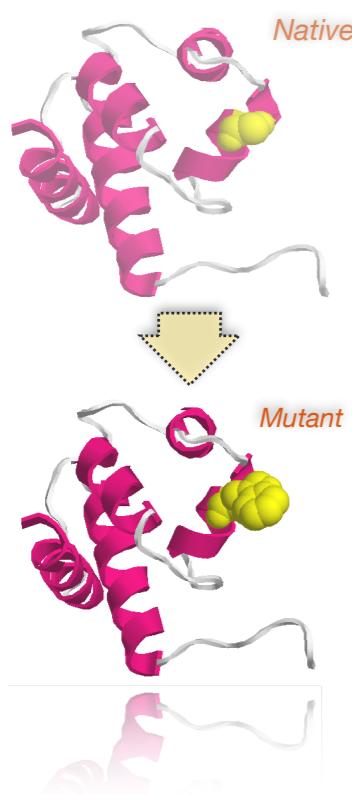
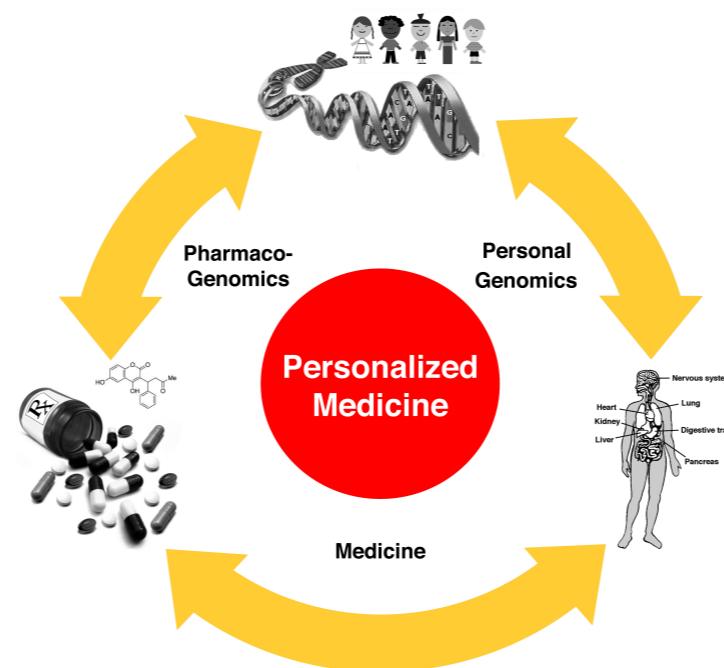


Predicting the impact of single point mutations



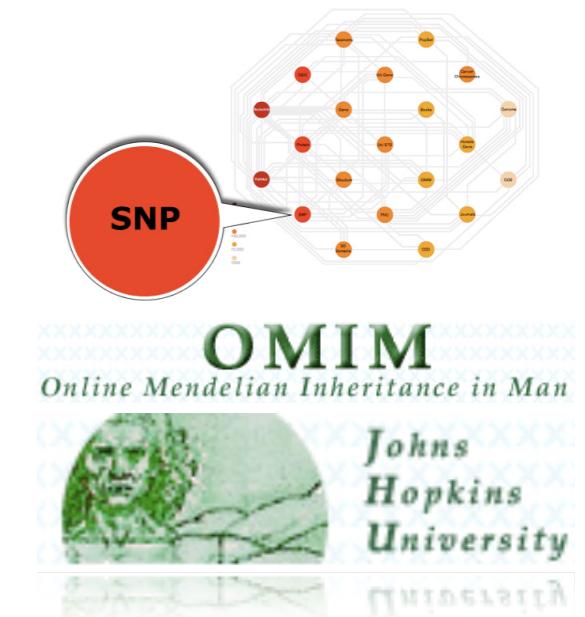
Università “La Sapienza”, Roma (Italy)
June 9, 2017



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



Biomolecules
Folding and
Disease



Department of Biological, Geological,
and Environmental Sciences (BiGeA)
University of Bologna



Research topics

- **Protein and RNA structure prediction:** developments of methods for the protein structure prediction by remote homology search. Implementation of new methods for RNA structure comparison, functional assignment and statistical-based potentials for model evaluation.
- **Protein Folding:** development of methods for the prediction of the protein folding kinetics rates and mechanisms using stochastic and machine learning approaches.
- **Predict the impact of genetic variations:** Machine learning methods for the prediction of the impact of single point mutations on protein stability and human health.

