MDLAS: MULTI-DIMENSION LEAKAGE-AWARE SPLITS FOR ANTIBODY–ANTIGEN STRUCTURAL BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing antibody–antigen structural benchmarks often treat leakage control primarily as a sequence-level property. We argue that this is insufficient: a split can prevent direct sequence overlap while retaining redundancy through shared antigens, epitopes, or paired antibody sequences, and can still exhibit substantial distributional imbalance. We present MDLAS (Multi-Dimension Leakage-Aware Split), a framework for constructing antibody–antigen structural benchmark splits that jointly controls cross-split redundancy and marginal distributional consistency. MDLAS builds a multi-relation graph over five redundancy signals spanning CDR-H3 identity (exact duplication and sequence clustering), antigen identity (shared antigen chains), epitope identity (shared binding footprint), and paired variable-domain similarity (co-occurring heavy/light chain pairs), assigns connected components indivisibly across train, validation, and test, and then optimizes the remaining feasible assignments across seven metadata axes. On a 15,557-entry protein/peptide-antigen working population, MDLAS produces an 80/10/10 split (12,445/1,556/1,556 entries) at the primary 70% CDR-H3 identity tier; independent auditing confirms zero cross-split overlap for the modeled redundancy relations, while distributional consistency is measured directly from the resulting assignments. This leakage control involves a substantial trade-off: because component assignment is indivisible, an antigen-dominated component containing 8,029 entries (51.6% of the working population) would overwhelm the balance of whichever split received it. Together with low-resolution-only components, it therefore cannot be assigned to the validation or test split. We further evaluate four CDR-H3 design baselines (MEAN, DiffAb, RefineGNN, and AgForce) under MDLAS. The models exhibit distinct performance profiles under MDLAS, with AgForce achieving better mean binding-interface performance than RefineGNN, consistent with their previously reported ordering on CHIMERA-Bench. These results motivate treating leakage structure, distributional composition, and residual within-test redundancy as distinct properties when designing antibody structural benchmarks.

1 INTRODUCTION

Antibody structure prediction, inverse folding, and de novo design methods are trained and evaluated almost exclusively on data drawn from the Structural Antibody Database (SAbDab). Analysis of an independently reconstructed, 20,968-row SAbDab snapshot reveals substantial compositional concentration: the top two heavy-chain germline families account for 66.7% of entries with a resolved germline call, 89.8% of antibodies share an exact CDR-H3 sequence with at least one other entry, and antigen-cluster concentration is substantial (Gini 0.72 across 2,262 antigen clusters). These observations motivate two design principles MDLAS follows directly: splitting by CDR-H3 cluster instead of PDB identifier, and using antigen-cluster-balanced splitting for antigen-disjoint evaluation.

Existing benchmark efforts curate SAbDab into task-specific splits with standardized metrics (Section 3). The most directly comparable prior work, CHIMERA-Bench (Ahmed et al., 2026), curates 2,922 complexes via quality filtering, sequence deduplication, and standardized numbering, then

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

offers three alternative single-axis leakage splits – each controlling one redundancy dimension at a time, without a paired distributional-consistency claim.

MDLAS instead builds one multi-relation graph over five redundancy signals and assigns connected components of that graph as indivisible units, guaranteeing zero cross-split overlap for all five modeled redundancy relations simultaneously while managing redundancy entirely through split assignment – no entry is discarded via similarity-based deduplication. This paper demonstrates MDLAS by constructing and validating its primary (70%-identity-tier) split: the extraction pipeline, the working-set restriction, the five-edge-type leakage graph, and the three-stage split-assignment algorithm used to produce a balanced validation and test set (Figure 1) – the same pipeline reconstructs equally-verified splits at three further identity tiers (Section 5.4).

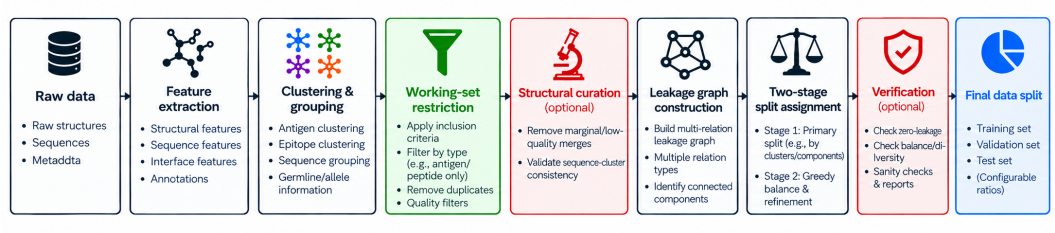

Figure 1: MDLAS pipeline. The pipeline proceeds from source data and feature extraction through clustering, working-set construction, structural curation, leakage-graph construction, and split assignment, followed by independent verification of leakage and distributional balance.

2 A Two-Part Criterion for Benchmark Quality

Existing antibody–antigen structural benchmark efforts, CHIMERA-Bench among them, report quality primarily as one property: leakage was prevented. CHIMERA-Bench offers epitope-disjoint, antigen-fold-disjoint, and temporal splits, each controlling one redundancy dimension at a time, without a paired claim that distribution shift was minimized. Satisfying the first gives no guarantee of the second: a split can be leakage-free on every axis it checks while still being, e.g., 95% one species in validation and 40% in training – exactly the risk SAbDab's own germline, species, and antigen-class concentration (Section 3) creates one level down from leakage itself.

We propose, and build MDLAS around, a two-part criterion: **Criterion 1 (Leakage-freedom).** No value of any defined redundancy signal – exact-duplicate group, sequence-cluster membership, antigen-cluster membership, epitope-cluster membership, or any further signal a given application defines – may appear in more than one of train, validation, or test. **Criterion 2 (Distributional consistency).** The per-axis marginal composition of validation and test, measured across the biologically relevant variables a source characterization study has shown to matter for the dataset in question, should not diverge substantially from the per-axis marginal composition of the full working population. This is a marginal, not joint, consistency claim: satisfying it does not establish that joint dependencies across axes are preserved.

Criterion 1 is a hard constraint: any violation is a leakage bug, not a tradeoff. Criterion 2 is a soft objective, optimized within the space of assignments Criterion 1 permits. The full working population is the descriptive reference for Criterion 2, while Stage A defines the feasible assignment space once the hard leakage and evaluation constraints are imposed – so the optimization seeks the best marginal match attainable without violating those constraints, not an unconstrained match to the global distribution. Every leakage axis CHIMERA-Bench's three splits check is a candidate redundancy signal under Criterion 1; MDLAS's five-edge graph (Section 4.4) checks several jointly and reports Criterion 2 alongside them as an independent, first-class claim, both evaluated separately in Sections 5.1 and 5.2.

3 Related Work

Structural and repertoire databases. SAbDab (Dunbar et al., 2014) aggregates antibody structures from the Protein Data Bank with consistent chain, species, and germline annotation, distributing each structure in pre-numbered Chothia and IMGT forms; MDLAS reads these directly (Section 4.2) instead of re-deriving numbering via ANARCI (Dunbar & Deane, 2016). The Observed Antibody Space (OAS) (Kovaltsuk et al., 2018; Olsen et al., 2022), orders of magnitude larger, is the reference point for whether SAbDab's composition reflects natural repertoire diversity or a biased sample of it. **Antibody-specific benchmarks.** RAbD (Adolf-Bryfogle et al., 2018) curates 60 complexes for CDR-H3 redesign, two orders of magnitude smaller than MDLAS's working population. CHIMERA-Bench (Ahmed et al., 2026), the most directly comparable prior work, curates 2,922 complexes via quality filtering (protein/peptide antigen only, resolution ≤ 4.0 Å, paired VH/VL), 95%-identity deduplication, and ANARCI-validated numbering, then offers epitope-disjoint, antigen-fold-disjoint, and temporal splits. MDLAS adopts its 4.5 Å contact cutoff and protein/peptide-antigen restriction directly, but applies the resolution filter as a confidence gate, not row exclusion and does not require paired VH/VL for inclusion. NbBench (Zhang & Tsuda, 2025) applies MMseqs2-based clustering for leakage-aware nanobody splits, combining leakage control with target-value stratification – the closest existing precedent to MDLAS's two-part criterion, though it checks a single redundancy relation for sequence/function prediction, not structural complexes. Each of these splits is a valid instance of Criterion 1; MDLAS's contribution is jointly enforcing several redundancy relations within one split, reported alongside Criterion 2. **General-purpose leakage-aware splitters.** DataSAIL (Joeres et al., 2025) and GraphPart (Teufel et al., 2023) partition data under similarity constraints while controlling label or property distributions, but reduce redundancy to a single similarity relation. Antibody–antigen redundancy does not reduce that way: a shared antigen chain, a shared epitope footprint, and a near-identical heavy/light pairing are distinct, non-metric relations that no single identity threshold expresses (Table 1). MDLAS unions them into one graph whose components are indivisible assignment units. We have not run a matched-constraint comparison against a domain-agnostic similarity matrix on this population, so the quantitative value of this relation set over DataSAIL/GraphPart remains unestablished. **Leakage-aware evaluation beyond antibodies.** Leakage has been found elsewhere: near-duplicate sequences inflate protein-protein interaction benchmarks (Bushuiev et al., 2024), and Graber et al. (2025) showed protein-ligand binding affinity benchmarks (PDBbind, CASF) suffer structure-based train/test redundancy, introducing a clustering-based resplitting procedure to correct it. Both parallel this paper's concern that redundancy invisible to a single coarse criterion can leak between splits, without an accompanying distributional-consistency claim.

4 Methods

4.1 Data source, schema, and working-set restriction

MDLAS operates on a 20,968-row SAbDab master table built from raw SAbDab structures and sequences. Every downstream annotation this paper reports – antigen classification and clustering, CDR-H3 clustering at all four tiers, epitope clustering, germline allele resolution, and the leakage graph and split themselves – is computed by MDLAS's own pipeline, via full structural re-extraction from SAbDab's raw bundles instead of repackaging the source study's own Cα-only, single-CDR-mask files. Each output combines a schema matching CHIMERA-Bench's fields verbatim with an additive `mdlas` sub-dictionary of split-relevant metadata (full field listing in Appendix A.1). Structural feature extraction (Section 4.2) runs over the full 20,968-row source pool without antigen-type filtering, so the resulting feature store remains reusable for tasks that do not require an antigen; 20,095 (95.8%) are successfully extracted. The benchmark reported in this paper, however, is built over a restricted working population: entries whose antigen is majority protein or peptide by residue composition – matching CHIMERA-Bench's own restriction, since antibody–antigen contact fields are undefined without one. Of the 20,095 successfully extracted entries, 15,745 (78.4%) have an antigen present of any chemistry; of these, 15,576 (98.9%) are majority protein or peptide (Figure 2).

A further 19 entries are excluded for pre-existing SAbDab chain-role-swap contamination (14 entries) and an undetected scFv fusion boundary (5 entries), leaving a final working population of **15,557 entries**, over which all graph and split statistics in this paper are computed. This is a task-

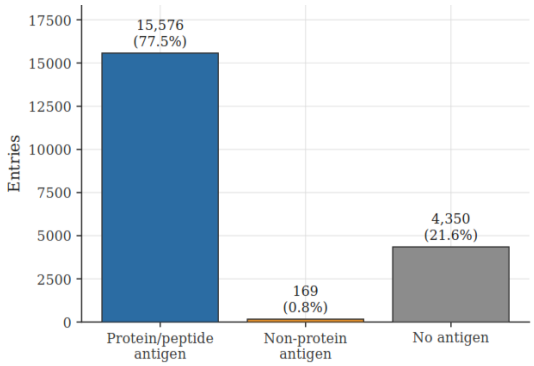

Figure 2: Composition of the full 20,968-row source pool by antigen chemistry before the final known-bad-complex exclusion. The final MDLAS working population contains 15,557 entries.

applicability filter, not deduplication: redundancy is handled entirely at the split-assignment level (Section 4.5), without removing any row.

4.2 Structural feature extraction

Heavy and light chain residues are extracted from SAbDab's pre-numbered IMGT and Chothia bundles, instead of re-deriving numbering via ANARCI, giving full atom14 coordinates, six CDR masks (IMGT boundaries), and a resolution-gated `structural_confidence` flag: entries at or below 4.0 Å resolution (18,684 of 20,968, 93.0%) receive full side-chain-dependent fields (contact pairs, epitope/paratope residues, antigen surface features); coarser entries (1,411, 7.0%) retain only backbone fields. Antigen surface features (128-point farthest-point-sampled physicochemical channels, following CHIMERA-Bench's schema shape) are computed for majority protein/peptide antigens. Full extraction mechanics are in Appendix A.1.

