How Much Do Antibody Benchmarks Depend on Related Training Examples? A Cross-Benchmark Audit

Anonymous authors
Paper under double-blind review

Abstract

Machine-learning antibody benchmarks can retain training examples related to antibodies in the test set, obscuring how much reported performance depends on that support. We audit four predictive benchmarks while keeping each test set fixed. Ordinary training O uses the original training pool, family-filtered training F removes examples related to fixed test antibodies under a benchmark-specific rule, and size-matched blind control S removes the same quantity without using family relationships, labels, or outcomes. In AVIDa, O–F is positive in all 160 robustness analyses and median O–S is approximately zero despite no exact mature-VHH recurrence. In AbDesign, scaffold-averaged O–F is 0.142 with a 95% bootstrap interval of [0.038, 0.259]. AlphaSeq shows large O–F values for AbLang2 and ESM-2, while MutAb shows smaller and more variable differences. The median auxiliary distribution-matched control remains closer to O than F in all four predictive benchmarks. In AbDesign, the same direction persists under IMGT CDRH3/CDRL3 and global-alignment relations fixed before evaluation that alter realized family memberships. Between ESM-2 and AbLang2, the preferred model remains unchanged across O, F, and S in three benchmarks and changes only in AlphaSeq. Antibody benchmarks therefore differ substantially in their dependence on related training support.

1 Introduction

Antibody benchmarks such as AVIDa-hIL6 and AlphaSeq are used to compare predictive models across large collections of antibody interactions (Tsuruta et al., 2023; Engelhart et al., 2022). Antibody-specific models including AbLang, AbLang2, and AntiBERTa learn representations from large sequence databases, while IgLM extends these approaches to sequence generation (Olsen et al., 2022; 2024; Leem et al., 2022; Shuai et al., 2023). Benchmark scores are therefore used to judge whether models generalize beyond their training examples.

Different antibody sequences are not necessarily independent. Training and test antibodies can share sequence similarity, parent backgrounds, scaffolds, or mutation neighborhoods without being identical. These relationships also differ across datasets. AVIDa-hIL6 contains VHH interactions across IL-6 contexts, AlphaSeq contains paired antibody interactions, and AbDesign and MutAb evaluate mutations within antibody systems (Tsuruta et al., 2023; Engelhart et al., 2022; Janusz et al., 2025; Chen et al., 2025). A split can therefore contain no exact train–test duplicates while still giving test examples substantial support from related training examples.

Prior work establishes that similarity structure can affect apparent generalization. Protein and drug-target studies report large performance changes as train–test similarity is reduced (Li & Yang, 2017; Zhang et al., 2025). BASE pairs fixed-test similarity filtering with size-matched random removal (Son et al., 2024), while SPECTRA evaluates performance across different levels of cross-split overlap (Ektefaie et al., 2024). Antibody-specific work also shows that training composition can alter predictive performance and learned biological patterns (Ursu et al., 2025). Our contribution is not a new similarity-aware splitting algorithm. Instead, we use a common fixed-test audit across structurally different antibody benchmarks to measure dependence on related supervised training support,

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.

distinguish it from training-set-size effects, test robustness to matched controls and alternative relatedness definitions, and evaluate score sensitivity separately from model-selection sensitivity.

Three questions drive the audit. First, we measure how much benchmark performance changes when related training examples are removed. Second, we ask how much of that change is reproduced by removing the same amount of data without using relatedness. Third, we test whether these interventions change which model a benchmark would select. We compare ordinary training O, family-filtered training F, and size-matched blind control S, then test whether the pattern persists under distribution-matched controls and alternative family definitions.

AVIDa, AbDesign, and AlphaSeq show large O–F differences while their primary S controls remain near O. MutAb shows a smaller and more variable pattern, and Ginkgo provides a structural boundary with no removable family-related support at the audited threshold. Distribution matching does not remove the qualitative pattern, and alternative relations fixed before evaluation preserve the AbDesign result. Model choice between AbLang2 and ESM-2 remains stable in three of four predictive benchmarks. The audit therefore separates related-support sensitivity, training-set-size sensitivity, and model-selection sensitivity.

2 Related Work

Biological benchmarks can overestimate generalization when related examples appear across training and test sets. Evolutionary sequence families have been used to construct benchmark splits with reduced cross-split relatedness (Petti & Eddy, 2022), while GraphPart and DataSAIL use similarity-aware partitioning for the same goal (Teufel et al., 2023; Joeres et al., 2025). Related train–test dependencies have also been shown to inflate performance in protein–ligand, protein-interaction, and drug-target benchmarks (Li & Yang, 2017; Bernett et al., 2024; Bushuiev et al., 2024; Graber et al., 2025; Zhang et al., 2025).

BASE is closest to our primary O/F/S contrast because it keeps the test set fixed, removes increasingly similar training examples, and compares the result with size-matched random removal (Son et al., 2024). SPECTRA instead evaluates performance across different levels of cross-split overlap (Ektefaie et al., 2024). GraphPart, DataSAIL, and SpanSeq construct or support similarity-aware partitions prospectively (Teufel et al., 2023; Joeres et al., 2025; Ferrer Florensa et al., 2024). NbBench standardizes nanobody evaluation across multiple tasks and model classes (Zhang & Tsuda, 2025). Our contribution is therefore not the filtering idea itself. We apply a common fixed-test audit across antibody benchmarks with different biological structures, use benchmark-specific family relations, add distribution-matched and alternative-relation robustness analyses, and ask separately whether score changes alter the preferred model.

Antibody benchmarks add structure through sequence families, parent antibodies, mutation neighborhoods, scaffolds, and shared experimental systems. Antibody-specific work has examined sequence, clonotype, paratope, structure, and learned-embedding similarity (Chomicz et al., 2024), and recent affinity work has evaluated sequence-identity-separated datasets (Hummer et al., 2025). AVIDa-hIL6, AlphaSeq, AbDesign, and MutAb-15 capture different forms of this structure (Tsuruta et al., 2023; Engelhart et al., 2022; Janusz et al., 2025; Chen et al., 2025). We compare the effect of removing related support with size-matched removal and evaluate score sensitivity separately from model-selection sensitivity.

3 Audit Design

3.1 Evaluation Design

For every benchmark, the test examples remain unchanged while the retained training examples vary. Relatedness is benchmark-specific because the datasets contain different biological structures, including sequence families, parent systems, scaffolds, and mutation neighborhoods. Section 4 gives the exact rules.

