Predicting the thermodynamic effect of mutations on a genome–wide scale

Alexey Strokach

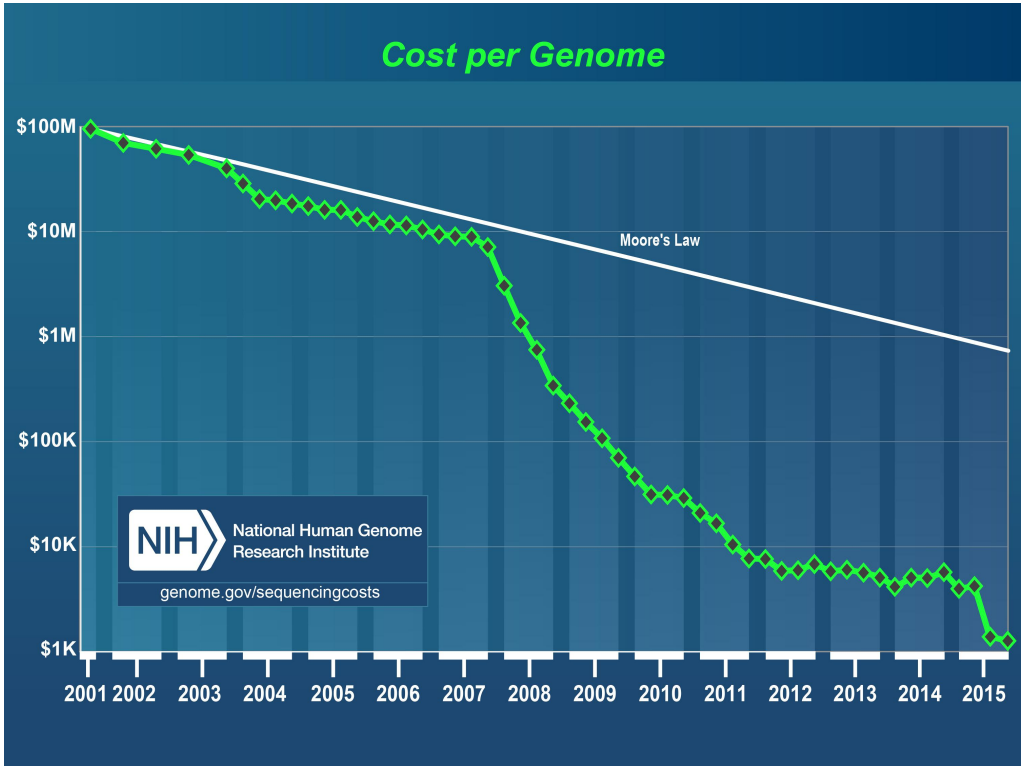
Overview

- Introduction
  - Rise in genome sequencing
  - Difficulty in interpreting discovered variants
  - ELASPIC (Berliner *et al.* 2014)
- Implementation
  - Domain definitions
  - Database backend
  - ELASPIC pipeline and CLI
  - Precalculating data
  - ELASPIC REST API
- Results
  - Grid-search (to find a good set of parameters)
  - Feature elimination
- Discussion
  - What have we learned?
  - Future directions
- Future directions
  - Better features / model
  - More training data

Introduction

Introduction - Rise in genome sequening

- Cost of genome sequencing has fallen dramatically.
- Interpreting genomic variants to yield meaningful and actionable results remains a challenge.



https://www.genome.gov/27541954/dna-sequencing-costs-data/

ARTICLE OPEN

Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine

Introduction - Variant calling / scoring

Variant calling

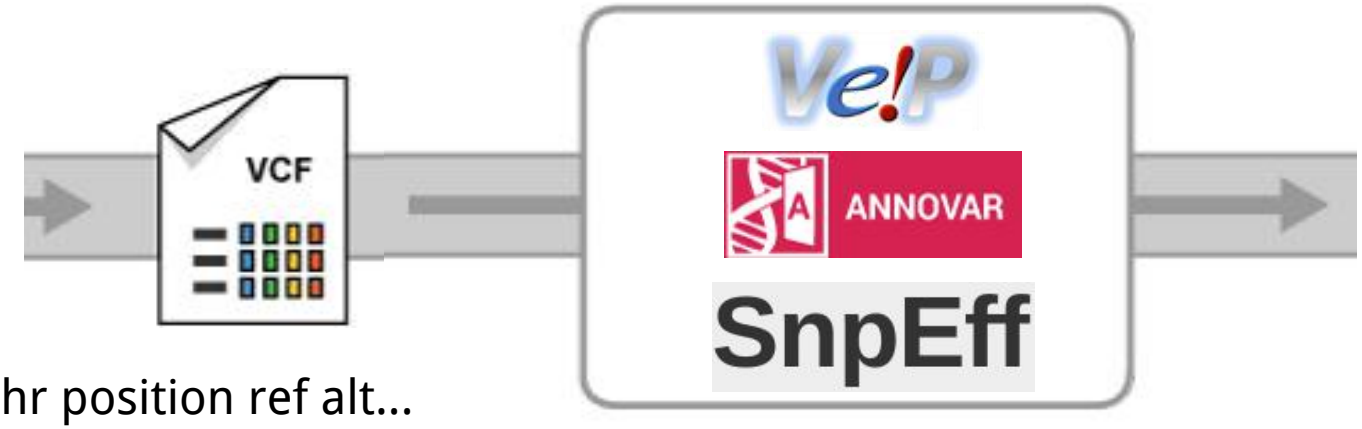
https://software.broadinstitute.org/gatk/

Introduction - Variant calling / scoring

Variant calling

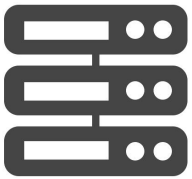
https://software.broadinstitute.org/gatk/

Variant scoring



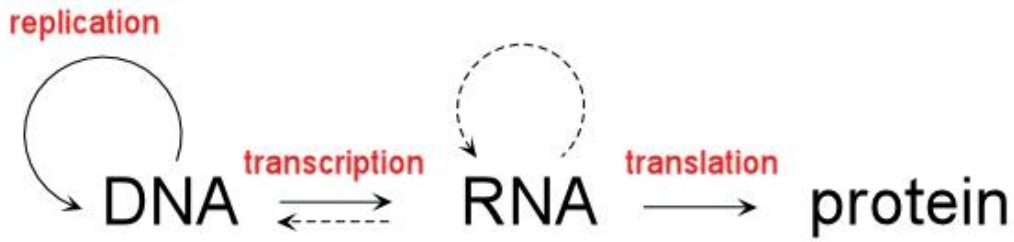
chr position ref alt...

dbNSFP

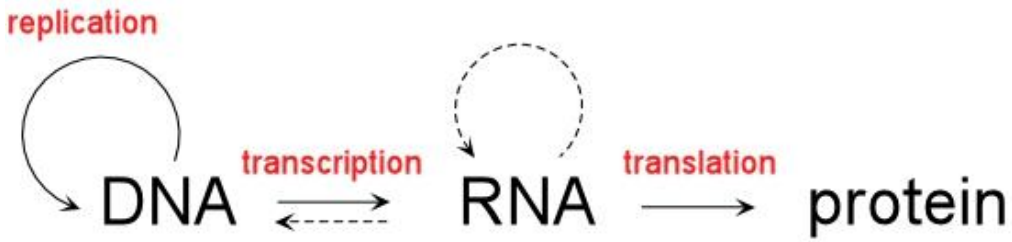


SIFT
Provean
GERP++
etc...

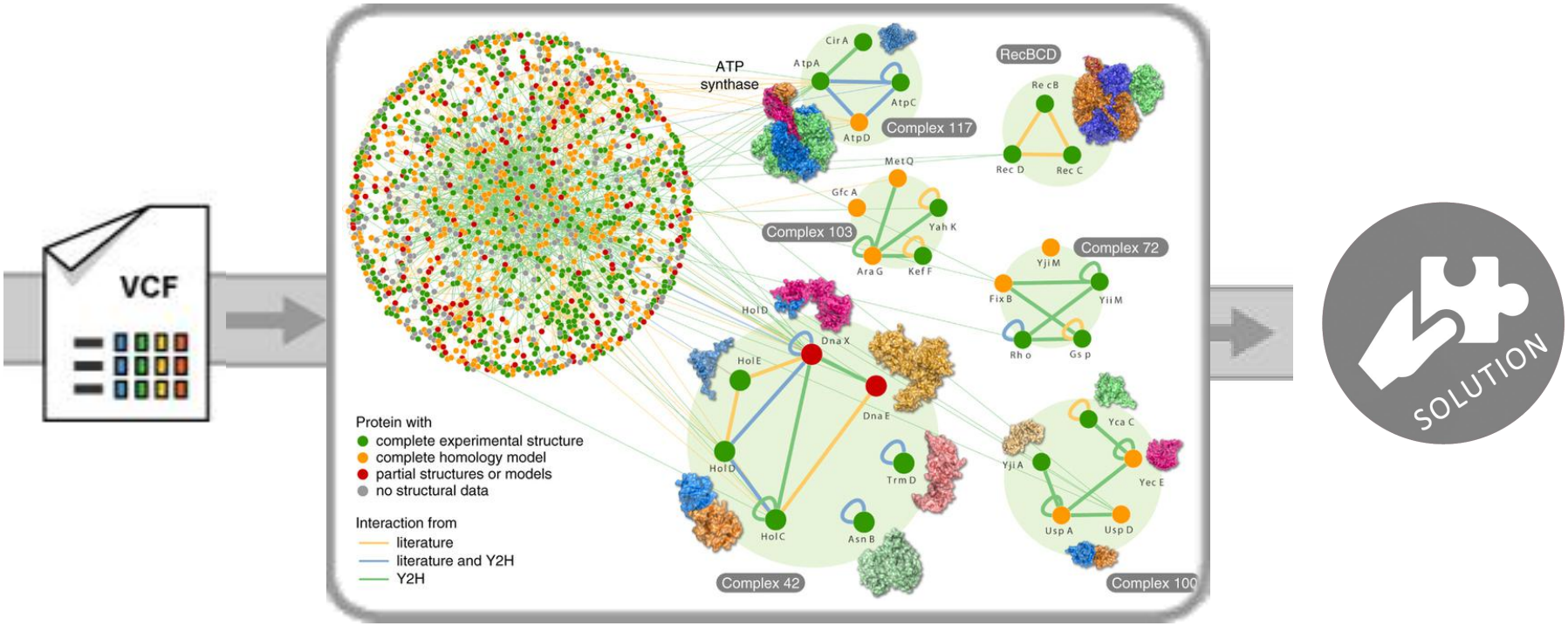
Introduction - Variant *analysis*



Introduction - Variant *analysis*



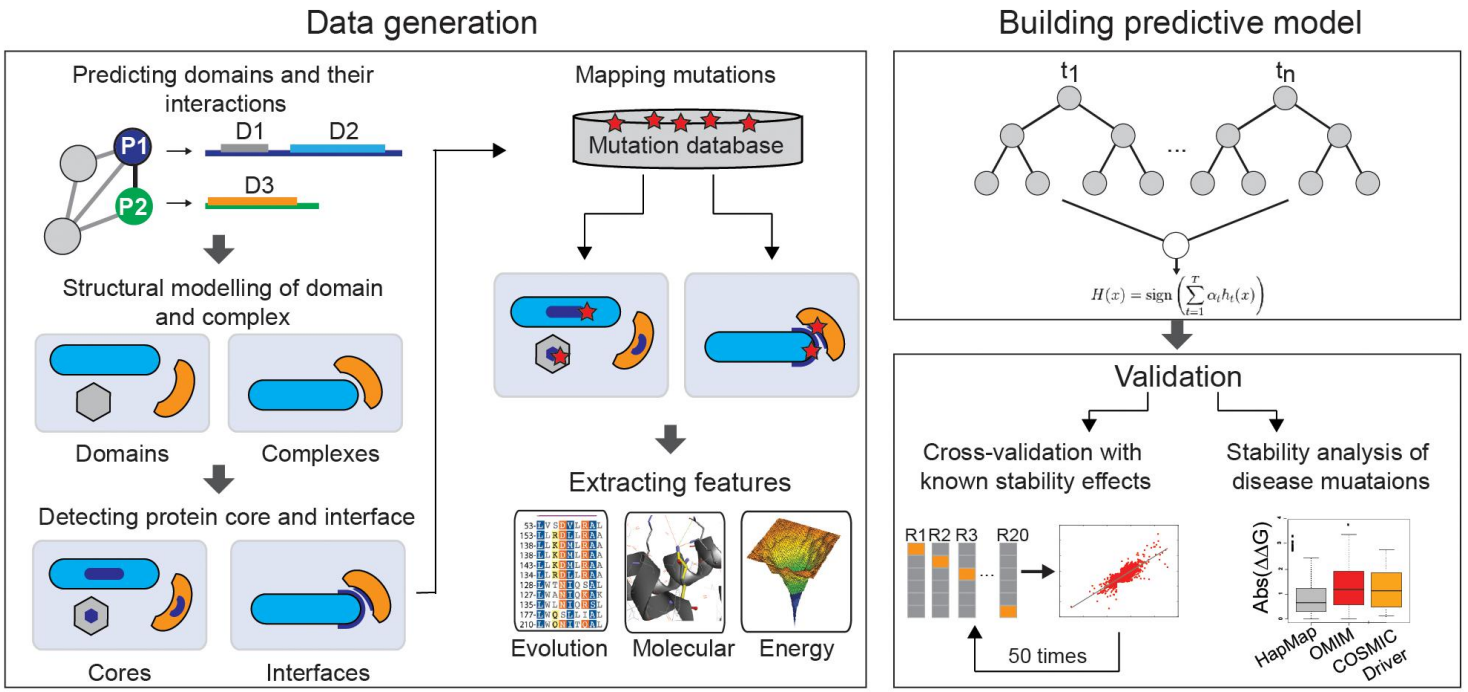
Variant interpretation



Rajagopala S. *et al*., Nature Biotech., 2014.

