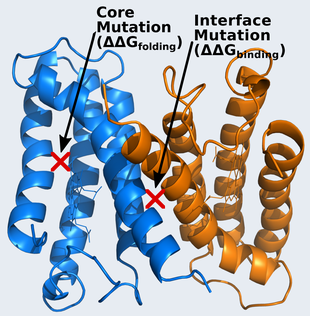
Predicting the Effect of Mutations on a Genome-Wide Scale

Alexey Strokach[1,2], Philip M. Kim[1,2,3]

[1]Department of Computer Science, University of Toronto
[2]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto
[3]Department of Molecular Genetics, University of Toronto

The Big Question



Advances in DNA sequencing technology have led to an enormous growth in the amount of available genomic data. Interpreting this data to produce meaningful and actionable results remains a challenge. Tools currently in use for annotating discovered variants rely on a sequence conservation score and provide little mechanistic insight to explain why a particular variant may be deleterious. Tools that exist for predicting the effect of mutations on the structure and function of a protein are laborious to use and require a crystal structure of the protein, severely limiting their coverage. ELASPIC, a pipeline recently developed in our lab, uses homology models instead of crystal structures to accurately predict the effect of a mutation on the stability of a protein and the affinity of one protein for another. In this work we extend ELASPIC to analyze the effect of mutations on a genome-wide scale.

Methods

1. Update and extend the set of sequential and structural features used by ELASPIC [1].
2. Select the best set of hyperparameters by evaluating the performance of $\Delta\Delta G$ predictors on the training and validation datasets.
3. Perform feature elimination to find the best set of features.
4. Train the final core and interface predictors and validate them on an independent test set.

Training the Core Predictor

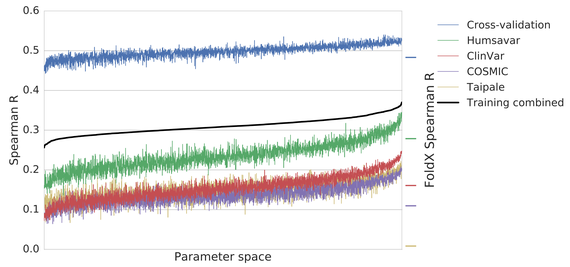


Figure 1: Core predictor hyperparameter optimization. The combined score (black line) corresponds to a weighted average of the validation scores. We selected hyperparameters producing a predictor with the highest combined score.

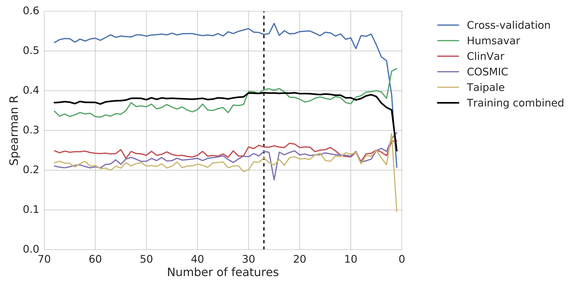


Figure 2: Performance of the core predictor at each step of feature elimination. Predictor with the highest combined score is indicated by the vertical dashed line.

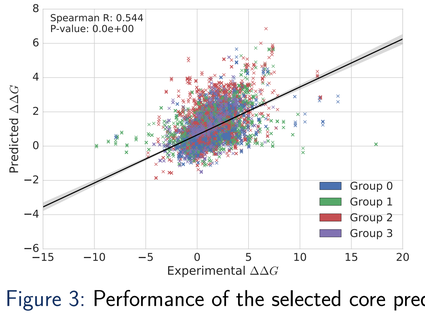


Figure 3: Performance of the selected core predictor on the training dataset, evaluated using four-fold cross-validation. Colours indicate different cross-validation bins.

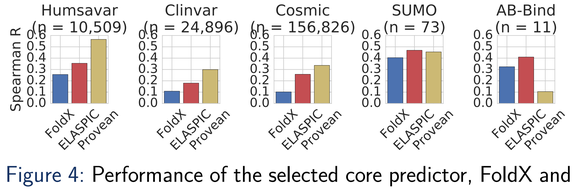


Figure 4: Performance of the selected core predictor, FoldX and Provean on the test datasets. There is no overlap in mutations (or proteins for Humsavar, ClinVar and COSMIC) between the test datasets, and the training and validation datasets.