We construct three primary training conditions. **Ordinary training O** uses the original training pool. **Family-filtered training F** removes training examples related to the fixed test examples under the

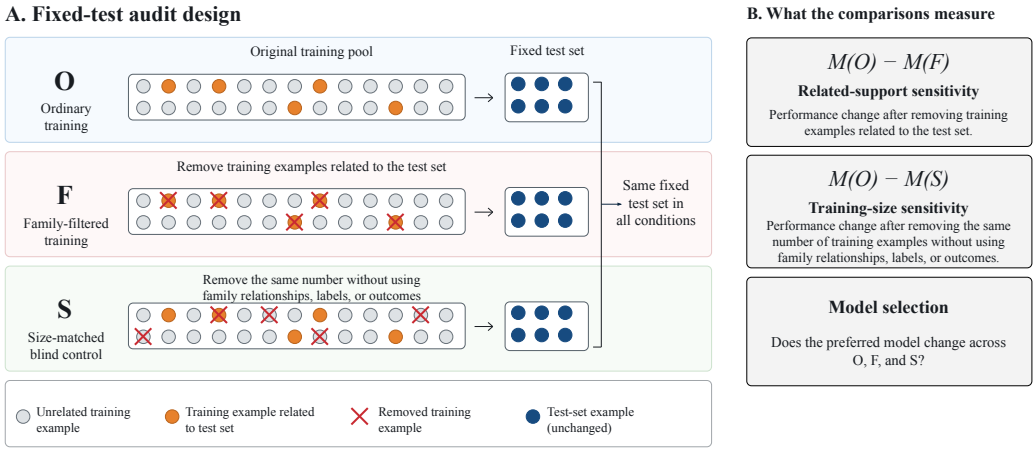


Figure 1: Fixed-test audit design. O uses the original training pool, F removes benchmark-specific related training examples, and S removes the same amount of data without using family relationships, labels, or outcomes. Blue, red, and green row shading denote O, F, and S, respectively. The test set is unchanged. We compare $M(O) - M(F)$, $M(O) - M(S)$, and model selection across conditions.

benchmark-specific relation. **Size-matched blind control S** removes the same number of examples as F without using family relationships, labels, or outcomes. Figure 1 summarizes the design.

Let $M(C)$ denote benchmark performance under condition $C \in \{O, F, S\}$. Each benchmark remains in its original evaluation metric. We define

$$\Delta_{\text{support}} = M(O) - M(F), \tag{1}$$
$$\Delta_{\text{size}} = M(O) - M(S), \tag{2}$$
$$\Delta_{\text{matched}} = M(S) - M(F). \tag{3}$$

Δ_{support} measures the change after removing related support, while Δ_{size} measures the change after removing the same amount of data blindly. We interpret Δ_{matched} diagnostically rather than causally. When $M(O) \approx M(S) > M(F)$, performance is more sensitive to which examples are removed than to the reduction in training size alone.

We also evaluate an auxiliary distribution-matched control S_{strat}. It retains exactly $|F|$ training examples while matching F within predefined non-family strata. AVIDa matches antigen context and training-label prevalence, AbDesign matches prior-exposure class and deterministic outer-half assignment, AlphaSeq matches replicate-availability pattern while excluding assay identity, and MutAb matches mutated chain. The auxiliary analysis uses one prespecified learner and analysis coordinate per benchmark rather than every robustness model and coordinate. These controls test whether obvious non-family distribution shifts induced by F reproduce the O–F gap. They do not replace the primary family-, label-, and outcome-blind S.

We do not numerically pool effects across benchmarks because AUROC and correlation-based metrics are measured on different scales. Cross-benchmark comparisons focus on changes within each benchmark.

3.2 Model-Selection Sensitivity

For each training condition, the preferred model is the model with the highest benchmark score. We restrict the common cross-benchmark comparison to AbLang2 and ESM-2 because both are evaluated across all four predictive benchmarks (Olsen et al., 2024; Lin et al., 2023). The same model pair is used under O, F, and S.

Table 1: Benchmark-specific evaluation setup. F removed is the fraction of audited training membership removed by family filtering at the primary relation. S is exactly size matched to F in the listed counting unit. Ginkgo is structural only.

Benchmark	Task	Metric and unit	Primary relation	S unit	F removed
AVIDa-hIL6	VHH binding	Context-macro AUROC, context	Equal-length mature VHHs with positional identity ≥ 0.90	Rows	47.1%
AbDesign	Mutation ranking	Spearman, scaffold	Paired VH and VL positional identity ≥ 0.90	Records	3.7%
AlphaSeq	Interaction ranking	Spearman, source lineage	Frozen paired-sequence lineage components at 0.90	Entities	28.6%
MutAb-15	Affinity-change ranking	Spearman, parent system	Paired VH and VL minimum-chain identity ≥ 0.90	Entities	8.3%
Ginkgo GDPa1	Structural boundary	No predictive metric	No cross-split family relation at 0.90	Not applicable	0.0%

3.3 Statistical Analysis

Statistical analysis is conducted at the natural unit of each benchmark rather than treating every measurement as an independent replicate. When confidence intervals are reported, resampling is performed at the corresponding benchmark-level unit using bootstrap methods (Efron & Tibshirani, 1993). When multiple related hypotheses are tested, p-values are adjusted using Holm's step-down procedure where appropriate (Holm, 1979). AVIDa robustness analyses reuse the same underlying benchmark data and are treated as dependent evaluations rather than independent replications.

4 Benchmarks and Experimental Setup

We evaluate AVIDa-hIL6, AbDesign, AlphaSeq, and MutAb-15 as predictive benchmarks and Ginkgo GDPa1 as a structural boundary case. O, F, and S use identical test examples within each predictive benchmark. Table 1 summarizes the primary setup and the fraction of audited training membership removed by F.

4.1 AVIDa-hIL6

AVIDa-hIL6 contains binary binding measurements between mature VHH sequences and wild-type or mutant IL-6 contexts (Tsuruta et al., 2023). For each deterministic split, the 31 antigen contexts and VHH sequences are partitioned independently. Twenty contexts are assigned to training, five to validation, and six to testing, while VHHs are assigned to training, validation, and testing at 70%, 15%, and 15%. O consists of observations at the intersection of training contexts and training-partitioned VHHs. The same six held-out contexts and test VHHs are used for O, F, and S.

Two mature VHHs are related when they have equal length and at least 0.90 positional identity. Deterministic single-linkage components are formed over the fixed entity universe. F removes every training row whose VHH component is represented among the test VHHs. S retains the same number of rows through sampling that does not use family relationships, labels, predictions, or outcomes.

AUROC is calculated within each of the six held-out IL-6 contexts and averaged equally. We evaluate ten deterministic splits and 20 S controls for each split and threshold. The primary robustness analysis evaluates amino-acid composition and contiguous 2-mer, 3-mer, and 4-mer linear models. AbLang2 and ESM-2 are evaluated separately for the common model-selection comparison.

4.2 AbDesign

AbDesign contains 658 paired VH/VL point-mutation records across 14 antibody–antigen scaffold contexts (Janusz et al., 2025). Within each scaffold, records are divided deterministically into two complementary halves, with each half used once as the test set. The two halves within a scaffold are not treated as independent biological units.

Under the primary relation, VH and VL must each have equal length to the corresponding candidate chain and at least 0.90 positional identity. Pair similarity is the lower identity of the two chains. Deterministic single-linkage components define F, and S retains the same number of records using 20 deterministic blind controls for each test half.

We also evaluate two alternative relations from a grid fixed before those predictions were run. FAM-B uses global alignment over the whole variable domains and combines VH and VL by minimum-chain identity. FAM-C uses IMGT positions 105–117 from CDRH3 and CDRL3 and again combines the two chains by minimum identity. FAM-B changes the full-variable alignment rule, whereas FAM-C restricts relatedness to the antigen-contact-associated CDR3 regions. Results below report FAM-B at 0.80 and FAM-C at 0.90 because these coordinates produce realized F memberships that differ from the primary relation.

