EvoStruct: Bridging Evolutionary and Structural Priors for Antibody CDR Design via Protein Language Model Adaptation

Anonymous Authors[1]

Abstract

Equivariant graph neural network (GNN) methods for antibody complementarity-determining region (CDR) design achieve the highest sequence recovery but suffer from severe vocabulary collapse. The current best GNN methods over-predict very few amino acids, such as tyrosine and glycine, while ignoring functionally important residues. We trace this failure to GNN encoders learning amino acid distributions *de novo* from limited structural data, discarding substitution patterns encoded in evolutionary databases. To resolve this, we propose EVOSTRUCT, which bridges a frozen protein language model (PLM) with 3D structural context from an E(3)-equivariant GNN via a cross-attention adapter. Unlike prior PLM-structure adapters for general protein design, EVOSTRUCT targets the vocabulary collapse problem specific to CDR design through progressive PLM unfreezing and R-Drop consistency regularization. On the CHIMERA-BENCH benchmark, EVOSTRUCT achieves the highest amino acid recovery and lowest perplexity among several antibody design methods, improving sequence recovery by 16% and reducing perplexity by 43% relative to the best GNN baselines, while recovering 2.3× greater amino acid diversity and the highest binding-pair correlation with ground truth.

1. Introduction

Antibodies bind antigens through their complementarity-determining regions (CDRs), six hypervariable loops whose sequence and structure determine binding specificity (Chothia & Lesk, 1987). Computational CDR design methods condition on antigen structure to generate sequences and backbone conformations for these loops (Luo et al., 2022; Kong et al., 2023a;b; Wu et al., 2025b). A growing body of evidence shows that existing methods largely fail to leverage antigen information. Predictions remain nearly unchanged when the antigen is removed (Li et al., 2025), and BLOSUM substitution matrices explain model outputs as well as learned likelihoods (Uçar & Sormanni, 2025; Chinery et al., 2024).

A systematic evaluation on the CHIMERA-BENCH benchmark (Ahmed et al., 2026) reveals that GNN methods with greedy decoding exhibit three linked failure modes. **Vocabulary collapse** reduces predicted amino acids to a fraction of the native diversity, with models overpredicting tyrosine and glycine while ignoring functionally important residues. This echoes the germline bias identified by Olsen et al. (2024) in antibody language models, but manifests more severely in structure-conditioned GNN methods. **Poor binding-pair learning** limits paratope-epitope amino acid pair correlation, indicating that no baseline captures the full binding-pair structure. **Contact position weakness** produces dramatically lower recovery at antigen-contacting positions compared to non-contacting positions, confirming that models learn general CDR statistics rather than antigen-specific binding preferences.

These failures share a common root. GNN encoders learn amino acid distributions *de novo* from limited structural training data (∼3,000 complexes), discarding the substitution patterns encoded in evolutionary sequence databases spanning hundreds of millions of proteins. Recent work on structure-informed PLM adaptation (Zheng et al., 2023) has shown that bridging frozen protein language models with structural encoders via cross-attention adapters preserves the PLM's vocabulary calibration for general protein design. We apply this paradigm to CDR design with EVOSTRUCT, which bridges a frozen ESM-2 (Lin et al., 2023) with an E(3)-equivariant GNN encoder through a cross-attention adapter that operates in ESM-2's representation space. Unlike general inverse folding adapters (Zheng et al., 2023; Shanehsazzadeh et al., 2023), EVOSTRUCT targets the vocabulary collapse problem specific to CDR design through progressive PLM unfreezing (Howard & Ruder, 2018) and R-Drop consistency regularization (Liang et al., 2021), and provides a systematic failure mode analysis quantifying the

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

diversity-accuracy tradeoff that constrains existing GNN methods.

Our contributions are:

1. A quantitative diagnosis of vocabulary collapse in CDR design using effective vocabulary (V_{eff}) and binding-pair correlation metrics, revealing that GNN methods recover only 20–35% of native amino acid diversity.
2. An adaptation of PLM-structure bridging for CDR design with progressive unfreezing and R-Drop regularization.
3. Empirical evidence that PLM vocabulary calibration transfers to CDR design, breaking the diversity-accuracy tradeoff with 16% higher sequence recovery, 43% lower perplexity, and 2.3× greater amino acid diversity than the best GNN baselines.

The rest of this paper is organized as follows. Section 2 discusses related work, and Section 3 defines the task, graph construction, and failure mode analysis. Section 4 describes the EVOSTRUCT architecture. Section 5 presents experimental results and failure mode resolution. Section 6 concludes.

2. Related Work

GNN-based CDR design. Equivariant GNN methods including MEAN (Kong et al., 2023a), dyMEAN (Kong et al., 2023b), and RAAD (Wu et al., 2025b) condition on antigen through spatial message passing and achieve the highest sequence recovery among all paradigms. However, these methods predict amino acids through simple linear or MLP heads operating in the GNN's learned embedding space, which suffers from severe vocabulary collapse under greedy decoding. Multiple studies document this and broader conditioning failures in CDR design methods: Li et al. (2025) show that predictions change minimally when the antigen is removed, Uçar & Sormanni (2025) demonstrate that BLOSUM matrices explain model outputs as well as learned likelihoods, and Chinery et al. (2024) find that simple computational methods can match deep learning approaches for generating diverse binder-enriched libraries. Olsen et al. (2024) identify a related germline bias in antibody language models where overprediction of germline residues limits design utility. We provide a systematic quantification of this vocabulary collapse in structure-conditioned CDR design and demonstrate that PLM adaptation is an effective solution.

Generative CDR design. Diffusion-based methods such as DiffAb (Luo et al., 2022), AbFlowNet (Abir et al., 2025), and AbMEGD (Chen et al., 2025) model CDR generation on SE(3), while AbODE (Verma et al., 2023) uses conjoined ODEs. RefineGNN (Jin et al., 2022b) generates CDRs autoregressively without antigen input. These sampling-based methods maintain higher vocabulary diversity but achieve substantially lower sequence recovery. FlowDesign (Wu et al., 2025a) addresses prior distribution quality for CDR design, targeting structural diversity rather than the sequence vocabulary collapse we identify. Among PLM-based approaches, AntiFold (Høie et al., 2025) fine-tunes ESM-IF1 for antibody inverse folding with CDR-weighted masking. AbEgDiffuser (Zhu et al., 2025) and RADAb (Wang et al., 2024) incorporate frozen ESM-2 features through additive injection or concatenation with GNN residue embeddings, but neither uses cross-attention adaptation or operates the sequence head in PLM embedding space.

Structure-informed PLM adaptation. LM-Design (Zheng et al., 2023) introduced the paradigm of inserting cross-attention adapters into frozen PLMs to inject structural context from GNN encoders for protein inverse folding. IgDesign (Shanehsazzadeh et al., 2023) applies a similar bottleneck adapter between IgMPNN and ESM2-3B specifically for antibody design, with in vitro experimental validation. Our work builds on this paradigm but differs in motivation and training strategy. Where LM-Design and IgDesign target general sequence recovery, we diagnose and address the vocabulary collapse problem specific to CDR design. We additionally employ progressive PLM unfreezing following the schedule of Howard & Ruder (2018) and R-Drop consistency regularization (Liang et al., 2021), which to our knowledge has not been applied in protein design. The shadow paratope loss from dyMEAN (Kong et al., 2023b) provides pairwise distance supervision for CDR-epitope geometry.