Introduction - Variant *analysis*

ELASPIC (Berliner *et al. 2014)*



Introduction - Goals

Goals

- Extend ELASPIC to work on a genome-wide scale
  - Profs for domain definitions
  - Use programs that work for *most* sequences / structures
- Implement ELASPIC as a back-end to a webserver
- Demonstrate the usefullness of ELASPIC in providing meaningful and actionable information about mutations.

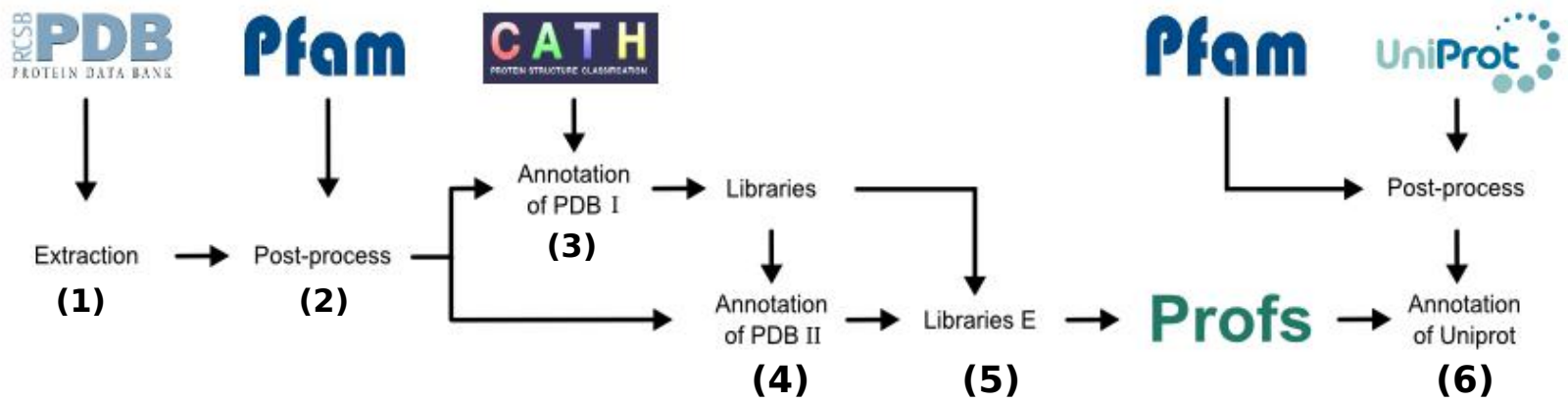
Implementation

Implementation

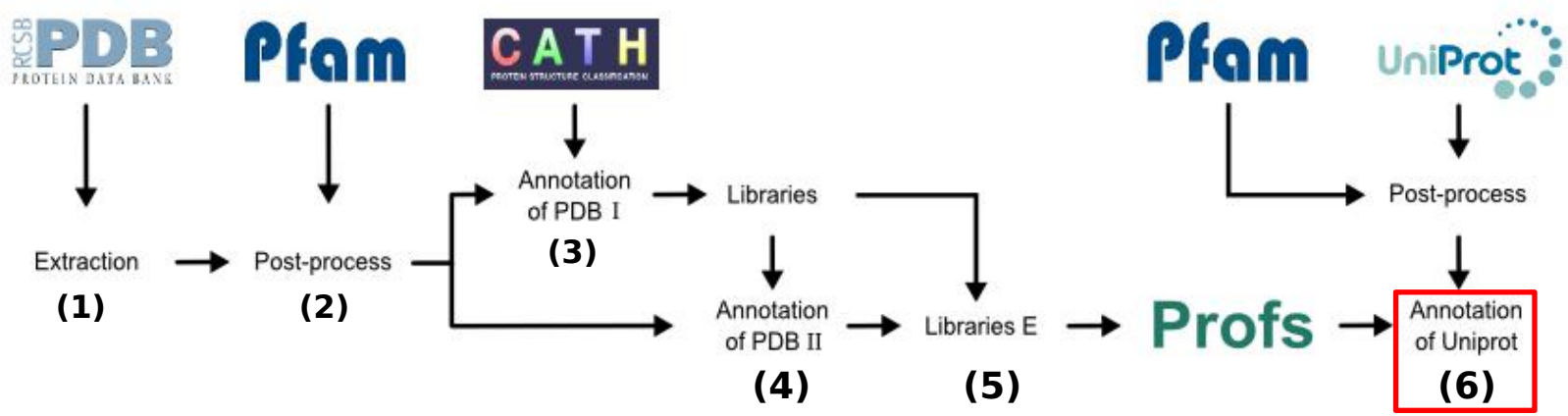
- Domain definitions
  - Profs (*Andres Felipe Giraldo-Forero*)
- Domain-domain interactions
- Database backend
- ELASPIC pipeline + CLI interface
- Precalculate Provean, homology models, and many mutations
- Webservice for submitting and monitoring jobs

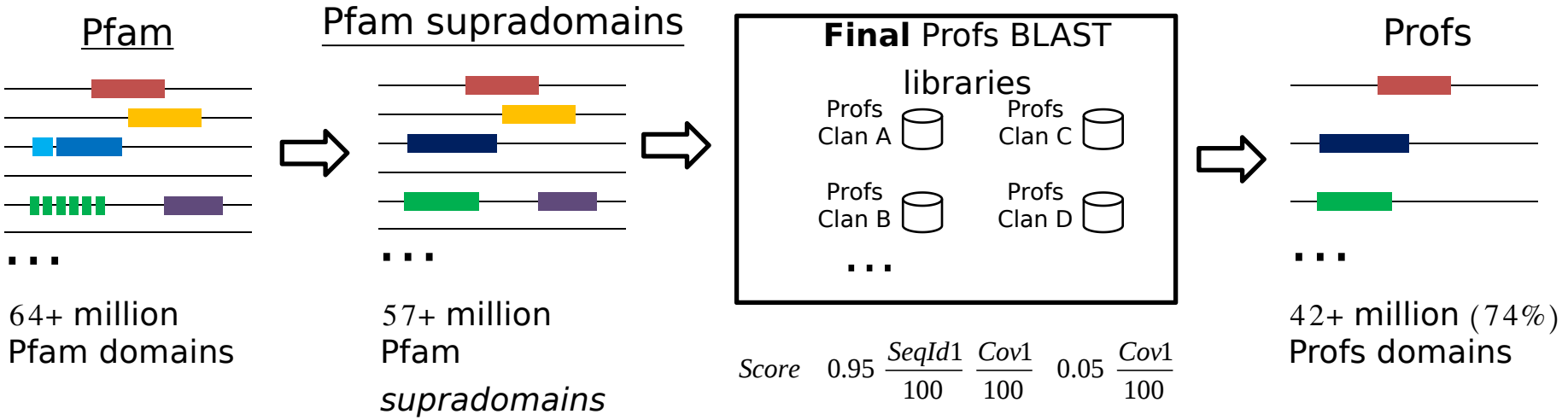
Implementation - Domain definitions

Implementation - Domain definitions

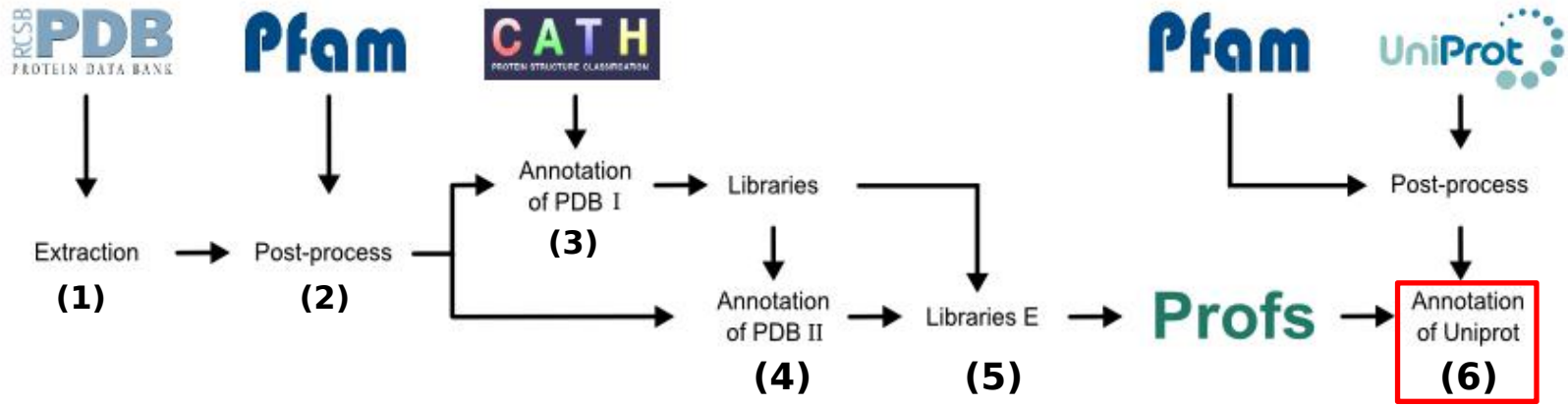


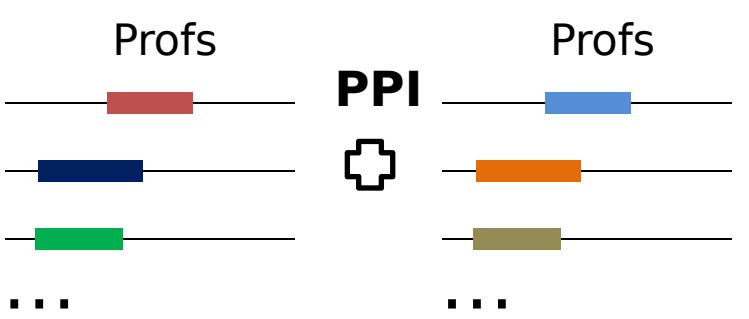
Implementation - Domain definitions





Implementation - Domain interactions

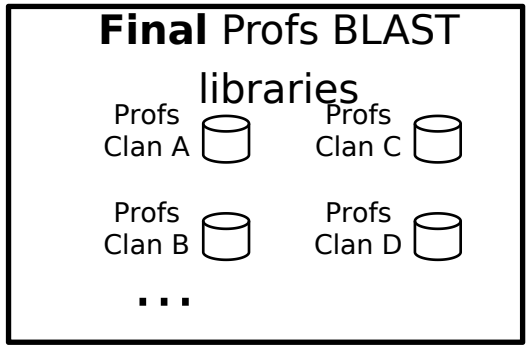




42+ million Profs domains

42+ million Profs domains

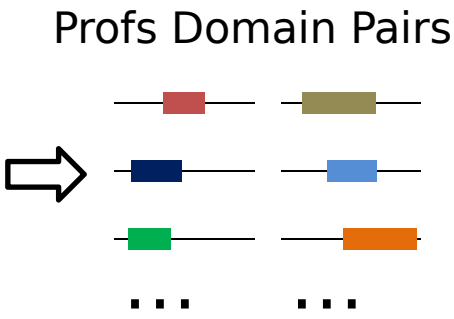
9+ million possible domain-domain combinations (mostly human)