Spearman correlation is computed separately for each half using the raw mutant-to-wild-type ELISA ratio. The two correlations are averaged within each scaffold, and the resulting 14 scaffold values are averaged equally. The primary learner uses chain-separated normalized contiguous 3-mer features with L2-regularized Ridge regression. AbLang2 and ESM-2 are evaluated separately for model selection.

4.3 AlphaSeq

We use MIT Lincoln Laboratory AlphaSeq Dataset 1 and restrict the analysis to interactions with `MIT_Target` (Engelhart et al., 2022). Antibodies are represented as ordered VH/VL pairs. Observations corresponding to the same canonical antibody, assay, and replicate are averaged before evaluation.

Four source library labels reduce to three source lineages because AAYL49 and AAYL50 originate from the same shared seed lineage. Each lineage is held out in turn while the test membership remains fixed across O, F, and S. F removes O entities belonging to the frozen 0.90 family component associated with the held-out lineage. S removes the same number of canonical entities without using family membership, assay information, outcomes, or predictions. Twenty S controls are evaluated for each lineage.

The affinity value is negated so that larger values indicate stronger interaction. For model fitting, training targets are centered separately within each assay–replicate context using only examples retained in the current training condition and then averaged equally for each entity. Test outcomes remain uncentered and are used only for evaluation. Spearman correlation is calculated within required assay–replicate cells, averaged within each lineage, and then averaged equally across the three lineages. AlphaSeq uses frozen AbLang2 and ESM-2 representations followed by Ridge regression. The original preregistered count rules were preserved as non-evaluable rather than silently replaced. The reported protocol was locked after harmonizing the historical evaluation semantics, so AlphaSeq is treated as a retrospective sensitivity analysis and receives no formal confidence interval or paired significance test.

4.4 MutAb-15

MutAb-15 contains experimentally measured affinity changes following antibody mutation (Chen et al., 2025). After structural mapping and filtering, the analysis contains 269 canonical paired VH/VL single-mutant entities across 12 parent antibody systems. The experimental $\Delta\Delta G$ value is the continuous target, with positive values corresponding to stronger mutant binding.

Within each parent system, entities are divided deterministically into complementary train and test halves. The training halves across all 12 systems form O. The primary family relation uses equal-length positional identity on the ordered VH/VL pair, with pair similarity given by the lower chain

identity at 0.90. For each held-out parent, F removes the training entities in components represented in that parent's test set. S retains the same number of entities using ten deterministic family-, label-, and outcome-blind controls.

Spearman correlation is computed within each parent system and averaged equally across the 12 systems. AbLang2 is the primary representation and ESM-2 150M is a prespecified robustness representation. Statistical comparisons use exact two-sided paired sign-flip tests over the 12 parent systems with Holm correction for the predefined O–F and S–F comparison family. A frozen global-alignment relation yields exactly the same family memberships as the primary relation and is treated only as implementation robustness.

4.5 Common Pretrained-Model Comparison

We evaluate AbLang2 and ESM-2 150M across all four predictive benchmarks (Olsen et al., 2024; Lin et al., 2023). For paired VH/VL datasets, each chain is embedded separately and the representations are concatenated before fitting the benchmark-specific frozen learner. Neither pretrained model is fine-tuned. Model selection is evaluated independently under O, F, and S using each benchmark's native metric.

4.6 Ginkgo GDPa1

Ginkgo GDPa1 provides a structural boundary case (van Niekerk et al., 2026). We examine a source split containing 246 paired VH/VL entities, with 54 in the held-out fold and 192 available for training. At the 0.90 paired positional relation, none of the 54 held-out entities belongs to a training-represented component. F therefore removes no training examples and $F = O$. We do not construct S or report a predictive O/F/S result.

5 Results

5.1 Dependence on Related Training Support Varies Across Benchmarks

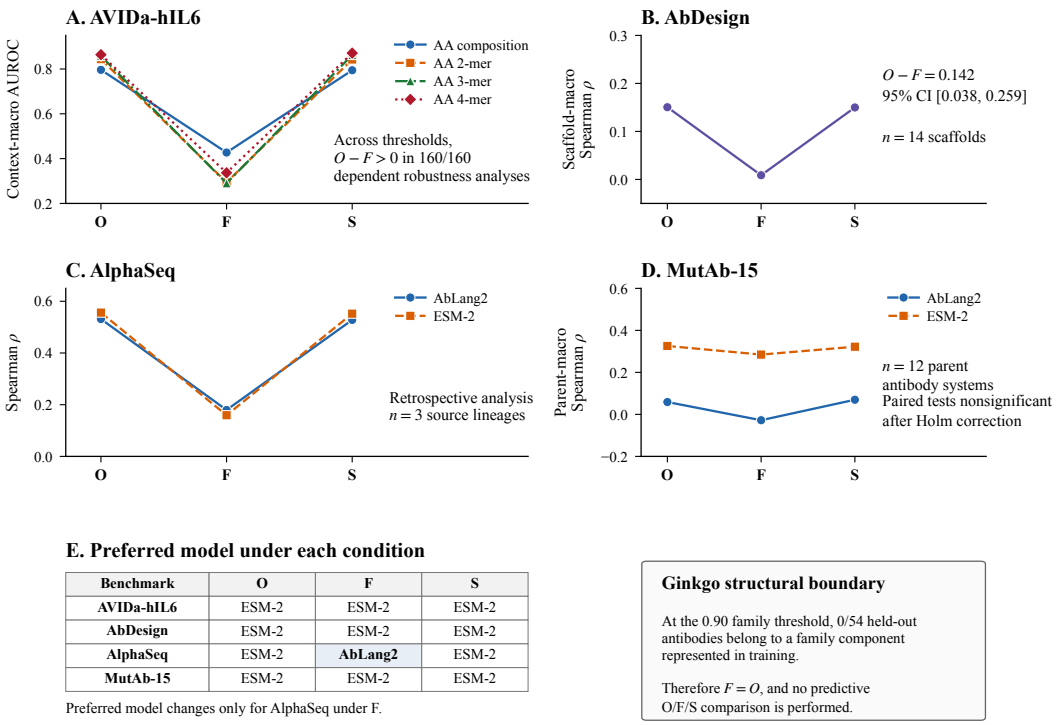
AVIDa, AbDesign, and AlphaSeq show large O–F differences while their size-matched controls remain near O. MutAb shows a smaller and less consistent pattern. Figure 2 reports the primary results in each benchmark's native metric.

AVIDa. Across four family thresholds, ten deterministic splits, and four simple model candidates, O–F is positive in all 160 dependent robustness analyses. At the primary 0.90 threshold, median O–F context-macro AUROC ranges from 0.367 to 0.570 across the four model candidates, while median O–S is approximately zero. The result occurs without exact mature-VHH recurrence between the training and test pools.

AbDesign. Scaffold-averaged Spearman correlation is 0.150 under O, 0.009 under F, and approximately 0.150 under S. O–F is 0.142 with a 95% paired-scaffold bootstrap interval of $[0.038,\ 0.259]$, and O exceeds F in 10 of 14 scaffold aggregates. Inference is performed at the scaffold level rather than treating the 28 held-out halves as independent evidence.

