

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

Alexey Strokach

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 6. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



Permission request regarding CAGI 6 recordings and hybrid presentations

I understand that my presentation as associated with CAGI 6 will be recorded and broadcasted via the internet through CAGI's designated technology provider.

As the copyright holder to this presentation entitled Alexey Strokach, I hereby agree that the copyright for the online broadcast and recording of my presentation be released to CAGI and my recording be uploaded on to the CAGI website.

CAGI6 Submissions Team "ELASPIC"

Alexey Strokach, Dr. Philip M. Kim

<https://ostrokach-slides.gitlab.io/2022-cagi6/>

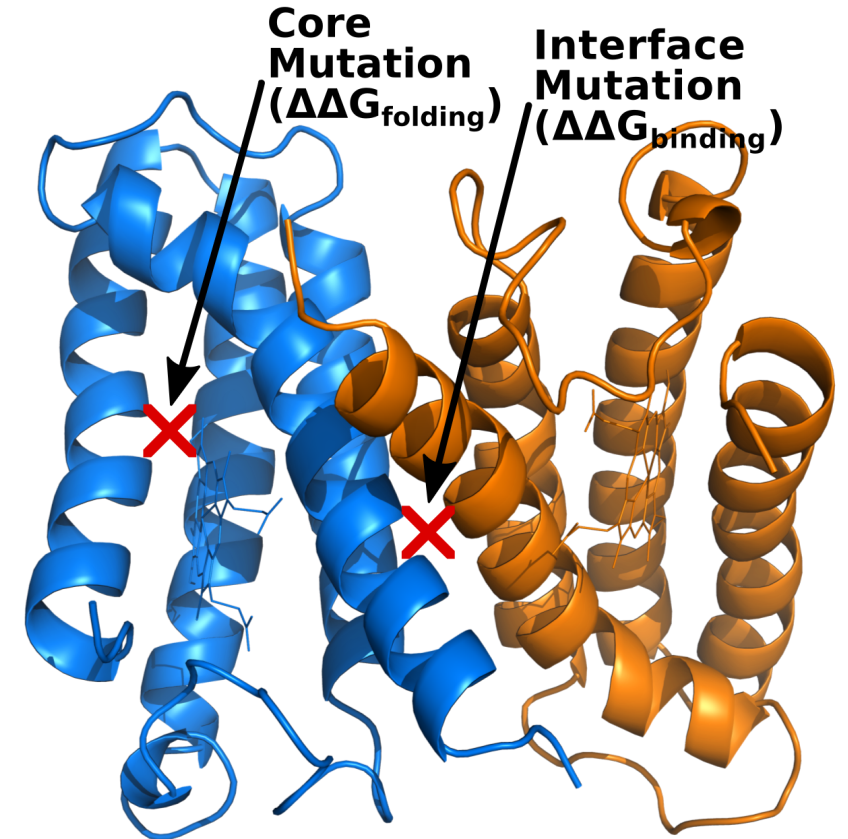
Predictions were made using different methods

- **ELASPIC2 (EL2)** — GBT model predicting the effect of mutations on protein stability and protein binding affinity using embeddings from ProteinSolver and ProtBERT and a *ranking objective function*.
- **ELASPIC3 (EL3)** — GBT model predicting deleterious and benign mutations using embeddings from ProteinSolver, ProtBERT, and AlphaFold.
- **ProteinSolver** — Graph neural network trained to reconstruct masked amino acids given the distagram describing the topology of the protein.
- **ProtBERT** — Deep neural network trained to reconstruct masked amino acids in millions of protein sequences.
- **Rosetta** — Tool for macromolecular modeling which uses on a semi-empirical energy function and MCMC sampling.

ELASPIC

Overview

- Machine learning model to predict the effect of mutations on protein folding and protein-protein interaction.
- Developed by Berliner et al. in 2015 and inherited by me.
- Integrates evolutionary and structural information to improve the accuracy of results.
- One of the best-performing methods in the CAGI 5 Frataxin challenge.



