



# Kung Fu Stability

POTENTE **SARA** - [sara.potente@outlook.IT](mailto:sara.potente@outlook.IT)

SODA **EMANUEL** - [emanuelsoda@gmail.com](mailto:emanuelsoda@gmail.com)

ZANGA **ALESSIO** - [alessio.zanga@outlook.it](mailto:alessio.zanga@outlook.it)

#PANDASTIC #PROTEIN #MACHINE\_LEARNING #BIOLOGY

#COPENHAGEN\_BIOINFORMATICS\_HACKATHON\_2021

# Our Team



- ▶ Sara Potente - Hi there! I'm a bioinformatician 🧬 at the University of Study and Politecnico of Milan focusing on genomics and transcriptomics 🦄.
- ▶ Alessio Zanga - Hi folks! I'm a MsC student 📖 from University of Milan - Bicocca, currently looking 👀 for a PhD.
- ▶ Emanuel Soda - Yo guys! I'm the prettiest of the of the group 🙌. Few information about me, I like bioinformatics, art and last but not least baking!

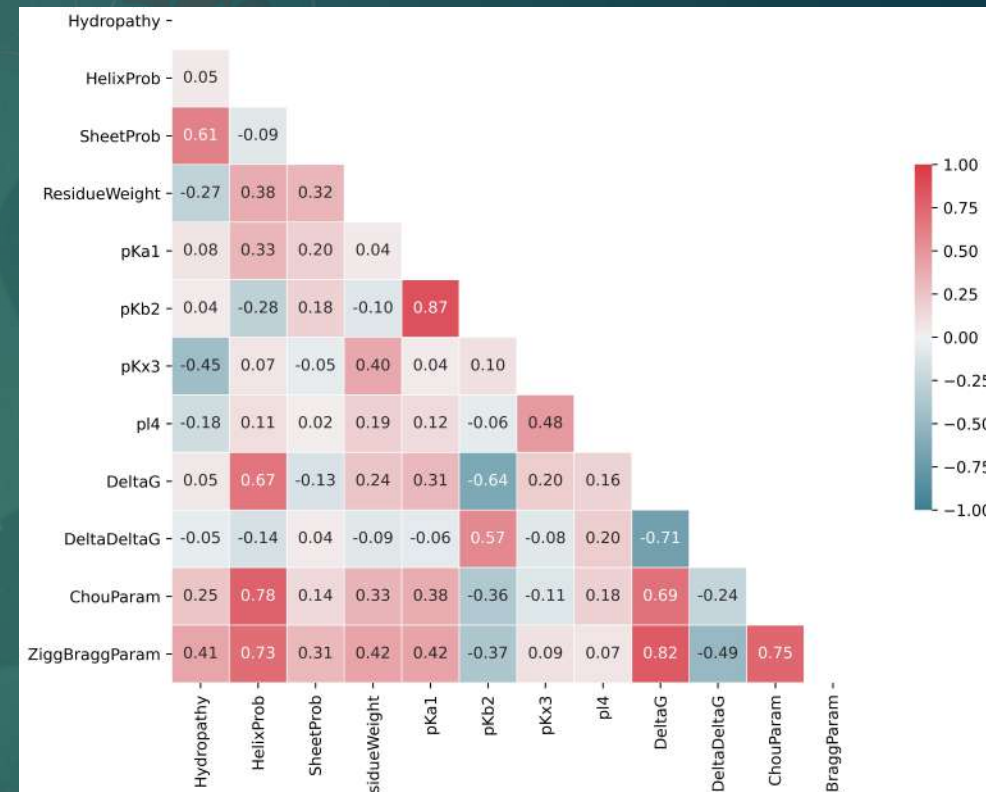


# Features extraction



## ► ChemicoPhysical Aminoacids Characteristics :

- Weight of residue
- Dissociation Constants (pKa1, pKb2, pKx3)
- pI4: pH at the isoelectric point
- deltaG: Contribution of the free energy transfer
- DeltaDeltaG: Change in free energy
- ChouParam: Preference for alpha helix
- ZiggBraggParam: Parameter from X-ray data
- HelixProb: Preference of aa to form alpha-helix
- SheetProb: Preference of aa to form beta-sheet
- Hydropathy: Index of hydrophobicity and hydrophilicity

