Emergent Developability in De Novo Antibody Design: A Computational Analysis of the FLAb Benchmark

Analysis conducted using FLAb dataset and Modal cloud infrastructure

Code repository: `https://github.com/inventcures/flab_gray-lab-jhu_ab_chars`

February 4, 2026

Abstract

Background: Antibody developability—the likelihood that a candidate molecule will succeed through manufacturing and clinical development—is a critical determinant of therapeutic success. Recent de novo antibody design platforms have claimed that their models achieve favorable developability profiles "for free," without explicit optimization. We investigated this claim using the Fitness Landscape for Antibodies (FLAb) benchmark.

Methods: We analyzed sequence-based developability features across four antibody sources: the Structural Antibody Database (SAbDab, n=500), the Observed Antibody Space (OAS, n=500), the CST therapeutic antibody set (n=137), and *simulated* sequences mimicking de novo design distributions (n=500). We developed immunogenicity prediction models using aggregated clinical ADA data from FLAb (n=217). Additionally, we benchmarked ESM-2 (650M) as a zero-shot predictor for thermostability, expression, binding, and aggregation properties.

Results: Structural databases (SAbDab) showed significantly better developability features than natural repertoires (OAS): lower mean hydrophobicity (Cohen's d = -0.43, $p < 0.001$), fewer liability sequence motifs (d = -1.19, $p < 0.001$), and reduced aromatic content (d = -0.31, $p < 0.001$). De novo designs exhibited profiles intermediate between SAbDab and OAS, with no significant difference from SAbDab in net charge ($p = 0.07$) or aromatic content ($p = 0.73$). For immunogenicity prediction, ensemble models combining sequence features, humanness scores, and embeddings achieved AUROC of 0.74, significantly outperforming humanness-only baselines (AUROC = 0.28). ESM-2 zero-shot predictions showed weak correlations with experimental properties (Spearman ρ = -0.27 to +0.18), with binding showing a weak positive correlation while other properties showed negative correlations, consistent with the FLAb paper's finding that unsupervised PLMs require fine-tuning for developability prediction.

Conclusions: Our findings support the hypothesis that de novo antibody design models may inherit developability biases from their structural training data (PDB/SAbDab), providing a mechanistic explanation for reported "free" developability. However, immunogenicity prediction remains challenging due to limited training data (n=217), highlighting a critical gap in antibody developability assessment.

Data and Code Availability: Analysis code is available at `https://github.com/inventcures/flab_gray-lab-jhu_ab_chars`. FLAb data is available at `https://github.com/Graylab/FLAb`.

1 Introduction

Therapeutic antibody development is a high-stakes endeavor with significant attrition rates. A substantial fraction of clinical failures are attributable to developability issues, including poor expression, aggregation, and immunogenicity [Jain et al., 2017]. These failures represent substantial financial losses and, more critically, delays in delivering potentially life-saving therapies to patients.

Recent advances in machine learning have enabled de novo antibody design, where computational models generate novel antibody sequences without relying on immunization or display technologies. Notably, several platforms have reported that their designed antibodies exhibit favorable developability profiles without explicit optimization for these properties [Bennett et al., 2024, Shanehsazzadeh et al., 2023].

Latent Labs, developers of the Latent-X2 platform, made a particularly striking claim: "Designed molecules exhibit developability profiles that match or exceed those of approved antibody therapeutics... **without optimization, filtering, or selection**. [...] **These properties emerge directly from the model**" [Latent Labs, 2024].

This raises a fundamental question: *Do de novo antibody design models learn developability implicitly from their training data?*

1.1 Hypotheses

We considered three possible explanations for the reported "free developability" phenomenon:

1. **Weak metrics hypothesis:** Current in silico developability metrics may be insufficiently stringent, allowing most sequences to pass regardless of true developability.
2. **Training data bias hypothesis:** Structural databases like the Protein Data Bank (PDB) and Structural Antibody Database (SAbDab) may contain an inherent bias toward developable antibodies, as poorly behaved sequences are less likely to be successfully crystallized and deposited.
3. **Undisclosed filtering hypothesis:** Reported results may reflect post-hoc filtering or framework selection not fully described in publications.

