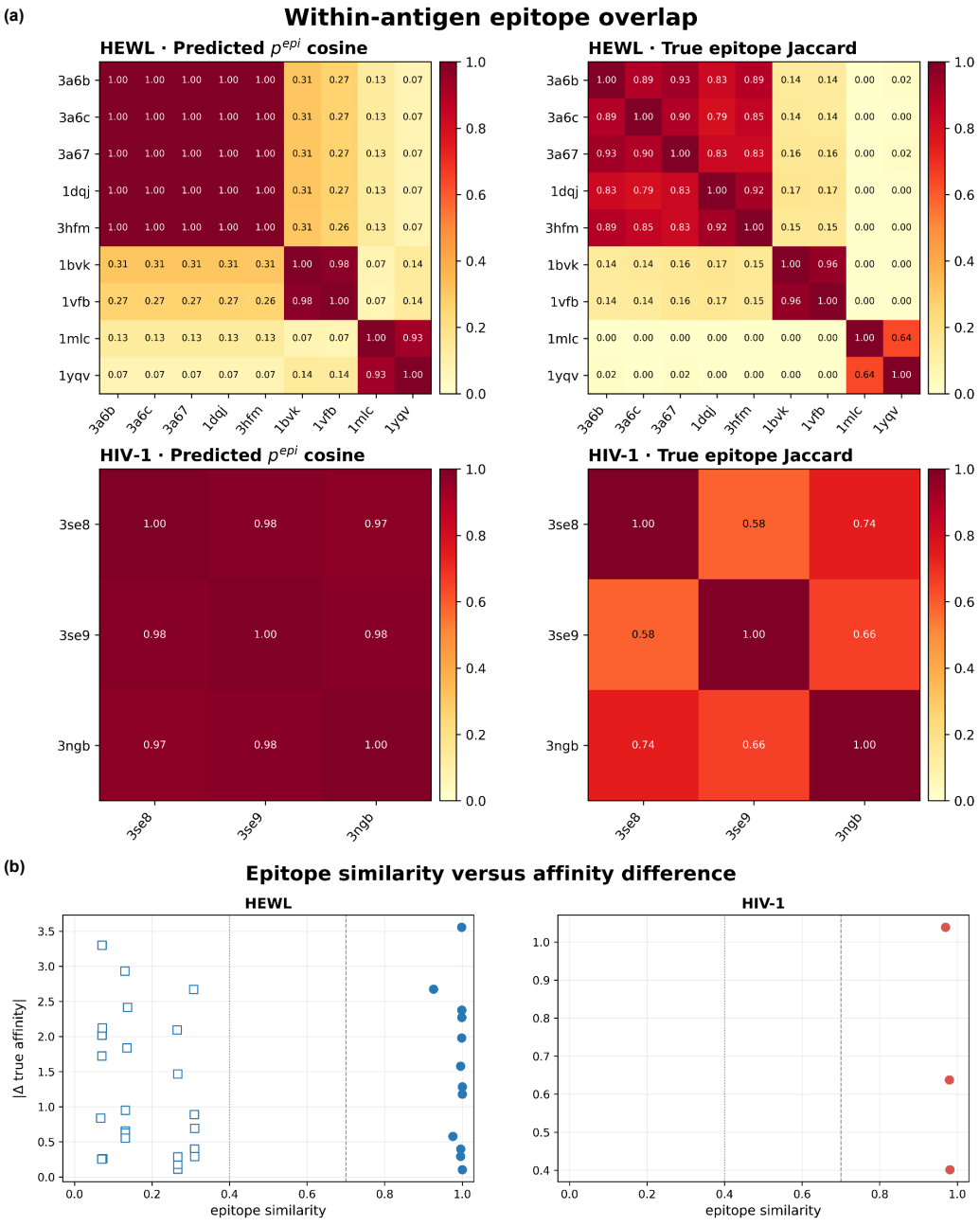


Figure A4: **Pairwise epitope-similarity analysis.** (a) Heatmaps show the cosine similarity between predicted epitope-probability profiles and the Jaccard similarity between experimentally annotated epitopes for each antigen group. (b) Scatter plots relate epitope similarity to the absolute difference in measured binding affinity. Filled circles and open squares represent antibody pairs with high and low epitope similarity, respectively, as defined by the vertical dashed and dotted threshold lines. HIV-1 antibody pairs exhibit consistently high epitope similarity, whereas HEWL antibody pairs span a broader range of similarities. Notably, high epitope similarity is not necessarily associated with similar binding affinities.

