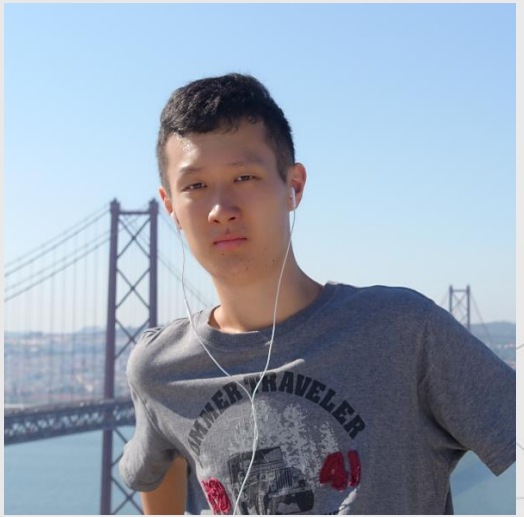
2021 Hackathon

Jiajun He, Zelin Li

0

Who we are

Spaghetti Vector Monster(SVM)



Jiajun He

Zelin Li

Major in Bioinformatics at UCPH

1

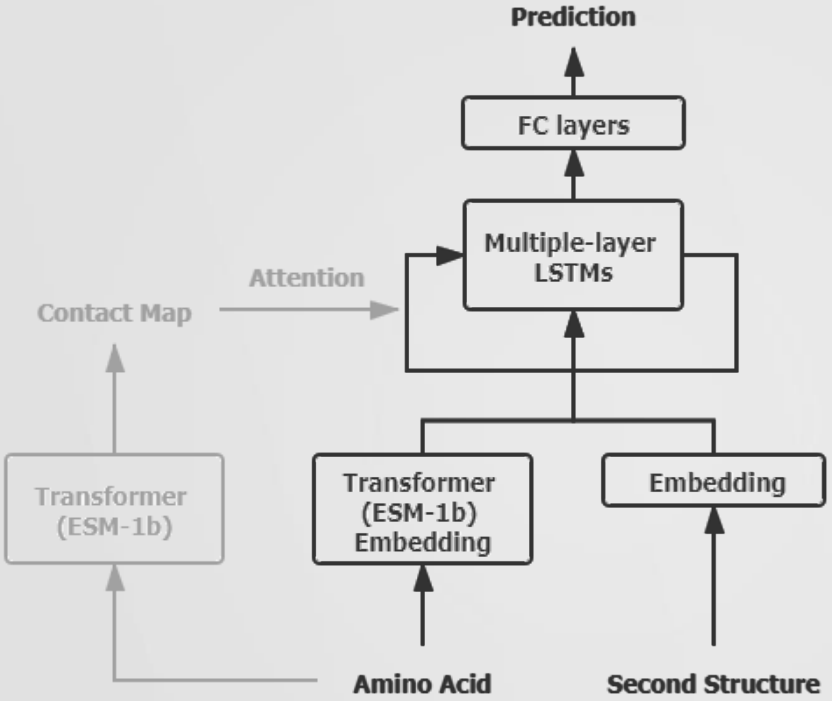
What we did

Models and Performances

Model	Single Mutation	Multiple Mutation
MLP*	0.8451	0.3177
Random Forest*	0.8136	0.3827
SVM*	0.8350	0.4089
Transformer Embedding + RNN	**0.8912**	**0.5940**

* Use one-hot encoding for amino acids.

Best Model

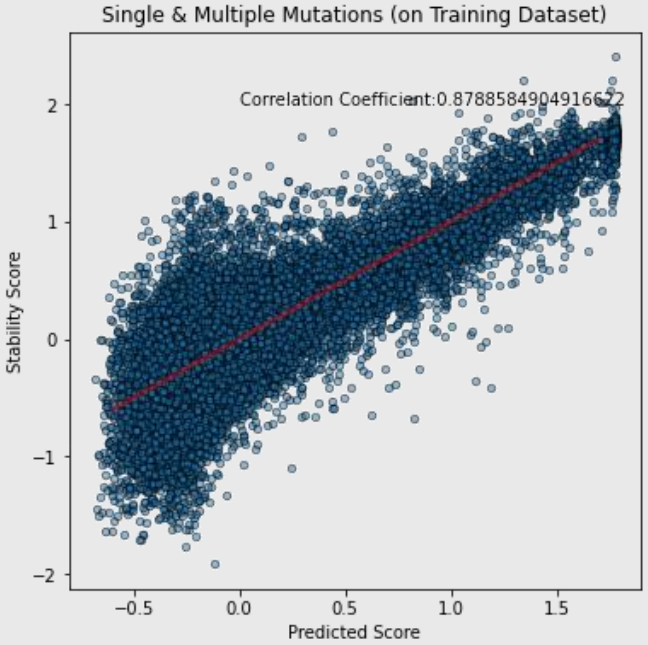


Embedding by Transformer (ESM-1b) → **RNN (LSTMs + FC layers)**

- One Model for Single and Multiple Mutations.