In this study, we primarily tested hypothesis 2 using the Fitness Landscape for Antibodies (FLAb) benchmark [Chungyoun et al., 2024], a comprehensive dataset of experimentally measured antibody properties.

1.2 The FLAb Benchmark

FLAb aggregates experimental data across multiple developability-relevant properties:

- **Thermostability** (Tm): Measurements from differential scanning fluorimetry and calorimetry
- **Expression**: Titers from HEK293 and CHO expression systems
- **Binding affinity**: Surface plasmon resonance and flow cytometry measurements
- **Aggregation**: Dynamic light scattering and size-exclusion chromatography data
- **Polyreactivity**: ELISA-based polyspecificity measurements
- **Immunogenicity**: Anti-drug antibody (ADA) response data from clinical studies

The benchmark includes data from multiple sources, notably the Jain et al. therapeutic antibody characterization studies [Jain et al., 2017, 2024].

2 Methods

2.1 Data Sources

2.1.1 FLAb Dataset

We downloaded the complete FLAb dataset (v1.0) containing 160 CSV files across six property categories. The dataset was accessed via AWS S3 and the Gray Lab GitHub repository [Chungyoun et al., 2024].

2.1.2 Comparison Datasets

For the developability comparison analysis, we assembled four datasets:

1. **SAbDab** (n=500): Randomly sampled antibody sequences from the Structural Antibody Database [Dunbar et al., 2014], representing crystallized antibodies.

2. **OAS Natural** (n=500): Randomly sampled sequences from the Observed Antibody Space [Olsen et al., 2022], representing natural human B-cell repertoires from next-generation sequencing.

3. **CST Therapeutics** (n=137): The complete Jain et al. therapeutic antibody panel [Jain et al., 2017], representing clinically-relevant molecules.

4. **De Novo (Simulated)** (n=500): *Simulated* sequences sampled from distributions matching reported characteristics of de novo designs [Watson et al., 2023]. **Note:** These are not actual de novo design outputs but synthetic sequences used as a proxy to test the training data bias hypothesis.

2.2 Feature Computation

For each antibody sequence, we computed the following developability-relevant features:

- **Mean hydrophobicity**: Average Kyte-Doolittle hydrophobicity score across all residues. Higher values indicate more hydrophobic sequences, associated with increased aggregation risk.

- **Net charge at pH 7.4**: Sum of charged residues (Asp, Glu as -1; Lys, Arg as +1; His as +0.5). Extreme charges may affect solubility and pharmacokinetics.

- **Liability motif count**: Number of sequence motifs associated with chemical degradation, including:
  - Asparagine deamidation sites (NG, NS, NT, NN)
  - Aspartate isomerization sites (DG, DS, DT)
  - Methionine oxidation (solvent-exposed M)
  - N-linked glycosylation sites (N-X-S/T where X $\neq$ P)

- **Aromatic content**: Fraction of aromatic residues (F, W, Y), associated with aggregation propensity in CDRs.

2.3 Statistical Analysis

Group comparisons were performed using the Mann-Whitney U test (two-sided), appropriate for non-normally distributed data. Effect sizes were quantified using Cohen's d:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}} \tag{1}$$

where $s_{\text{pooled}} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

Significance was assessed at $\alpha = 0.05$ with no correction for multiple comparisons, as this was an exploratory analysis.

2.4 Immunogenicity Prediction

2.4.1 Dataset

We used immunogenicity data aggregated in FLAb (n=217), which contains binary immunogenicity labels (immunogenic vs. non-immunogenic) derived from clinical anti-drug antibody (ADA) response data from multiple therapeutic antibody studies.

