Augmenting Low-Activity Activation Domain Sequences Using In Silico Directed Evolution and Machine Learning

Authors: Jasleen, Skye Leng, Grayson, Qamil Mirza

Introduction

Activation domains (ADs) are short, disordered protein fragments capable of enhancing gene expression by recruiting components of the transcriptional machinery. Despite their importance, the relationship between AD sequence and function remains poorly understood due to the lack of conserved motifs and their intrinsic disorder. Traditional experimental screening approaches for identifying functional ADs are time-intensive and may fail to uncover subtle, high-dimensional sequence-function patterns. This project investigates whether machine learning models can be used alongside feature embeddings generated by large protein language models, in particular the Evolutionary Scale Model 2 (ESM2) developed by Facebook Artificial Intelligence Research (FAIR), to accurately predict AD activity and consequently be used to guide the design of synthetic ADs with higher functional output.

Our central research question is: Does leveraging pre-trained protein language model embeddings improve the prediction of AD activity? Following this, are we then able to use this to discover subtle sequence-function relationships for designing these 40 amino acid protein fragments?

To address this, we focused on a dataset of 40 amino acid protein fragments with experimentally measured activity in both a *Saccharomyces cerevisiae* glucose (SCglucose) and *Saccharomyces cerevisiae* galactose (SCgalactose) reporter system. Many of these sequences exhibit low or

undetectable activity, challenging sequence-based prediction models. Our work trains regression models on three primary types of features:

(1) Pre-Computed Features

(2) Pre-Computed Features + ESM2 Feature Embeddings

(3) ESM2 Feature Embeddings

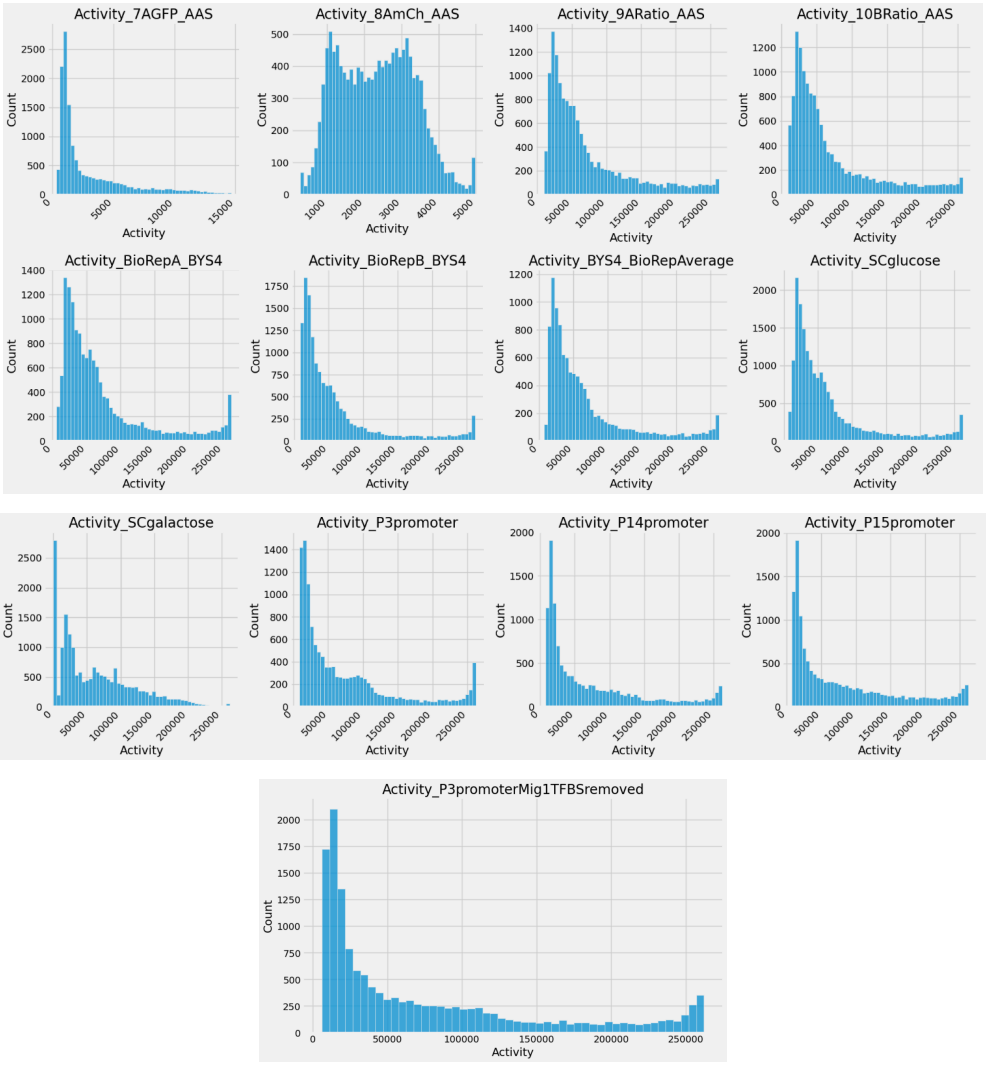
We then select the best-performing model and perform in silico directed evolution to carry out pointwise mutagenesis to identify new candidate AD sequences predicted to have higher AD activity. We then provide a further analysis of the pointwise changes made, the changes in net charge, and net hydrophobicity. Additionally, we provide a residue disorder analysis computed with MetaPredict to ensure the intrinsic disordered nature of the protein is maintained. (Qamil Mirza, Grayson, Jasleen)

Warm Up Analyses

Activity_SCglucose Calculation

The Activity_SCglucose column is calculated by taking the average of the Activity_BioRepA_BYS4 and the Activity_BioRepB_BYS4. In the case that one of the columns is missing, we just take the value of the column with a value for Activity_SCglucose.

Activity Histograms



The above histograms show that the activities are right-skewed and that most samples have low activity. It also shows that replicates A and B have good reproducibility and are consistent across the activities.

Feature Engineering on Amino Acid Sequences

Net Charge

The net charge of each tile is calculated by iterating through the AD sequence amino acid string, and cross-referencing the current amino acid with a pre-defined charges dictionary. The charges

dictionary adds 1 to the net charge each time a K or an R is encountered and subtracts 1 when a D or an E is encountered.

Amino Acid Counts

The amino acid counts are formed by iterating through the AD sequence amino acid string, and building a dictionary that keeps track of how many times each amino acid is encountered.

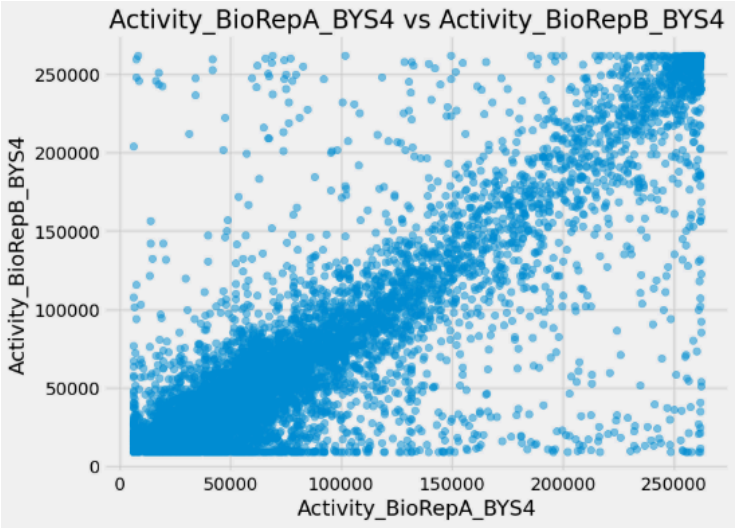
Motif Counts

Motif counts are computed using regex pattern matching. Given a list of motif counts, we look for matches in each AD sequence and increment the count for that motif if it is found. (Grayson, Qamil Mirza)

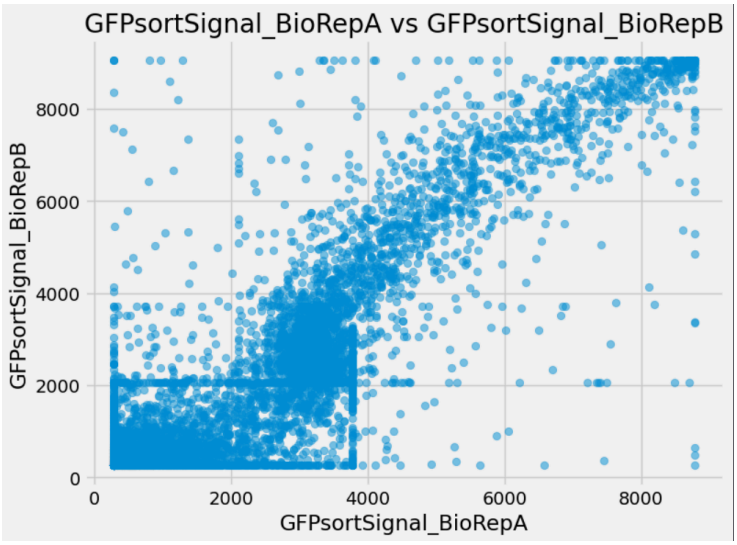
Summary of Biological Replicate Entries

# Measured in Both BioReps	# Measured only in BioRep A	# Measured only in BioRep B	# Missing From Both	# of Tiles
10852	4177	3915	94	19038

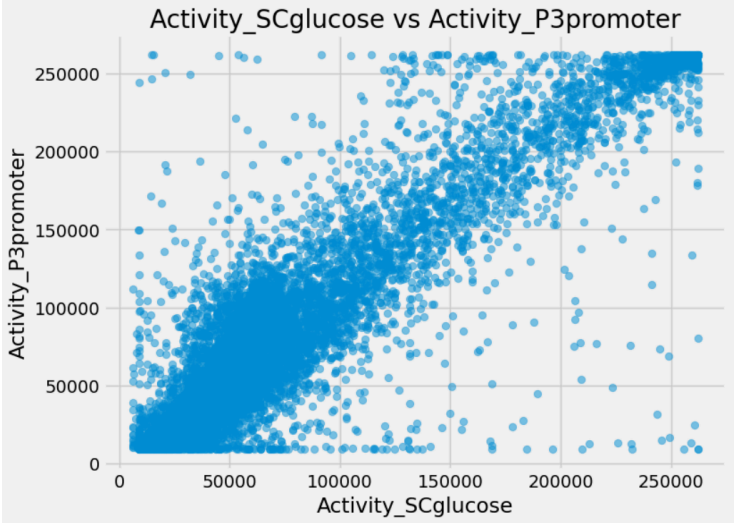
Activity_BioRepA_BYS4 vs Activity_BioRepB_BYS4



