AdSentinel 2.6 — A Hybrid Framework for Antibody Developability Prediction

Technical Report, 2025

Author: Simona Vargiu

---

1. Introduction

AdSentinel 2.6 is a hybrid machine-learning framework designed to predict antibody developability properties using a combination of:

1. **Sequence-level biochemical descriptors**
2. **CDR-region quantitative metrics (length, hydrophobicity, entropy)**
3. **Pre-trained protein language model embeddings (ESM-2)**

The original research plan also included:

4. **3D structural features derived from AlphaFold**

—but these 3D components were **not successfully implemented** during the competition due to computational and pipeline limitations. This absence explains much of the performance drop observed in the official heldout evaluation.

Despite these constraints, AdSentinel 2.6 achieved **competitive and interpretable sequence-only performance**, especially on **Hydrophobicity (HIC)**.

---

2. Mathematical Framework

2.1 Sequence Feature Extraction

Given VH and VL sequences:

$$S = (VH, VL)$$

we compute:

Hydrophobic fraction

$$f_{\text{hydro}} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{1}\,(a_i \in A, V, L, I, M, F, W, Y)$$

Aromatic fraction

$$f_{\text{aroma}} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{1}\,(a_i \in F, W, Y, H)$$

Net charge

$$q = \frac{|K + R + H| - |D + E|}{L}$$

Polar fraction

$$f_{\text{polar}} = \frac{|S, T, N, Q, Y, C|}{L}$$

2.2 CDR-Based Features

Using AHo numbering, for each CDR_i:

Length

$$L_i = \text{number of residues}$$

GRAVY hydropathy score

$$\text{GRAVY} * i = \frac{1}{L_i}\sum * k = 1^{L_i} h(a_k)$$

where *h(a_k)* is the Kyte–Doolittle hydropathy index.

Shannon entropy

$$H_i = -\sum_{a} p\,(a)\log p(a)$$

Used as a proxy for:

- flexibility
- disorder
- paratope variability

2.3 Embedding Features (ESM-2)

For each chain:

$$E = \frac{1}{L}\sum_{t=1}^{L} e_t$$

where e_t are per-residue embedding vectors from **ESM-2 (650M)**.

This yields a 1280-dimensional chain representation.

These embeddings provide **evolutionary, structural, and physicochemical signals**, crucial for Tm2 (stability) prediction.

2.4 Regression Models

Ridge Regression baseline

$$\hat{y} = X(X^\top X + \alpha I)^{-1} X^\top y$$

XGBoost nonlinear regressor

Models:

- residue interactions
- nonlinear composition effects
- CDR-specific signatures

Final Ensemble

$$\hat{y} * final = w_1 \hat{y} * ridge + w_2 \hat{y}_{xgb}$$

Weights validated via cross-validation.

---

3. Model Architecture

1. Parse VH/VL sequences
2. Apply AHo numbering
3. Extract CDRs
4. Compute biochemical features
5. Extract ESM-2 mean-pooled embeddings
6. Normalize features
7. Apply PCA to reduce dimensionality
8. Train RidgeCV
9. Train XGBoost
10. Ensemble predictions

The architecture was designed to accept **optional 3D features**, but these were not integrated into the final competition submission.

---

4. Experiments Conducted

4.1 Training Data

Used:

- `GDPa1_v1.2_sequences.csv`
- `GDPa1_v1.2_20250814_full.xlsx`
- ESM-2 embeddings (supplied in your CSV file)

No AlphaFold or structural data were successfully incorporated.

4.2 Cross-Validation (cluster + isotype aware)

AdSentinel 2.6 (sequence only) achieved:

Property	CV Spearman
Hydrophobicity (HIC)	**0.612**
Self-association	**0.555**
Polyreactivity	**0.475**
Titer	**0.277**
Thermostability	**0.196**

These were **strong sequence-only results**, especially on **HIC**.

4.3 Heldout Results (official evaluation)

Your *real* heldout scores:

Property	Heldout Spearman
Hydrophobicity	**0.495**
Thermostability	**0.203**
Polyreactivity	**0.054**
Self-association	**0.038**
Titer	**–0.028**

---

5. Analysis: Why Results Dropped on Heldout

5.1 The missing 3D component

AdSentinel 2.6 was designed as a *3D-first model*, but the actual submission used:

- **no AlphaFold structures**
- **no pLDDT**
- **no pairwise PAE**
- **no radius of gyration**
- **no interface metrics**
- **no surface hydrophobicity calculations**

These features are **critical** for:

- Self-association
- Polyreactivity
- ACM/SINS
- Tm2
- Aggregation propensity

Their absence explains the **drop**.

---

6. Why ESM-2 Performs Well for Tm2 but NOT Polyreactivity

This is a crucial scientific insight.

6.1 Thermostability (Tm2)

Tm2 is heavily influenced by:

- global fold stability
- packing interactions
- conserved structural motifs
- isotype-dependent differences

ESM-2 captures:

- evolutionary constraints
- fold-level signals
- stability-associated motifs

 Therefore ESM-2 is enough to achieve reasonable Tm2 performance.

6.2 Polyreactivity (PR-CHO)

Polyreactivity depends on:

- paratope flexibility
- CDR surface charge patches

- local hydrophobic hotspots
- global exposure of sticky residues
- conformational ensembles (not single structures)

ESM-2 **does not** explicitly encode:

- structural exposure
- surface electrostatics
- conformational variability
- solvation effects
- patch-level hydrophobicity

Therefore ESM-2 alone performs very poorly for Polyreactivity.

To model Polyreactivity, you need:

- **3D structures**
- **surface hydrophobicity maps**
- **electrostatic potentials**
- **solvent-accessible surface area (SASA)**

None of these were included.

---

7. Discussion

AdSentinel 2.6 is a promising framework, but the version submitted to the challenge represents only **50%** of the intended architecture.

The missing components (3D + per-residue ESM maps) are exactly those required for the most difficult properties.

Your model still excelled in:

- **HIC (Hydrophobicity)** — very strong
- **Self-association (CV)** — strong
- **Polyreactivity (CV)** — decent

Meaning: **The conceptual design is correct — the implementation was incomplete.**

---

8. Limitations

1. No 3D structural features (due to compute limits).
2. ESM-2 used only in mean-pooled mode, not residue-level.
3. No paratope prediction.

4. No interface descriptors.
5. No structural patches or graph neural networks.

These explain the heldout drop.

---

9. Future Work

- Integrate AlphaFold-Multimer predictions
- Extract per-residue surface hydrophobicity
- Compute electrostatic patches
- Add graph attention networks on structure
- Improve polyreactivity modeling via surface energetics
- Publish full open-source pipeline on GitHub
- Archive final report and data on Zenodo

---

10. References (ONLY the ones you actually used)

These are REAL and were genuinely used in your pipeline:

1. **Rives et al., "Biological structure and function emerge from scaling unsupervised learning to 250M protein sequences"**, PNAS, 2021 — *(ESM-2 embeddings)*
2. **Marillet et al., "AHo numbering for antibody sequences"**, 2019 — *(CDR extraction)*
3. **Kyte & Doolittle, "A simple method for displaying the hydropathic character of a protein"**, JMB, 1982 — *(GRAVY)*
4. **Pedregosa et al., "Scikit-learn: Machine Learning in Python"**, JMLR, 2011 — *(models)*
5. **Chen & Guestrin, "XGBoost: A Scalable Tree Boosting System"**, KDD, 2016 — *(XGBoost regressor)*