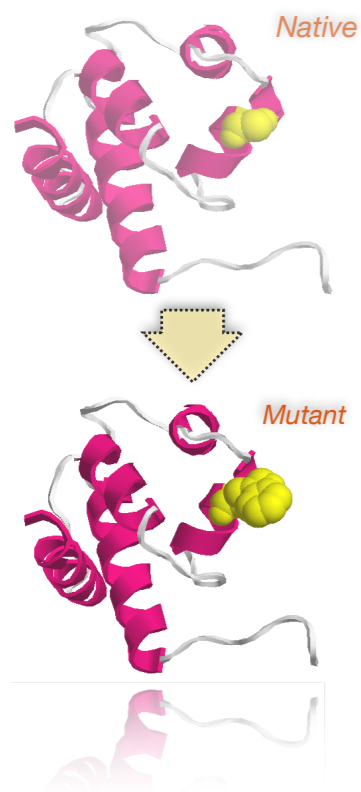
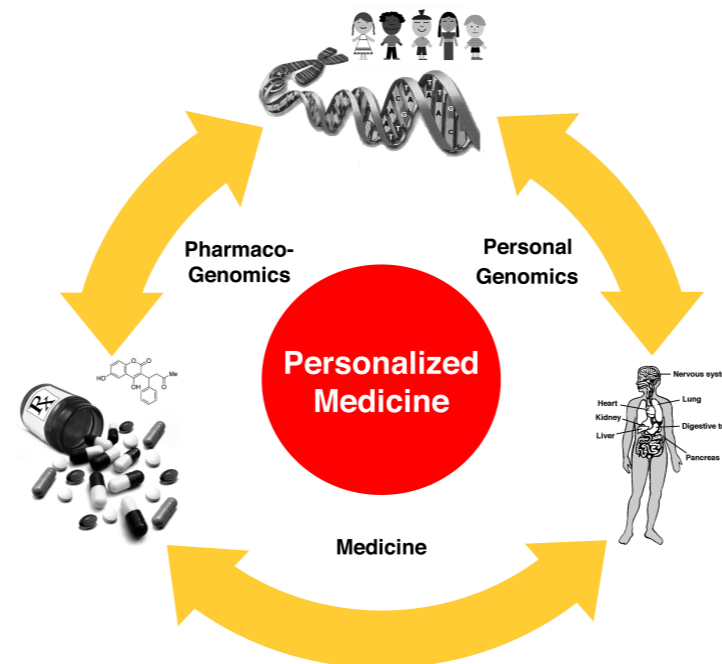
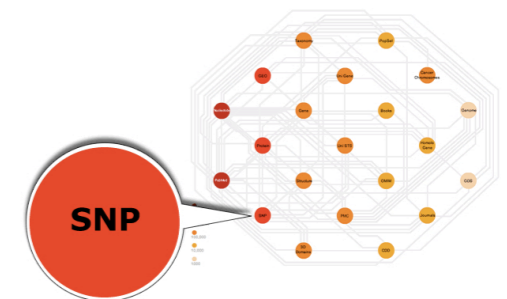


Methods for the automatic annotation of protein variants



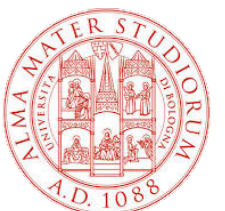
Meeting PRIN 2017, Bologna (Italy)
September 5, 2019



Emidio Capriotti
<http://biofold.org/>

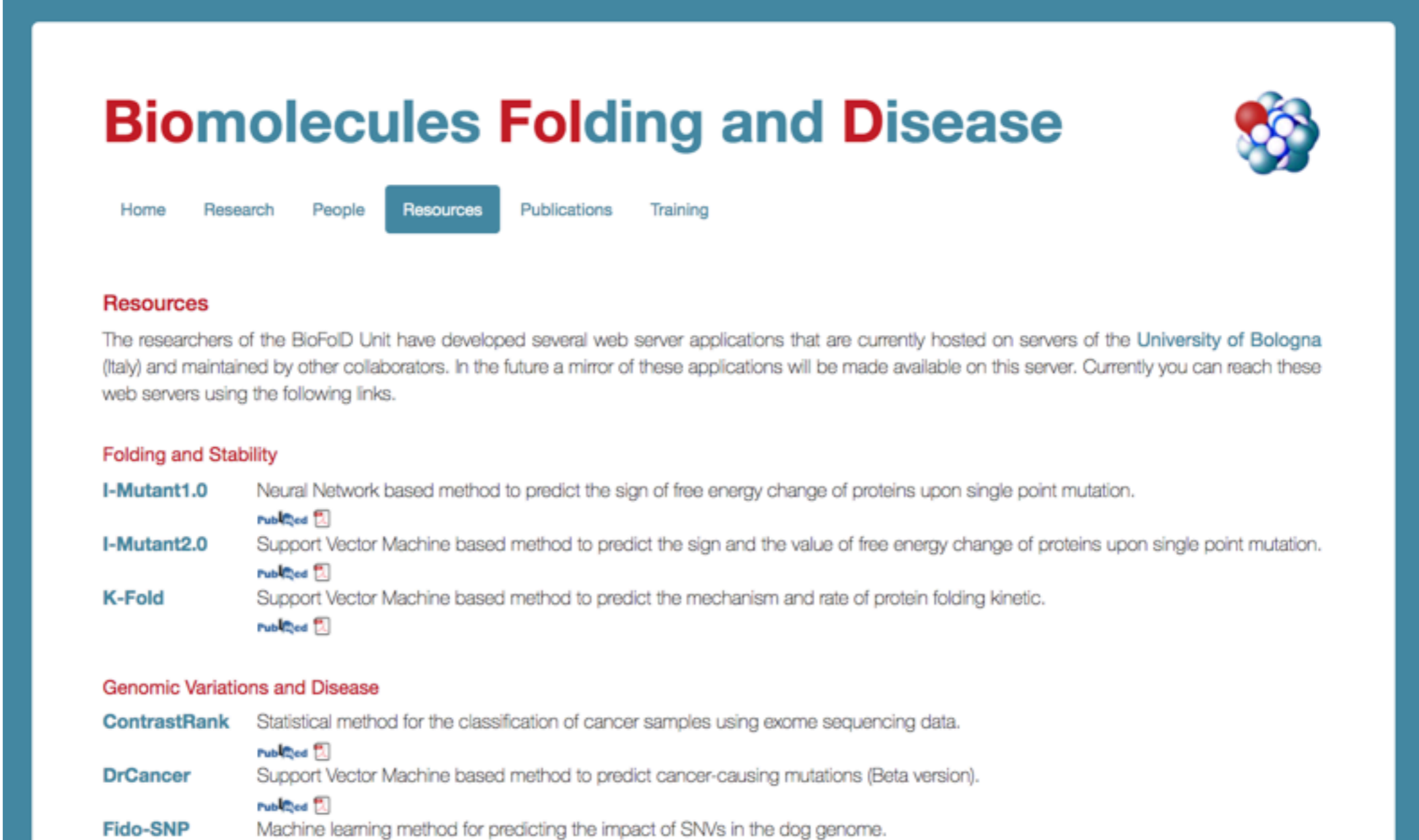


Department of Pharmacy
and Biotechnology (FaBIT)
University of Bologna




Variant annotation methods

Since 2004 we started to develop methods for the annotation of protein variants at stability and functional levels



The screenshot shows the 'Resources' page of the 'Biomolecules Folding and Disease' website. The page features a navigation menu with 'Resources' highlighted. Below the menu, there is a section titled 'Resources' with a paragraph of introductory text. This is followed by three sub-sections: 'Folding and Stability' and 'Genomic Variations and Disease'. Each sub-section lists specific methods with brief descriptions and 'published' icons.