3. Preliminaries

3.1. Task Definition

Following CHIMERA-BENCH (Ahmed et al., 2026), we formulate the CDR design task as follows. Given an antigen structure $A = \{(s_j, \mathbf{x}_j) \mid j \in V_A\}$, an epitope specification $E \subseteq V_A$, and an antibody framework $F = \{(s_i, \mathbf{x}_i) \mid i \in V_{\text{FR}}\}$, we design CDR residues $R = \{(s_k, \mathbf{x}_k) \mid k \in V_{\text{CDR}}\}$ that maximize the conditional likelihood subject to epitope contact constraints:

$$R^* = \arg\max_R \; p_\theta\big(R \mid A, E, F\big), \quad \text{s.t. } \mathcal{C}(R, A) \neq \emptyset \quad (1)$$

where each residue has amino acid type $s_k \in \{1, \ldots, 20\}$ and Cα coordinate $\mathbf{x}_k \in \mathbb{R}^3$. The contact set $\mathcal{C}(R, A) = \{j \in V_A \mid \exists\, k \in V_{\text{CDR}} : \|\mathbf{x}_k - \mathbf{x}_j\| < d_c\}$ denotes antigen residues contacted within cutoff d_c. We focus on CDR-H3, the most variable loop and primary determinant of antigen specificity (Chothia & Lesk, 1987).

3.2. Graph Construction

We represent the antibody-antigen complex as a heterogeneous graph $\mathcal{G} = (V, \mathcal{E})$. The node set $V = V_{\text{HC}} \cup V_{\text{LC}} \cup V_A \cup V_{\text{glob}}$ contains residue nodes from the heavy chain, light chain, and antigen, plus three global delimiter tokens ($V_{\text{glob}} = \{\text{BOH}, \text{BOL}, \text{BOA}\}$). Each residue node i carries amino acid type $s_i \in \{1, \ldots, 20\}$ and four backbone atom coordinates $\mathbf{X}_i = [\mathbf{x}_i^{\text{N}}, \mathbf{x}_i^{\text{C}\alpha}, \mathbf{x}_i^{\text{C}}, \mathbf{x}_i^{\text{O}}] \in \mathbb{R}^{4\times 3}$.

We partition the edge set $\mathcal{E}$ into 8 typed subsets that capture different structural relationships. Within each chain, we construct *radial edges* connecting all pairs within a Cα distance threshold, *sequential edges* linking residues separated by one or two positions in primary sequence, and *KNN edges* connecting each residue to its K nearest spatial neighbors. Across chains, we add *inter-chain radial edges* and *inter-chain KNN edges* that enable direct communication between antibody and antigen residues. Global delimiter tokens connect to all residues in their respective chains via *global-to-chain edges*. Each edge (i, j) of type t carries a feature vector $\mathbf{e}_{ij} \in \mathbb{R}^{d_e}$ encoding edge type, relative position in local coordinate frames, pairwise distance RBFs across four backbone atom pairs, a quaternion encoding of relative backbone orientation, and local frame direction features.

3.3. Failure Mode Analysis

We evaluate several CDR-H3 design methods on the CHIMERA-BENCH benchmark and identify three systematic failure modes that motivate our approach.

Vocabulary Collapse. We define the effective vocabulary as $V_{\text{eff}} = \exp(-\sum_a p(a) \log p(a))$, the exponentiated Shannon entropy of the predicted amino acid distribution. Native CDR-H3 sequences exhibit $V_{\text{eff}} \approx 15.5$, indicating near-uniform usage of all 20 amino acids. GNN methods with greedy decoding collapse to V_{eff} of 3.0–5.5. RAAD (Wu et al., 2025b) overpredicts tyrosine by 34% and glycine by 21%, while rare but biochemically critical residues (tryptophan, cysteine, methionine) appear at near-zero frequency (Figure 3). Sampling-based methods approach native diversity (V_{eff} 11.7–14.9) but at the cost of much lower accuracy.

Binding-Pair Failure. Binding specificity requires that the model learn which paratope amino acids pair with which epitope amino acids. We compute the correlation between predicted and ground-truth paratope-epitope amino acid pair frequencies. No baseline exceeds $r = 0.69$ (Figure 4), and DiffAb (Luo et al., 2022) shows anti-correlation ($r = -0.29$), systematically predicting the wrong amino acids at interface positions. This indicates that existing conditioning mechanisms route negligible antigen information to the sequence prediction head.

Contact Position Weakness. All methods show dramatically lower amino acid recovery at antigen-contacting positions (8–23%) compared to non-contacting positions (23–51%). Contact positions are precisely where amino acid identity is most constrained by the binding partner, yet they are where models struggle to perform well. The gap reveals that models learn general CDR positional statistics rather than antigen-specific binding physics, and that the antigen input contributes little to prediction quality even at the positions where it matters most.

4. Method

EVOSTRUCT processes the antibody-antigen complex through two parallel pathways and fuses their outputs via a cross-attention adapter (Figure 1), following the PLM-structure bridging paradigm of Zheng et al. (2023). A relation-aware E(3)-equivariant GNN encodes the full 3D complex graph, capturing spatial geometry. Separately, a frozen protein language model produces CDR embeddings that encode evolutionary substitution patterns learned from hundreds of millions of protein sequences. A structural adapter then cross-attends the PLM's CDR embeddings to the GNN's structural context, and a sequence head operating in the PLM's representation space produces the final amino acid predictions. The entire sequence prediction pathway runs in the PLM's embedding space, preserving its calibrated amino acid vocabulary, while the GNN supplies structural context *into* PLM representations.

4.1. PLM Backbone

The sequence pathway leverages ESM-2 (Lin et al., 2023), a protein language model trained on over 250 million protein sequences via masked language modeling. ESM-2 encodes amino acid substitution patterns, evolutionary conservation, and biochemical properties in its high-dimensional embeddings. We extract CDR representations by passing the full heavy-chain sequence through ESM-2 with CDR positions replaced by the mask token, then selecting the embeddings at CDR positions:

$$\mathbf{H}_{\text{esm}} = \text{ESM-2}(\text{mask}(s_{\text{HC}}, V_{\text{CDR}}))\big|_{V_{\text{CDR}}} \in \mathbb{R}^{L\times d_{\text{esm}}} \quad (2)$$

where $L = |V_{\text{CDR}}|$ is the CDR length and d_{esm} is the PLM embedding dimension. Masking CDR positions prevents trivial copying of the ground-truth sequence during training. The ESM-2 model remains frozen during early training and is progressively unfrozen in later phases following Howard & Ruder (2018), preventing catastrophic forgetting of evolutionary priors.

