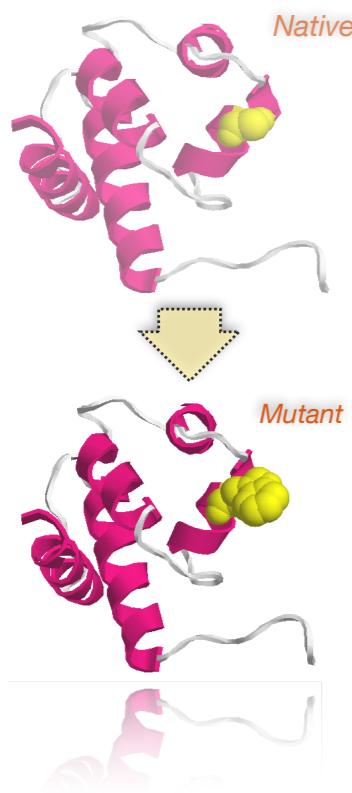
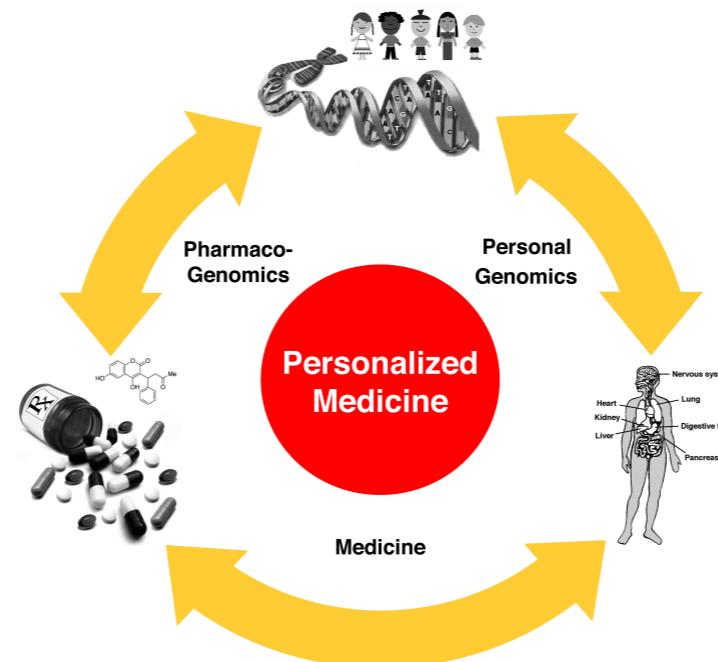


# Methods for the automatic annotation of protein variants



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

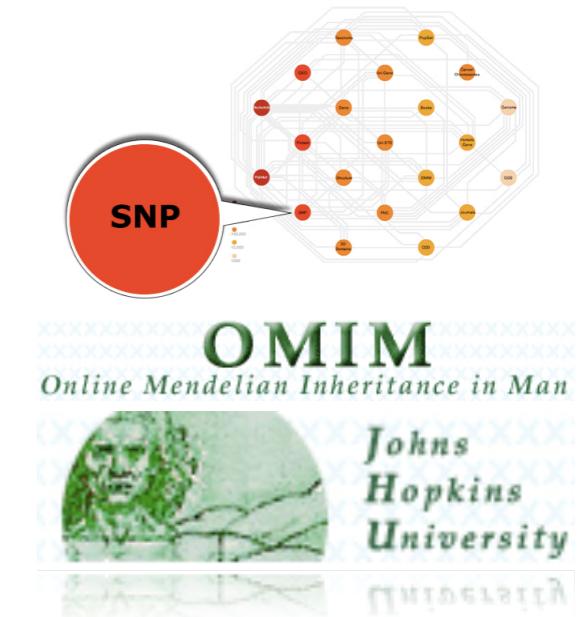


Emidio Capriotti

<http://biofold.org/>



Biomolecules  
Folding and  
Disease



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

## Biomolecules Folding and Disease



Home   Research   People   **Resources**   Publications   Training

### Resources

The researchers of the BioFold 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.  


