Chimera-Bench: A Benchmark Dataset for Epitope-Specific Antibody Design

Anonymous authors
Paper under double-blind review

Abstract

Computational antibody design has seen rapid methodological progress, with dozens of deep generative methods proposed in the past three years, yet the field lacks a standardized benchmark for fair comparison and model development. These methods are evaluated on different SAbDab snapshots, non-overlapping test sets, and incompatible metrics, and the literature fragments the design problem into numerous sub-tasks with no common definition. We introduce Chimera-Bench (**C**DR **M**odeling with **E**pitope-guided **R**edesign), a unified benchmark built around a single canonical task: *epitope-conditioned CDR sequence–structure co-design*. Chimera-Bench provides three components. The first is a curated, deduplicated dataset of **2,922** antibody–antigen complexes with epitope and paratope annotations. The second is a set of three biologically motivated splits that test generalization to unseen epitopes, unseen antigen folds, and prospective temporal targets. The third is a comprehensive evaluation protocol with five metric groups, including novel epitope-specificity measures. We benchmark eleven methods spanning six generative paradigms and report results across all splits. Chimera-Bench is the largest dataset of its kind for the antibody design problem, allowing the community to develop and test novel methods and evaluate their generalizability.

1 Introduction

Antibodies are among the most important classes of biotherapeutics (Norman et al., 2020). Their binding specificity is largely determined by six complementarity-determining regions (CDRs), particularly CDR-H3, which makes these loops the primary targets for computational design. Deep generative models have recently transformed this space, with diffusion models (Luo et al., 2022; Martinkus et al., 2023), flow matching (Tan et al., 2025), equivariant graph neural networks (Wu et al., 2025; Kong et al., 2022; 2023), autoregressive models (Jin et al., 2021), and foundation models (Wang et al., 2025) all proposed for CDR sequence–structure co-design. Yet despite this rapid progress, there is no standardized way to compare these methods. They are trained on different snapshots of the Structural Antibody Database (SAbDab) (Dunbar et al., 2014) with varying filtering criteria, evaluated on non-overlapping test sets from the RAbD dataset (Adolf-Bryfogle et al., 2018) to hold-outs, and scored with incompatible metrics computed under different definitions. For example, contact cutoffs vary from 4.5 to 6.6 Å across methods, and RMSD is computed with or without Kabsch alignment. The methods also require inputs in different and incompatible formats, making head-to-head comparison impractical without building separate data pipelines for each method.

A more fundamental issue is the absence of a common task definition. The literature fragments antibody design into numerous sub-tasks, including inverse folding, structure prediction, co-design, docking, affinity optimization, de novo generation, and epitope-conditioned design, and individual methods address different subsets. This makes it unclear what is being compared, even when two papers report the same metric name. We argue that in many therapeutic settings, the target epitope is specified and propose *epitope-conditioned CDR sequence–structure co-design* as the canonical formulation. Given an antigen structure, an epitope specification, and an antibody framework, the task is to design CDR residues that are structurally valid, contact the target epitope, and avoid off-target binding. This formulation subsumes the sub-tasks above as special cases and accommodates all baseline methods. Methods that natively support epitope conditioning

map directly to this definition, while the remaining methods can be evaluated as-is, with epitope-specificity metrics revealing whether their designs contact the intended site.

We introduce Chimera-Bench (**C**DR **M**odeling with **E**pitope-guided **R**edesign), a benchmark dataset for the antibody design task. Our contributions are:

- A **canonical problem definition** that unifies seven sub-tasks from the literature into a single epitope-conditioned CDR co-design formulation.
- A **curated dataset** of deduplicated, quality-filtered antibody–antigen complexes from SAbDab, CDR masks, epitope/paratope annotations, and contact maps.
- **Three evaluation splits** (epitope-group, antigen-fold, and temporal), each with cluster-level leakage prevention, testing generalization along distinct biological axes.
- A **comprehensive evaluation protocol** with five metric groups covering sequence quality, structural accuracy, binding interface quality, epitope specificity, and designability, including novel epitope-specificity metrics not reported by any existing method.
- The **first head-to-head retraining** of eleven methods spanning six generative paradigms under identical data, splits, and metrics, together with the empirical findings this comparison exposes.

2 Related Work

Antibody CDR design methods. Early learning-based approaches to CDR design were autoregressive. RefineGNN (Jin et al., 2021) generates CDR residues left-to-right while iteratively refining the predicted structure, and AbDockGen (Jin et al., 2022) combines hierarchical equivariant refinement with antibody–antigen docking. MEAN (Kong et al., 2022) recast the problem as E(3)-equivariant graph translation, and dyMEAN (Kong et al., 2023) extended this to end-to-end full-atom antibody design. RAAD (Wu et al., 2025) later introduced relation-aware equivariance for settings where the epitope is unknown. Diffusion models have since become the dominant paradigm, beginning with DiffAb (Luo et al., 2022), which pioneered antigen-conditioned CDR diffusion over sequences, coordinates, and orientations simultaneously. AbX (Zhu et al., 2024) augmented this framework with evolutionary and physical constraints, and LEAD (Yao et al., 2025) introduced property-guided sampling through black-box optimization in a shared latent space. dyAb (Tan et al., 2025) applies flow matching to flexible antibody design using AlphaFold-predicted pre-binding antigen conformations, and AbODE (Verma et al., 2023) formulates CDR generation through conjoined ODEs. RFAntibody (Bennett et al., 2025) fine-tunes RFdiffusion on ∼8,100 PDB antibody structures for de novo design of complete variable regions (VHHs and scFvs) conditioned on a target epitope. In the broader protein design space, BoltzGen (Stark et al., 2025) is an all-atom diffusion model that unifies binder design and structure prediction across modalities, including nanobodies, though it targets general binder design rather than CDR-level co-design. ProteinMPNN (Dauparas et al., 2022), while also not antibody-specific, serves as a strong inverse folding baseline for CDR sequence design given a fixed backbone.

Existing benchmarks and datasets. SAbDab (Dunbar et al., 2014) is the primary structural database for antibodies but provides no standardized splits, evaluation protocol, or task definition for generative design and contains redundant complexes that require preprocessing for training neural networks. The RAbD benchmark (Adolf-Bryfogle et al., 2018) defines 60 antibody–antigen complexes for CDR-H3 redesign but lacks epitope annotations and modern generative metrics such as DockQ or diversity. SKEMPI v2 (Jankauskaitė et al., 2019) provides experimental binding affinity measurements for protein complexes including approximately 53 antibody entries, but targets mutation-level affinity prediction rather than generative design. AbBiBench (Zhao et al., 2025) benchmarks antibody binding affinity maturation with 184,500+ experimental measurements across 14 antibodies, but focuses on affinity prediction rather than generative CDR design. In the broader protein domain, ProteinGym (Notin et al., 2023) benchmarks fitness prediction across 250+ deep mutational scanning assays and ATOM3D (Townshend et al., 2020) defines molecular learning tasks on 3D structures, but neither addresses the antibody-specific challenges of CDR co-design, epitope conditioning, or binding interface evaluation. To our knowledge, Chimera-Bench is the first benchmark to combine stan-

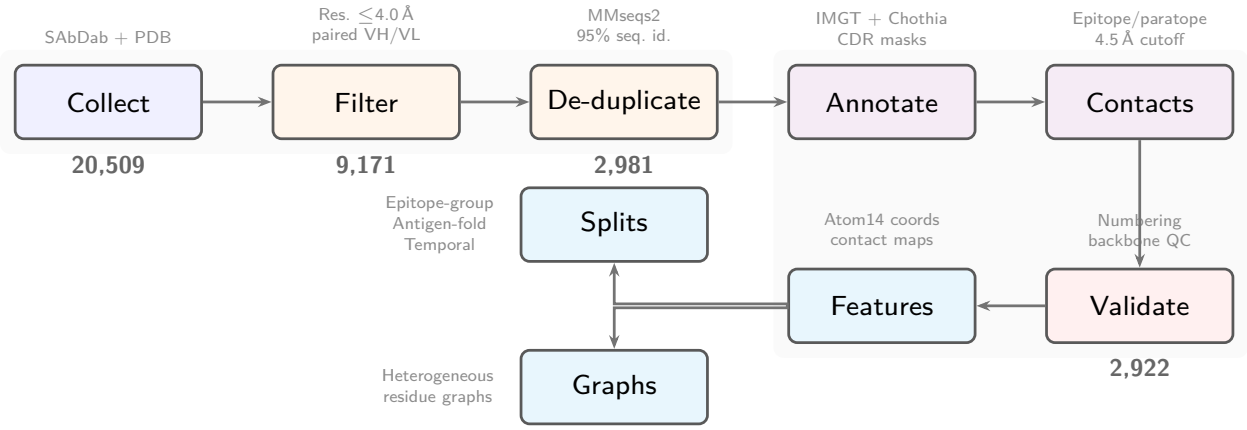

Figure 1: The Chimera-Bench data curation pipeline that collects antigen-antibody complexes from SAbDab and produces filtered and annotated complexes for the antibody design task.

dardized data curation, biologically motivated splits, epitope-specificity metrics, and cross-method format compatibility for computational antibody design.

Table 1 positions Chimera-Bench against prior antibody and protein design benchmarks (RAbD (Adolf-Bryfogle et al., 2018), SAbDab (Dunbar et al., 2014), SKEMPI v2 (Jankauskaitė et al., 2019), AsEP (Liu et al., 2024), AbBiBench (Zhao et al., 2025), ProteinGym (Notin et al., 2023)). Chimera-Bench is the largest curated benchmark of its kind and the first to combine standardized annotations, multiple evaluation splits, and cross-method format compatibility.

Table 1: Comparison of Chimera-Bench with prior antibody design benchmarks and datasets.

	Chimera-Bench	RAbD	SAbDab	SKEMPI v2	AsEP	AbBiBench	ProteinGym
Domain	Ab design	Ab design	Ab structures	Affinity	Epitope pred.	Affinity mat.	Fitness
Complexes	2,922	60	7,000+	53 (Ab)	1,723	14 Ab / 9 Ag	–
Epitope labels	✓	–	–	–	✓	–	–
Paratope labels	✓	–	–	–	✓	–	–
CDR annotations	✓	–	–	–	–	–	–
Multiple splits	3	1	–	1	2	3	3
Leakage prevention	✓	–	–	–	✓	✓	✓
Affinity data	–	–	Partial	✓	–	✓	✓

3 The Chimera-Bench Dataset

3.1 Dataset Construction

The Chimera-Bench dataset is constructed from SAbDab (Dunbar et al., 2014) through an eight-step pipeline (Figure 1). Starting from 20,509 SAbDab entries, we apply quality filters (protein/peptide antigens only, paired VH/VL chains, resolution ≤4.0 Å), cluster at 95% sequence identity using MMseqs2 (Steinegger & Söding, 2017), and validate each complex for ANARCI (Dunbar & Deane, 2016) numberability, conserved residue presence, CDR completeness, and backbone integrity. Residues are numbered under both the IMGT (Lefranc et al., 2003) and Chothia (Chothia & Lesk, 1987) schemes, and epitope/paratope residues are identified at a 4.5 Å contact cutoff. The full processing details, validation criteria, and exclusion breakdown are provided in Appendix A.1.

The final dataset contains **2,922 complexes** from 2,721 unique PDBs (2,485 protein-antigen, 437 peptide-antigen, median resolution 2.72 Å), with on average 17.9 epitope and 20.5 paratope residues per complex. The graph construction details are in Appendix A.7.