Operating in ESM-2's embedding space rather than the GNN's preserves the PLM's vocabulary calibration. ESM-

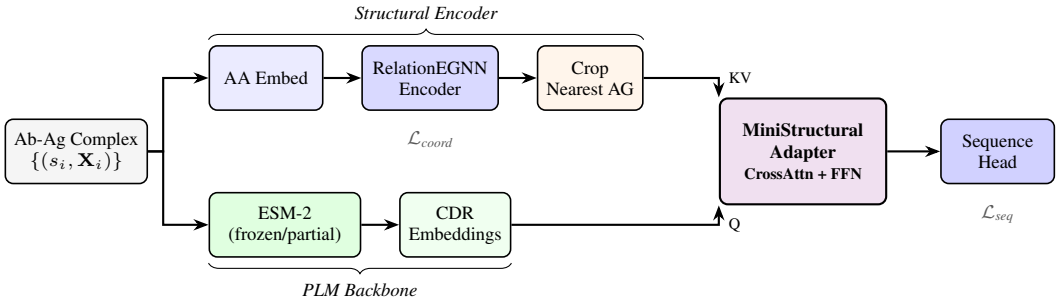

Figure 1. **EvoStruct architecture.** Two parallel paths process the antibody-antigen complex. A Relational-EGNN encoder produces structural embeddings over the full complex graph, while frozen ESM-2 produces evolutionary embeddings for CDR positions. The Mini Structural Adapter cross-attends ESM-2 CDR queries to GNN-encoded structural keys/values (CDR + nearest antigen residues), whereas the Sequence Head produces the final CDR sequence designs.

2's output distribution, learned from hundreds of millions of sequences, naturally assigns non-zero probability to all 20 amino acids at each position, with probabilities reflecting evolutionary substitution patterns. GNN methods that learn their own amino acid embeddings from a few thousand training complexes lack this calibration and collapse to a handful of frequent amino acids under greedy decoding (Section 3.3). This is the same principle underlying LM-Design (Zheng et al., 2023) for general protein design, which we apply here to resolve the vocabulary collapse specific to CDR design.

4.2. Structural Encoder

The structural pathway encodes the full antibody-antigen complex through multiple layers of a relation-aware E(3)-equivariant graph neural network (Satorras et al., 2021; Wu et al., 2025b) operating on graph $\mathcal{G}$ (Section 3.2). Each residue is initialized with a learned amino acid embedding, projected to the GNN hidden dimension via a linear layer. CDR residue embeddings are masked to zero during training. Epitope residues receive an additional learnable embedding that signals their membership in the designated binding site.

Each GNN layer updates node features and coordinates simultaneously. For edge (i, j) of type t, the message function concatenates the sender and receiver embeddings, the outer product of coordinate displacements, and the edge features:

$$\mathbf{m}_{ij}^{(l)} = \text{MLP}_{\text{msg}}^{(l)}\Big([\mathbf{h}_i^{(l)},\ \mathbf{h}_j^{(l)},\ \text{vec}\big(\Delta\mathbf{x}_{ij}(\Delta\mathbf{x}_{ij})^\top\big),\ \mathbf{e}_{ij}]\Big) \tag{3}$$

where $\Delta\mathbf{x}_{ij} = \mathbf{x}_i^{(l)} - \mathbf{x}_j^{(l)}$ and $\text{vec}(\cdot)$ flattens the 3×3 outer product into a 9-dimensional vector. The outer product entries are dot products of displacement components, ensuring E(3)-invariance of the message function. Node features aggregate messages from all edge types with type-specific linear projections and a skip connection for training stability:

$$\mathbf{h}_i^{(l+1)} = \mathbf{h}_i^{(l)} + \text{MLP}_{\text{node}}^{(l)}\left(\left[\mathbf{h}_i^{(l)},\ \sum_t \mathbf{W}_t^{(l)} \sum_{j \in \mathcal{N}_t(i)} \mathbf{m}_{ij}^{(l)}\right]\right) \tag{4}$$

where $\mathbf{W}_t^{(l)}$ is a type-specific projection matrix and $\mathcal{N}_t(i)$ denotes the neighbors of node i under edge type t. Coordinates are updated equivariantly by adding a weighted sum of displacement vectors:

$$\mathbf{x}_i^{(l+1)} = \mathbf{x}_i^{(l)} + \sum_t \frac{1}{|\mathcal{N}_t(i)|} \sum_{j \in \mathcal{N}_t(i)} \Delta\mathbf{x}_{ij} \cdot \text{MLP}_t^{\text{coord},(l)}(\mathbf{m}_{ij}^{(l)}) \tag{5}$$

The product $\Delta\mathbf{x}_{ij} \cdot \text{scalar}$ is equivariant by construction, because any rotation $\mathbf{R}$ transforms the displacement as $\mathbf{R}\Delta\mathbf{x}_{ij}$ while leaving the scalar factor unchanged. After L_{enc} layers, the encoder produces per-residue embeddings $\mathbf{h} \in \mathbb{R}^{N \times d_{\text{gnn}}}$ and updated backbone coordinates $\hat{\mathbf{X}} \in \mathbb{R}^{N \times 4 \times 3}$.

Antigen context cropping. Rather than providing the full antigen to the adapter, we crop to the K_{ag} antigen residues nearest to the CDR by Cα distance. This bounds memory usage while preserving the most relevant epitope information, since binding contacts necessarily involve spatially proximal residues. The structural context passed to the adapter is the concatenation of CDR and cropped antigen embeddings:

$$\mathbf{H}_{\text{ctx}} = [\mathbf{h}_k \mid k \in V_{\text{CDR}}] \,\|\, [\mathbf{h}_j \mid j \in \text{top}_{K_{\text{ag}}}(V_A)] \in \mathbb{R}^{(L+K_{\text{ag}}) \times d_{\text{gnn}}} \tag{6}$$

4.3. Mini Structural Adapter

The adapter bridges the two representation spaces by projecting ESM-2's CDR embeddings into a shared intermediate space, cross-attending to GNN-encoded structural context, and projecting back to the PLM's embedding dimension. This adapter contains a single cross-attention block followed by a feed-forward network (FFN), with residual connections and layer normalization at each stage.

We first project both representations into a shared adapter dimension d_a:

$$\mathbf{Q} = \mathbf{H}_{\text{esm}}\,\mathbf{W}_{\text{down}} \in \mathbb{R}^{L\times d_a} \tag{7}$$

$$\mathbf{K},\mathbf{V} = \mathbf{H}_{\text{ctx}}\,\mathbf{W}_{\text{gnn}} \in \mathbb{R}^{(L+K_{\text{ag}})\times d_a} \tag{8}$$

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d_{\text{esm}}\times d_a}$ projects ESM-2 embeddings down and $\mathbf{W}_{\text{gnn}} \in \mathbb{R}^{d_{\text{gnn}}\times d_a}$ projects GNN embeddings up. The queries are derived from ESM-2 (evolutionary prior), while keys and values are derived from the GNN (structural context). This asymmetry encodes the design principle that the PLM representation queries "what structural environment am I in?" and the GNN provides the answer.

The multi-head cross-attention with H heads computes attention scores and a weighted aggregation over the structural context:

$$\text{CrossAttn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \Big\|_{h=1}^{H} \text{softmax}\left(\frac{\mathbf{Q}^{(h)}(\mathbf{K}^{(h)})^\top}{\sqrt{d_a/H}}\right)\mathbf{V}^{(h)} \tag{9}$$

We then apply a residual connection and layer normalization after the cross-attention:

$$\mathbf{Z} = \text{LayerNorm}(\mathbf{Q} + \text{CrossAttn}(\mathbf{Q},\mathbf{K},\mathbf{V})) \tag{10}$$

A two-layer FFN with SiLU activation, also with residual connection and layer normalization, further processes the representations:

$$\mathbf{Z}' = \text{LayerNorm}(\mathbf{Z} + \text{FFN}(\mathbf{Z})) \tag{11}$$

Finally, the adapter projects back to the PLM's embedding dimension:

$$\mathbf{H}_{\text{refined}} = \mathbf{Z}'\,\mathbf{W}_{\text{up}} \in \mathbb{R}^{L\times d_{\text{esm}}} \tag{12}$$

The refined representations inhabit ESM-2's embedding space but are now informed by the 3D structural context of the antibody-antigen complex, including the relative geometry of CDR and antigen residues and the edge-type-specific interactions encoded by the GNN.