$$Score1 \quad 0.95 \frac{SeqId1}{100} \frac{Cov1}{100} \quad 0.05 \frac{Cov1}{100}$$

$$Score2 \quad 0.95 \frac{SeqId2}{100} \frac{Cov2}{100} \quad 0.05 \frac{Cov2}{100}$$

$$Score \quad Score1 \; Score2$$



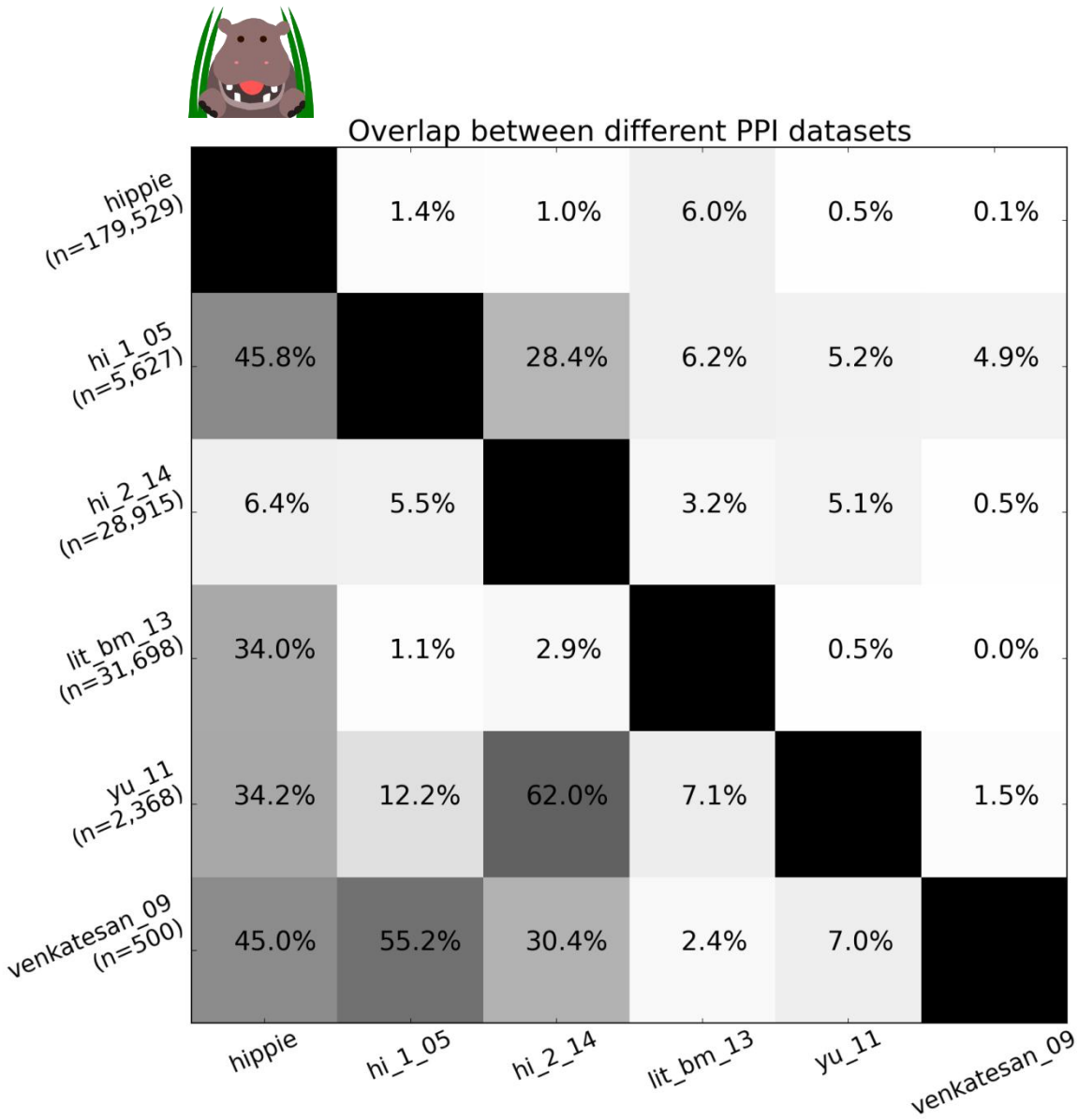
466,010 domain pairs with structural templates

Implementation - Protein interactions

HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores, Schaefer *et al.*, PLOS ONE, 2012

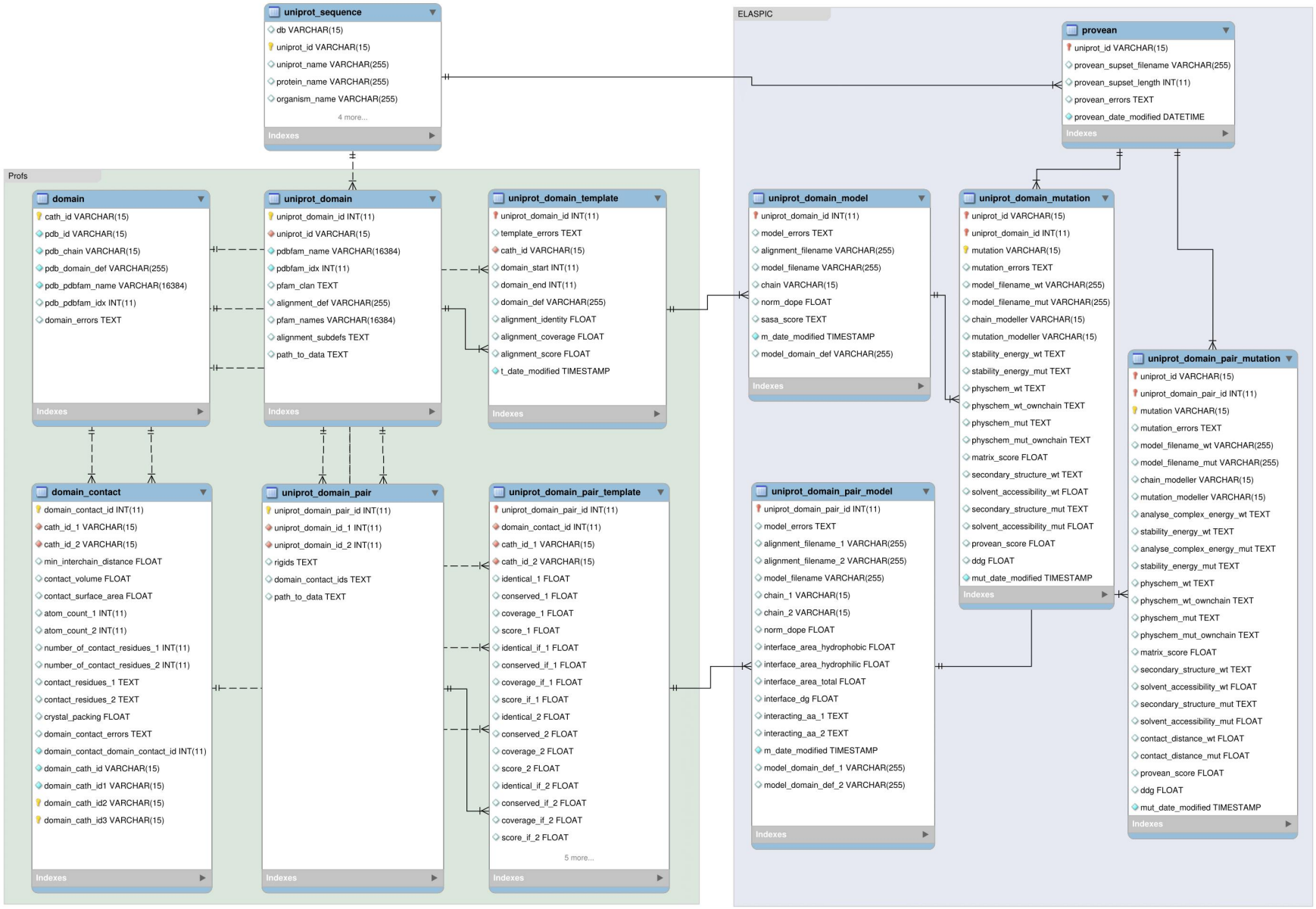
A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome, Stelzl *et al.*, Cell, 2005

A proteome-scale map of the human interactome network, Rolland *et al.*, Cell, 2014