ELASPIC2 (EL2)

Overview

- Predicting the effect of mutations on protein stability and protein binding affinity using *pretrained neural networks* and a *ranking objective function*.
 - Use embeddings output by ProteinSolver and ProtBERT instead of hand-crafted features.
 - Use a *ranking objective function* to leverage data obtained using different modalities, including *in vitro* $\Delta\Delta G$ measurements and deep mutation scanning experiments.
- Strokach et al. (2021). *ELASPIC2 (EL2): Combining Contextualized Language Models and Graph Neural Networks to Predict Effects of Mutations*.
<https://doi.org/10.1016/j.jmb.2021.166810>

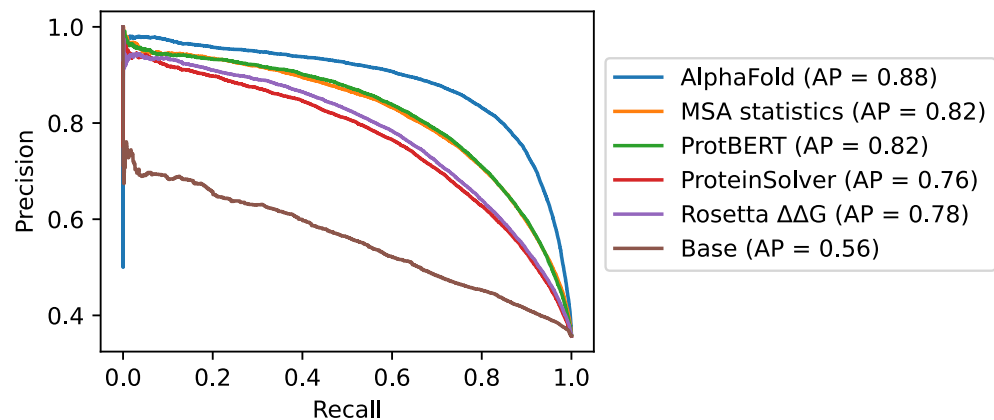
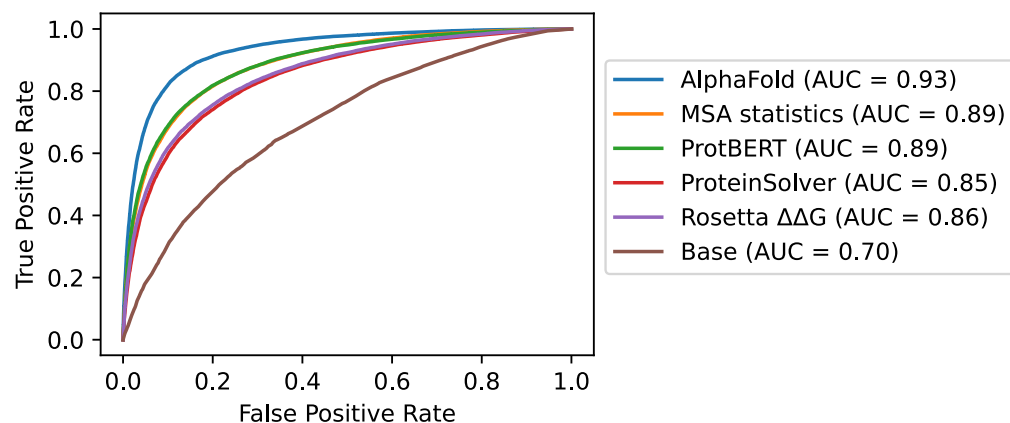
ELASPIC3 (EL3)

