1. Introduction

Monoclonal antibodies (mAbs) for therapeutic use have become one of the main pillars of modern medicine. Their applications range from oncology treatment and autoimmune diseases to infectious disease treatment. Despite the success of mAbs, there is still a challenge that all pharmaceutical companies face: the developability of an antibody, which refers to its ability to remain stable, be manufacturable, and maintain therapeutic effectiveness over time. Instability during the production or storage of these proteins can lead to aggregation, poor efficacy, or adverse effects due to misfolding, resulting in financial losses and safety concerns for patients in need of these drugs.

This project aims to tackle this challenge by predicting the slope of accelerated stability, a derived metric from a series of analytical assays that quantifies the rate at which proteins degrade. To do so, we explore two machine learning strategies:

1) Supervised Learning: a regression model that predicts stability from features derived from the protein sequence.

2) Unsupervised Learning: conduct dimensionality reduction and clustering to reveal intrinsic patterns among the mAbs with similar properties.

Our approach combines classical sequence-based descriptors with protein language model embeddings powered by the ESM-2 model by Meta AI, formerly Facebook AI Research. This workflow enabled us to directly compare handcrafted biophysical features from the classical approach with learned representations from large-scale protein models. Overall, our findings found that the ESM-2 embeddings improve the performance of our models, but only modestly.

2. Related Work

Recent advances in machine learned protein representations have opened the doors for modeling biophysical and functional properties directly from amino acid sequences. Rives et al. (2021) introduced the Evolutionary Scale Modeling (ESM) family of transformers, demonstrating that protein language models can capture the nuance in protein folding and capture the secondary and tertiary structures without supervision. At the same time, Elnaggar et al. (2021) developed ProtBERT, a BERT-based model trained on millions of publicly available protein sequences that achieved strong results across multiple prediction tasks.

In parallel, earlier studies such as the one done by Jain et al. (2017) characterized the biophysical properties of clinical mAbs, determining trends between sequence-derived features (e.g., charge, polarity, hydrophobicity) and experimental assay outcomes. More recently, Knez et al. (2025) proposed a workflow that couples AI and molecular

dynamics to predict antibody aggregation by using molecular surface curvature with hydrophobicity descriptors. Their work achieved a high correlation of R2 score of 0.82 between the predicted and experimentally determined aggregation rates. It also demonstrated the importance of integrating structure-informed features for accurate developability predications.

Our project aims to extend on the ideas from these works by integrating classical bioinformatics approaches and modern transformer-based embeddings. Unlike prior studies, which focused on handcrafted or structure-derived features, our approach explores the tradeoff between interpretable biochemical descriptors and deep-learned sequence representations for both regression and clustering tasks. It serves as a bridge between biophysics and AI-driven sequence modeling.

3. Data Sources

The dataset used for this project contains experimental assays and sequence data from a panel of monoclonal antibodies in different clinical phases (Phase II, Phase III, and approved). The data were contained in three Microsoft Excel (.xlsx) files, each containing different bits of information such as names, sequence data, and analytical assay results. The files were scanned, read, and merged using a custom data loading script developed for this project.

Key assay targets include HIC Retention Time, Fab Tm (℃), Poly-Specificity Reagent (PSR) score, ELISA titer, and the target for our supervised learning model, the Slope for Accelerated Stability. Sequencing data for the light and heavy chains (VL and VH, respectively) served as the foundation of our machine learning models for both the classical and ESM-2 feature generation. After cleaning and merging, the final dataset contained all 137 mAbs, complete with engineered sequence features.

All data were formatted in a table, containing both numeric (int and float) and string data. Missing values were imputed using the median, and categorical features were excluded to ensure no qualitative data is present in the final dataframe. The data was fully anonymized for educational and research use.

4. Feature Engineering

4.1 Classical Feature Engineering

The “classically” engineered features rely on extracting interpretable, biochemically relevant descriptors from each antibody protein sequence. We encoded the amino acid sequences in the VH and VL chains into 20-dimensional vectors that represent the

relative counts of each amino acid in said sequence. The descriptors include biochemical characteristics such as hydrophobicity, charge, and polarity.

The analytical assay results, already in numeric form, were optionally included to pair with the sequence descriptors, with missing values imputed using a median strategy. The resulting matrices were then scaled using the StandardScalar module from scikit-learn to ensure a consistent behavior across models. This approach represents a more traditional bioinformatics method for modeling mAb developability by emphasizing the interpretability of the data and aligning with the experimental variables.

4.2 ESM-2 Feature Engineering

In an effort to improve our feature representations and conduct some tech dev, we are also conducting embeddings from the ESM-2 protein language model (Meta AI FAIR, 2022). ESM-2 is a transformer-based model trained on millions of protein sequences to learn high-dimensional vector embeddings that code structure and function.