3.2 Dataset Statistics

Structural properties. The final dataset comprises 2,922 complexes drawn from 2,721 unique PDB entries at a median resolution of 2.72 Å. Of these, 2,485 (85%) target protein antigens and 437 (15%) target peptide antigens, and the set is dominated by X-ray diffraction (1,929) and cryo-EM (989) structures. Antigen lengths vary widely (288±278 residues), with longer antigens corresponding to viral surface proteins and shorter ones to peptide epitopes. Antibody chains are far more uniform: the IMGT-numbered variable domains span 121±5 (VH) and 108±3 (VL) residues, within full heavy and light chains.

Interface properties. Epitope sizes range from 1 to over 50 residues (mean 17.9), and paratope sizes are slightly larger (mean 20.5), consistent with the observation that paratope residues span multiple CDR loops while epitope residues cluster on a single antigen surface patch. The mean of 45.6 residue-level contact pairs per complex provides a rich signal for evaluating binding interface reconstruction.

CDR statistics. Under IMGT numbering, CDR-H3 is the most variable loop in both length and sequence (mean 14.6±4.3 residues, range 3–63), with a heavy right tail reflecting the diversity of VDJ recombination. CDR-H1 (8.2±0.7) and CDR-H2 (7.9±0.7) show narrower distributions consistent with their more conserved canonical structures. Light chain CDRs are generally more uniform in length: L1 (7.4±2.1), L2 (3.0±0.4, nearly fixed-length under IMGT), and L3 (9.4±1.3). In our dataset, light-chain residues account for a mean of 35% of paratope contacts (IQR 25–46%), consistent with prior analyses (Ramaraj et al., 2012), and are important for overall interface quality. The full CDR length distributions, together with chain-length and interface-size distributions, are shown in Figure 2.

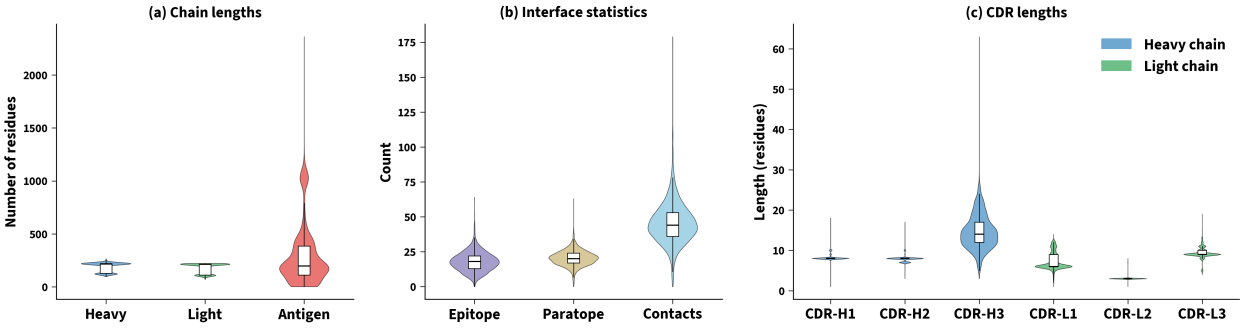

Figure 2: Distributions across the Chimera-Bench dataset. (a) Heavy, light, and antigen chain lengths. (b) Interface statistics, namely epitope size, paratope size, and number of residue contacts per complex. (c) CDR loop lengths under IMGT numbering, where CDR-H3 shows the widest variability consistent with its role as the primary determinant of antigen specificity.

Temporal and species distributions. PDB deposition dates range from the mid-1990s to mid-2024, with a strong concentration in 2019–2023 reflecting the surge in COVID-related antibody structures. The majority of antigens originate from viral proteins or human self-antigens (Figure 3).

3.3 Epitope Diversity

A distinctive feature of Chimera-Bench is that many antigens are targeted by multiple antibodies binding to distinct epitope regions. Of the 2,470 unique antigen sequences in the dataset, 258 (10.4%) are bound at two or more distinct epitopes, collectively accounting for 710 of the 2,922 complexes (24.3%). The most promiscuous antigen is targeted at 30 distinct epitope sites (Figure 4). This epitope diversity is central to Chimera-Bench's evaluation. A successful model must not merely reconstruct a plausible binding interface but must do so for the *correct* epitope, which makes epitope-conditioned generation essential. The epitope-group split explicitly tests this capability by holding out entire epitope clusters at test time.

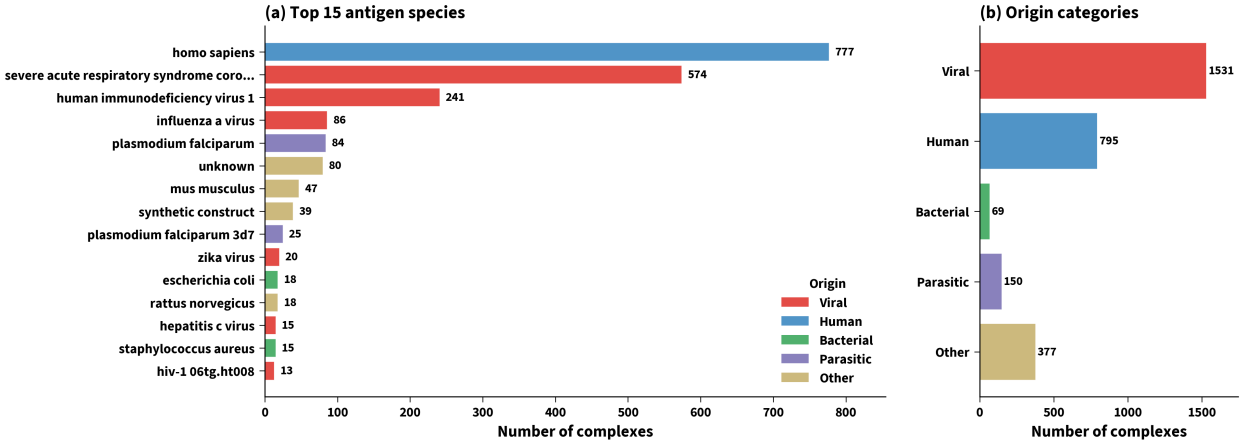

Figure 3: Antigen species and origin distributions. (a) Top 15 species by complex count. (b) Breakdown by biological origin category.

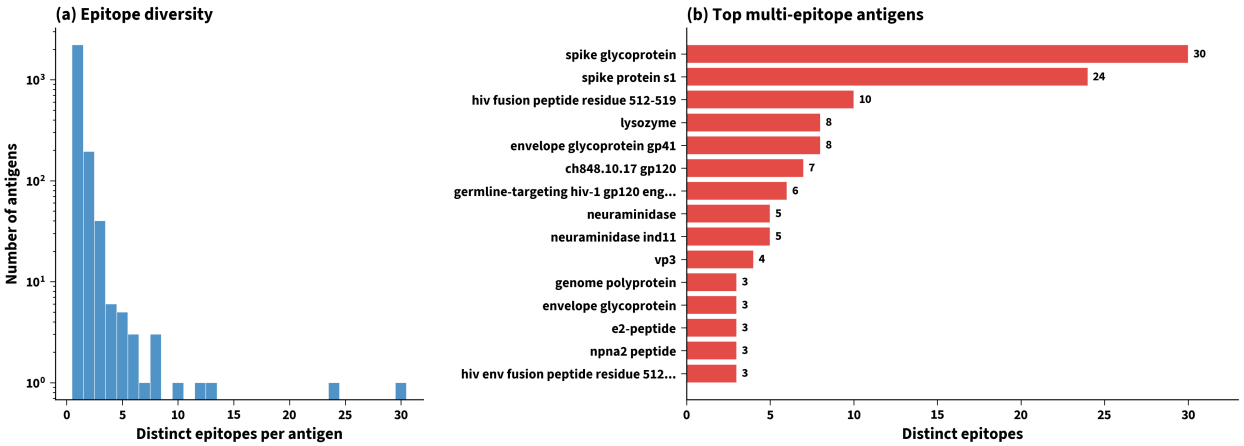

Figure 4: Epitope diversity in Chimera-Bench. (a) Distribution of distinct epitopes per antigen sequence (log scale); most antigens have a single epitope, but a long tail extends to 30. (b) Top 15 multi-epitope antigens ranked by number of distinct epitope sites.

3.4 Data Splits

A central design choice in Chimera-Bench is the use of three biologically motivated splits, each testing a different generalization axis while enforcing cluster-level separation to prevent data leakage (Table 2). All three splits assign whole clusters to a single partition, so no cluster spans training and test. We pack clusters greedily, placing each cluster in the partition with the largest remaining deficit relative to an 80/10/10 target, which keeps the ratios close even when individual clusters are large.

The **epitope-group** split clusters complexes by the three-dimensional shape of their epitope, so the test set contains epitope geometries never seen during training. We represent each epitope by the Cα coordinates of its contacting antigen residues and measure the distance between two equal-sized epitopes by their Kabsch-aligned RMSD. For epitopes of differing size, where superposition is undefined, we fall back to a size-and-spread surrogate that compares residue counts and the radial spread about the patch centroid. We cluster the resulting distance matrix by average-linkage hierarchical clustering at a 3 Å cutoff and assign complexes with fewer than three resolved epitope residues to singleton clusters.

The **antigen-fold** split clusters antigen chains by structural fold with Foldseek (van Kempen et al., 2022) (`easy-cluster`, TM-score ≥ 0.5, bidirectional coverage $\geq 50\%$, ignoring sequence identity), so the test set contains antigen folds absent from training. Structural clustering reaches near-complete coverage of the dataset, whereas CATH superfamily labels cover only about 30% of PDB chains.

The **temporal** split orders all complexes by PDB deposition date and places the oldest in training and the most recent in test, with quantile cutoffs chosen to hit the 80/10/10 ratio. This simulates prospective deployment on targets deposited after the training period. Complexes without a parseable date default to training.

Table 2: Summary of Chimera-Bench evaluation splits.

Split	Train	Val	Test	Generalization target
Epitope-group	2,338	292	292	Unseen epitope patterns
Antigen-fold	2,338	292	292	Unseen antigen topologies
Temporal	2,337	292	293	Prospective (post-2023)

We report the per-partition distributions of the covariates that govern design difficulty, namely CDR-H3 length, epitope size, and antigen size, to contextualize cross-split comparisons (Figure 5). The epitope-group and temporal splits are closely matched across partitions, whereas the antigen-fold split shifts toward smaller antigens and epitopes at test time, a direct consequence of holding out entire structural folds.