AlphaSeq. For AbLang2, Spearman correlation changes from 0.531 under O to 0.180 under F, while S remains at 0.528. For ESM-2, the corresponding scores are 0.556, 0.159, and 0.552. O–F is therefore 0.352 and 0.397, whereas O–S is 0.003 and 0.004. These results are treated as retrospective sensitivity evidence because only three source lineages are available.

MutAb-15. Under AbLang2, aggregate Spearman correlation is 0.059 under O, -0.028 under F, and 0.070 under S, giving O–F $= 0.087$. Under ESM-2, the corresponding values are 0.326, 0.285, and 0.322, giving O–F $= 0.041$. Parent-level effects vary across the 12 systems, and none of the predefined paired comparisons is statistically significant after correction.



E. Preferred model under each condition

Benchmark	O	F	S
AVIDa-hIL6	ESM-2	ESM-2	ESM-2
AbDesign	ESM-2	ESM-2	ESM-2
AlphaSeq	ESM-2	**AbLang2**	ESM-2
MutAb-15	ESM-2	ESM-2	ESM-2

Preferred model changes only for AlphaSeq under F.

Ginkgo structural boundary

At the 0.90 family threshold, 0/54 held-out antibodies belong to a family component represented in training.

Therefore $F = O$, and no predictive O/F/S comparison is performed.

Axes differ across panels. Scores are shown in each benchmark's native metric and are not numerically pooled across benchmarks.

Figure 2: Benchmark sensitivity to related training support and model selection. (A) AVIDa has $O - F > 0$ in all 160 dependent robustness analyses across thresholds. (B) AbDesign has $O - F = 0.142$ with a 95% CI of $[0.038, 0.259]$. (C) AlphaSeq is retrospective. (D) MutAb-15 paired tests are nonsignificant after Holm correction. (E) Preferred AbLang2 or ESM-2 model under O, F, and S. Model choice changes only for AlphaSeq under F. Ginkgo is a structural boundary case with $F = O$. Axes use each benchmark's native metric.

Table 2: Distribution-matched control using the prespecified S_{strat} analysis for each benchmark. Values use each benchmark's native metric.

Benchmark	O	F	S	S_{strat}
AVIDa-hIL6	0.931	0.595	0.929	0.890
AbDesign	0.150	0.009	0.150	0.141
AlphaSeq	0.531	0.180	0.528	0.529
MutAb-15	0.059	-0.028	0.070	0.066

Ginkgo. At the 0.90 threshold, none of the 54 held-out entities belongs to a family component represented in training, so $F = O$. This is a structural boundary rather than evidence of a predictive null effect.

5.2 Distribution Matching Does Not Reproduce Family Filtering

The primary S control matches only the number of retained training examples. Table 2 reports the prespecified S_{strat} analysis for each benchmark. For AVIDa, this control uses the AbLang2 Ridge analysis rather than the simple-model robustness grid in Figure 2. In all four predictive benchmarks, the aggregate S_{strat} estimate remains closer to O than F and the qualitative conclusion is unchanged.

The AVIDa distribution-matched control moves below ordinary S by 0.040 but remains 0.295 above F. AbDesign and AlphaSeq remain close to O after distribution matching. MutAb has wider replicate variation, but its aggregate S_{strat} remains close to O and above F. These controls therefore do not

support class prevalence, replicate availability, the matched AbDesign prior-exposure and outer-half margins, or mutated-chain composition as sufficient explanations for the primary O–F gaps.

5.3 AbDesign Persists Under Alternative Relatedness Definitions

The AbDesign result is not restricted to the primary whole-variable positional relation. FAM-C at 0.90 changes the component member set for all 658 entities and changes F membership in 21 of 280 frozen evaluation cells. Across 38 of 40 evaluable model-seed units, median O, F, and S Spearman correlations are 0.042, -0.020, and 0.041, and both O and S exceed F in all 38 evaluable units.

FAM-B at 0.80 changes component membership for 122 of 658 entities and changes F membership in 60 of 280 cells. Across 34 of 40 evaluable model-seed units, median O, F, and S are 0.039, -0.027, and 0.033, with O and S above F in all 34 evaluable units. The two relations change different aspects of family construction while preserving the direction of the AbDesign result.

5.4 Score Sensitivity and Model Selection Are Different

Within the common AbLang2–ESM-2 comparison, ESM-2 remains preferred under O, F, and S for AVIDa, AbDesign, and MutAb. AlphaSeq is the only benchmark in which the preferred model changes, with ESM-2 preferred under O and S and AbLang2 under F. Large score changes therefore need not change the ordering of this model pair.

6 Discussion

Dependence on related training support varies across antibody benchmarks. AVIDa, AbDesign, and AlphaSeq show large changes after removing test-related training examples, MutAb shows a smaller and more heterogeneous pattern, and Ginkgo contains no removable support under the audited relation. Reporting only a conventional test score does not reveal where a benchmark lies on this spectrum.

AVIDa shows that exact duplicate removal is insufficient for eliminating dependence on training support in the same single-linkage family components under the evaluated relation. Its O–F gap persists despite the absence of exact mature-VHH matches between the training and test pools. Test sequences can therefore be unseen while still belonging to family components represented in training.

The control analyses narrow two alternative explanations. Ordinary S shows that training-set size alone does not reproduce the largest O–F gaps. The aggregate S_{strat} estimate further matches predefined non-family structure and remains closer to O than F in every predictive benchmark. In AbDesign, CDRH3/CDRL3-focused and global-alignment relations fixed before evaluation alter realized family memberships without changing the direction of the result. The observed dependence is therefore not explained solely by count reduction, the matched non-family margins tested here, or one AbDesign family definition.

Score sensitivity and model-selection sensitivity remain distinct. AVIDa and AbDesign show large score changes while preserving the preferred model in the AbLang2–ESM-2 comparison, whereas AlphaSeq changes model ordering under F. Benchmark reports should therefore distinguish dependence of absolute performance on related training support from changes in which model performs best.

An O/F/S audit complements prospective similarity-aware split design. Prospective partitioning asks how a benchmark should be constructed. The fixed-test audit asks how much an already reported benchmark score depends on related training support. Reporting the family definition, fraction of training support removed, performance after removing it, and an appropriate size-matched reference would make that dependence visible.

7 Limitations

No single relation captures every biologically meaningful connection between antibody examples. Primary analyses use deterministic single-linkage components, which can connect entities through

chains. The AbDesign FAM-B and FAM-C analyses change the relation and realized F memberships but retain component-based filtering, so they do not exhaust alternative linkage rules or structure-based definitions.

In AbDesign, AlphaSeq, and MutAb-15, the primary component structure is closely tied to scaffold, source-lineage, and parent-system organization. Family filtering in these benchmarks should therefore be interpreted as removing training support associated with those benchmark-specific biological or experimental groupings under the stated sequence relation, rather than as isolating sequence similarity from all other system structure. This overlap is part of the benchmark dependence being audited, but it also means that F is not one uniform biological intervention across benchmarks.

The O/F/S and S_{strat} contrasts are diagnostic rather than causal. Family filtering can alter the training distribution in ways not captured by training-set size or the selected strata. The auxiliary strata also differ by benchmark. AVIDa intentionally matches training-label prevalence, while AlphaSeq excludes assay identity because it closely tracks lineage. The distribution-matched analysis uses one prespecified learner and coordinate per benchmark rather than the full robustness grid. The controls therefore test specific alternative explanations rather than defining one universal matched intervention.