4.3 Clustering: sequence, antigen, epitope, and germline

Sequence, antigen, and epitope. CDR-H3 sequences are clustered via MMseqs2 (80% bidirectional coverage) at four identity tiers, yielding 5,220 clusters at the primary 70% tier (full per-tier counts in Appendix A.4, Table 7); antigen-cluster identifiers (2,262 clusters at 70% identity) are computed directly over MDLAS's own extracted antigen sequences. Two complexes share an epitope cluster only if they share an antigen cluster and their contact residues map to identical positions on that cluster's representative sequence, via a sequence-aligned identifier (full procedure in Appendix A.1); this resolves to 9,951 unique epitope clusters across 14,248 of 14,289 eligible entries (99.7%). **Germline allele.** Allele-level assignment, restricted to human and mouse per available curated coverage, is computed by BLOSUM62-aligning each entry against a fresh OGRDB (Lees et al., 2020) reference set; 14,282 of the 20,968-row source pool (68.1%) receive a chain-level assignment (full distribution in Appendix A.1).

4.4 Multi-relation leakage graph

Five edge types are combined via union–find (Algorithm 1, Appendix A.2) over the 15,557-entry working population: exact CDR-H3 duplication, 70%-tier sequence-cluster membership, antigen-cluster membership, epitope-cluster membership, and paired heavy/light-chain sequence-cluster membership (90% identity, both chains simultaneously). CDR-H3 is the primary signal, since MDLAS targets CDR-H3-conditioned generation; paired heavy/light-chain clustering is the most conservative of several full-variable-domain signals evaluated (Section A.3.1), catching redundancy CDR-H3 clustering misses without over-merging. Edges are applied in a fixed order, and the *marginal* merge count per type is recorded under that order (Table 1) – an order-conditional contribution, not a causal decomposition, since the final partition does not depend on edge order.

Antigen-cluster edges contribute the most beyond exact duplication (2,861 new merges, over $10\times$ the sequence-cluster contribution), underscoring how much antigen-level redundancy a CDR-H3-

Table 1: Marginal contribution of each edge type to the leakage graph's connected components over the 15,557-entry working population, applied in the order shown, primary tier (70% CDR-H3 identity). The largest single component after all edge types is 8,029 entries (Section A.3.2).

Edge type added	Components after	New merges (marginal)
(none)	15,557	—
Exact CDR-H3 duplicate	4,561	10,996
+ Sequence cluster (70%)	4,286	275
+ Antigen cluster	1,425	2,861
+ Epitope cluster	1,425	0
+ Paired heavy/light-chain cluster (90%)	1,285	140

only signal would leave unaddressed. Epitope-cluster edges contribute exactly zero marginal merges, a structural guarantee: epitope clusters are defined within antigen clusters (Section 4.1), so any two entries sharing one already share the antigen cluster that merged them a step earlier. Paired heavy/light-chain-cluster edges contribute 140 additional merges – entries whose full variable-domain pair is nearly identical despite different CDR-H3 and antigen clusters (Section A.3.1). Two further candidate edge types – broader full-variable-domain similarity and a Kabsch-RMSD structural edge – were evaluated and excluded (full characterization in Appendix A.3.2); after the five retained types, the largest connected component contains 8,029 entries (51.6% of the working population), too large to fit inside a 1,556-entry evaluation split without dominating it.

4.5 Split assignment

Each connected component is treated as an indivisible unit and assigned to train, validation, or test in three stages, over the 15,557-entry working population (full mechanism, pseudocode, and worked numbers in Appendix A.2 and Appendix A.2.1).

Stage A (hard constraints) forces a component to training, regardless of balance, if it exceeds a configurable fraction of the smaller of the validation/test targets (this alone forces the antigen-dominated mega-component, Section A.3.2, to training) or if it contains any `structural_confidence: low_resolution` entry, so evaluation is restricted to structures passing the predefined structural-confidence gate while training retains every entry regardless of resolution. Combined, these constraints force 118 components (9,914 entries, 63.7% of the working population) to training for the primary split; the resolution constraint is by far the larger of the two, forcing 117 low-resolution-only components (1,885 entries) beyond the mega-component alone (Section 2 discusses how this constrains Criterion 2's feasible assignment space). **Stage B (soft balance)** assigns remaining components to validation or test via a strict, never-overshoot greedy pass, ranking candidates lexicographically by fill-ratio bucket then by Jensen–Shannon divergence contribution across seven metadata axes (`antigen_class` weighted $4\times$, reflecting the largest observed marginal imbalance among the seven balance axes); training is the residual sink for whatever does not fit. **Stage C (exact-fill)** closes any residual gap via an exact-subset-sum search over training's components, failing loudly instead of reporting an off-target size (full mechanism in Appendix A.2.1); it did not need to trigger for any split reported here, since Stage B's own greedy pass lands exactly on target throughout.

5 Results

Split composition: The primary (70%-tier) split over the 15,557-entry protein/peptide-antigen working population contains 12,445 training, 1,556 validation, and 1,556 test entries, exactly matching the 80/10/10 target with no size tolerance.

5.1 Zero-leakage verification (Criterion 1)

Cross-split leakage was checked explicitly across all five edge-defining keys for each tier's own clustering (exact-duplicate group, sequence cluster, antigen cluster, epitope cluster, paired-Fv cluster), over all 15,557 working-population entries. No value of any key appears in more than one split, at any tier. This establishes disjointness of the stored cluster identifiers, which is a weaker property than a pairwise identity bound: MMseqs2 assigns each sequence to a single representative under

a greedy set-cover, so two sequences above the tier threshold can be placed in different clusters and therefore routed to different splits without any identifier ever colliding. We therefore bound the residual directly, not by inferring it from cluster membership, with an exhaustive all-versus-all alignment of every evaluation CDR-H3 against every training CDR-H3, clustering bypassed entirely (Appendix A.2.3 gives the exact commands). This surfaces real cross-split identity above the nominal 70% tier: 1,134 test–train pairs and 681 val–train pairs, maximum identity 81.8% (test) and 90.0% (val); the highest-identity crossings occur at ordinary CDR-H3 lengths (11–15 residues), not the short-loop regime where percentage identity is a coarse statistic by construction, so this is a real residual of the greedy clustering step, not a length artifact. As a further diagnostic, every split was also cross-checked against the 90%-tier CDR-H3 clustering, a commonly reported fixed reference point: clean at 40% and 70%; at 95%, 2 values (11 entries) share a 90%-tier cluster across splits, expected at a looser working tier than 90% itself.

5.2 Balance and diversity (Criterion 2)

Diversity within validation and test was checked directly, not just inferred from aggregate balance, since a split with acceptable marginal-distribution balance could in principle still be dominated by a small number of large components. Validation spans 294 distinct components (largest contributing 97 of 1,556 entries, 6.2%), 411 distinct antigen clusters, 1,117 distinct epitope clusters, 782 distinct PDB identifiers, and 12 distinct germline families. Test spans 332 distinct components (largest contributing 84 of 1,556, 5.4%), 416 distinct antigen clusters, 1,115 distinct epitope clusters, 894 distinct PDB identifiers, and 12 distinct germline families.

Table 2: Jensen–Shannon divergence (bits) between each split's composition and the working population's global distribution, across seven metadata axes.

Split	Species	Germ. (H)	Germ. (L)	Ag class	Therap.	CDR-H3 len.	Struct. conf.
Train	0.0056	0.0016	0.0035	0.0037	0.0000	0.0025	0.0009
Val	0.0754	0.0307	0.0552	0.0695	0.0000	0.0605	0.0429
Test	0.0641	0.0341	0.0580	0.0677	0.0003	0.0362	0.0429

To give the reported JS divergence a quantitative reference, we compared the balance-optimized assignment against 100 random feasible assignments under the same Stage A constraints, with Stage B's greedy pass replaced by random shuffle-and-fill. MDLAS achieves an unweighted JS divergence of 0.04553 – better than 63 of 100 draws (63rd percentile) and their mean (0.04695), but not extremal (best draw: 0.03476): Stage B's greedy pass is a substantial, not maximal, improvement over chance. It also improves rare-category coverage relative to a substantial fraction of the draws (full analysis in Appendix A.2.2).

5.3 Hyperparameter and seed robustness

Stage B's three tunable parameters (antigen-class tie-breaking weight, fill-ratio bucket size, oversized-component threshold) were swept over a 24-configuration grid, and split assignment was independently rerun across 21 random seeds, checking that the reported balance reflects the configuration, not a lucky setting or seed. Every configuration and seed passes zero-leakage verification identically – Criterion 1 is invariant to Stage B tuning by construction – and unweighted mean JS divergence stays within a narrow band in both sweeps (0.0439–0.0466 across the grid; mean 0.04554, std 0.00420 across seeds), with defaults landing within, not at the edge of, this range (full grids and per-seed analysis in Appendix A.4).

5.4 Robustness across CDR-H3 identity tiers

The leakage graph and split were reconstructed at all four CDR-H3 identity tiers (40%, 70%, 90%, 95%; full tables, figure, and seed/hyperparameter sweeps in Appendix A.4). All four converge to the same split sizes, pass zero-leakage verification independently, and the antigen-dominated mega-component shrinks monotonically as the tier tightens (53.3% at 40% to 47.6% at 95%). The 70% tier achieves the best balance score of all four and is reported as primary; the other three are fully built and equally verified for labs preferring a different stringency.

6 BASELINE MODEL EVALUATION

We next evaluate whether the MDLAS split produces meaningful differences in downstream model performance. We train and evaluate four representative CDR-H3 sequence-structure co-design baselines against the primary MDLAS split, spanning distinct generative paradigms and including both antigen-conditioned and antigen-blind models. We compare the resulting model ordering with results reported on the differently constructed CHIMERA-Bench split to assess whether relative model rankings are sensitive to split construction.

6.1 TASK, BASELINES, AND EVALUATION PROTOCOL

The downstream task is CDR-H3 sequence-structure co-design: given the available antibody framework, light-chain information when present, and antigen structure, generate the CDR-H3 loop's sequence and backbone conformation jointly, with the native loop removed from the input. CDR-H3 is typically the most sequence-diverse antibody CDR and a major contributor to antigen specificity, making it a useful task for testing whether benchmark construction changes comparative model performance. The four baselines span the distinct generative families represented in prior CDR-H3 design benchmarks, not variations on one architecture:[1] **MEAN** (Kong et al., 2023) (equivariant graph translation), **DiffAb** (Luo et al., 2022) (the first diffusion model for joint sequence-structure design on this task), **RefineGNN** (Jin et al., 2022) (autoregressive refinement, no conditioning on absolute spatial placement), and **AgForce** (Ahmed & Patterson, 2026) (an encoder-decoder built to counter *antigen blindness* – predictions ignoring antigen identity despite nominal conditioning, a failure mode common to the other three).

All four are trained and scored under CHIMERA-Bench's own evaluation protocol (Ahmed et al., 2026), applied against MDLAS's split instead: sequence recovery (AAR, CAAR), structural accuracy (Kabsch-aligned RMSD, TM-score), binding-interface accuracy (fraction of native contacts, interface RMSD, DockQ, epitope-F1), composite scores (Chimera-S, Chimera-B), and novelty. For baseline training only, we further restrict the 12,445-entry primary training set to its 11,152 full-confidence entries, applying Section 4.5's resolution flag to training as well, though the split itself retains every training entry regardless of resolution (Section 7.1); validation and test are unaffected, already 100% full-confidence by construction. Interface-dependent metrics are computed under each metric's own optimal superposition, not a single shared placement normalization: interface RMSD superposes the model's interface atoms onto the native interface atoms before measuring, and the antibody-level term superposes the predicted antibody onto the native antibody. Native coordinates of the atoms being scored therefore enter the transform in every case, so no metric reported here measures absolute placement of the designed loop in a frame fixed independently of the prediction. This follows the CAPRI convention for interface RMSD, but departs from the DockQ convention for the ligand-RMSD term, which superposes on the receptor and measures ligand displacement without re-fitting the ligand; the value reported here is correspondingly smaller than a receptor-anchored equivalent. All four baselines are scored under identical transforms, so their relative standing is unaffected; the absolute interface values are within-benchmark quantities and are not directly comparable to published DockQ or iRMSD numbers obtained under receptor-anchored protocols. Metric implementations and versions are pinned in the released evaluation package.

6.2 RESULTS

The four baselines separate cleanly instead of clustering together – early evidence that MDLAS's split is discriminative, not uniformly easy or hard. MEAN leads sequence recovery (AAR 0.359), significant against all three other baselines (Appendix A.5), by a wide margin against DiffAb and RefineGNN (0.206/0.240, $r = 0.91/0.84$) and a narrower, still-significant margin against AgForce (0.339, $r = 0.21$). This matches the AAR 0.35–0.37 range independently reported for MEAN-family GNN methods on CHIMERA-Bench's own, smaller epitope-group split (2,338/292/292 train/val/test) (Ahmed & Patterson, 2026), suggesting sequence recovery is a stable property of the method, not an artifact of either split. MEAN nonetheless trails AgForce on every binding-interface and composite metric in Table 4 (all significant, same appendix).

[1] Four of CHIMERA-Bench's eleven, chosen for paradigm coverage: testing whether split construction changes relative standing does not require retraining every baseline a prior benchmark reports, and Table 4 already covers the most directly comparable pair (RefineGNN vs. AgForce, antigen-blindness).