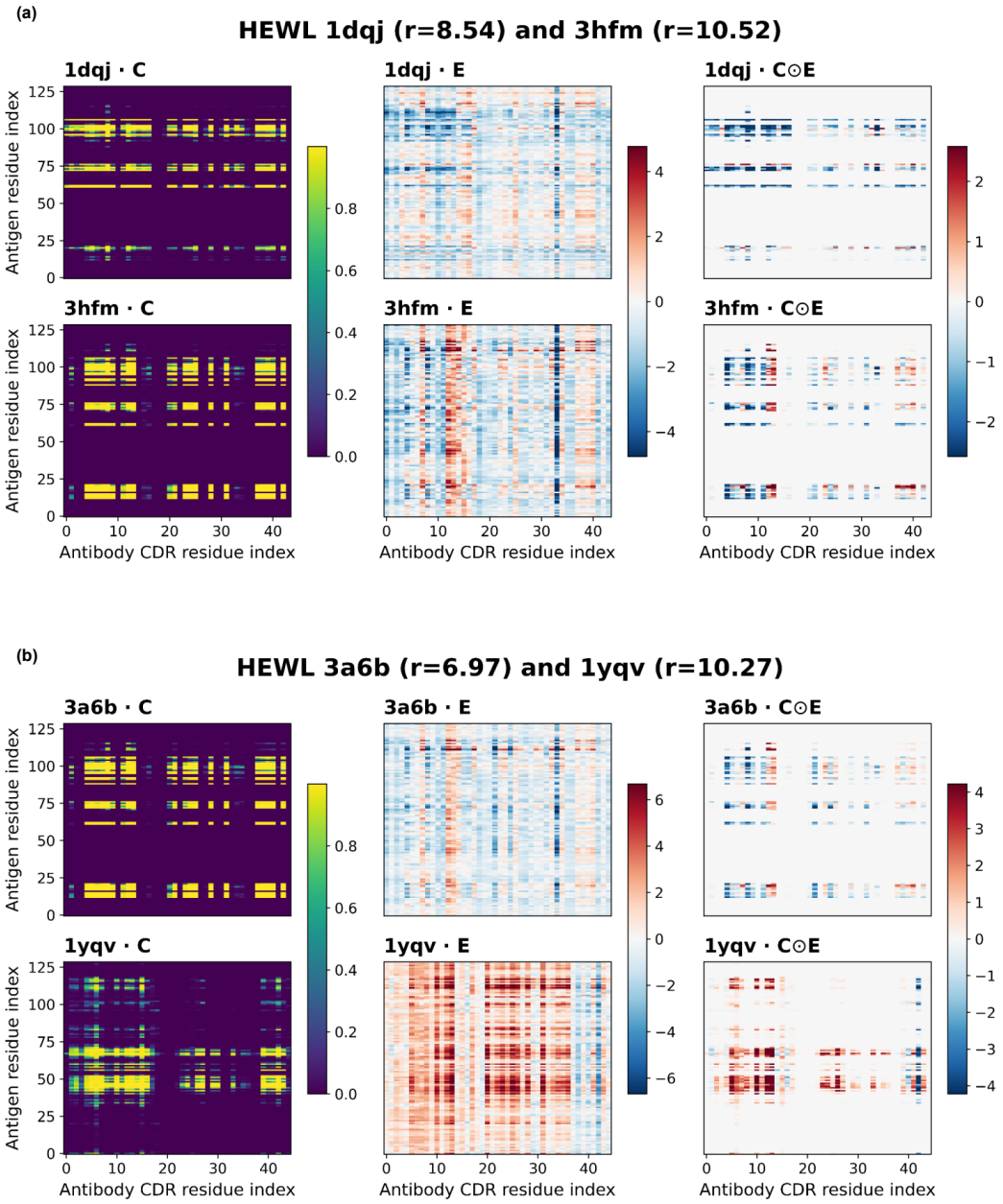


Figure A5: **Visualization of the learned interaction fields for two Ag–Ab pairs of HEWL antigen group.** (a) The upper block compares 1dqj and 3hfm, which bind to similar regions of HEWL, whereas (b) the lower block compares 3a6b and 1yqv, which exhibit substantially different binding interfaces. For each pair, the three columns show the contact field C, the energy-like field E, and the element-wise product $C \odot E$, respectively. Antibody pairs with similar binding sites display similar contact fields but distinct energy-like patterns, supporting the use of complementary contact and energetic representations.