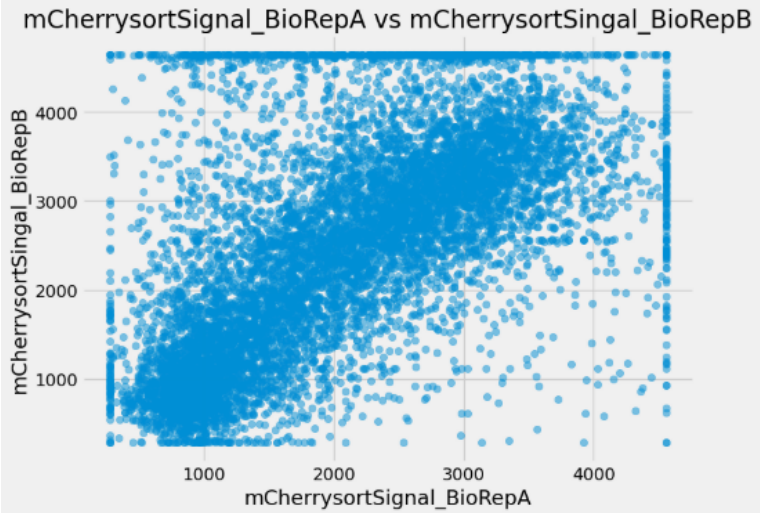
GFPsortSignal_BioRepA vs GFPsortSignal_BioRepB



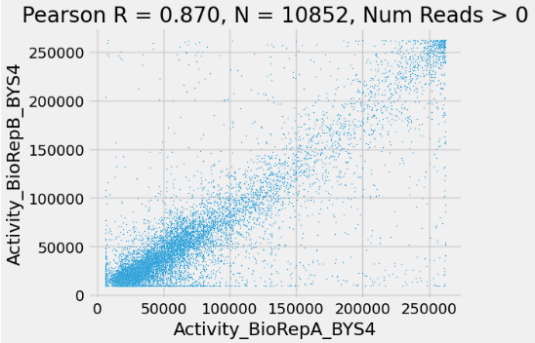
Activity_SCglucose vs Activity_P3promoter

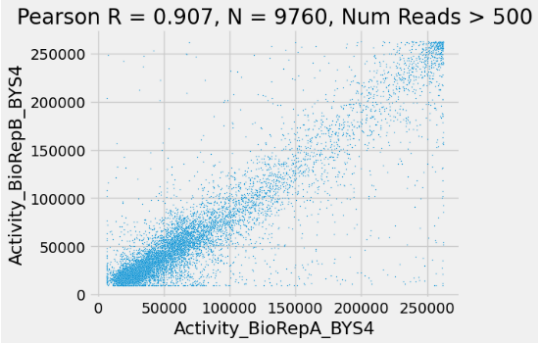


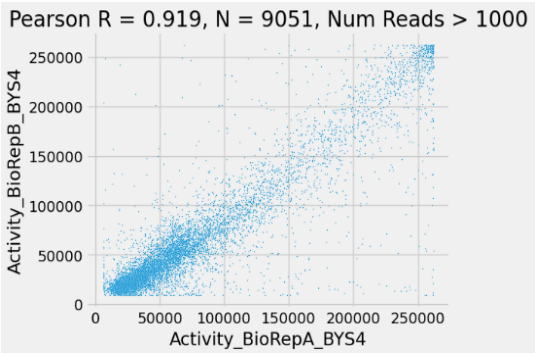
mCherrysortSignal_BioRepA vs mCherrysortSingal_BioRepB

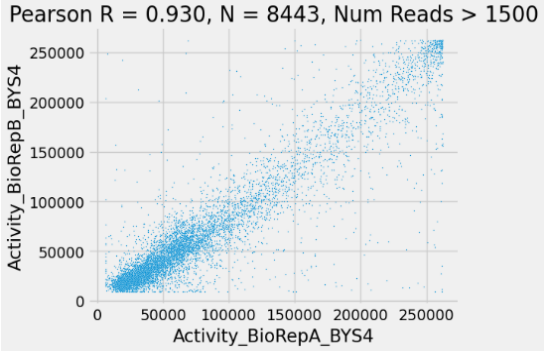


TotalReads_BioRepA_BYS4 and TotalReads_BioRepB_BYS4 Read Thresholding Correlation









Part I: Activity Predictors
Exploratory Data Analysis

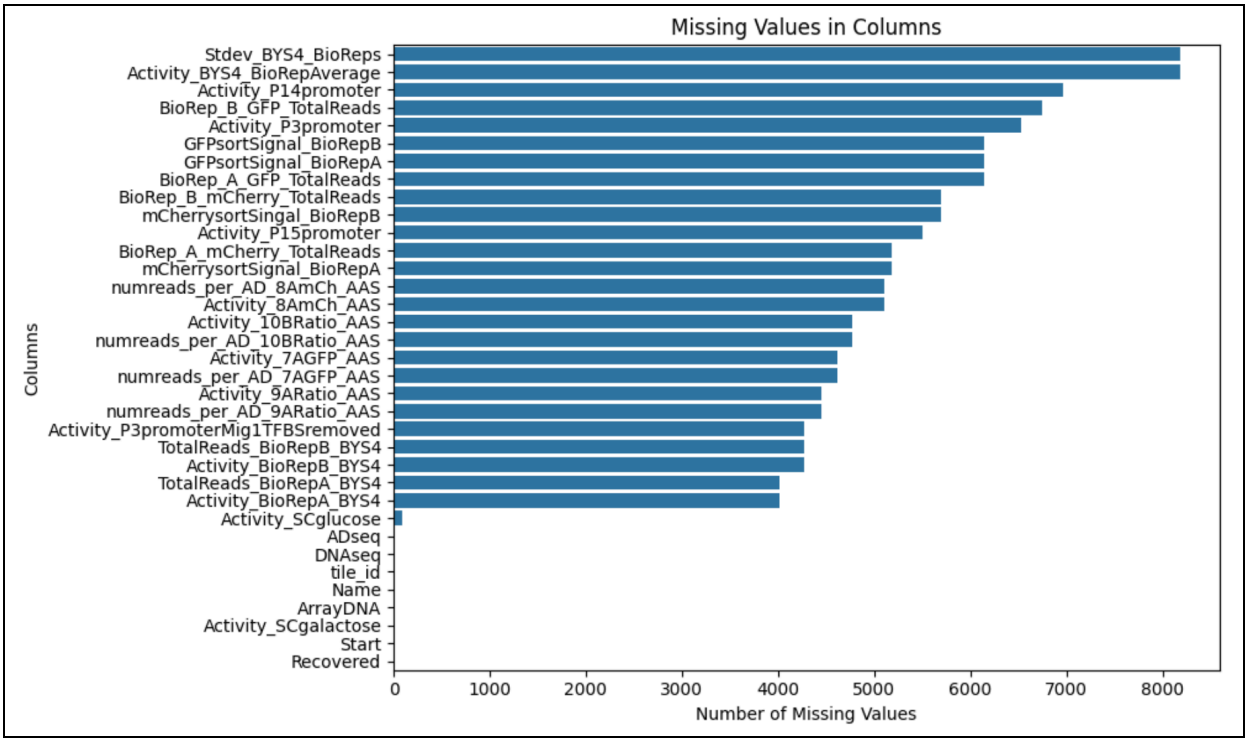


Figure 1: Missing Values in Columns

We start off by analyzing the dataset to determine the number of missing values for each column in the dataset. The x-axis refers to the number of missing values, and the y-axis represents the various columns in our dataset. We utilize a horizontal bar chart to sort the counts in descending order. It is important to recognize that these zero-activity values reflect activities that were not measured at all, rather than reflecting functionally silent activities. From this plot we can obtain and utilize the rows with higher sequencing reads to avoid noise when we begin our modeling. Based on the above graph, we chose to focus mainly on Activity_SCglucose and Activity_SCgalactose as our regressor targets, as they have the greatest amount of measured activity relative to other activities in the dataset. Below, we also create a correlation matrix to understand which features are highly associated with each other. (Qamil)