**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.  


**K-Fold** Support Vector Machine based method to predict the mechanism and rate of protein folding kinetic.  


#### Genomic Variations and Disease

**ContrastRank** Statistical method for the classification of cancer samples using exome sequencing data.  

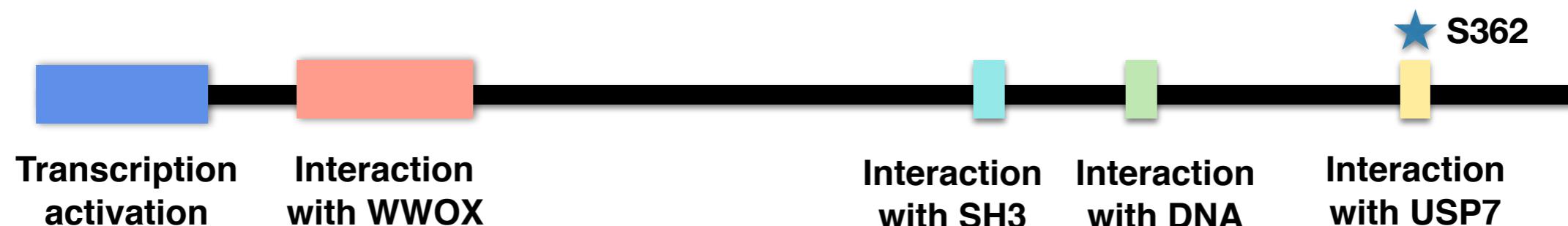

**DrCancer** Support Vector Machine based method to predict cancer-causing mutations (Beta version).  


**Fido-SNP** Machine learning method for predicting the impact of SNVs in the dog genome.

# Sequence, Structure & Function

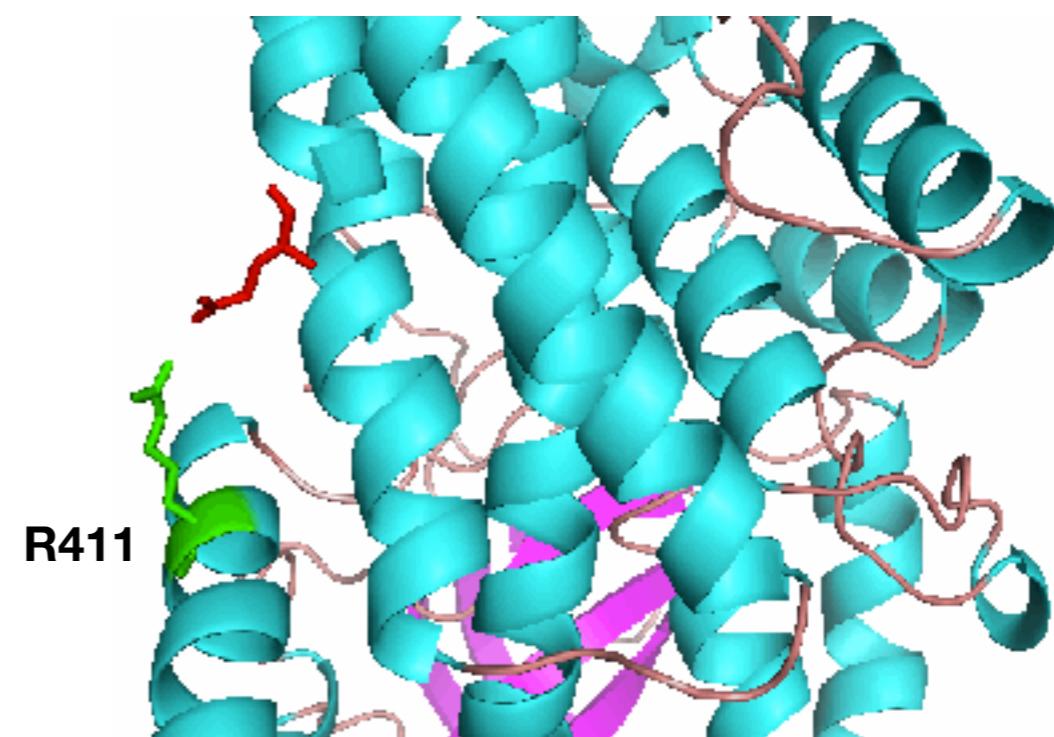
Genomic variants in sequence motifs could affect protein function.

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.



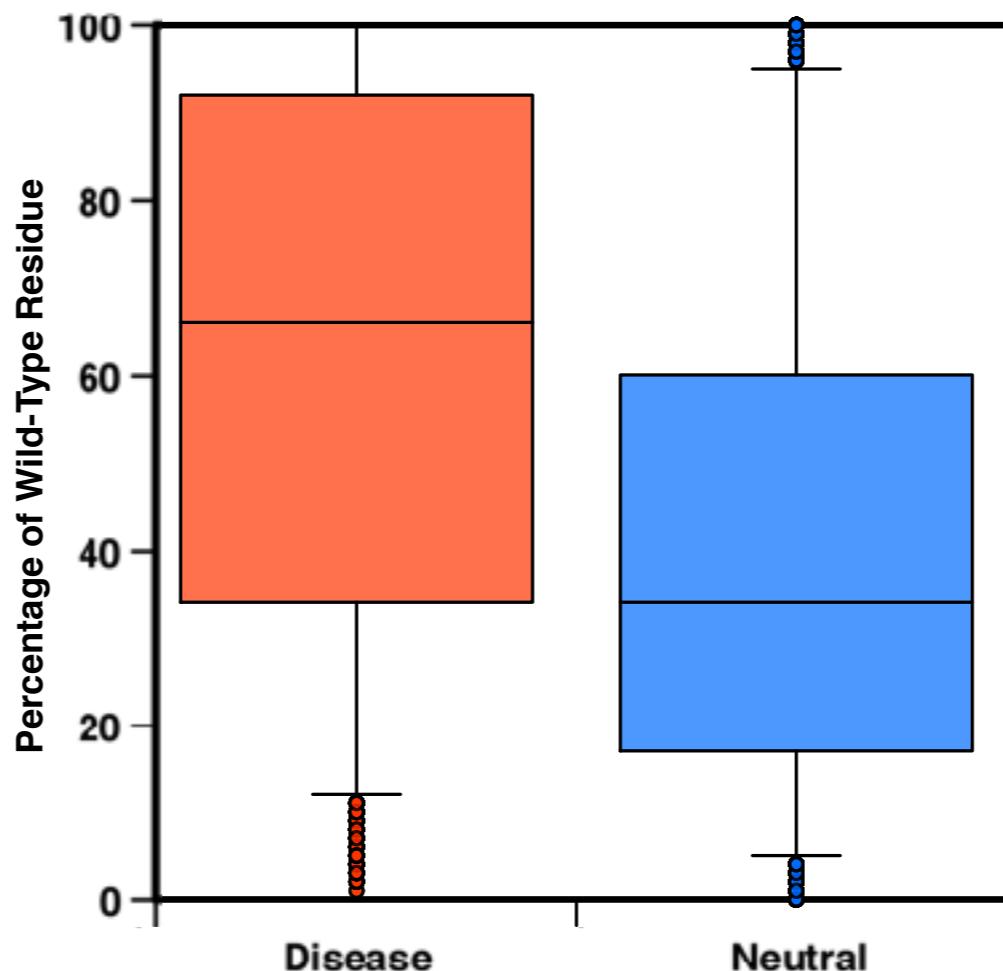
# Conserved or not?