C CONTACT FIELD EVALUATION

We evaluated whether the pairwise interaction fields learned without explicit contact-level supervision capture structurally plausible Ag–Ab contacts. To this end, we compared the predicted residue-pair scores with structure-derived contact labels on the BSP test set for CoBALT and its two single-task variants, CoBALT-BSP-Single and CoBALT-BAR-Single.

A residue pair was labeled as a contact based on the Euclidean distance between the corresponding C_α atoms. We considered two distance cutoffs: 4.5 Å as a stringent contact criterion and 6.0 Å as a more permissive criterion that includes a broader range of spatially proximal residue pairs. Because positive contacts constitute only a small fraction of all residue pairs, performance was evaluated using micro-averaged AUROC and AUPRC enrichment. AUPRC enrichment was calculated relative to the empirical prevalence of positive pairs, such that an enrichment of $1\times$ corresponds to random-ranking performance. The results are presented in Appendix Figures A6 and A7.

At the 4.5 Å cutoff, CoBALT achieved a micro-averaged AUROC of 95.25% and an AUPRC enrichment of approximately $38\times$. In comparison, CoBALT-BSP-Single achieved an AUROC of 68.49% and an enrichment of $1.8\times$, whereas CoBALT-BAR-Single achieved an AUROC of 43.39% and an enrichment of $0.83\times$ (Appendix Figure A6(b) and (c)). Similar results were observed at the 6.0 Å cutoff: CoBALT achieved an AUROC of 94.45% and an enrichment of approximately $39\times$, compared with 67.92% and $1.8\times$ for CoBALT-BSP-Single and 48.58% and $0.93\times$ for CoBALT-BAR-Single (Appendix Figure A7(b) and (c)). These results highlight the substantial benefit of joint training for learning structurally meaningful pairwise interaction fields. CoBALT identified structurally proximal residue pairs considerably more accurately than either single-task variant. CoBALT-BSP-Single captured a modest contact-related signal, whereas CoBALT-BAR-Single exhibited near-random pairwise discrimination.

For direct visual comparison, Appendix Figures A6(a) and A7(a) show the interaction fields predicted by the three models for the same representative complex, 4ydk, with structure-derived contacts indicated by red crosses. CoBALT produces a two-dimensionally structured interaction field that varies across both antigen and antibody residues, with most labeled contacts falling within or near high-scoring regions. This correspondence remains apparent when the contact cutoff is relaxed from 4.5 to 6.0 Å, with many of the additional contacts also located in regions assigned relatively high scores by CoBALT. In contrast, the two single-task variants produce predominantly vertical stripe-like patterns: their scores vary substantially across antibody residues but show limited specificity with respect to antigen residues. Consequently, they do not localize individual contacting residue pairs as clearly, although CoBALT-BSP-Single retains some contact-relevant signal, consistent with its above-random aggregate performance.