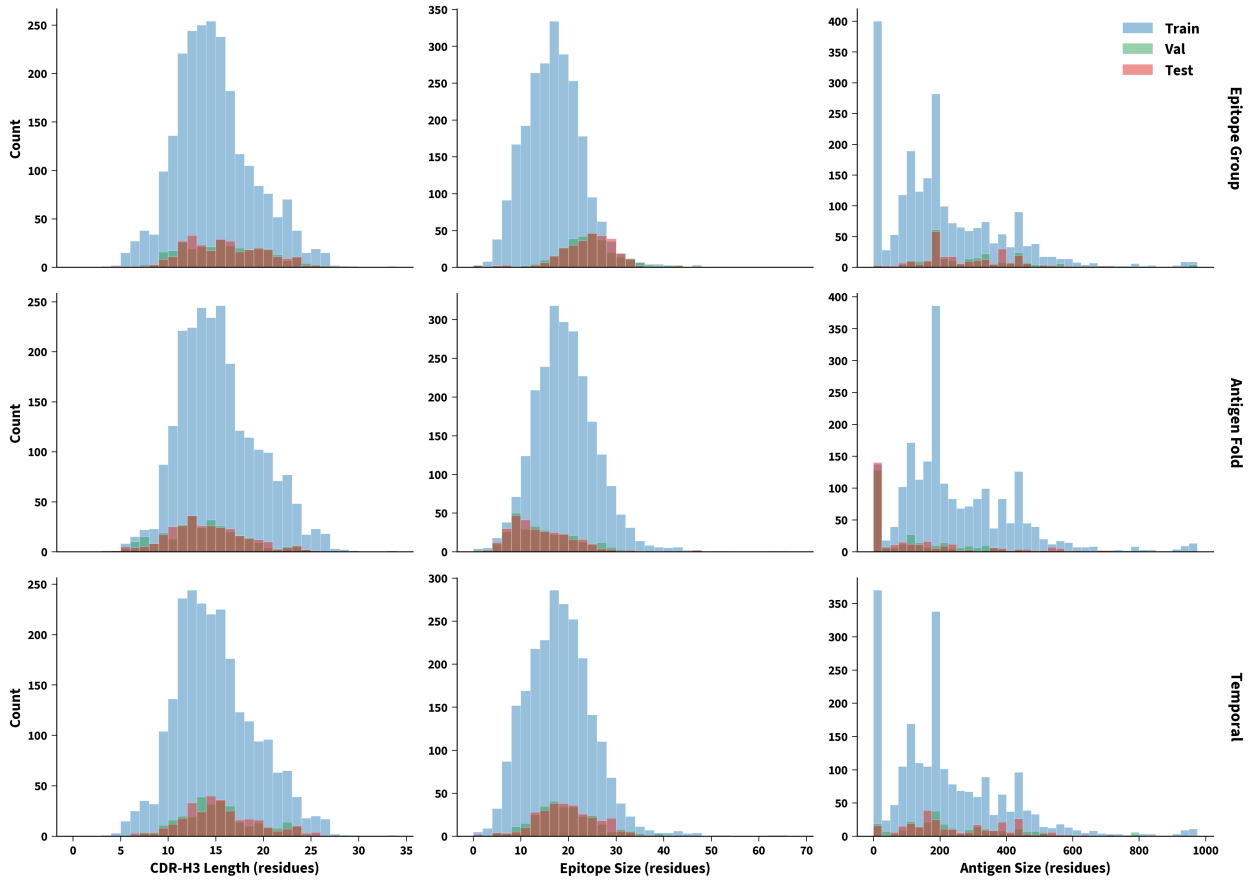

Figure 5: Distribution of CDR-H3 length, epitope size, and antigen size (columns) across the train, validation, and test partitions of the three splits (rows).

4 Task and Evaluation Protocol

4.1 Task Definition

As discussed above, the antibody design literature fragments the problem into numerous sub-tasks, and individual methods address different subsets. CHIMERA-BENCH consolidates these into a single canonical task, *epitope-conditioned CDR sequence–structure co-design*, motivated by the observation that in many therapeutic settings the target epitope is specified. Given an antigen structure $A = \{(s_j, \mathbf{x}_j) \mid j \in V_A\}$, an epitope $E \subseteq V_A$, and an antibody framework $F = \{(s_i, \mathbf{x}_i) \mid i \in V_{\text{FR}}\}$, the task is to design CDR residues that maximize the learned conditional distribution while satisfying epitope contact and precision constraints. Following Luo et al. (2022), each residue is represented as a tuple of amino acid type, Cα coordinate, and local frame orientation. The remaining sub-tasks from the literature are subsumed as special cases: inverse folding fixes the backbone and designs sequence only (the ProteinMPNN setting), structure prediction fixes the sequence, unconditional co-design sets $E = V_A$ (i.e., the entire antigen surface is treated as the target, imposing no epitope constraint), and docking quality is captured implicitly through DockQ and iRMSD.

The task can be expressed concisely as:

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

where $R = \{(s_k, \mathbf{x}_k, \mathbf{O}_k)\}_{k \in V_{\text{CDR}}}$ denotes the designed CDR residues with amino acid type s_k, Cα coordinate $\mathbf{x}_k$, and local frame orientation $\mathbf{O}_k$, and $\mathcal{C}(R, A) = \{j \in V_A \mid \exists\, k \in V_{\text{CDR}} : \|\mathbf{x}_k - \mathbf{x}_j\| < d_c\}$ is the set of antigen residues contacted by the designed CDRs within a cutoff distance d_c.

4.2 Evaluation Protocol

We evaluate designs across five metric groups that collectively assess sequence quality, structural accuracy, binding interface quality, epitope specificity, and designability. **Sequence quality** is measured by amino acid recovery (AAR) and contact AAR (CAAR), which restricts this calculation to paratope residues within 4.5 Å of the antigen, along with perplexity (PPL) for models that produce sequence likelihoods. **Structural quality** is assessed by Cα RMSD after Kabsch alignment and TM-score (Zhang & Skolnick, 2004). We standardize on Kabsch-aligned RMSD throughout, as existing methods use inconsistent implementations. The **binding interface quality** is captured by the fraction of native contacts recovered (Fnat), interface RMSD (iRMSD) restricted to contact residues, and the DockQ composite score (Basu & Wallner, 2016), which combines Fnat, iRMSD, and ligand RMSD into a single metric. **Epitope specificity** is a novel evaluation axis introduced by CHIMERA-BENCH. We compute epitope precision, recall, and F1, measuring whether a design contacts the intended binding site. **Designability** counts known sequence liability motifs associated with manufacturing issues. We use two contact definitions for two distinct purposes. Dataset annotations and the paratope mask use a strict 4.5 Å heavy-atom cutoff, which identifies physical contacts for graph construction and CAAR. The interface and epitope metrics (Fnat, iRMSD, DockQ, EpiF1) instead use an 8 Å Cα–Cα cutoff computed symmetrically on both native and predicted structures, a definition that tolerates the Cα-level coordinate error of generative models without rewarding spurious atom-level clashes.

We further define two evaluation tracks.

1. **CDR-H3 co-design** is mandatory for all methods, since CDR-H3 is the most variable loop and the primary determinant of antigen specificity.

2. **All-CDR co-design** evaluates methods that support multi-loop generation, designing all six CDRs (H1–H3, L1–L3) simultaneously.

Few baseline methods natively support epitope conditioning and map directly to the canonical definition. Our epitope-specificity metrics then reveal whether their designs contact the intended binding site. The formal definitions are provided in Appendix A.2.

5 Experiments

We benchmark eleven antibody design methods spanning six generative paradigms: equivariant GNNs (RAAD (Wu et al., 2025), MEAN (Kong et al., 2022), dyMEAN (Kong et al., 2023)), diffusion models and their extensions (DiffAb (Luo et al., 2022), AbMEGD (Chen et al., 2025), RADAb (Wang et al., 2024), AbFlowNet (Abir et al., 2025)), flow matching (dyAb (Tan et al., 2025)), autoregressive generation (RefineGNN (Jin et al., 2021)), hierarchical equivariant networks (AbDockGen (Jin et al., 2022)), and conjoined ODEs (AbODE (Verma et al., 2023)). Each method is retrained on Chimera-Bench using the authors' released code with default hyperparameters. The training details are in Appendix A.6.

5.1 CDR-H3 Design Results

Table 3 reports CDR-H3 co-design results on the epitope-group test split, our primary evaluation setting since it holds out entire epitope clusters. Full results on the antigen-fold and temporal splits are deferred to Appendix A.4 (Tables 8 and 9), and Figure 6 visualizes the core metrics across all three splits. All values are computed using the standardized Chimera-Bench evaluation pipeline with Kabsch-aligned RMSD. Perplexity is reported as "–" for diffusion and flow-based models that do not produce sequence likelihoods. RADAb's large CDR-H3 RMSD and iRMSD with std larger than the mean reflect a minority of diffusion sampling failures that place the loop far off native.

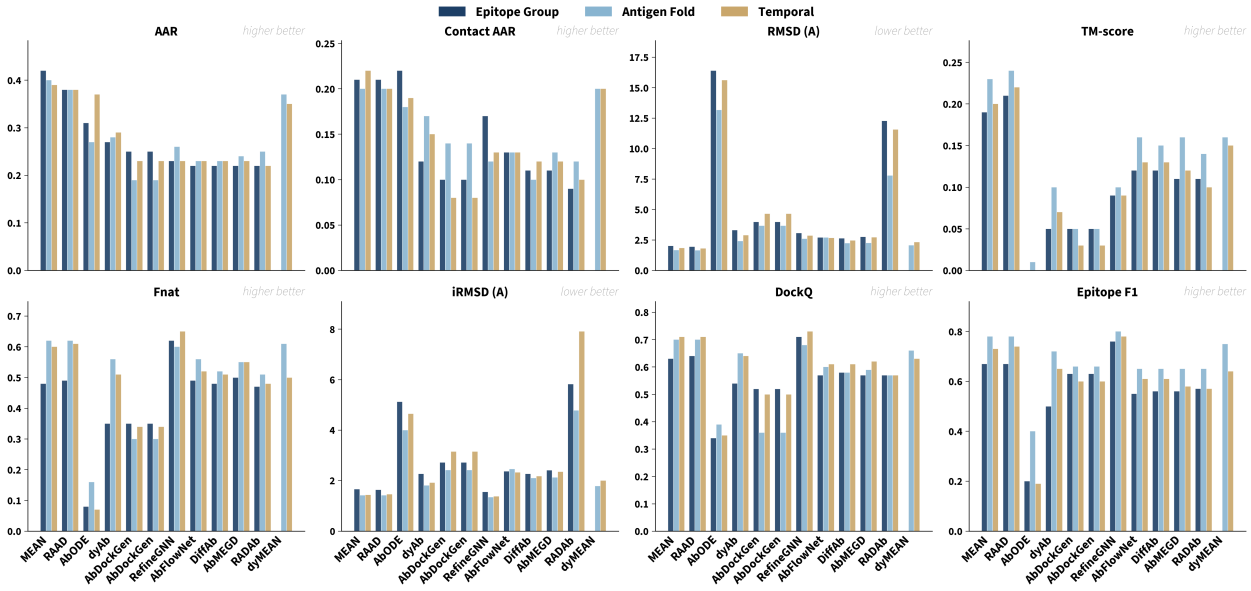

Figure 6: CDR-H3 design metrics across all methods on the three test splits (EG: epitope-group, AF: antigen-fold, TP: temporal).

Table 3: CDR-H3 design results on the **epitope-group** test split. Best values are in **bold**, second-best are underlined. "–" indicates a metric that is not applicable (PPL for diffusion and flow models).