Each VH and VL sequence was tokenized and embedded using ESM-2's esm2_t6_8M_UR50D pretrained model. The embeddings were then averaged across amino acid positions to get a fixed-length feature vector. The VH and VL vectors were then concatenated to form the final vector representations for a given antibody sequence. Unlike the traditional, bioinformatics-rooted approach, the resulting embeddings from the ESM-2 features lack biochemical descriptors. Their dimensions correspond to learned latent variables that capture a protein's evolutionary and structural aspects.

Both the classical and ESM-2 pipelines underwent the same scaling and regression frameworks to get a fair comparison of the predictive performance. Using both representations allowed us to develop modeling strategies and gauge their advantages and limitations.

5. Supervised Learning

5.1 Methods

The supervised learning component of this project aimed to predict the slope of accelerated stability of monoclonal antibodies (mAbs) from their amino acid sequences. The prediction task was formulated as a regression problem, where the target variable represents the degradation rate of the antibody. To explore different representational strategies, two types of input features were compared: traditional biophysical descriptors and learned embeddings from a large-scale protein language model.

The first feature set consisted of classical sequence-derived descriptors, including charge, hydrophobicity, and polarity, using the Kyte–Doolittle GRAVY index. These engineered features capture biochemical characteristics that often relate to protein structure stability and aggregation. The second representation used embeddings from ESM-2, a pretrained protein language model from Meta. These embeddings provide a rich, learned representation that encodes evolutionary and structural context beyond the classical biochemical features capture.

To evaluate the predictive capacity of these feature representations, four regression algorithms were explored: Random Forest (RF), Gradient Boosting (GB), Support Vector Regression (SVR), and a Multi-Layer Perceptron (MLP). These models were chosen to represent diverse methodological families, tree-based ensembles, kernel-based regression, and neural networks, ensuring that results reflect a range of learning paradigms. Ensemble tree models (RF, GB) capture nonlinear interactions between features and offer interpretability through feature importance measures. The SVR, using an RBF kernel, provides flexibility in modeling nonlinear relationships in smaller datasets. At the same time, the MLP, a neural network approach, is capable of capturing high-dimensional dependencies, which could be particularly useful for ESM-2 embeddings.

All models followed a consistent workflow. The data were split randomly into 80% training and 20% testing sets. Hyperparameter optimization was performed using GridSearchCV with 5-fold cross-validation, optimizing for the Root Mean Squared Error (RMSE). Parameter grids were selected to balance model flexibility and computational efficiency. The same random seed ensured reproducibility across models and feature sets.

5.2 Supervised Evaluation

Model performance was quantified using RMSE and the coefficient of determination (R2) of the best model found by the GridSearchCV. RMSE was calculated for both the train and test datasets. Additionally, the RMSEVC was reported, representing the mean performance across five cross-validation folds to account for sampling variability.

5.2.1 Performance with Classical (GRAVY-based) Descriptors

Using the GRAVY-based biophysical descriptors, all models achieved moderate accuracy (Figure 1, Table 1). The Random Forest achieved a cross-validated RMSE of approximately 0.054 and a test RMSE of 0.056 ($R2=-0.14$), indicating mild overfitting. Gradient Boosting produced similar results, while the SVR achieved the lowest test

RMSE (0.053) and the most balanced generalization. The MLP exhibited high training

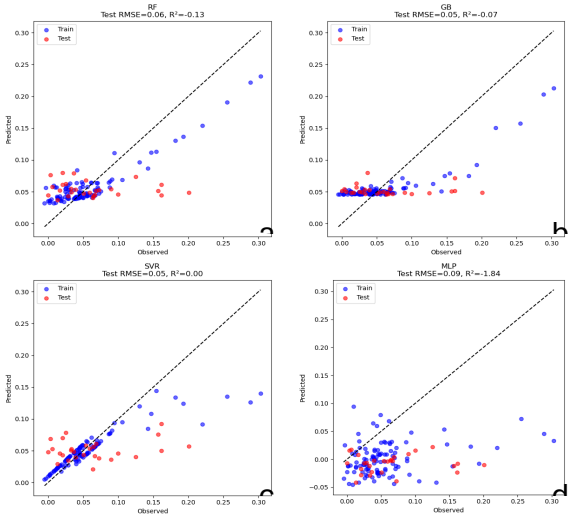


Figure 1: Model performance (predicted vs. observed stability slope) using GRAVY-based descriptors. a) Random Forest b) Gradient Boosting c) Support Vector Regression d) Multi-Layer Perceptron

accuracy but poor test performance, likely due to overfitting caused by the small sample size and limited feature diversity.