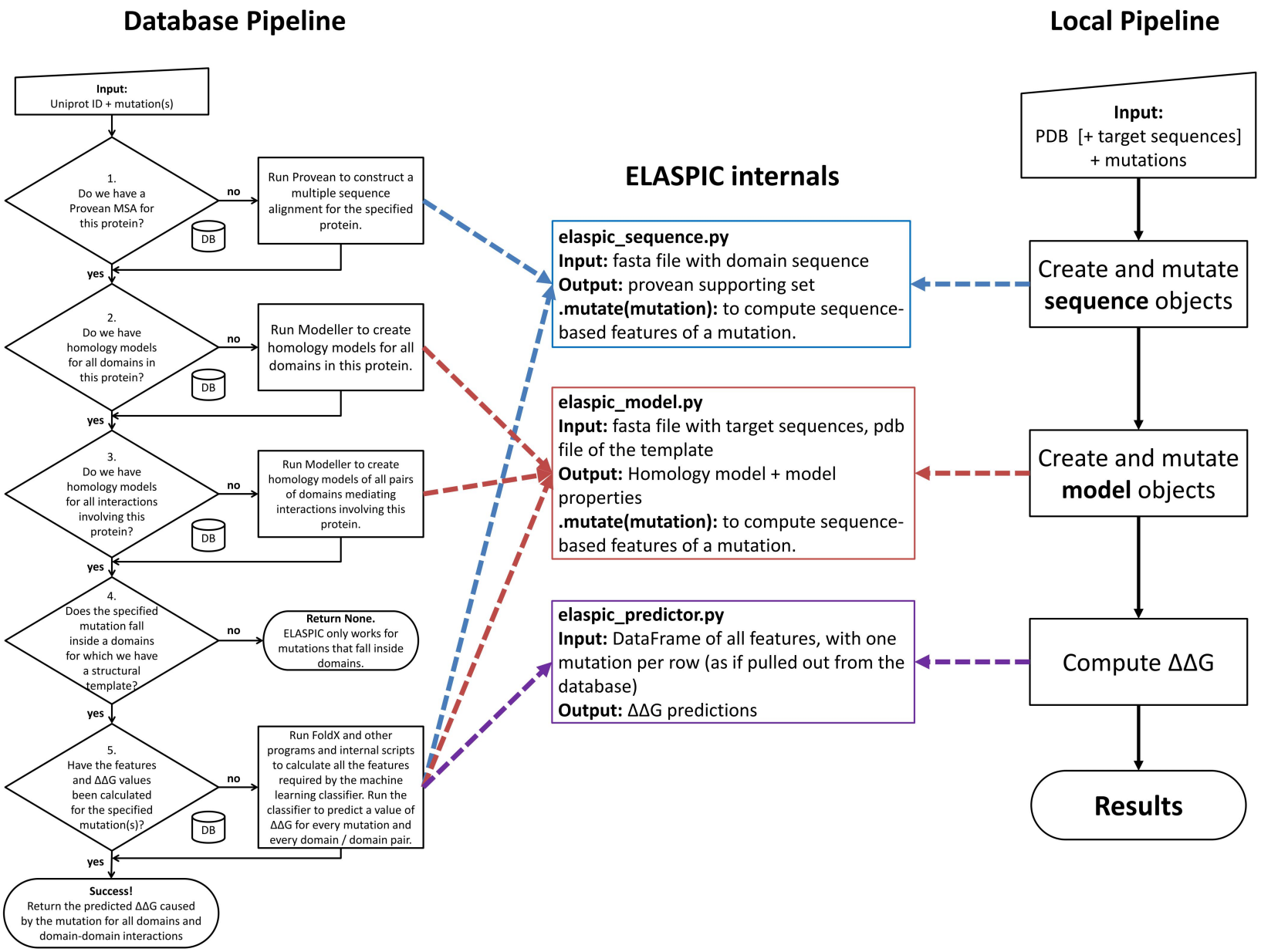
Implementation - Database backend

Implementation - Database backend



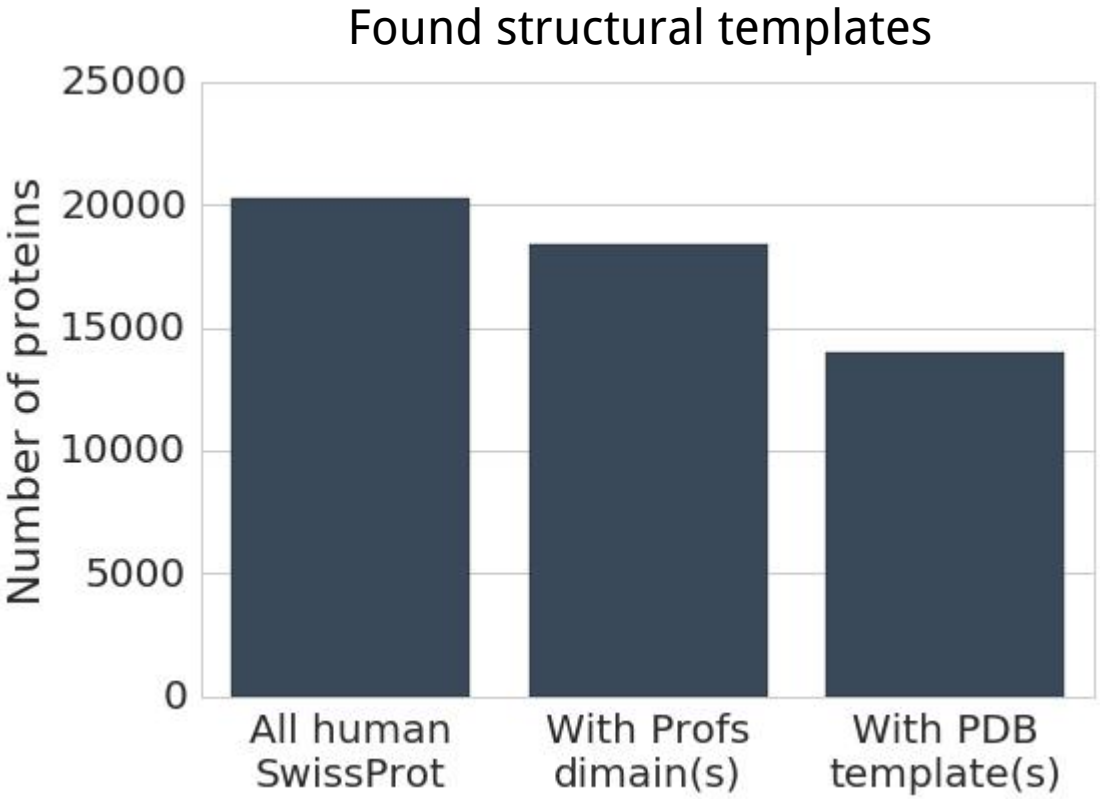
Implementation - ELASPIC pipeline and CLI

Implementation - ELASPIC pipeline and CLI

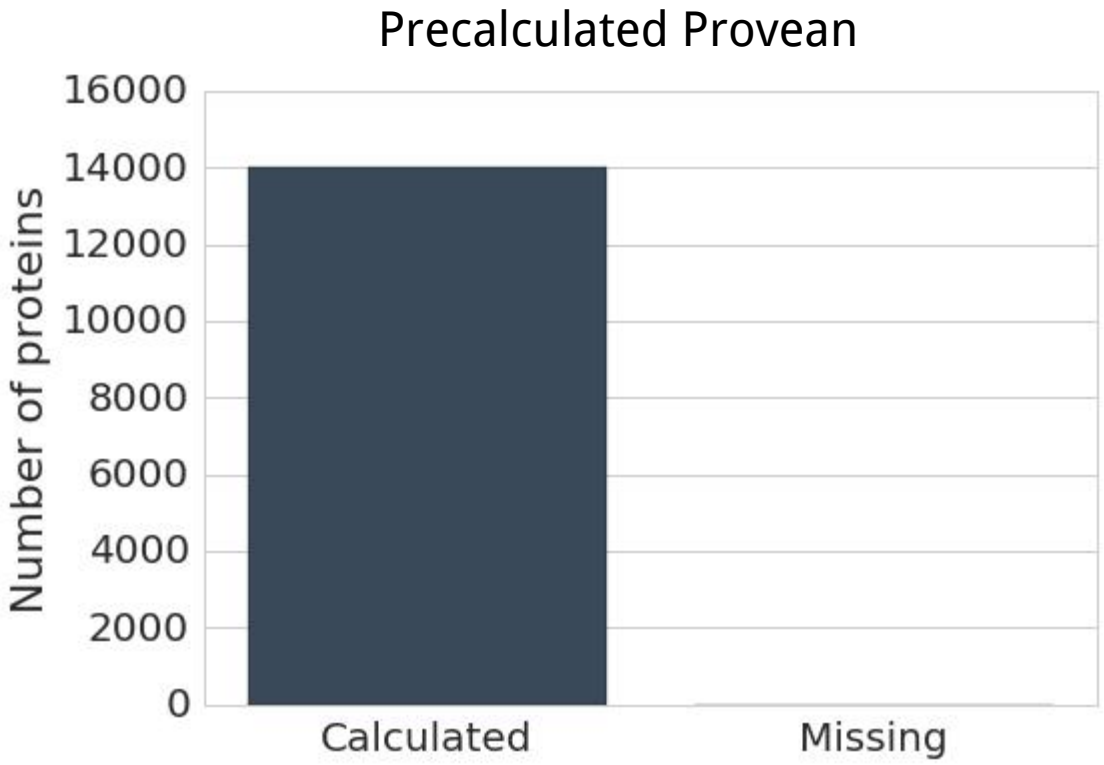


Implementation - Precalculated data

Implementation - Precalculated data

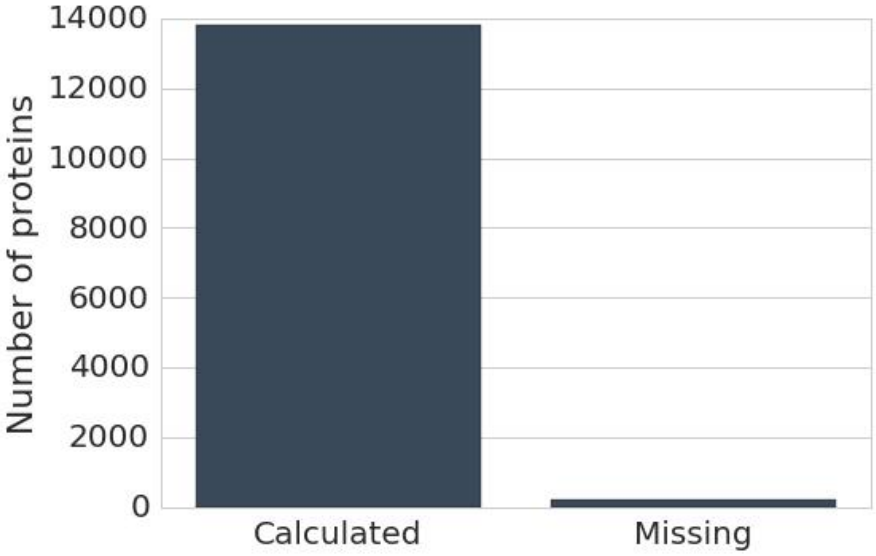


Implementation - Precalculated data

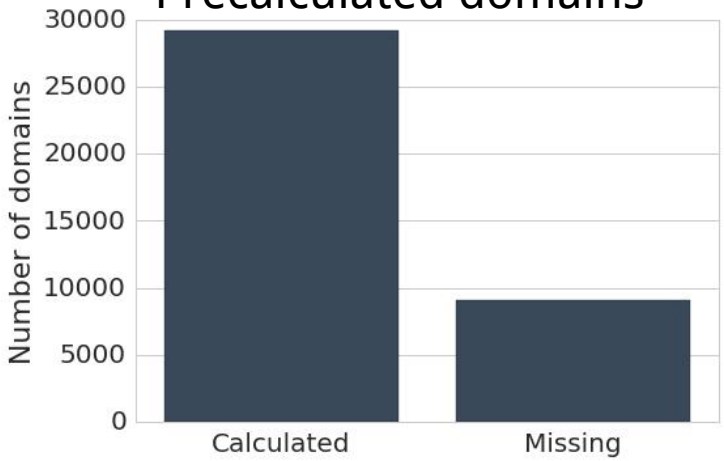


Implementation - Precalculated data

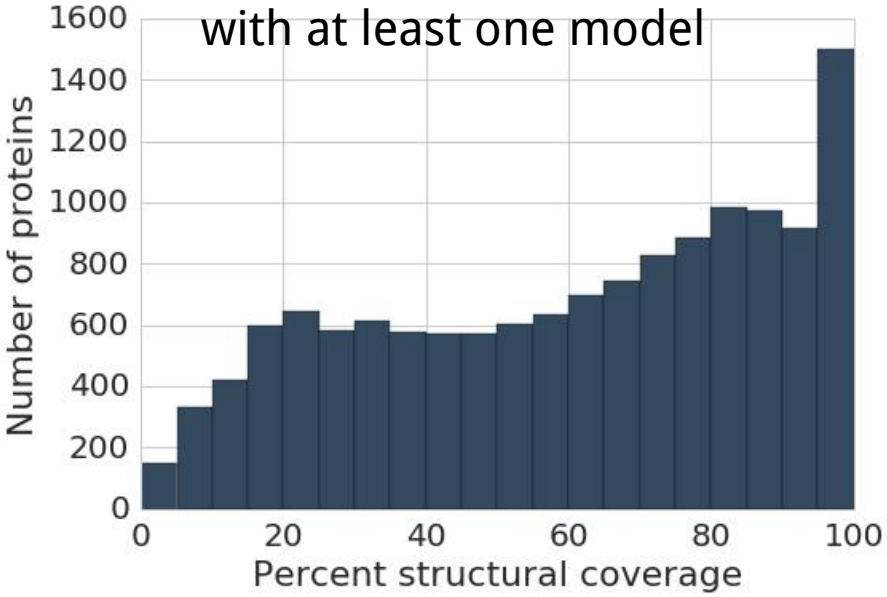
Precalculated *at least one* model



Precalculated domains

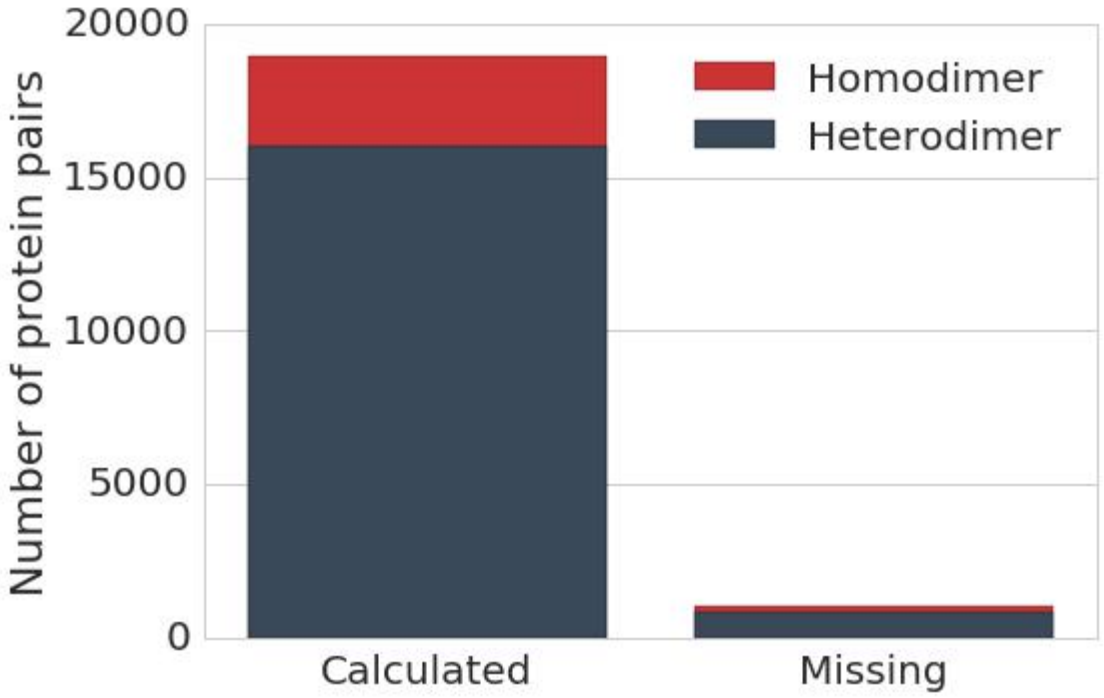


Structural coverage for proteins with at least one model



Implementation - Precalculated data

Precalculated *interfaces* (excluding splice variants)