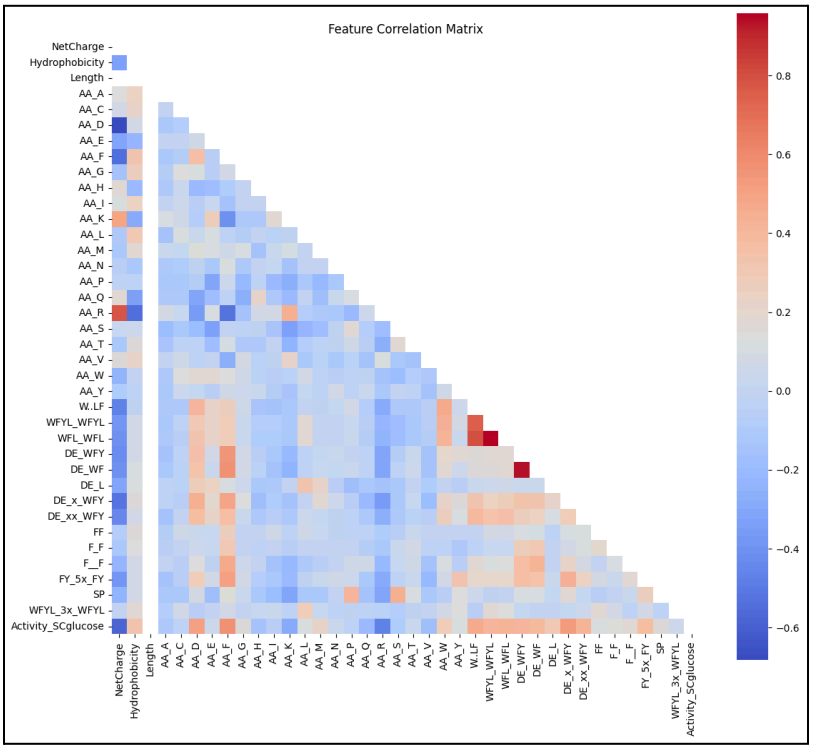


Figure 2A: Feature Correlation Matrix for Glucose

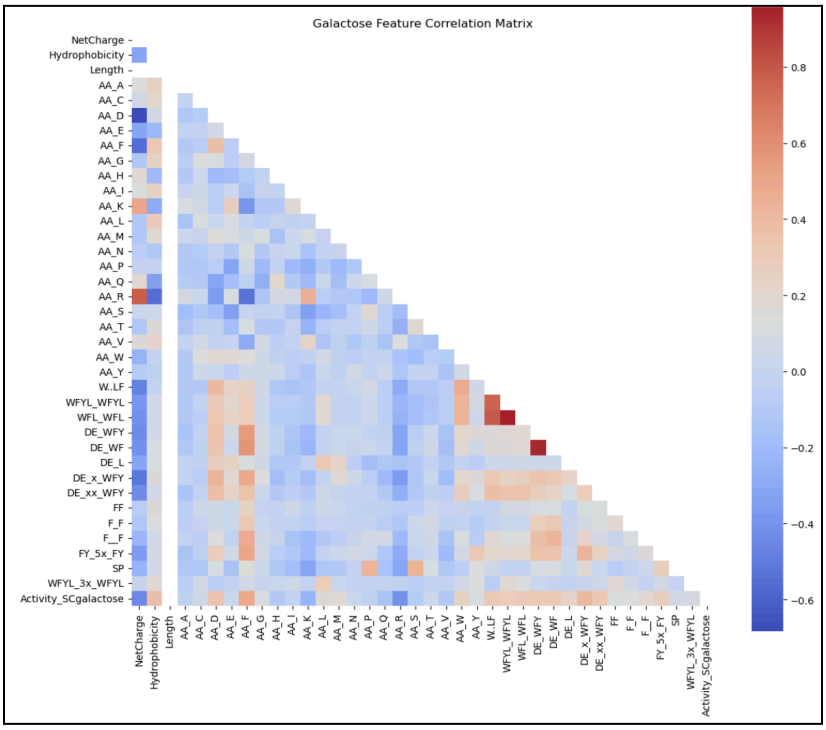


Figure 2B: Feature Correlation Matrix for Galactose

Figure 2: Feature Correlation Matrices

To explore the correlation between numerical values, we computed two Pearson correlation matrices and visualized them using heat maps. These matrices include the relationship between “Activity_SCglucose” and the numerical features derived from amino acid counts and predefined sequence motifs, as well as such relationship for “Activity_SCgalactose”. Each column and row in the matrices represents one of these features (Activity_SCglucose/ Activity_SCgalactose, an amino acid, or a motif frequency). The matrices visualize pairwise correlations, where values closer to +1 indicate a strong, positive relationship between the two variables and values closer to -1 indicate a strong, negative relationship between the two variables.

We chose all variables shown above as our raw features. But through these two figures we are able to identify the specific motifs or amino acids that have the strongest correlation with “Activity_SCglucose” (**Figure 2A**) and “Activity_SCgalactose” (**Figure 2B**). Ultimately, this helps uncover the features that hold most functional importance when analyzing the “Activity_SCglucose” and “Activity_SCgalactose”. (Qamil)

Feature Engineering

We compute two types of features: **sequence-inferred descriptors** and **protein language model embeddings.**

Sequence-inferred Features

These are hand-engineered biological features derived from the amino acid sequence:

1. **Net Charge:** Calculated based on the counts of positively and negatively charged residues at a physiological pH of 7.
2. **Hydrophobicity:** Computed using the Kyte-Doolittle scale, capturing the overall hydrophobic character of the peptide.
3. **Amino acid composition**: We include raw counts of each of the 20 standard amino acids.
4. **Motif counts**: Frequency of specific biologically relevant motifs which may correlate with functional activity.

These features incorporate known biophysical priors that can help with interpretation and ensure the model is learning biologically relevant features.

ESM-2 Embeddings

We use the ESM-2 protein language model with 8 million parameters to generate per-sequence embedding vectors. ESM-2 is trained on millions of evolutionarily diverse protein sequences to learn contextualized amino acid representations. We incorporate these embeddings to capture information such as:

1. Long-range dependencies in the sequence
2. Structural or functional similarity that may extend beyond simple motif detection
3. Evolutionary priors that are not easily engineered by hand

We then extracted the mean-pooled embedding for each sequence to obtain a fixed-length representation suitable for downstream modelling

Feature Combination

Finally, both types of features are concatenated and input into our predictive models. This hybrid approach enables the model to leverage hand-engineered heuristics while also utilizing more complex learned representations for a more robust prediction. (Qamil Mirza, Grayson, Skye, Jasleen)

Activity Predictor Model Development and Results

We then train 6 regression models for each of the 3 types of features we computed above. Our work found that the Random Forest Regressor model generally performed the best for both targets. Furthermore, Random Forest performed the best when using only the pre-computed features for “Activity_SCglucose”, and adding ESM2-generated features to the dataset slightly degraded model performance for the “Activity_SCglucose” target. In contrast, we found that adding ESM2-generated features to the dataset for the “Activity_SCgalactose” improves the model performance slightly relative to only using the pre-computed features. Overall, utilizing only the ESM2-generated embeddings for training yielded poorer model performance across both targets. (Skye, Grayson)

Part I: More Figures & Tables

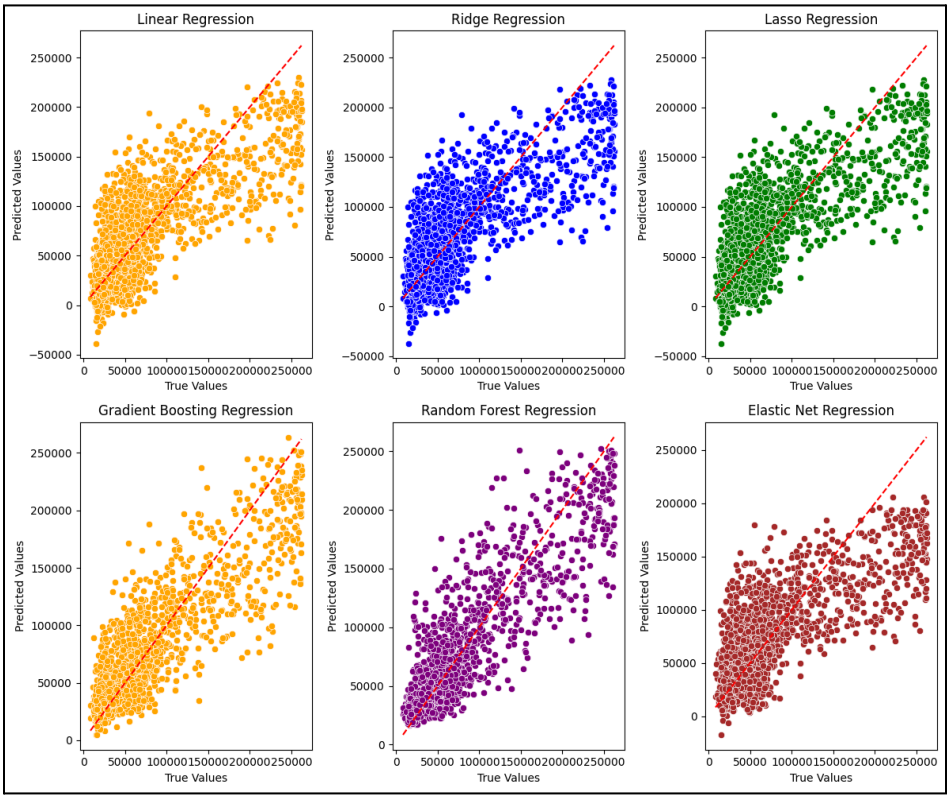


Figure 3A: Regression Models for Glucose with Only Raw Features

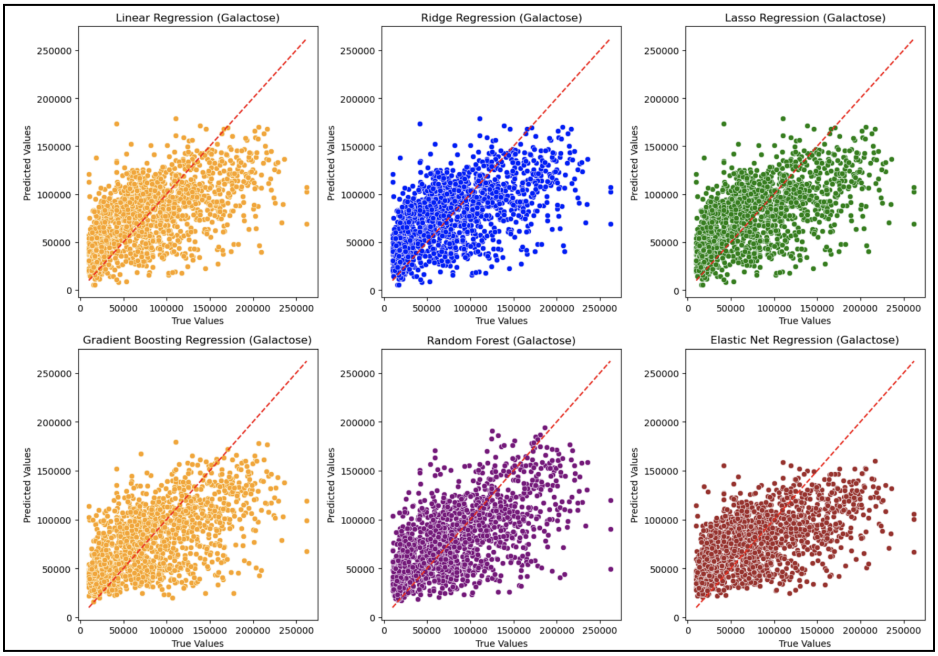


Figure 3B: Regression Models for Galactose with Only Raw Features

Figure 3: Regression Models (Only Raw Features)

We split our data into a training set (80%) and a test set (20%). Since the target values are much higher than our engineered features, we used Z-score normalization to scale it. Then we built 6 multiple baseline models, and visualized their performances. Each red line in the figure means that the predicted activity value is equal to the actual value.

From **Figure 3A** (glucose), we can roughly see that Random Forest and Gradient Boosting make more accurate predictions close to the real activities (more dots close to the red line), and Elastic Net Regression does not work well on our dataset. However, more numerical results should be illustrated (**Table 1A**). And we considered Random Forest to be the best, based on its higher R^2, Pearson and lower MAE, RMSE.