Table 3: Sequence recovery, structural accuracy, and novelty, all four baselines against MDLAS's primary (70%-tier) test split. AAR/CAAR: amino-acid recovery, overall and restricted to antigen-contacting residues. RMSD: Kabsch-aligned Cα RMSD (Å) against the true loop. Novelty: sequence divergence from the training set, reported without a best-value convention since higher is not unambiguously better. Best value per column in bold. Mean $\pm$ std over the test split.

Model	AAR	CAAR	RMSD (Å)	TM-score	Novelty
MEAN	$\mathbf{0.359 \pm 0.144}$	$\mathbf{0.184 \pm 0.210}$	1.754 ± 0.710	0.206 ± 0.109	0.309 ± 0.105
DiffAb	0.206 ± 0.129	0.126 ± 0.176	2.242 ± 2.193	0.174 ± 0.151	0.449 ± 0.083
RefineGNN	0.240 ± 0.128	0.107 ± 0.170	3.703 ± 1.007	0.066 ± 0.034	0.414 ± 0.103
AgForce	0.339 ± 0.154	0.159 ± 0.186	$\mathbf{1.591 \pm 0.807}$	$\mathbf{0.267 \pm 0.140}$	0.351 ± 0.107

Table 4: Binding-interface accuracy and CHIMERA-Bench composite scores, same split and baselines as Table 3. Fnat: fraction of native contacts recovered. iRMSD: interface RMSD (Å). DockQ, Epitope-F1, Chimera-S, Chimera-B: CHIMERA-Bench's own metric definitions. Best value per column in bold.

Model	Fnat	iRMSD (Å)	DockQ	Epitope-F1	Chimera-S	Chimera-B
MEAN	0.632 ± 0.281	1.447 ± 0.645	0.708 ± 0.183	0.773 ± 0.221	0.505 ± 0.093	0.576 ± 0.174
DiffAb	0.628 ± 0.294	2.050 ± 1.494	0.667 ± 0.186	0.703 ± 0.230	0.472 ± 0.114	0.540 ± 0.176
RefineGNN	0.646 ± 0.280	1.689 ± 0.774	0.686 ± 0.180	0.781 ± 0.203	0.415 ± 0.078	0.584 ± 0.161
AgForce	$\mathbf{0.684 \pm 0.272}$	$\mathbf{1.306 \pm 0.644}$	$\mathbf{0.740 \pm 0.188}$	$\mathbf{0.815 \pm 0.199}$	$\mathbf{0.536 \pm 0.103}$	$\mathbf{0.613 \pm 0.160}$

Because baseline implementations have model-specific input-compatibility failures, aggregate metrics are computed on each model's available test subset instead of forcing an artificial common subset. Pairwise statistical comparisons are therefore reported on the common subset specified in Appendix A.5. We emphasize effect sizes and consistency of the observed differences alongside statistical significance.

The two benchmarks agree on binding accuracy, under splits sharing none of their construction choices. On CHIMERA-Bench's own split, AgForce – built to counter antigen blindness – is reported ahead of RefineGNN on the binding-interface metrics common to both evaluations (Ahmed & Patterson, 2026).[2] That ordering reproduces under MDLAS: AgForce outperforms RefineGNN on all six binding-interface and composite metrics (Table 4), each significant (Holm-corrected $p < 10^{-9}$, jointly with every other pairwise comparison – Appendix A.5) and consistent per complex (AgForce wins 71.7% on DockQ). Sequence recovery does not track the same ordering: RefineGNN's AAR (0.240) sits at the upper edge of its 0.20–0.23 CHIMERA-Bench range instead of below it, while DiffAb (0.206) is the weakest baseline here instead. The two splits differ in scale and composition beyond antigen-disjointness alone, so this establishes that the ordering is stable under a substantially different construction – strengthening antigen separation does not, on its own, change which model scores higher at the binding interface. DiffAb, the only pure-diffusion model, trails on most metrics (not TM-score or Chimera-S, where RefineGNN scores lowest instead) and shows by far the largest variance – RMSD std 2.194 Å, iRMSD std 1.494 Å, against RefineGNN's next-highest 1.007 Å/0.774 Å. This is driven substantially by a single-complex near-total collapse (`9g13_GNA_agH_m0`; Appendix A.8): excluding it drops RMSD std by 46% and iRMSD std by 34%, so one isolated failure accounts for much, though not all, of the spread.

7 Discussion

7.1 Tradeoffs of antigen-disjoint, high-confidence evaluation

The oversized-component threshold routes the antigen-dominated mega-component (8,029 entries, 51.6% of the working population; Section A.3.2) to training by construction: components are indivisible under Criterion 1 (Section 4.5), so no subset can be reassigned to validation or test. Validation and test therefore contain essentially no representation of this redundancy, since any component large enough to evaluate on would overflow the split's budget or dominate its composition –

[2] AgForce and CHIMERA-Bench share two authors (Ahmed, Patterson); the comparison here is therefore a replication of that ordering on an independently constructed split, not a first report of it.

an inherent consequence of antigen-disjoint evaluation, independent of CDR-H3 tier (Section 5.4). Components that stay within budget still range from singletons to 84 members; Appendix A.13 shows baselines score measurably higher on binding-interface metrics for complexes with more same-component siblings, most acutely for the antigen-blind RefineGNN baseline. The resolution-based hard constraint (Section 4.5) carries a second, larger cost: validation and test contain no low-resolution structures, not the working population's global $\sim$8.3% share, forcing 117 further components (1,885 entries, 12.1%) to training beyond the mega-component alone, with training retaining the full population regardless of resolution.

8 Conclusion

Leakage-freedom and distributional consistency are independent properties of a benchmark split, and antibody benchmark studies have generally prioritized leakage control over distributional balance. MDLAS unions several antibody-specific redundancy relations – CDR-H3 duplication and clustering, antigen identity, epitope footprint, paired-Fv similarity – into a single graph whose connected components are indivisible assignment units, and optimizes marginal balance across seven metadata axes within the feasible space those components leave, while retaining every working-population entry. Against general-purpose similarity-aware splitters (Section 3), what is specific here is this relational construction and what it exposes downstream, not the pairing of the two objectives. The two-part criterion itself (Section 2) is not antibody-specific in principle, though every claim here is validated only on the antibody–antigen case; porting MDLAS to another domain (e.g. protein-ligand affinity, protein-protein interaction benchmarks) would still require identifying that domain's own redundancy signals and metadata axes.

9 Limitations and Future Work

MDLAS's guarantees are scoped to the five modeled redundancy relations. The epitope identifier is defined within antigen clusters, so conserved epitopes on only-distantly-related antigens may not be linked; similarly, full-antibody, framework, or other sequence similarities not represented in the graph may remain. VH/VL and framework similarity were evaluated as candidate relations but produced a component covering most entries and were therefore not retained as general-purpose edges (Section A.3.1).

Criterion 2 measures marginal consistency across seven predefined axes instead of joint or conditional representativeness. Moreover, oversized and low-resolution-only components constrain the feasible assignment space, so the greedy assignment does not certify an optimal balance; some individual random feasible assignments achieve better aggregate scores (Appendix A.2.2). These constraints should therefore be interpreted as defining the evaluation population instead of guaranteeing universal distributional representativeness.

Finally, baseline evaluation is affected by heterogeneous input coverage: 510 of 1,556 test complexes (32.8%) lack a light chain, and all four evaluated models perform worse on this subset (Appendix A.10). We also have not compared MDLAS against general-purpose similarity-aware splitters under matched constraints (Joeres et al., 2025; Teufel et al., 2023); consequently, the value of the antibody-specific relation set relative to a domain-agnostic similarity matrix remains unquantified.

Data and Code Availability

MDLAS is constructed from the publicly available SAbDab database, using the 28 May 2026 SAbDab snapshot documented in the reproducibility statement. The MDLAS pipeline (structural feature extraction, sequence/antigen/epitope/germline clustering, multi-relation leakage-graph construction, and three-stage split assignment), the primary 70%-tier split and three alternate-tier reconstructions, and the four baseline models' training, inference, and evaluation scripts are provided as supplementary submission artifacts and will be released publicly upon acceptance as a self-contained, installable Python package. The original SAbDab data are not redistributed and can be obtained from the public SAbDab resource.

AI USE STATEMENT

AI tools are used only as writing and language-assistance tools during the preparation of this manuscript. Their use was limited to tasks such as grammar and spelling correction, improving sentence clarity and readability, refining academic phrasing, identifying typographical or formatting inconsistencies, and assisting with the organization and polishing of written text. All scientific content, including research questions, methodology, algorithmic design, experimental setup, data processing decisions, analysis, interpretations, and conclusions, was developed and verified by the authors. AI tools were not used as part of the scientific or computational pipeline, were not used to generate or process experimental data, nor to perform experiments, model training, evaluation, statistical analysis, or other substantive research tasks. The authors remain fully responsible for the accuracy, originality, and integrity of the manuscript and its contents.

REPRODUCIBILITY STATEMENT

The data-extraction, leakage-graph construction, and split-assignment pipeline is described in Sections 4.2–4.5 and Appendix A.1–A.2; the evaluation protocol and metric implementations used in Sections 5.1–6.2 and the four baseline models' training and inference configurations are described in Section 6.1. Split-assignment hyperparameters (oversized-component threshold, fill-ratio bucket size, axis weights) and the seeds used for the primary split and its robustness sweeps are reported in Section 5.3 and Appendix A.4.1–A.4.2. Code and split files are covered by the Data and Code Availability statement above. For this work we used a SAbDab summary file downloaded on 28 May 2026. (`sabdab_summary_all.tsv`): SHA-256 f7cc4972c0da4cce4cb1e55ff8894b1 42c199825ced03ea928228d93d4527b27

REFERENCES

Jared Adolf-Bryfogle, Oleks Kalyuzhniy, Michael Kubitz, Brian D. Weitzner, Xiaozhen Hu, Yumiko Adachi, William R. Schief, and Roland L. Dunbrack. Rosettaantibodydesign (rabd): A general framework for computational antibody design. *PLoS Computational Biology*, 14(4):e1006112, 2018. doi: 10.1371/journal.pcbi.1006112.

Mansoor Ahmed and Murray Patterson. Agforce enables antigen-conditioned generative antibody design. *arXiv preprint arXiv:2605.21610*, 2026. doi: 10.48550/arXiv.2605.21610.

Mansoor Ahmed, Nadeem Taj, Imdad Ullah Khan, Hemanth Venkateswara, and Murray Patterson. Chimera-bench: A benchmark dataset for epitope-specific antibody design. *arXiv preprint arXiv:2603.13431*, 2026. doi: 10.48550/arXiv.2603.13431.

Anton Bushuiev, Roman Bushuiev, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, and Josef Sivic. Revealing data leakage in protein interaction benchmarks. *arXiv preprint arXiv:2404.10457*, 2024. doi: 10.48550/arXiv.2404.10457.

James Dunbar and Charlotte M. Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016. doi: 10.1093/bioinformatics/btv552.

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. doi: 10.1093/nar/gkt1043.

David Graber, Peter Stockinger, Fabian Meyer, Siddhartha Mishra, Claus Horn, and Rebecca Buller. Resolving data bias improves generalization in binding affinity prediction. *Nature Machine Intelligence*, 7:1713–1725, 2025. doi: 10.1038/s42256-025-01124-5.

Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi Jaakkola. Iterative refinement graph neural network for antibody sequence-structure co-design. In *International Conference on Learning Representations*, 2022. arXiv:2110.04624.

Roman Joeres, David B. Blumenthal, and Olga V. Kalinina. Data splitting to avoid information leakage with DataSAIL. *Nature Communications*, 16, 2025. doi: 10.1038/s41467-025-58606-8.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3d equivariant graph translation. In *International Conference on Learning Representations*, 2023. arXiv:2208.06073.

Aleksandr Kovaltsuk, Jinwoo Leem, Sebastian Kelm, James Snowden, Charlotte M. Deane, and Konrad Krawczyk. Observed antibody space: A resource for data mining next-generation sequencing of antibody repertoires. *The Journal of Immunology*, 201(8):2502–2509, 2018. doi: 10.4049/jimmunol.1800708.

William D. Lees, Christian E. Busse, Martin Corcoran, Mats Ohlin, Cathrine Scheepers, Frederick A. Matsen, Gur Yaari, Corey T. Watson, Andrew Collins, and Adrian J. Shepherd. Ogrdb: a reference database of inferred immune receptor genes. *Nucleic Acids Research*, 48(D1):D964–D970, 2020. doi: 10.1093/nar/gkz822.

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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9754–9767, 2022.

Tobias H. Olsen, Fergus Boyles, and Charlotte M. Deane. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022. doi: 10.1002/pro.4205.

Felix Teufel, Magnus Halldor Gislason, Jose Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, and Henrik Nielsen. GraphPart: homology partitioning for biological sequence analysis. *NAR Genomics and Bioinformatics*, 5(4):lqad088, 2023. doi: 10.1093/nargab/lqad088.