Biomolecules Folding and Disease






- Home
- Research
- People
- Resources**
- Publications
- Training



Resources

The researchers of the BioFoD Unit have developed several web server applications that are currently hosted on servers of the [University of Bologna](#) (Italy) and maintained by other collaborators. In the future a mirror of these applications will be made available on this server. Currently you can reach these web servers using the following links.

Folding and Stability

- I-Mutant1.0** Neural Network based method to predict the sign of free energy change of proteins upon single point mutation.
[published](#) 
- I-Mutant2.0** Support Vector Machine based method to predict the sign and the value of free energy change of proteins upon single point mutation.
[published](#) 
- K-Fold** Support Vector Machine based method to predict the mechanism and rate of protein folding kinetic.
[published](#) 

Genomic Variations and Disease

- ContrastRank** Statistical method for the classification of cancer samples using exome sequencing data.
[published](#) 
- DrCancer** Support Vector Machine based method to predict cancer-causing mutations (Beta version).
[published](#) 
- Fido-SNP** Machine learning method for predicting the impact of SNVs in the dog genome.

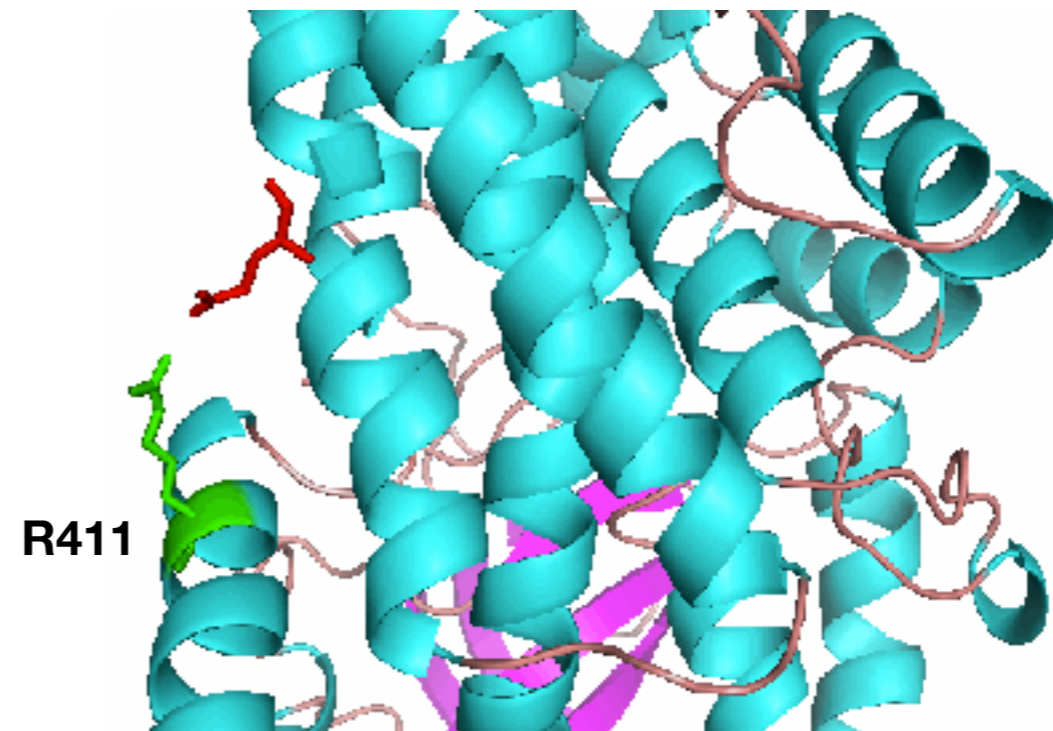
Sequence, Structure & Function

Genomic **variants in sequence motifs could affect protein function.**
Mutation S362A of P53 affect the interaction with hydrolase USP7 and the deubiquitination of the protein.



Nonsynonymous variants responsible for **protein structural changes and cause loss of stability** of the folded protein.

Mutation R411L removes the salt bridge stabilizing the structure of the IVD dehydrogenase.



Conserved or not?

In positions 66 the Glutamic acid is highly conserved Asparagine in position 138 is mutated Threonine or Alanine