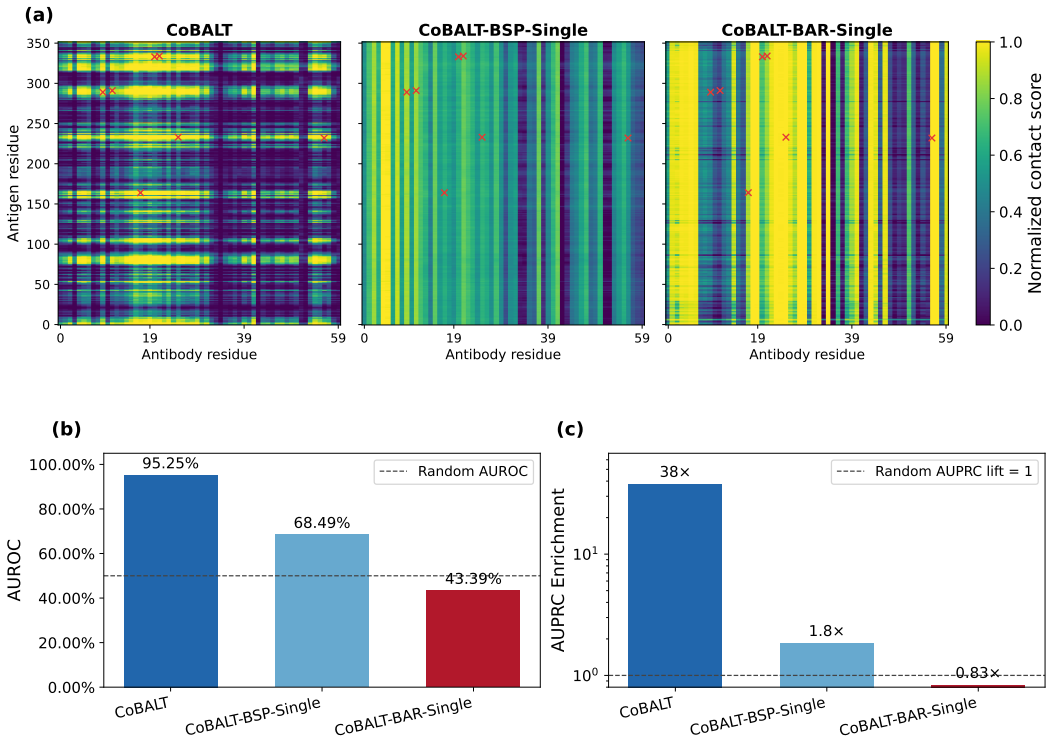


Figure A6: **Contact-field evaluation of CoBALT and its single-task variants under the distance cutoff of 4.5Å.** (a) Predicted residue-pair contact fields. Red crosses mark ground-truth contacts defined using a 4.5 Å C_α–C_α distance cutoff. (b) Micro-averaged AUROC on the BSP test set. The dashed line denotes random performance. (c) AUPRC enrichment relative to the empirical positive prevalence. The dashed line denotes random performance, corresponding to an enrichment of 1.

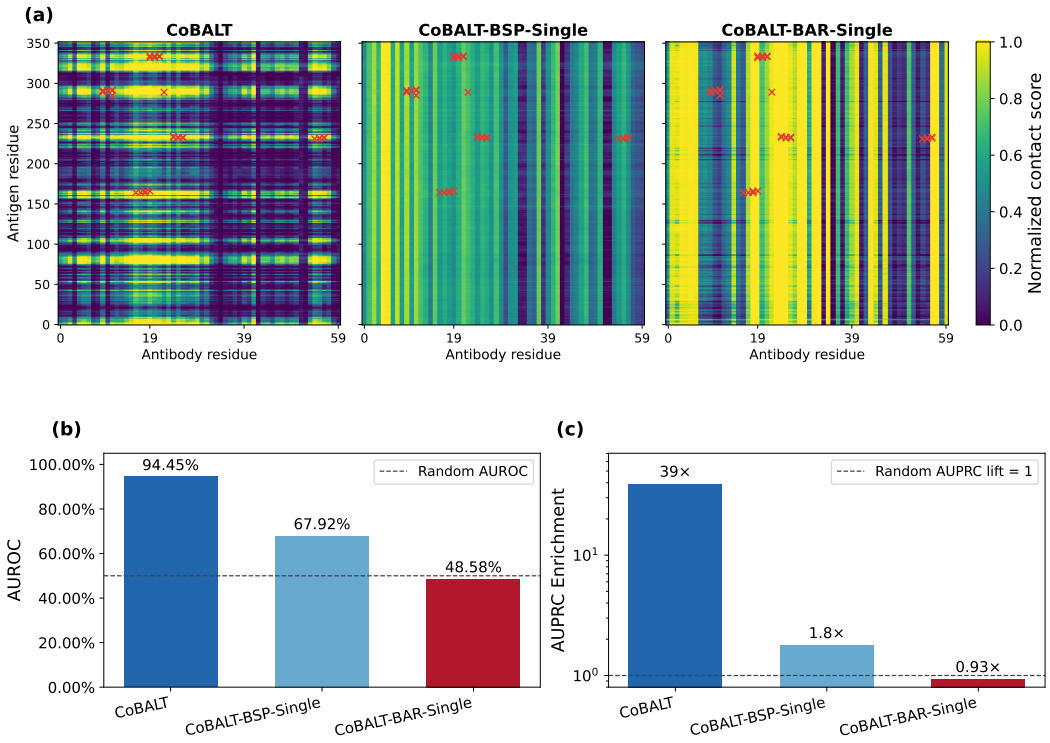


Figure A7: **Contact-field evaluation of CoBALT and its single-task variants under the distance cutoff of 6Å.** (a) Predicted residue-pair contact fields. Red crosses mark ground-truth contacts defined using a 6.0 Å C_α–C_α distance cutoff. (b) Micro-averaged AUROC on the BSP test set. The dashed line denotes random performance. (c) AUPRC enrichment relative to the empirical positive prevalence. The dashed line denotes random performance, corresponding to an enrichment of 1.