Yiming Zhang and Koji Tsuda. Nbbench: Benchmarking language models for comprehensive nanobody tasks. *Machine Learning: Science and Technology*, 6:040502, 2025.

A Appendix

This appendix contains extended methodological detail, full algorithm pseudocode, and extended validation and robustness results supporting the main text.

A.1 Extended structural feature extraction

Full schema. Each output `.pt` file is keyed by each row's unique `filename_stem` identifier and combines two namespaces: a set of fields matching CHIMERA-Bench's schema verbatim (`complex_id`, `heavy_sequence`, `light_sequence`, `antigen_sequence`, `heavy_atom14_coords`, `heavy_ca_coords`, `epitope_residues`, `paratope_residues`, `contact_pairs`, `numbering`, `cdr_masks`, `ag_surface_points`, `ag_surface_chemical_feats`), and an additional `mdlas` sub-dictionary carrying species, germline (family and, for human/mouse, allele-level), therapeutic status, exact-duplicate group, per-tier CDR-H3 cluster identifiers, antigen- and epitope-cluster identifiers, an antigen-chemistry flag, deposition year, and resolution.

Numbering and atom14 coordinates. Heavy and (where present) light chain residues are extracted from the IMGT-numbered structure bundle, with residue numbers read directly from the PDB residue identifier (fractional encoding for insertion codes); Chothia numbering comes from parsing the sibling `chothia/` bundle with identical logic, matched to the IMGT-numbered residues by chain and residue count – numbering is taken from these pre-numbered SAbDab bundles instead of re-derived via ANARCI, as CHIMERA-Bench does, trading its independent numbering cross-validation for one fewer external dependency. Atom14 coordinates are extracted per residue using the standard 14-slot heavy-atom ordering (backbone N, C_α, C, O followed by residue-specific side-chain atoms), validated against expected residue-specific heavy-atom occupancy. CDR masks follow IMGT boundaries (H1: 27–38.9, H2: 56–65.9, H3: 105–117.9, and the equivalent light-chain ranges), encoded as -1 for framework and 0–5 for H1–H3/L1–L3, matching CHIMERA-Bench's encoding convention.

Interface and resolution thresholds. Two thresholds govern downstream fields, kept separate: an interface cutoff of 4.5 Å (heavy-atom distance) defines epitope/paratope contact whenever contacts are computed, matching CHIMERA-Bench's definition, and a resolution gate of 4.0 Å governs *whether* side-chain-dependent fields (full atom14 beyond backbone, contact pairs, epitope/paratope residues, antigen surface features) are computed at all. Entries coarser than 4.0 Å are not excluded from the feature store, but have these fields set to null with an explicit `structural_confidence: low_resolution` flag; backbone fields are populated to the extent the deposited structure itself resolves them. For most low-resolution entries this is a complete backbone (N, C_α, C, O), but the most extreme cases provide only a C_α trace, a property of the source deposition itself: a 22.0 Å cryo-EM deposition, for example, contains exclusively `CA` atom records for its heavy chain, with no $N/C/O$ records in the file to extract. This is captured via a per-atom `atom14_mask` instead of fabricated; complete backbone geometry should not be assumed from `structural_confidence` alone without checking this mask. Across the full source pool, 18,684 entries (93.0%) retain full structural confidence and 1,411 (7.0%) are flagged low-resolution. This flag is a confidence gate at extraction time, not a row-exclusion filter; Section 4.5 describes a second, stricter use of the same flag specifically for validation and test eligibility.

Antigen surface features. Antigen residues are retained based on a curated exclusion list (water, common ions, and crystallization or buffer additives) instead of a standard-amino-acid-only filter, so that structurally resolved nucleic-acid and carbohydrate antigens are captured as legitimate antigen chemistry at extraction time instead of dropped. Each antigen-bound entry's chemistry is then classified by majority residue composition (Section 4.1); the protein-oriented six-channel antigen surface featurization (128 points via farthest-point sampling over antigen Cα positions, following CHIMERA-Bench's schema shape, with per-residue Kyte–Doolittle hydrophobicity, formal charge, hydrogen-bond donor/acceptor flags, aromaticity, and polarity) is defined only for majority protein/peptide entries, since that feature scheme has no defined analogue for nucleic-acid or carbohydrate chemistry.

A.1.1 An unresolvable CDR-H3 boundary case

One entry, `6do1_FNA_agB_m0` (an unpaired, heavy-chain-only antibody), is excluded from all CDR-H3-based clustering: its resolved heavy chain terminates at `...YYCAAV`, immediately at the conserved motif preceding the CDR-H3 loop, before any H3 residues are resolved. This is caught directly from the residue range and flagged `h3_resolved: false`, not left to surface later as a downstream clustering artifact.

A.1.2 Epitope-cluster alignment procedure

CHIMERA-Bench's epitope-cluster definition – two complexes belong to the same cluster if and only if their sorted epitope residue identifiers (chain, position) are identical – requires globally comparable residue identifiers. A naive (chain, author residue number) key is insufficient for this: both are local to a single PDB deposition, so this identifier systematically undercounts true epitope redundancy across differently-numbered or differently-chain-labeled depositions of the same biological epitope. MDLAS instead computes a sequence-aligned identifier: within each antigen cluster (Section 4.1), MDLAS recovers the cluster's representative antigen sequence – embedded in the antigen-cluster identifier itself – and aligns every other member's antigen sequence to it via local BLOSUM62 alignment (the same method used for germline allele assignment, Section 4.3), then translates each entry's epitope residue indices (defined relative to its own antigen sequence) into reference-sequence coordinates via that alignment. The epitope-cluster identifier is then a hash of (antigen-cluster identifier, sorted reference-sequence positions); two entries share an epitope cluster if and only if they share an antigen cluster *and* their contact residues map to identical positions on the cluster's reference sequence, an exact-match rule matching CHIMERA-Bench's own. This is computed only for protein/peptide-antigen entries within the resolution confidence gate, since BLOSUM62 alignment is undefined for nucleic-acid or carbohydrate sequences; a small number of entries pass the antigen-class check but still contain characters outside the aligner's supported alphabet and are excluded by a second, defensive check.

Of 14,289 antigen-bound, full-confidence, protein/peptide entries with at least one contact residue, 14,248 (99.7%) receive a computed epitope-cluster identifier – 3,463 via an identity fast-path (entries that are themselves their cluster's representative sequence) and 10,826 via real alignment – resolving

to 9,951 unique epitope clusters; the remaining 41 map to an empty position set after alignment. Entries without an epitope-cluster identifier are excluded for one of four disjoint reasons: 4,350 have no antigen cluster (no antigen present at all), 102 have a non-protein/peptide `antigen_class`, 26 use characters outside the aligner's supported alphabet, and 1,328 have zero epitope contact residues despite passing every upstream filter.

A.1.3 Germline allele resolution: full distribution

Germline family assignment (Section 4.1) is coarse by design, resolving each entry only to a heavy- or light-chain V-gene family. Allele-level germline assignment is computed directly by MDLAS's own pipeline: a fresh OGRDB (Lees et al., 2020) reference set is fetched (queried through `receptor_utils`) and BLOSUM62-aligned against each entry's own extracted heavy and light chain sequences, restricted to human and mouse entries, for which the available curated references support confident allele-level calls, and joined into MDLAS's schema as an additive `mdlas` field, keyed on `filename_stem`.

14,282 of the 20,968-row source pool (68.1%) receive a chain-level germline assignment; the remainder falls outside human/mouse scope (llama, alpaca, rabbit, macaque, rat, camel, chicken, and others) or has no successfully extracted structure to assign against (873 entries). Restricted to human entries with a confident call, 9,688 receive a V-family-resolved heavy-chain allele, with ranking IGHV3$(4{,}565)$ > IGHV1$(3{,}113)$ > IGHV4$(1{,}270)$ > IGHV5(324) > IGHV2(201) > IGHV7(143) > IGHV6(72); IGHV3 and IGHV1 together account for 79.3% of these (Figure 3). Mouse allele names (e.g. `IGHV0-24BS*00`) are OGRDB strain-inferred placeholders instead of official IMGT designations, so no family-level rollup is computed for mouse alleles.

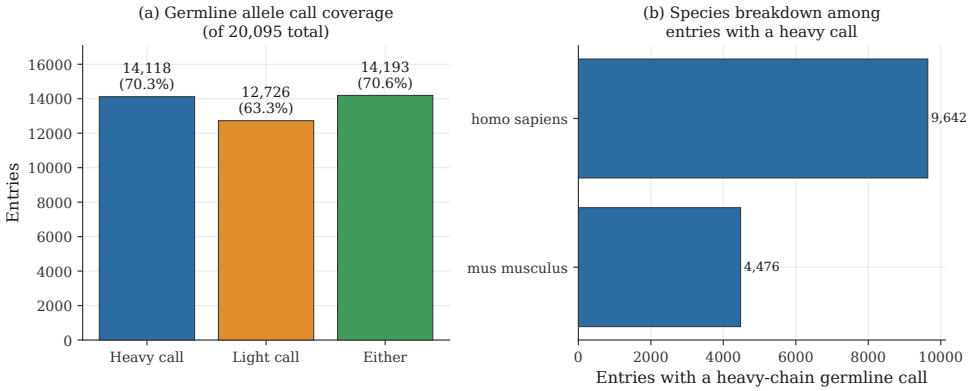

Figure 3: (a) Germline allele coverage using the confident-V-gene-call definition (not the looser species-match-attempted definition – see text). (b) Species breakdown among entries with a confident heavy-chain call, showing that the resulting confident calls are concentrated in human and mouse entries, as expected from the restricted reference set.

A.2 Algorithms

A.2.1 Split assignment: full mechanism

Each connected component is treated as an indivisible unit and assigned to train, validation, or test in three stages, over the 15,557-entry working population.

Stage A (hard constraints). Two independent hard constraints force a component to training regardless of its balance contribution. First, any component exceeding a configurable fraction of the smaller of the validation and test targets cannot fit in a small evaluation split without either exceeding that split's entire size budget or dominating its composition. For validation and test targets of 1,556 entries each (an 80/10/10 train/validation/test split of the 15,557-entry working population) and the reported oversized threshold of 50% of that target (778 entries), this constraint alone forces exactly 1 component – the antigen-dominated mega-component (Section A.3.2) – to training.

Algorithm 1 Multi-relation leakage graph construction

Require: Working-population entries E, each with fields `exact_duplicate_group_id`, `cdr_h3_cluster_id` (primary tier), `antigen_cluster_id`, `epitope_cluster_id`, `paired_fv_cluster_id` (90% tier)
Ensure: Component assignment ϕ : complex_id $\rightarrow$ representative complex_id

```
Initialize union–find U over { complex_id(e) : e ∈ E }
for key k ∈ (exact_duplicate_group_id, cdr_h3_cluster_id, antigen_cluster_id, epitope_cluster_id, paired_fv_cluster_id) do ▷ fixed order, for marginal-contribution accounting only
    Group entries of E by value of k, discarding entries with no value
    for each group g with |g| ≥ 2 do
        for each m ∈ g \ {g_1} do
            U.UNION(g_1, m)
        end for
    end for
end for
C ← U.COMPONENTS()                ▷ final partition does not depend on edge-processing order
for each component c ∈ C do
    r ← min(c)                    ▷ deterministic representative
    for each id ∈ c do
        φ(id) ← r
    end for
end for
return φ
```

Second, any component containing at least one `structural_confidence: low_resolution` entry (Section 4.2) is also forced to training, so that validation and test contain only full-confidence structural ground truth. Because components are indivisible under Criterion 1, excluding only the low-resolution member would require splitting a leakage-connected component and would therefore violate the hard leakage constraint; the resolution rule consequently operates at the component level instead of the entry level. This is also a deliberately stricter standard than the confidence gate applies to the feature store as a whole: scoring a model's predictions against a low-resolution reference structure conflates genuine model error with noise in the reference itself, which is a cost evaluation cannot absorb the way training can. Training benefits from retaining every entry regardless of resolution – a model can be robust to label noise in training examples in a way an evaluation reference cannot be robust to noise in itself – so nothing is removed from the feature store or from training; only the pool eligible for validation/test membership is restricted. Combined, the two constraints force 118 components (9,914 entries, 63.7% of the working population) to training for the primary 70%-tier split: 0 components for being oversized only (the oversized threshold at 778 entries is high enough that no non-mega component reaches it), 117 components (1,885 entries) for containing a low-resolution entry only, and 1 component (8,029 entries – the mega-component itself, which is both oversized and contains low-resolution entries) for both. Unlike the oversized constraint, whose cost is fixed once the mega-component is identified, the resolution constraint's cost scales directly with the working population's own low-resolution share, and is by far the larger of the two.