From **Figure 3B** (galactose), we can see that the scatter is large for all models, as the points are mostly distant from red lines. The performance is also visually slightly better for the tree-based models (Gradient Boosting Regression and Random Forest Regression), especially the Random Forest model as more points clustered along the red line. These visualizations are also analyzed with numerical results (**Table 1B**). Similar to the results for glucose, the Random Forest model has the highest R^2, strongest Pearson correlation, and lowest MAE and RMSE. These results converge to the conclusion that the Random Forest model is the best among all six. (Grayson)

Table 1: Only Raw Features

Model	R^2	MAE	RMSE	Pearson
Linear Regression	0.6195	29117.0950	39797.4656	0.7875
Ridge Regression	0.6201	28986.7738	39766.8844	0.7876
Lasso Regression	0.6201	28987.1442	39767.0803	0.7876
Gradient Boosting	0.7537	22520.3700	32021.1280	0.8700
Random Forest	**0.7965**	**20034.6822**	**29109.0641**	**0.8931**
Elastic Net	0.5959	29412.5192	41017.2696	0.7758

Table 1A: Performance comparison of training models with only raw features (Glucose)

Model	R^2	MAE	RMSE	Pearson
Linear Regression	0.3900	32843.5195	42034.3811	0.6252
Ridge Regression	0.3900	32844.0871	42034.6595	0.6252
Lasso Regression	0.3900	32843.9924	42034.3889	0.6252
Gradient Boosting	0.4080	32322.8506	41409.9851	0.6394
Random Forest	**0.4289**	**31196.1936**	**40673.9173**	**0.6553**
Elastic Net	0.3708	34093.2495	42689.9072	0.6131

Table 1B: Performance comparison of training models with only raw features (Galactose)

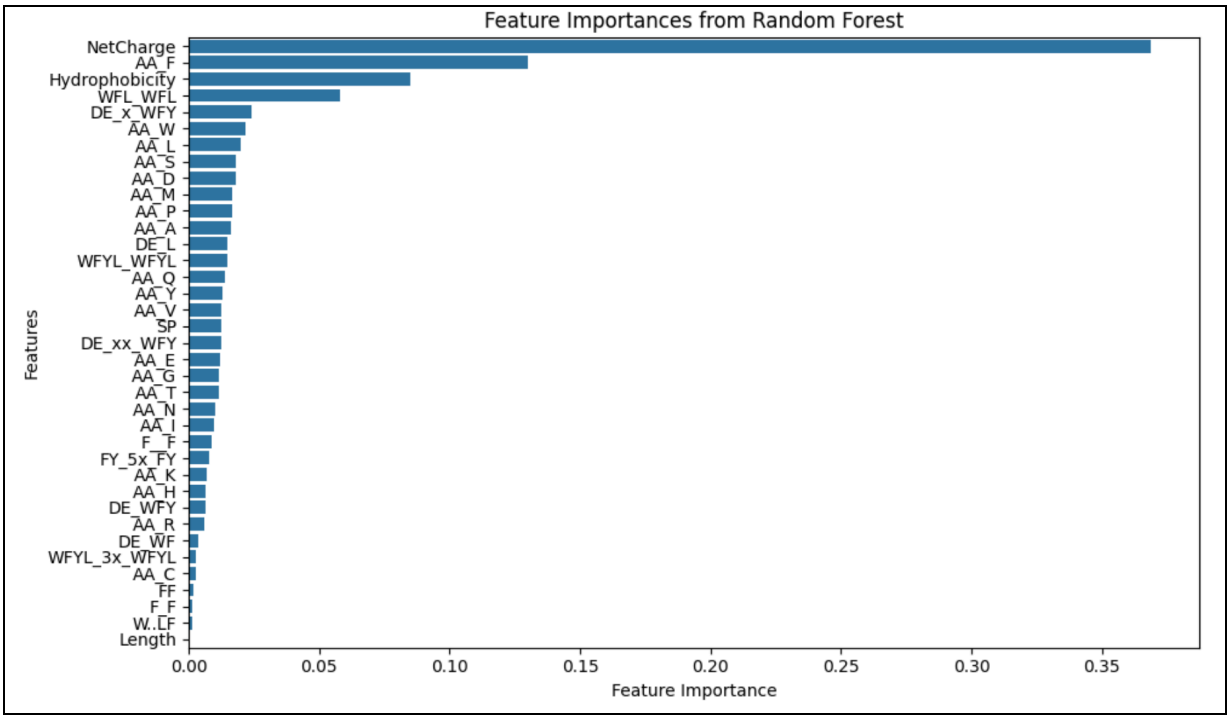


Figure 4A: Feature Importance from Random Forest for Glucose (Raw Features)

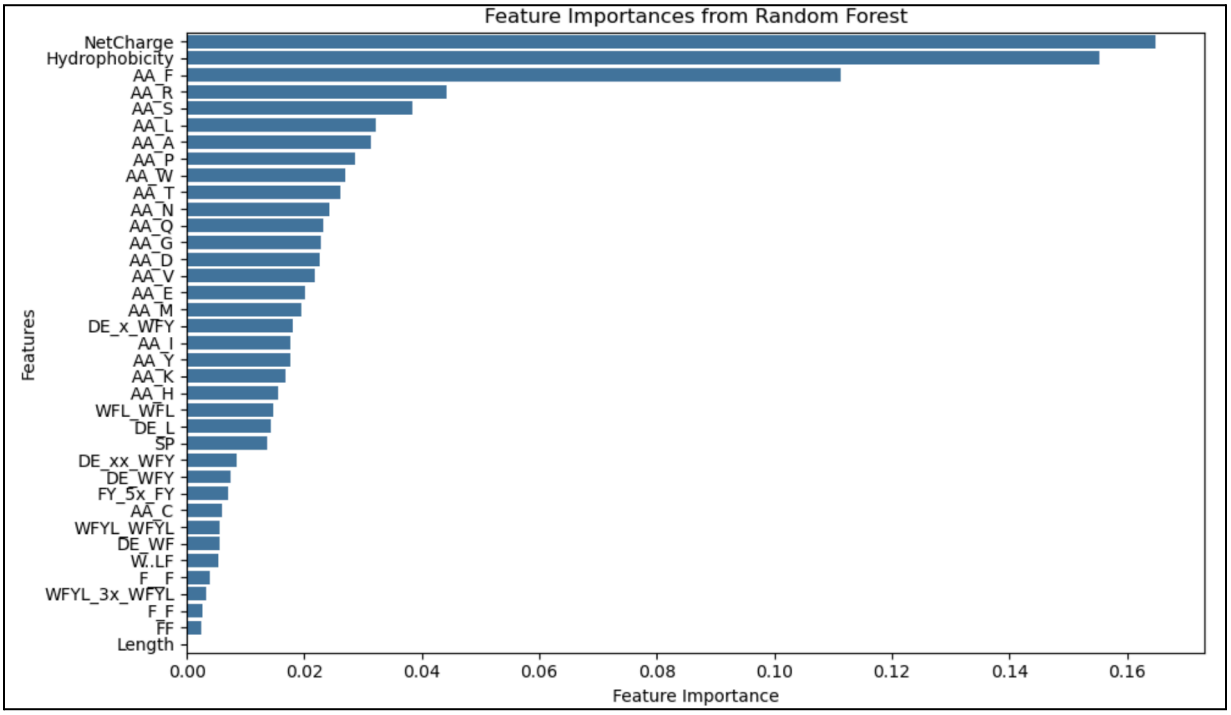


Figure 4B: Feature Importance from Random Forest for Glucose (Raw Features)

Figure 4: Feature Importance (Only Raw Features)

This portion of the analysis visualizes which features of the protein sequence, such as amino acid frequencies and motif counts, are the most predictive of SCglucose/SCgalactose using a Random Forest Regression model. The input data included amino acid counts, structural predictions, and predefined motif sequences. After we trained our Random Forest Regression model, feature importance scores were obtained to quantify the importance of each feature with regards to the model's predictions. These features were then displayed in descending order of importance using a horizontal bar plot. The x-axis represents feature importance and the y-axis represents the features themselves.

Figure 4A supports the conclusion that certain features have a stronger predictive power for "Activity_SCglucose", the "Net Charge" and "AA_F" features depicting the highest feature importance in this case. Similarly, **Figure 4B** supports that certain features have a stronger predictive power for "Activity_SCglucose:" the "Net Charge" and "Hydrophobicity" features depicting the highest feature importance in this case. (Skye, Jasleen)

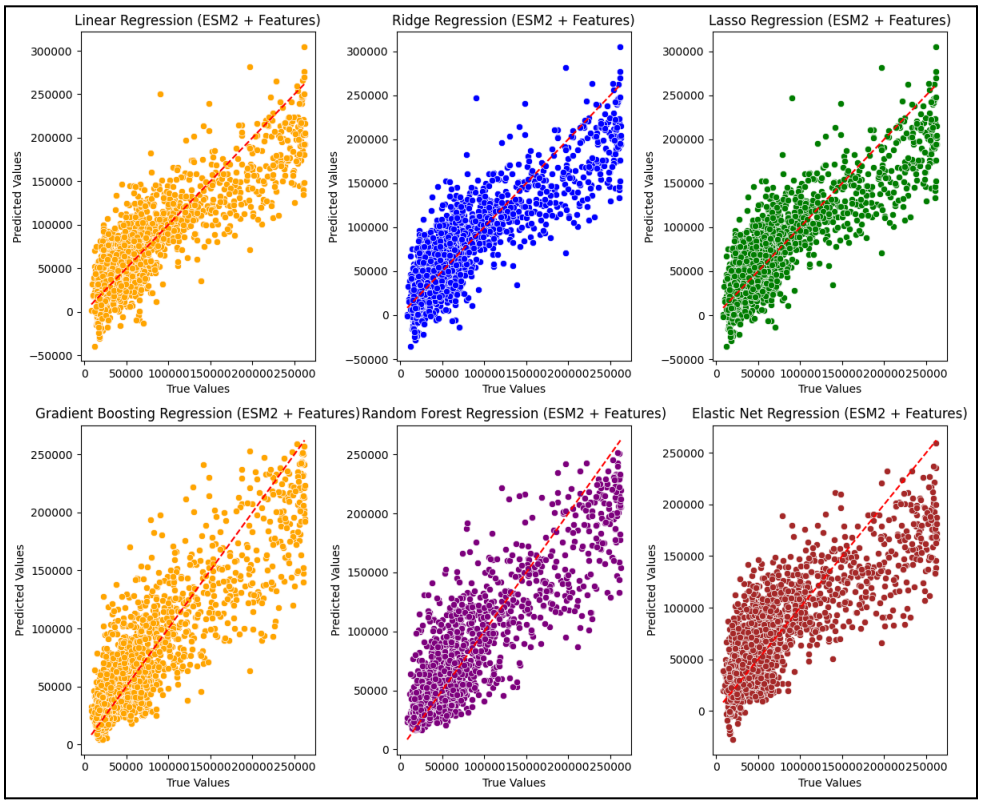


Figure 5A: Regression Models for Glucose with ESM2 and Features

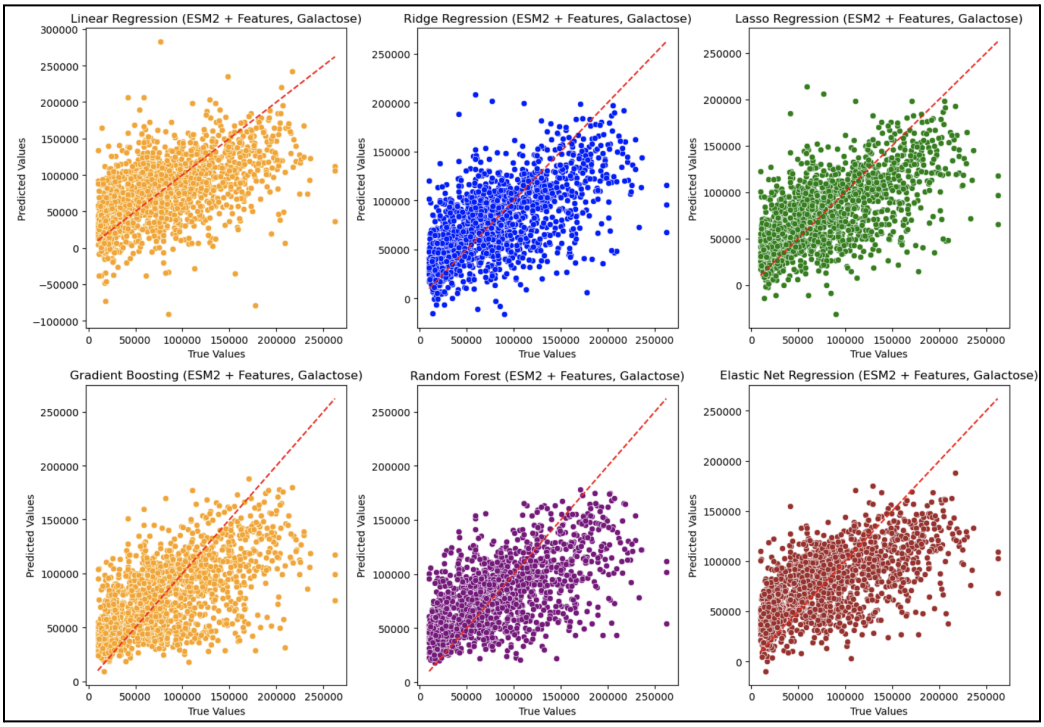


Figure 5B: Regression Models for Galactose with ESM2 and Features

Figure 5: Regression Models (ESM2 + Features)