Overview

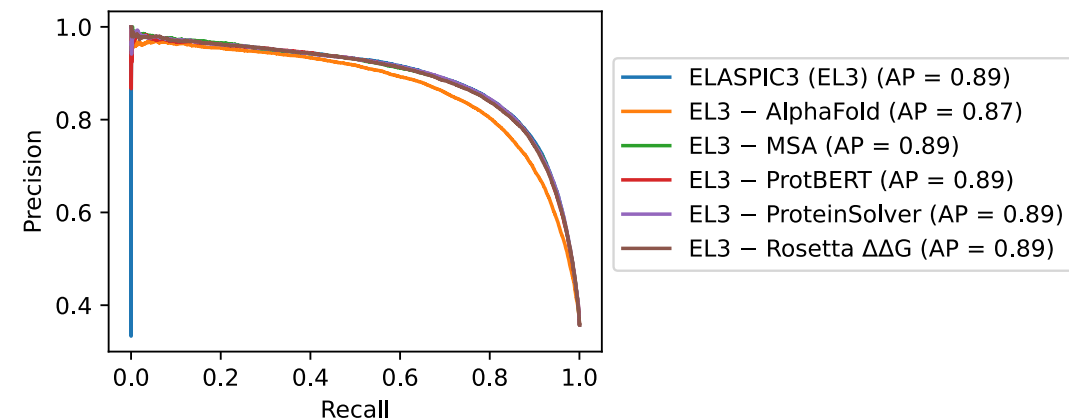
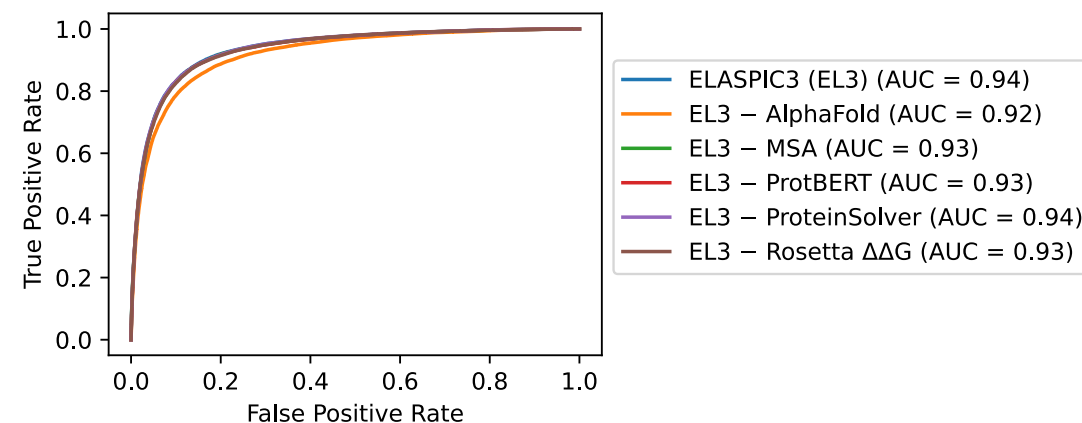
- Predicting the effect of mutations on phenotype, i.e. whether a given mutation is deleterious or benign, using *pretrained neural networks*.
 - Use embeddings output by ProteinSolver, ProtBERT, *and AlphaFold*, instead of hand-crafted features.
 - Use the humsavar and CAGI6 Sherlock training datasets and AlphaFold models of the human proteome.
 - Achieved third place on the CAGI6 Sherlock challenge leaderboard, with an AUC of 0.946 and a recall at 0.8 precision of 0.866.
- Code for training the model: <https://github.com/elaspic/cagi6-submission>.