Stage B (soft balance). Remaining components are assigned to validation or test via a strict, never-overshoot greedy pass; training is a residual sink for whatever does not fit in either, since its target size is defined as $n_{\text{total}} - \text{val} - \text{test}$ instead of as an independent balance objective. For each remaining component, its Jensen–Shannon divergence contribution against the working population's global distribution is computed across seven metadata axes: heavy-chain species, heavy- and light-chain germline family, antigen class, therapeutic status, CDR-H3 length, and structural confidence. Candidates are compared lexicographically, first by a coarse fill-ratio bucket (5% of target), then by JS divergence within that bucket, not by an additive score. Although base-2 JS divergence is bounded to $[0, 1]$ bit regardless of category count, high-cardinality axes such as CDR-H3 length ($\sim$33 categories) can produce larger finite-sample fluctuations in the JS estimate when candidate splits are sparse. More importantly, fill ratio and distributional similarity serve different optimization goals: the former enforces the desired split size, whereas the latter refines the choice among candidates

Algorithm 2 Three-stage split assignment

Require: Components (Algorithm 1's output, grouped) with per-axis metadata counts; targets $T_{\text{train}}, T_{\text{val}}, T_{\text{test}}$; oversized fraction ρ; EXCLUDELOWRES flag; fill-bucket size β; axis weights w

Ensure: Split membership $\{\text{train}, \text{val}, \text{test}\}$, sizes exactly $(T_{\text{train}}, T_{\text{val}}, T_{\text{test}})$; per-axis JS divergence report

Stage A (hard constraints)

$\tau \leftarrow \rho \cdot \min(T_{\text{val}}, T_{\text{test}})$ ▷ oversized threshold
for each component c **do**
  if $|c| > \tau$ **or** (EXCLUDELOWRES **and** c contains a `low_resolution` entry) **then**
    Assign $c \rightarrow$ train; update train's per-axis counters
  else
    Add c to R (remaining components)
  end if
end for

Stage B (soft balance, strict feasibility)

Order R by a fixed-seed (42) uniform random permutation ▷ not size-sorted – avoids confounding processing order with antigen-class-correlated component size
for each component $c \in R$, in that order **do**
  candidates $\leftarrow \emptyset$
  for split $s \in \{\text{val}, \text{test}\}$ **do**
    if $\text{size}(s) + |c| > T_s$ **then**
      continue ▷ infeasible, skip – no size tolerance
    end if
    bucket $\leftarrow \text{ROUND}\big((\text{size}(s)/T_s)/\beta\big)$
    js $\leftarrow \sum_{a \in \text{axes}} w_a \cdot \text{JS}\big(\text{counter}_s[a] + c[a],\ \text{global}[a]\big)$
    candidates $\leftarrow$ candidates $\cup \{(\text{bucket}, \text{js}, s)\}$
  end for
  if candidates $\neq \emptyset$ **then**
    $s^* \leftarrow \arg\min$ over candidates, lexicographic on $(\text{bucket}, \text{js})$ ▷ fullest-behind split wins; JS breaks ties within a bucket
  else
    $s^* \leftarrow$ train ▷ no room in val/test; falls to the residual sink
  end if
  Assign $c \rightarrow s^*$; update s^*'s per-axis counters
end for

Stage C (exact-fill)

for split $s \in \{\text{val}, \text{test}\}$ **do**
  $d \leftarrow T_s - \text{size}(s)$ ▷ residual deficit, $d \geq 0$ by Stage B's strict feasibility check
  if $d > 0$ **then**
    Find an exact subset $S \subseteq$ train's components with $\sum_{c \in S} |c| = d$, via subset-sum DP, ties broken by balance-quality score
    if no exact subset exists **then**
      fail ▷ loudly – no silent off-target split
    end if
    Move S from train to s; update both splits' per-axis counters
  end if
end for
return split membership (sizes now exactly $T_{\text{train}}, T_{\text{val}}, T_{\text{test}}$); per-axis, per-split JS divergence vs. global distribution

with comparable fill. A lexicographic objective therefore avoids arbitrary weight tuning while ensuring that small improvements in distribution matching never outweigh the primary requirement of achieving the target split size. Components are processed in fixed-seed (42) uniform random order instead of by size, since component size correlates with antigen class – redundant classes (`viral`, `other_protein`) disproportionately form larger components – so a size-ordered traversal would bias which classes are available to route at any point in the assignment. Antigen class is weighted

$4\times$ in the tie-breaking score, reflecting the largest observed imbalance among the seven axes (Section 5.3). Structural confidence remains one of the seven axes for consistency of implementation, but is no longer a genuine degree of freedom for Stage B once the Stage A resolution constraint is active: every remaining candidate component is already full-confidence by construction, so validation and test necessarily diverge from the working population's global low-resolution share by a fixed amount (Section 5.2), not something Stage B can optimize toward zero.

Stage C (exact-fill). Stage B's strict feasibility check (never assign a component if doing so would exceed a split's target) generally leaves validation and/or test slightly short of their exact target once no remaining component fits the residual gap – an expected consequence of assigning whole, variably-sized components instead of individual entries. Rather than accept this residual gap or relax Stage B's feasibility check with a size tolerance, a third stage closes it exactly: an exact-subset-sum search over training's current components finds a subset whose sizes sum to precisely the outstanding deficit, moving that subset from training into the deficient split; where multiple exact subsets exist, ties are broken by the same balance-quality score Stage B uses, so the specific subset chosen still improves distributional match instead of being arbitrary. If no exact subset exists, split construction fails loudly instead of silently reporting an off-target split size. Stage B's own greedy pass lands exactly on target for every split reported in this paper (the primary tier and all three alternate tiers, Section 5.4), verified from each run's log. Stage C was not needed for the reported runs. Where it is needed, an exact-subset-sum search over leftovers alone can fail even when a valid repair exists: in the configurations we examined, every eligible leftover component individually exceeded both residual gaps, so no subset of them sums to either deficit, while reassignment among already-allocated components could still close it. A repair procedure for that case is a local search over already-assigned eligible components – pairwise swaps with backtracking, preserving Stage A's oversized- and low-resolution-component locks as hard constraints – instead of a larger subset-sum. This does not affect any split reported in this paper, all of which land on target in Stage B.

A.2.2 Balance versus random-feasible assignment: full analysis

To give the reported JS divergence a quantitative meaning, we compared it against a distribution of 100 random feasible assignments: Stage A's oversized- and low-resolution-component forcing applied identically, with Stage B's balance-optimizing greedy assignment replaced by pure random shuffle-and-fill, ignoring composition entirely. Random-feasible assignment achieves mean unweighted JS divergence 0.04695 (std 0.00390, range [0.03476, 0.05532]) against the working population; MDLAS's actual balance-optimized split achieves 0.04553, better than 63 of the 100 random-feasible draws (63rd empirical percentile) and better than their mean, but 0.01077 above the best individual draw (0.03476). Stage B's greedy heuristic therefore captures most, but not all, of the achievable gain over chance within the Stage A-constrained feasible space; we have not separately quantified how much of the residual gap reflects categories no feasible assignment under these constraints can balance further, versus suboptimality specific to the greedy heuristic (Section 4.5) that a stronger assignment procedure could in principle close.

Rare-category coverage tells a consistent story. Across the 100 random draws, 117 (axis, category) combinations were missing from validation or test in at least one draw – categories too rare, given the target split sizes, for any single assignment to reliably capture. The real split is missing 74 of these same combinations, but none exclusively: every category absent from the real split is also absent from at least one random draw, and the real split additionally succeeds in representing 45 categories that random assignment sometimes fails to capture. The real split therefore improves rare-category coverage relative to a substantial fraction of random feasible assignments while retaining the same hard feasibility constraints.

A.2.3 Exhaustive pairwise cross-split identity audit

Cluster-identifier disjointness (Section 5.1) certifies no cluster label is shared across splits, not that the closest cross-split pair falls below the identity threshold, since MMseqs2's greedy set-cover assigns each sequence to exactly one representative. We removed clustering from the question entirely: CDR-H3 sequences were extracted per split (`cdr_masks` loop index 2 over `heavy_sequence`)

and searched all-versus-all with an exhaustive, unclustered `mmseqs2` search (sensitivity 7.5, no hit cap, no identity floor, 80% bidirectional coverage).[3]

Across the primary (70%-tier) split, 1,134 of the 1,556×12,445 test–train CDR-H3 pairs and 681 of the 1,621×12,445 val–train pairs reach or exceed 70% identity (maximum 81.8% and 90.0% respectively), concentrated at typical CDR-H3 lengths instead of the sub-10-residue regime – none of the twenty highest-identity pairs involve a loop shorter than 11 residues. This is the residual MMseqs2's greedy clustering misses (Section 5.1): real, not negligible, and disclosed directly instead of only implied by cluster-ID disjointness.

A.3 Extended leakage-graph validation

A.3.1 Full variable-domain similarity is not a viable general-purpose edge

Beyond CDR-H3, four broader sequence-redundancy signals were computed and evaluated directly as candidate leakage-graph edges: heavy-chain variable-domain (VH), light-chain variable-domain (VL), framework-region (residues outside all six CDR loops), and paired heavy+light-chain (paired-Fv) sequence-cluster membership, each at multiple identity thresholds.

Applied unscoped at 90% identity, combining all four signals produces a single 14,619-entry component – 93.9% of the 15,577-entry working population. This is not genuine redundancy: the component spans 740 distinct antigen clusters, and its internal germline-family and antigen-class composition closely matches the overall population's instead of showing the concentration a real redundancy cluster would. Raising the identity threshold to 95% shrinks the component only to 12,354 entries (79.3%) – still large enough to make evaluation on the complement meaningless. A second variant, scoping these four signals to operate only within existing antigen clusters, is provably useless before it is even run: any two entries sharing an antigen cluster are already merged by the antigen-cluster edge applied earlier in the fixed order (Table 1), so an antigen-scoped version of any later edge is a strict subset of an already-fully-merged relation and is mathematically guaranteed to contribute zero marginal structure.

A likely explanation is a broader property of antibody sequence space instead of a feature unique to this dataset: variable-domain sequences share a small set of common germline scaffolds across the entire repertoire, related antibodies or not, so a similarity threshold applied globally over VH, VL, or framework sequence percolates – transitively merging distantly related sequences into one giant component – in a way CDR-H3 clustering and antigen clustering, both scoped to much more specific biological identity, do not. Paired-Fv clustering is the most conservative of the four signals precisely because it requires both chains to match simultaneously instead of either chain independently, and it is consistently the smallest-clustering signal at every threshold tested: the largest paired-Fv cluster is 80 entries at 90% identity and 24 at 95%, versus 84–403 for VH, VL, or framework clustering alone at the same thresholds. MDLAS accordingly retains only paired-Fv clustering as a hard, split-blocking edge (Table 1); VH, VL, and framework clustering are computed and available in the schema but are not applied as default leakage-graph edges, since doing so would collapse nearly the entire working population into one component. Criterion 1's zero-leakage guarantee is therefore scoped to the five edge types in Table 1, not to full antibody sequence redundancy in general (Section 9).

A.3.2 The largest component is antigen-dominated concentration

After all five edge types are applied, the largest connected component contains 8,029 entries – 51.6% of the 15,557-entry working population – computed directly over the restricted working population (Figure 4 shows the full component-size distribution).

Within it, 430 distinct antigen clusters are represented, the three largest individually accounting for 1,367, 1,225, and 841 members – together 42.8% of the component's own membership, a meaningfully concentrated core instead of a diffuse tail. Sampled sequences from the single largest antigen cluster (1,367 members) match coronavirus spike glycoprotein, including at least one member iden-

[3] `mmseqs easy-search` with `-s 7.5 --max-seqs 100000 -e 10000 -c 0.8 --cov-mode 0 --alignment-mod -e 10000` specifically prevents short-sequence pairs being discarded on E-value before their identity is reported.

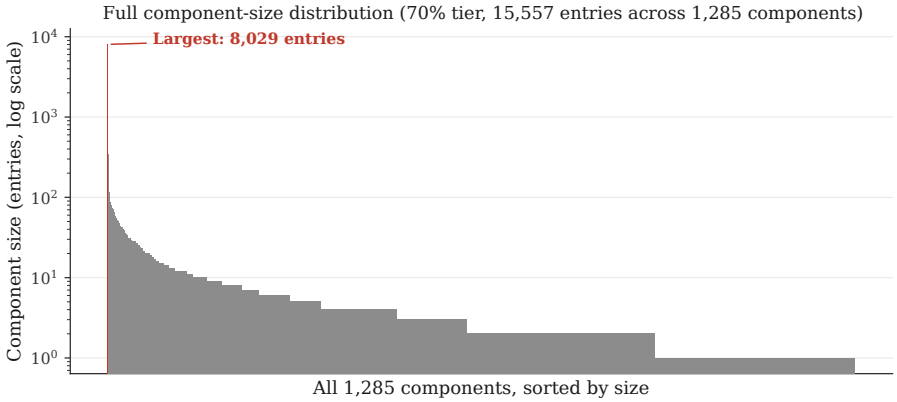

Figure 4: Full component-size distribution, all 1,285 components, sorted descending, log scale. One component dominates the entire distribution by more than an order of magnitude over the next-largest.

tical to a solved SARS-CoV spike/S230-antibody complex; SARS-CoV and SARS-CoV-2 spike receptor-binding domains are close homologs and plausibly cluster together at this identity threshold. The component spans 4,276 distinct PDB identifiers – an entries-per-PDB density (1.878) close to the working population's own overall ratio ($1.886\times$, Section A.3.4) – and 1,746 distinct 70%-tier CDR-H3 sequence clusters, supporting the interpretation that this is broad antigen-associated concentration instead of a narrow sequence-level artifact. The component is 60.9% `viral`-class by the source table's coarse antigen-class annotation. A single component this large cannot fit inside a 1,556-entry evaluation split without dominating it, motivating its exclusion from validation and test under a genuine antigen-disjointness constraint (Section 7.1).