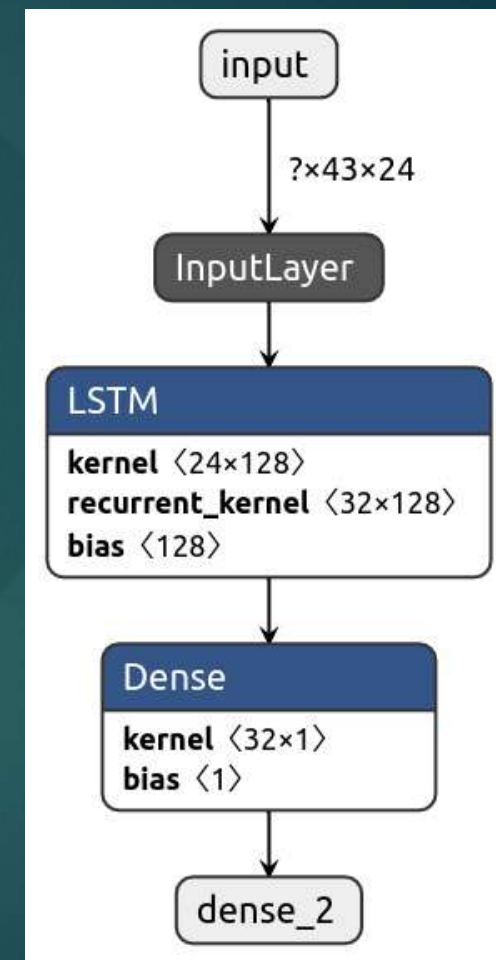




# Model Architecture



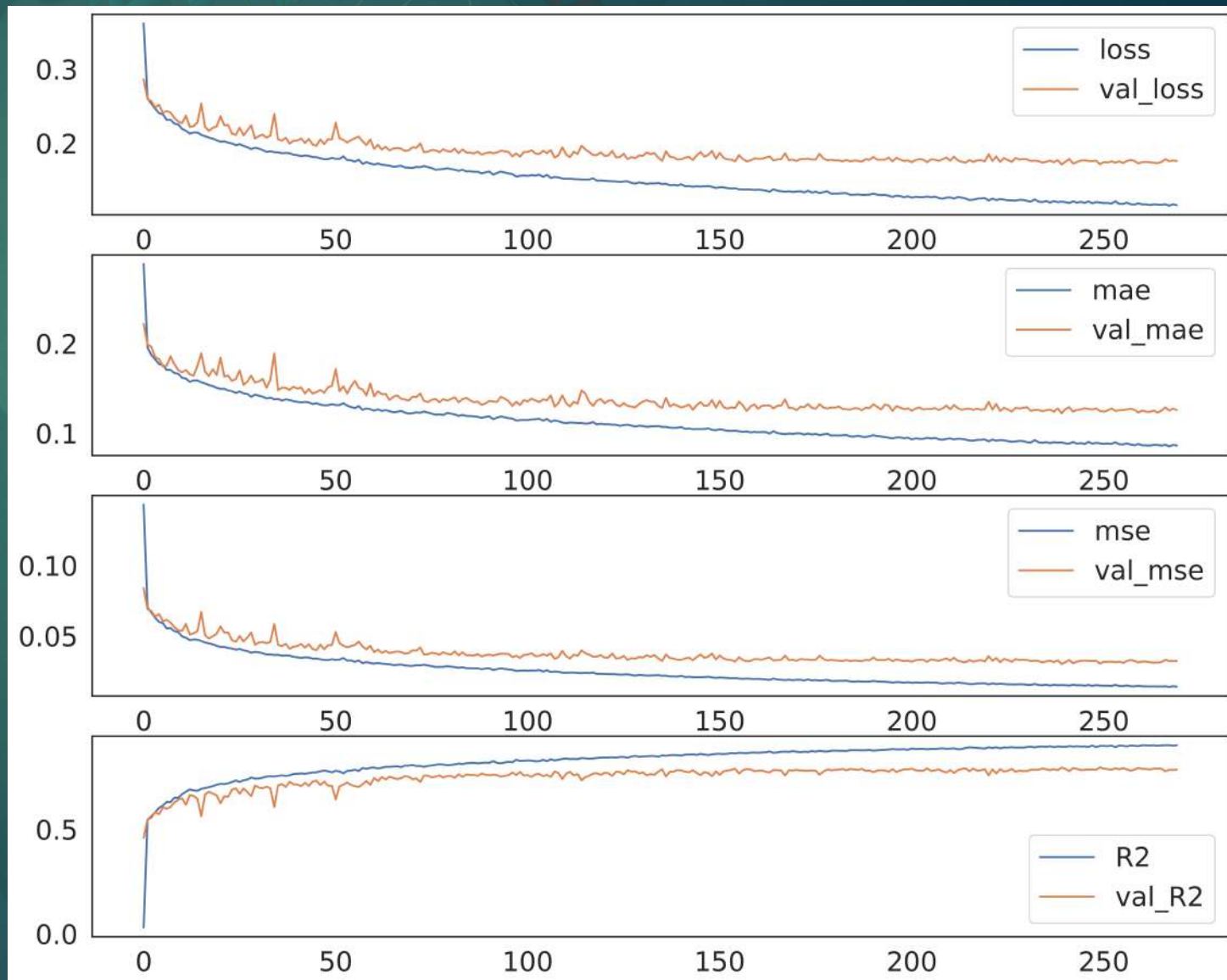
- ▶ We defined one model architecture for both single- and multi-mutant scenarios and trained it separately on the two training sets.
- ▶ The final network architecture is composed by a first LSTM layer, which focus on the sequence properties as a whole, and a out dense layer, that is responsible for the regression step.
- ▶ We tried bigger and more complex models, but in the end there was not so much gain compared to the increased training time.