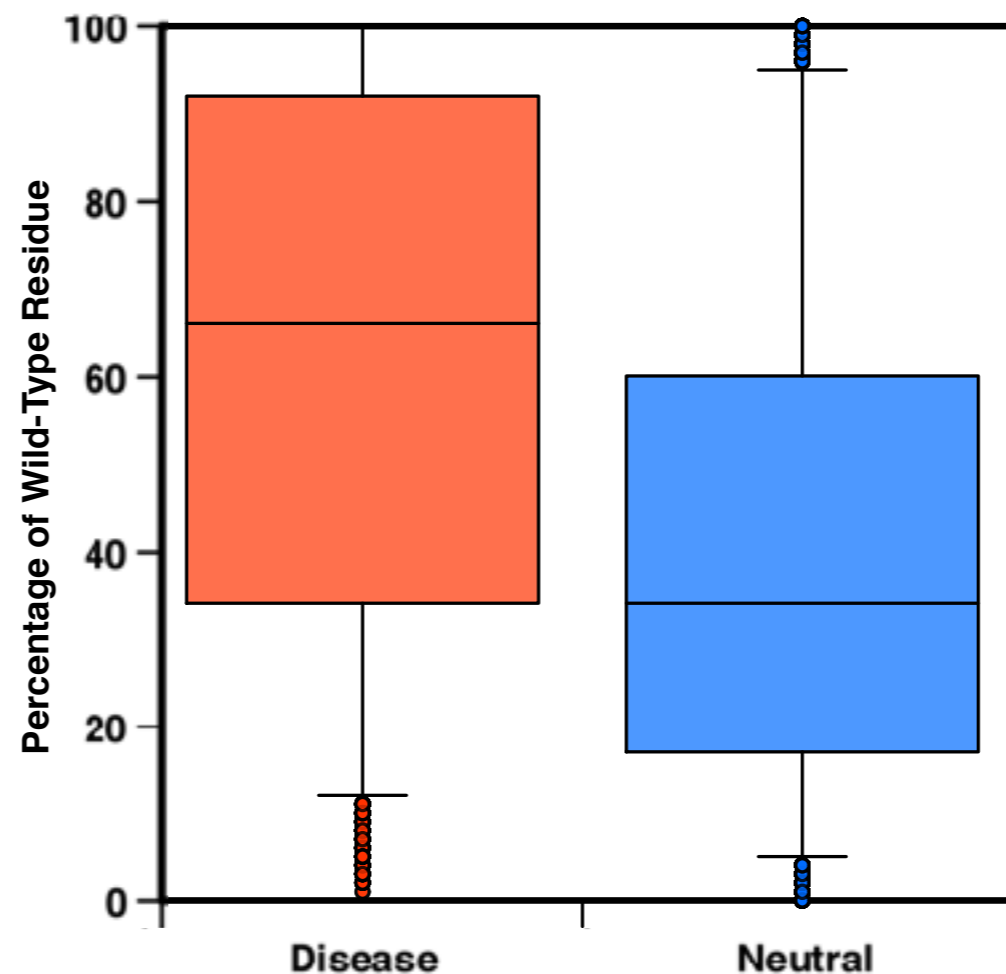
	bits	E-value	N	100.0%		1	:	80			
1 P11686	400	1e-110	1	100.0%	MDVGSKEVLMESPPDYSAAPRGRFGIPC	CPVHLK	RLLIIVVVVVVLI	VVVIVGALLMGLHMSQKHTE	MVLEMSIGAPEAQQ		
2 P15783	280	3e-74	1	80.6%	MDVGSKEVLMESPPDYSAAPRGRFGIPC	CPVHLK	RLLIIVVVVVVLI	VVVIVGALLMGLHMSQKHTE	MVLEMSIGAPEAQQ		
3 P21841	276	6e-73	1	78.7%	MDVGSKEVLMESPPDYTAVPGGRLLIPCC	PNIKR	LLIIVVVVVVLI	VVVIVGALLMGLHMSQKHTE	MVLEMSITGPEAQQ		
4 P22398	270	3e-71	1	78.2%	MDMSKEVLMESPPDYSAGPRSQFRIPCC	PVHLK	RLLIIVVVVVVLI	VVVIVGALLMGLHMSQKHTE	MVLEMSIGAPETQK		
5 Q1XFL5	268	1e-70	1	80.2%	MDMGSKAELMESPPDYSAAPRGRFGIPC	CPVHLK	RLLIIVVVVVVLI	VVVIVGALLMGLHMSQKHTE	MVLEMSIGAPEVQQ		
6 UPI0000E219B8	261	1e-68	1	89.4%	MDVGSKEVLMESPPDYSAAPRGRFGIPC	CPVHLK	RLLIIVVVVVVLI	VVVIVGALLMGLHMSQKHTE	MVLEMSIGAPEAQQ		
7 UPI00005A47C8	259	6e-68	1	78.2%	MDVGSKEVLIESPpdYSAAPRGRFGIPC	FPSSLK	RLLIIVVVIVLI	VVVIVGALLMGLHMSQKHTE	MVLEMSMGGPEAQQ		
8 Q3MSM1	206	8e-52	1	83.4%	MDVGSKEVLMESPPDYSAVPGGRRLRIPCC	PVNLK	RLLIIVVVVVVLI	VVVIVGALLMGLHMSQKHTE	MVLEMSLAGPEAQQ		
9 Q95M82	85	3e-15	1	82.4%	-----	-----	-----	-----	VLEMSIGGPEAPQ		
10 UPI000155C160	84	4e-15	1	48.9%	-----	-----	-----	-----	-----		
11 UPI0001555957	82	1e-14	1	83.6%	-----	KVRADSP	PDYSVAPRGR	LGI	CCPFHLKRLLIIVVVVVLI	VVVVLGALLMGLHMSQKHTE	M-----
12 B3DM51	81	4e-14	1	34.8%	-----	-----	-----	-----	HMSQKHTE	TIFQMSL-----	QD
.....											
.....											

	bits	E-value	N	100.0%		81	1	:	160		
1 P11686	400	1e-110	1	100.0%	RLALSEHLVTTATFSIGSTGLVVYDYQQLLI	AYKPAPGTCCYIMKIAPESIPSLEALNR	KVHNFQMECSLQAKPAVPTSK				
2 P15783	280	3e-74	1	80.6%	RLALSEHLVTTATFSIGSTGLVVYDYQQLLI	AYKPAPGTCCYIMKIAPESIPSLEALNR	KVHNFQMECSLQAKPAVPTSK				
3 P21841(Mouse)	276	6e-73	1	78.7%	RLALSERVGTATFSIGSTGTVVYDYQRLLI	AYKPAPGTCCYIMKMAPQNI	PSLEALTRKLNQF-----	QAKPQVPSSK			
4 P22398	270	3e-71	1	78.2%	RLAPSERADTIATFSIGSTGIVVYDYQRLLI	AYKPAPGTCCYIMKMAPESIPSLEAF	ARKLNQF-----	RAKPSTPTSK			
5 Q1XFL5	268	1e-70	1	80.2%	RLALSEWAGTTATFPIGSTGIVTCDYQRLLI	AYKPAPGTCCYLMKMAPDSIPSLEAL	ARK-----	FQANPAEPPTQ			
6 UPI0000E219B8	261	1e-68	1	89.4%	RLALSEHLVTTATFSIGSTGLVVYDYQQLLI	AYKPAPGTCCYIMKIAPESIPSLEALNR	KVHNFQMECSLQAKPAVPTSK				
7 UPI00005A47C8	259	6e-68	1	78.2%	RLALQERVGTATFSIGSTGIVVYDYQRLLI	AYKPAPGTCCYIMKMPENIPSLEAL	TRKFQDFQV-----	KPAVSTSK			
8 Q3MSM1	206	8e-52	1	83.4%	RLALSEHLVTTATFSIGSTGLVVYDYQQLLI	AYKPAPGTCCYIMKMAPQNI	PSLEALTRKLNQF-----	QAKPQVPSSK			
9 Q95M82	85	3e-15	1	82.4%	RLALRGRADTTATFSIGSTGIVVYDYQRLLI	AYKPAPG-----	-----	-----			
10 UPI000155C160	84	4e-15	1	48.9%	-----	-----	-----	-----	RLLIAYQPSPGATCYVTKMAPENIPSLDAITRE	---FQ---SYQAKPSMPATK	
11 UPI0001555957	82	1e-14	1	83.6%	-----	-----	-----	-----	-----	-----	
12 B3DM51	81	4e-14	1	34.8%	GSSTGAHGTGVATfgINSASVVYDYSKLLI	IGTRPRPGHACYITRMDPEQVQSLETIAESV	-----	-----	-----	-----	LSK