A.3.3 Structural redundancy ablation

A fifth candidate edge type, backbone structural similarity, was evaluated directly instead of assumed, using the same method applied elsewhere to this shared dataset: for every pair of equal-length CDR-H3 loops within the same 90%-tier sequence cluster, Cα coordinates are superposed via the Kabsch algorithm and RMSD computed, restricted to within-cluster pairs and never across clusters. The Kabsch implementation was validated against synthetic rotated, translated, and mirrored point clouds before being applied to real data.

Run over MDLAS's own atom14 coordinates, this gives 393,984 pairs compared, 87.48% under 1 Å, and 98.27% under 2 Å, closely matching the profile expected of genuine within-cluster duplication and consistent with the Kabsch implementation's own synthetic-data validation above. Applying the 387,158 pairs falling under the 2 Å cutoff as an additional candidate edge type produces exactly zero marginal new merges, because the sequence-cluster edge has already unioned every member of a shared cluster regardless of RMSD: any structural edge built this way is a strict subset of a relation the graph has already fully merged. This validates the sequence-cluster edge – within-cluster sequence redundancy in this dataset is, essentially without exception, also structural redundancy – but leaves open whether structurally similar loops exist *across* different sequence clusters, examined next.

A.3.4 Cross-cluster structural similarity reflects conformational-class overlap, not redundancy

We computed pairwise Kabsch RMSD, using the same validated implementation, for all same-length CDR-H3 pairs spanning distinct 90%-tier clusters, over the full extracted population with a resolved CDR-H3 and cluster assignment (20,092 entries – broader than the protein/peptide-antigen working population used elsewhere in this paper, since conformational diversity of the CDR-H3 loop is not itself an antigen-dependent property): 13,800,879 total same-length pairs, of which 13,407,042 (97.1%) are cross-cluster. At the same 2.0 Å cutoff used for the within-cluster ablation, 2,851,618 pairs (21.3%) naively qualify as similar, but two properties indicate this is not genuine redundancy:

the RMSD distribution is inverted relative to the within-cluster ablation's true-duplicate profile (mean/median RMSD 1.492/1.581 Å at this cutoff, versus mean 0.504 Å for genuine within-cluster duplicates), and the degree distribution is saturated instead of sparse (17,557 entries have at least one qualifying partner, mean degree 324.8, median 169.0, max 1,412, only 2.23% with a single partner).

Re-examining the same 2.85M scored pairs at progressively tighter thresholds (Table 5) tests whether a sparse, duplicate-like core survives beneath this loose cutoff. It does not: tightening the cutoff by $4\times$ shrinks the qualifying set by 98.6% and the maximum degree from 1,412 to 91, but the median entry still retains 7 qualifying partners even at 0.5 Å – tighter than the threshold at which the great majority of true within-cluster duplicates already fall – with no discontinuity anywhere in the progression suggestive of a hidden sparse core.

Threshold	Qualifying pairs	% of 2.0 Å set	Mean/median RMSD	Entries w/ partner	Mean/median degree	% degree=1
2.0 Å (original)	2,851,618	100%	1.492 / 1.581	17,557	324.8 / 169.0	2.23%
1.0 Å	382,721	13.42%	0.756 / 0.796	11,498	66.6 / 30.0	6.18%
0.5 Å	41,035	1.44%	0.388 / 0.410	5,538	14.8 / 7.0	13.87%

Table 5: Cross-cluster CDR-H3 structural similarity at progressively tighter RMSD thresholds. Tightening the cutoff by $4\times$ shrinks the qualifying set by 98.6% and the maximum degree from 1,412 to 91, but the median entry retains 7 qualifying partners even at the strictest threshold tested – never approaching the near-singleton profile expected of genuine duplication.

We conclude this reflects CDR-H3's well-documented limited canonical conformational diversity instead of a leakage-relevant redundancy signal: many same-length CDR-H3 loops occupy one of a small number of recurring backbone conformations regardless of sequence-cluster membership, so a large fraction of cross-cluster pairs will always superpose closely by chance. Absolute Cα-RMSD, unconditioned on sequence identity, does not provide a sparse, duplicate-like redundancy signal for this population at any threshold tested, and this candidate edge type is therefore not added to the leakage graph.

CHIMERA-Bench reports its final dataset spans 2,721 unique PDB identifiers across 2,922 complexes (a $1.074\times$ ratio), reflecting how much redundancy their content-based (MMseqs2) deduplication already removed. The same statistic on MDLAS's undeduplicated 15,557-entry working population gives 8,247 unique PDB identifiers, a $1.886\times$ ratio (Figure 5), expected since MDLAS does not remove rows and stable across all four CDR-H3 identity tiers, since it does not depend on CDR-H3 clustering. More informative is that 619 of 1,285 components (48.2%) span more than one distinct PDB identifier at the primary 70% tier, a fraction that stays essentially flat across all four tiers checked (40%: 47.6%; 70%: 48.2%; 90%: 48.3%; 95%: 48.4%): roughly half of the working population's leakage-relevant redundancy is cross-PDB regardless of tier, which a provenance-based deduplication rule addresses only indirectly, since it operates on provenance instead of content.

Deduplicating to one representative per exact-duplicate group (Section 4.4) instead of retaining every row, as CHIMERA-Bench does, would remove 70.7% of the working population (15,557 to 4,561 entries) and reduce antigen-cluster concentration substantially (a 23.9% relative Gini reduction, 0.7201 to 0.5479) – exactly the axis MDLAS already controls at the split level – while changing CDR-H3 length and germline diversity only modestly.

A.3.5 Content-based redundancy control, and why rows are not deduplicated

A natural alternative to MDLAS's content-based leakage graph is a PDB-provenance-based rule (e.g. one complex per PDB identifier); this under-corrects for the dominant redundancy mode above – the same antibody–antigen pair re-deposited across many distinct PDB identifiers – while over-correcting against genuinely distinct antibodies solved in the same deposition, since leakage is a content property, not a provenance one. MDLAS applies this at the row level: instead of deduplicating before splitting, as CHIMERA-Bench's 2,922-complex curated dataset does, every entry of the 15,557-entry working population lands in some split (deduplication cost quantified above), since a deduplicated training subset is trivially derivable from the per-entry cluster identifiers already in the schema.

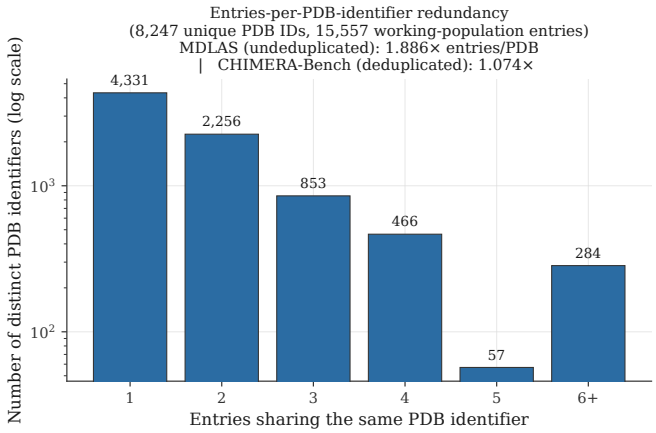

Figure 5: Distribution of entries per PDB identifier over the working population, on a log scale. Most PDB identifiers contribute a single entry, but a long tail contributes several – this shape is exactly the redundancy MDLAS's leakage graph handles at the split-assignment level instead of removing via PDB-level deduplication.

A.4 Extended results

A.4.1 Hyperparameter sensitivity grid

Stage B exposes three tunable parameters: the `antigen_class` tie-breaking weight (reported default $4\times$), the fill-ratio bucket size (reported default 5%), and the oversized-component threshold (reported default 50%). To check that the reported configuration was not narrowly tuned, we swept the weight and bucket size jointly against two oversized thresholds (10%, 30%) – 4 weight settings (unweighted, $2\times$, $4\times$, $8\times$) $\times$ 3 bucket sizes (2%, 5%, 10%) $\times$ 2 oversized thresholds, a 24-configuration grid, over the same primary-tier leakage graph, production seed, and every other parameter (including the low-resolution exclusion, Section 4.5) at its reported value.

Every configuration passed zero-leakage verification identically, confirming Criterion 1 is invariant to Stage B tuning by construction. Unweighted mean Jensen–Shannon divergence across all seven axes (validation and test combined) ranged from 0.0439 to 0.0466 over the full 24-configuration grid (Table 6). The oversized threshold (10% vs. 30%) has exactly zero effect at every one of the 12 weight/bucket combinations tested: with the low-resolution exclusion now active, Stage A's forcing is dominated by the 117 low-resolution-only components (1,885 entries) instead of by size alone, and no component in this dataset falls between the two oversized thresholds while also being full-confidence, so varying this parameter no longer changes which components reach Stage B.

Table 6: Hyperparameter sensitivity: unweighted mean Jensen–Shannon divergence (bits, val+test, all seven axes), averaged over the 3×2 fill-bucket-size/oversized-threshold sub-grid for each `antigen_class` weight setting (the oversized-threshold axis of this sub-grid is degenerate at 12 of 24 points, see text). Differences across all four settings are small (0.04516–0.04551, a 0.78% relative spread).

`antigen_class` weight	Mean JS across sub-grid (bits)
Unweighted	0.04516
$8\times$	0.04524
$4\times$ (reported default)	0.04549
$2\times$	0.04551

The four weight settings differ negligibly on this grid (0.78% relative range, tighter than the 90%-tier grid's 1.42%), so the sub-grid ranking does not drive the choice of weight. We report $4\times$ as the default because `antigen_class` is independently established (Table 1) as the largest single source of leakage-relevant redundancy in this dataset after exact duplication – weighting it reflects that finding directly, independent of this sensitivity result. Any of the four settings works well at

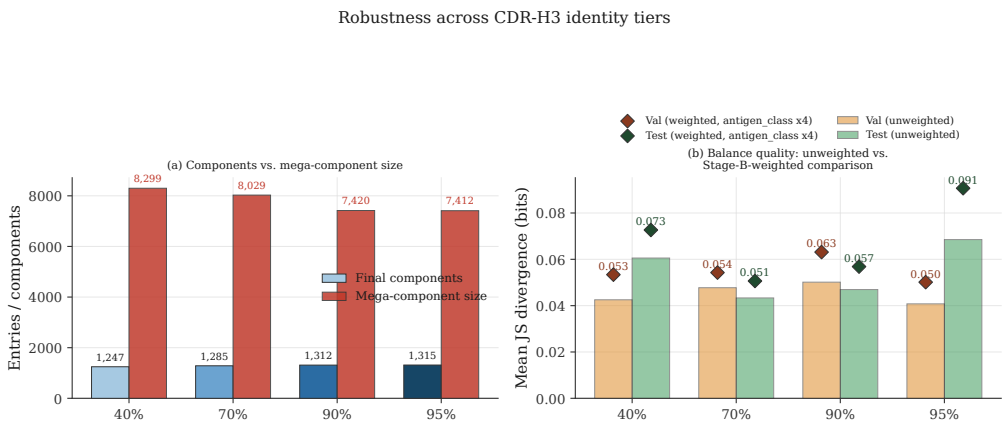

Figure 6: Same data as Table 7. (a) Looser identity thresholds produce fewer, larger components, and grow the antigen-dominated mega-component further. (b) Balance quality shown two ways: unweighted mean JS (bars) treats all seven axes equally; weighted mean JS (diamond markers) applies Stage B's own antigen_class-$\times 4$ weighting. Both metrics favor 70%; see text.

this tier: all pass zero-leakage verification identically, and a lab that prefers a different point on this comparatively flat tradeoff can adopt any of them directly.

A.4.2 Seed robustness of the balance objective

Stage B's greedy component assignment depends on a random ordering of components (Section 4.5), fixed at seed 42 for the split reported throughout this paper. To characterize how much of the reported balance depends on this specific choice, we reran split assignment at the same hyperparameters (Appendix A.4.1) across 21 seeds: 0–19 plus the production seed. Every seed produced a leakage-clean split. Mean unweighted JS divergence (val+test, averaged across all seven balance axes) was reasonably clustered across seeds: mean 0.04554, standard deviation 0.00420, range [0.03514, 0.05099], with the production seed's own value (0.04554) landing essentially at the distribution's mean. Pairwise Jaccard overlap of validation-set membership across seed pairs averaged 0.161 (std 0.026, range [0.109, 0.248]): which specific entries land in validation varies substantially across seeds, an expected consequence of assigning whole components instead of individual entries via a greedy, order-sensitive algorithm, in which different shuffle orders route different components into validation versus test even as the balance objective consistently steers the aggregate composition toward a similar target. The balance conclusion in Section 5.2 is therefore a property of the configuration instead of of the specific seed. This does not imply that individual component membership is seed-invariant: exact split membership remains dependent on the fixed random seed and should therefore be treated as one reproducible realization of the assignment procedure.