# Learning Curve



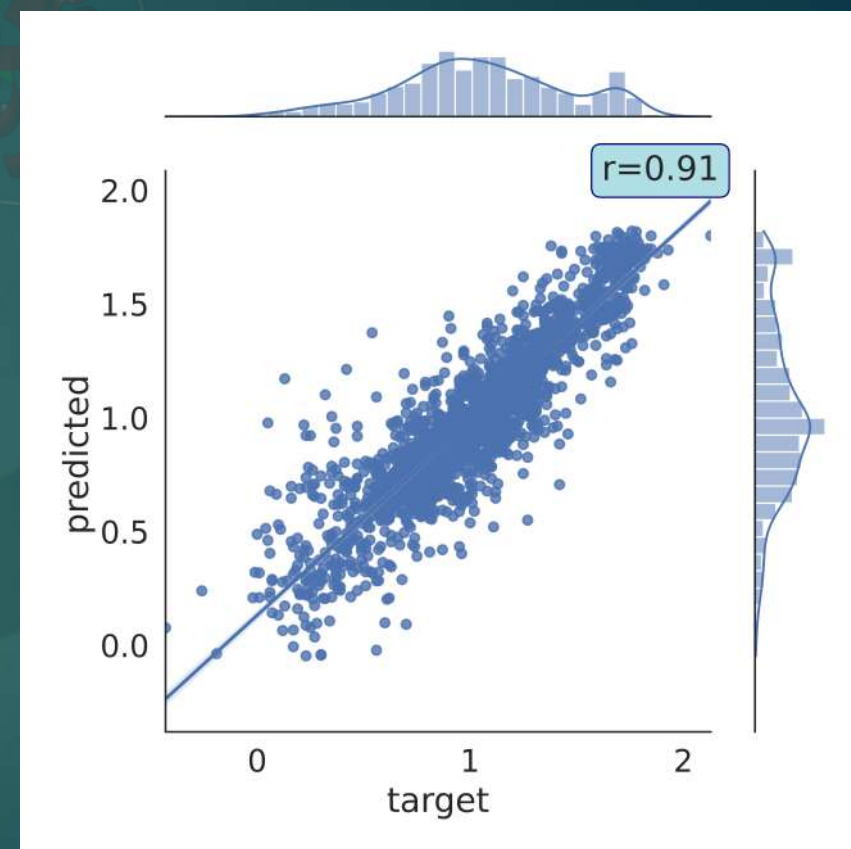
- ▶ The model was trained using an early stopping approach on the loss of validation split.
- ▶ As reported in figures, there is no-to-little overfitting.