Sequence profile

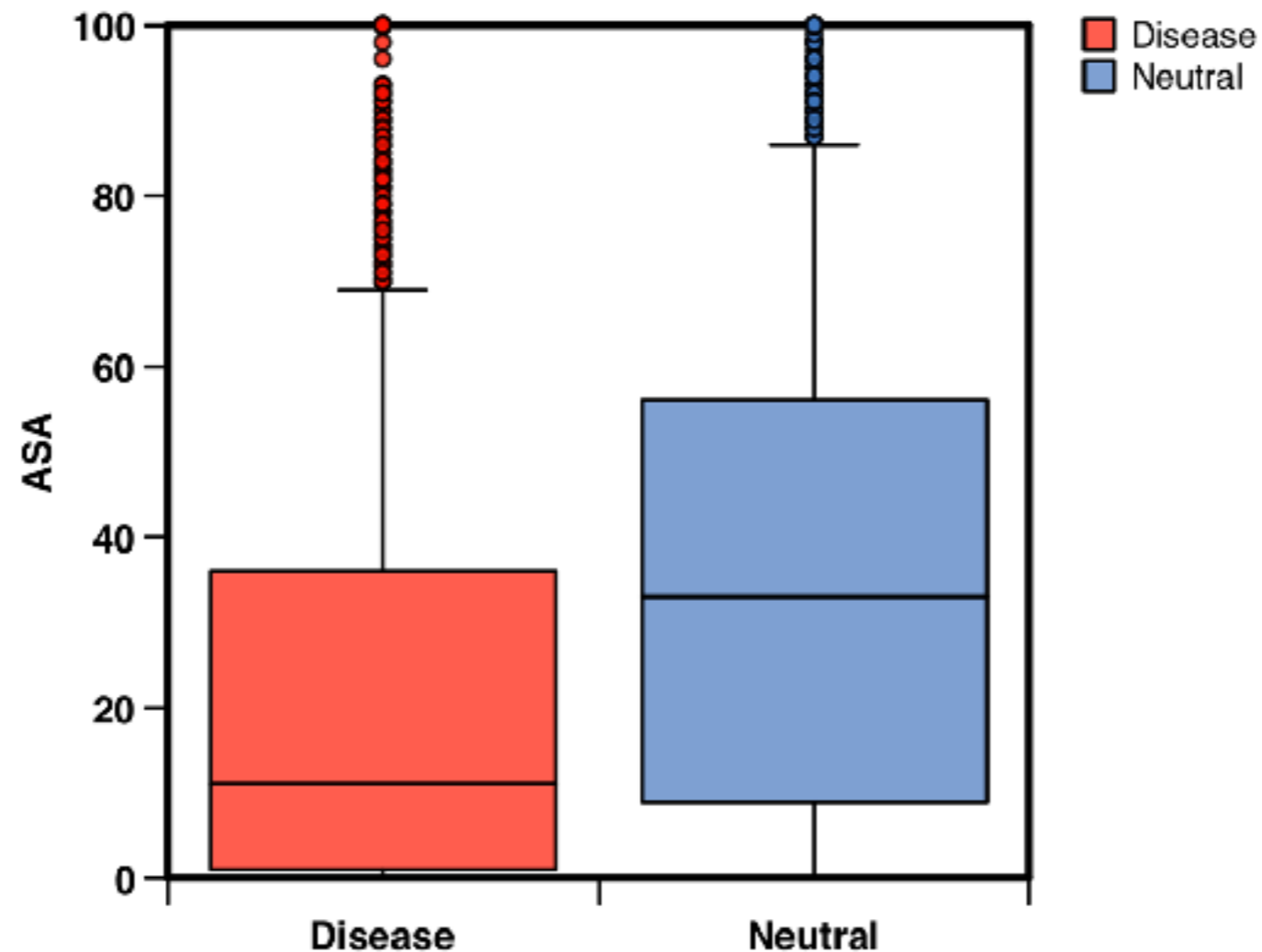
The protein **sequence profile** is calculated running **BLAST on the UniRef90** dataset and selecting only the hits with e-value $< 10^{-9}$.

The **frequency distributions of the wild-type residues** for disease-related and neutral variants are significantly different (KS p-value=0).



Structure environment

There is a **significant difference** (KS p-value = 2.8×10^{-71}) between the **distributions of the Relative Accessible Solvent Area for disease-related and neutral variants**. Their mean values are respectively 20.6 and 35.7.



Protein variant databases

Many database are available collecting data about protein variations. Some of them are not longer updated



swissvar



ProTherm

Thermodynamic Database
for Proteins and Mutants



Protein Mutant Database

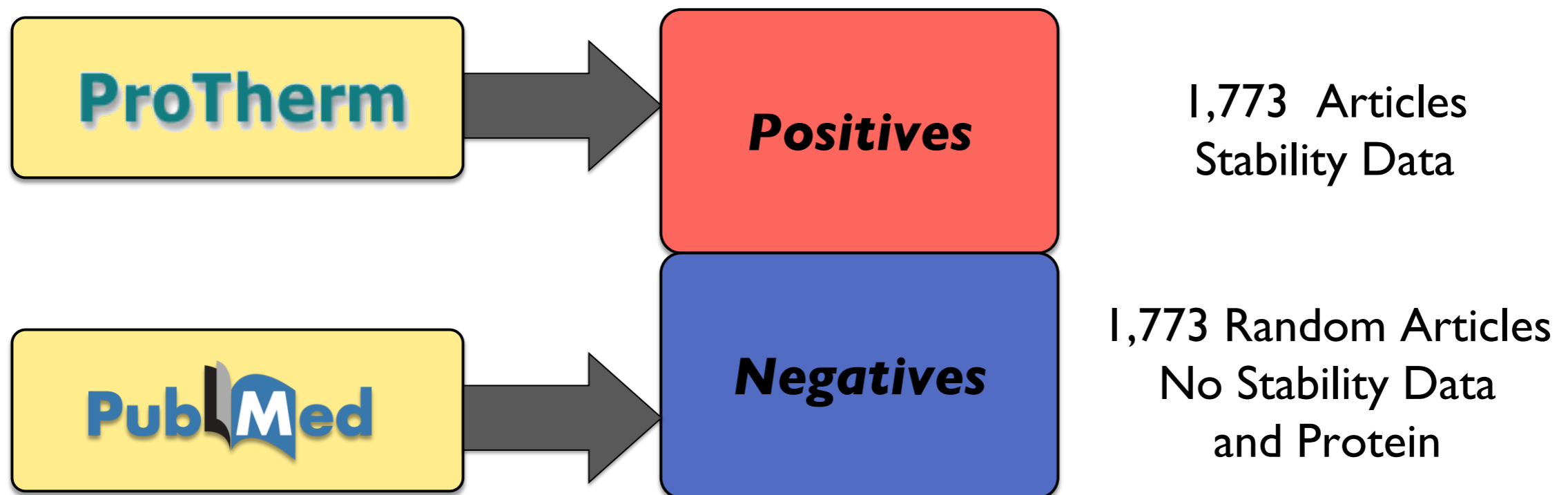
Center for Information Biology and DNA Data Bank of Japan
National Institute of Genetics

Open challenges

- **Automatic retrieval** of the data from literature
- **Integration of different sources of data** in a dedicated repository
- **Benchmarking** of new computational methods
- Understanding the **relationships between protein stability, function and disease**

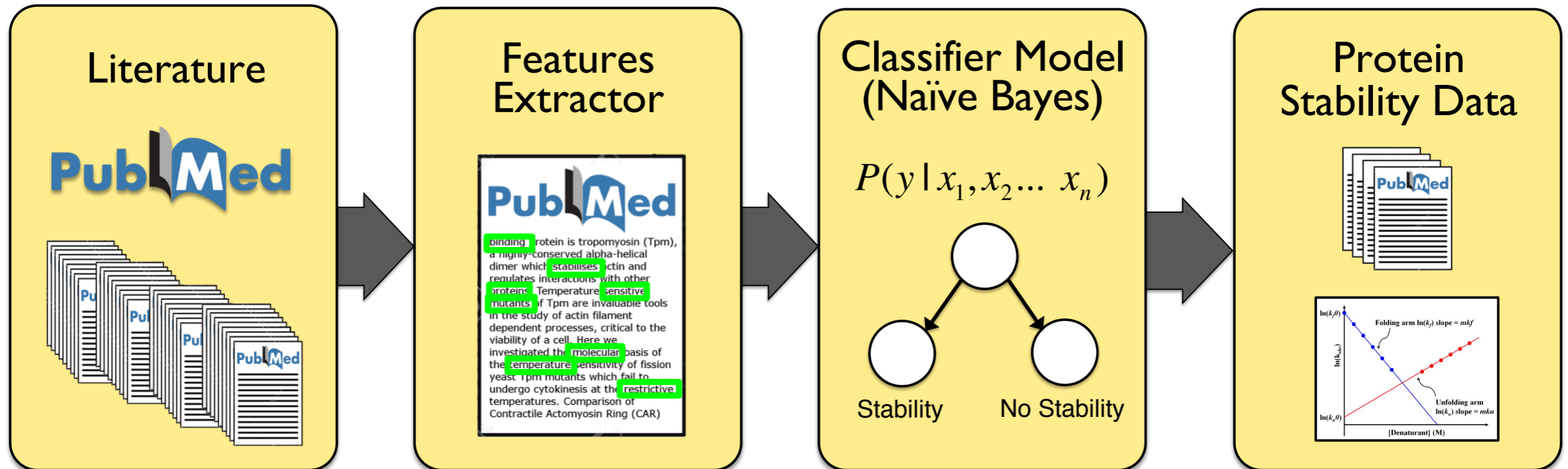
Literature dataset

We collected a **set of articles from ProTherm database** and use them for training a machine learning method. Negative set are selected from PubMed.



