BIOE20: AI-guided Computational Experimental Design for Active Data Selection in Protein Developability Models

Arinze Ezeifeka, Blake Gilbert, Gabriel Lipman, Janet Mwebi
Advisors: Lan Ma, PhD; Valentin Stanev, PhD

BIOCOMPUTATIONAL ENGINEERING

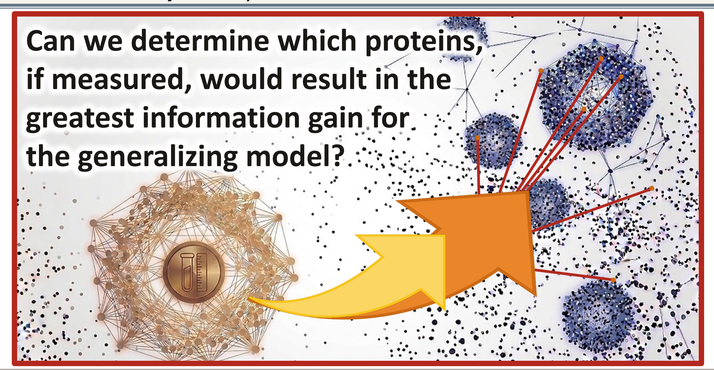
A. JAMES CLARK SCHOOL OF ENGINEERING

INTRODUCTION

- **Antibody development.** Key to medical advancements, but antibodies are very costly to develop, even in small amounts for testing.
- **Machine learning.** Used to predict developability. Accurate models require a large, descriptive and high-quality training data. Most data sets are **not publicly available** and **costly to create from scratch**.

Measuring 10K sequences = $1.5-$8 million[1]

- **Generalization.** Following R&D of one protein, it and nearby sequences are well defined, however this data must then be used to generalize to the next unknown region. Beyond previous research, sequences are **added randomly** as they become available.

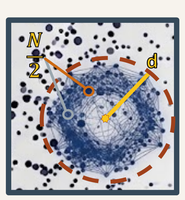


METHODOLOGY

- **Dataset.** 134,302 unique nanobody sequences, categorized as low or high polyreactive based on binding to a polyspecificity reagent.[2]
- **Computational Experimental Design.** Dry-lab simulation of experimental data acquisition. Generate artificial 'clusters' of sequences to mimic real initial data sets, then batch add remainders to examine information gain.
- **Levenstein Distance (L.D.) Matrix.** Precalculated between all sequences within 10K experimental subsets. Normalized.

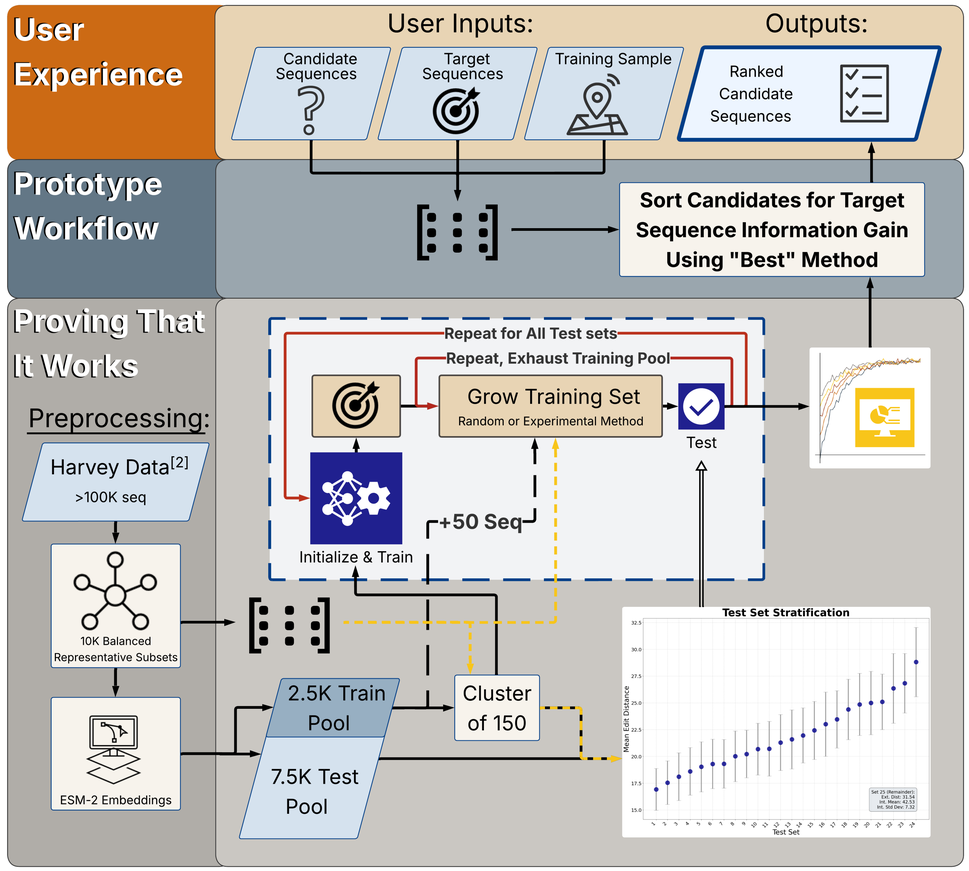
AGPHQT
AAPHQT

- **"Cluster" Generation.** For a seed seq, a neighborhood of ***d*** L.D. such that N/2 sequences of each class exist. A sample of N is taken randomly from neighborhood.