Before training with new features, we utilized the 8M parameter version **Evolutionary Scale Model 2** to generate more embeddings to our sequences. This model captures evolutionary and structural features from protein sequences, and we hoped it enhanced the overall feature representation. Then we used 6 multiple baseline models to learn these new features again. Each red line in **Figure 5A** means that the predicted activity value is equal to the actual value. From the figure above, we can roughly see that each model shows good performance on predicting the activities of new protein sequences. But according to the numerical results (**Table 2A**), we finally chose Random Forest based on its higher R^2, Pearson and lower MAE, RMSE. In **Figure 5B**, the predictions visually aligned more closely with the diagonal line, indicating improved performance across most models. According to the numeric metrics shown in **Table 2B**, Random Forest again outperformed the other models with the highest R^2, strongest Pearson correlation, and the lowest MAE and RMSE. These results suggest that adding domain-specific features helped all models improve slightly, and re-confirmed Random Forest as the most effective regression model. (Grayson, Skye)

Table 2: ESM2 + Features

Model	R^2	MAE	RMSE	Pearson
Linear Regression	0.7491	24150.9297	32317.0708	0.8656
Ridge Regression	0.7512	24034.4200	32185.2868	0.8782
Lasso Regression	0.7511	24038.6443	32190.0767	0.8667
Gradient Boosting	0.7704	21776.7985	30919.2553	0.8782
Random Forest	**0.7835**	**21023.8345**	**30019.5091**	**0.8880**
Elastic Net	0.6830	26313.43846	36329.06839	0.8285

Table 2A: Performance comparison of training models with the raw features and ESM embeddings (Glucose)

Model	R^2	MAE	RMSE	Pearson
Linear Regression	0.2354	35825.2727	47062.1330	0.5585
Ridge Regression	0.4182	31288.1602	41051.0628	0.6510
Lasso Regression	0.4180	31299.8515	41059.5701	0.6510
Gradient Boosting	0.4199	31707.2751	40991.1811	0.6487
Random Forest	**0.4332**	**31497.8912**	**40519.4916**	**0.6593**
Elastic Net	0.4072	32442.1349	41439.7177	0.6394

Table 2B: Performance comparison of training models with the raw features and ESM embeddings (Galactose)

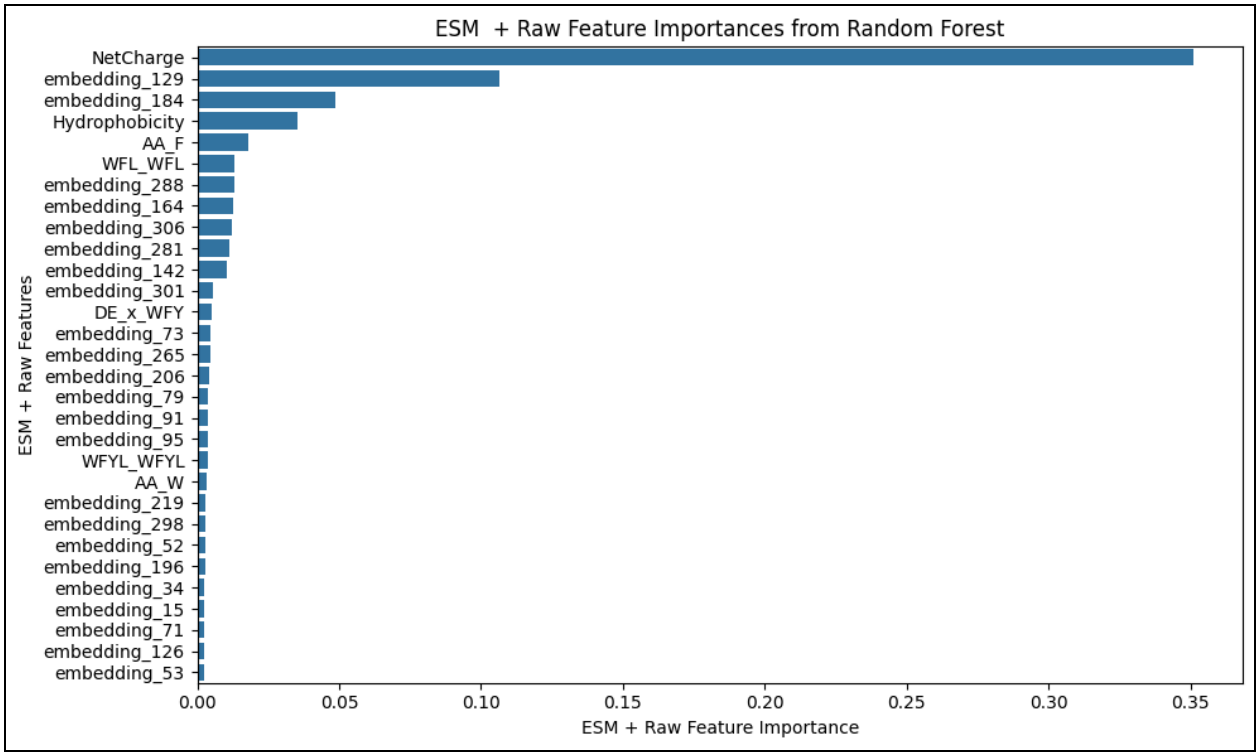


Figure 6A: Feature Importance from Random Forest for Glucose (ESM + Raw Features)

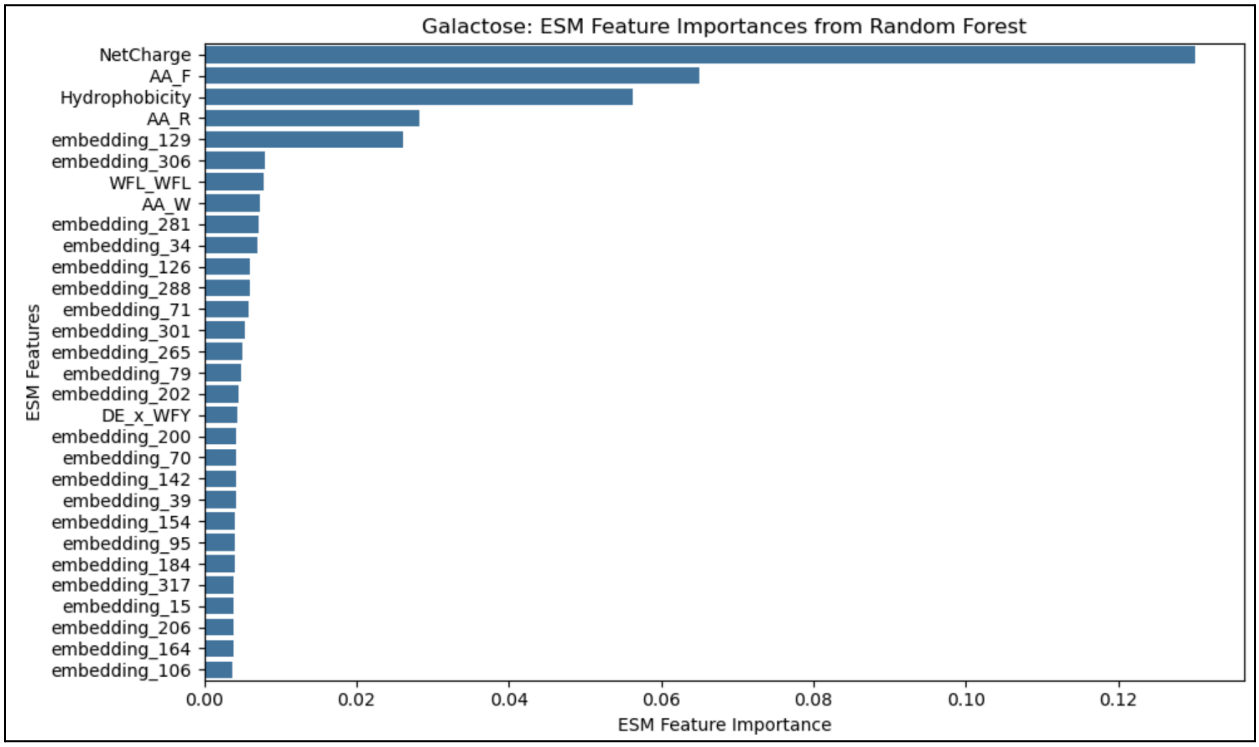


Figure 6B: Feature Importance from Random Forest for Galactose (ESM + Raw Features)

Figure 6: Feature Importance (ESM2 + Features)

This portion of the analysis visualizes which ESM2-derived (Evolutionary Scale Modeling) features, in conjunction with amino-acid level features, are the most predictive of SCglucose using a Random Forest Regression model. The input data includes protein sequence embeddings generated by the ESM2 model, which takes the protein sequence and returns embedding vectors corresponding to that sequence. In this case, embedding vectors involves biochemical and structural properties of the sequences. After we trained our Random Forest Regression model, feature importance scores were obtained to quantify the importance of each feature with regard to the model's predictions. The top 30 features were then displayed in descending order of importance using a horizontal bar plot. The x-axis represents ESM + Raw feature importance, and the y-axis represents the features themselves.