4.4. Sequence Head and Coordinate Prediction

The sequence head maps the refined PLM-dimensional representations to amino acid logits through a two-layer MLP with layer normalization, SiLU activation, and dropout:

$$\boldsymbol{\ell}_i = \text{MLP}_{\text{seq}}(\text{LayerNorm}(\mathbf{h}_i^{\text{refined}})) \in \mathbb{R}^{|\mathcal{A}|} \tag{13}$$

where $|\mathcal{A}|$ is the amino acid vocabulary size. At inference, we predict the amino acid at each position via greedy decoding, $\hat{s}_i = \arg\max_a \ell_i^a$.

The GNN encoder simultaneously produces updated Cα coordinates $\hat{\mathbf{x}}_k$ for CDR positions through the equivariant coordinate update (Equation (5)). These predicted coordinates provide the structural component of the co-design output.

4.5. Training Objective

We train EvoStruct with five loss terms combined with an R-Drop consistency regularizer (Liang et al., 2021).

Sequence loss. We minimize the per-position cross-entropy on predicted CDR amino acids:

$$\mathcal{L}_{\text{seq}} = -\frac{1}{L}\sum_{i=1}^{L}\log\frac{\exp(\ell_i^{y_i})}{\sum_a \exp(\ell_i^a)} \tag{14}$$

where y_i is the ground-truth amino acid at CDR position i.

Coordinate loss. We apply a smooth-ℓ_1 (Huber) loss on the GNN's predicted versus true Cα coordinates for CDR positions:

$$\mathcal{L}_{\text{coord}} = \frac{1}{L}\sum_{k\in V_{\text{CDR}}}\text{smooth}_{\ell_1}\left(\hat{\mathbf{x}}_k^{\text{C}\alpha} - \mathbf{x}_k^{\text{C}\alpha,\text{true}}\right) \tag{15}$$

Pairing loss. We employ an InfoNCE-style contrastive loss (Chen et al., 2020) that matches mean-pooled CDR and antigen GNN embeddings within the batch, encouraging the encoder to produce representations that distinguish cognate from non-cognate antibody-antigen pairs:

$$\mathcal{L}_{\text{pair}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\bar{\mathbf{h}}_i^{\text{cdr}}\cdot\bar{\mathbf{h}}_i^{\text{ag}}/\tau_p)}{\sum_{k=1}^{B}\exp(\bar{\mathbf{h}}_i^{\text{cdr}}\cdot\bar{\mathbf{h}}_k^{\text{ag}}/\tau_p)} \tag{16}$$

where B is the batch size and τ_p is a temperature parameter.

Docking loss. We penalize the minimum predicted Cα distance from each CDR residue to epitope atoms when it exceeds a cutoff, encouraging the predicted CDR backbone to dock near the epitope surface:

$$\mathcal{L}_{\text{dock}} = \frac{1}{L}\sum_{k\in V_{\text{CDR}}}\max\left(0,\ \min_{j\in E}\|\hat{\mathbf{x}}_k - \mathbf{x}_j\| - d_{\text{dock}}\right) \tag{17}$$

Shadow paratope loss. Following dyMEAN (Kong et al., 2023b), we penalize deviation of the predicted CDR-epitope pairwise distance matrix from ground truth. This provides a richer geometric supervisory signal than coordinate loss alone, because it supervises the relative arrangement of CDR and epitope residues rather than absolute positions:

$$\mathcal{L}_{\text{shadow}} = \frac{1}{L|E|}\sum_{k\in V_{\text{CDR}}}\sum_{j\in E}\left|\|\hat{\mathbf{x}}_k - \mathbf{x}_j\| - \|\mathbf{x}_k^{\text{true}} - \mathbf{x}_j\|\right| \tag{18}$$

R-Drop regularization. Each training step performs two forward passes with different dropout masks, producing two sequence losses $\mathcal{L}_{\text{seq}}^{(1)}$ and $\mathcal{L}_{\text{seq}}^{(2)}$. Inspired by R-Drop (Liang

et al., 2021), we penalize inconsistency between the two passes. We use a squared loss difference rather than the symmetric KL divergence of the original formulation, as this avoids computing per-position KL terms while still encouraging consistent predictions:

$$\mathcal{L}_{\text{R-Drop}} = \alpha_{\text{rd}} \cdot \left(\mathcal{L}_{\text{seq}}^{(1)} - \mathcal{L}_{\text{seq}}^{(2)}\right)^2 \tag{19}$$

The consistency penalty encourages robust structural context injection regardless of which cross-attention connections are dropped. The total training loss averages the two base losses and adds the consistency penalty:

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{\text{base}}^{(1)} + \mathcal{L}_{\text{base}}^{(2)}\right) + \mathcal{L}_{\text{R-Drop}} \tag{20}$$

where $\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{seq}} + \lambda_{\text{coord}}\mathcal{L}_{\text{coord}} + \lambda_{\text{pair}}\mathcal{L}_{\text{pair}} + \lambda_{\text{dock}}\mathcal{L}_{\text{dock}} + \lambda_{\text{shadow}}\mathcal{L}_{\text{shadow}}$.

5. Experiments

5.1. Experimental Setup

Dataset and splits. We evaluate on the CHIMERA-BENCH benchmark (Ahmed et al., 2026), which contains 2,922 antibody-antigen complexes from the Structural Antibody Database (SAbDab) after deduplication at 95% sequence identity. We use the epitope-group split, which clusters complexes by epitope similarity to test generalization to unseen binding sites.

Baselines. We compare against eleven CDR-H3 design methods spanning five architectural families. (1) *GNN-based:* RAAD (Wu et al., 2025b), MEAN (Kong et al., 2023a), dyMEAN (Kong et al., 2023b). (2) *Diffusion:* DiffAb (Luo et al., 2022), AbFlowNet (Abir et al., 2025), AbMEGD (Chen et al., 2025). (3) *Retrieval-augmented:* RADAb (Wang et al., 2024). (4) *Flow/ODE:* dyAb (Tan et al., 2025), AbODE (Verma et al., 2023). (5) *Autoregressive:* RefineGNN (Jin et al., 2022b), AbDockGen (Jin et al., 2022a). All baselines are retrained on CHIMERA-BENCH using the authors' released code with default hyperparameters.

Metrics. We retrained the baseline methods on the CHIMERA-BENCH dataset and evaluated based on multiple metrics. *Sequence quality:* amino acid recovery (AAR), contact amino acid recovery (CAAR), and perplexity (PPL). *Structural quality:* Cα RMSD. *Interface quality:* fraction of native contacts (fnat) and DockQ. *Epitope awareness:* epitope F1. *Safety:* number of sequence liabilities (n_{liab}).