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

# Sequence profile

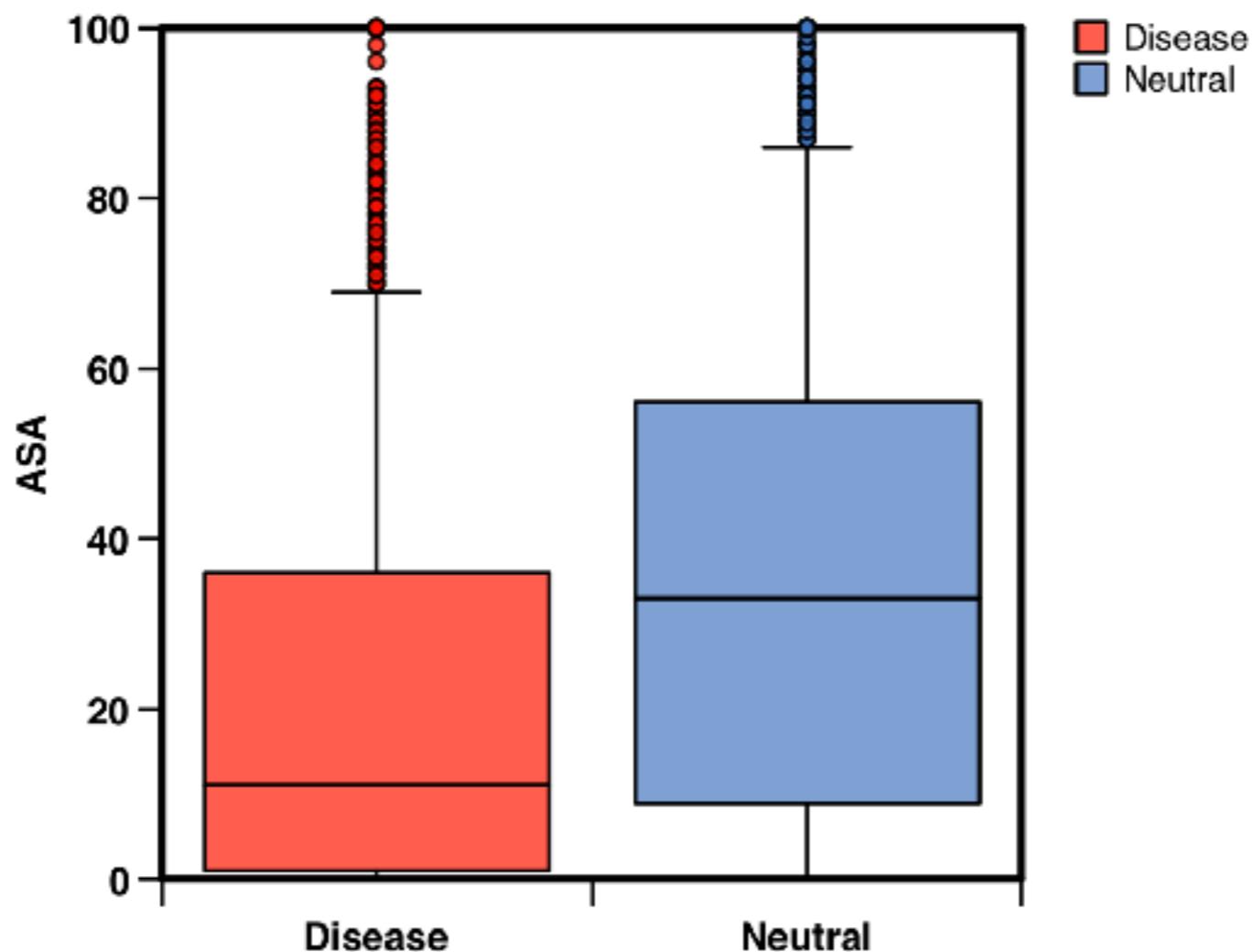
The protein **sequence profile** is calculated running **BLAST** on the **UniRef90** dataset and selecting only the hits with  $e\text{-value} < 10^{-9}$ .

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



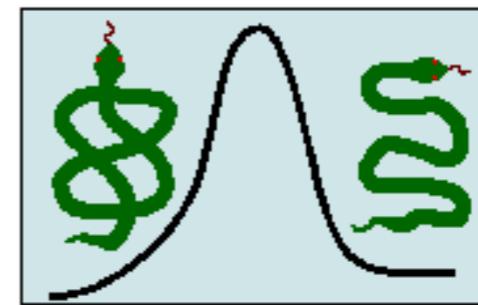
# Structure environment

There is a **significant difference** (KS p-value =  $2.8 \times 10^{-71}$ ) 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