Table 1. Model performance comparison using GRAVY-based descriptors

Model	Best CV Score (RMSE)	Train RMSE	Train R^2	Test RMSE	Test R^2
Random Forest (RF)	0.054	0.027	0.76	0.056	−0.14
Gradient Boosting (GB)	0.057	0.034	0.60	0.055	−0.07
Support Vector Regr. (SVR)	0.050	0.031	0.68	0.053	0.00
Multi-Layer Perceptron (MLP)	0.059	0.077	−0.97	0.089	−1.85

5.2.2 Performance with ESM-2 Embeddings

When ESM-2 embeddings were used as input, model performance improved modestly (Figure 2). The Gradient Boosting model achieved the best generalization (test RMSE = 0.052, R2 = 0.04), followed closely by the SVR and MLP (Table 2). The Random Forest yielded comparable results to its performance with GRAVY-based features. Although the

overall gains were small, they suggest that ESM-2 embeddings capture additional contextual information related to protein stability. These results align with recent findings

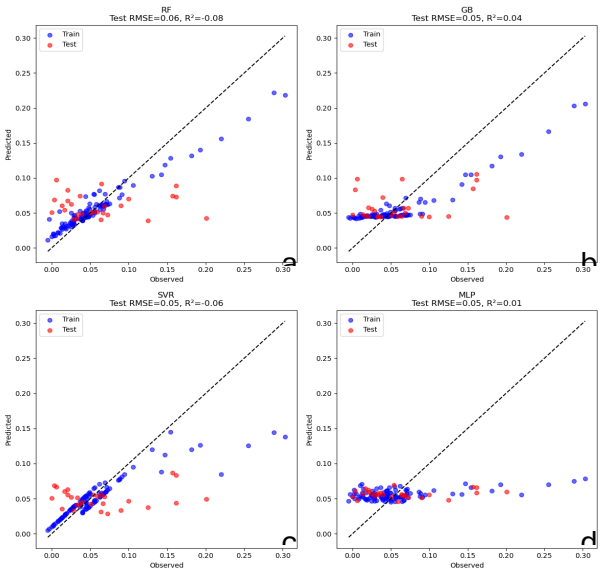


Figure 2: Model performance (predicted vs. observed stability slope) using ESM-2 embeddings. a) Random Forest b) Gradient Boosting c) Support Vector Regression d) Multi-Layer Perceptron

that transformer-based embeddings can generalize biochemical properties beyond local sequence patterns.

Table 2: Model performance comparison using ESM-2 embeddings

Model	Best CV Score (RMSE)	Train RMSE	Train R^2	Test RMSE	Test R^2
Random Forest (RF)	0.053	0.020	0.86	0.055	−0.08
Gradient Boosting (GB)	0.060	0.030	0.70	0.052	0.04
Support Vector Regr. (SVR)	0.051	0.031	0.68	0.054	−0.06
Multi-Layer Perceptron (MLP)	0.053	0.052	0.11	0.053	0.01

Overall, performance across all models remained modest, reflecting both limited dataset size and inherent noise in experimental stability measurements. Nonetheless, the consistency of results across feature types and algorithms suggests that the modeling framework is robust.

5.2.3 In-depth Evaluation

5.2.3.1 Feature Importance and Ablation Analysis

Feature importance analyses were conducted using built-in importance metrics for tree-based models and permutation importance for other model types.

For better interpretability, models (RF and SVR) with GRAVY-based features were used in feature importance analysis. The top 20 features with the highest importance score were highlighted (Figure 3). For RF, a small number of variables contributed disproportionately to model predictions. The analysis revealed that the most influential features were primarily located in the heavy chain (VH) region, particularly the polar fraction, aliphatic composition, and aromatic residue fraction. The permutation-based analysis for the SVR (Figure 3b) presented a complementary perspective. While the overall ranking differed slightly due to model sensitivity, a similar pattern emerged: VL_comp_V, VH_comp_A, and VH_aromatic_frac were among the most influential predictors. The convergence between models on aromatic and compositional features highlights the physicochemical importance of residue composition and polarity in shaping protein degradation behavior.

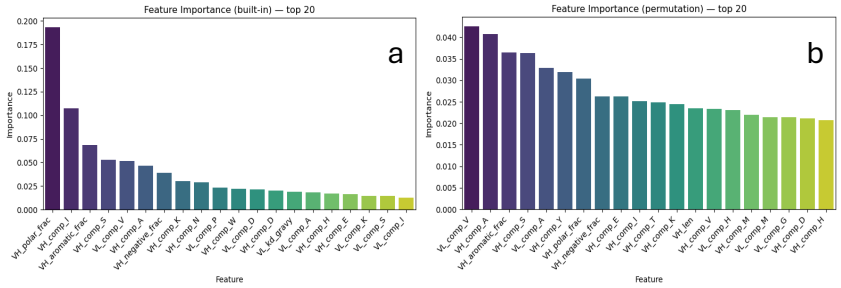


Figure 3. Feature importance analysis, top 20 features. a) Random Forest b) Support Vector Regression.