# Performance Single Mutant



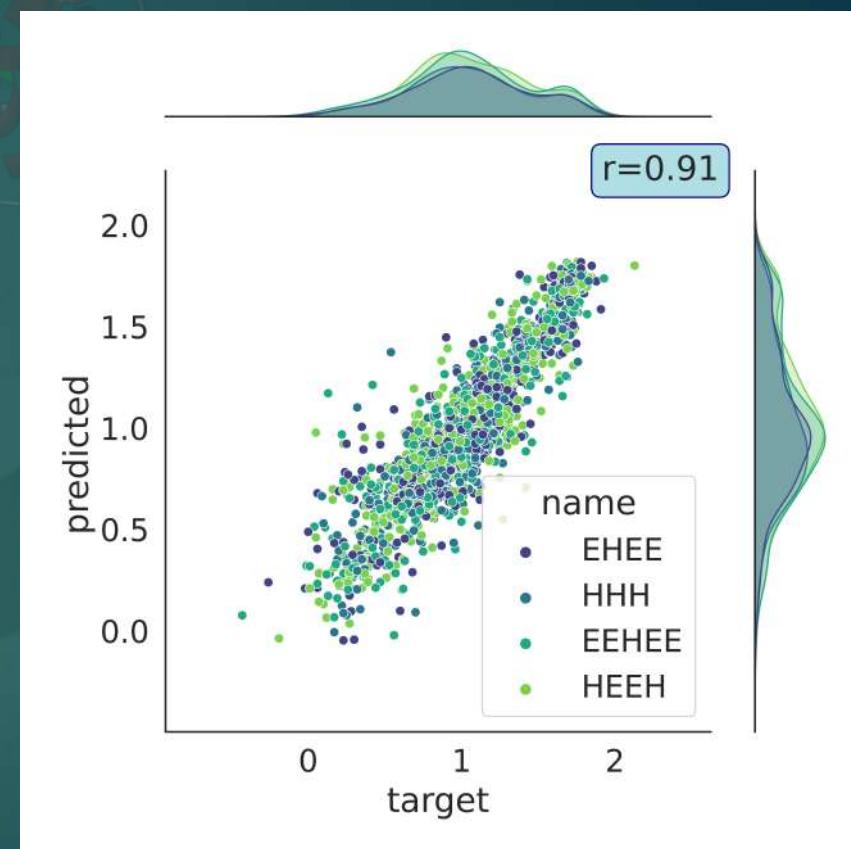
- ▶ On the X-axis the stability measured experimentally.
- ▶ On the Y-axis the predicted one.
- ▶ The model shows Pearson correlation of 0.91.



# Performance Single Mutant



- ▶ All the class shows the same bimodal like distribution of the correlation coefficient.
- ▶ Class EHEE seems to be the most clustered

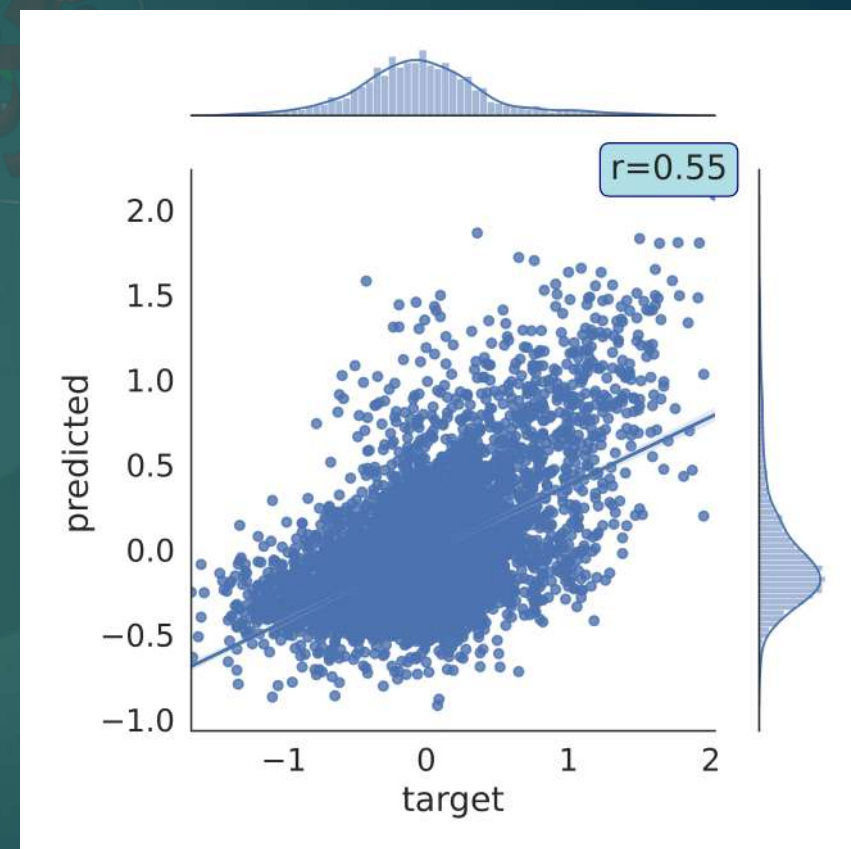




# Performance Multi Mutant



- ▶ As for the literature the prediction of the multi mutant shows a lower Pearson correlation of 0.55.
- ▶ The number of outliers seems to be lower compared to the literature.