Data manipulation

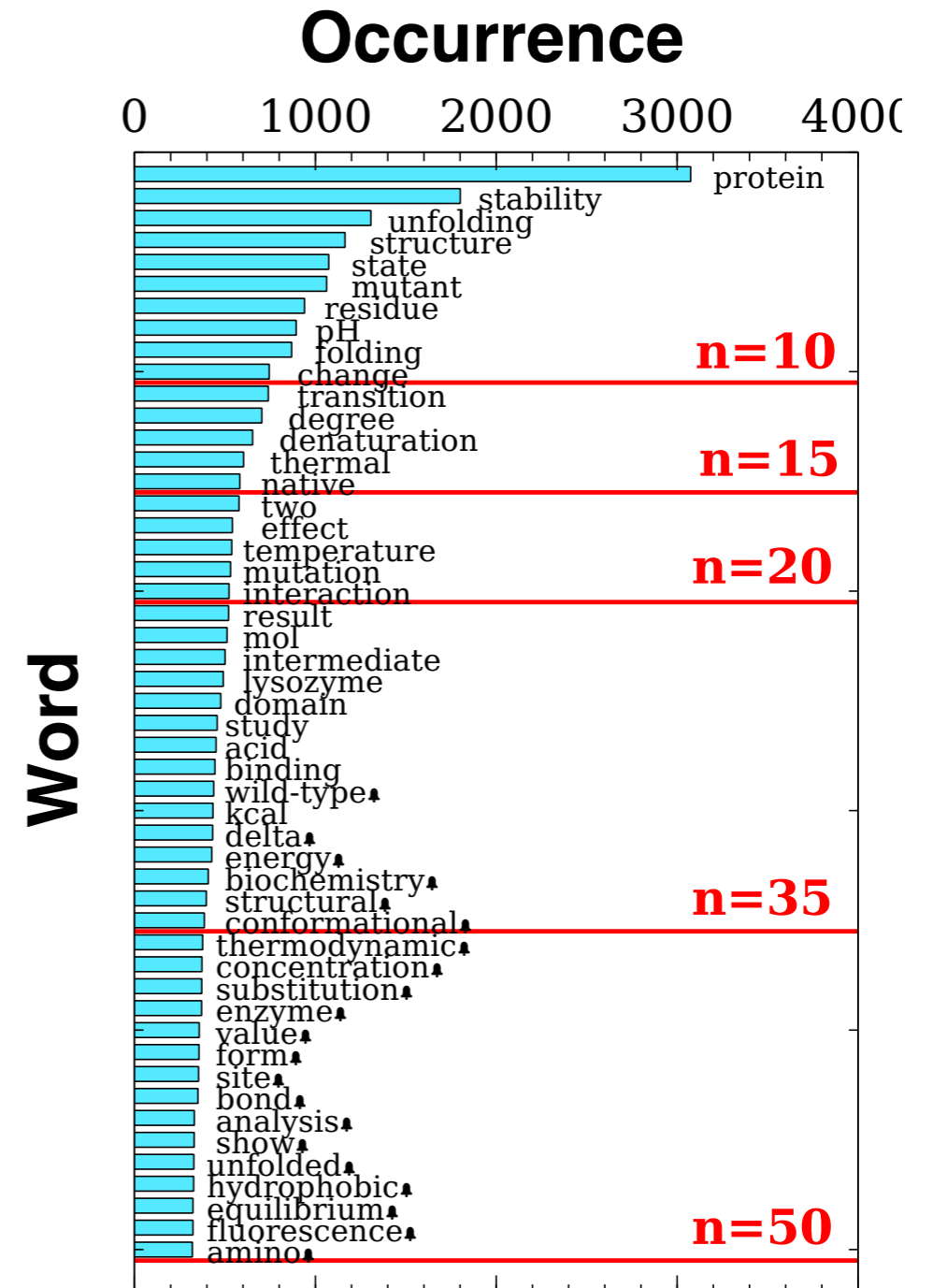
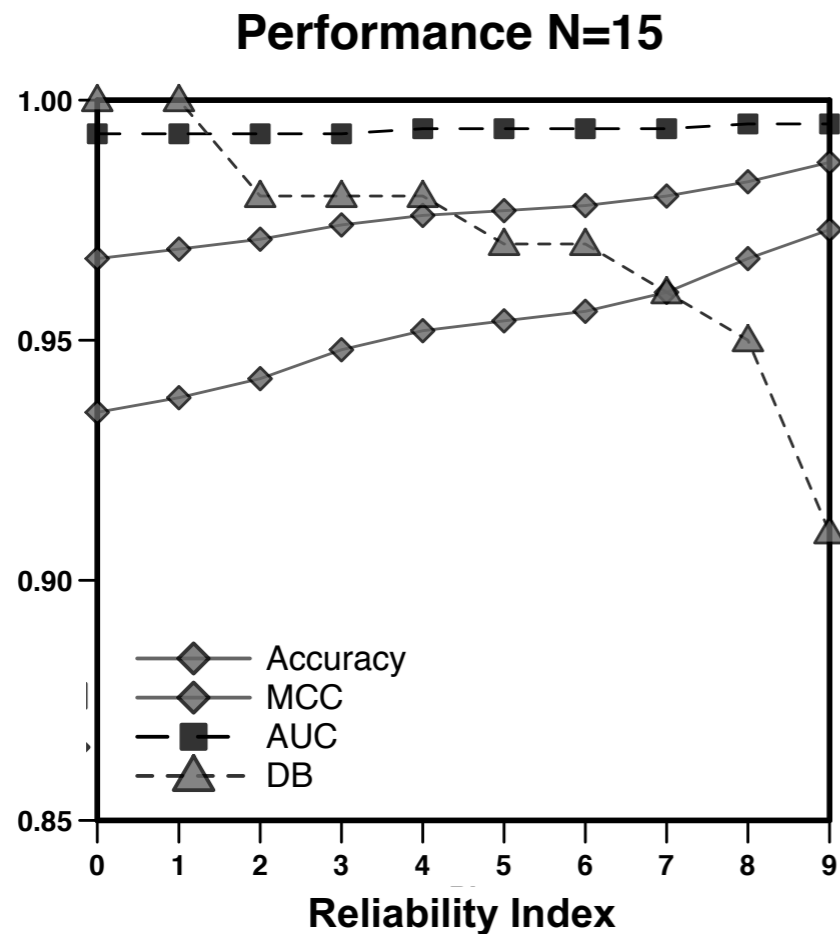
We extracted the words from the **full text version of the manuscript** and build a Naive-Bayes classifier based on the occurrence of the different words.