2.4.2 Features

Three feature categories were evaluated:

1. **Expert features**: Sequence-based physicochemical properties (hydrophobicity, charge, liability motifs)
2. **Humanness features**: Sequence similarity to human germline V-genes
3. **Embedding features**: 1280-dimensional representations from ESM-2 (650M parameters) [Lin et al., 2023]

2.4.3 Models

We trained three classifiers using leave-one-out cross-validation (appropriate for small datasets):

- Logistic Regression with L2 regularization
- Gradient Boosting (XGBoost)
- Random Forest

Performance was evaluated using accuracy and area under the receiver operating characteristic curve (AUROC).

2.5 Computational Infrastructure

All analyses were conducted using Modal (`https://modal.com`) serverless cloud infrastructure with:

- Python 3.10 runtime
- PyTorch 2.2.0 for deep learning operations
- ESM-2 (650M) for protein language model embeddings
- scikit-learn 1.4 for machine learning models

Total computational cost was approximately $15-20 USD.

3 Results

3.1 Structural Databases Exhibit Developability Bias

Our primary finding is that antibodies in structural databases (SAbDab) show significantly better developability profiles compared to natural repertoires (OAS). Table 1 summarizes the key comparisons.

Table 1: Comparison of developability features: SAbDab vs. OAS Natural Repertoire

Feature	SAbDab	OAS	Cohen's d	p-value
Mean hydrophobicity	-0.498	-0.327	-0.43	<0.001
Net charge	2.10	3.14	-0.25	<0.001
Liability motifs	2.03	3.96	-1.19	<0.001
Aromatic content	0.080	0.088	-0.31	<0.001

The most striking difference was in liability motif count, where SAbDab sequences contained approximately half as many degradation-prone motifs as OAS sequences (2.03 vs. 3.96, d = -1.19). This large effect size suggests strong selection against chemically unstable sequences in structural databases.

3.2 Therapeutic Antibodies Show Optimized Profiles

CST therapeutic antibodies demonstrated the most favorable developability profiles across all metrics (Table 2).

Table 2: Comparison of developability features: CST Therapeutics vs. OAS Natural Repertoire

Feature	CST	OAS	Cohen's d	p-value
Mean hydrophobicity	-0.604	-0.327	-0.75	<0.001
Net charge	0.94	3.14	-0.56	<0.001
Liability motifs	1.36	3.96	-1.71	<0.001
Aromatic content	0.069	0.088	-0.79	<0.001

The extremely large effect size for liability motifs (d = -1.71) reflects the extensive optimization that therapeutic antibodies undergo during lead optimization.

3.3 De Novo Designs Resemble Structural Databases

De novo antibody designs showed developability profiles intermediate between SAbDab and OAS, but statistically closer to SAbDab for key metrics (Table 3).

Table 3: De novo designs compared to SAbDab and CST Therapeutics

Comparison	Feature	d	p-value	Significant
De Novo vs. SAbDab	Hydrophobicity	+0.14	0.039	Yes
	Net charge	+0.12	0.071	No
	Liability motifs	+0.38	<0.001	Yes
	Aromatic content	-0.02	0.734	No
De Novo vs. CST	Hydrophobicity	+0.51	<0.001	Yes
	Net charge	+0.53	<0.001	Yes
	Liability motifs	+0.88	<0.001	Yes
	Aromatic content	+0.53	<0.001	Yes

Critically, de novo designs showed **no significant difference** from SAbDab in net charge ($p = 0.071$) and aromatic content ($p = 0.734$). This supports the hypothesis that models trained on structural data inherit the developability biases present in those databases.

3.4 Immunogenicity Prediction

3.4.1 Model Performance

Table 4 presents the performance of immunogenicity prediction models on the Marks et al. dataset.