A.4.3 Full multi-tier table and figure

Table 7: Leakage graph and split summary across all four CDR-H3 identity tiers. "Unweighted" is the plain average of the seven per-axis Jensen–Shannon divergences (Table 2); "weighted" applies Stage B's own tie-breaking weights (antigen_class $\times 4$, all others $\times 1$) before averaging – the metric the split-assignment algorithm actually optimizes for. "Avg" is the mean of the val and test columns. All four splits pass zero-leakage verification on all five edge-defining keys and are built from independently reconstructed leakage graphs.

					Unweighted			Weighted (antigen_class $\times 4$)		
Tier	Components	Train	Val	Test	Val	Test	Avg	Val	Test	Avg
40%	1,247	12,445	1,556	1,556	0.0425	0.0606	0.0516	0.0534	0.0727	0.0630
70% (primary)	1,285	12,445	1,556	1,556	0.0477	0.0433	**0.0455**	0.0543	0.0506	**0.0525**
90%	1,312	12,445	1,556	1,556	0.0501	0.0469	0.0485	0.0631	0.0569	0.0600
95%	1,315	12,445	1,556	1,556	0.0408	0.0685	0.0546	0.0501	0.0907	0.0704

Final component count increases monotonically with identity tier (1,247 at 40% up to 1,315 at 95%), the expected direction: a looser identity threshold groups more sequences into the same cluster, producing fewer, larger components. The antigen-dominated mega-component (Section A.3.2) shrinks monotonically as the tier tightens – 8,299 entries (53.3% of the working population) at 40%, 8,029 (51.6%) at 70%, 7,420 (47.7%) at 90%, and 7,412 (47.6%) at 95% – the same direction as the component-count trend: looser identity thresholds sweep more entries together via CDR-H3 similarity alone, prior to antigen-cluster membership (sequence-cluster edges are applied before antigen-cluster edges, Section 4.4).

Table 7 reports both the unweighted and weighted balance score at all four CDR-H3 identity tiers. The 70% tier achieves the lowest score on both metrics – unweighted 0.0455 and weighted 0.0525. A 21-seed robustness sweep and a 24-configuration hyperparameter sensitivity grid (Appendix A.4.1), each run at both 70% and 90% for direct comparison, support this result: the unweighted- and weighted-JS rankings both hold on average across seeds (unweighted mean 0.0455 vs. 0.0477; weighted mean 0.0774 vs. 0.0825), and 70%'s hyperparameter-grid relative spread (0.78%) is tighter than 90%'s (1.42%), indicating comparable or better robustness to configuration choice at 70%.

A.5 Full pairwise significance testing across baseline metrics

Section 6.2 highlights specific pairwise comparisons in prose (MEAN's sequence-recovery lead, and the AgForce–RefineGNN binding-accuracy gap); this appendix reports the same paired-significance treatment applied uniformly to every model pair on every non-novelty metric in Tables 3 and 4, not only the comparisons discussed in the main text. For each of the 10 metrics (novelty excluded, since it has no best-value convention) and each of the $\binom{4}{2} = 6$ model pairs, we run a paired Wilcoxon signed-rank test on the common test-complex set shared by all four baselines (1,514 of 1,556 complexes; Appendix A.12 diagnoses the small per-model coverage gap), computing the matched-pairs rank-biserial correlation directly from signed rank sums (sign convention: positive means the first-named model's raw metric value is higher; for RMSD and iRMSD, where lower is better, a positive value therefore indicates the first-named model performing *worse* – e.g. iRMSD R–A is $+0.62$ in Table 8 because RefineGNN's raw iRMSD, 1.689, exceeds AgForce's, 1.306, which favors AgForce, not RefineGNN) instead of from the unsigned statistic `scipy.stats.wilcoxon` returns. Holm–Bonferroni correction is applied across all 60 tests jointly. These row-level tests do not account for dependence among entries within the same leakage component; multiplicity correction does not address that dependence.

56 of the 60 tests are significant at $\alpha = 0.05$ after correction (worst-case $p_{\text{holm}} = 0.0154$; smallest significant effect size $|r| = 0.10$). The four non-significant pairs are MEAN vs. RefineGNN on Fnat, Epitope-F1, and Chimera-B, and MEAN vs. DiffAb on Fnat – in each case among the closest-scoring pairs on that specific metric, with effect sizes near zero ($|r| \leq 0.07$) instead of a large true difference the test lacked power to detect.

Table 8: Signed rank-biserial effect size r for every model pair, every metric in Tables 3–4. Column pairs: M=MEAN, D=DiffAb, R=RefineGNN, A=AgForce; e.g. "M–D" is MEAN vs. DiffAb. Sign follows the raw metric value: positive means the first-named model scores numerically higher, which favors it on higher-is-better metrics (AAR, CAAR, TM-score, Fnat, DockQ, Epitope-F1, Chimera-S, Chimera-B) and against it on the two lower-is-better metrics (RMSD, iRMSD). *: significant after Holm correction across all 60 tests jointly ($\alpha = 0.05$).

Metric	M–D	M–R	M–A	D–R	D–A	R–A
AAR	+0.91*	+0.84*	+0.21*	-0.30*	-0.81*	-0.73*
CAAR	+0.41*	+0.50*	+0.13*	+0.15*	-0.28*	-0.37*
RMSD	-0.59*	-1.00*	+0.53*	-0.97*	+0.79*	+1.00*
TM-score	+0.47*	+0.98*	-0.70*	+0.83*	-0.77*	-0.99*
Fnat	+0.05	-0.07	-0.39*	-0.10*	-0.32*	-0.23*
iRMSD	-0.74*	-0.41*	+0.40*	+0.37*	+0.86*	+0.62*
DockQ	+0.43*	+0.25*	-0.42*	-0.18*	-0.69*	-0.54*
Epitope-F1	+0.38*	-0.04	-0.36*	-0.44*	-0.62*	-0.27*
Chimera-S	+0.52*	+0.91*	-0.60*	+0.62*	-0.79*	-0.96*
Chimera-B	+0.28*	-0.05	-0.37*	-0.31*	-0.56*	-0.26*

Two patterns beyond the main-text comparisons are visible directly in Table 8. First, MEAN's sequence-recovery lead is significant against all three other baselines (M–D, M–R, M–A all positive on AAR and CAAR), but MEAN itself trails AgForce, significantly, on every Table 4 metric (M–A negative on all five higher-is-better rows and positive on iRMSD, both directions favoring AgForce under the sign convention above) – the same "leads on sequence, trails on binding" pattern the main text reports for MEAN specifically now holds with the same statistical treatment as the AgForce–RefineGNN comparison. Second, AgForce is the significant winner (or tied-winner) of every pairwise comparison on every binding-interface and composite metric, against all three other baselines simultaneously – the broadest pattern in the table, and the strongest quantitative support for MDLAS's split producing a discriminative instead of degenerate baseline ranking. These row-level tests treat the 1,514 common test complexes as independent, which residual within-test component structure (Appendix A.13) violates; Appendix A.6 reruns every comparison at the leakage-component level and finds the same conclusions hold.

A.6 Component-level reanalysis: are the 1,514 test complexes independent replicates?

The 1,514 common test complexes are not independent replicates: complexes sharing a leakage-graph component (Section 4.4) are not independent draws, and within-test component redundancy is substantial (Appendix A.13). We recompute every Table 8 comparison two ways that respect this: a macro mean, where each of the 324 leakage components represented among the common test complexes contributes its within-component mean once regardless of member count, and a cluster bootstrap (10,000 resamples of whole components, with replacement) giving a 95% CI on that macro mean.

Table 9: Component-level macro mean difference for every metric × pair comparison in Table 8, each leakage component contributing once. †: bootstrap 95% CI (10,000 resamples, 324 components; 316 for iRMSD, Section A.6) crosses zero. Column pairs as in Table 8.

Metric	M–D	M–R	M–A	D–R	D–A	R–A
AAR	+0.156	+0.121	+0.028	-0.035	-0.128	-0.093
CAAR	+0.072	+0.089	+0.041	+0.017†	-0.031	-0.048
RMSD	-0.504	-1.947	+0.180	-1.443	+0.685	+2.128
TM-score	+0.030	+0.134	-0.065	+0.104	-0.095	-0.199
Fnat	-0.013†	-0.005†	-0.067	+0.008†	-0.054	-0.062
iRMSD	-0.633	-0.240	+0.155	+0.393	+0.788	+0.395
DockQ	+0.044	+0.025	-0.040	-0.019	-0.084	-0.065
Epitope-F1	+0.089	-0.006†	-0.047	-0.095	-0.136	-0.041
Chimera-S	+0.034	+0.091	-0.035	+0.056	-0.069	-0.125
Chimera-B	+0.041	-0.005†	-0.044	-0.045	-0.084	-0.039

Conclusions are unchanged: of 60 comparisons, 54 retain a bootstrap 95% CI excluding zero, and the 6 that cross zero – CAAR D–R, FNAT M–D/M–R/D–R, EPITOPE-F1 M–R, CHIMERA-B M–R – are exactly the near-tied pairs Table 8's row-level test already flags as small-effect ($|r| \leq 0.07$) or non-significant. The iRMSD row draws on 316 of the 324 components instead of all 324: a small number of complexes have no defined interface RMSD (no resolved interface contacts to superpose on) for at least one of the two models being compared. Full bootstrap CIs and pooled (non-macro) means are released alongside the evaluation package (Data and Code Availability).

A.7 Robustness of the AgForce–RefineGNN binding-accuracy gap across antigen classes

Section 6.2 and Appendix A.5 report AgForce's binding-interface advantage over the pooled 1,556-complex test split; a pooled result can in principle be driven by a single dominant class instead of holding generally. We re-run the same paired-significance treatment (Wilcoxon signed-rank, rank-biserial effect size, Holm correction) separately within each of the test split's eight antigen classes, correcting each class's own family of tests (6 metrics × 3 opposing baselines = 18 tests against AgForce, or 36 against all six pairs) independently instead of jointly across classes, since each class

is its own question. Classes with fewer than 15 common complexes are skipped instead of tested underpowered; only `nucleic_acid_or_carbohydrate` ($n = 5$) falls below this bar.

Table 10: AgForce vs. each other baseline on the six binding-interface/composite metrics, significant-metric count (of 6) per antigen class after within-class Holm correction. n: complexes in that class (union across models); test counts per metric vary slightly with per-model coverage.

Antigen class	n	vs. MEAN	vs. DiffAb	vs. RefineGNN
Other protein	1,076	6/6	6/6	6/6
Viral	226	6/6	6/6	6/6
Enzyme	79	2/6	6/6	6/6
Immune receptor	86	4/6	5/6	2/6
Bacterial	49	0/6	2/6	3/6
Cancer-associated	16	0/6	3/6	1/6
Unknown/other	19	0/6	0/6	1/6

The full 6-of-6 sweep against all three other baselines simultaneously – the pooled-test-set pattern Appendix A.5 reports – holds cleanly within the two largest classes alone (other protein and viral, together 1,302 of 1,556 complexes, 83.7% of the test set). Enzyme ($n = 79$) sweeps DiffAb and RefineGNN fully but reaches significance against MEAN specifically on only 2 of 6 metrics, consistent with MEAN and AgForce scoring closer to each other on this class than elsewhere. The four smallest classes (immune receptor, bacterial, cancer-associated, unknown/other – 170 complexes combined, 10.9% of the test set) lose significance on most metrics under the same strict per-class correction; effect direction still favors AgForce throughout except a handful of near-zero, non-significant exceptions confined to classes under 50 complexes (e.g. iRMSD vs. RefineGNN in cancer-associated, chimera-B vs. RefineGNN in unknown/other). This pattern is consistent with reduced statistical power at these sample sizes instead of a genuine per-class contradiction of the pooled finding.

A.8 Replicate structures within a component: a systematic-vs-stochastic failure distinction

Section 6.2 traces DiffAb's elevated RMSD/iRMSD variance to a single complex, `9g13_GNA_agH_m0`. A systematic scan across every metric (not only RMSD) surfaces a related pattern on the same PDB deposition affecting a different model. PDB 9g13 contributes exactly four entries to the working population – four chain-copies of a single-domain camelid nanobody (no light chain) bound to isoform tau-f of microtubule-associated protein tau, an intrinsically disordered antigen resolvable by crystallography only via nanobody stabilization – and the leakage graph places all four in one small, otherwise-isolated component, per `component_assignment.json`; the atomic-component invariant (Section 4.5) accordingly keeps all four together in the test split.

Table 11: RMSD (Å) on all four replicate copies of PDB 9g13, one row per model. Each model's own dataset-wide median RMSD is shown for reference.