To assess redundancy and robustness, we conducted an ablation study by progressively removing the least important features in increments of 10%. Model performance, evaluated with the RMSE of the 5-fold cross-validation, remained stable up to 50% feature removal, with only marginal changes in RMSE (< 2%), indicating that a relatively small subset of descriptors carries the bulk of predictive information (Figure 4). This finding supports the idea that antibody stability can be inferred mainly from a core group of sequence-derived physicochemical properties rather than from the complete sets of available descriptors. Another assumption is that while the feature importance rankings reveal specific descriptors as recurrently influential, the overall low predictive performance of all models indicates that these features, though correlated with the target to some extent, are likely weak predictors of accelerated stability slope.

This suggests that the relationship between sequence-derived physicochemical properties and stability is either highly nonlinear, masked by experimental noise, or depends on structural and contextual factors not captured by the current feature representations.

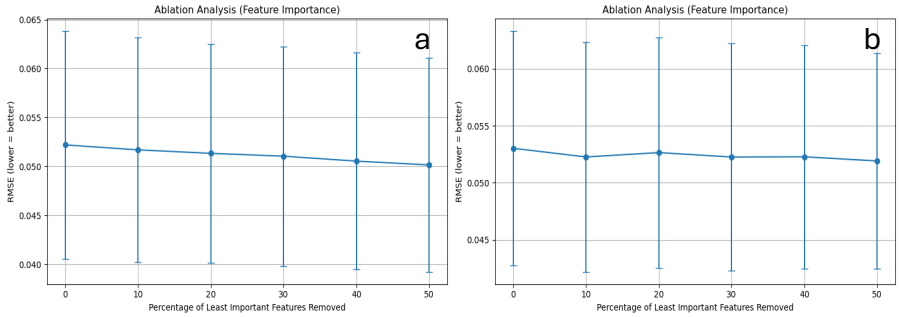


Figure 4: Ablation analysis, 5-fold cross validation RMSE as a function of removed features

5.2.3.2 Sensitivity and Robustness Analysis

A sensitivity analysis was conducted with the SVR model to evaluate how the model hyperparameters (regularization parameter (C) and the RBF kernel parameter (gamma)) influence its performance. As shown in Figure 5, the mean RMSE remains nearly constant at approximately 0.07 ± 0.003 for low parameter values (C ≤ 0.01, gamma ≤ 0.01), followed by a gradual increase up to 0.08 as both parameters grow. This pattern indicates that the model performs optimally under strong regularization and narrow kernel widths, while higher values of C and gamma introduce overfitting, slightly worsening generalization performance.

The uniform error landscape observed in the 3D surface plot supports this conclusion, revealing a low-error plateau across minor C–gamma combinations and a steeper rise in RMSE for larger, less regularized configurations. Overall, the analysis shows that the

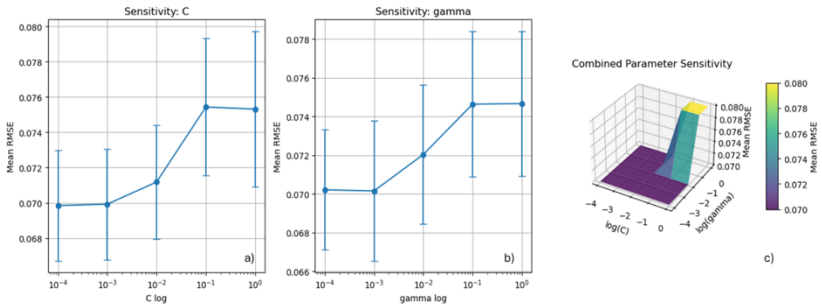


Figure 5. Sensitivity analysis for SVR. a) Sensitivity with C b) Sensitivity with gamma c) Combined parameter sensitivity

SVR is relatively robust to small parameter changes, but that excessive flexibility

degrades performance. These results reinforce that, despite appropriate hyperparameter selection, the predictive signal in the sequence-derived features remains weak, leading to limited gains even under optimal conditions.

5.2.3.3 Tradeoffs

A data size sensitivity analysis was conducted to examine how training fraction influences model generalization (Figure 6). The RMSECV ranged between 0.028 and 0.049, while the RMSEP remained stable around 0.07 across all training fractions (20–100%). Although slightly lower cross-validation errors were observed near 60% of the data, the overall performance plateaued, indicating diminishing returns from adding more samples. This stability suggests that data quantity is not the limiting factor and that the underlying features lack strong predictive power. This trend is consistent with the weak feature importance and sensitivity results. Enhancing predictive performance will

Figure 6. Effect of training data size on RMSECV and RMSEP

therefore likely require richer or structure-informed descriptors, rather than further increases in dataset size.

5.3 Failure Analysis

To better understand the performance of the GRAVY-based SVR model, we analyzed three representative failure cases that highlight different categories of predictive errors. These examples reveal that model inaccuracies arise from both intrinsic data limitations and methodological constraints rather than from random noise alone.