Method	AAR↑	CAAR↑	PPL↓	RMSD↓	TM↑	Fnat↑	iRMSD↓	DockQ↑	EpiF1↑	n_liab.↓
RAAD	0.38±0.13	0.21±0.23	3.31±0.47	**1.95±0.87**	**0.21±0.13**	0.49±0.30	1.64±0.86	0.64±0.21	0.67±0.27	0.71±0.62
MEAN	**0.42±0.15**	0.21±0.23	**3.00±0.61**	2.01±0.89	0.19±0.12	0.48±0.31	1.66±0.86	0.63±0.21	0.67±0.29	0.90±0.92
dyMEAN	0.35±0.12	0.17±0.20	3.38±0.46	2.60±0.94	0.12±0.07	0.41±0.30	2.23±0.84	0.56±0.19	0.55±0.29	0.83±0.38
DiffAb	0.22±0.12	0.11±0.18	–	2.64±1.19	0.12±0.12	0.48±0.32	2.27±0.98	0.58±0.23	0.56±0.27	0.62±0.77
AbFlowNet	0.22±0.12	0.13±0.18	–	2.70±1.24	0.12±0.12	0.49±0.33	2.37±1.01	0.57±0.22	0.55±0.28	0.46±0.68
AbMEGD	0.22±0.11	0.11±0.18	–	2.76±1.26	0.11±0.11	0.50±0.32	2.41±1.04	0.57±0.22	0.56±0.27	0.47±0.72
RADAb	0.22±0.12	0.09±0.16	–	12.28±75.62	0.11±0.10	0.47±0.32	5.82±31.86	0.57±0.22	0.57±0.28	0.83±0.90
dyAb	0.27±0.10	0.12±0.17	–	3.31±0.84	0.05±0.03	0.35±0.28	2.27±0.86	0.54±0.18	0.50±0.32	0.98±0.15
RefineGNN	0.23±0.12	0.17±0.23	7.80±2.73	3.07±0.72	0.09±0.06	**0.62±0.20**	**1.55±0.58**	**0.71±0.09**	**0.76±0.10**	0.81±0.68
AbDockGen	0.23±0.11	0.09±0.16	7.89±2.94	3.93±1.26	0.05±0.03	0.34±0.25	2.64±1.03	0.52±0.17	0.62±0.27	0.60±0.74
AbODE	0.31±0.13	**0.22±0.21**	9.99±3.50	16.40±4.67	0.00±0.00	0.08±0.12	5.12±2.24	0.34±0.08	0.20±0.19	**0.08±0.27**

5.2 All-CDR Design Results

Beyond CDR-H3, several methods design all six CDR loops simultaneously. Table 4 reports per-CDR AAR, RMSD, and EpiF1 on the epitope-group split. Only DiffAb, AbFlowNet, AbMEGD, RADAb, and dyAb produce all six loops on this split, while the remaining methods are limited to the heavy-chain loops. A clear difficulty hierarchy emerges across CDR types, where H3 is hardest (lowest AAR, highest RMSD) and H1 is easiest.

Table 4: Per-CDR results on the **epitope-group** test split. Mean AAR, RMSD, and EpiF1 are shown per CDR type. "–" indicates the method does not support the given CDR. Best values are in **bold**, second-best are underlined.

	AAR↑						RMSD↓						EpiF1↑					
Method	H1	H2	H3	L1	L2	L3	H1	H2	H3	L1	L2	L3	H1	H2	H3	L1	L2	L3
RAAD	**0.72**	**0.67**	0.38	–	–	–	**0.59**	**0.51**	**1.95**	–	–	–	**0.92**	**0.89**	0.67	–	–	–
MEAN	0.70	0.59	**0.42**	–	–	–	0.85	0.83	2.01	–	–	–	0.86	0.85	0.67	–	–	–
dyMEAN	0.67	0.65	0.35	–	–	–	1.07	0.99	2.60	–	–	–	0.80	0.79	0.55	–	–	–
DiffAb	0.57	0.29	0.22	0.55	0.50	**0.43**	0.87	0.67	2.64	1.03	2.03	**0.96**	0.83	0.84	0.56	0.80	**0.93**	0.66
AbFlowNet	0.56	0.31	0.22	0.54	**0.51**	0.40	0.79	0.65	2.70	**0.99**	0.49	0.96	0.84	0.84	0.55	**0.82**	0.92	0.65
AbMEGD	0.56	0.33	0.22	0.52	0.47	0.38	0.85	0.71	2.76	1.00	**0.48**	1.09	0.82	0.83	0.56	0.82	0.92	0.64
RADAb	0.61	0.34	0.22	**0.64**	0.50	0.42	0.95	0.81	12.28	4.50	1.49	1.10	0.83	0.83	0.57	0.79	0.91	0.67
dyAb	0.48	0.32	0.27	0.24	0.46	0.26	1.87	1.82	3.31	1.64	1.11	1.37	0.75	0.72	0.50	0.81	0.89	**0.83**
RefineGNN	0.67	0.47	0.23	–	–	–	0.63	1.74	3.07	–	–	–	0.92	0.84	**0.76**	–	–	–
AbDockGen	–	–	0.23	–	–	–	–	–	3.93	–	–	–	–	–	0.62	–	–	–
AbODE	0.53	0.56	0.31	–	–	–	4.76	6.69	16.40	–	–	–	0.45	0.54	0.20	–	–	–

5.3 Analysis

Sequence recovery and structural quality decouple. Equivariant GNN methods (RAAD, MEAN, dyMEAN) lead sequence recovery on the epitope-group split, with CDR-H3 AAR between 0.35 and 0.42, well above the diffusion and flow-matching methods (DiffAb, AbMEGD, AbFlowNet, dyAb) that cluster between 0.22 and 0.27. Structural accuracy follows a different ordering. RAAD and MEAN also produce the most accurate CDR-H3 backbones (RMSD near 2.0 Å), whereas AbDockGen is the weakest method that still returns well-formed loops (RMSD=3.93 Å). AbODE is a degenerate case, with CDR-H3 RMSD of 16.40 Å and TM-score near zero across all three splits, a failure we trace to its polar-coordinate loop parameterization (§5.4). Notably, AbODE still reaches the highest contact-residue recovery on this split (CAAR=0.22), so its sequence predictions are reasonable even though its coordinates are not, a clear decoupling of sequence and structure.

Interface quality does not track sequence recovery. The interface metrics reorder the methods again. RefineGNN attains the best native-contact recovery and docking quality (Fnat=0.62, DockQ=0.71, EpiF1=0.76 on the epitope-group split) despite modest sequence recovery (AAR=0.23), while RAAD and MEAN trail on Fnat (0.48–0.49) even though they lead on AAR. At the other end, dyAb and AbODE recover few native contacts (Fnat=0.35 and 0.08). Sequence recovery alone is therefore a poor proxy for binding-interface quality, which motivates reporting the two metric groups separately.

Novel epitopes expose a generalization gap. A single random split would hide where these methods break. Sequence recovery is nearly invariant to the split, for example RAAD reaches AAR=0.38 on all three, but interface and epitope metrics degrade sharply on the epitope-group split, which holds out entire epitope clusters. RAAD's Fnat falls from 0.62 on antigen-fold to 0.49 on epitope-group, and its EpiF1 falls from 0.78 to 0.67. The same direction of change holds for nearly every method (Tables 8 and 9), and Table 10 summarizes this epitope-group versus antigen-fold gap side by side. Binding-interface quality, not sequence recovery, is the axis that separates fitting seen epitopes from generalizing to novel ones, which is exactly what the epitope-group split and the epitope-specificity metrics are designed to expose.

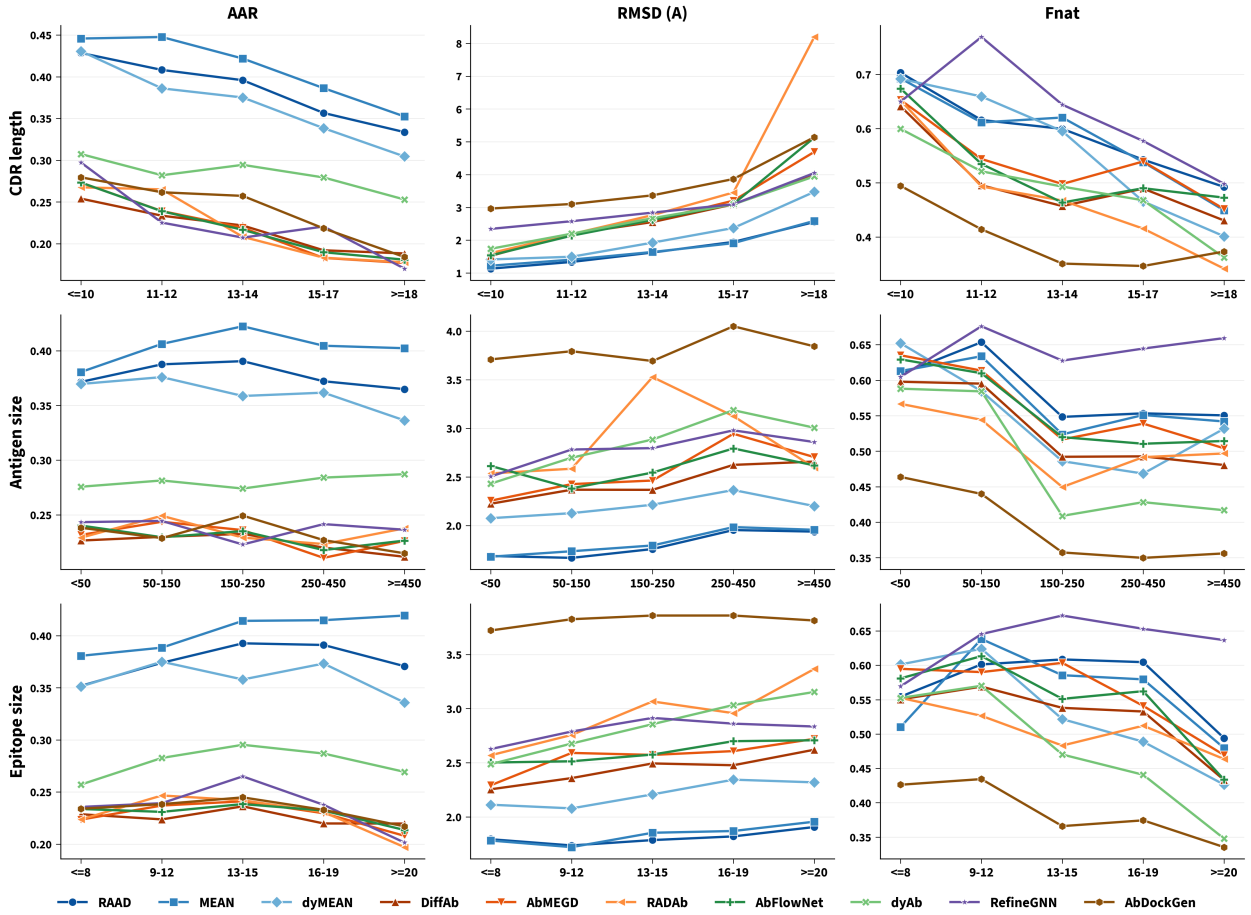

Figure 7: CDR-H3 design quality on the epitope-group split, stratified by three covariates (rows: CDR-H3 length, antigen size, epitope size) across three metrics (columns: sequence recovery AAR, backbone accuracy RMSD, native-contact recovery Fnat). Loop length is the dominant difficulty axis, with all three metrics degrading as the loop lengthens, while antigen and epitope size leave AAR nearly flat.