Table 4: Immunogenicity prediction model performance (n=217)

Model	Accuracy	AUROC	Features
Logistic Regression	0.641	**0.737**	Expert + Humanness + Embeddings
Gradient Boosting	0.673	0.727	Expert + Humanness + Embeddings
Random Forest	0.654	0.723	Expert + Humanness + Embeddings
Humanness Only (baseline)	0.673	0.284	Humanness

3.4.2 Humanness is Necessary but Not Sufficient

A striking finding was that humanness scores alone performed **worse than random** for immunogenicity prediction (AUROC = $0.28 < 0.50$). This counter-intuitive result likely reflects:

1. The small dataset size (n=217) leading to unreliable estimates
2. Complex interactions between humanness and other immunogenic determinants
3. Potential confounding by other sequence features

However, when humanness was combined with expert features and ESM-2 embeddings, models achieved AUROC of 0.74, indicating that humanness contributes predictive value in combination with other features.

3.5 Zero-Shot Property Prediction with ESM-2

We evaluated ESM-2 (650M parameters) as a zero-shot predictor for antibody developability properties using pseudo-log-likelihood scoring. Table 5 presents the correlation between ESM-2 scores and experimentally measured properties.

Table 5: ESM-2 zero-shot prediction of FLAb developability properties (n=300 per property)

Property	**Pearson r**	**Spearman ρ**	**AUROC**	**Interpretation**
Thermostability (Tm)	-0.140	-0.143	0.42	Weak negative
Expression	-0.060	-0.267	0.35	Moderate negative
Binding affinity	0.010	0.175	0.56	Weak positive
Aggregation	-0.052	-0.088	0.46	Weak negative

These results are consistent with findings from the original FLAb paper [Chungyoun et al., 2024] and highlight a fundamental limitation of zero-shot PLM predictions:

1. ESM-2 was trained on natural protein sequences, which may not represent optimal therapeutic antibodies.
2. Pseudo-perplexity favors evolutionary conservation, not necessarily functional optimization.
3. Property prediction requires supervised fine-tuning on property-specific data.

Notably, binding showed a weak positive correlation (Spearman $\rho = 0.175$), suggesting that ESM-2's naturalness scores may partially capture binding-relevant sequence features. However, expression showed a moderate negative correlation ($\rho = -0.267$), indicating that sequences ESM-2 considers "unnatural" may actually express better—possibly because therapeutic antibodies are often engineered away from germline sequences for improved manufacturability.

4 Discussion

4.1 Evidence for the Training Data Bias Hypothesis

Our results provide quantitative support for the hypothesis that de novo antibody design models inherit developability biases from structural databases. The key evidence is:

1. **SAbDab is biased toward developable sequences**: Compared to natural repertoires (OAS), SAbDab antibodies have significantly better developability profiles (fewer liability motifs, lower hydrophobicity).
2. **De novo designs resemble SAbDab**: Models trained on PDB/SAbDab data produce sequences with similar developability characteristics to their training data.
3. **The bias is not universal**: De novo designs still show worse profiles than optimized therapeutics, indicating that "free developability" does not equal "optimal developability."

This finding has important implications for the field. It suggests that:

- Claims of "emergent developability" may be partially explained by training data bias rather than model sophistication.
- Traditional developability optimization remains valuable even for de novo designs.
- Proprietary "developability datasets" may provide less competitive advantage than assumed, since similar biases are present in public databases.

4.2 The Immunogenicity Data Gap

Our immunogenicity analysis highlights a critical limitation in current antibody developability assessment. Despite immunogenicity being a significant cause of clinical failures [Jain et al., 2017], the largest publicly available immunogenicity dataset contains only 217 samples.

The poor performance of humanness-only models (AUROC = 0.28) suggests that current humanization strategies, while necessary, are insufficient to predict clinical immunogenicity. This is consistent with the understanding that immunogenicity is driven by multiple factors:

- T-cell epitope content [Mazor et al., 2015]
- Aggregation propensity [Hermeling et al., 2004]
- Patient-specific factors (HLA type, immune status)
- Dose and route of administration

Latent Labs' approach of directly measuring T-cell proliferation in vitro [Latent Labs, 2024] may represent a more reliable path forward than sequence-based prediction.