Figure 6A supports the conclusion that certain ESM + Raw features have a stronger predictive power for "Activity_SCglucose", the "Net Charge" and "embedding_129" features depicting the highest feature importance in this case. Similarly, **Figure 6A** supports that certain features have a stronger predictive power for "Activity_SCglucose:" the "Net Charge" and "AA_F" features depicting the highest feature importance in this case. (Jasleen)

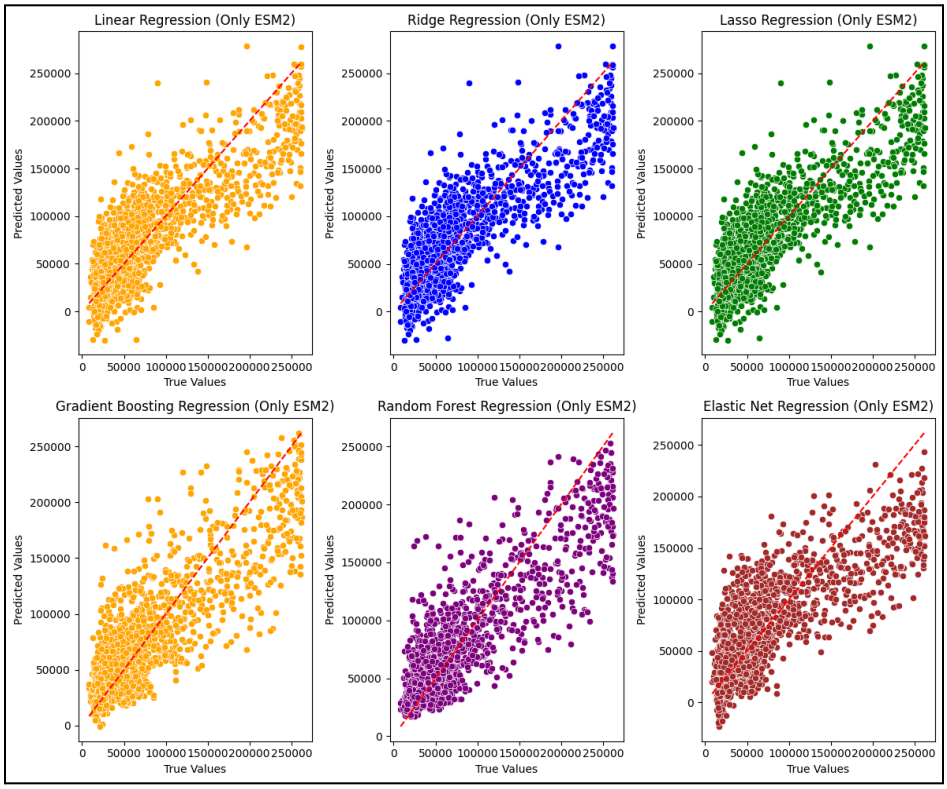


Figure 7A: Regression Models for Glucose with only ESM2

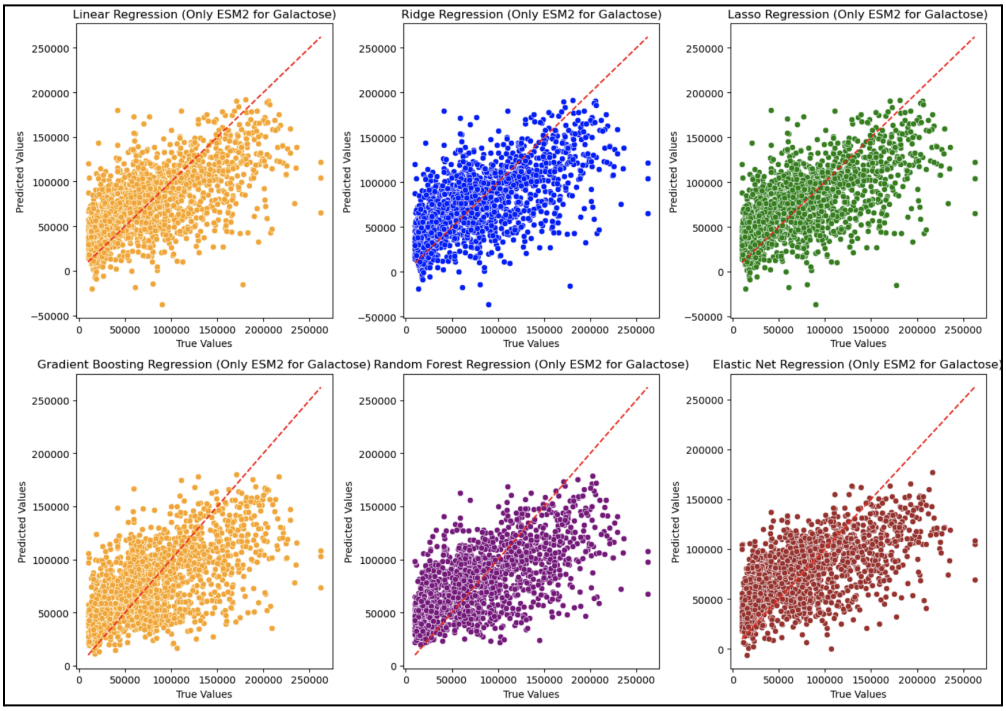


Figure 7B: Regression Models for Galactose with only ESM2

Figure 7: Regression Models (Only ESM2)

This portion of the project aims to explore the effect of solely using the ESM2 embeddings for model training. Similar to previous parts, we use six different regression models to make predictions.

Each red line in **Figure 7A** means that the predicted activity value is equal to the actual value. From the figure above, we can roughly see that each model shows good performance in predicting the activities of new protein sequences. Based on numerical results (**Table 3A**), we finally chose Random Forest based on its higher R^2, Pearson, and lower MAE, RMSE. By comparing **Table 3A** to **Table 1A**, we could conclude that only considering ESM2 also does not perform as well as only considering the raw features for glucose.

In **Figure 7B**, the regression plots reveal similar levels of scatter around the red reference line across all models, making it visually challenging to distinguish performance differences when using only ESM2 embeddings. However, the quantitative results in **Table 3B** show that Random Forest once again achieved the best overall performance, with the highest R^2, lowest MAE and RMSE, and the strongest Pearson correlation. Compared to the results in Table 2B (with raw features considered), all models perform slightly worse across all metrics. This decline highlights that relying solely on pretrained embeddings limits predictive accuracy, therefore, it is important to both integrate ESM2 and domain-specific features for optimal model performance. (Jasleen, Skye)

Table 3: Only ESM2 Embeddings

Model	R^2	MAE	RMSE	Pearson
Linear Regression	0.7348	24932.2937	33226.3968	0.8573
Ridge Regression	0.7352	24905.9094	33199.5024	0.8575
Lasso Regression	0.7351	24916.3528	33211.3067	0.8574
Gradient Boosting	0.7441	22890.4082	32636.9725	0.8633
Random Forest	**0.7624**	**21832.5965**	**31449.5534**	**0.8780**
Elastic Net	0.6530	27943.1754	38005.2545	0.8111

Table 3A: Performance comparison of training models on the test set with only ESM embeddings (Glucose)

Model	R^2	MAE	RMSE	Pearson
Linear Regression	0.4111	31659.8418	41303.3636	0.6459
Ridge Regression	0.4118	31641.6869	41276.8357	0.6464
Lasso Regression	0.4116	31646.2998	41283.4232	0.6463
Gradient Boosting	0.4096	32104.8712	41353.3909	0.6409
Random Forest	**0.4276**	**31798.4132**	**40720.3787**	**0.6557**
Elastic Net	0.3956	32797.0896	41840.0459	0.6306

Table 3B: Performance comparison of training models on the test set with only ESM embeddings (Galactose)

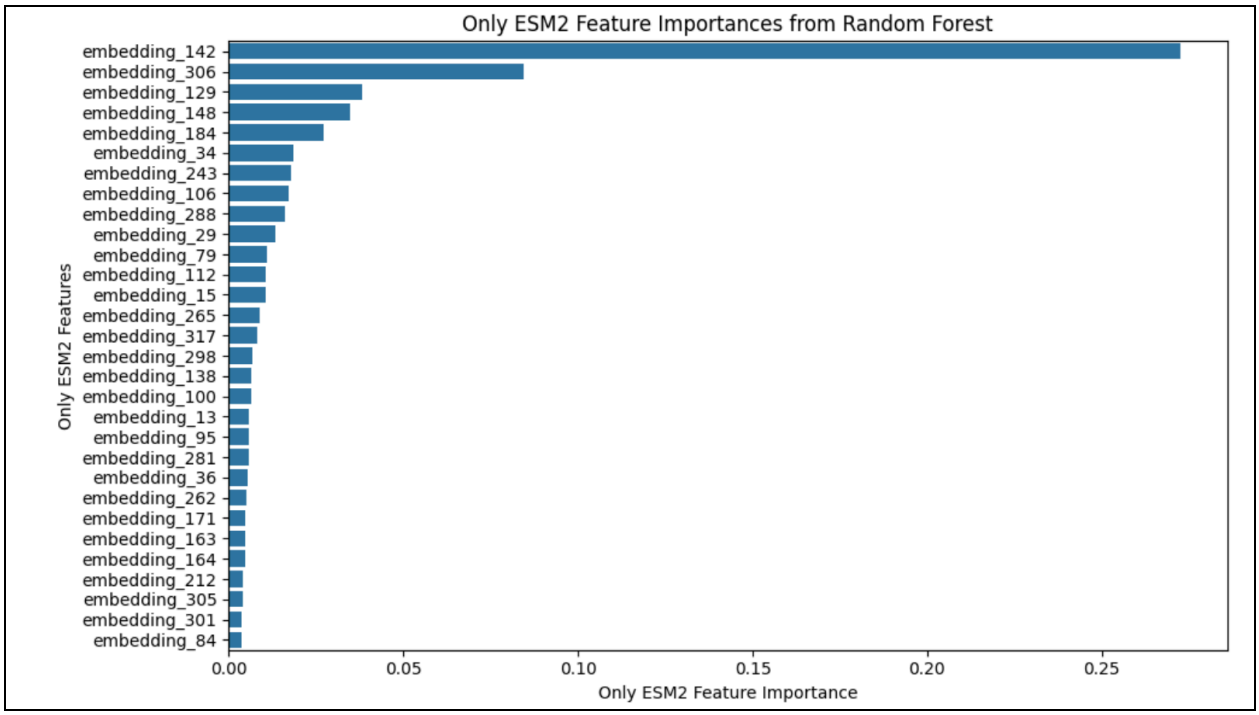


Figure 8A: Feature Importance from Random Forest for Glucose (Only ESM)