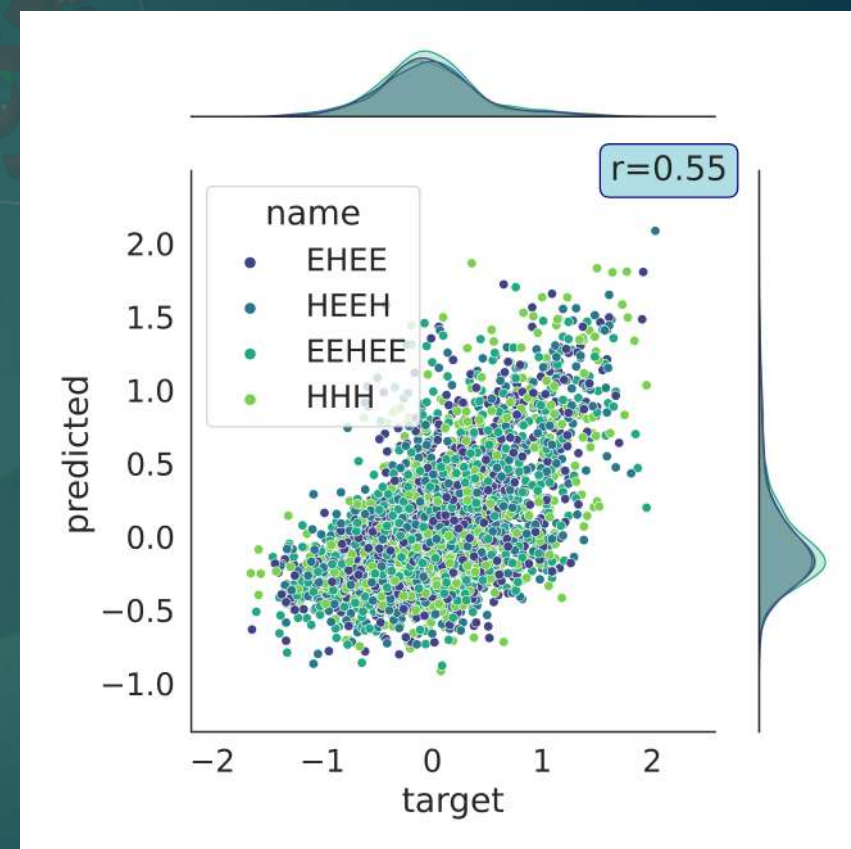




# Performance Multi Mutant



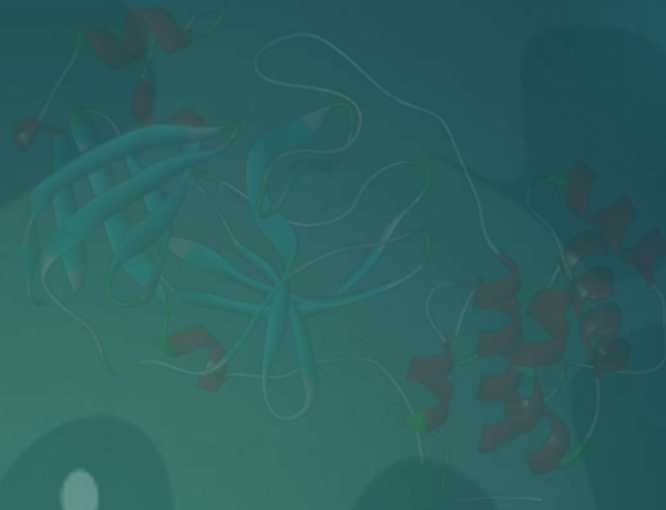
- ▶ All the class shows the same unimodal distribution of the correlation coefficient.
- ▶ No particular class shows a clusterization.



# Bibliography



- ▶ <https://www.sigmaaldrich.com/life-science/metabolomics/learning-center/amino-acid-reference-chart.html>
- ▶ [https://en.wikipedia.org/wiki/Amino\\_acid](https://en.wikipedia.org/wiki/Amino_acid)
- ▶ Lehninger Principles of Biochemistry
- ▶ "Generation and evaluation of dimension-reduced..." (Meiler, 2001)
- ▶ "A Thermodynamic Scale for the Helix-Forming Tendencies of ..." (O'Neil, Degrado - 1990)
- ▶ "Everyone Is a Protagonist: Residue Conformational Preferences..." (Ligabue-Braun et al., 2018)



# Thanks all for the attention

FEEL FREE TO REACH OUT IF YOU HAVE ANY QUESTION