**ProTherm**

Thermodynamic Database  
for Proteins and Mutants



**Protein Mutant Database**

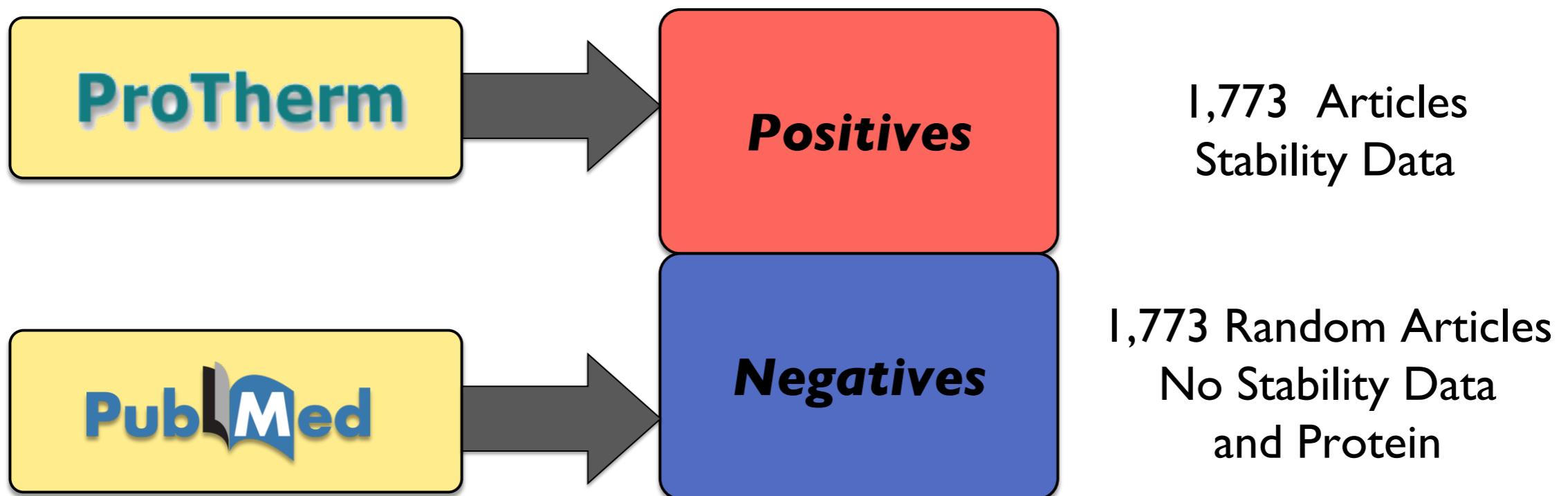
Center for Information Biology and DNA Data Bank of Japan  
National Institute of Genetics