Evidence differs in strength across benchmarks. AbDesign is a frozen but narrow replication across 14 scaffolds. AlphaSeq is retrospective with only three source lineages. MutAb inference uses 12 parent systems, its paired tests are nonsignificant, and the distribution-matched replicates are variable. Ginkgo is structural only and should not be interpreted as a predictive null result.

The model-selection analysis covers only AbLang2 and ESM-2 under frozen representations and no fine-tuning, so stability should not be generalized to other encoders or training regimes. The pretrained representations may also contain sequence information related to benchmark antibodies. Similarity to protein-language-model pretraining data can affect downstream performance estimates (Hermann et al., 2024). Our filtering therefore measures sensitivity to supervised benchmark-training support conditional on information already present in the representation.

Because tasks and metrics differ, we do not numerically pool effect sizes. Cross-benchmark conclusions concern within-benchmark sensitivity.

8 Conclusion

Antibody benchmarks differ substantially in their dependence on related training examples. AVIDa, AbDesign, and AlphaSeq show large changes after related support is removed while size-matched blind removal stays much closer to ordinary training. MutAb is smaller and more variable, and Ginkgo has no removable support under the main relation. At the prespecified S_{strat} analysis coordinate, the aggregate distribution-matched result preserves the qualitative pattern in all four predictive benchmarks, and alternative family definitions fixed before evaluation preserve the AbDesign result. These score changes do not necessarily change model selection. The results separate dependence on related training support, sensitivity to training-set size, and sensitivity of model selection.

AI use statement

In this work, we used generative AI tools to provide feedback on research questions, experimental design, and research methodology, assist with data preprocessing and implementation of analysis and evaluation code, and support interpretation of experimental results. We did not use generative AI to generate synthetic datasets, formulate or prove mathematical claims, perform translation, or conduct qualitative or thematic data analysis. Other required-disclosure tasks were either not applicable or are covered by the uses described above.

Additionally, we used generative AI tools to assist and aid writing and editing, literature review and identification of relevant work, and software development and debugging, including assistance with a script used to generate a scientific figure from verified analysis outputs. An AI-based paper-review tool was also used to identify potential weaknesses in the experimental design and reporting.

We reviewed all AI-assisted work. AI-assisted experimental suggestions were evaluated before use, generated or edited code was tested against the intended analysis and frozen artifacts, reported nu-

merical results were verified against checksum-bound source outputs, and AI-assisted literature suggestions and citations were checked against source records. AI-assisted figure code was used for visualization, and reported values were supplied from verified analysis outputs. We take responsibility for the final content of this work, including text, claims, code, analyses, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All reported experiments use deterministic train and test memberships, fixed relatedness definitions, and archived result manifests. The final paper assembly verifies source hashes before reading values and does not run learner, prediction, embedding, or membership-generation code. Appendices A–H document dataset preprocessing, split construction, family definitions, S and S_{strat} sampling, statistical units and tests, software versions, analysis chronology, and core artifact checksums. Figure 2 is generated from checksum-verified data rather than embedded scientific constants. AlphaSeq is labeled retrospective and only corrected S_{strat} v2 is used. AbDesign alternative-family evidence is bound to prospectively frozen artifacts, MutAb FAM-B is treated only as implementation robustness, and Ginkgo is structural only. Missing upstream revisions or runtime versions are reported as unavailable rather than inferred.

REFERENCES

Judith Bernett, David B. Blumenthal, Dominik G. Grimm, Florian Haselbeck, Roman Joeres, Olga V. Kalinina, and Markus List. Guiding questions to avoid data leakage in biological machine learning applications. *Nature Methods*, 21:1444–1453, 2024. doi: 10.1038/s41592-024-02362-y.

Anton Bushuiev, Roman Bushuiev, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, and Josef Sivic. Revealing data leakage in protein interaction benchmarks. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2024. URL https://openreview.net/forum?id=ORMXYUK5IY.

Zhengshan Chen, Song He, Xiangyang Chi, and Xiaochen Bo. Predicting antibody affinity changes upon mutation based on unbound protein structures. *International Journal of Molecular Sciences*, 26(3):1343, 2025. doi: 10.3390/ijms26031343.

Dawid Chomicz, Jarosław Kończak, Sonia Wróbel, Tadeusz Satława, Paweł Dudzic, Bartosz Janusz, Mateusz Tarkowski, Piotr Deszyński, Tomasz Gawłowski, Anna Kostyn, Marek Orłowski, Tomasz Klaus, Lukas Schulte, Kyle Martin, Stephen R. Comeau, and Konrad Krawczyk. Benchmarking antibody clustering methods using sequence, structural, and machine learning similarity measures for antibody discovery applications. *Frontiers in Molecular Biosciences*, 11:1352508, 2024. doi: 10.3389/fmolb.2024.1352508.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York, 1993. ISBN 978-0-412-04231-7.

Yasha Ektefaie, Andrew Shen, Daria Bykova, Maximillian G. Marin, Marinka Zitnik, and Maha Farhat. Evaluating generalizability of artificial intelligence models for molecular datasets. *Nature Machine Intelligence*, 6:1512–1524, 2024. doi: 10.1038/s42256-024-00931-6.

Emily Engelhart, Ryan Emerson, Leslie Shing, Chelsea Lennartz, Daniel Guion, Mary Kelley, Charles Lin, Randolph Lopez, David Younger, and Matthew E. Walsh. A dataset comprised of binding interactions for 104,972 antibodies against a SARS-CoV-2 peptide. *Scientific Data*, 9: 653, 2022. doi: 10.1038/s41597-022-01779-4.

Alfred Ferrer Florensa, José Juan Almagro Armenteros, Henrik Nielsen, Frank Møller Aarestrup, and Philip Thomas Lanken Conradsen Clausen. SpanSeq: similarity-based sequence data splitting method for improved development and assessment of deep learning projects. *NAR Genomics and Bioinformatics*, 6(3):lqae106, 2024. doi: 10.1093/nargab/lqae106.

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.

Leon Hermann, Tobias Fiedler, Hoang An Nguyen, Melania Nowicka, and Jakub M. Bartoszewicz. Beware of data leakage from protein LLM pretraining. In *Proceedings of the 19th Machine Learning in Computational Biology Meeting*, volume 261 of *Proceedings of Machine Learning Research*, pp. 106–116. PMLR, 2024. URL https://proceedings.mlr.press/v261/hermann24a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Alissa M. Hummer, Constantin Schneider, Lewis Chinery, and Charlotte M. Deane. Investigating the volume and diversity of data needed for generalizable antibody–antigen $\Delta\Delta G$ prediction. *Nature Computational Science*, 5:635–647, 2025. doi: 10.1038/s43588-025-00823-8.

Bartosz Janusz, Dawid Chomicz, Samuel Demharter, Marloes Arts, Jurrian de Kanter, Yano Wilke, Helena Britze, Sonia Wrobel, Tomasz Gawłowski, Pawel Dudzic, Kärt Ukkivi, Lauri Peil, Roberto Spreafico, and Konrad Krawczyk. AbDesign: database of point mutants of antibodies with associated structures reveals poor generalization of binding predictions from machine learning models. *mAbs*, 17(1):2567319, 2025. doi: 10.1080/19420862.2025.2567319.