Training the Interface Predictor

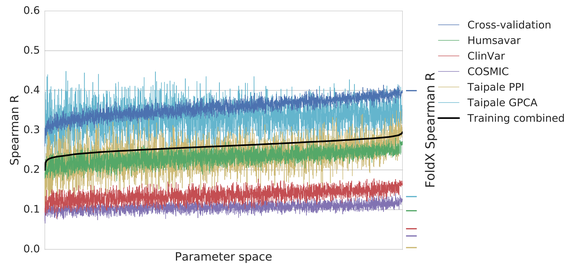


Figure 5: Interface predictor hyperparameter optimization. The combined score (black line) corresponds to a weighted average of the validation scores. We selected hyperparameters producing a predictor with the highest combined score.

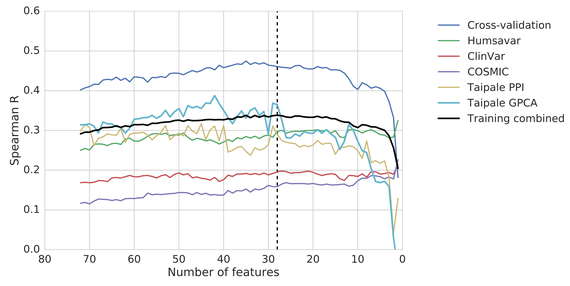


Figure 6: Performance of the interface predictor at each step of feature elimination. Predictor with the highest combined score is indicated by the vertical dashed line.

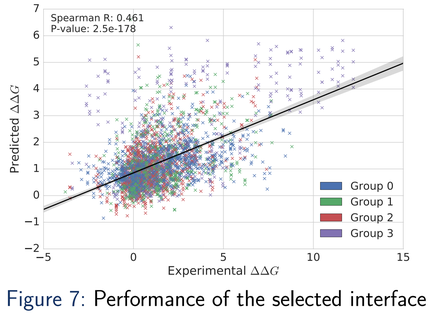


Figure 7: Performance of the selected interface predictor on the training dataset, evaluated using four-fold cross-validation. Colours indicate different cross-validation bins.

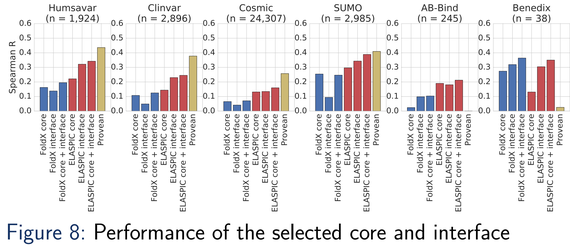


Figure 8: Performance of the selected core and interface predictors, FoldX and Provean on the interface test datasets.

ELASPIC Webserver

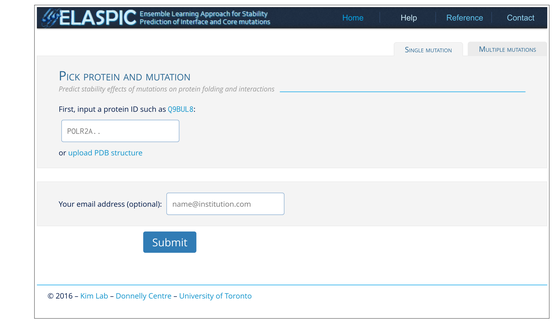


Figure 9: Screenshot from the ELASPIC webserver: `http://elaspic.kimlab.org`. The webserver was created by Daniel Witvliet et al. [2].

ELASPIC CLI

ELASPIC can be installed on most Linux-based operating systems, and offers an intuitive command-line interface (CLI):

```
# Install ELASPIC
$ conda config channels --append kimlab
$ conda install elaspic
# Run ELASPIC
$ elaspic --help
usage: elaspic [-h] [-v] {run,database,train} ...

optional arguments:
  -h, --help            show this help message and exit
  -v, --verbose         Specify verbosity level

command:
  {run,database,train}
    run                 Run ELASPIC
    database            Perform database maintenance tasks
    train               Train the ELASPIC classifiers
```

The ELASPIC CLI can be used to evaluate the effect of mutations on individual structures. It can also automatically construct homology models of domains and domain-domain interactions for any protein in the SwissProt database. Homology models of all *human* proteins and protein-protein interactions have been precalculated and are available from the elaspic website: `http://elaspic.kimlab.org/static/download`.

Conclusions

- ELASPIC accurately predicts the structural effect of mutations for the majority of proteins in a genome.
- Structural information does *not* substantially improve our ability to predict whether a mutation is deleterious.

References

[1] N. Berliner, J. Teyra, R. Colak, S. G. Lopez, P. M. Kim; PLOS ONE (2014) **9** (9): e107353.

[2] D. K. Witvliet, A. Strokach, A. F. Giraldo-Forero, J. Teyra, R. Colak, P. M. Kim; Bioinformatics (2016) **32** (10): 1589-1591.