# 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

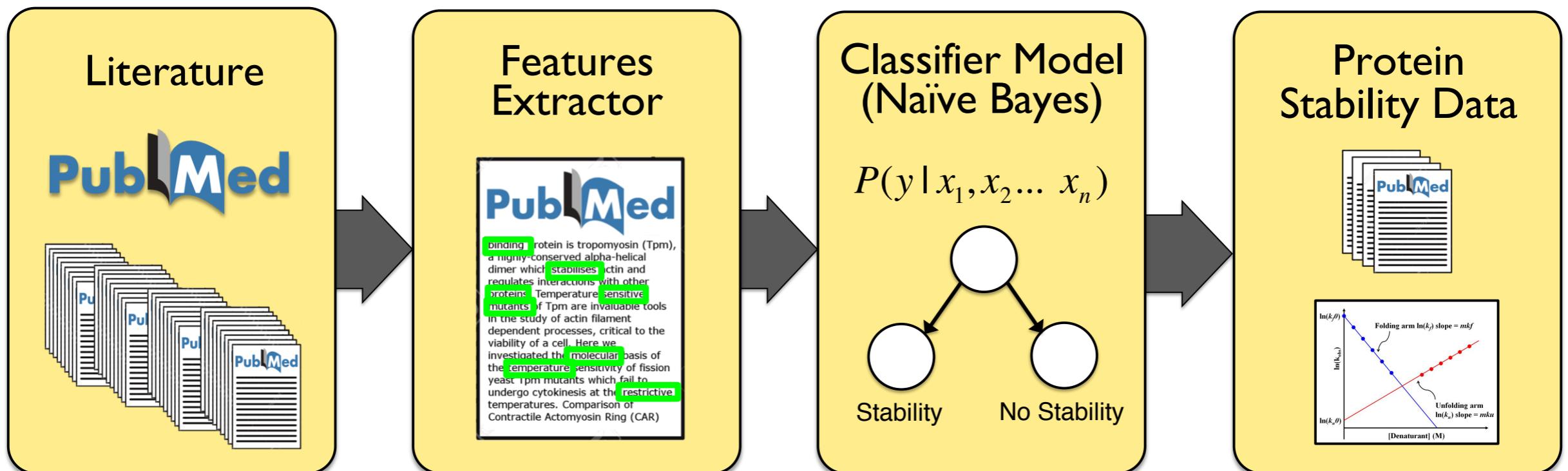
# Literature dataset

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

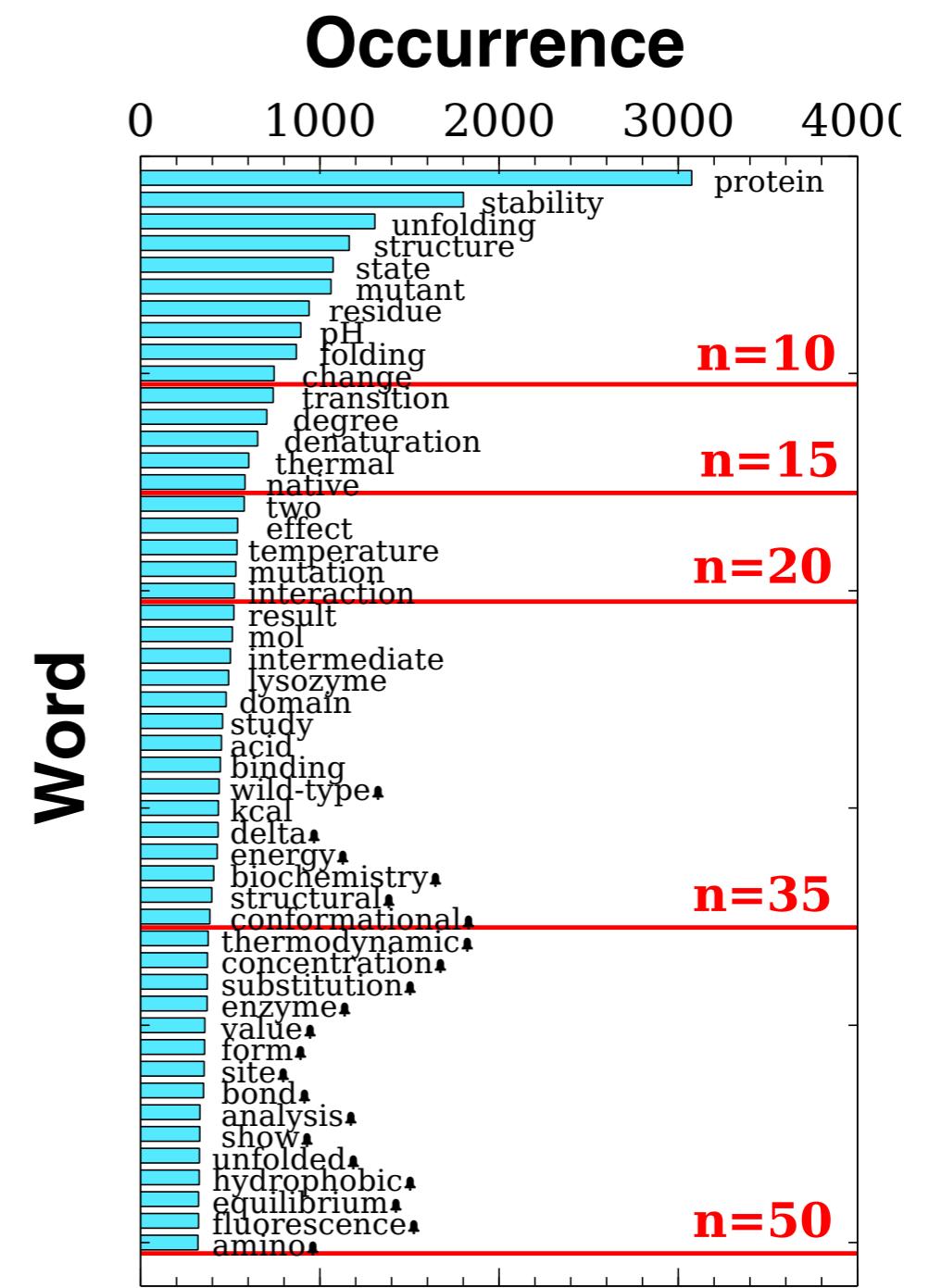
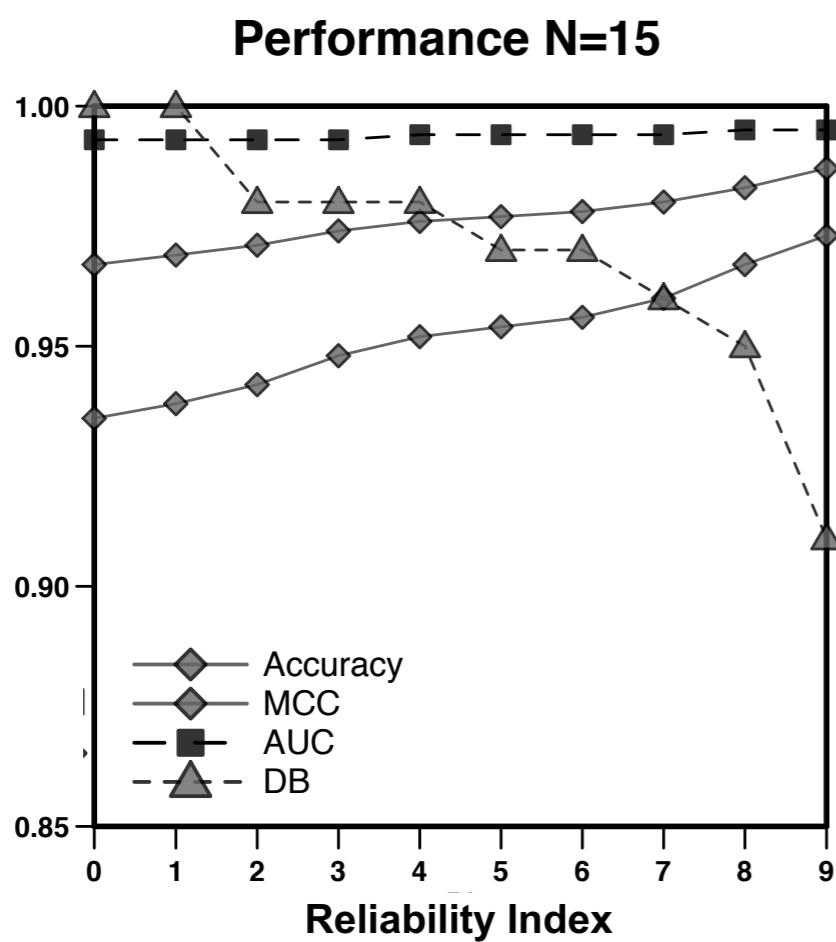
We extracted the words from the **full text version of the manuscript** and build a Naive-Bayes classifier based on the occurrence of the different words.