Implementation - Precalculated data

991,533 core mutations

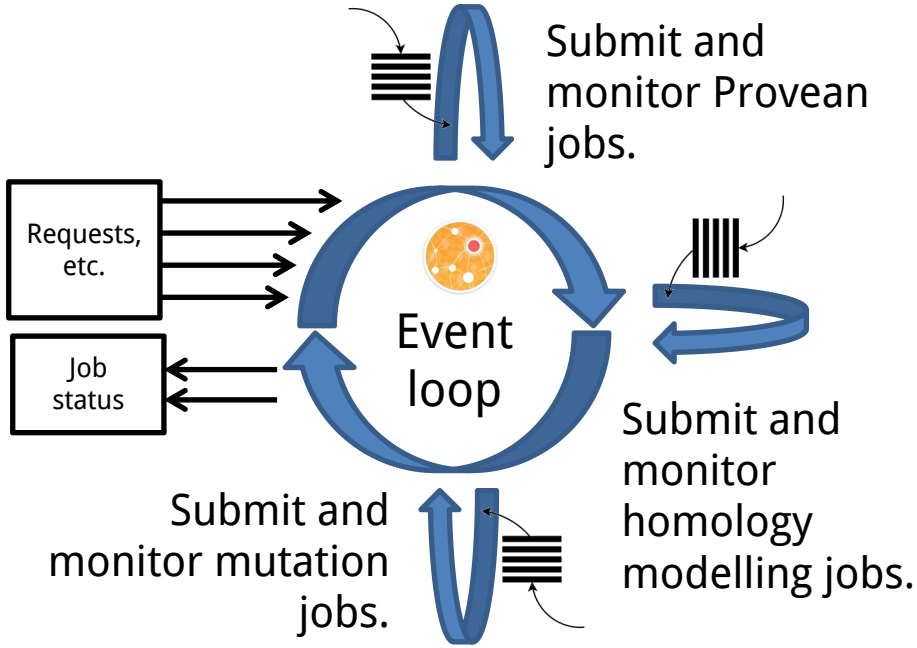
593,106 interface mutations

Includes mutations from popular databases such as UniProt (humsavar.txt), ClinVar, COSMIC.

Implementation - Webservice

Implementation - Webservice

- Asyncronous web server for submitting and monitoring ELASPIC jobs.
- Can have thousands of mutations, so having a thread monitoring each mutation would not be a good design.
- Need to update job status to "done" when all mutations are done, in order to send the "Job complete" email.
- **No blocking calls!**



Python 3.5 + aiohttp

Tornado

Results

Results

- Core mutation predictor
  - Retrain core predictor
  - Feature elimination to thrown away unimportant features
  - Validate on an independent, non-overlapping datataset

- Interface mutation predictor
  - Retrain core predictor
  - Feature elimination to thrown away unimportant features
  - Validate on an independent, non-overlapping datataset

Results - Datasets

Training:

- **Protherm**: Database of changes in the Gibbs free energy of protein folding caused by mutations.
- **Skempi**: Database of changes in the Gibbs free energy of protein folding caused by mutations.

Training / Validation:

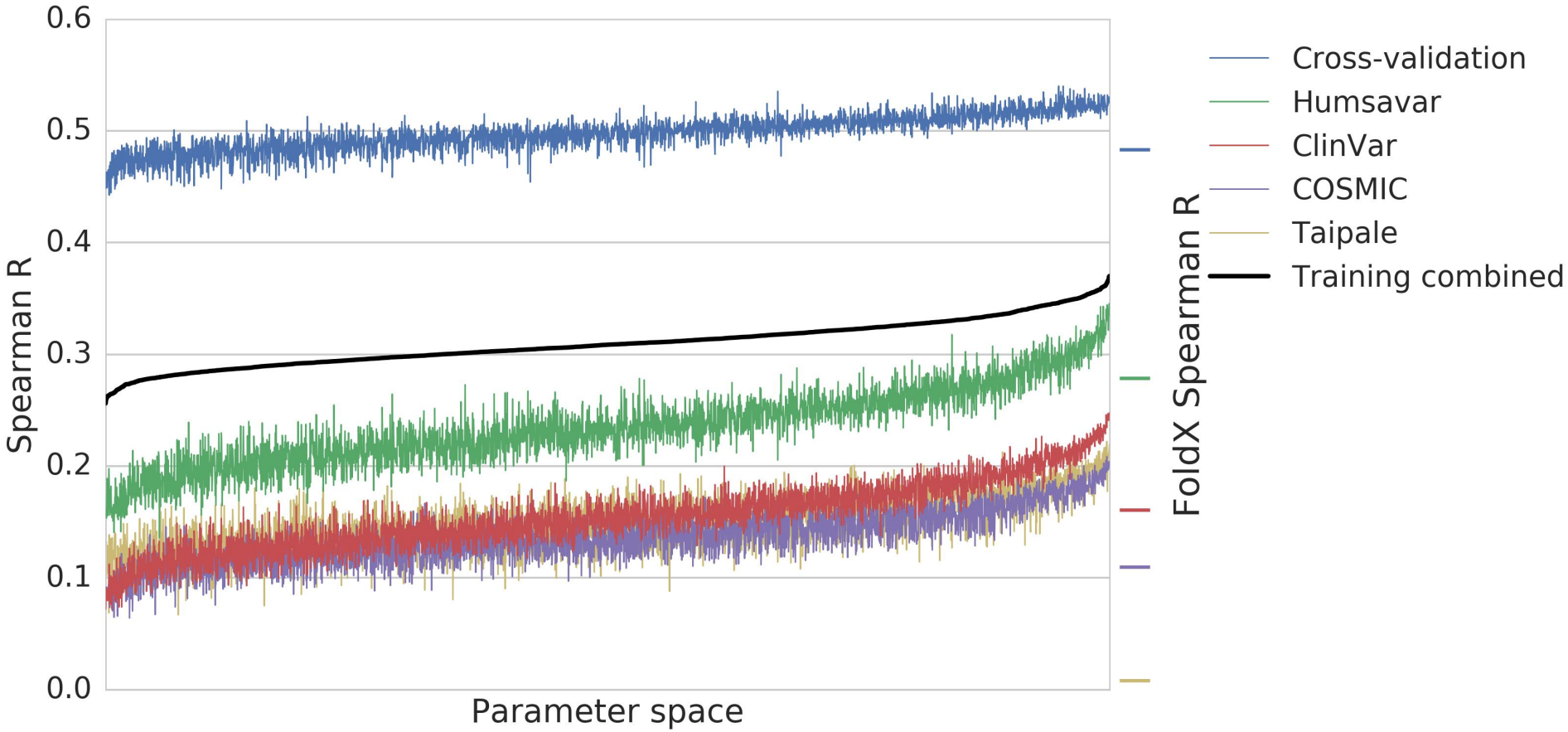
- **Taipale**: Chaperone interaction assay measuring protein stability.
- **Taipale PPI**: "Yiest two hybrid studies measuring the effect of mutations on the presence / abscence of interactions."
- **Taipale GPCA:** "Gaussia princeps luciferase protein complementation assay" measuring the effect of mutations on protein affinity.

Test:

- **Humsavar**: Disease-causing mutations vs. polymorphisms. Mostly OMIM, old ClinVar, old COSMIC.
- **Clinvar**: Disease-causing mutations with a weaker inheritence link than OMIM.
- **COSMIC**: Mutations found in cancers. Use high-confidence FATHMM predictions.
- **SUMO Ligase**: Mutations affecting the activity of SUMO ligase, measured using a cell viability assey.
- **AB-Bind**: Antibody affinity maturation experiments.
- **Benedix *et al.***: Alanine scanning of the TEM1-BLIP (β-lactamase - β-lactamase-inhibitor) complex.

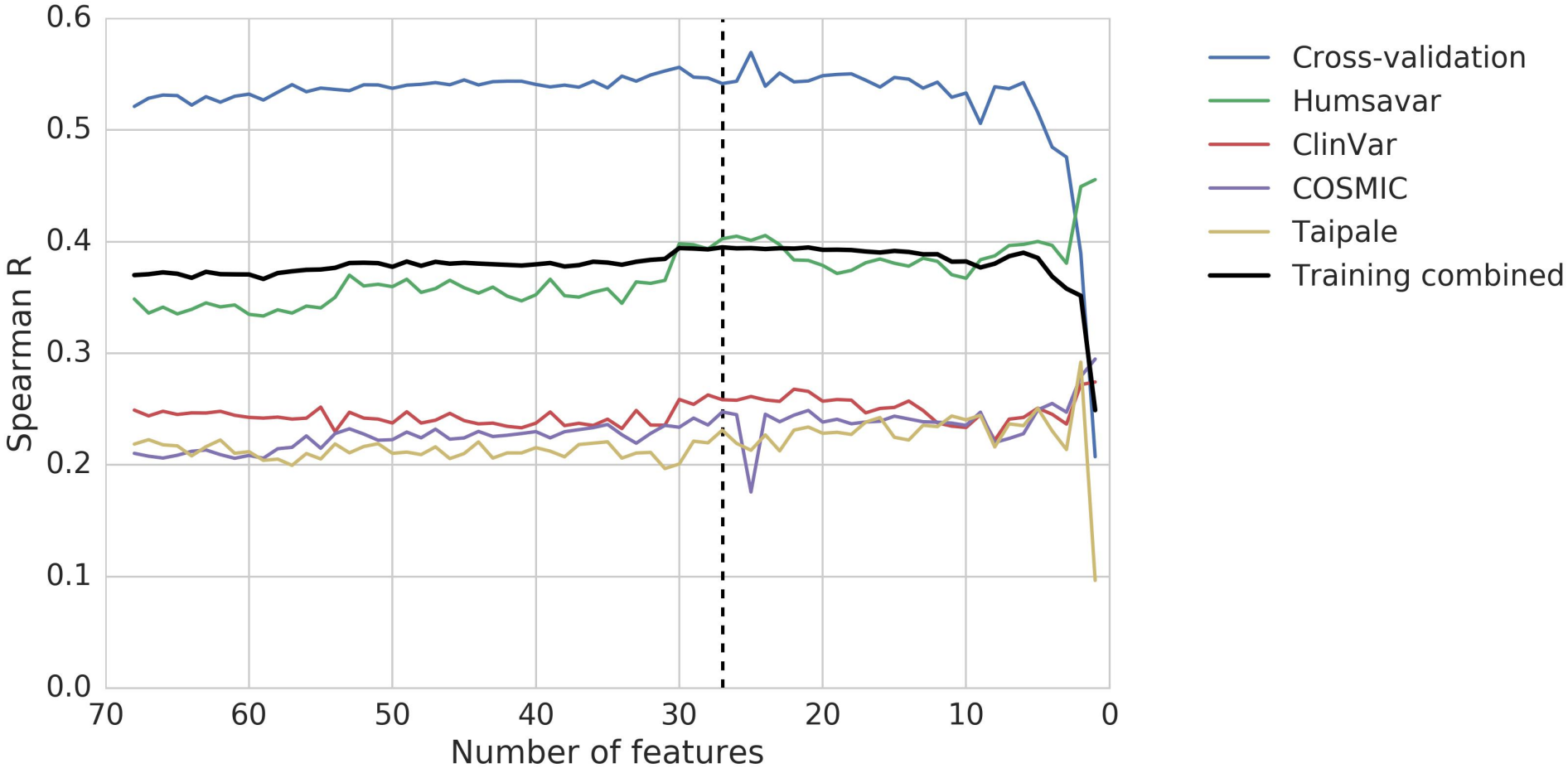
Results - Core predictor

Results - Core parameter search



Training combined = 3 × *Cross-validation* +
1 × (*Humsavar* + *ClinVar* + *COSMIC)* +
1 × *Taipale*