Difficulty scales with CDR-H3 length. Loop length is the dominant difficulty axis. The top row of Figure 7 stratifies the epitope-group test set by CDR-H3 length and shows that sequence recovery, backbone accuracy, and native-contact recovery all degrade as the loop grows. The methods separate by how much. The one-shot denoising and equivariant-translation methods lose the most native-contact recovery from short to long loops, while RefineGNN degrades least and leads the long-loop regime despite its modest sequence recovery. Its autoregressive, structure-conditioned formulation appears to preserve binding contacts as the loop extends. By contrast, neither antigen size nor epitope size moves AAR appreciably (middle and bottom rows of Figure 7), so long CDR-H3 loops are the regime where new methods stand to gain the most.

Sequence recovery is insensitive to antigen and epitope properties. We further stratify the epitope-group CDR-H3 results by antigen size, epitope size, and antigen type. Sequence recovery (AAR) is insensitive to all three. CDR-H3 AAR is nearly flat across antigen sizes and epitope sizes (middle and bottom rows of Figure 7) and varies by at most 0.03 between protein and peptide antigens. The interface metrics are the ones that move. Native-contact recovery is non-monotonic in epitope size, peaking on medium epitopes of 7 to 15 residues and dropping on both smaller epitopes, which offer too few contacts to recover, and larger epitopes, which offer too many to place correctly (bottom row of Figure 7). Peptide antigens are systematically easier on the interface, with most methods gaining 0.03 to 0.15 Fnat over protein antigens, so the protein-antigen

subset is the harder and more informative slice. As in the length analysis, RefineGNN is the most robust to these covariates on the interface metrics.

Capacity and compute do not predict quality. Larger models and longer training do not yield better designs on Chimera-Bench. The smallest trainable model, MEAN at 0.7M parameters, matches or beats every diffusion model on CDR-H3 across all three splits, including RADAb, which adds a frozen 650M-parameter ESM-2 encoder and a retrieval stage for no measurable gain over plain DiffAb. The gap is more prominent in cost-adjusted terms. MEAN and RAAD train in under 4 GPU-hours in total, and yet match or beat the diffusion models that each need 6 to 11 hours (Table 5). Inference cost spans three orders of magnitude, from under a minute per split for the lightest GNNs to roughly two to three hours per split for RADAb, whose MSA construction and retrieval stage dominate its runtime. Our analysis shows that architecture and inductive bias drive performance in this benchmark more than the model capacity.

Table 5: Computational resources for all evaluated methods. Params are total and trainable parameter counts. Train is total wall-clock training time in GPU-hours summed over all CDR models a method trains, and Infer is total test-set inference time in minutes, both reported per split (epitope group: EG, antigen fold: AF, temporal: TP).

		Params		Train (GPU-h)			Infer (min)		
Method	Type	Total	Train.	EG	AF	TP	EG	AF	TP
MEAN	Equiv. GNN	0.7M	0.7M	3.5	2.6	2.4	0.5	0.5	0.6
RAAD	Equiv. GNN	6.8M	6.8M	2.8	3.0	2.1	0.8	0.7	0.5
dyMEAN	Equiv. GNN	2.0M	2.0M	15.3	14.5	16.5	0.7	1.4	0.6
RefineGNN	Autoreg. GNN	5.8M	5.8M	3.5	3.2	3.1	2.8	3.2	3.2
dyAb	Flow matching	2.5M	2.5M	1.1	0.7	0.7	2.2	1.6	1.6
AbDockGen	Hier. ENN	8.8M	8.8M	2.3	1.6	1.6	2.9	1.7	2.0
AbODE	Neural ODE	0.24M	0.24M	13.9	7.0	9.0	0.9	0.9	0.9
DiffAb	Diffusion	4.0M	4.0M	7.4	6.0	6.6	8.5	7.9	10.0
AbFlowNet	Flow matching	4.0M	4.0M	8.3	4.7	6.1	6.6	10.0	11.0
AbMEGD	Diffusion	5.8M	5.8M	9.8	10.2	11.2	7.8	7.6	7.8
RADAb	Retr.+Diff.	661M	10.1M	6.5	7.4	10.7	119.6	113.2	204.4

No single method dominates. The best method depends on both the split and the metric group. RAAD and MEAN lead sequence recovery and backbone accuracy on every split, while RefineGNN leads the interface and epitope metrics on the held-out epitope-group and temporal splits. No method wins more than four of the eight CDR-H3 metrics on any one split, which is why we report each split and metric group separately rather than collapsing them into a single ranking. The practical consequence is clear. The open problem is generalization to novel epitopes, not raw sequence recovery, and the highest-value regime for new methods is long CDR-H3 loops evaluated with interface metrics, where structure-conditioned autoregressive inductive biases outperform the largest models.

5.4 Qualitative CDR-H3 Structure Comparison

Figure 8 shows the predicted CDR-H3 backbone structures from seven methods overlaid on the native loop for three complexes selected to illustrate different design regimes. ***7JKS*** (H3 length 14) is a typical case where equivariant GNN methods (RAAD, dyMEAN) recover the native sequence well (AAR$\sim$0.36) with moderate RMSD ($\sim$1.6 Å), while diffusion-based methods achieve lower AAR ($\sim$0.17–0.33) with slightly higher RMSD ($\sim$2.1–2.7 Å). ***4J4P*** (H3 length 17) represents a case where equivariant GNNs and diffusion methods disagree. dyMEAN achieves the highest AAR (0.71) while diffusion methods recover substantially fewer residues (AAR$\sim$0.13–0.27), illustrating how longer loops amplify both sequence and structural prediction errors. ***4XMP*** (H3 length 25) is a challenging long-loop case where all methods struggle. AAR drops below 0.18 for every method, and the diffusion-based methods produce particularly large structural deviations on this long loop. AbODE returns a degenerate global placement here, consistent with its catastrophic aggregate RMSD in the main results.

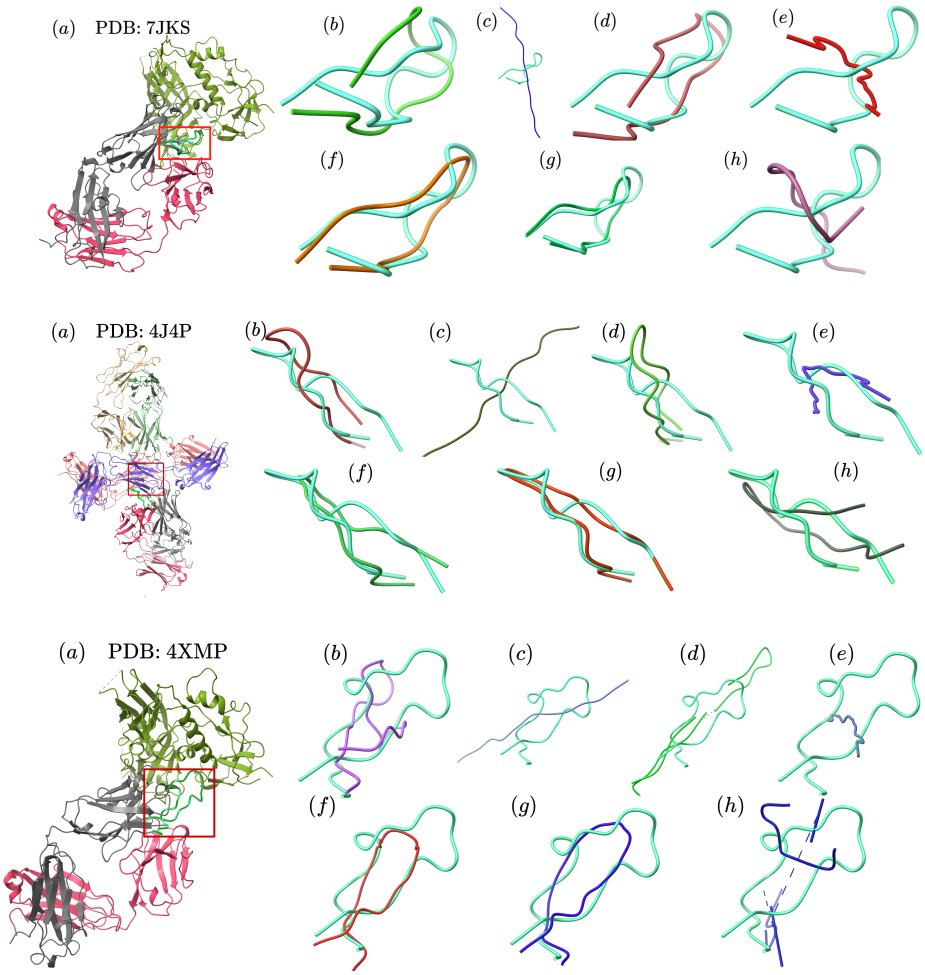

Figure 8: Native (a) and predicted CDR-H3 backbone structures for three complexes of increasing difficulty: 7JKS (H3 length 14), 4J4P (H3 length 17), and 4XMP (H3 length 25). Panels show predictions from (b) AbFlowNet, (c) AbODE, (d) DiffAb, (e) dyAb, (f) dyMEAN, (g) RAAD, (h) RefineGNN. The native CDR H3 regions are shown in cyan colors.

5.5 Limitations

Chimera-Bench evaluates designs against static crystallographic structures and does not model conformational flexibility, solvent effects, or experimental binding affinity, so high benchmark scores are necessary but not sufficient evidence of a viable binder, and the dataset inherits the biases of SAbDab and the PDB toward viral surface antigens, human-derived antibodies, and X-ray crystallography (66% of complexes). The coverage is also incomplete, as several methods could not be evaluated under leakage control, including RFAntibody (Bennett et al., 2025), whose weights are trained on the full PDB without a disclosed split, and ProteinMPNN (Dauparas et al., 2022), whose inverse-folding benchmarking could be the future work.

6 Conclusion

We have presented Chimera-Bench, a unified benchmark for epitope-specific antibody CDR design. Chimera-Bench provides a canonical task formulation, a curated dataset of 2,922 complexes with biologically motivated splits, and a standardized evaluation protocol, including novel epitope-specificity measures. Our evaluation of eleven baseline methods spanning different generative paradigms establishes comparisons under consistent conditions and reveals patterns invisible under a single random split.