Single Nucleotide Variants

Single Nucleotide Variants (SNVs)

is a DNA sequence variation occurring when a single nucleotide A, T, C, or G in the genome differs between members of the species.

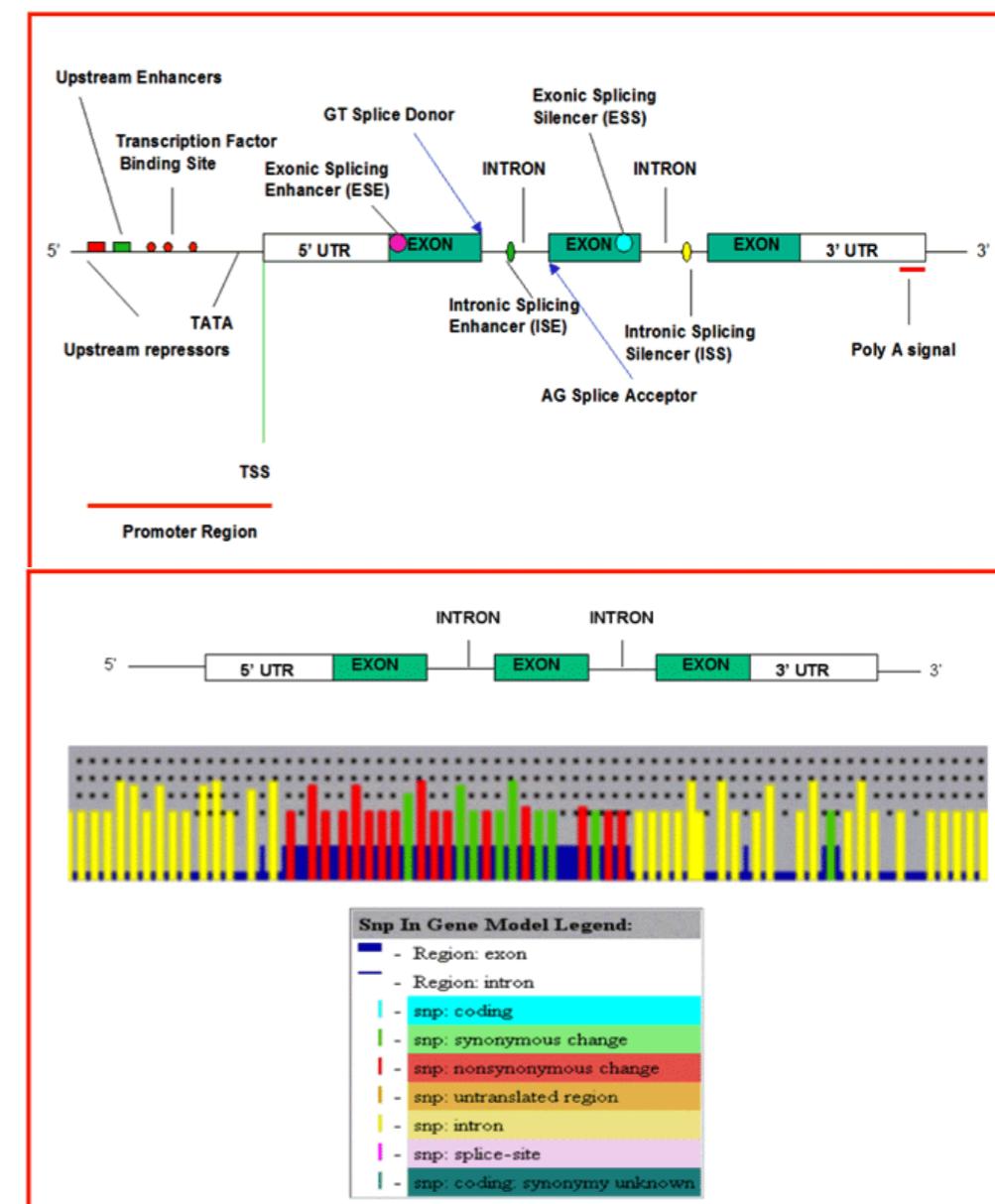
It is used to refer to Polymorphisms when the population frequency is $\geq 1\%$

SNVs occur at any position and can be classified on the base of their locations.