5.2. Implementation Details

We use ESM-2 with 650M parameters ($d_{\text{esm}} = 1280$) as the PLM backbone. The RelationEGNN encoder has 5 layers

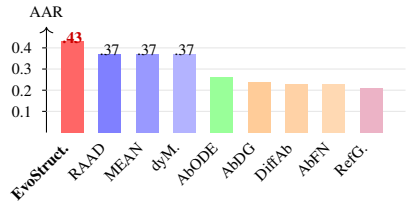

Figure 2. **Amino acid recovery (AAR) comparison.** EVOSTRUCT (red) achieves 0.43 AAR, a 16% relative improvement over the best GNN baselines (0.37). Blue: GNN methods. Orange: diffusion/flow. Green: ODE. Purple: autoregressive.

with hidden dimension $d_{\text{gnn}} = 256$, amino acid embedding dimension 32, and 8 edge types. We crop the $K_{\text{ag}} = 128$ nearest antigen residues as structural context. The adapter operates in dimension $d_a = 640$ with $H = 8$ attention heads and an FFN expansion ratio of 2. The sequence head projects from 1280 to 640 to 25 dimensions.

We train with Adam optimizer, gradient clipping at 0.5, dropout rate 0.2, and batch size 4, using early stopping with patience 10 on validation loss. The loss weights are determined by hyperparameter sweep ($\lambda_{\text{coord}} = 1.376$, $\lambda_{\text{pair}} = 0.525$, $\lambda_{\text{dock}} = 0.5$, $\lambda_{\text{shadow}} = 0.3$, $\alpha_{\text{rd}} = 1.0$). Training follows a three-phase progressive unfreezing schedule (Howard & Ruder, 2018). Phase 1 (up to 50 epochs, lr $= 10^{-4}$) trains the adapter, GNN, and sequence head with ESM-2 frozen. Phase 2 (40 epochs, lr $= 5\times10^{-5}$) unfreezes the top 4 ESM-2 transformer layers. Phase 3 (30 epochs, lr $= 10^{-5}$) fine-tunes all parameters at a low learning rate. All phases use exponential LR decay ($\gamma = 0.9$). It takes approximately 3 hours to train the model on a single H100 GPU.

5.3. Main Results

Table 1 presents the main comparison. EVOSTRUCT achieves the highest amino acid recovery (0.43) among all eleven methods, a 16% relative improvement over the best GNN baselines (RAAD, MEAN, dyMEAN at 0.37). The improvement is accompanied by a 43% reduction in perplexity (1.88 vs. 3.27 for RAAD), indicating substantially more confident and calibrated predictions. Structurally, EVOSTRUCT produces CDR backbones with RMSD 1.84 Å, comparable to RAAD (1.75 Å) and MEAN (1.84 Å). Interface metrics (fnat 0.61, DockQ 0.70) are competitive with the best GNN baselines, and epitope F1 (0.73) matches the GNN range.

RefineGNN achieves the highest fnat (0.65) and DockQ (0.73) despite not conditioning on the antigen, confirming the finding from prior analyses (Ahmed et al., 2026) that antigen conditioning does not currently improve binding quality in existing architectures. EVOSTRUCT narrows this gap while simultaneously leading in sequence recovery, suggesting that PLM-derived representations offer a path to-

Table 1. CDR-H3 design results on the CHIMERA-BENCH benchmark. We report mean ± std across complexes. **Bold** indicates the best value, underline indicates second best. "–" indicates the method does not produce perplexity scores.

Method	AAR ↑	CAAR ↑	PPL ↓	RMSD ↓	fnat ↑	DockQ ↑	EpiF1 ↑	n_{liab} ↓
EvoStruct	**0.43±0.19**	0.22±0.27	**1.88±0.43**	1.84±0.82	0.61±0.31	0.70±0.19	0.73±0.24	0.70±0.84
RAAD	0.37±0.12	0.21±0.22	3.27±0.48	**1.75±0.77**	0.56±0.30	0.70±0.15	0.72±0.25	0.57±0.65
MEAN	0.37±0.13	**0.24±0.23**	3.10±0.47	1.84±0.75	0.57±0.31	0.69±0.15	0.72±0.25	0.58±0.49
dyMEAN	0.37±0.13	0.22±0.23	3.29±0.40	2.22±0.97	0.53±0.31	0.65±0.15	0.64±0.28	0.83±0.38
AbODE	0.26±0.12	0.20±0.22	11.70±4.34	14.64±3.21	0.11±0.21	0.37±0.15	0.27±0.25	1.50±1.90
DiffAb	0.23±0.12	0.14±0.19	–	2.49±1.28	0.59±0.31	0.65±0.16	0.64±0.25	0.52±0.71
AbFlowNet	0.23±0.11	0.14±0.18	–	2.38±1.22	0.60±0.31	0.66±0.16	0.65±0.25	0.38±0.63
AbMEGD	0.21±0.12	0.12±0.16	–	2.44±1.29	0.56±0.29	0.64±0.14	0.64±0.25	0.49±0.67
RADAb	0.20±0.12	0.11±0.17	–	5.33±17.22	0.49±0.32	0.60±0.17	0.60±0.27	0.56±0.74
dyAb	0.19±0.08	0.09±0.14	–	2.34±0.87	0.14±0.21	0.45±0.09	0.24±0.31	**0.00±0.00**
RefineGNN	0.21±0.11	0.10±0.14	8.46±3.28	2.86±0.87	**0.65±0.28**	**0.73±0.14**	**0.76±0.22**	0.71±0.80
AbDockGen	0.24±0.12	0.12±0.18	8.04±2.65	4.67±1.32	0.41±0.26	0.55±0.14	0.62±0.22	0.70±0.79

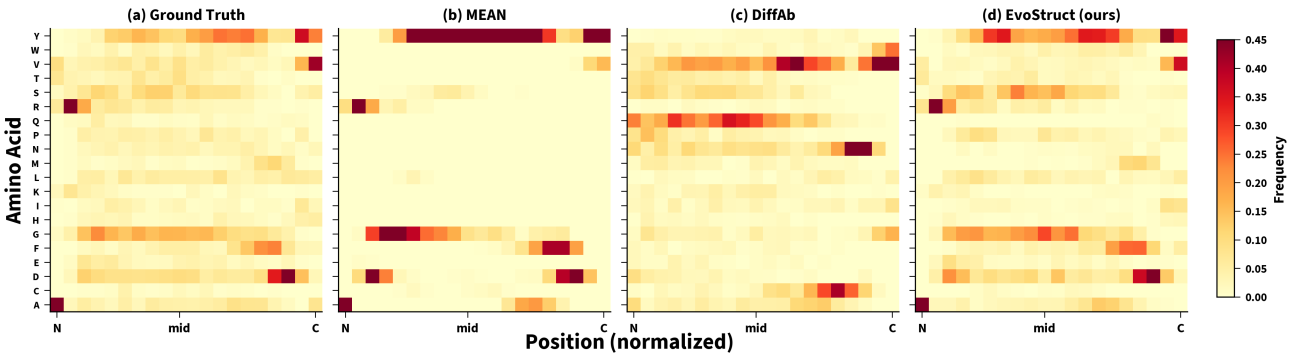

Figure 3. **Vocabulary collapse.** Per-position amino acid frequency heatmaps. EvoStruct preserves near-native diversity (V_{eff}=12.4).

ward methods that excel in both sequence and structural quality.