References

Abrar Rahman Abir, Haz Sameen Shahgir, Md Rownok Zahan Ratul, Md Toki Tahmid, Greg Ver Steeg, and Yue Dong. AbFlowNet: Optimizing antibody-antigen binding energy via diffusion-GFlowNet fusion. *arXiv preprint arXiv:2505.12358*, 2025. 8, 17, 18, 19, 20, 21

Jared Adolf-Bryfogle, Oleks Kalyuzhniy, Michael Kubitz, Brian D Weitzner, Xiaozhen Hu, Yumiko Adachi, William R Schief, and Roland L Dunbrack Jr. Rosetta Antibody Design (RAbD): A general framework for computational antibody design. *PLoS computational biology*, 14(4):e1006112, 2018. 1, 2, 3

Mansoor Ahmed, Sarwan Ali, Avais Jan, Imdad Ullah Khan, and Murray Patterson. Improved graph-based antibody-aware epitope prediction with protein language model-based embeddings. *bioRxiv*, 2025. doi: 10.1101/2025.02.12.637989. 21

Sankar Basu and Björn Wallner. DockQ: a quality measure for protein-protein docking models. *PloS one*, 11(8):e0161879, 2016. 7, 17

Nathaniel R Bennett, Joseph L Watson, Robert J Ragotte, Andrew J Borst, DéJenaé L See, Connor Weidle, Riti Biswas, Yutong Yu, Ellen L Shrock, Russell Ault, et al. Atomically accurate de novo design of antibodies with RFdiffusion. *Nature*, pp. 1–11, 2025. 2, 12

Jiameng Chen, Xiantao Cai, Jia Wu, and Wenbin Hu. Antibody design and optimization with multi-scale equivariant graph diffusion models for accurate complex antigen binding. *arXiv preprint arXiv:2506.20957*, 2025. 8, 18, 19, 20, 21

Cyrus Chothia and Arthur M Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of molecular biology*, 196(4):901–917, 1987. 3, 18

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. 2, 12, 20

James Dunbar and Charlotte M Deane. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016. 3

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. 1, 2, 3

Joël Janin, Kim Henrick, John Moult, Lynn Ten Eyck, Michael JE Sternberg, Sandor Vajda, Ilya Vakser, and Shoshana J Wodak. CAPRI: a critical assessment of predicted interactions. *Proteins: Structure, Function, and Bioinformatics*, 52(1):2–9, 2003. 17

Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019. 2, 3

Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi Jaakkola. Iterative refinement graph neural network for antibody sequence-structure co-design. *arXiv preprint arXiv:2110.04624*, 2021. 1, 2, 8, 17, 18, 19, 20

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Antibody-antigen docking and design via hierarchical structure refinement. In *International Conference on Machine Learning*, pp. 10217–10227. PMLR, 2022. 2, 8, 18, 19, 20, 21

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3D equivariant graph translation. *arXiv preprint arXiv:2208.06073*, 2022. 1, 2, 8, 17, 18, 19, 20

Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-end full-atom antibody design. *arXiv preprint arXiv:2302.00203*, 2023. 1, 2, 8, 17, 18, 19, 20

Marie-Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental & Comparative Immunology*, 27(1):55–77, 2003. 3, 18

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. 21

Chunan Liu, Lilian Denzler, Yihong Chen, Andrew Martin, and Brooks Paige. AsEP: Benchmarking deep learning methods for antibody-specific epitope prediction. *arXiv preprint arXiv:2407.18184*, 2024. 3

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022. 1, 2, 7, 8, 17, 18, 19, 20

Karolis Martinkus, Jan Ludwiczak, Wei-Ching Liang, Julien Lafrance-Vanasse, Isidro Hotzel, Arvind Rajpal, Yan Wu, Kyunghyun Cho, Richard Bonneau, Vladimir Gligorijevic, et al. AbDiffuser: full-atom generation of in-vitro functioning antibodies. *Advances in Neural Information Processing Systems*, 36:40729–40759, 2023. 1

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

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023. 2, 3

Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *elife*, 3:e02030, 2014. 17

Thiru Ramaraj, Thomas Angel, Edward A Dratz, Algirdas J Jesaitis, and Suvobrata Bhattacharyya. Antigen–antibody interface properties: Composition, residue interactions, and features of the antigen. *Biochimica et Biophysica Acta (BBA) – Proteins and Proteomics*, 1824(3):520–532, 2012. 4

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. 21

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, et al. BoltzGen: Toward universal binder design. *bioRxiv*, pp. 2025–11, 2025. 2

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. 3, 16

Cheng Tan, Yijie Zhang, Zhangyang Gao, Yufei Huang, Haitao Lin, Lirong Wu, Fandi Wu, Mathieu Blanchette, and Stan Z Li. dyAb: Flow matching for flexible antibody design with AlphaFold-driven pre-binding antigen. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 782–790, 2025. 1, 2, 8, 17, 18, 19, 20, 21

Raphael JL Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon Anderson, Stephan Eismann, et al. ATOM3D: Tasks on molecules in three dimensions. *arXiv preprint arXiv:2012.04035*, 2020. 2

Talip Uçar, Cedric Malherbe, and Ferran Gonzalez. Exploring log-likelihood scores for ranking antibody sequence designs. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024. 20

Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *Biorxiv*, pp. 2022–02, 2022. 6

Yogesh Verma, Markus Heinonen, and Vikas Garg. AbODE: Ab Initio antibody design using conjoined ODEs. In *International Conference on Machine Learning*, pp. 35037–35050. PMLR, 2023. 2, 8, 18, 19, 20, 21

Rubo Wang, Fandi Wu, Jiale Shi, Yidong Song, Yu Kong, Jian Ma, Bing He, Qihong Yan, Tianlei Ying, Peilin Zhao, et al. A generative foundation model for antibody design. *bioRxiv*, pp. 2025–09, 2025. 1, 20

Zichen Wang, Yaokun Ji, Jianing Tian, and Shuangjia Zheng. Retrieval augmented diffusion model for structure-informed antibody design and optimization. *arXiv preprint arXiv:2410.15040*, 2024. 8, 17, 18, 19, 20, 21

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. 1, 2, 8, 17, 18, 19, 20, 21

Xuezhi Xie, Jin Sub Lee, Dongki Kim, Jaehyeong Jo, Jisun Kim, and Philip M Kim. Antibody-SGM: Antigen-specific joint design of antibody sequence and structure using diffusion models. In *2023 ICML Workshop Comput Biol*, 2023. 20

Li C Xue, Drena Dobbs, Alexandre MJJ Bonvin, and Vasant Honavar. Computational prediction of protein interfaces: A review of data driven methods. *FEBS letters*, 589(23):3516–3526, 2015. 17

Yinghua Yao, Yuangang Pan, and Xixian Chen. Generative co-design of antibody sequences and structures via black-box guidance in a shared latent space. *arXiv preprint arXiv:2508.11424*, 2025. 2, 20

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. 7, 17

Xinyan Zhao, Yi-Ching Tang, Akshita Singh, Victor J Cantu, KwanHo An, Junseok Lee, Adam E Stogsdill, Ibraheem M Hamdi, Ashwin Kumar Ramesh, Zhiqiang An, et al. AbBiBench: A benchmark for antibody binding affinity maturation and design. *arXiv preprint arXiv:2506.04235*, 2025. 2, 3

Tian Zhu, Milong Ren, and Haicang Zhang. Antibody design using a score-based diffusion model guided by evolutionary, physical and geometric constraints. In *Forty-first International Conference on Machine Learning*, 2024. 2, 20

Yibo Zhu, Xiumin Shi, Jingjuan Zhang, Weizhong Sun, and Lu Wang. AbEgDiffuser: Antibody sequence-structure codesign with equivariant graph neural networks and diffusion models. *Journal of Chemical Theory and Computation*, 21(21):11307–11317, 2025. 20

A Appendix

A.1 Processing Details

The six sequential quality filters applied during dataset construction are: (1) restriction to protein and peptide antigens, excluding carbohydrates, nucleic acids, and haptens, (2) requirement for both heavy (VH) and light (VL) chains, (3) requirement for an annotated antigen chain, (4) a crystallographic resolution cutoff of 4.0 Å, (5) verification that the structure file was successfully downloaded, and (6) removal of duplicate rows sharing the same PDB, heavy chain, light chain, and antigen chain identifiers. Sequence deduplication concatenates VH and VL sequences for each complex and clusters at 95% identity with 80% coverage using MMseqs2 (Steinegger & Söding, 2017), retaining the highest-resolution representative per cluster, yielding 2,981 deduplicated complexes.

The validation step verifies baseline compatibility by checking four criteria: successful ANARCI numbering for both VH and VL chains, presence of conserved residues at IMGT positions 23 (Cys), 41 (Trp), and 104 (Cys), identifiability of all six CDR regions under IMGT numbering, and no missing Cα atoms in any chain. This excludes 59 complexes, of which 27 have missing backbone atoms (mostly peptide antigen termini), 12 contain atypical V-genes that ANARCI cannot number, 11 have engineered conserved-residue substitutions, and 9 have incomplete CDR or conserved-position annotations (including 3 lambda chains lacking CDR-L2 and 6 with missing IMGT anchor positions or CDR regions). The excluded complexes and their reasons are recorded for transparency.

A.2 Metric Definitions

We provide formal definitions for all metrics used in Chimera-Bench. Let $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_N)$ and $s = (s_1, \ldots, s_N)$ denote the predicted and native CDR sequences of length N, $\hat{\mathbf{X}}, \mathbf{X} \in \mathbb{R}^{N\times 3}$ the predicted and native Cα coordinates, $m \in \{0,1\}^N$ the paratope contact mask, and $\mathcal{C}, \hat{\mathcal{C}}$ the native and predicted sets of antibody–antigen contact pairs. We group these metrics into five categories as follows:

A.2.1 Sequence quality.

These metrics measure how well a design recovers the native CDR sequence.

Amino Acid Recovery (AAR) is the fraction of CDR positions whose predicted residue matches the native residue. It ranges from 0 to 1, and higher is better.

$$\mathrm{AAR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\hat{s}_i = s_i] \tag{2}$$

Contact Amino Acid Recovery (CAAR) restricts AAR to the paratope contact positions, isolating recovery at the binding-relevant residues that drive affinity. It ranges from 0 to 1, and higher is better.

$$\mathrm{CAAR} = \frac{\sum_{i=1}^{N} m_i \cdot \mathbb{1}[\hat{s}_i = s_i]}{\sum_{i=1}^{N} m_i} \tag{3}$$

Perplexity (PPL) is the exponentiated per-residue negative log-likelihood the model assigns to the native sequence, measuring how confidently and fluently a model reproduces it. Lower is better. It is defined only for likelihood-based models and is reported as not applicable for diffusion and flow models.

$$\mathrm{PPL} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(s_i \mid \text{context})\right) \tag{4}$$

A.2.2 Structural quality.

These metrics measure how well a design reproduces the native CDR backbone geometry.

Root Mean Square Deviation (RMSD) is the average Cα distance between the predicted and native loop after optimal SVD-based Kabsch superposition, reported in Å with lower being better. Here R^* is the optimal rotation from the SVD of $\hat{\mathbf{X}}_c^\top \mathbf{X}_c$ on centered coordinates.

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|R^* \hat{\mathbf{x}}_{c,i} - \mathbf{x}_{c,i}\|^2} \tag{5}$$

Template Modeling score (TM) (Zhang & Skolnick, 2004) is a length-normalized measure of global fold similarity on the Kabsch-aligned structures, less sensitive to local outliers than RMSD. It ranges from 0 to 1, and higher is better. Here d_i is the distance between aligned Cα atoms and L is the target length.

$$\text{TM} = \frac{1}{L}\sum_{i=1}^{L} \frac{1}{1 + (d_i/d_0)^2}, \quad d_0 = \max\bigl(1.24\sqrt[3]{L-15} - 1.8,\ 0.5\bigr) \tag{6}$$

A.2.3 Binding interface quality.

These metrics measure how well a design reproduces the native antibody–antigen interface. Two residues in different chains interact if their C_α atoms are within 8 Å (Xue et al., 2015; Ovchinnikov et al., 2014). We follow this definition as a proxy for docking evaluation by CAPRI (Critical Assessment of Predicted Interactions) Janin et al. (2003) since some baseline methods output side chains (Luo et al., 2022; Abir et al., 2025; Wu et al., 2025; Wang et al., 2024) while some other methods output only the backbone atoms (Kong et al., 2023; Tan et al., 2025; Jin et al., 2021; Kong et al., 2022).