- Early stopping, Drop out;
- Kaiming Initialization for FC layers, Orthogonal Initialization for LSTMs.

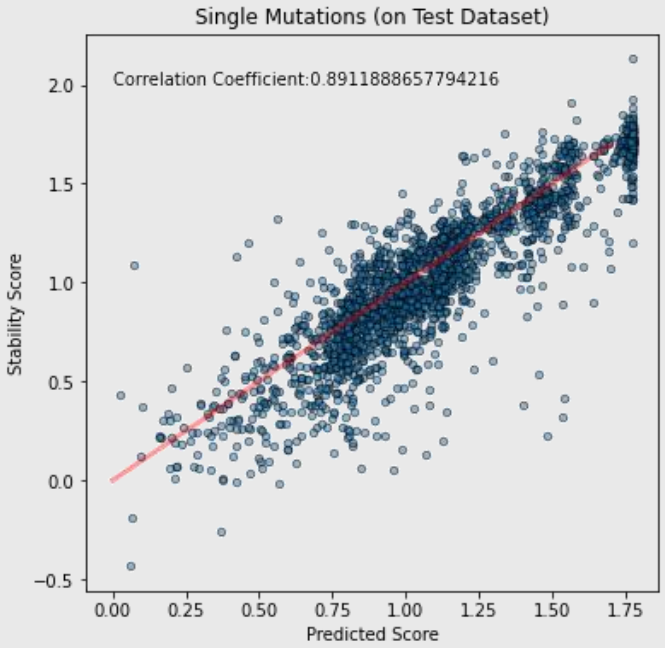
Model Performances: Training & Test Set



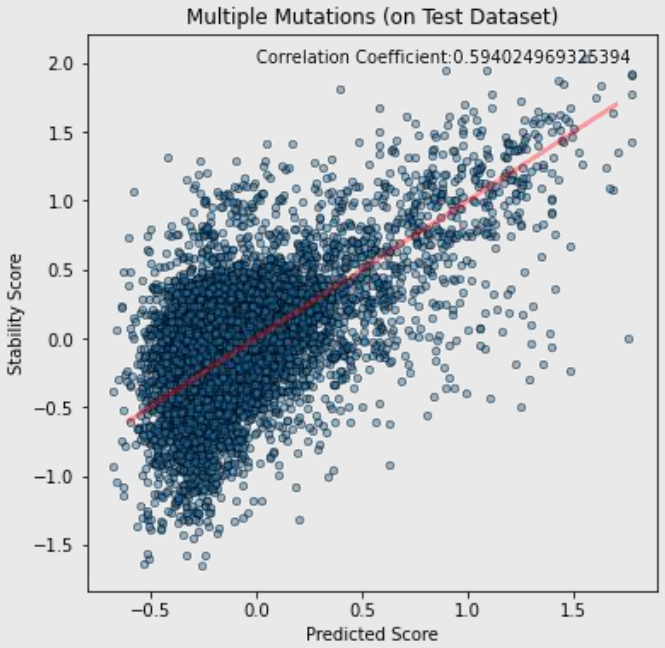
Training: 0.879

Test: 0.839

Model Performances: Single & Multiple Mutations



Single: 0.891



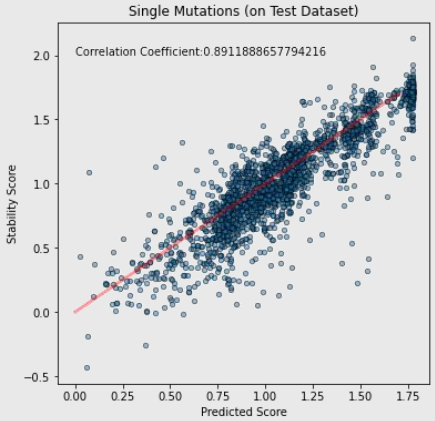
Multiple: 0.594

Conclusion and Discussion

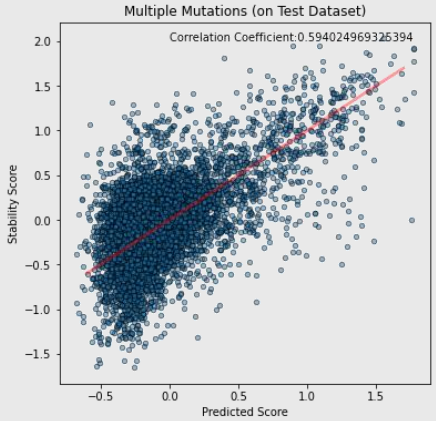
- Better feature engineering yields better results.
- Multiple mutation data is harder to predict than single mutation data, especially those protein with a negative score.

Is the Secondary Structure Necessary?

With SS:

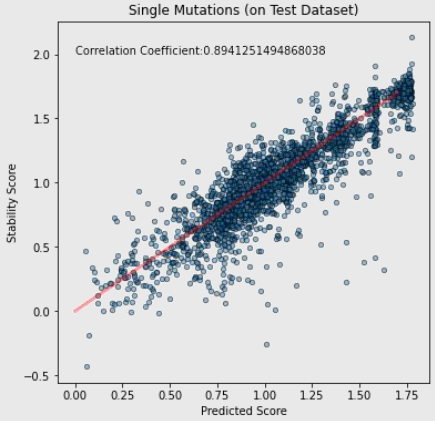


Single: 0.891

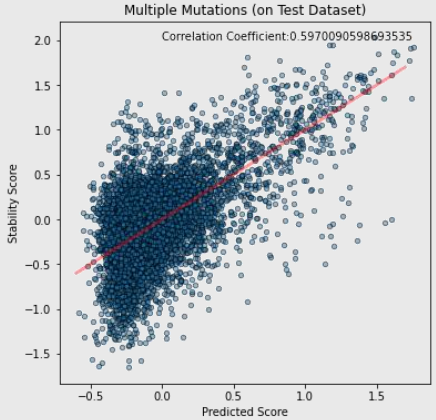


Multiple: 0.594

Without SS:



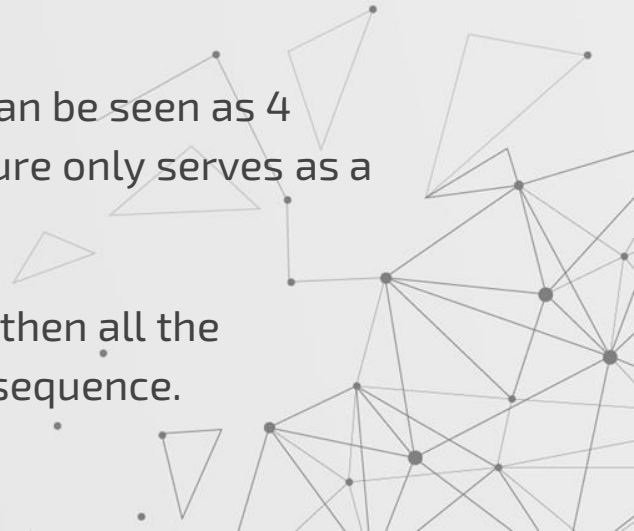
Single: 0.894



Multiple: 0.597

Conclusion and Discussion

- Better feature engineering yields better results.
- Multiple mutation data is harder to predict than single mutation data, especially those protein with a negative score.
- Secondary structure is almost redundant for our task.
    - We have only 4 original sequences. So our task can be seen as 4 individual regressions, and the secondary structure only serves as a category label.
    - If the original energies are thought to be similar, then all the information is stored in the mutated amino acid sequence.



Room for Improvements

- Fine tuning for each dataset respectively.
- Better feature engineering, e.g., considering the chemical properties of amino acids.
- Better architecture, e.g., transfer Learning by Transformer, etc.
- Use more proteins to collect mutation data. (better from different organisms and environments)

Thank you

Our code will be updated on our team page and on our GitHub by tonight ☺

https://biolib.com/SVM2/Spaghetti-Vector-Monster-2/
https://github.com/hejj16
https://github.com/lzlniu