5.4. Failure Mode Resolution

We evaluate EvoStruct on the three failure modes identified in Section 3.3.

Vocabulary diversity. EvoStruct achieves $V_{\text{eff}} = 12.4$, recovering 80% of the ground-truth amino acid diversity (Figure 3). By contrast, the best GNN baselines (RAAD, MEAN, dyMEAN) collapse to V_{eff} of 4.7–5.3, recovering only 30–34% of native diversity. EvoStruct produces 282 unique bigrams and 1,214 unique trigrams (vs. 364/1,818 in the ground truth and 52/110 for RAAD), covering all 20 of the top ground-truth motifs. This demonstrates that operating in ESM-2's embedding space preserves the PLM's calibrated vocabulary, breaking the diversity-accuracy tradeoff that constrains GNN-only methods.

Binding-pair correlation. EvoStruct achieves the highest interface enrichment correlation with ground truth ($r = 0.81$ vs. next-best 0.68 for RAAD) and the highest paratope-epitope pair correlation ($r = 0.73$ vs. next-best 0.69 for AbFlowNet), as shown in Figure 4. This indicates that the structural adapter successfully routes antigen information to the sequence head, enabling EvoStruct to learn which amino acids pair with which epitope residues.

Positional recovery. EvoStruct improves AAR at both contact positions (22.6% vs. 20.6% for RAAD) and non-contact positions (51.0% vs. 43.5% for RAAD), as shown in Figure 5. The larger gain at non-contact positions (+7.5pp) reflects ESM-2's strong framework priors, which provide better predictions at conserved positions not constrained by the binding partner. The per-position AAR profile along CDR-H3 shows that EvoStruct dominates at anchor positions (71.2% vs. 60.7% for RAAD), where ESM-2's evolutionary priors are strongest, and improves at the hypervariable apex (22.4% vs. 19.0%). Contact AAR remains the hardest challenge for all methods, suggesting that deeper antigen conditioning mechanisms beyond cross-attention may be needed.

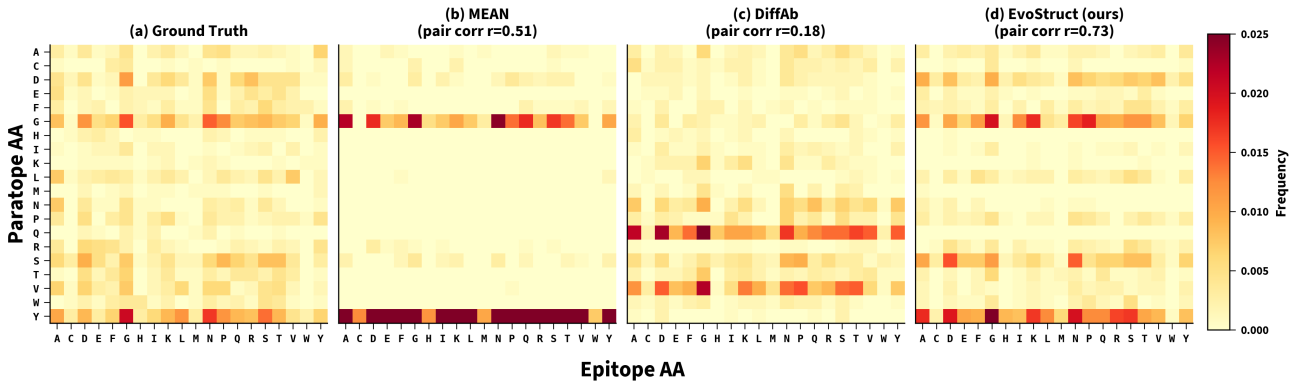

Figure 4. **Binding-pair preferences.** Paratope-epitope AA pair heatmaps. EVOSTRUCT achieves the highest pair correlation (r=0.73) with ground truth, capturing binding-specific preferences.

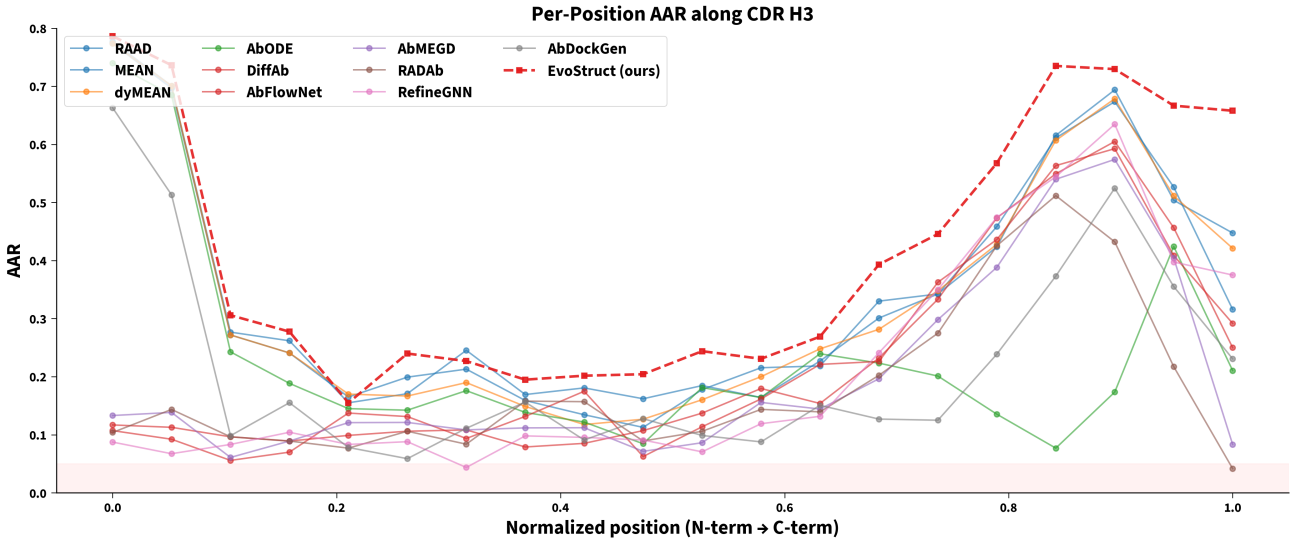

Figure 5. **Positional AAR profile.** Per-position recovery along CDR-H3. EVOSTRUCT (red dashed) leads at both anchor positions and the hypervariable apex.

5.5. Discussion

The results confirm that the PLM-structure adapter paradigm (Zheng et al., 2023), originally developed for general protein inverse folding, transfers effectively to CDR design. The vocabulary calibration of ESM-2 transfers directly through the adapter, without requiring the model to relearn amino acid substitution patterns from limited structural data. GNN methods must discover these patterns from ∼3,000 training complexes, which leads them to converge on the few most frequent amino acids per position. EVOSTRUCT avoids this by inheriting ESM-2's distribution over all 20 amino acids, achieving higher diversity and higher accuracy simultaneously.

The adapter architecture enables a clear separation of concerns. The GNN encoder focuses on encoding 3D geometry and spatial relationships, while ESM-2 handles amino acid vocabulary calibration and evolutionary substitution patterns. The cross-attention mechanism allows structural context to modulate evolutionary priors without overwriting them.