Coding SNVs can be subdivided into two groups:

Synonymous: when single base substitutions do not cause a change in the resultant amino acid

Non-synonymous or Single Amino Acid Variants (SAVs): when single base substitutions cause a change in the resultant amino acid.



Sequence, Structure & Function

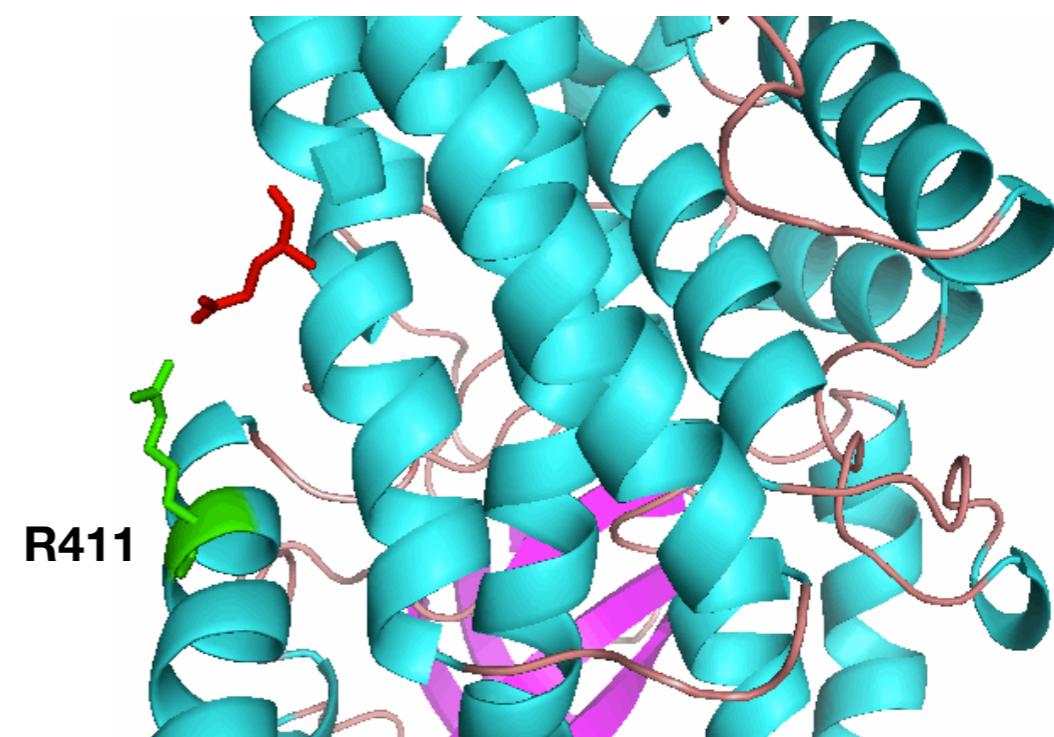
Genomic variants in sequence motifs could affect protein function.

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.