- **Classifier Models.** Logistic regression and Random Forest models were used for binary classification of polyreactivity due to fast training and to verify our *prototype is model agnostic*.
- **ESM-2 Embeddings.** 650M; 1280-dimension vectorization of full sequences. Designed to create biologically meaningful representations for the models.

WORKFLOW



DISCUSSION

Random Sampling (Baseline)
[Passive selection: samples are added in fixed-size batches without model feedback.]

Selection rule: $x_i \sim Uniform(X_{unlabeled})$

Active Neighborhood Selection (Uncertainty + Locality)
[Prioritizes samples the model is most uncertain about]

Scoring function: $score(x) = uncertainty(x) + \lambda \cdot d(x, medoid)$

Selection rule: $x^* = arg_{x \in X_{unlabled}} \, min \, score(x)$

Hybrid Neighborhood Selection (Uncertainty + Global Diversity)
[Balances exploitation (uncertainty) with exploration (diversity)]

Scoring function: $score(x) = \alpha \cdot uncertainty(x) + \beta \cdot d(x, X_{Train})$

Selection rule: $x^* = arg_{x \in X_{unlabled}} \, max \, score(x)$

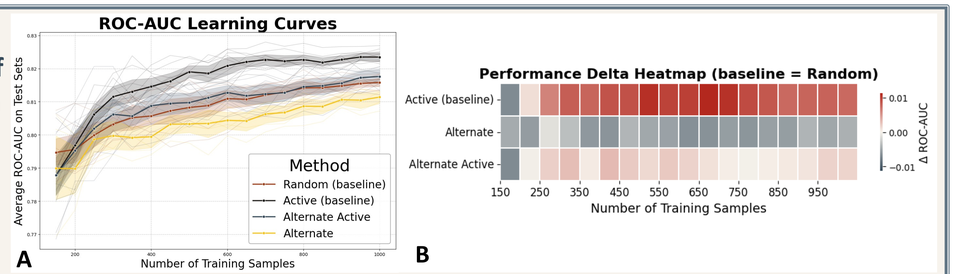
Random sampling with **dynamic learning**, where the model iteratively selects the most informative new vs. 50 random points. At each step, it identifies **uncertain predictions** and uses clustering (e.g., KMeans or KMedoids) to ensure the selected samples are also **diverse and non-redundant**. This leads to faster learning and better performance by focusing on data points that improve the model the most while covering different regions of the dataset. Diversity determined using a combination of cosine similarity & LD.

FUTURE WORK

- **More Runs**. For statistical validation. Model types. Selection methods. Verify combinations.
- **Parameter Variations.** Allow user to enter features such as initial training set size and desired output size for result optimization.
- **Piecewise Methods.** Combine methods in a piecewise schedule by initial size and sequence additions.
- **ML Selection.** Create data frame of points added and run Information Gain to as input for a ML model to identify deeper biological and structural selection criteria.

RESULTS

Performance comparison of data selection strategies against a random baseline.



(A) *ROC-AUC learning curves*. Over increasing training set sizes. Solid lines represent the mean performance over multiple seeds (individual runs shown as semi-transparent lines)

(B) *Delta heatmap.* Difference in mean ROC-AUC relative to the Random control (Δ = method – Random). Positive values indicate improvement over random sampling. Negative values indicate degradation. Neutral cream diverges at zero to strong dark slate for large positive deltas. Results demonstrate that Neighborhood Active and Hybrid Selection consistently outperform the baseline at different data set sizes.

ETHICS

Open Source | Low Cost | Healthcare | Transparency

ACKNOWLEDGEMENTS

Our Group would like to acknowledge the conception of this research question by Dr. Valentin Stanev of AstraZeneca and thank him for his time and guidance. We would also like to thank Dr. Lan Ma, our program's faculty advisor, as well as the other faculty of the Biocomputational Program who supported us through the completion of this project.

REFERENCES

[1] Subas Satish, Hema Preethi, et al. "NAb-Seq: An Accurate, Rapid, and Cost-Effective Method for Antibody Long-Read Sequencing in Hybridoma Cell Lines and Single B Cells." MAbs, vol. 14, no. 1, 13 Aug. 2022, https://doi.org/10.1080/19420862.2022.2106621.
[2] E. P. Harvey *et al.*, "An *in silico* method to assess antibody fragment polyreactivity," *Nat. Commun.*, vol. 13, no. 7554, 2022, doi: 10.1038/s41467-022-35276-4.
AI Use Statement: Google Gemini was used to generate elements used in images for this poster

AstraZeneca