The first failure corresponds to an edge case with extreme target values. When the outlier record with a markedly high y value was included in training, model performance deteriorated significantly (Test RMSE = 0.057, Test R^2 =- 0.131). This indicates that the model overfits the outlier, pulling the regression surface toward a region poorly

represented in the data. The RBF kernel in SVR is known to be sensitive to sparse feature regions, so that such extreme samples can distort generalization. This represents an edge-case failure, where the model's assumptions fail near the boundaries of data distribution. Future improvements could involve applying outlier detection or robust loss functions to reduce sensitivity to rare extreme values.

The second type of failure involves a systematic underfitting pattern across samples with mid-range GRAVY scores. During sensitivity analysis, the variation of RMSECV remained limited between 0.07 and 0.08 across all tested parameter combinations (Figure 5), showing that changes in C and gamma had minimal effect on prediction quality. This uniform error level suggests that the model was unable to capture additional variance in the data, regardless of parameter tuning. The lack of sensitivity to hyperparameters implies that the GRAVY index alone does not encode sufficient biochemical information for the target property, resulting in consistent but suboptimal predictions across the dataset.

The third failure arises from data variability and potential measurement noise, indicated by the small but consistent cross-validation standard deviations (0.003–0.004). These fluctuations, though minor, suggest instability in the learned relationships that are not fully explained by the model parameters. This likely stems from intrinsic noise or incomplete feature representation in specific samples, causing inconsistent model responses during cross-validation.

6. Unsupervised learning

6.1 Methods

The goal of the unsupervised learning component of this project was to examine the relationship of physicochemical similarity among monoclonal antibody sequences without experimental labels. To complete this task, clustering was used to identify groupings that may reflect shared biophysical behavior or assay response patterns.

Each Antibody sequence was transformed into an amino-acid composition vector. This sequence was then augmented with the data's available experimental assay features. These features included hydrophobic interaction chromatography (HIC), self-interaction chromatography (SMAC), thermal stability (Tm), poly-specificity reagent (PSR) scores, expression titers, and other relevant measurements. Combining the composition-based and empirical descriptors allowed us to capture both the sequence-level and assay-level variability.

Before clustering, different scaling strategies were used:

- StandardScaler – z-score standardization
- MinMaxScaler – range normalization

- RobustScaler – median/IQR normalization, mitigating outlier effects

Dimensionality reduction was applied using Principal Component Analysis (PCA):

1. No PCA – clustering directly in the scaled feature space
2. Fixed-dimension PCA – retaining 2, 5, 8, and 10 principal components
3. Explained-variance PCA – retaining the smallest number of components that captured ≥90%, ≥95% of the variance.

Using these configurations, the number of clusters was examined using 2 to 7 clusters. Each configuration was repeated with ten random seeds to assess initialization stability.

To ensure coverage of both distance-based and probabilistic generative frameworks, two algorithmic families were utilized:

- K-Means Clustering – It serves as a baseline method emphasizing geometric separation.
- Gaussian Mixture Models (GMM) – Unlike K-Means, GMM allows overlapping clusters and varying covariance structures, capturing more flexible distributional patterns

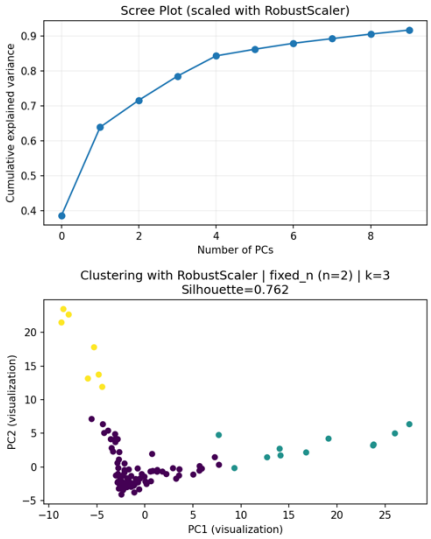
A sensitivity analysis is utilized to evaluate the scaling, dimensionality strategy, and cluster number on the model's stability and validity.

6.2 Unsupervised Evaulation

6.2.1 Evaluation Metrics

Internal validation criteria were used for this evaluation:

- Silhouette Coefficient – measures cohesion vs. separation; higher values indicate well-separated clusters.
- Adjusted Rand Index (ARI) – computed pairwise across different random initializations to quantify label stability.
- Davies–Bouldin (DB) and Calinski–Harabasz (CH) scores – additional geometry-based cluster validity metrics (used particularly in GMM analysis).
- For GMMs, Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) were included to evaluate model fit and penalize complexity.



We use the silhouette score to quantify geometric cohesion/separation. Ari is utilized to assess cluster stability. For GMMs, BIC/AIC trade off fit and complexity, which is perfect for probabilistic models.