Results - Core feature elimination



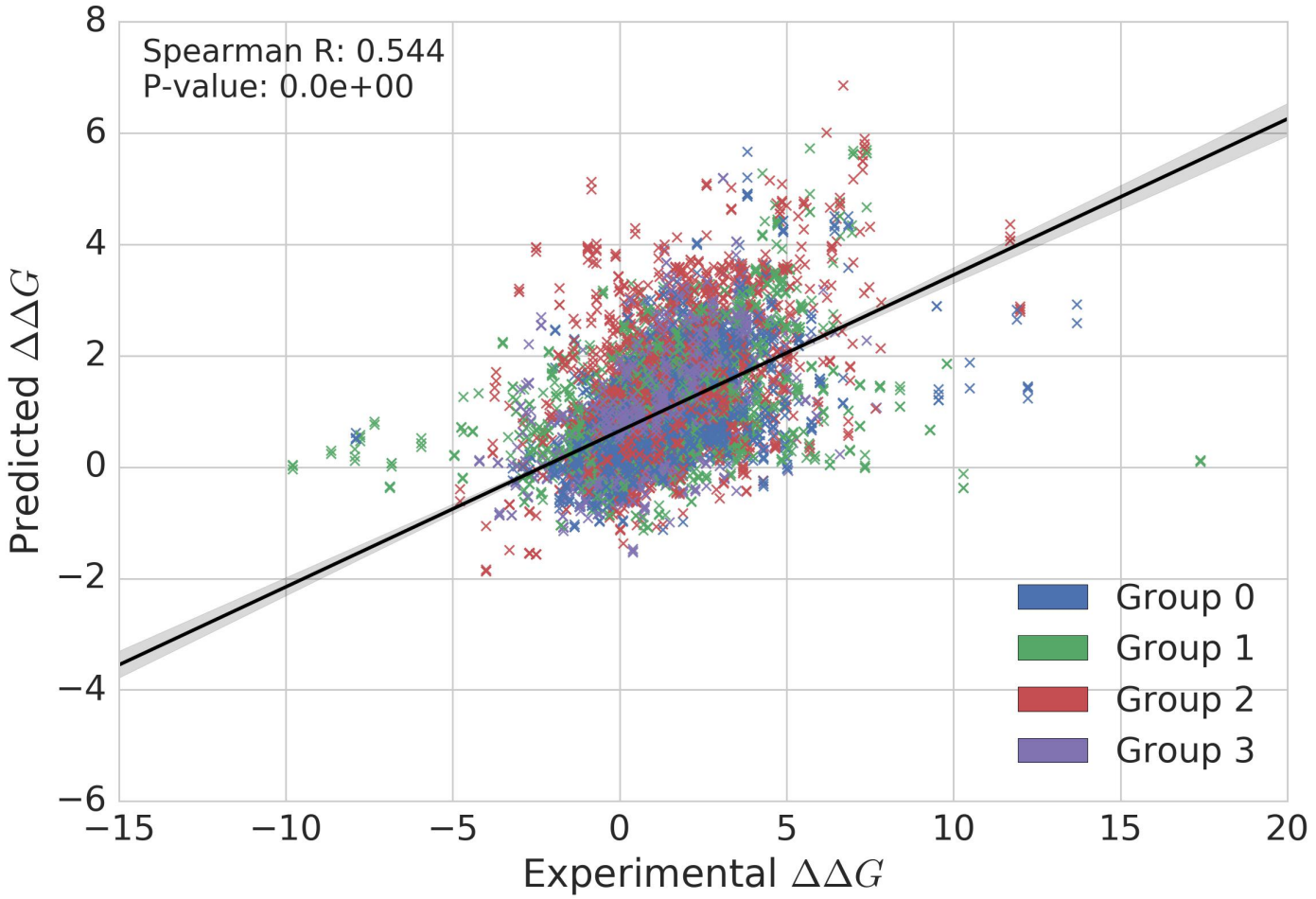
Training combined = 3 × *Cross-validation* +
1 × (*Humsavar* + *ClinVar* + *COSMIC)* +
1 × *Taipale*

Results - Core important features

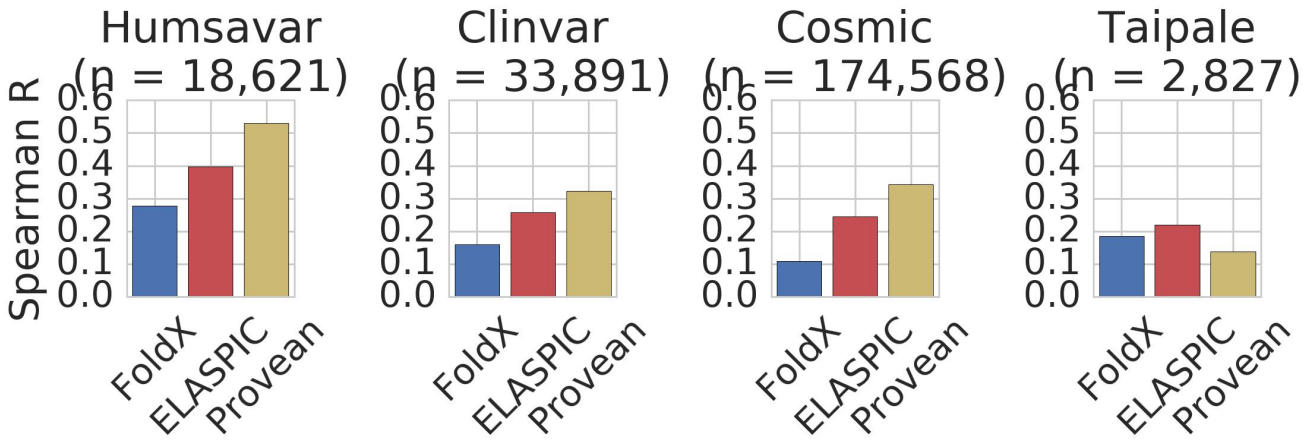
Feature name	**Feature description**	* / **
alignment_coverage	Alignment quality	
alignment_identity	Alignment quality	
alignment_score	Alignment quality	
backbone_hbond_change	FoldX	
backbone_hbond_wt	FoldX	
cis_bond_wt	FoldX	
disulfide_wt	FoldX	
electrostatic_kon_change	FoldX	
electrostatics_change	**FoldX**	*
entropy_mainchain_change	FoldX	
helix_dipole_wt	FoldX	
matrix_score	Sequence conservation	
pcv_hbond_change	Physico-chemical features	
pcv_hbond_self_change	Physico-chemical features	
pcv_salt_equal_change	Physico-chemical features	
pcv_salt_equal_self_wt	Physico-chemical features	
pcv_salt_equal_wt	Physico-chemical features	
pcv_salt_opposite_change	Physico-chemical features	
pcv_vdw_self_change	Physico-chemical features	
provean_score	**Sequence conservation**	**
sloop_entropy_wt	FoldX	
solvation_hydrophobic_change	**FoldX**	*
solvation_polar_change	**FoldX**	**
solvent_accessibility_wt	**FoldX**	*
torsional_clash_change	FoldX	
van_der_waals_clashes_change	**FoldX**	*
water_bridge_wt	FoldX	

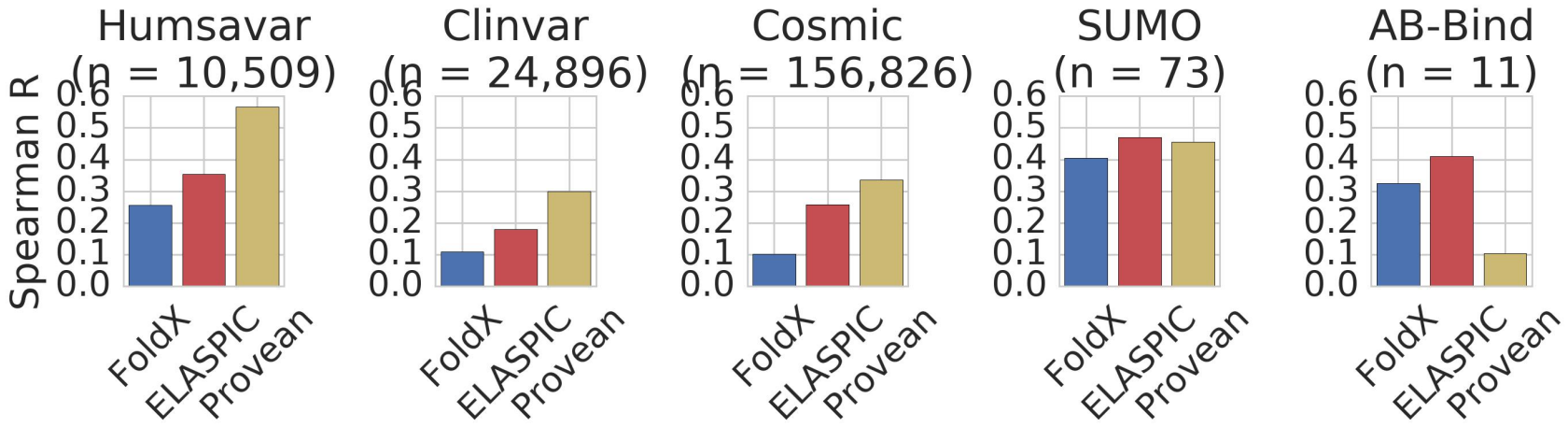
* top 6 ** top 6 interface

Results - Core cross-validation



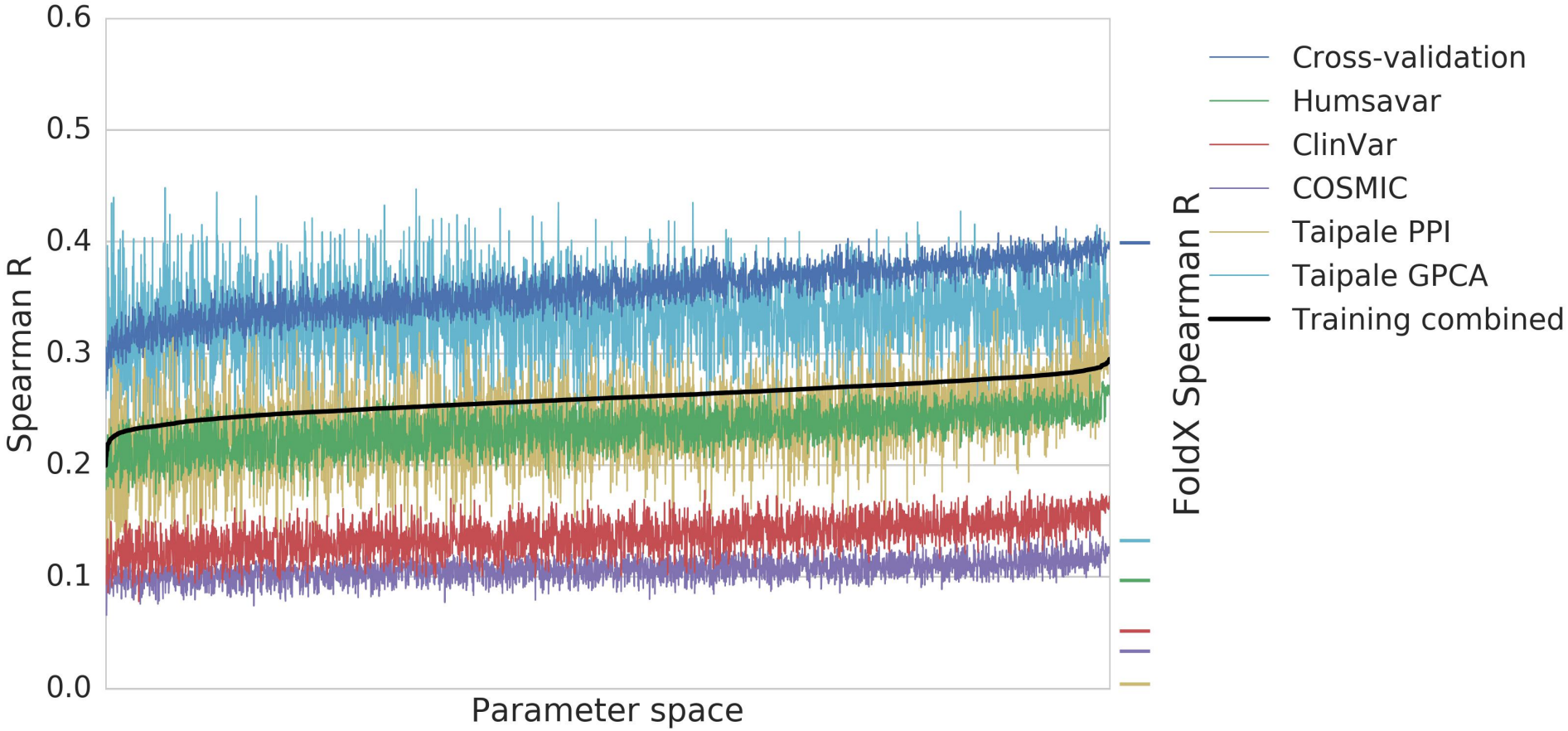
Results - Core validation/test performance