# Method performance

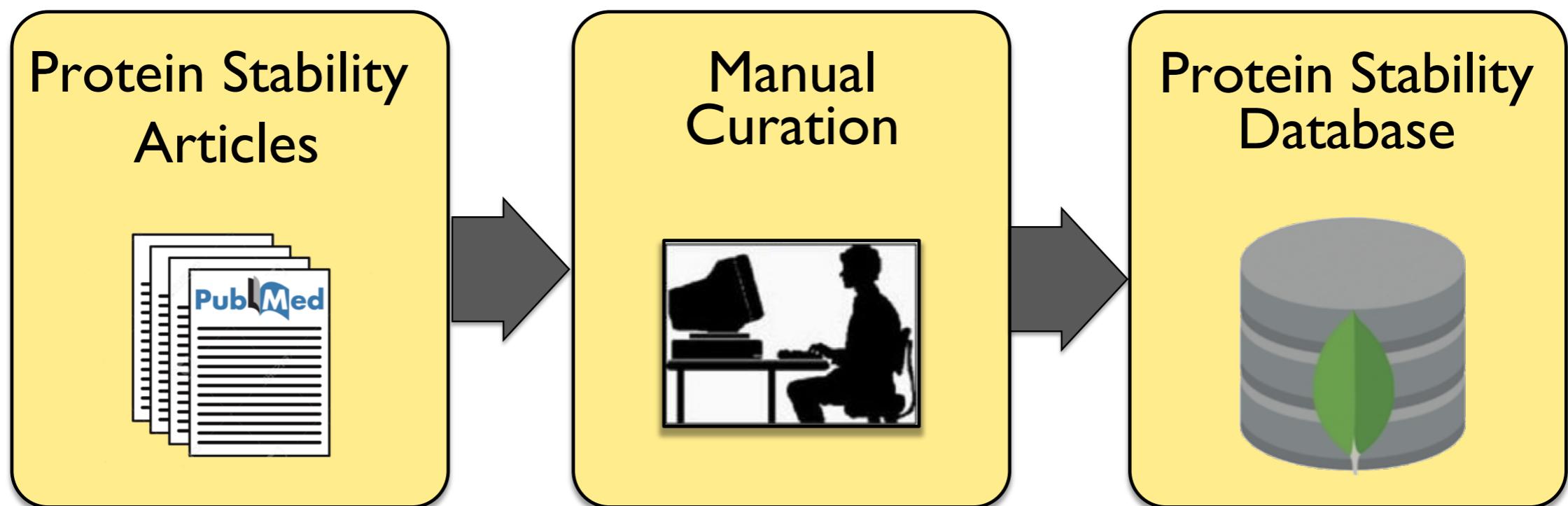
We performed different tests using different number of words as prediction features

| Features  | Accuracy    | MCC         | AUC         |
|-----------|-------------|-------------|-------------|
| 10        | 0.95        | 0.90        | 0.99        |
| <b>15</b> | <b>0.97</b> | <b>0.94</b> | <b>0.99</b> |
| 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.

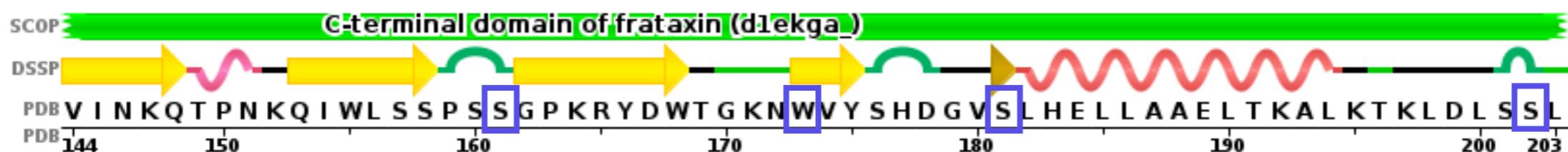
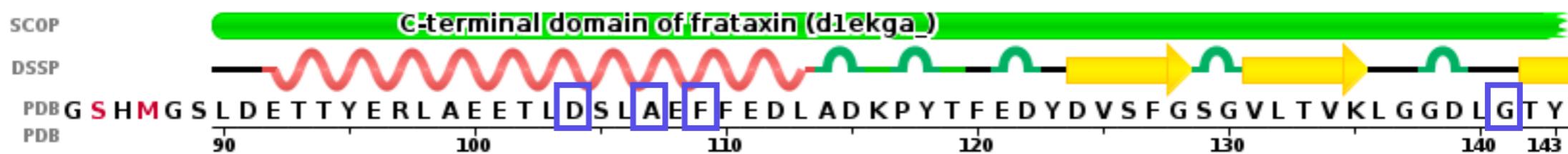
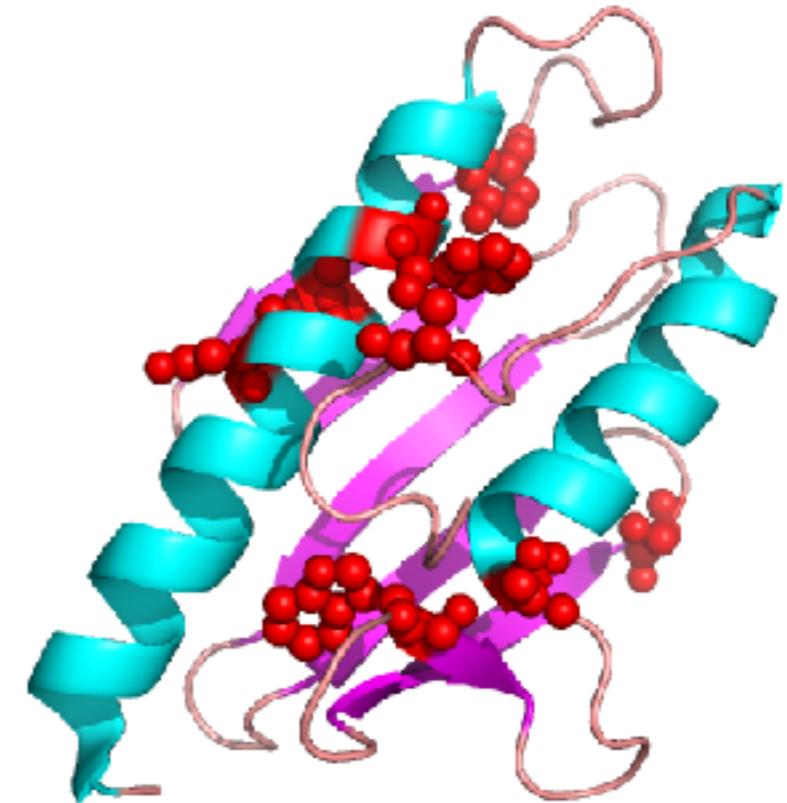
Please distribute this information widely and follow our Twitter feed @CAGInews and the web site for updates. For more information on the CAGI experiment, see the [Overview](#).