Method performance

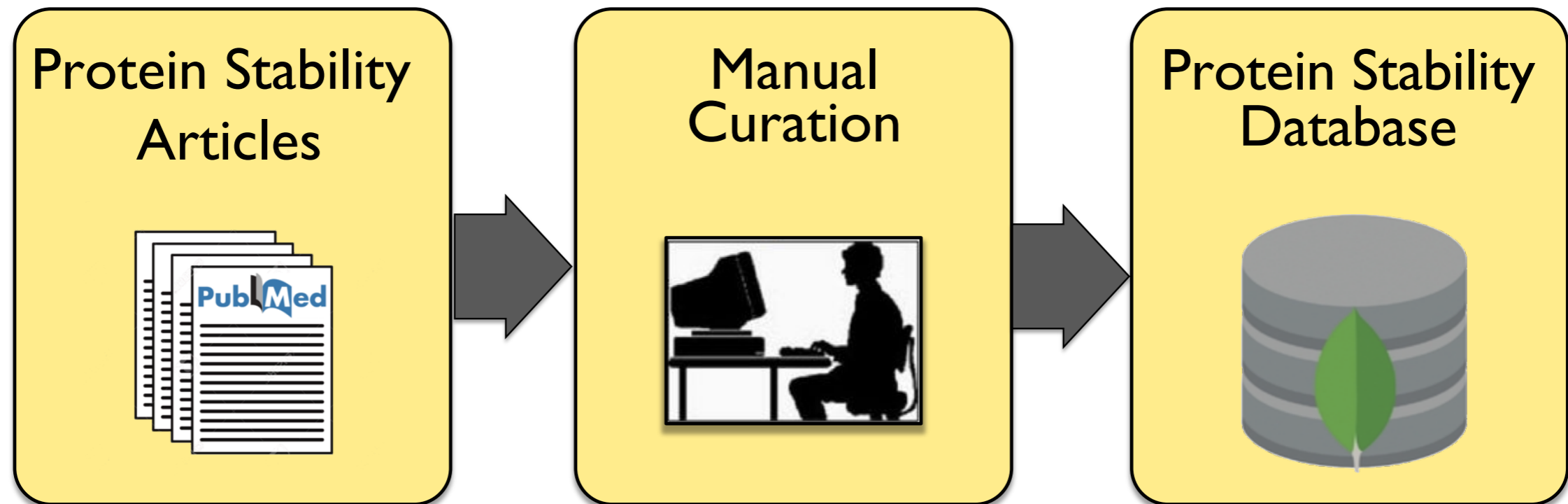
We performed different tests using different number of **words as prediction features**

Features	Accuracy	MCC	AUC
10	0.95	0.90	0.99
15	0.97	0.94	0.99
20	0.97	0.94	0.99
35	0.97	0.94	1.00
50	0.97	0.94	1.00



Data retrieval

The implementation of the previous machine learning method will be useful to **simplify the manual curation process** preselecting a set of manuscripts with high probability of including stability data.



Meta prediction

One approach that we previously tested for the prediction of functionally deleterious variants is the **development of a meta prediction approach**.

- We can **integrate different prediction methods** for protein stability change
- We can train methods on **different type of variant datasets**

The **final goal** consists in selecting highly-reliable predictions

CAGI experiments

The Critical Assessment of Genome Interpretation is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation.

Hi emidio, welcome back. [Your account](#) [Sign out](#)

CAGI

[Home](#) [Data Use Agreement](#) [FAQ](#) [Organizers](#) [Contact](#) [CAGI 4](#) [Previous CAGIs](#)

CAGI 4

- [Overview](#)
- [CAGI Presentations](#)
- [Challenges](#)
 - [Bipolar exomes](#)
 - [Crohn's exomes](#)
 - [eQTL causal SNPs](#)
 - [Hopkins clinical panel](#)
 - [NAGLU](#)
 - [NPM-ALK](#)
 - [PGP](#)
 - [Pyruvate kinase](#)
 - [SickKids clinical genomes](#)
 - [SUMO ligase](#)
 - [Warfarin exomes](#)
- [Conference](#)

Welcome to the CAGI experiment!

The CAGI 4 Conference

The Fourth Critical Assessment of Genome Interpretation (CAGI 4) prediction season has closed. Eleven challenges were released beginning on 3 August 2015, and the final challenge closed on 1 February 2016. Independent assessment of the predictions has been completed.

The CAGI 4 Conference was held 25-27 March 2016 in Genentech Hall on the UCSF Mission Bay campus in San Francisco, California. Conference presentations (remixable slides and video) are provided on the [CAGI 4 conference program page](#) and also on each challenge page.

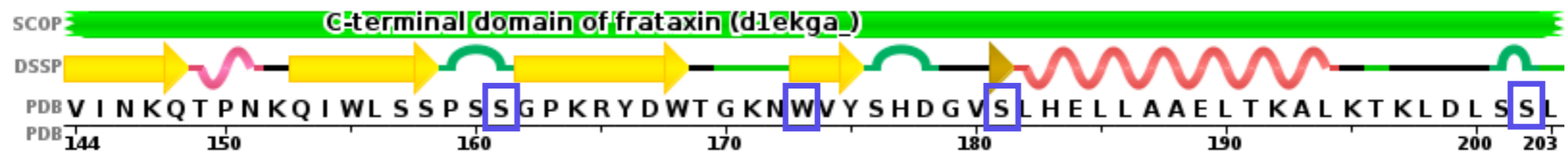
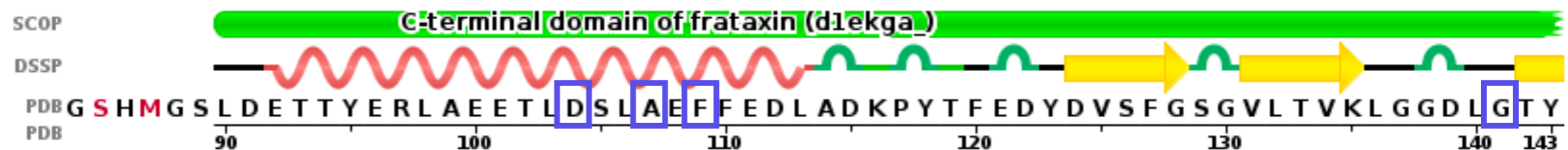
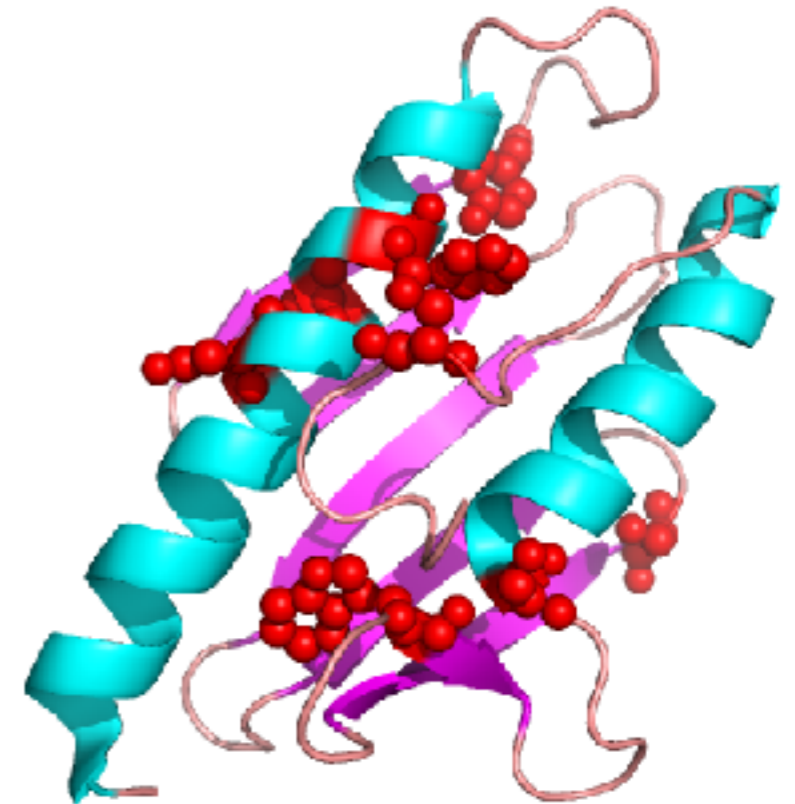
Please distribute this information widely and follow our Twitter feed @CAGInews and the web site for updates. For more information on the CAGI experiment, see the [Overview](#).