Fraction of Native Contacts (Fnat) is the proportion of native antibody–antigen contact pairs that the design recovers. It ranges from 0 to 1, and higher is better.

$$\text{Fnat} = \frac{|\hat{\mathcal{C}} \cap \mathcal{C}|}{|\mathcal{C}|} \tag{7}$$

Interface RMSD (iRMSD) is the Cα RMSD computed over the interface residues alone, in Å with lower being better. $\mathbf{X}_{\text{iface}}$ denotes the Cα coordinates of residues participating in contacts.

$$\text{iRMSD} = \text{RMSD}\bigl(\hat{\mathbf{X}}_{\text{iface}},\ \mathbf{X}_{\text{iface}}\bigr) \tag{8}$$

Docking Quality (DockQ) (Basu & Wallner, 2016) is a single CAPRI-style score that combines Fnat, iRMSD, and the Kabsch-aligned full-antibody Cα RMSD (LRMSD). It ranges from 0 to 1, and higher is better. Both native and predicted contact sets are computed symmetrically from Cα–Cα distances at the 8 Å cutoff.

$$\text{DockQ} = \frac{1}{3}\left(\text{Fnat} + \frac{1}{1 + (\text{iRMSD}/1.5)^2} + \frac{1}{1 + (\text{LRMSD}/8.5)^2}\right) \tag{9}$$

A.2.4 Epitope specificity.

These metrics measure whether the design contacts the intended binding site. Let $\hat{E}$ be the set of antigen residues contacted by the designed antibody (within 8 Å Cα–Cα distance) and E the true epitope residues.

Epitope Precision (EpiPrec) is the fraction of contacted antigen residues that belong to the true epitope, and **Epitope Recall (EpiRec)** is the fraction of true epitope residues that the design contacts. **Epitope F1 (EpiF1)** is their harmonic mean and is the headline epitope metric. All three range from 0 to 1, and higher is better.

$$\text{EpiPrec} = \frac{|\hat{E} \cap E|}{|\hat{E}|}, \quad \text{EpiRec} = \frac{|\hat{E} \cap E|}{|E|}, \quad \text{EpiF1} = \frac{2 \cdot \text{EpiPrec} \cdot \text{EpiRec}}{\text{EpiPrec} + \text{EpiRec}} \tag{10}$$

A.2.5 Designability.

Sequence liability count (n_liab.) is the number of known developability liability motifs (NG, DG, DS, DD, NS, NT, M) present in the designed CDR sequence. It is a non-negative integer, and lower is better.

A.3 Pipeline Configuration

Table 6 lists the values of all thresholds and processing parameters used throughout this work. Table 7 provides the CDR boundary definitions under both numbering schemes.

Table 6: Pipeline configuration parameters.

Parameter	Description	Value
Max resolution	Structure quality threshold	4.0 Å
Contact cutoff	Epitope/paratope definition	4.5 Å
Sequence identity	MMseqs2 clustering threshold	95%
Coverage	MMseqs2 alignment coverage	80%
Numbering schemes	Residue numbering	IMGT, Chothia
Graph k-NN	Nearest neighbors for intra-chain edges	10
Graph spatial cutoff	Intra-chain spatial edge threshold	8.0 Å
Inter-chain cutoff	Inter-chain spatial edge threshold	12.0 Å
Train/Val/Test ratio	Split proportions	80/10/10
Temporal split	Date-based quantile cutoffs	80/10/10

Table 7: CDR boundary definitions used in CHIMERA-BENCH.

Scheme	H1	H2	H3	L1	L2	L3
IMGT (Lefranc et al., 2003)	27–38	56–65	105–117	27–38	56–65	105–117
Chothia (Chothia & Lesk, 1987)	26–32	52–56	95–102	24–34	50–56	89–97

A.4 Additional Results

Tables 8 and 9 report the full CDR-H3 results on the antigen-fold and temporal test splits, complementing the epitope-group results in Table 3. Figure 9 compares the methods across all evaluation metrics for each heavy-chain CDR.

Table 8: CDR-H3 design results on the **antigen-fold** test split. Best values are in **bold**, second-best are underlined.

Method	AAR↑	CAAR↑	PPL↓	RMSD↓	TM↑	Fnat↑	iRMSD↓	DockQ↑	EpiF1↑	n_liab.↓
RAAD (Wu et al., 2025)	0.38±0.13	**0.20±0.21**	3.06±0.49	**1.65±0.70**	**0.24±0.13**	**0.62±0.31**	1.42±0.62	**0.70±0.21**	0.78±0.25	0.59±0.60
MEAN (Kong et al., 2022)	**0.40±0.14**	0.20±0.23	**2.79±0.57**	1.66±0.64	0.23±0.11	0.62±0.32	1.42±0.61	0.70±0.21	0.78±0.26	0.38±0.70
dyMEAN (Kong et al., 2023)	0.37±0.13	0.20±0.23	3.35±0.66	2.06±0.83	0.16±0.09	0.61±0.31	1.79±0.82	0.66±0.20	0.75±0.25	0.21±0.41
DiffAb (Luo et al., 2022)	0.23±0.12	0.10±0.15	–	2.24±1.02	0.15±0.13	0.52±0.34	2.10±0.89	0.58±0.25	0.65±0.30	0.70±0.77
AbFlowNet (Abir et al., 2025)	0.23±0.12	0.13±0.18	–	2.70±5.36	0.16±0.14	0.56±0.34	2.46±4.31	0.60±0.26	0.65±0.30	0.61±0.81
AbMEGD (Chen et al., 2025)	0.24±0.13	0.13±0.20	–	2.25±1.06	0.16±0.15	0.55±0.34	2.13±0.94	0.59±0.25	0.65±0.30	0.64±0.76
RADAb (Wang et al., 2024)	0.25±0.13	0.12±0.17	–	7.79±41.92	0.14±0.13	0.51±0.33	4.78±22.78	0.57±0.26	0.65±0.31	0.57±0.79
dyAb (Tan et al., 2025)	0.28±0.11	0.17±0.21	–	2.40±0.81	0.10±0.05	0.56±0.33	1.81±0.81	0.65±0.21	0.72±0.29	0.92±0.27
RefineGNN (Jin et al., 2021)	0.26±0.13	0.12±0.16	8.17±3.99	2.59±0.65	0.10±0.06	0.60±0.32	**1.35±0.54**	0.68±0.22	**0.80±0.26**	0.59±0.90
AbDockGen (Jin et al., 2022)	0.24±0.13	0.10±0.16	7.89±3.45	3.65±0.81	0.05±0.03	0.46±0.27	2.41±0.84	0.57±0.17	0.75±0.23	0.59±0.68
AbODE (Verma et al., 2023)	0.27±0.13	0.18±0.20	10.78±3.13	13.17±4.81	0.01±0.01	0.16±0.22	4.00±1.79	0.39±0.16	0.40±0.36	**0.20±0.51**

A.5 Cross-Method Metric Comparison

Table 11 summarizes which metrics each of the surveyed methods originally reports, illustrating the inconsistency that CHIMERA-BENCH standardizes. Abbreviations: DP = dynamic-programming RMSD (no Kabsch), K = Kabsch-aligned RMSD, sc = self-consistency RMSD (re-fold then compare), SeqID = BLOSUM62 alignment identity, PLL = pseudo-log-likelihood, LL/NLL = model (negative) log-likelihood.

Table 9: CDR-H3 design results on the **temporal** test split. Best values are in **bold**, second-best are underlined.

Method	AAR↑	CAAR↑	PPL↓	RMSD↓	TM↑	Fnat↑	iRMSD↓	DockQ↑	EpiF1↑	n_liab.↓
RAAD (Wu et al., 2025)	0.38±0.13	0.20±0.22	**3.24±0.44**	**1.81±0.76**	**0.22±0.12**	0.61±0.28	1.46±0.70	0.71±0.16	0.74±0.23	0.77±0.65
MEAN (Kong et al., 2022)	**0.39±0.13**	**0.22±0.23**	3.31±0.53	1.84±0.68	0.20±0.09	0.60±0.29	1.44±0.66	0.71±0.16	0.73±0.23	0.66±0.66
dyMEAN (Kong et al., 2023)	0.35±0.13	0.20±0.22	3.28±0.42	2.32±0.97	0.15±0.10	0.50±0.30	2.00±0.92	0.63±0.17	0.64±0.26	0.99±0.71
DiffAb (Luo et al., 2022)	0.23±0.11	0.12±0.17	–	2.46±1.05	0.13±0.10	0.51±0.30	2.18±0.95	0.61±0.19	0.61±0.26	0.74±0.84
AbFlowNet (Abir et al., 2025)	0.23±0.12	0.13±0.19	–	2.67±2.79	0.13±0.11	0.52±0.32	2.33±1.98	0.61±0.20	0.61±0.26	0.45±0.70
AbMEGD (Chen et al., 2025)	0.23±0.11	0.12±0.17	–	2.72±2.42	0.12±0.10	0.55±0.31	2.35±1.67	0.62±0.20	0.58±0.26	0.63±0.83
RADAb (Wang et al., 2024)	0.22±0.12	0.10±0.16	–	11.57±49.30	0.10±0.09	0.48±0.33	7.91±34.02	0.57±0.23	0.57±0.28	0.41±0.59
dyAb (Tan et al., 2025)	0.29±0.11	0.15±0.20	–	2.89±0.88	0.07±0.03	0.51±0.29	1.92±0.82	0.64±0.16	0.65±0.28	**0.00±0.00**
RefineGNN (Jin et al., 2021)	0.23±0.14	0.13±0.18	7.91±3.20	2.85±0.88	0.09±0.05	**0.65±0.27**	**1.38±0.77**	**0.73±0.16**	**0.78±0.20**	0.64±0.79
AbDockGen (Jin et al., 2022)	0.23±0.12	0.09±0.15	7.55±3.06	3.89±0.96	0.05±0.02	0.37±0.25	2.52±1.00	0.55±0.14	0.68±0.22	0.71±0.75
AbODE (Verma et al., 2023)	0.37±0.13	0.19±0.22	9.26±4.09	15.63±4.55	0.00±0.00	0.07±0.12	4.65±2.22	0.35±0.10	0.19±0.22	0.83±0.39

Table 10: Generalization gap on CDR-H3. For each method, we report the epitope-group (EG, unseen epitope clusters) and antigen-fold (AF, unseen antigen folds) test splits. Sequence recovery (AAR) is nearly identical across the two splits, while the interface and epitope metrics (Fnat, DockQ, EpiF1) are consistently lower on the harder epitope-group split.