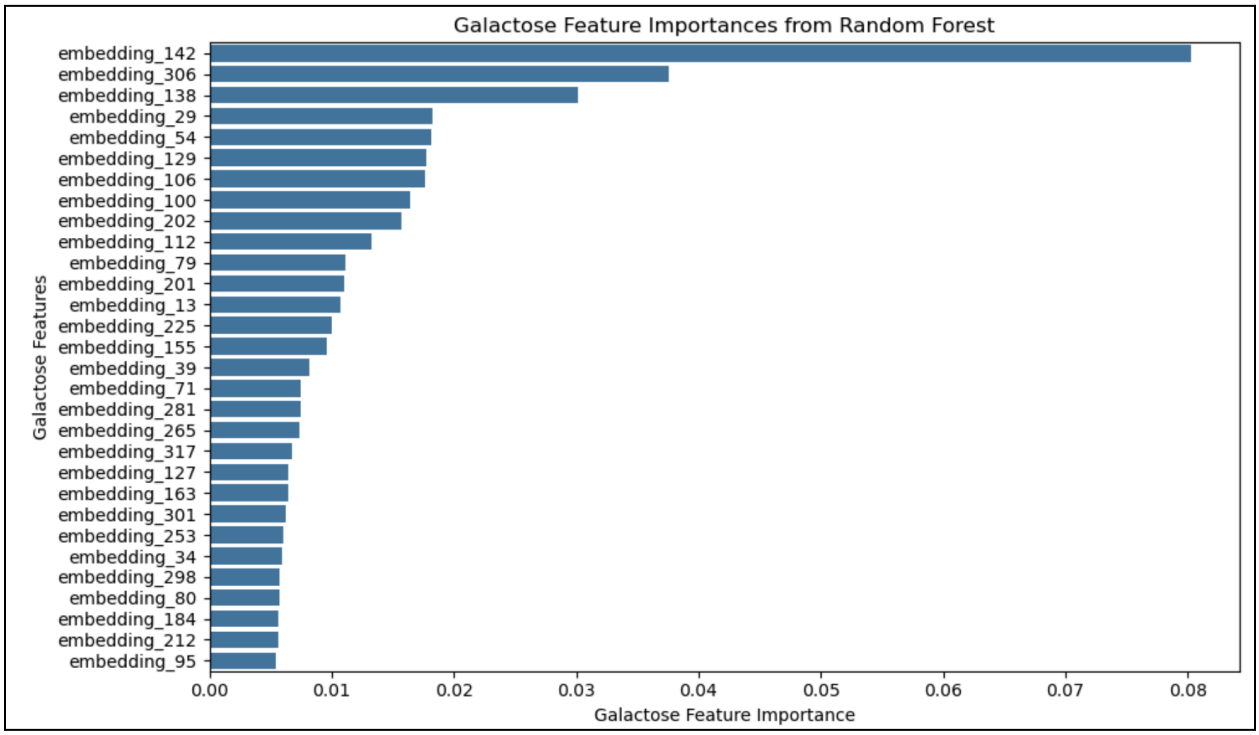


Figure 8B: Feature Importance from Random Forest for Galactose (Only ESM)

Figure 8: This portion of the analysis visualizes which ESM-derived (Evolutionary Scale Modeling) features, without the addition of amino-acid level features, are the most predictive of SCglucose/SCgalactose using a Random Forest Regression model. The input data includes protein sequence embeddings generated by the ESM2 model, which takes the protein sequence and returns embedding vectors corresponding to that sequence. In this case, embedding vectors involves biochemical and structural properties of the sequences. After we trained our Random Forest Regression model, feature importance scores were obtained to quantify the importance of each feature in the model's predictions. The top 30 features were then displayed in descending order of importance using a horizontal bar plot. The x-axis represents ESM2 feature importance, and the y-axis represents the features themselves.

Figure 8A and **Figure 8B** support the conclusion that certain ESM2 features have a stronger predictive power for "Activity_SCglucose" and "Activity_SCgalactose." The "embedding_142" and "embedding_306" features depict the highest feature importance in both cases. (Jasleen)

Part II: In Silico Directed Evolution

We implemented an *in silico* directed evolution pipeline to iteratively optimize sequences based on their intrinsic disorder properties. This procedure aims to mimic natural selection by applying random pointwise mutations across generations and selecting for favorable phenotypes, which in our case is high disorder. Starting from an initial low-activity sequence, we introduce random single-residue substitutions at randomly selected positions using a fixed list of the 20 standard amino acids. Each generation then produces a pool of unique mutant sequences, which are then evaluated for intrinsic disorder using MetaPredict, a neural network-based tool that predicts the per-residue disorder from the input amino acid sequences.

Furthermore, to maintain selective pressure, only sequences whose average disorder exceeds a threshold of 0.5 are retained (Emenecker, 2021). Among these, we keep only the top 5,000 mutants per generation to prevent efficiency bottlenecks due to our limited computational resources. We then compute the hand-engineered features for these filtered sequences and propagate them to the next generation for further mutation. Again, due to computational bottlenecks, our work only runs this simulation for 5 generations. The final dataset produced includes feature-annotated, evolved sequences stratified by generation, capturing the evolutionary trajectory from low disorder to high disorder. (Qamil Mirza)

Pseudocode Overview

Input:

- initial_sequence: str
- aa_list: list of amino acids
- add_features_fn: feature annotation function
- num_generations: number of evolution steps
- num_mutations_per_generation: number of mutations to sample per generation
- max_sequences_per_gen: max mutants to keep each generation
- min_disorder_threshold: filter cutoff for average disorder

Initialize:

- current_sequences ← [initial_sequence]
- all_featured_sequences ← empty list

For generation in 1 to num_generations:

1. Randomly sample mutation sites and replacement residues
2. Apply single-point mutations to generate mutant sequences
3. Filter mutants:
   - Predict disorder using MetaPredict
   - Keep mutants with mean disorder ≥ min_disorder_threshold
4. Limit total to max_sequences_per_gen mutants
5. Compute features using add_features_fn
6. Save annotated mutants
7. current_sequences ← filtered mutants

Return all_featured_sequences as the final dataset

Disorder Profile & Mutational Analysis

To evaluate the effectiveness of our in silico directed evolution on intrinsic disorder, we compare the per-residue disorder profiles of original low-activity sequences to those of their top candidate mutants. Using MetaPredict, we generated a disorder score curve for the top candidate sequence and the original sequence and analyzed whether the evolved variants maintained or enhanced disorder across the activation domain.

Beyond overall disorder, we also examined specific sequence-level changes introduced by mutagenesis. By comparing the original and top candidate sequences, we identified key substitutions and quantified changes in net charge and hydrophobicity. This analysis allowed us to assess not only whether the disordered nature of the sequence was conserved, but also whether these biochemical shifts might explain the improved predicted activity. (Grayson, Qamil Mirza)

Glucose

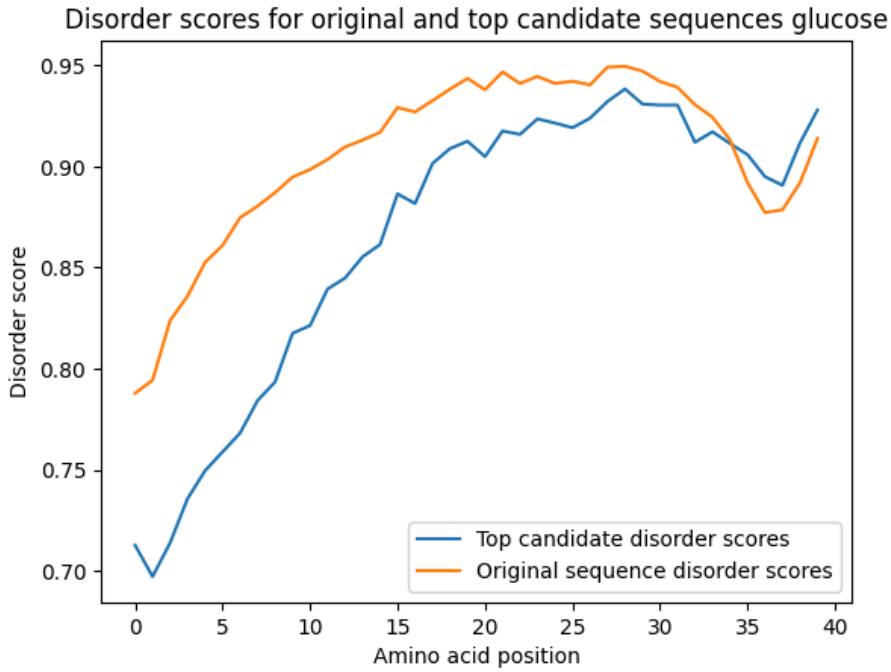


Figure 9A: MetaPredict predicted disorder score comparison between a chosen sequence and the top mutated candidate of that sequence for “Activity_SCglucose”. This figure shows that in silico directed evolution with the optimality constraint, keeping only mutations that have a mean disorder score of above 0.5, is met. We wanted to ensure that after mutagenesis, the intrinsic disorder of the activation domain sequences is maintained. This figure concludes that we successfully enforced disorder into our proposed candidate mutations. (Skye)

Attribute	Original Sequence	Top Candidate Sequence	Difference
Sequence	DFVLFDSPQPQRTT VNRPSSVPSNSAA PFGSLQSNTTSTN	DFVLFDSPQPQRD TVDRPSSPPSNSAA PFGDLFSNTTSTN	-
Point Mutations	-	**pos 12:** T → D **pos 15:** N → D **pos 20:** V → P **pos 30:** S → D **pos 32:** Q → F	5 substitutions
Net Charge	0	-3	-3
Hydrophobicity	-26.6	-31.6	-5.0
Experimental Activity	6138	-	-
Predicted Activity	-	122, 001.8	+ 115, 863.8

Table 4A: Comparison between original and top candidate AD sequences. Point mutations led to significant predicted increases in activity, alongside shifts in net charge and hydrophobicity.

Galactose

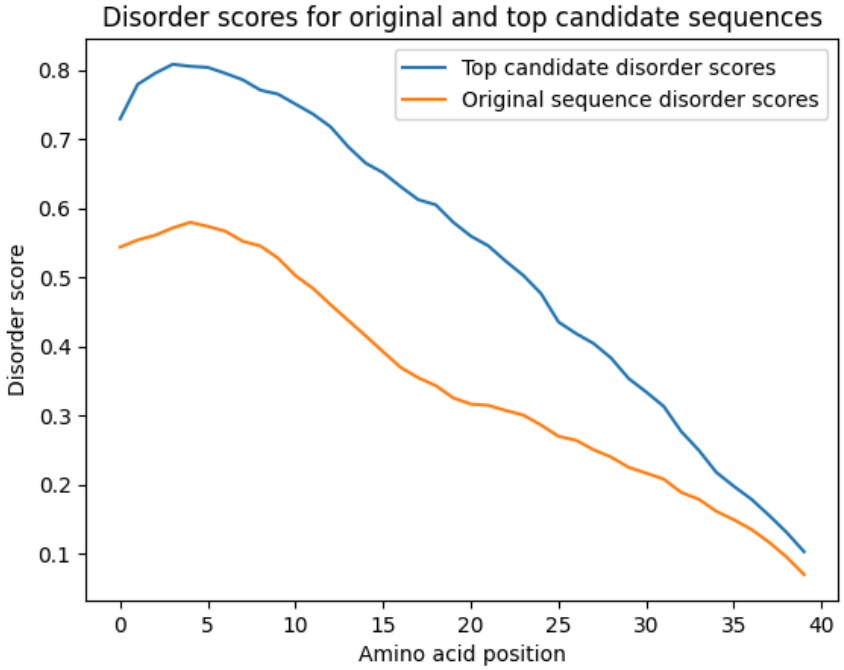


Figure 9B: MetaPredict predicted disorder score comparison between a chosen sequence and the top mutated candidate of that sequence for "Activity_SCgalactose." Similar to the glucose case, the top candidate's disorder profile somewhat mirrors the disorder trajectory's original sequence with minimal deviation. This similarity arises because the original sequence already satisfied the disorder threshold of 0.5, leaving limited room for improvement under the constraint. Consequently, the evolutionary algorithm retained sequences with similar disorders rather than exploring more aggressive mutations that might increase the disorderedness. (Skye)

Attribute	Original Sequence	Top Candidate Sequence	Difference
Sequence	DPNDTVAMKRAR NTLAARKSRQRK MQRFDELEDKIAK LEA	DENDTLAMKRRR NTSAARKSRQRK MQRFDELEDKIVK LEA	-
Point Mutations	-	**Position 1: P -> E** **Position 5: V -> L**	2 substitutions

		Position 10: A -> R **Position 14: L -> S** **Position 35: A -> V**	
Net Charge	4	5	1
Hydrophobicity	-51.3	-62.1	-10.8
Experimental Activity	0.0	-	-
Predicted Activity	-	158, 936.9	+158, 936.9

Table 4B: Comparison between original and top candidate AD sequences. Point mutations led to significant predicted increases in activity, alongside shifts in net charge and hydrophobicity.