Model	_ANA_agB	_CNA_agD	_ENA_agF	_GNA_agH	Own median (all test)
MEAN	2.602	2.505	2.654	2.794	1.617
DiffAb	3.220	3.730	3.478	**74.267**	2.030
RefineGNN	13.233	13.484	13.087	12.762	3.600
AgForce	3.590	3.823	3.936	3.144	1.426

The four models exhibit distinct failure profiles on this replicated component. RefineGNN's RMSD is elevated on all four copies alike (12.8–13.5 Å, $\sim$3.6$\times$ its own dataset-wide median) – a consistent, architecture-level weakness on this target instead of a one-off. DiffAb's three non-`GNA` copies are only moderately elevated (3.2–3.7 Å, $\sim$1.6–1.8$\times$ its own median, comparable to MEAN's own 1.6–1.7$\times$ elevation on this same component) but the fourth collapses completely (74.3 Å) – a discrete, stochastic failure instead of a consistent one. AgForce sits in between ($\sim$2.2–2.8$\times$ its own median across all four) but shows no single-copy collapse. The distinction is visible here because MDLAS retains naturally occurring near-duplicate structures within the same atomic split component instead of deduplicating them away. The resulting replicated cases therefore expose whether an observed failure is consistent across related structures or reflects a more isolated stochastic event.

A.9 BASELINE EVALUATION AT THE 90%-TIER SPLIT: DOES TIER CHOICE CHANGE THE PICTURE?

Section 5.4 shows the split's own leakage-freedom and balance properties hold across all four CDR-H3 identity tiers; this asks the same question of the baseline results in Section 6.2. All four baselines are retrained and re-evaluated, identically to Section 6.1, against the 90%-tier split's own train/val/test sets – a genuinely different 1,556-complex test population from the primary tier's (only 409 of 1,556 complexes, 26.3%, overlap between the two, despite both landing at the same size by construction, Table 7).

Table 12: AAR and DockQ, all four baselines, primary (70%) tier vs. the 90% tier, same evaluation protocol (Section 6.1) applied to each tier's own test split.

Model	AAR (70%)	AAR (90%)	DockQ (70%)	DockQ (90%)
MEAN	0.359	0.372	0.708	0.711
DiffAb	0.206	0.212	0.667	0.643
RefineGNN	0.240	0.180	0.686	0.718
AgForce	0.339	0.333	0.740	0.751

Both of Section 6.2's headline findings hold at both tiers. MEAN leads AAR significantly against all three other baselines at 90% too (all $p_{\text{holm}} < 10^{-15}$), same as at 70%. AgForce sweeps RefineGNN on all six binding-interface and composite metrics at 90% as well (all significant, $p_{\text{holm}} < 10^{-5}$) – the same ordering Section 6.2 reports at 70%; the DockQ win rate specifically is 61.5% at 90% vs. 71.7% at 70%, and the common-test-complex set for pairwise testing is 1,527 of 1,556 at 90% vs. 1,514 of 1,556 at 70% (57 of 60 pairwise tests significant at 90%, 56 of 60 at 70%) – consistently strong, though not numerically identical, at either tier.

Not every metric is this stable individually: RefineGNN's own AAR moves from 0.180 at 90% to 0.240 at 70% – the difference between being the weakest sequence-recovery baseline of the four (90%) and not (70%, where DiffAb is weakest instead, Section 6.2). This is consistent with Section 5.4's finding that the split's own composition is not identical across tiers (different specific complexes, same target sizes): a baseline's absolute standing on any one metric can shift with the specific 1,556-complex population it is scored against, even while the qualitative pattern – which model wins, and by roughly how much – stays stable.

A.10 LIGHT-CHAIN EFFECT: FULL BREAKDOWN

Section 9 reports that all four baselines score significantly worse on the 510 no-light-chain test complexes, on AAR, DockQ, Fnat, and Chimera-B alike. Figure 7 shows this directly, per model, per metric. Preprocessing of the missing chain differs across baselines: MEAN, DiffAb, and RefineGNN omit it entirely; AgForce inserts a placeholder token at the coordinate origin instead. Either way, the model receives genuinely shorter input, not a fabricated chain – and the direction of the effect (worse without a light chain) is the same for all four regardless of which convention they use.

A.11 COMPLEX-LEVEL DISCRIMINATIVENESS: DO THE BASELINES FAIL ON THE SAME COMPLEXES?

Section 6.2's aggregate comparisons establish that the four baselines separate cleanly on mean performance; a split could still fail to be discriminative at the complex level if the same complexes were uniformly easy or hard for every model, in which case per-complex rankings would carry little information beyond one shared difficulty axis. We check this directly: pairwise Spearman correlation of per-complex DockQ across the four baselines, over the 1,514-complex common set (Appendix A.5).

Correlations are moderate (0.46–0.71 across the six pairs) – clearly positive, so complexes hard for one model tend to be at least somewhat harder for others, but far from 1.0, so model identity meaningfully changes which complexes score well. RefineGNN correlates most weakly with the other three (0.46–0.55, versus 0.52–0.71 among MEAN/DiffAb/AgForce), consistent with Section 6.2's antigen-blindness story: RefineGNN's difficulty profile differs from the three antigen-

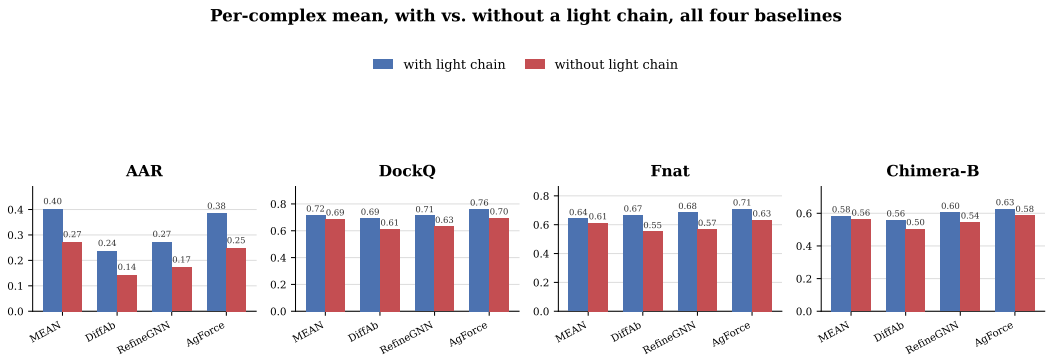

Figure 7: Per-complex mean, with vs. without a light chain, all four baselines, on the four metrics Section 9 reports. The gap is consistent in direction across every model and every metric – no baseline's architecture is spared – though largest in absolute terms for AAR and smallest for DockQ. 84.1% of the no-light-chain subset are camelid single-domain constructs.

Table 13: Spearman correlation of per-complex DockQ between each pair of baselines, common test-complex set ($n = 1{,}514$).

	MEAN	DiffAb	RefineGNN	AgForce
MEAN	1.000	0.607	0.552	0.706
DiffAb	0.607	1.000	0.457	0.620
RefineGNN	0.552	0.457	1.000	0.517
AgForce	0.706	0.620	0.517	1.000

conditioned models specifically because it does not use the signal they do. The same pattern – moderate, RefineGNN-weakest correlations – holds on Fnat and Epitope-F1 as well (correlations 0.39–0.65).

A variance decomposition of the $1{,}514 \times 4$ DockQ matrix attributes 78.3% of total variance to complex identity (a shared "how hard is this complex" component common to all four models) and the remaining 21.7% to model-specific deviation from that shared difficulty (57.3% and 65.8% respectively on Epitope-F1 and Fnat, both still comfortably majority complex-driven). The split carries real, complex-level difficulty structure instead of pure noise, while retaining enough model-specific variance for the pairwise comparisons in Appendix A.5 to be meaningful rather than an artifact of a handful of uniformly hard complexes dominating every model's score alike.

A.12 Coverage-gap diagnosis: why baseline coverage falls short of 1,556

Tables 3–4 and the coverage figures cited throughout Section 6.2 and Appendix A.5 report each baseline against fewer than the full 1,556-complex test split (MEAN 1,547, DiffAb 1,521, RefineGNN 1,551, AgForce 1,555; the union across all four recovers the full 1,556, so every test complex succeeds under at least one baseline). This gap has not previously been diagnosed beyond its size; cross-referencing each model's missing-complex set against the master metadata table isolates three distinct, model-specific root causes instead of one shared one.

Table 14: Tier-70 test-split coverage and the dominant diagnosed cause of each baseline's gap.

Model	Present	Missing	Dominant cause
MEAN	1,547	9	shared-PDB failure (6 of 9, also missing in DiffAb)
DiffAb	1,521	35	scFv format (18), shared-PDB (6), cryo-EM (8 of remaining 11)
RefineGNN	1,551	5	light-chain-absent input (5 of 5)
AgForce	1,555	1	no pattern (single complex)

RefineGNN fails exclusively on light-chain-absent complexes. All five of RefineGNN's missing complexes lack a light chain, against a 32.8% base rate for light-chain-absent complexes in the test split (Section 9) – a hypergeometric enrichment with $p = 0.0037$. Four of the five are separate chain-copies of a single deposition (`5o0w`). This is consistent with a preprocessing step that assumes a paired heavy/light input frame (Section 6.1) and does not degrade gracefully when a single-domain construct supplies only one chain.

MEAN and DiffAb share a common failure, not two independent ones. Six of MEAN's nine missing complexes are also missing from DiffAb – against an expected overlap of 0.2 under independence, this co-occurrence has $p = 6.7 \times 10^{-9}$ (hypergeometric). Both gaps also cluster by PDB deposition well above the test split's own baseline redundancy (1.74 entries per PDB identifier, computed directly over the 1,556-complex union): 67% of MEAN's missing entries and 63% of DiffAb's belong to a deposition contributing two or more missing chain-copies at once (e.g. `7nvo`, `8hgi`), not failing independently. This is the signature of a shared upstream structural-file issue at specific depositions, not two unrelated per-model quirks.

DiffAb's remaining gap is dominated by a single input format. Excluding the six complexes shared with MEAN leaves 29 DiffAb-only failures. Eighteen of these (62.1%) are scFv-format antibodies – heavy and light chain fused into one polypeptide – against a 2.8% population base rate ($p = 7.4 \times 10^{-23}$), and this signal is entirely DiffAb-specific: no scFv complex appears in any other baseline's missing set. The remaining 11 are enriched for cryo-EM depositions specifically (8 of 11, 72.7%, vs. 29.3% among non-scFv test complexes; $p = 0.0035$), though resolution itself only trends in the same direction (median 3.23 Å vs. 2.76 Å) without reaching significance at this sample size ($p = 0.114$).

AgForce's single missing complex carries none of these signals and is treated as noise.

A.13 Residual component redundancy and evaluation difficulty

Zero cross-split leakage (Criterion 1) does not by itself imply that all test complexes have comparable levels of within-test redundancy: components below the oversized threshold (Section 4.5) still range from singletons up to 84 members (Section 5.2), and every member of a given component – being an indivisible split-assignment unit – lands in the same split together. This appendix asks whether baselines score systematically better on test complexes that have more same-component siblings, using each test complex's full leakage-graph component size (Section 4.4) as the redundancy measure.

Table 15: Relative lift in binding-interface/composite metrics for test complexes in the largest-component quartile (component size 18–84, $n \approx 388$ per model) vs. the remaining three quartiles ($n \approx 1{,}150$–$1{,}166$), tier-70. All 12 comparisons significant after Holm–Bonferroni correction across the full table (one-sided Mann–Whitney; worst $p_{\text{holm}} = 0.0044$).

Model	DockQ	Fnat	Chimera-B
MEAN	+5.8%	+14.0%	+9.9%
DiffAb	+5.3%	+5.8%	+10.5%
RefineGNN	+11.2%	+25.9%	+15.1%
AgForce	+3.2%	+6.0%	+5.8%

The lift is consistent in direction and significant for all four baselines on all three binding-interface/composite metrics (Table 15). It is also specific to binding accuracy instead of general: the same top-quartile-vs-rest comparison on AAR shows no consistent effect (+1.1% to +7.3%, none surviving Holm correction across the four models, worst $p_{\text{holm}} = 0.06$) – sequence recovery is tied to CDR-H3 identity, which the leakage graph's primary edge type already controls directly (Section 4.4), while binding-interface accuracy is not itself a leakage-graph redundancy signal and so retains a residual association with it.

The size of the effect tracks how the four baselines are already known to differ. RefineGNN – the one baseline conditioned on no antigen information at all (Section 6.1) – shows the largest lift on every metric (+11.2% to +25.9%), while AgForce – built specifically to counter antigen blindness – shows the smallest (+3.2% to +6.0%). This is consistent with, though does not by itself establish,

the possibility that RefineGNN's binding-interface performance benefits disproportionately from the more redundant portion of the test set, complementing Section 6.2's finding that RefineGNN's binding-interface strength does not survive MDLAS's stricter, antigen-disjoint evaluation as it does on CHIMERA-Bench's own split.