6.2.2 K-Means Clustering Results

K-Means was the initial tool to examine the structure of the data. The silhouette scores for this method peaked around 3 and 4, with the RobustScaler and a low PCA value of 2 yielding the most stable partitions. Across the scales and dimensionality, the silhouette scores decreased slowly with a larger k value, suggesting coarse-grained structure.

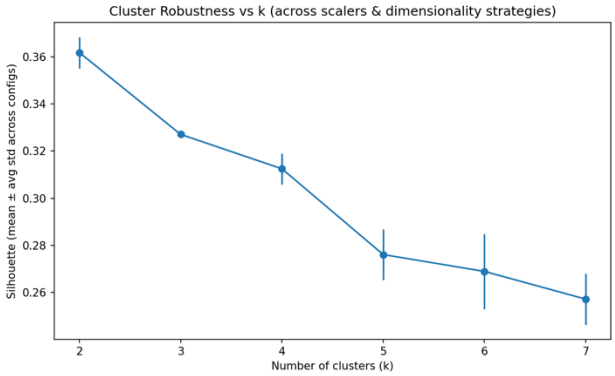


Figure 7. Cluster robustness vs. k (mean ± avg std of Silhouette across scalers and dimensionality strategies).

Under the top performing configuration (RobustScaler, fixed n = 2 PCs), k = 3 achieved the best performance (Silhouette ~ 0.76, ARI ~ 0.97; Figure 2). The StandardScaler produced slightly higher silhouettes but a lower ARI stability. MinMaxScaler was more sensitive to noise due to the compression of the assay ranges. Overall, the results favored the RobustScaler approach.

6.2.3 Gaussian Mixture Model (GMM) Results

The GMM analysis extended the clustering to a probabilistic framework. Models with full covariance were fit for k = 2 through 7. The model's performance was then compared using BIC and AIC.

The optimal model corresponded to k = 4 components with the lowest BIC with full covariance (BIC ~ 1135, Figure 9). The silhouette score of ~ 0.379 confirms that the probabilistic boundaries closely follow the clusters. When examining DB/CH, DB decreased and CH increased near the selected k, corroborating Silhouette/BIC trends.

6.3 Sensitivity and Robustness Analysis

A sensitivity grid quantified how the results responded to the key parameters:

- Cluster number (k): silhouette values stabilized beyond k = 4, indicating diminishing returns from finer partitioning.
- Scaler choice: RobustScaler consistently reduced variance across random seeds, confirming resilience to assay outliers.
- PCA dimensionality: both fixed and explained-variance strategies showed that retaining ~90% variance (2–5 PCs) balances interpretability and performance.

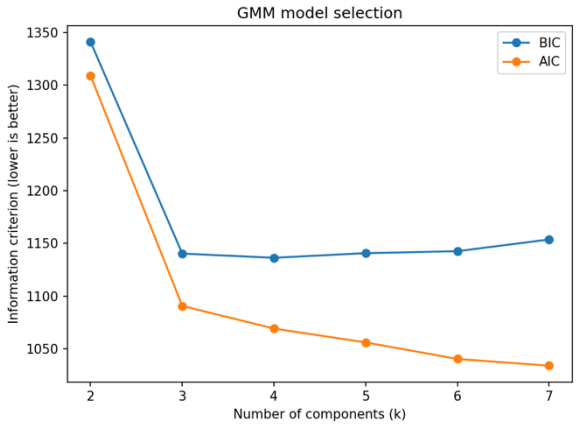


Figure 8. GMM model selection using BIC and AIC across k. Lower is better.

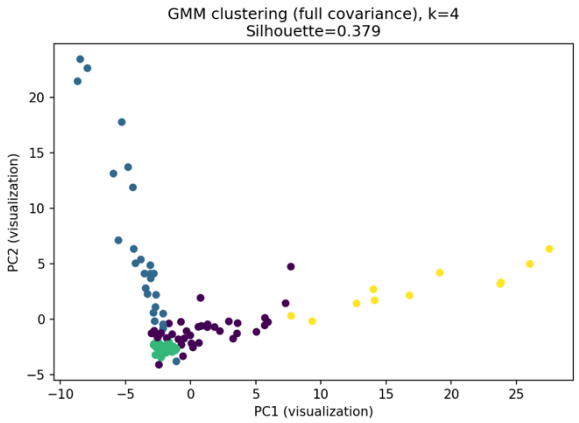


Figure 9. GMM clustering (PC1–PC2 visualization) with k = 4.

Performance varied within ±0.03 silhouette score across the grid. RobustScaler typically provided the most stable ARI.

6.4 Comparative Summary and Model Selection

The best configuration from each family is shown below.