ELASPIC3 (EL3)

Single embedding performance



All-but-one embedding performance

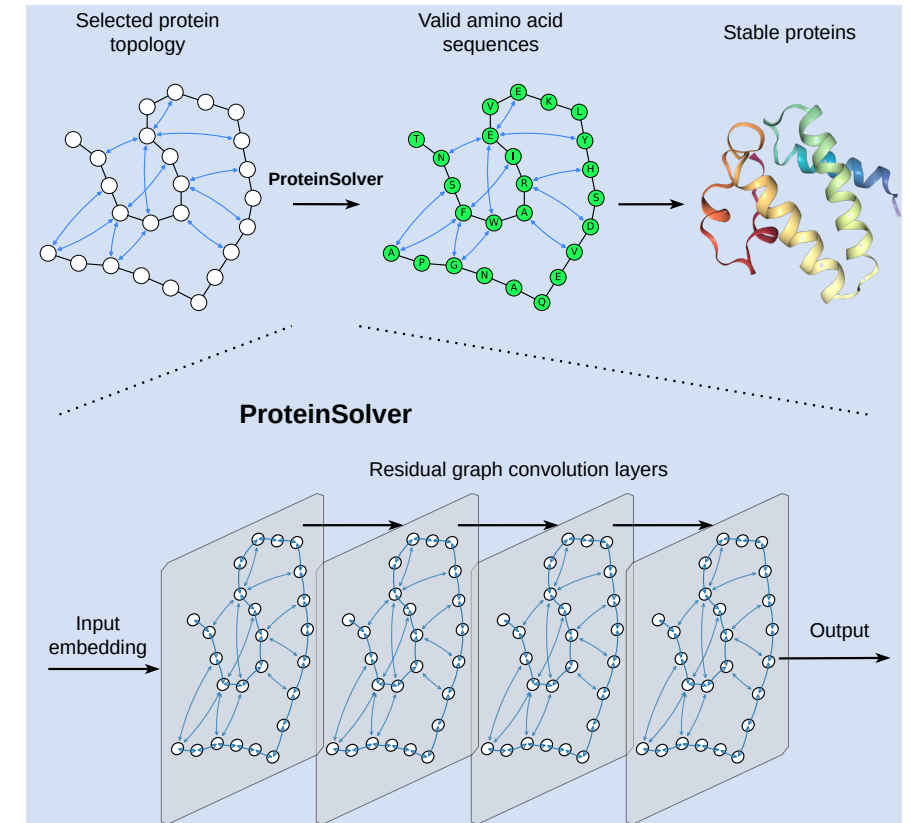


ProteinSolver

Overview

- A graph neural network trained to reconstruct masked amino acids given the distagram describing the topology of the protein.
- The model provides a probability and an embedding for a given residue.
- Strokach et al. (2020). *Fast and Flexible Protein Design Using Deep Graph Neural Networks*.

<https://doi.org/10.1016/j.cels.2020.08.016>



ProtBERT

Overview

- A BERT language model trained on millions of protein sequences using a masked language modeling objective.
- The model provides a probability and an embedding for a given residue in the protein sequence.
- Elnaggar et al. (2020). *ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing*.
<https://doi.org/10.1101/2020.07.12.199554>

Rosetta $\Delta\Delta G$

Overview

- Rosetta uses a semi-empirical energy function to evaluate the stability of the wild type and the mutant proteins.
- We use the `cartesian_ddg` protocol to obtain energy terms that are averaged over different protein conformations.
- Park et al. (2016). *Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules*.
<https://doi.org/10.1021/acs.jctc.6b00819>

Conclusions

Future directions

- Incorporate AlphaFold features into the ELASPIC2 model.
- Check if fine-tuning the neural networks using a ranking objective function results in better accuracy than using LightGBM.
- Compare the performance of our submissions across all CAGI6 challenges and check for patterns.
- Take inspiration and learn from all the other teams participating in CAGI6!

Acknowledgements

Supervisor

- Philip M. Kim

Committee Members

- Hannes Röst
- Quaid Morris

Members of kimlab

- Albert Perez-Riba
- Carles Corbi-Verge
- David Becerra
- Fan Shen
- Han Wen
- Mac Mok
- Michael Garton
- Osama Abdin
- Pedro Alberto Valiente Flores
- Qiao Fang
- Satra Nim
- Vincent Lee
- Yuju Lee