**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 ( $\Delta\Delta G_{H_2O}$ ).
- 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  $\Delta\Delta G_{H_2O}$  is considered for the assessment

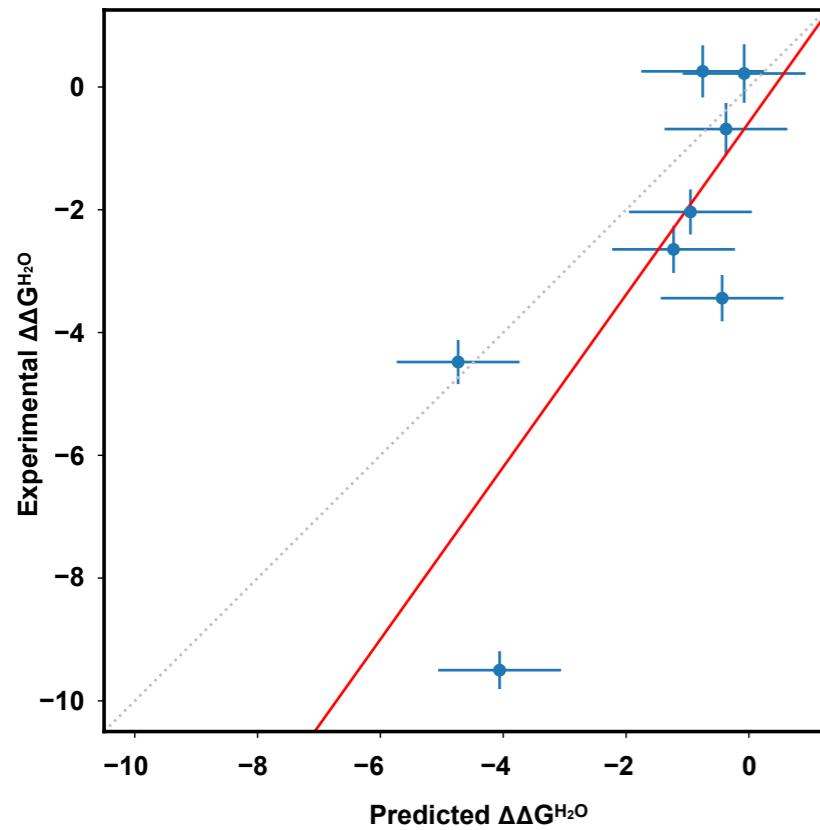


# Fold-X and I-Mutant

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

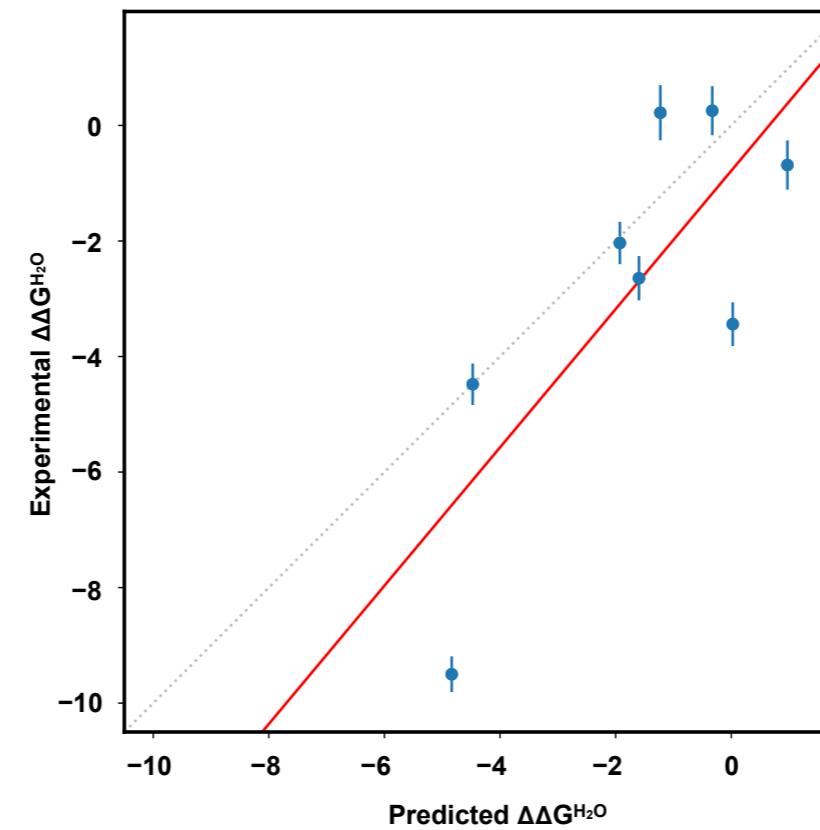
Reverse Group\_6