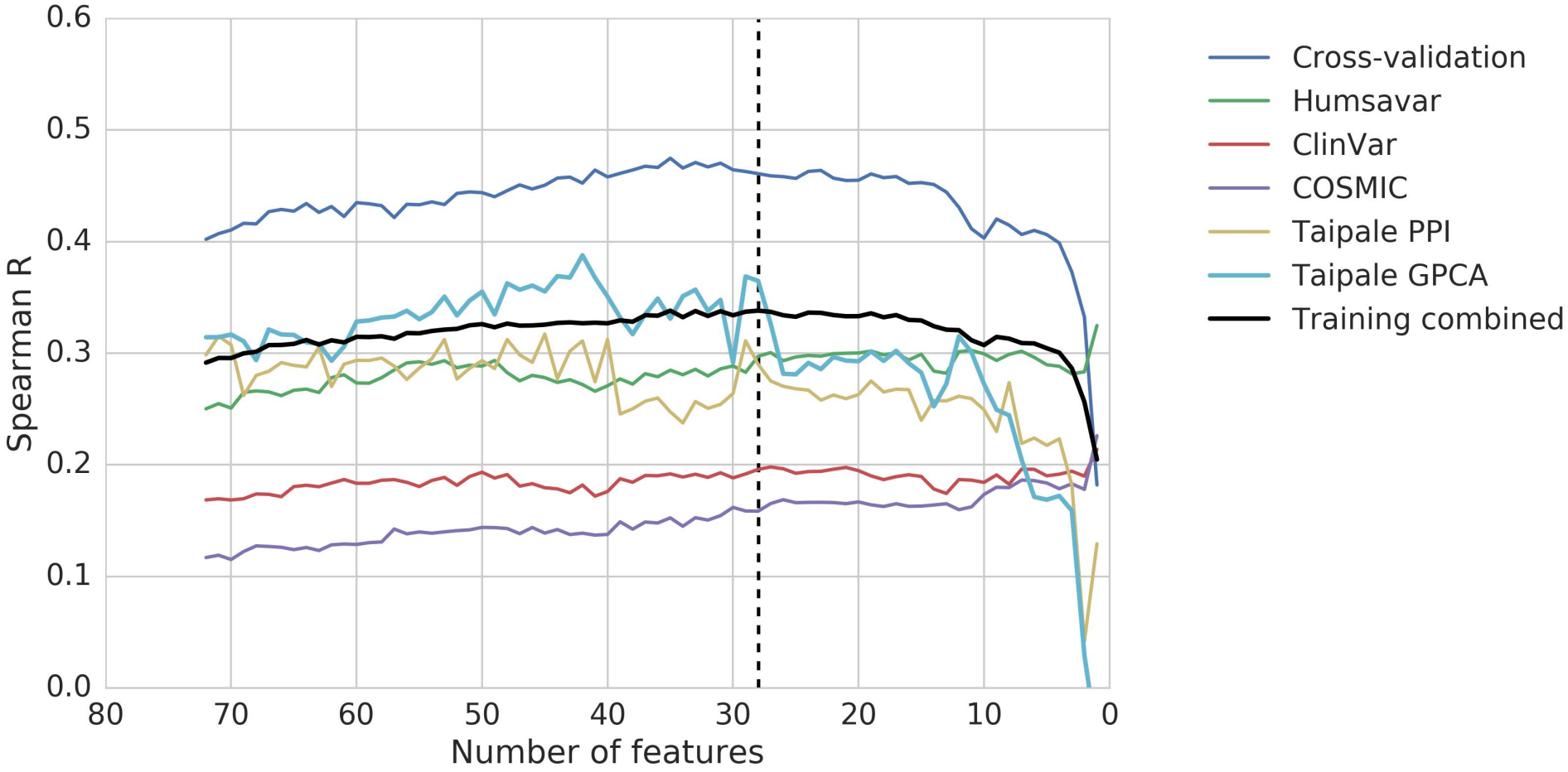
Results - Interface predictor

Results - Interface parameter search



Training combined = 3 × *Cross-validation* +
1 × (*Humsavar* + *ClinVar* + *COSMIC)* +
1/4 × (*Taipale PPI* + *Taipale GPCA*)

Results - Interface feature elimination



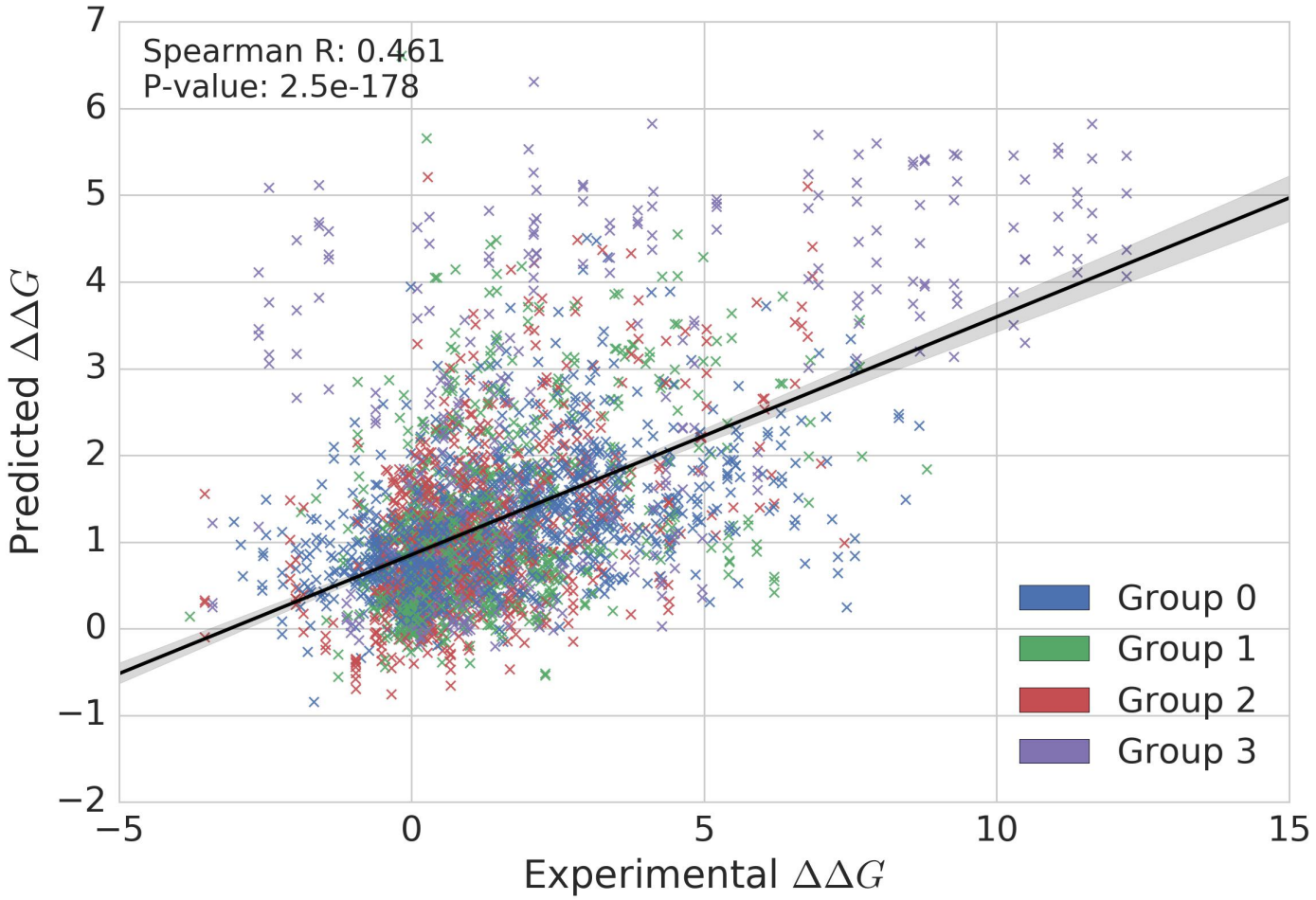
Training combined = 3 × *Cross-validation* +
1 × (*Humsavar* + *ClinVar* + *COSMIC)* +
1/4 × (*Taipale PPI* + *Taipale GPCA*)

Results - Interface important features

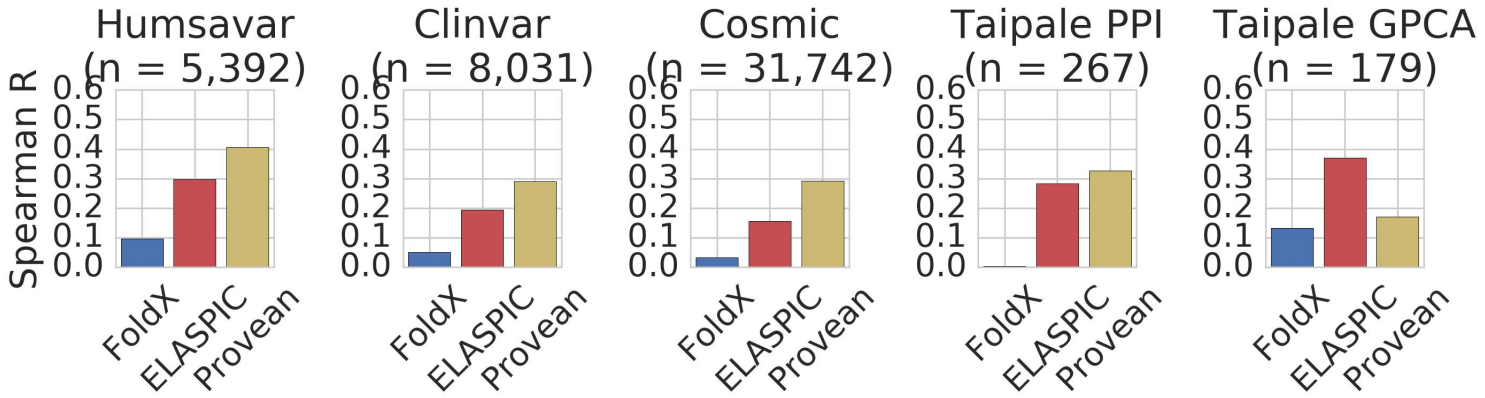
Feature name	**Feature description**	* / **
alignment_score	Alignment quality	
backbone_clash_change	FoldX	
backbone_clash_wt	FoldX	
backbone_hbond_change	FoldX	
cis_bond_wt	FoldX	
electrostatic_kon_wt	FoldX	
energy_ionisation_wt	FoldX	
entropy_complex_change	FoldX	
entropy_sidechain_change	**FoldX**	*
intraclashes_energy_2_change	FoldX	
partial_covalent_bonds_wt	**FoldX**	*
pcv_hbond_self_change	Physico-chemical features	
pcv_hbond_wt	Physico-chemical features	
pcv_salt_equal_self_change	Physico-chemical features	
pcv_salt_equal_wt	Physico-chemical features	
pcv_salt_opposite_change	Physico-chemical features	
pcv_salt_opposite_self_change	Physico-chemical features	
pcv_salt_opposite_self_wt	Physico-chemical features	
pcv_vdw_self_change	Physico-chemical features	
pcv_vdw_self_wt	**Physico-chemical features**	*
pcv_vdw_wt	**Physico-chemical features**	*
provean_score	**Sequence conservation**	**
sloop_entropy_change	FoldX	
solvation_hydrophobic_change	FoldX	
solvation_polar_change	**FoldX**	**
solvation_polar_wt	FoldX	
torsional_clash_change	FoldX	
water_bridge_change	FoldX	