Roman Joeres, David B. Blumenthal, and Olga V. Kalinina. Data splitting to avoid information leakage with DataSAIL. *Nature Communications*, 16:3337, 2025. doi: 10.1038/s41467-025-58606-8.

Jinwoo Leem, Laura S. Mitchell, James H. R. Farmery, Justin Barton, and Jacob D. Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7):100513, 2022. doi: 10.1016/j.patter.2022.100513.

Yang Li and Jianyi Yang. Structural and sequence similarity makes a significant impact on machine-learning-based scoring functions for protein–ligand interactions. *Journal of Chemical Information and Modeling*, 57(4):1007–1012, 2017. doi: 10.1021/acs.jcim.7b00049.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Tobias H. Olsen, Iain H. Moal, and Charlotte M. Deane. AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022. doi: 10.1093/bioadv/vbac046.

Tobias H. Olsen, Iain H. Moal, and Charlotte M. Deane. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics*, 40(11):btae618, 2024. doi: 10.1093/bioinformatics/btae618.

Samantha Petti and Sean R. Eddy. Constructing benchmark test sets for biological sequence analysis using independent set algorithms. *PLOS Computational Biology*, 18(3):e1009492, 2022. doi: 10.1371/journal.pcbi.1009492.

Richard W. Shuai, Jeffrey A. Ruffolo, and Jeffrey J. Gray. IgLM: Infilling language modeling for antibody sequence design. *Cell Systems*, 14(11):979–989.e4, 2023. doi: 10.1016/j.cels.2023.10.001.

Hyojin Son, Sechan Lee, Jaeuk Kim, Haangik Park, Myeong-Ha Hwang, and Gwan-Su Yi. BASE: a web service for providing compound-protein binding affinity prediction datasets with reduced similarity bias. *BMC Bioinformatics*, 25:340, 2024. doi: 10.1186/s12859-024-05968-3.

Felix Teufel, Magnús Halldór Gíslason, José 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.

Hirofumi Tsuruta, Hiroyuki Yamazaki, Ryota Maeda, Ryotaro Tamura, Jennifer Wei, Zelda E. Mariet, Poomarin Phloyphisut, Hidetoshi Shimokawa, Joseph R. Ledsam, Lucy Colwell, and Akihiro Imura. AVIDa-hIL6: A large-scale VHH dataset produced from an immunized alpaca for predicting antigen-antibody interactions. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1823. Datasets and Benchmarks Track.

Eugen Ursu, Aygul Minnegalieva, Puneet Rawat, Maria Chernigovskaya, Robi Tacutu, Geir Kjetil Sandve, Philippe A. Robert, and Victor Greiff. Training data composition determines machine learning generalization and biological rule discovery. *Nature Machine Intelligence*, 7:1206–1219, 2025. doi: 10.1038/s42256-025-01089-5.

Lood van Niekerk, Joshua Moller, Seth Ritter, Porfirio Quintero-Cadena, Rich Cohen, Georgia Channing, Michael Chungyoun, Laura Rand, Alexander Smith, Aanal Bhatt, Yolaine Pierre, Blake Harris, Xiang Ao, Lucia Grippo, Maximilian Schwenk, Adam Rosenbaum, Olga Allen, Nimra Asi, Jiang Zhu, Aviral Singh, Daksh Sammi, Rushikesh Jadhav, Antonín Dušek, Shyam Chandra, Valentin Badea, Nels Thorsteinson, Nathaniel Blalock, Jeonghyeon Kim, Oliver M. Turnbull, Ameya Kulkarni, Vivek Kohar, Netsanet Gebremedhin, Charlotte M. Deane, Peter M. Tessier, and Ammar Arsiwala. 2025 ginkgo datapoints antibody developability competition outcomes: limited model performance and a call for data standardization. *mAbs*, 18(1):2634216, 2026. doi: 10.1080/19420862.2026.2634216.

Chenbin Zhang, Zhiqiang Hu, Chuchu Jiang, Wen Chen, Jie Xu, and Shaoting Zhang. Rethinking the generalization of drug target affinity prediction algorithms via similarity aware evaluation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7172e147d916eef4cb1eb30016ce725f-Abstract-Conference.html.

Yiming Zhang and Koji Tsuda. NbBench: Benchmarking language models for comprehensive nanobody tasks. *Machine Learning: Science and Technology*, 6(4):040502, 2025. doi: 10.1088/2632-2153/ae20ec.

APPENDIX

A ADDITIONAL AUDIT CONSTRUCTION

A.1 DETERMINISTIC CONSTRUCTION OF O, F, AND S

The test membership is fixed within every predictive benchmark. O is the ordinary training membership. F removes O examples whose benchmark-specific family component is represented in the fixed test set. S retains exactly the same number of training examples as F but is sampled without family relationships, labels, outcomes, or predictions. Every stored predictive S membership satisfies $|S| = |F|$.

Family components are deterministic single-linkage connected components over the relevant frozen entity universe. Stable hash-based identifiers are used for split assignment, control sampling, and membership records. No test outcome or model prediction enters component construction or ordinary S sampling.

A.2 AUXILIARY DISTRIBUTION-MATCHED CONTROL

The auxiliary S_{strat} control also retains exactly $|F|$ examples but matches F within declared non-family strata. It is a robustness diagnostic rather than a replacement for the primary blind S. The final v2 membership manifest contains 760 membership records and verifies subset, cardinality, per-stratum quota, uniqueness, and sampling-without-replacement invariants for every planned membership. Existing O, F, and S models were not refit when S_{strat} was evaluated.

Table 3: Distribution-matched control construction. The auxiliary analysis uses one prespecified learner and coordinate per benchmark.

Benchmark	Matching strata	Replicates	Learner
AVIDa-hIL6	Antigen context and training-label prevalence	20	AbLang2 Ridge
AbDesign	Prior-exposure class and deterministic outer-half assignment	20	3-mer Ridge
AlphaSeq	Replicate-availability multiset with assay identity excluded	20	AbLang2 Ridge
MutAb-15	Mutated chain H or L within parent system	10	AbLang2 Ridge

B BENCHMARK IMPLEMENTATION DETAILS

B.1 AVIDA-HIL6

Input sequences are uppercased and restricted to standard amino acids. The 22-residue VHH signal peptide and terminal His tag are removed, as are the 29-residue antigen signal peptide and terminal His tag. Duplicate mature-VHH by antigen-context keys are quarantined. Deterministic split seeds 37037 through 37046 independently partition antigen contexts and VHH entities. Twenty of 31 antigen contexts are assigned to training, five to validation, and six to testing. VHH entities are assigned 70 percent to training, 15 percent to validation, and 15 percent to testing.

The primary FAM-A relation requires equal mature-VHH length and positional identity of at least 0.90. Robustness thresholds are 0.80, 0.85, 0.90, and 0.95. Stage 38 evaluates amino-acid composition and normalized contiguous 2-mer, 3-mer, and 4-mer features with L2 logistic regression using $C = 1$ and the `lbfgs` solver. AUROC is computed separately in each of the six held-out antigen contexts and then averaged equally. The 160 threshold by split by model evaluations are dependent computational configurations rather than independent biological replications.