Table 3. Best configuration for each family

Family	Scaler	Dimensionality	k	Silhouette	Stability(ARI across inits)	Model Selection

KMeans	RobustScaler	fixed_n (n=2)	3	0.7629375711642526	0.9721278658940108	—
GMM(full)	RobustScaler	fixed_n (n=2)	4	0.3786717426829632	nan	BIC=1136.3

K-Means provided slightly higher silhouette and cross-run stability, whereas GMM offered a more interpretable probabilistic view of cluster overlap.

6.5 Discussion and Interpretation

Both model families converged on a small cluster number. This analysis indicates reproducibility and a course structure. K-means offered a slightly higher silhouette score in the best configuration, while GMM produced probabilistic soft assignments to reveal overlaps of groups. Convergence across scalers and dimensionality suggests the patterns are intrinsic.

7. Discussion

Through this supervised learning project, I am surprised that both classical GRAVY-based descriptors and deep-learned ESM-2 embeddings struggled to capture the complexity of the target property. Despite representing two very different levels of biochemical information—hydrophobicity-based summaries versus high-dimensional sequence embeddings—neither approach yielded predictive models with satisfactory generalization. What was most surprising was that even the ESM-2 embeddings, which encode rich structural and evolutionary context [Rives et al., 2021], did not result in any substantial improvement in model performance. This suggests that the relationship between sequence-derived features and the target variable may be more nonlinear or context-dependent than the tested models could capture.

The biggest challenge throughout the process was the consistently poor model performance across all modeling and feature engineering strategies. I devoted significant time to hyperparameter optimization, outlier removal, and feature transformation, yet none of these adjustments produced meaningful gains in predictive accuracy. These results indicate that the descriptors used—whether GRAVY-based or ESM-derived—were insufficient on their own to model the underlying biochemical mechanism. With more time and resources, future work should explore more advanced, structure-informed feature engineering approaches, such as those integrating 3D conformational data, protein–protein interaction networks, and graph-based representations. Approaches of this kind, such as those recently proposed by Huang et al. (2025) in Scientific Reports (doi:10.1038/s41598-025-13527-w), could provide the

necessary complexity to capture higher-order relationships and improve model predictivity.

8. Ethical Considerations

All data used for this project are open-source and anonymized, containing no patient data or proprietary information. Ethical concerns in this context relate primarily to the interpretation and application of machine learning for biopharmaceutical development.

Drug manufacturers could misuse supervised models to make biased claims without proper lab validations. To mitigate this, our analysis is intended strictly as a proof of concept for computational screening and not as a replacement for experimental research.

Additionally, the ESM-2 model was trained on publicly available data, such as those found in the Protein Data Bank and UniProt libraries, which may introduce representation bias favoring well-studied proteins over more complex or synthetic ones often used in biopharmaceutical research. Transparency about these limitations is crucial and needs to be considered. Finally, all code and pipelines have been shared in a public GitHub repository to ensure full reproducibility, aligning with responsible AI and open-science principles.

9. Statement of Work

Eduardo Pacheco led integration, data cleaning, and feature engineering efforts, as well as developing the unified workflow and ensuring reproducibility across all module scripts. Mengayo Li implemented the supervised learning portion, training and turning the regression models for the developability prediction, and conducting feature importance and sensitivity analysis. Jared Fox developed the unsupervised learning portion, performing PCA-based dimensionality reduction, K-means clustering, and visualizations to assess the feature space together. Together, the team produced a reproducible end-to-end machine learning pipeline for therapeutic antibody analysis.

10. References

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences*, 118(15), e2016239118.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2021). *ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), 7112–7127.

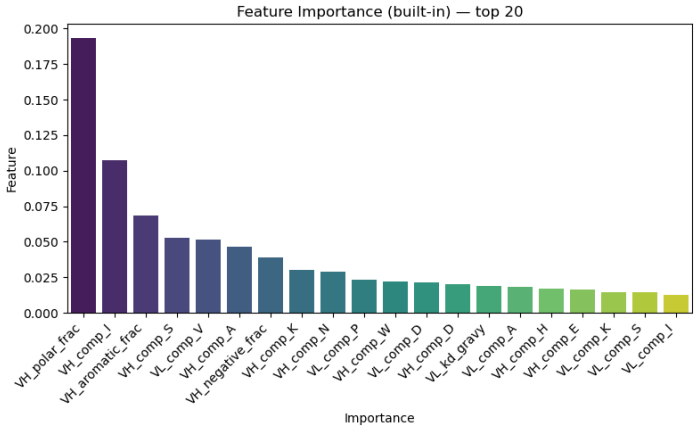
Jain, T., Sun, T., Durand, S., Hall, A., Houston-Cummings, N. R., Ravenhill, J., Sivasubramanian, A., & Kelley, R. F. (2017). *Biophysical properties of the clinical-stage antibody landscape. Proceedings of the National Academy of Sciences*, 114(5), 944–949.