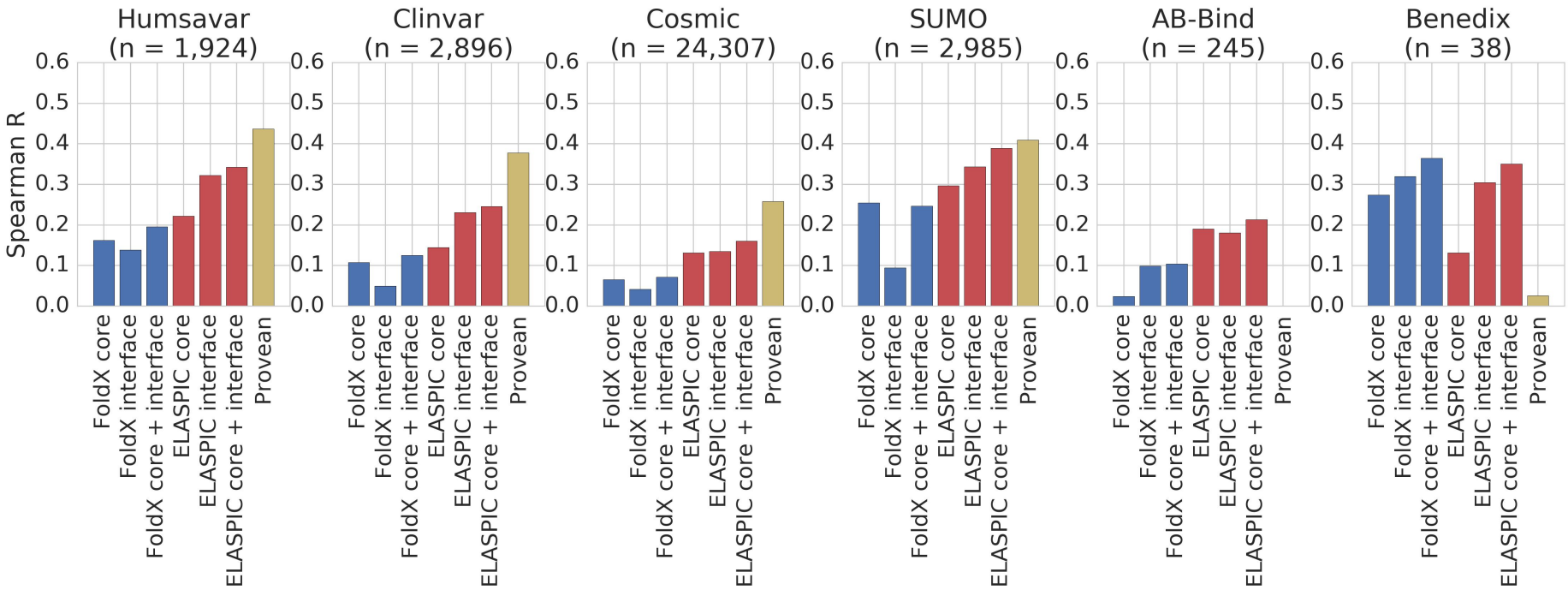
* top 6 ** top 6 core

Results - Interface cross-validation



Results - Interface validation/test performance





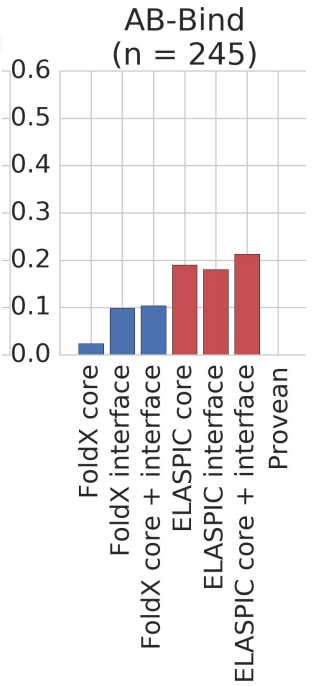
Discussion

Discussion

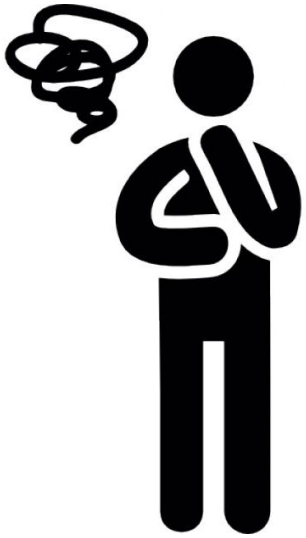
- ‘ΔΔG’ is a worse predictor of mutation deleteriousness than conservation score.
- This holds true even for interface mutations.
- No evidence that ‘edgetic’ mutations are more likely to be disease causing.
- Most features are not very informative.
- The only dataset where we do better than Provean is AB-Bind.
- …
- Need to look at deleterious mutations with a small ΔΔG and benign mutations with a high ΔΔG…

Discussion

- Don’t think predicting ΔΔG on a genome-wide scale is very useful...
- ...until we start thinking about building a thermodynamic model of the cell?
- Some features could be useful:
  - SASA scores
  - Whether or not the residue is involved in an interaction



- The only dataset where we do better than Provean is AB-Bind...
  - Maybe there is a way to ‘repurpose’ the project for protein design?
  - See future directions...



Future directions

Future directions

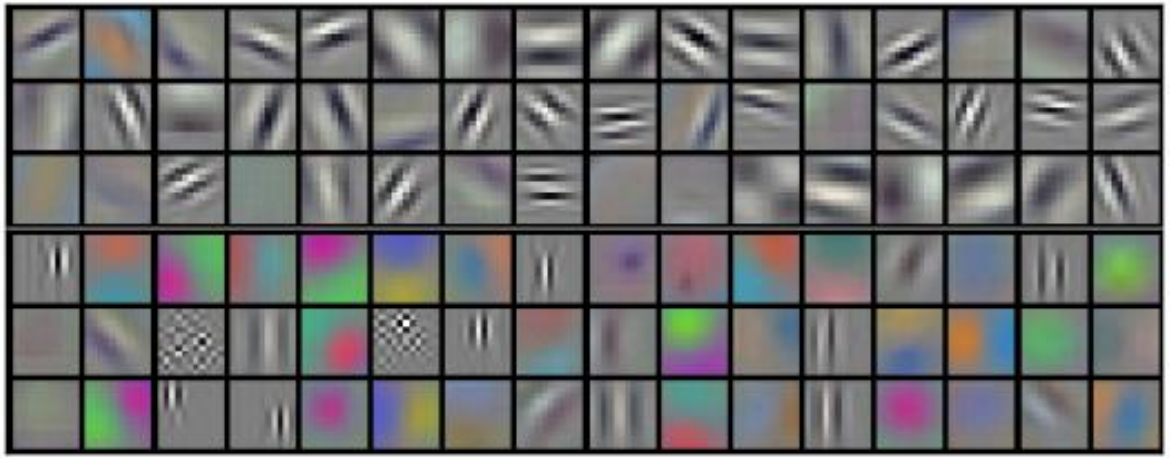
- Better features
- More training data
- ELASPIC v2.0

Future directions - Better features

Future directions - Better features

Lessons from convolutional neural networks:

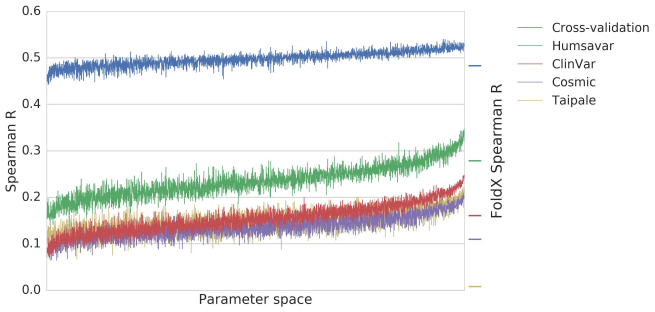
- Pretrain a deep ConvNet on a large dataset (e.g. ImageNet)
- Use the first layer(s) in other applications with a smaller training set



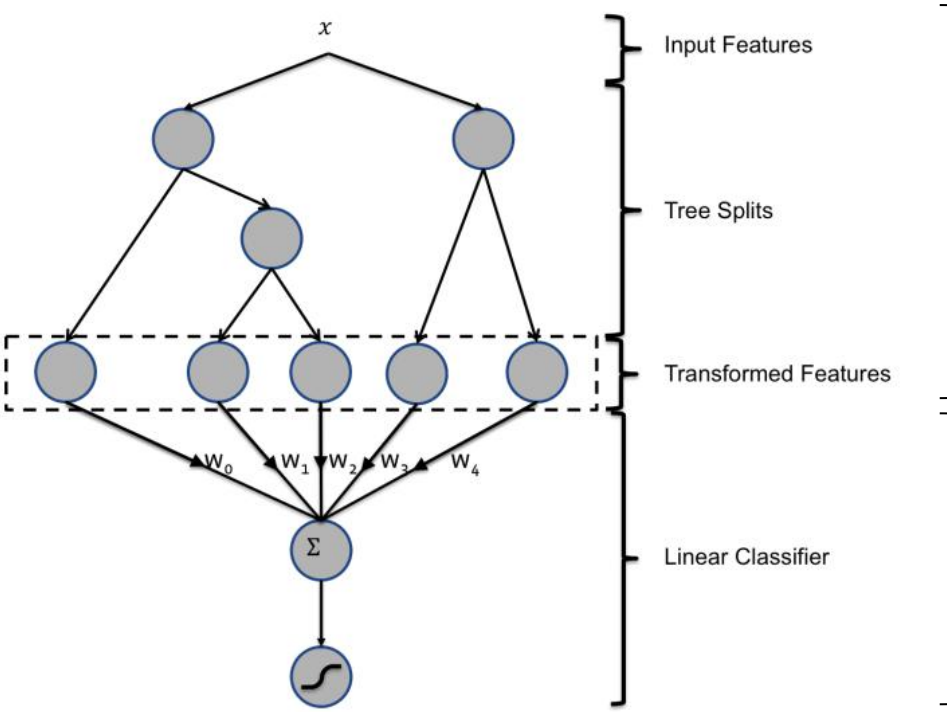
Future directions - Better features

Use the same approach for ELASPIC:

- Feature extraction using a Gradient Boosting Classifier (GBC) trained to predict deleteriousness scores
  - *With provean:* learn features with better mixing of sequence and structural information?
  - *Without provean:* learn features approximating backbone wiggle in the absence of sequence information.
- Logistic regression using trained GBC features as input to predict mutation.
- At least good performance telling you which mutations are bad...



Future directions - Better features



Train to predict the deleteriousness of mutations.

Train to predict ΔΔG.

Figure 1: Hybrid model structure. Input features are transformed by means of boosted decision trees. The output of each individual tree is treated as a categorical input feature to a sparse linear classifier. Boosted decision trees prove to be very powerful feature transforms.

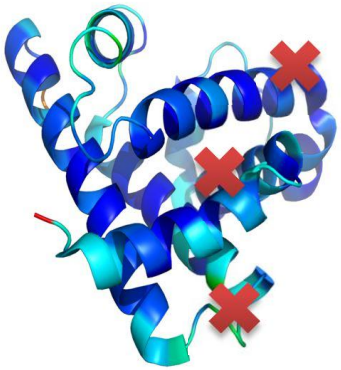
Practical Lessons from Predicting Clicks on Ads at Facebook. He Xinran *et al.* International Workshop on Data Mining for Online Advertising (ADKDD). August 24, 2014.

Future directions - More data

Future directions - More data

Include *multi-residue mutations*

- Bootstrap ΔΔG scores of constituent mutations by normalizing the ΔΔG given by ELASIC by the divergence between the ELASPIC predicted and experimental values.



Actual: 3.3 kcal / mol

1st: 3.3 kcal / mol

2nd: 2.2 kcal / mol

2nd: 1.1 kcal / mol

Total: 6.6 kcal / mol

Normalize (divide by 2 in this case) and add to the ELASPIC training data.

Future directions - Even more data

Include other datasets following the same "bootstrap" procedure.

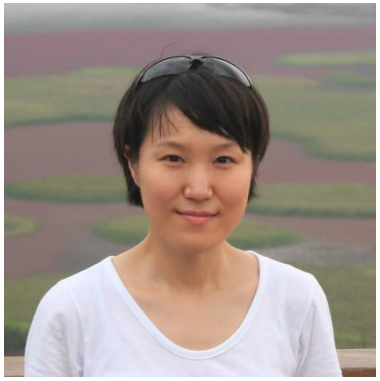
- For the *core* predictor, create a dataset of mutations to thermophilic organism coming from nearest ancesstor.
- For the *interface* predictor, convert phage-display drop-off scores to ΔΔG and add to ELASPIC training set (maybe need to log-transform or perform other correctionns).

Acknowledgements

Supervisor:
Philip M. Kim

Members of kimlab:
Recep Colak
Carles Corbi
Michael Garton
Clare Juhyun Jeon
Mark Sun
Joan Teyra

Project students:
Niklas Berliner
Andres Felipe Giraldo Forero
Hanif Jetha
Sebastian Garcia Lopez
Daniel Witvliet



Clare Juhyun Jeon

Joan Teyra