The S_{strat} analysis uses the prespecified AbLang2 Ridge coordinate at threshold 0.90 and seed 37037. It retains the F quota within antigen context by training-label strata. Training labels are used only for this declared prevalence-matching control.

B.2 AbDesign

The frozen cohort contains 658 ELISA mutant records across 14 antibody-antigen scaffolds. Records must represent one heavy- or light-chain substitution, contain complete variable VH and VL sequences, map to a valid parent, and pass sequence-mutation consistency checks. Each scaffold is divided into deterministic complementary halves. Each half is used once as test, and the two halves are averaged before scaffold-level aggregation.

FAM-A requires equal VH and VL lengths and at least 0.90 positional identity in both chains, with pair similarity defined by the lower chain identity. The primary learner uses chain-separated normalized contiguous 3-mer frequencies and Ridge regression with $\alpha = 1$ and the `lsqr` solver. Training targets are standardized using the retained training data before fitting. The evaluation target is the raw mutant-to-wild-type ELISA ratio. Spearman correlation is computed in each half, averaged within scaffold, and then macro-averaged across 14 scaffolds.

The S_{strat} control matches F quotas within prior-exposure class by deterministic outer-half assignment for each of the 28 scaffold-half cells. Exact scaffold matching is not used because it would reproduce the family-filtered intervention rather than provide an independent control.

B.3 AlphaSeq

The analysis uses MIT Lincoln Laboratory AlphaSeq Dataset 1 and the MIT Target measurements. AAYL49 through AAYL52 are retained. AAYL49 and AAYL50 share one source lineage, leaving three parent lineages. Ordered VH and VL pairs are canonical entities. Duplicate entity by assay by replicate measurements are averaged and negative controls are excluded. Affinity values are negated so that larger values indicate stronger interaction. For each training condition, training targets are centered separately within each assay–replicate context using only retained training observations and then averaged equally for each entity. Test outcomes remain uncentered and are used only for evaluation.

Within each lineage, a label-blind hash ordering defines the fixed test subset. F removes the held-out lineage component from O. Ordinary S removes the same number of canonical entities without family membership, assay information, outcomes, or predictions. Frozen AbLang2 or ESM-2 representations are standardized using training data only before Ridge regression with $\alpha = 1$ and the `lsqr` solver. Spearman correlation is calculated within required assay by replicate cells, averaged within lineage, and then averaged equally across the three lineages.

The original Stage 44B preregistered execution is preserved as non-evaluable. The final harmonized protocol was locked before its reported execution and is treated as retrospective sensitivity evidence. The first S_{strat} membership construction is invalidated because it collapsed a replicate-availability multiset to a set and therefore erased multiplicity such as $(1, 1, 2, 2, 3, 3)$. Only the corrected v2 memberships and results are used. The v2 control matches replicate-availability multiplicity without using assay identity or affinity magnitude.

B.4 MutAb-15

The final MutAb-15 analysis contains 269 paired VH and VL single-mutant entities across 12 parent antibody systems. Canonical structure selection and mutation mapping require the expected chain role, author position, wild-type residue, valid paired numbering, and exactly one variable-domain substitution. Within each parent, a deterministic hash ordering defines complementary train and test halves.

FAM-A uses equal-length whole-variable positional identity with pair similarity given by the lower VH or VL identity at threshold 0.90. AbLang2 is the primary representation and ESM-2 150M is the prespecified representation robustness analysis. Both are frozen. The benchmark learner is Ridge regression with $\alpha = 1$ and `solver=auto`. Spearman correlation is computed within each parent and macro-averaged across the 12 parent systems. Constant predictions score zero.

The S_{strat} control matches F quotas by mutated chain within parent. A frozen global-alignment relation yields exactly the same family memberships and controls as FAM-A and is therefore treated only as implementation robustness.

B.5 GINKGO GDPA1

The structural audit uses 246 paired VH and VL entities from the GDPa1 source split, with 54 in the inherited held-out fold and 192 available for training. At the paired positional identity threshold of 0.90, none of the 54 held-out entities belongs to a component represented in training. F therefore equals O and removes no training entity. No S control, learner, predictive metric, confidence interval, or significance test is defined for Ginkgo.

C TRAINING REMOVAL AND COMPONENT STRUCTURE

Table 4: Aggregate training membership under O and F. Counts for AbDesign, AlphaSeq, and MutAb-15 are membership placements summed across repeated fixed-test fits rather than unique entities. Every stored predictive S control exactly matches the corresponding F size.

Benchmark	O	F	Removed	Removed percent	S
AVIDa-hIL6	151579	80198	71381	47.1	80198
AbDesign	17766	17108	658	3.7	17108
AlphaSeq	201126	143660	57466	28.6	143660
MutAb-15	1644	1507	137	8.3	1507
Ginkgo GDPa1	192	192	0	0.0	–

Table 5: Primary family-component structure in the frozen entity universe.

Benchmark	Entities	Components	Median size	Max size	Singleton percent
AVIDa-hIL6	38099	14516	1	2615	79.9
AbDesign	658	14	46	66	0.0
AlphaSeq	71830	3	25689	32629	0.0
MutAb-15	269	12	25	41	0.0
Ginkgo GDPa1	246	241	1	2	97.9

For AVIDa, the component summary uses the full frozen family-component manifest, including bridge VHH entities used for deterministic component construction. This component universe is larger than the predictive VHH union used for model fitting.

The fraction of training membership removed is not itself treated as an explanatory variable in the main analysis. A post-outcome descriptive check gives Spearman correlations of 0.471 for AbDesign and 0.459 for MutAb-15 between unit-level removal fraction and O minus F under the primary representation. The AVIDa association is undefined because the removal fraction is constant across its six evaluation units. AlphaSeq has only three lineages and is too small for a stable association. No confidence interval, p-value, or causal interpretation is assigned to these exploratory correlations.

D DISTRIBUTION-MATCHED CONTROL RESULTS

Table 6: Full aggregate S_{strat} results. The range is the minimum and maximum aggregate score across the stored distribution-matched replicates.

Benchmark	O	F	S	Median S_{strat}	$S_{\mathrm{strat}} - F$	Range
AVIDa-hIL6	0.931	0.595	0.929	0.890	0.295	[0.879, 0.898]
AbDesign	0.150	0.009	0.150	0.141	0.132	[0.128, 0.151]
AlphaSeq	0.531	0.180	0.528	0.529	0.349	[0.526, 0.532]
MutAb-15	0.059	-0.028	0.070	0.066	0.094	[-0.062, 0.122]

The median S_{strat} estimate remains closer to O than F for every predictive benchmark. AVIDa shows the largest shift from ordinary S to S_{strat}, with O minus S_{strat} equal to 0.041 while $S_{\mathrm{strat}} - F$ remains 0.295. AbDesign and AlphaSeq remain close to O. MutAb-15 has substantial replicate

variation, so the control is interpreted only at the aggregate level and receives no new significance test.

E ALTERNATIVE RELATEDNESS DEFINITIONS

E.1 RELATION DEFINITIONS

AbDesign FAM-B uses Biopython 1.83 global alignment over the full variable domains. Match score is 1, mismatch score is 0, gap-open score is -1, and gap-extension score is -0.1. Identity is exact matches divided by all aligned columns including gaps, and paired similarity is the lower of the VH and VL identities.