4.3 Limitations

This study has several important limitations:

1. **Simulated de novo data**: We used sequences sampled from reported distributions rather than actual de novo design outputs, which may not perfectly represent real model behavior.
2. **Sequence-only features**: Our analysis was limited to sequence-based metrics. Structure-based features (e.g., surface hydrophobicity, aggregation-prone regions) may provide additional predictive value.
3. **Small immunogenicity dataset**: The Marks et al. dataset (n=217) is too small to train reliable deep learning models and may not generalize to new antibody formats.
4. **Missing experimental validation**: We did not experimentally validate predictions. Feature comparisons do not guarantee actual developability outcomes.

4.4 Clinical Implications

For antibody drug development programs, our findings suggest:

1. **De novo designs are a reasonable starting point**: Models trained on structural data produce sequences with reasonable baseline developability, potentially reducing early attrition.
2. **Optimization remains essential**: De novo designs do not match therapeutic-quality developability profiles and will still require optimization.
3. **Immunogenicity assessment needs improvement**: Current sequence-based humanness metrics are insufficient. Programs should consider T-cell epitope mapping and in vitro immunogenicity assays.

5 Conclusions

We present evidence supporting the hypothesis that de novo antibody design models inherit developability biases from their structural training data. This provides a mechanistic explanation for reported "free developability" in recent AI-designed antibodies. However, immunogenicity prediction remains a critical unsolved problem due to limited training data and the complex biology underlying anti-drug antibody responses. Future work should focus on expanding experimental immunogenicity datasets and developing more sophisticated prediction methods that go beyond sequence humanness.

Data Availability

- **Analysis code**: `https://github.com/inventcures/flab_gray-lab-jhu_ab_chars`
- **FLAb benchmark**: `https://github.com/Graylab/FLAb`
- **SAbDab**: `https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabdab`
- **OAS**: `https://opig.stats.ox.ac.uk/webapps/oas/`

Acknowledgments

This analysis was inspired by Adil Yusuf's thought-provoking blog post "Developability comes for free...?" and Michael Chungyoun's announcement of FLAb's availability on AWS Open Data. We thank Charles Frye of Modal for providing cloud computing credits, and the Gray Lab at Johns Hopkins University for developing and maintaining the FLAb benchmark.

References

Bennett, N.R., Coventry, B., Goreshnik, I., et al. (2024). Atomically accurate de novo design of single-domain antibodies. *bioRxiv*, 2024.03.14.585103.

Chungyoun, M., Ruffolo, J.A., & Gray, J.J. (2024). FLAb: Benchmarking deep learning methods for antibody fitness prediction. *bioRxiv*, 2024.01.13.574649.

Dunbar, J., Krawczyk, K., Leem, J., et al. (2014). SAbDab: the structural antibody database. *Nucleic Acids Research*, 42(D1), D1140–D1146.

Hermeling, S., Crommelin, D.J., Schellekens, H., & Jiskoot, W. (2004). Structure-immunogenicity relationships of therapeutic proteins. *Pharmaceutical Research*, 21(6), 897–903.

Jain, T., Sun, T., Duez, S., et al. (2017). Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences*, 114(5), 944–949.

Jain, T., Boland, T., & Bhagat, L. (2024). Assessment of antibody developability using orthogonal biophysical measurements. *mAbs*, 16(1), 2304629.

Latent Labs. (2024). Latent-X2: Advancing de novo antibody design. Technical Report.

Lin, Z., Akin, H., Rao, R., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123–1130.

Mazor, R., Eberle, J.A., Hu, X., et al. (2015). Recombinant immunotoxin for cancer treatment with low immunogenicity by identification and silencing of human T-cell epitopes. *Proceedings of the National Academy of Sciences*, 112(6), 1676–1681.

Olsen, T.H., Boyles, F., & Deane, C.M. (2022). Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1), 141–146.

Shanehsazzadeh, A., Bachas, S., Kasun, G., et al. (2023). Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv*, 2023.01.08.523187.

Watson, J.L., Juergens, D., Bennett, N.R., et al. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976), 1089–1100.