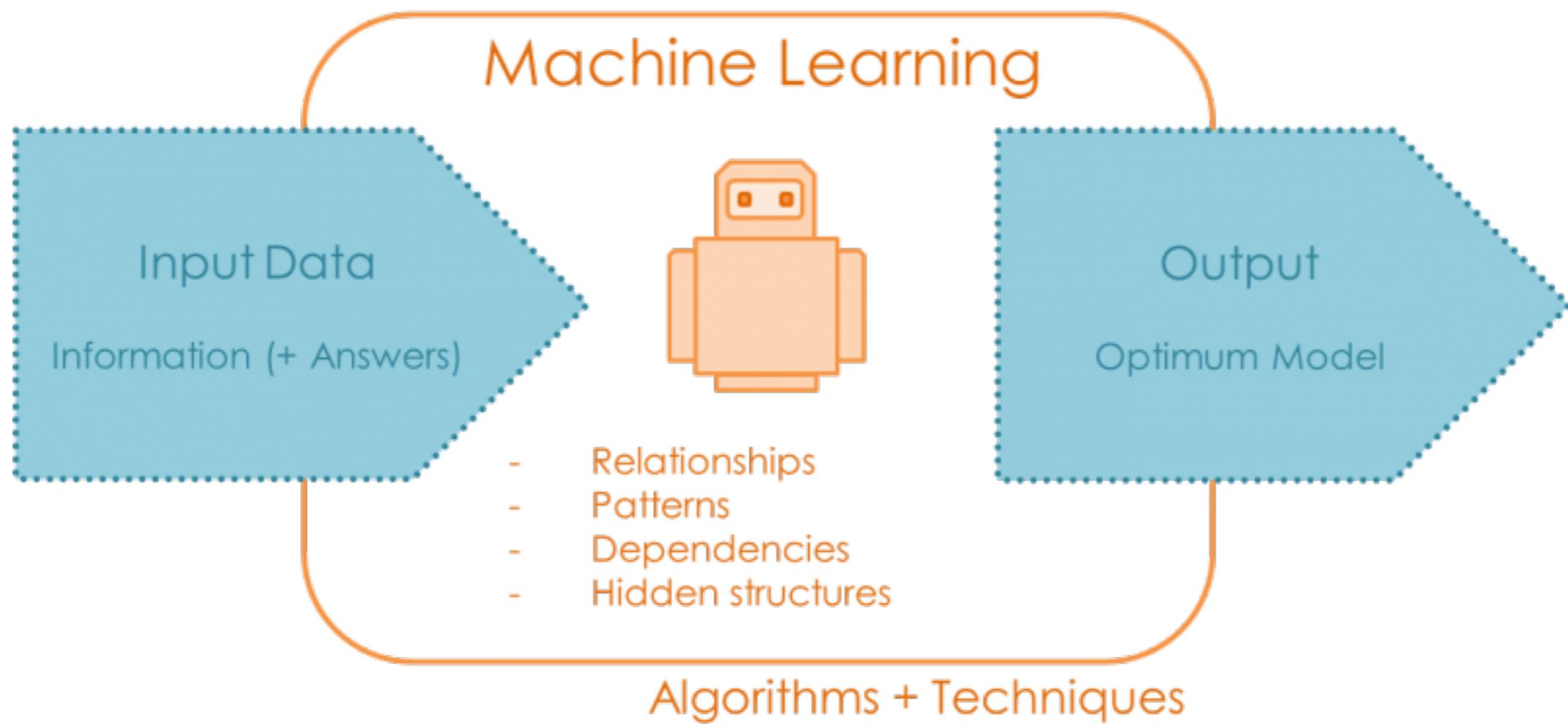


Machine learning

- Computational approach to build models based on the analysis of empirical data.
- Machine learning algorithms are suitable to address problems for which analytic solution does not exists and large amount of data are available.
- They are implemented selecting a representative set of data that are used in a training step and then validated on a test set with data “*not seen*” during the training.
- Most popular machine learning approaches are in computational biology are Neural Networks, Support Vector Machines and Random Forest.