CAGI Lead Scientist or Postdoctoral Researcher position open!

Take the lead of the CAGI experiment! We are searching for a CAGI Lead Scientist or Postdoctoral Researcher to join us in early 2016. Roger Hoskins will lead the CAGI 4 experiment to its completion, but he is unable to continue in the role beyond mid-2016. He will overlap with the new CAGI leader to ensure a seamless transition. Job descriptions posted at <http://compbio.berkeley.edu/jobs>

Frataxin challenge at CAGI

- Participants were asked to submit **predictions of the variation of the unfolding free energy change** upon mutation at concentration 0 of denaturant ($\Delta\Delta G_{H20}$).
- Data providers at the University of Roma experimentally determined the **unfolding free energy of the wild-type and 8 mutants using CD and Fluorescence**. The average value between the two $\Delta\Delta G_{H20}$ is considered for the assessment

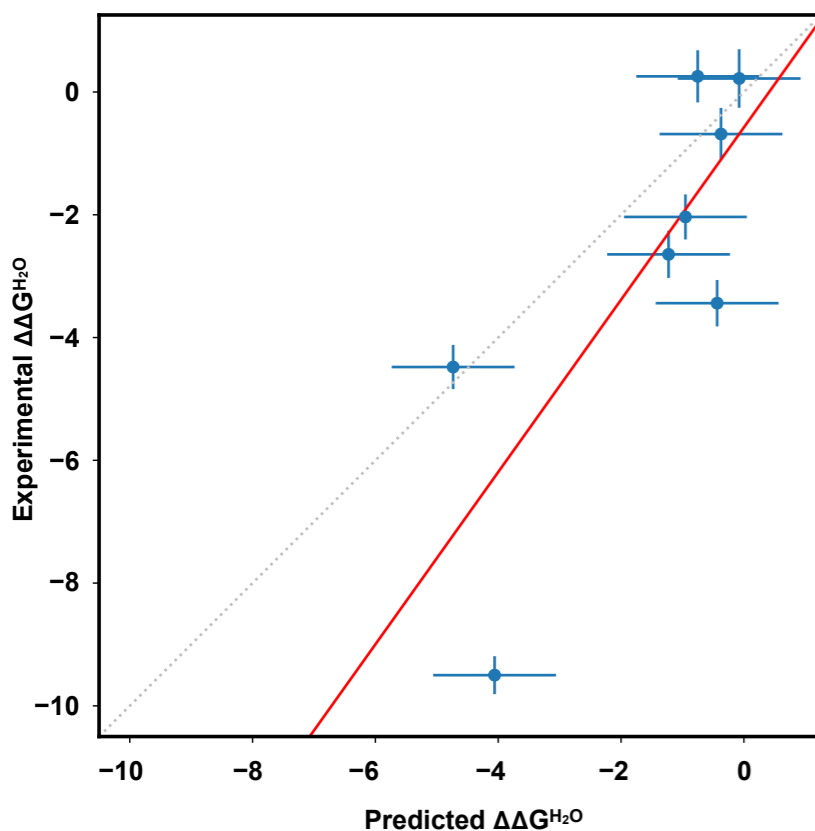


Fold-X and I-Mutant

The performance of **Fold-X** is comparable with the negative of predictions from **Group 6**