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



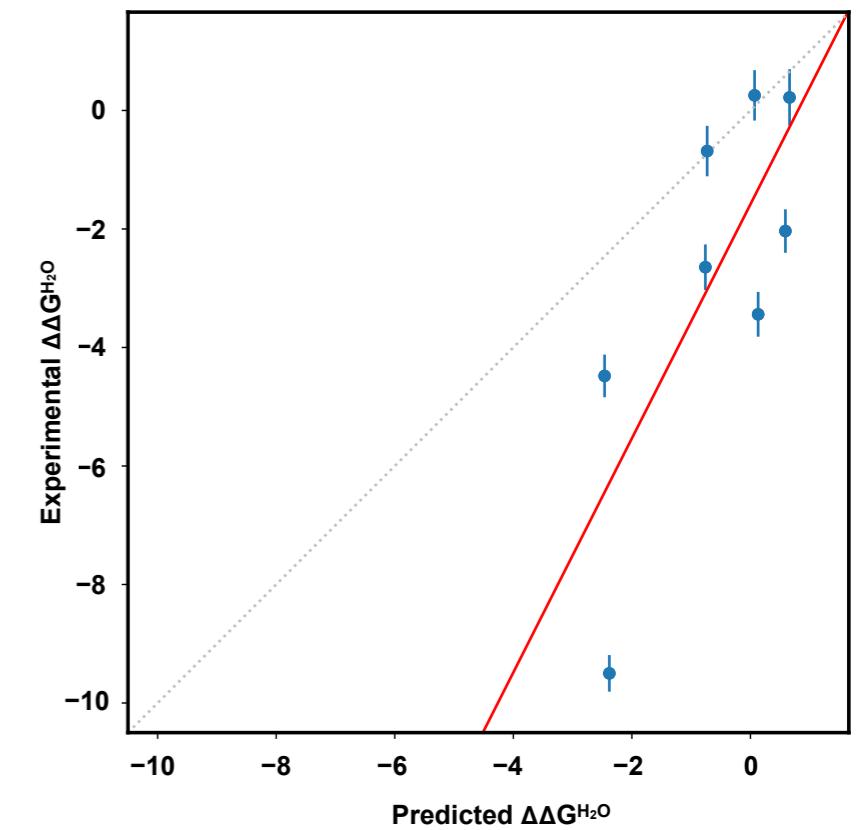
Fold-X

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



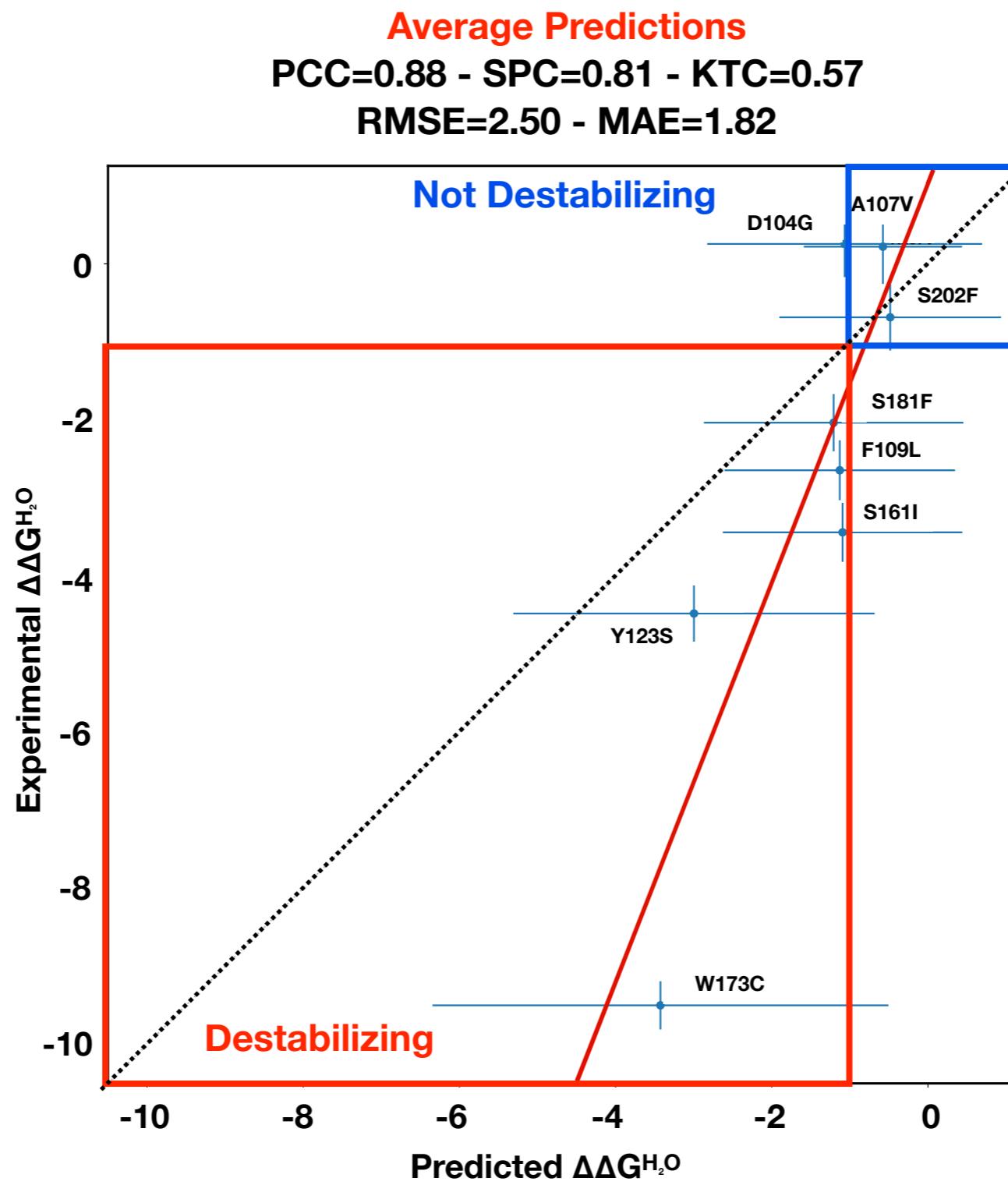
I-Mutant

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, Rutger University, NJ  
Hannah Carter, UCSD, CA  
Francisco Melo, Universidad Católica, 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/>