AbDesign FAM-C uses ANARCI 2020.04.23 with IMGT positions 105 through 117 including insertion codes. CDRH3 and CDRL3 identities are computed separately with equal-length positional identity and paired similarity is the lower of the two. No V or J restriction is used.

Table 7: AbDesign alternative-relatedness interventions. Component differences compare each alternative with the primary FAM-A relation. F cells changed reports fixed evaluation cells in which the retained F membership differs from FAM-A.

Relation	Threshold	Entities with changed component	F cells changed	Training decisions changed
FAM-C	0.90	658 of 658	21 of 280	569 of 177660
FAM-B	0.80	122 of 658	60 of 280	4880 of 177660

Table 8: Frozen AbDesign predictive results under the two highlighted alternative relations. Values are medians across evaluable model-seed units.

Relation	O	F	S	O minus F	S minus F	Evaluable units
FAM-C 0.90	0.042	-0.020	0.041	0.071	0.070	38 of 40
FAM-B 0.80	0.039	-0.027	0.033	0.063	0.062	34 of 40

O exceeds F and S exceeds F in every evaluable unit for both highlighted relations. FAM-C 0.95 produces no removal and is therefore a structural no-op rather than evidence for or against the predictive effect. MutAb FAM-B is membership-identical to FAM-A and is not counted as an independent biological robustness relation.

F STATISTICAL PROCEDURES AND ANALYSIS CHRONOLOGY

Table 9: Natural statistical units and inferential procedures.

Benchmark	Statistical unit	Procedure
AVIDa-hIL6	Held-out antigen context, $n = 6$	Primary Stage 38 O/F/S matrix is descriptive. Thresholds, seeds, models, and rows are dependent computational configurations.
AbDesign	Scaffold, $n = 14$	Two halves are averaged within scaffold. A paired scaffold bootstrap with 10000 draws gives percentile 95 percent intervals for O minus F and S minus F.
AlphaSeq	Source parent lineage, $n = 3$	Descriptive only. No formal confidence interval or p-value is reported.
MutAb-15	Parent antibody system, $n = 12$	Exact two-sided paired sign-flip tests enumerate all 2^{12} signs for O minus F and median-S minus F. Holm correction covers those two contrasts within each representation and family analysis.
Ginkgo GDPa1	One inherited held-out fold	Structural audit only. No predictive inference is performed.

The evidence classes are preserved throughout the paper. AVIDa Stage 38 is retrospective exploratory frozen. AbDesign Stage 36F is prospectively frozen and treated as a narrow replication. AlphaSeq is retrospective harmonized historical-semantics sensitivity evidence. MutAb-15 is preregistered confirmatory with prespecified robustness. Ginkgo is structural only. The S_{strat} v2 memberships and AbDesign alternative-relation grid were fixed before their reported executions.

G SOFTWARE AND DETERMINISTIC EXECUTION

Recorded frozen environments use Python 3.10.20 with NumPy 1.26.4, SciPy 1.11.4, and scikit-learn 1.7.2 for the analyses where those versions are bound. Pretrained-model analyses use transformers 4.45.2 and frozen AbLang2 or ESM-2 representations. AlphaSeq and MutAb record torch 2.2.0, pandas 2.3.3, tokenizers 0.20.3, huggingface-hub 0.24.7, safetensors 0.4.5, regex 2026.7.19, ablang2 0.1.1, einops 0.8.2, and rotary-embedding-torch 0.6.5. AbDesign alternative relations additionally use Biopython 1.83 and ANARCI 2020.04.23. Exact Stage 38 AVIDa package versions and the Stage 41C Ginkgo runtime are unavailable and are not inferred.

Key deterministic entry points are listed below.

Table 10: Primary reproduction entry points. Paths are repository-relative.

Benchmark	Entry points
AVIDa-hIL6	`scripts/run_stage37b_execution_amendment.py`, `scripts/run_stage38_systematic_robustness.py`, `scripts/run_stage43d_pretrained_robustness.py`
AbDesign	`scripts/run_stage36f_abdesign_replication.py`, `scripts/stage43c_local_run.sh`
AlphaSeq	`data/experiments/alphaseq-harmonized-sensitivity-LOCKED/code/freeze_alphaseq_harmonized.py`, `data/experiments/alphaseq-harmonized-sensitivity-LOCKED/code/run_alphaseq_harmonized.py`
MutAb-15	`scripts/run_mutab15_structural_audit.py`, `scripts/freeze_mutab15_predictive.py`, `scripts/run_mutab15_predictive.py`
Ginkgo GDPa1	`scripts/run_stage41c_structural_audit.py`
All predictive benchmarks	`scripts/run_sstrat_controls.py`, `scripts/build_ofs_training_size_audit.py`
Final reporting	`scripts/assemble_final_paper.py`, `figure2.py`

H FINAL ASSEMBLY AND PROVENANCE

The final reporting pipeline is read-only with respect to scientific evidence. The assembler verifies 30 authoritative input hashes before reading values. It does not fit models, recompute predictions, generate embeddings, or regenerate memberships. Repeated assembly is byte-identical under the fixed timestamp policy. The final evidence ledger contains 153 numeric evidence rows, and the Figure 2 data file contains 52 plotted scientific values or labels. The renderer verifies the Figure 2 data checksum before plotting and contains no embedded scientific result values.

The final assembly excludes AlphaSeq S_{strat} v1, keeps Ginkgo structural only, and excludes MutAb FAM-B from independent alternative-relation evidence. All supplied headline values match the frozen reports with no discrepancies.

The assembler code hash is `78aa5ea099c2944387218756f016b84a9bd14aefe52ef19f55a1009f6795fd19`. The Figure 2 renderer hash is `779ccdc820d445082ea107cf09743fa0875436a9da2801bd7984181f4821f060`. The manifest canonical self-checksum is `bbec6452a2a766b222912fcd53f4c78c690799278a02e0ca9365223746f87ee7`.

Upstream provenance is reported only when frozen evidence exists. The exact AVIDa source release revision, the exact AbDesign upstream archive revision, the Ginkgo repository revision and runtime

Table 11: Core final-assembly artifacts and SHA-256 checksums.

Artifact	SHA-256
ASSEMBLY_MANIFEST.json	ccf4dbe4fc1959f799e3f61b29242c60 f384627cee0668e4cc19d15e9c219d36
final_paper_evidence.csv	8dd8828b82661acaaf141d2e79b2dfd1 96178fb2765fd00c32ceab0f6d568ade
final_table_benchmark_setup.csv	af73c12b04cc7af3eac697f79c7116da d653d0c8634b9834e219a78470757df0
final_table_robustness.csv	ac9bed37b1a79e3d3aff230d99ad8ed3 9383f16d930bc04bb1d10deedb0887a6
final_table_alternative_relations.csv	6ea75179b841e0b2720ca00ecfa5640c 1946c463a8d0bc19374b2a36e6276c18
figure2_data.csv	9a78722f620896c3d89d738e29e12059 feefe9a7a1f3169976a3776239d15976

versions, and possible pretrained-corpus overlap for AbLang2 and ESM-2 remain unavailable or unresolved. These values are not inferred.