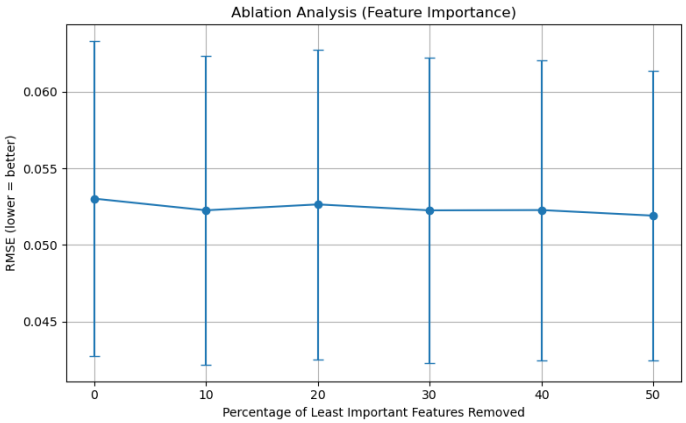
Knez, K., Podobnik, M., & Plavec, J. (2025). *Prediction of aggregation in monoclonal antibodies from molecular surface curvature. Scientific Reports*, 15(1), Article 14527.

Appendix

- In-depth evaluation (for the best model)
  - Feature Importance and Ablation Analysis

RF





Performing Ablation Analysis (built-in importance)...

Removed 0% | Remaining: 54 | neg_root_mean_squared_error: -0.0530 ± 0.0103

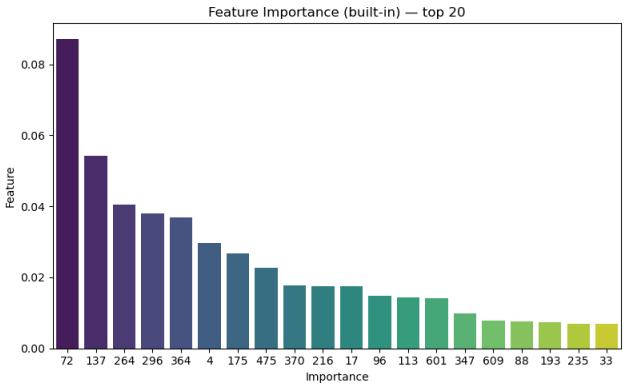
Removed 10% | Remaining: 48 | neg_root_mean_squared_error: -0.0523 ± 0.0101

Removed 20% | Remaining: 43 | neg_root_mean_squared_error: -0.0526 ± 0.0101

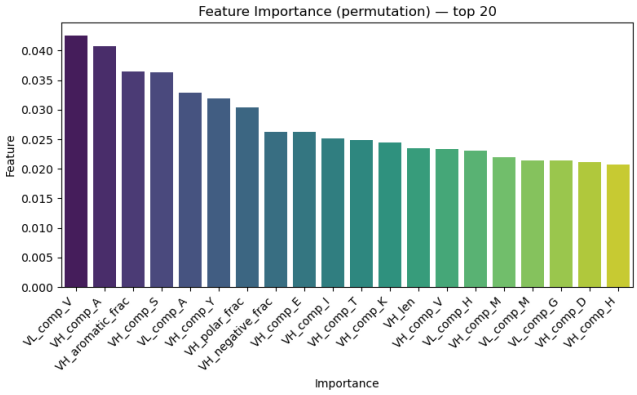
Removed 30% | Remaining: 37 | neg_root_mean_squared_error: -0.0523 ± 0.0100

Removed 40% | Remaining: 32 | neg_root_mean_squared_error: -0.0523 ± 0.0098

Removed 50% | Remaining: 27 | neg_root_mean_squared_error: -0.0519 ± 0.0095



SVR



Performing Ablation Analysis (permutation importance)...

Removed 0% | Remaining: 54 | neg_root_mean_squared_error: -0.0522 ± 0.0117

Removed 10% | Remaining: 48 | neg_root_mean_squared_error: -0.0517 ± 0.0115

Removed 20% | Remaining: 43 | neg_root_mean_squared_error: -0.0513 ± 0.0112

Removed 30% | Remaining: 37 | neg_root_mean_squared_error: -0.0510 ± 0.0112

Removed 40% | Remaining: 32 | neg_root_mean_squared_error: -0.0505 ± 0.0111

Removed 50% | Remaining: 27 | neg_root_mean_squared_error: -0.0501 ± 0.0110

- Failure analysis
  - 3 specific examples of prediction fails

Without outlier removal

=== SVR ===

Best Params: {'C': 0.1, 'epsilon': 0.001, 'gamma': 0.1, 'kernel': 'rbf'}

Best Score: 0.09213519880806656

Train RMSE: 0.110, Train R^2: 0.266

Test RMSE: 0.057, Test R^2: -0.131