The main limitation is that EVOSTRUCT does not achieve the highest binding metrics. RefineGNN, which does not condition on the antigen at all, achieves fnat 0.65 and DockQ 0.73 compared to EVOSTRUCT's 0.61 and 0.70. This suggests that the cross-attention adapter, while effective for sequence quality, does not fully exploit antigen structural information for binding interface prediction. Contact AAR remains at 22.6% across methods, pointing to a fundamental difficulty in predicting antigen-specific amino acid identity at binding positions that may require more expressive conditioning mechanisms or alternative training objectives. Future work could explore richer adapter architectures, alternative sequence objectives that encourage antigen-specific

predictions at contact positions, and integration with explicit contact prediction mechanisms.

6. Conclusion

In this paper, we diagnose vocabulary collapse as a systematic failure mode in GNN-based CDR design and resolve it by adapting the PLM-structure bridging paradigm for CDR design. EVOSTRUCT cross-attends ESM-2 CDR embeddings to GNN-encoded structural features, keeping the sequence prediction pathway in the PLM's calibrated representation space. Several experiments on the CHIMERA-BENCH benchmark demonstrate that EVOSTRUCT breaks the diversity-accuracy tradeoff, achieving the highest sequence recovery and vocabulary diversity among eleven baselines with the strongest binding-pair correlation.

Impact Statement

This paper presents work whose goal is to advance computational antibody design. The designed sequences require extensive experimental validation before any therapeutic application. We see no specific negative societal consequences that must be highlighted.

References

Abir, A. R., Shahgir, H. S., Ratul, M. R. Z., Tahmid, M. T., Steeg, G. V., and Dong, Y. Abflownet: Optimizing antibody-antigen binding energy via diffusion-gflownet fusion. *arXiv preprint arXiv:2505.12358*, 2025.

Ahmed, M., Taj, N., Khan, I. U., Venkateswara, H., and Patterson, M. CHIMERA-bench: A benchmark dataset for epitope-specific antibody design. In *ICLR 2026 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2026. URL https://openreview.net/forum?id=PyZvVIJbSy.

Chen, J., Cai, X., Wu, J., and Hu, W. Antibody design and optimization with multi-scale equivariant graph diffusion models for accurate complex antigen binding. *arXiv preprint arXiv:2506.20957*, 2025.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Chinery, L., Hummer, A. M., Mehta, B. B., Akbar, R., Rawat, P., Slabodkin, A., Le Quy, K., Lund-Johansen, F., Greiff, V., Jeliazkov, J. R., and Deane, C. M. Simple computational methods can outperform deep learning in designing diverse, binder-enriched antibody libraries. *bioRxiv*, 2024. doi: 10.1101/2024.03.26.586756.

Chothia, C. and Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of molecular biology*, 196(4):901–917, 1987.

Høie, M. H., Hummer, A. M., Olsen, T. H., Aguilar-Sanjuan, B., Nielsen, M., and Deane, C. M. Antifold: improved structure-based antibody design using inverse folding. *Bioinformatics Advances*, 5(1):vbae202, 2025.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 328–339, 2018.

Jin, W., Barzilay, R., and Jaakkola, T. Antibody-antigen docking and design via hierarchical structure refinement. In *International Conference on Machine Learning*, pp. 10217–10227. PMLR, 2022a.

Jin, W., Wohlwend, J., Barzilay, R., and Jaakkola, T. Iterative refinement graph neural network for antibody sequence-structure co-design. In *International Conference on Learning Representations*, 2022b.

Kong, X., Huang, W., and Liu, Y. Conditional antibody design as 3D equivariant graph translation. In *International Conference on Learning Representations*, 2023a.

Kong, X., Huang, W., and Liu, Y. End-to-end full-atom antibody design. In *International Conference on Machine Learning*, pp. 17409–17429. PMLR, 2023b.

Li, Y., Lang, Y., Xu, C., Zhou, Y., Pang, Z., and Greisen, P. J. Benchmarking inverse folding models for antibody CDR sequence design. *PLOS ONE*, 20(6):e0324566, 2025. doi: 10.1371/journal.pone.0324566.

Liang, X., Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W., Zhang, M., and Liu, T.-Y. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., and Ma, J. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35: 9754–9767, 2022.

Olsen, T. H., Moal, I. H., and Deane, C. M. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics*, 40(11): btae618, 2024.

Satorras, V. G., Hoogeboom, E., and Welling, M. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Shanehsazzadeh, A., Alverio, J., Kasun, G., Levine, S., Khan, J. A., Chung, C., Diaz, N., Luton, B. K., Tarter, Y., McCloskey, C., et al. In vitro validated antibody design against multiple therapeutic antigens using generative inverse folding. *BioRxiv*, pp. 2023–2012, 2023.

Tan, C., Zhang, Y., Gao, Z., Huang, Y., Lin, H., Wu, L., Wu, F., Blanchette, M., and Li, S. Z. dyab: Flow matching for flexible antibody design with alphafold-driven prebinding antigen. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 782–790, 2025.

Uçar, T. and Sormanni, P. Blosum is all you learn—generative antibody models reflect evolutionary priors. *bioRxiv*, pp. 2025–10, 2025.

Verma, Y., Heinonen, M., and Garg, V. Abode: Ab initio antibody design using conjoined odes. In *International Conference on Machine Learning*, pp. 35037–35050. PMLR, 2023.

Wang, Z., Ji, Y., Tian, J., and Zheng, S. Retrieval augmented diffusion model for structure-informed antibody design and optimization. *arXiv preprint arXiv:2410.15040*, 2024.

Wu, J., Kong, X., Sun, N., Wei, J., Shan, S., Feng, F., Wu, F., Peng, J., Zhang, L., Liu, Y., and Ma, J. Flowdesign: Improved design of antibody cdrs through flow matching and better prior distributions. *Cell Systems*, 2025a. doi: 10.1016/j.cels.2025.101270.

Wu, L., Lin, H., Huang, Y., Gao, Z., Tan, C., Liu, Y., Wu, T., and Li, S. Z. 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, 2025b.

Zheng, Z., Deng, Y., Xue, D., Zhou, Y., Ye, F., and Gu, Q. Structure-informed language models are protein designers. In *International Conference on Machine Learning*, volume 202, pp. 42317–42338. PMLR, 2023.

Zhu, Y., Shi, X., Zhang, J., Sun, W., and Wang, L. Abegdiffuser: Antibody sequence-structure codesign with equivariant graph neural networks and diffusion models. *Journal of Chemical Theory and Computation*, 21(21):11307–11317, 2025.

7. Appendix

7.1. Detailed Results

Below, we provide the detailed results for our benchmarking of the baseline methods and compare them with EVOSTRUCT. Figure 6 shows the vocabulary collapse failure mode, while Figure 7 shows the failures of the baselines with regard to their binding pairs preference.