	AAR↑		Fnat↑		DockQ↑		EpiF1↑	
Method	EG	AF	EG	AF	EG	AF	EG	AF
RAAD	0.38	0.38	0.49	0.62	0.64	0.70	0.67	0.78
MEAN	0.42	0.40	0.48	0.62	0.63	0.70	0.67	0.78
DiffAb	0.22	0.23	0.48	0.52	0.58	0.58	0.56	0.65
AbFlowNet	0.22	0.23	0.49	0.56	0.57	0.60	0.55	0.65
AbMEGD	0.22	0.24	0.50	0.55	0.57	0.59	0.56	0.65
RADAb	0.22	0.25	0.47	0.51	0.57	0.57	0.57	0.65
dyAb	0.27	0.28	0.35	0.56	0.54	0.65	0.50	0.72
RefineGNN	0.23	0.26	0.62	0.60	0.71	0.68	0.76	0.80
AbDockGen	0.23	0.24	0.34	0.46	0.52	0.57	0.62	0.75
AbODE	0.31	0.27	0.08	0.16	0.34	0.39	0.20	0.40

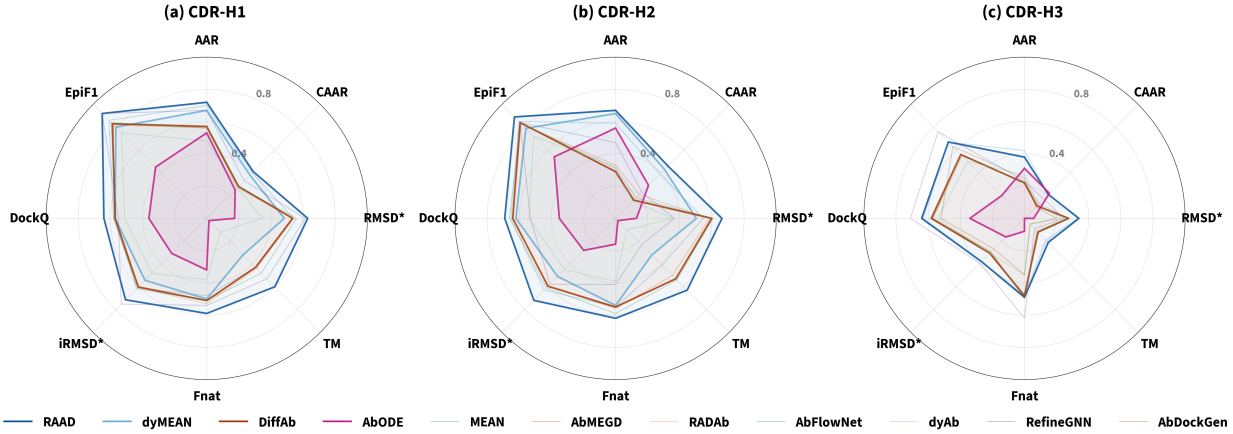

Figure 9: Per-CDR comparison of methods across all evaluation metrics for (a) CDR-H1, (b) CDR-H2, and (c) CDR-H3. Each axis is scaled so that a larger radius is better; RMSD and iRMSD (marked with an asterisk) are inverted accordingly.

A.6 Baseline Retraining on Chimera-Bench

All baselines are retrained on Chimera-Bench using the authors' released code with default hyperparameters. We provide a shared training and evaluation framework that handles data loading in each method's native input format, split-aware partitioning, early stopping, model checkpointing, and performs full evaluation. At test time, per-complex predictions are evaluated against native structures using all metrics defined

Table 11: Metrics originally reported by each method (prior to CHIMERA-BENCH standardization).

Method	AAR	RMSD	TM	lDDT	DockQ	PPL	dG/IMP	CAAR
MEAN (Kong et al., 2022)	AAR	K	✓	–	–	✓	ddG	–
dyMEAN (Kong et al., 2023)	AAR	K	✓	✓	✓	–	ddG	✓
RAAD (Wu et al., 2025)	AAR	K	✓	–	–	✓	–	–
RefineGNN (Jin et al., 2021)	AAR	K	–	–	–	✓	–	–
AbODE (Verma et al., 2023)	AAR	K	–	–	–	✓	–	–
AbDockGen (Jin et al., 2022)	AAR	K	–	–	✓	✓	–	–
DiffAb (Luo et al., 2022)	SeqID	DP	–	–	–	–	IMP,dG	–
dyAb (Tan et al., 2025)	AAR	K	✓	✓	✓	–	ddG	✓
RADAb (Wang et al., 2024)	AAR	sc	–	–	–	PLL	ddG	–
AbFlowNet (Abir et al., 2025)	SeqID	DP	–	–	–	–	IMP,dG	–
AbX (Zhu et al., 2024)	AAR	K	–	–	–	PLL	IMP	–
AbEgDiffuser (Zhu et al., 2025)	AAR	DP	–	–	–	–	IMP	–
AbMEGD (Chen et al., 2025)	SeqID	DP	–	–	–	–	IMP	–
AbSGM (Xie et al., 2023)	–	K	–	–	–	–	–	–
LEAD (Yao et al., 2025)	AAR	DP	–	–	–	–	ddG	–
ProteinMPNN (Dauparas et al., 2022)	AAR	–	–	–	–	NLL	–	–
IgGM (Wang et al., 2025)	AAR	K	✓	✓	✓	–	–	–
DiffAbXL (Uçar et al., 2024)	–	–	–	–	–	LL	–	–

in Section 4.2. All training and inference is performed on an NVIDIA H100 GPU. Below we describe the method-specific adaptations required for each baseline.

DiffAb. DiffAb (Luo et al., 2022) operates under the Chothia numbering scheme and generates all six CDRs simultaneously by randomly masking between one and six CDRs during training. At inference, all CDRs are masked and regenerated in a single forward pass, producing one set of predictions per complex from which per-CDR metrics are extracted. We train with Adam at a learning rate of 10^{-4} and a batch size of 16, using 100 diffusion steps with equal loss weights on rotation, position, and sequence terms.

RAAD. RAAD (Wu et al., 2025) generates one CDR at a time, requiring three training runs per split. We use the Adam optimizer with a learning rate of 10^{-3} and a batch size of 8. The model employs four relation-aware equivariant message-passing layers and constructs multi-relational graphs with radial, k-nearest neighbor, sequential, and global edge types. The structure loss weight is set to α=0.8, placing greater emphasis on coordinate accuracy compared to MEAN.

MEAN. Because MEAN (Kong et al., 2022) generates a single CDR loop per forward pass, we train three separate models per split to cover H1, H2, and H3. Each model is optimized with Adam at a learning rate of 10^{-3} and a batch size of 16. The structure loss is weighted by α=0.05 relative to the sequence loss, and the model performs five multi-channel equivariant attention iterations per forward pass to progressively refine the generated CDR.

dyMEAN. dyMEAN (Kong et al., 2023) supports flexible CDR selection and performs three iterative refinement rounds per forward pass. The learning rate decays exponentially from 10^{-3} to 10^{-4} with a batch size of 4. A cosine annealing schedule linearly decreases the fraction of native template information provided to the model, gradually forcing reliance on its own predictions. The paratope is defined using a 6.6 Å contact threshold under IMGT numbering.

RefineGNN. RefineGNN (Jin et al., 2021) generates CDR residues autoregressively from left to right while iteratively refining the predicted backbone structure after each residue is placed. Like MEAN and RAAD, it generates one CDR per forward pass, so we train three separate models per split for H1, H2, and H3. The model uses four message-passing layers with a hidden dimension of 256 and constructs k-nearest-neighbor graphs with k=9. A bidirectional GRU encodes framework context, and the structural refinement step runs

at every position (update frequency 1), predicting pairwise distances that are converted to coordinates via eigendecomposition. During training, teacher forcing provides the ground-truth sequence at each step, while at inference, the model samples from the predicted distribution autoregressively.

AbDockGen. AbDockGen (Jin et al., 2022) is a hierarchical equivariant refinement network that generates CDR-H3 conditioned on the epitope surface, making it applicable only to **Track 1** evaluation. The model encodes the top 20 nearest epitope residues using an E(3)-equivariant graph neural network and generates the CDR-H3 sequence and structure autoregressively with iterative coordinate refinement. The architecture consists of four hierarchical EGNN layers with a hidden dimension of 256 and k=9 nearest neighbors, operating at both atom and residue levels. Explicit clash avoidance constraints enforce minimum distances of 3.8 Å between backbone atoms and 1.5 Å between side-chain atoms during coordinate updates. We train with Adam at a learning rate of 10^{-3}, dynamic batching at 100 tokens, and gradient clipping at norm 1.0, for at most 10 epochs with a learning rate anneal factor of 0.9.

AbODE. AbODE (Verma et al., 2023) formulates CDR generation as a conjoined ordinary differential equation that jointly evolves sequence logits and backbone coordinates. Like MEAN and RAAD, it generates one CDR at a time and requires three training runs per split. The model consists of four TransformerConv layers and we train with Adam at a learning rate of 10^{-3} and a batch size of 1 due to variable-size fully-connected graphs.

AbFlowNet. AbFlowNet (Abir et al., 2025) extends DiffAb's diffusion architecture with a GFlowNet trajectory balance objective to optimize binding energy. It uses Chothia numbering and generates all six CDRs simultaneously. The training-time backward loss is disabled for CHIMERA integration. We train with Adam at a learning rate of 10^{-4} and a batch size of 16.

AbMEGD. AbMEGD (Chen et al., 2025) is a multi-CDR diffusion model that shares DiffAb's Chothia-based preprocessing pipeline and generates all six CDRs simultaneously. We train with the authors' default hyperparameters using Adam optimization.

RADAb. RADAb (Wang et al., 2024) is a retrieval-augmented diffusion model that combines DiffAb's diffusion framework with sequence retrieval from a FASTA database. It uses ESM-2 (650M parameters) for sequence embeddings and an MSA transformer for retrieved sequence alignment. We train under Chothia numbering with Adam optimization.

dyAb. dyAb (Tan et al., 2025) applies flow matching to flexible antibody design under IMGT numbering, generating all six CDRs simultaneously. The model includes OpenMM-based structure relaxation and a DDG prediction module. We train with the authors' default configuration.

A.7 Protein Graph Construction

For each complex, we construct a heterogeneous multi-relational residue graph (Wu et al., 2025) stored as a PyG `HeteroData` object. The graph contains three node types corresponding to heavy chain, light chain, and antigen residues.

Node features (105D). Each residue node carries a 105-dimensional feature vector consisting of a residue type one-hot encoding (20D), sinusoidal positional encoding along the chain (16D), backbone dihedral angle sin/cos encodings (12D for 6 angles), radial basis function (RBF) encodings of Cα-to-backbone-atom distances (48D for 3 atoms with 16 RBF bins each), and a local coordinate frame derived from N, Cα, C atoms via Gram–Schmidt orthogonalization (9D). Each node additionally stores the residue type index, backbone atom coordinates (N, Cα, C, O), CDR region labels under IMGT numbering, a binary flag marking designable CDR positions, and binary epitope or paratope labels. Since pre-trained protein language model (PLM) embeddings are proven to capture meaningful evolutionary features of proteins (Ahmed et al., 2025), we utilize ESM2 (Lin et al., 2023) to compute embeddings for antigens and AntiBERTy (Ruffolo et al., 2023) for antibodies.

Intra-chain edges. Within each chain, four relation types encode complementary structural relationships. Sequential ± 1 edges connect adjacent residues along the backbone, sequential ± 2 edges provide broader backbone context, k-nearest neighbor edges (k=10) capture local 3D proximity regardless of sequence position, and spatial edges connect all residue pairs within 8.0 Å not already covered by the preceding types. Each intra-chain edge carries a 39-dimensional feature vector consisting of edge type one-hot (4D), relative positional encoding (16D), Cα–Cα distance RBF encoding (16D), and normalized direction vector (3D).

Inter-chain edges. Two types of inter-chain edges connect antibody and antigen nodes. Contact edges are derived from the annotated contact pairs (4.5 Å heavy-atom cutoff) and directly encode the binding interface by linking heavy-to-antigen and light-to-antigen residue pairs. Spatial edges use a broader 12 Å Cα–Cα cutoff with 16-bin RBF distance features, providing the spatial context used by MEAN, dyMEAN, dyAb, and RAAD for message passing across chains. Heavy-to-light spatial edges are also included for intra-antibody context.

The resulting graphs are saved as a single dictionary mapping complex identifiers to `HeteroData` objects. These precomputed graphs can be used directly by graph-based methods or converted to homogeneous representations as needed.