Future Directions

To further refine our in silico directed evolution framework and improve candidate quality, we propose the following future extensions:

Extend Evolutionary Depth

Due to computational bottlenecks, we were unable to investigate the effect of increasing the number of generations and keeping a greater number of mutations. A deeper evolutionary trajectory may reveal more diverse sequences and reveal paths towards highly disordered yet functional mutations.

Leverage Larger Protein Language Models

We utilized the 8M parameter ESM-2 model due to our limited computational resources. Thus we believe upgrading from the 8M parameter model to larger variants may be able to enrich sequence embeddings with deeper evolutionary and structural context. These higher-capacity models may better capture more subtle functional signals that are not easily hand-engineered

Optimize For Disorder

Our current optimality constraint for the *in silico* directed evolution algorithm uses a filtering-based constraint that ensures the mean disorder of our mutations is greater than some arbitrary threshold. Instead, we would like to explore the use of a soft optimization objective that aims to maximize disorder scores across generations. (Qamil Mirza, Grayson)

Conclusions

Our work demonstrates the utility of combining protein language model embeddings with traditional biochemical features to improve the prediction of activation domain activity. By leveraging ESM-2-generated representations in combination with biochemical descriptors, we manage to enhance model performance just slightly for “Activity_SCgalactose”, and we found that model performance degraded slightly when used for “Activity_SCglucose”.

Using the best-performing model, we applied *in silico* directed evolution to generate novel AD candidates through random pointwise mutagenesis. These evolved sequences were selected not only for predicted activity but also filtered based on their intrinsic disorder, to ensure consistency with the inherent disordered nature of activation domains. Our analyses revealed how small localized changes in net charge, hydrophobicity, and residue composition can contribute to higher predicted activity.

Together, these findings suggest that while protein embeddings from large-scale models like ESM-2 can encode useful sequence-level information, their predictive utility for activation domain activity seems to remain context-dependent. In our case, ESM-2 embeddings provided only marginal improvements for one reporter system and decreased performance in another. Nonetheless, the integration of learned embeddings with biochemical features remains a promising direction, especially when combining domain-specific descriptors and constraints like intrinsic disorder. More work is needed to explore the role of large protein language modes in guiding the design and screening of functional disordered protein fragments. (Qamil Mirza)

References

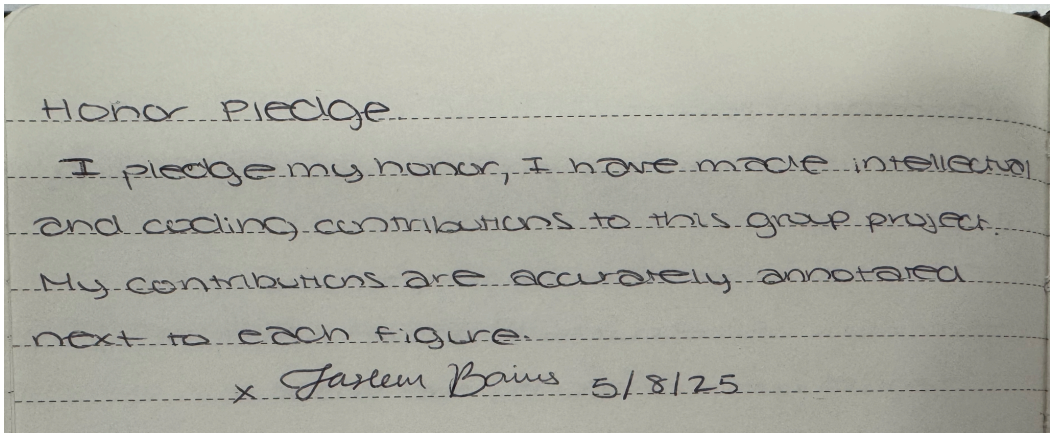
Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & others. (2022). *Language models of protein sequences at the scale of evolution enable accurate structure prediction*. *bioRxiv*. https://doi.org/10.1101/2022.07.20.500902

Emenecker, R. J., Griffith, D., & Holehouse, A. S. (2021). *Metapredict: A fast, accurate, and easy-to-use predictor of consensus disorder and structure*. *Biophysical Journal, 120*(20), 4312–4319. https://doi.org/10.1016/j.bpj.2021.08.025

Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J., & Holehouse, A. S. (2024). *Direct prediction of intrinsically disordered protein conformational properties from sequence*. *Nature Methods, 21*, 465–476. https://doi.org/10.1038/s41592-024-02165-9

Lotthammer, J. M., Hernández-García, J., Griffith, D., Weijers, D., Holehouse, A. S., & Emenecker, R. J. (n.d.). *Metapredict enables accurate disorder prediction across the Tree of Life*. *bioRxiv*. https://doi.org/10.1101/2023.10.03.561139

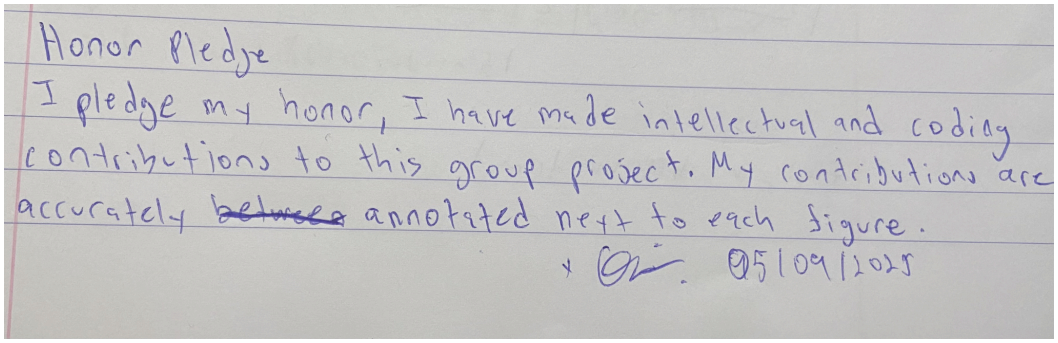
Honor Pledges



Honor Pledge

I pledge my honor, I have made intellectual and coding contributions to this group project. My contributions are accurately annotated next to each figure.

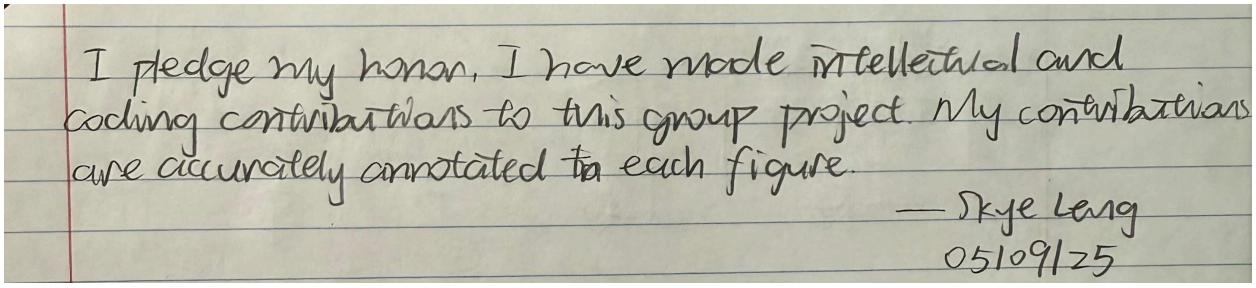
x Jaslem Bains 5/8/25



Honor Pledge

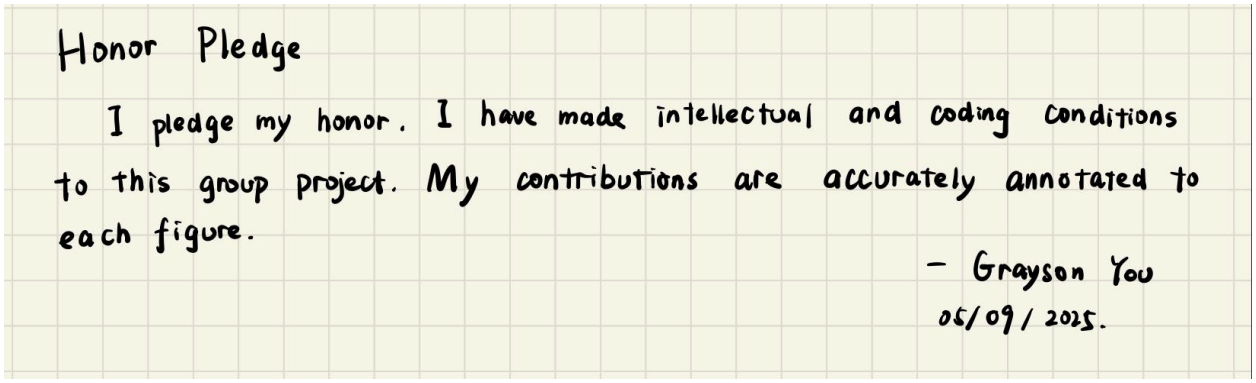
I pledge my honor, I have made intellectual and coding contributions to this group project. My contributions are accurately ~~between~~ annotated next to each figure.

x 05/09/2025



I pledge my honor, I have made intellectual and coding contributions to this group project. My contributions are accurately annotated to each figure.

— Skye Leung
05/09/25



Honor Pledge

I pledge my honor. I have made intellectual and coding conditions to this group project. My contributions are accurately annotated to each figure.

- Grayson You
05/09/2025.