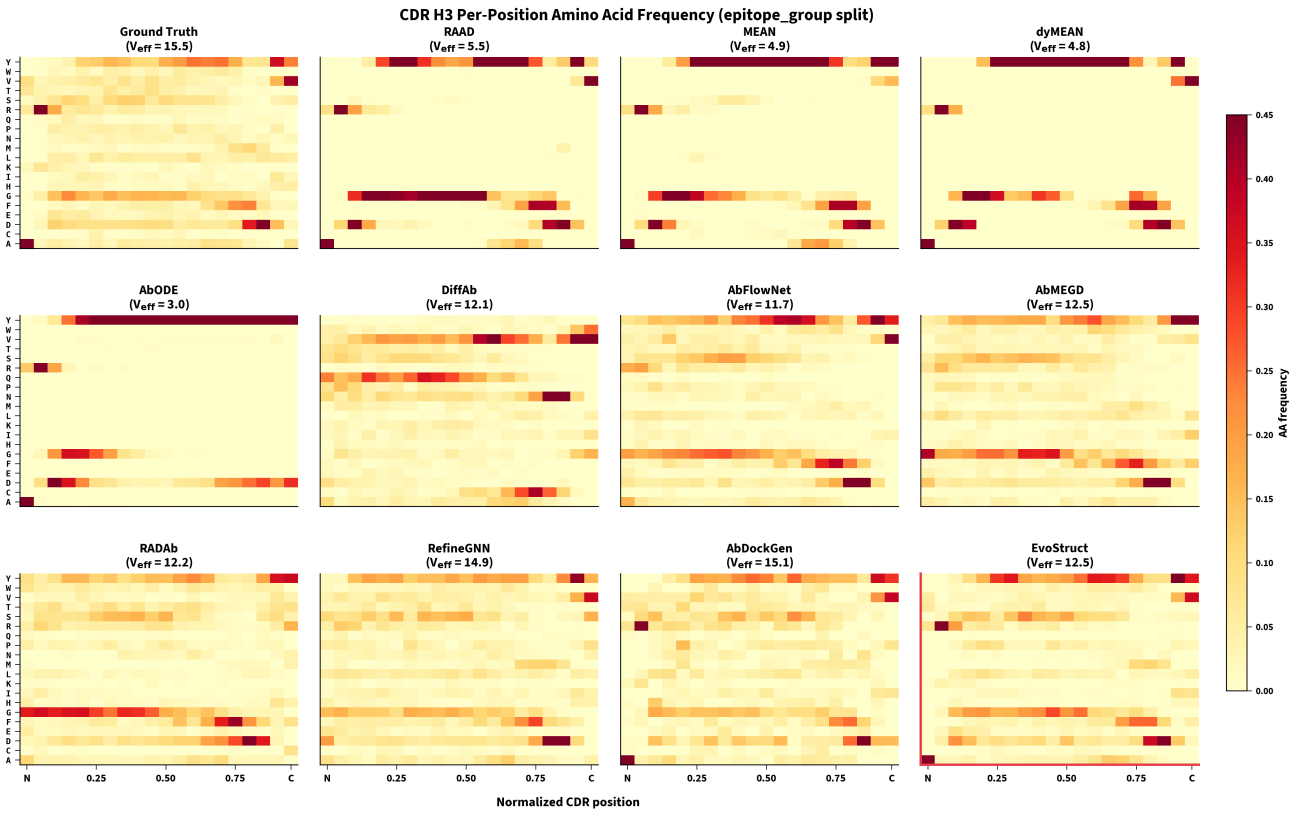

Figure 6. **Vocabulary collapse.** Per-position amino acid frequency heatmaps across all the benchmarked baseline methods.

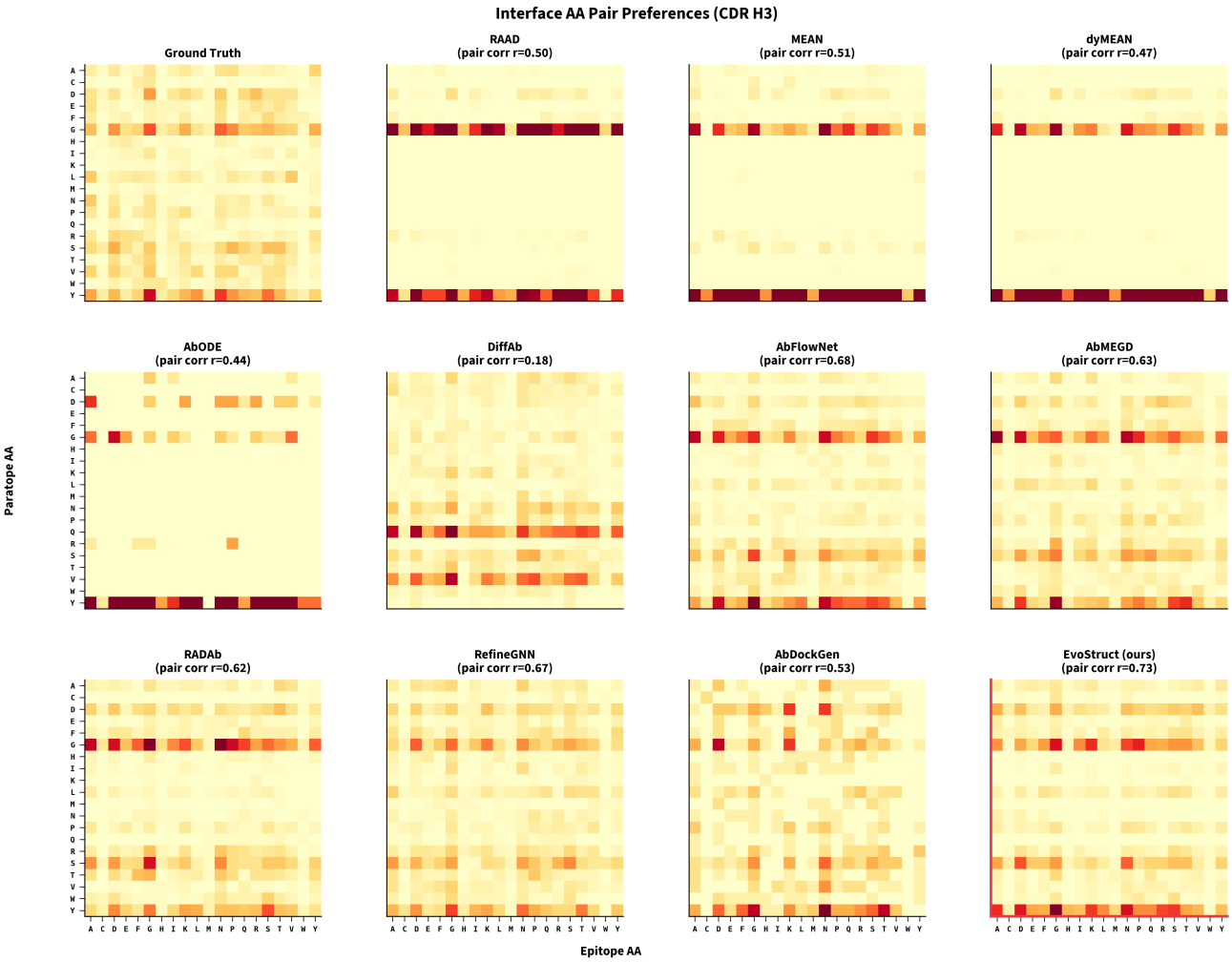

Figure 7. **Binding-pair preferences.** Paratope-epitope AA pair heatmaps across all the benchmarked baseline methods.