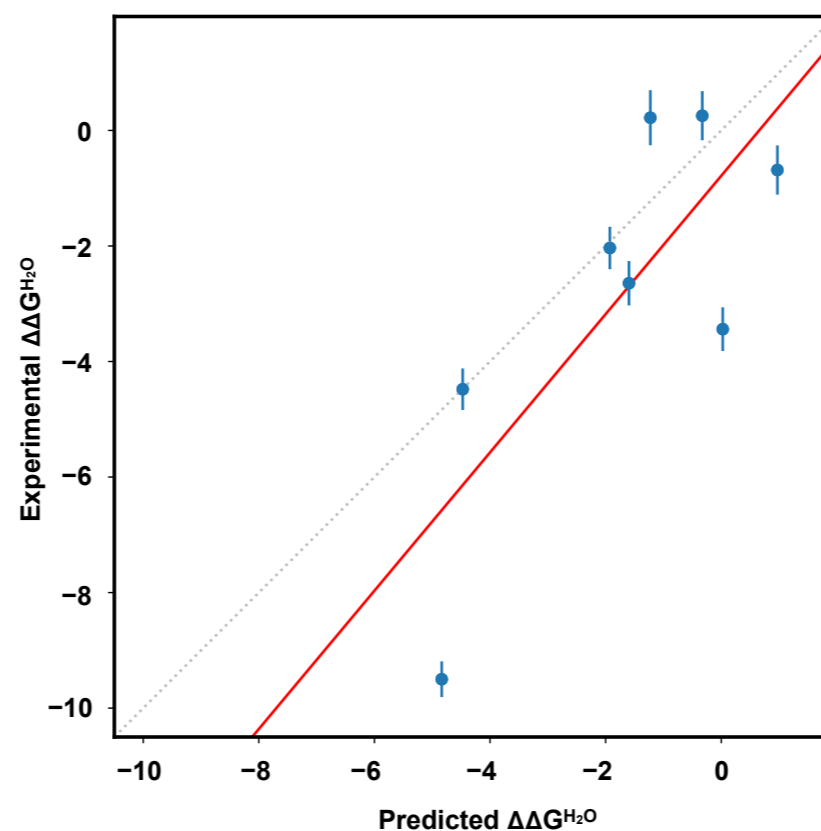
Reverse Group_6

PCC=0.78 - SPC=0.71 - KTC=0.57
RMSE=2.32 - MAE=1.60



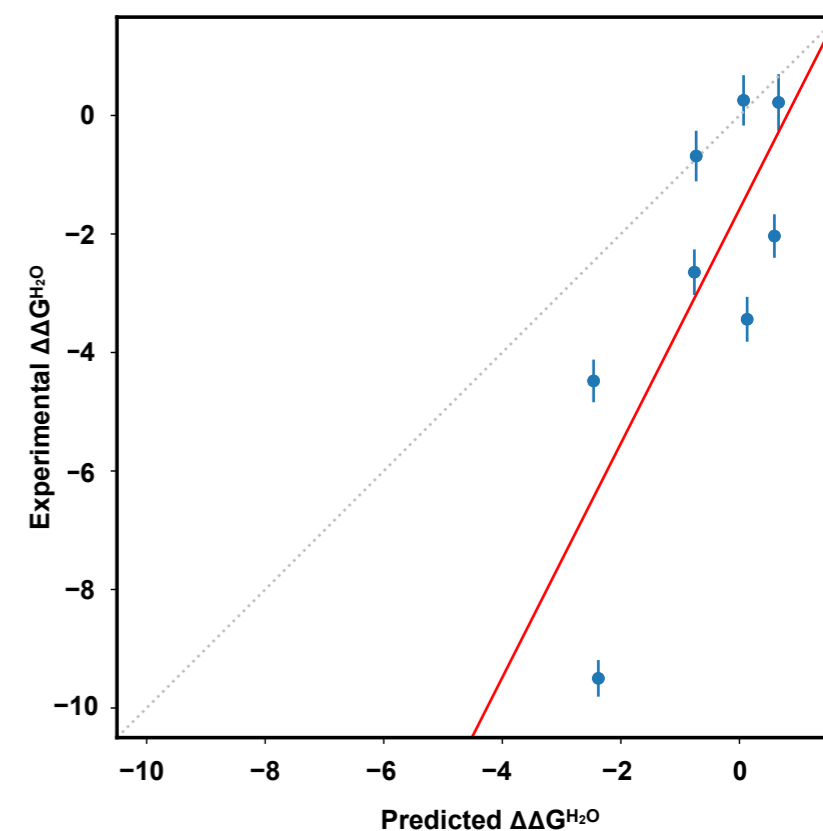
Fold-X

PCC=0.77 - SPC=0.62 - KTC=0.50
RMSE=2.24 - MAE=1.62



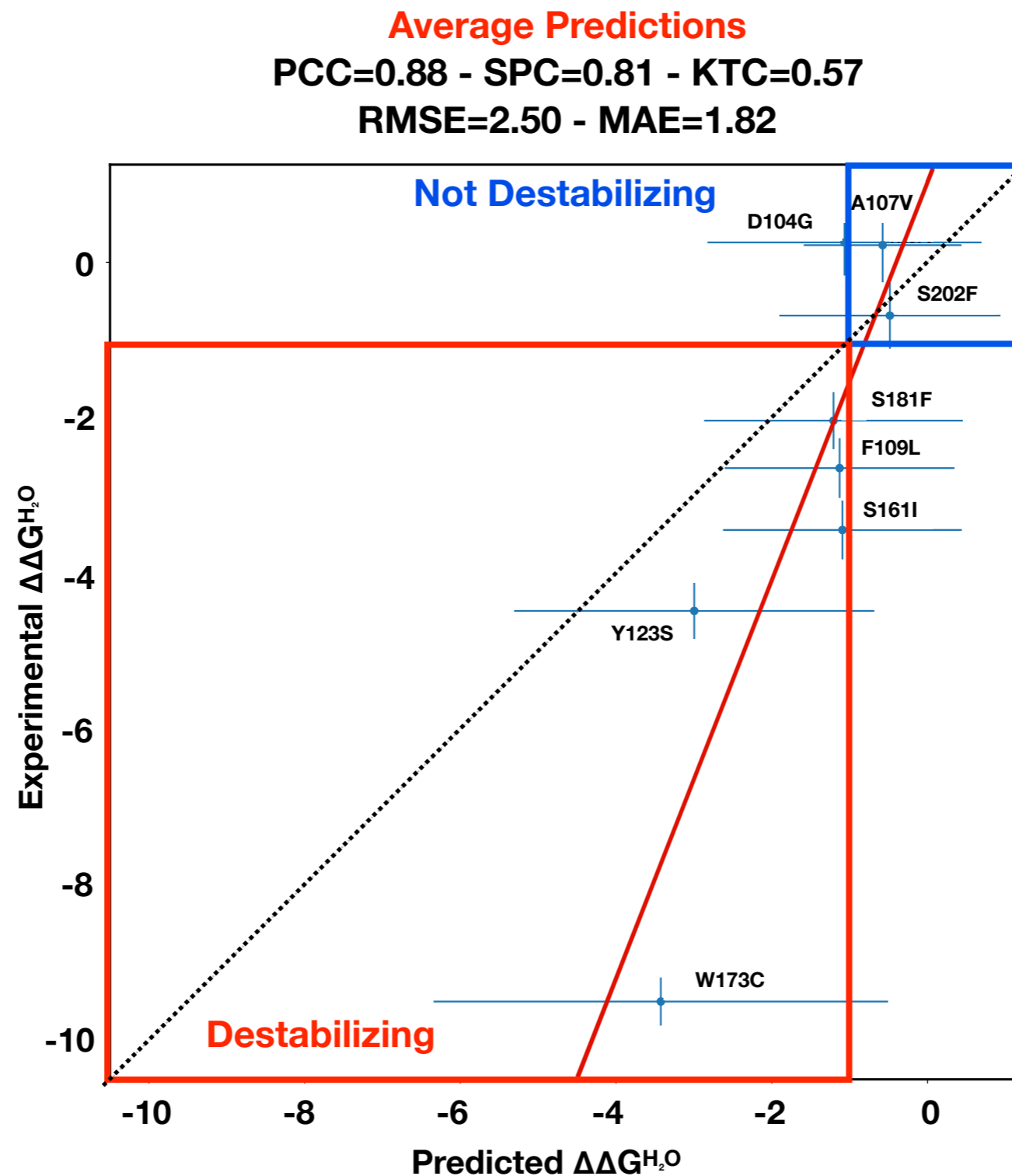
I-Mutant

PCC=0.76 - SPC=0.64 - KTC=0.50
RMSE=3.13 - MAE=2.24



Averaged predictions

For **W173C** we found the highest difference between predicted and experimental $\Delta\Delta G$ s



Acknowledgments

Structural Genomics @CNAG

Marc A. Marti-Renom
Francois Serra

Computational Biology and Bioinformatics Research Group (UIB)

Jairo Rocha

Division of Informatics at UAB

Malay Basu
Division Clinical Immunology
& Rheumatology
Harry Schroeder
Mohamed Khass

Helix Group (Stanford University)

Russ B. Altman
Jennifer Lahti
Tianyun Liu
Grace Tang

Bologna Biocomputing Group

Rita Casadio
Pier Luigi Martelli
Paola Turina

University of Torino

Piero Fariselli

University of Camerino

Mario Compiani

Mathematical Modeling of Biological Systems (University of Düsseldorf)

Markus Kollmann
Linlin Zhao

Other Collaborations

Yana Bromberg, Rutgers University, NJ
Hannah Carter, UCSD, CA
Francisco Melo, Universidad Catolica, Chile
Sean Mooney, Buck Institute, Novato
Cedric Notredame, CRG Barcelona
Gustavo Parisi, Universidad de Quilmes
Frederic Rousseau, KU Leuven
Joost Schymkowitz, KU Leuven

FUNDING

Italian MIUR: PRIN 2017
NIH: 1R21 AI134027- 01A1
Italian MIUR: FFABR 2017
UNIBO: International Cooperation
Startup funding Dept. of Pathology UAB
NIH:3R00HL111322-04S1 Co-Investigator
EMBO Short Term Fellowship
Marie Curie International Outgoing Grant
Marco Polo Research Project
BIOSAPIENS Network of Excellence
SPINNER Consortium

Biomolecules, Folding and Disease



<http://biofold.org/>