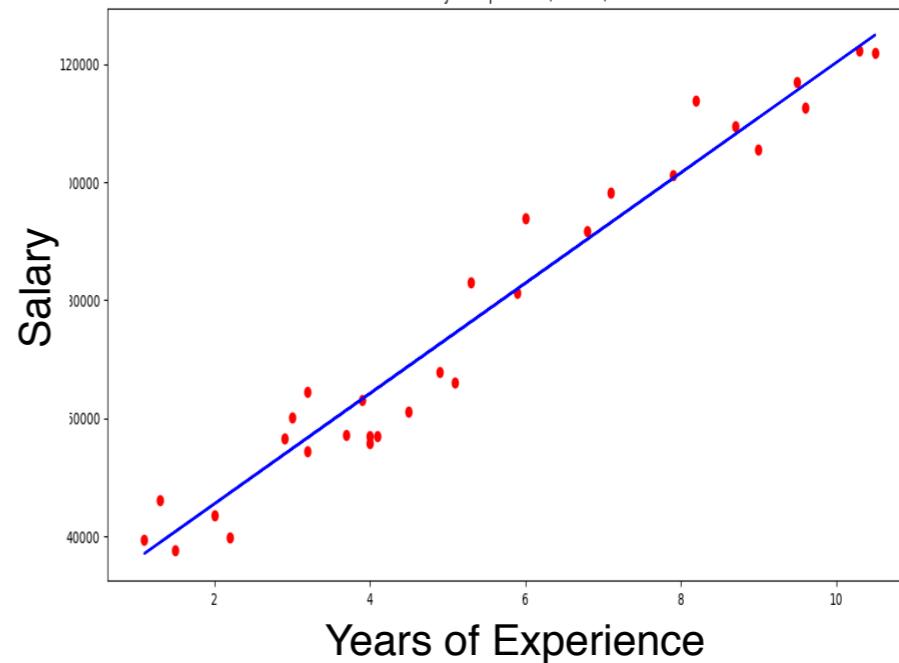
Input and Output

A machine learning algorithm takes in input a set of variables (features) and returns a numerical or discrete output

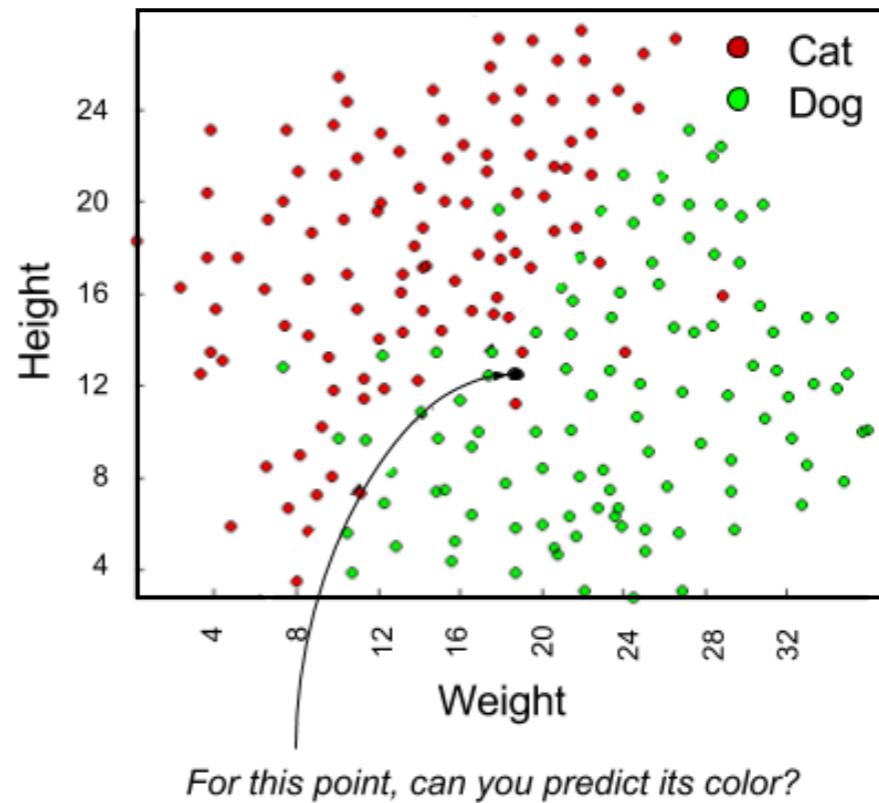


Types of Predictions

- Regression is used to predict continuous values.



- Classification is used to predict which class a data point is part of (discrete value).



Regression Evaluation

Compare predicted and real values using different correlation tests
and the Root Mean Square Error

Pearson Correlation

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Root Mean Square Error

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (P_i - O_i)^2}{n}}$$

Classification Evaluation

Overall Accuracy

$$Q2 = \frac{TP + TN}{TP + FN + TN + FP}$$

Sensitivity

$$S = \frac{TP}{TP + FN}$$

Precision

$$P = \frac{TP}{TP + FP}$$

Matthews Correlation

$$C = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

		Actual values	
		Positive	Negative
Predicted values	Positive	TP	FP
	Negative	FN	TN

ROC Curve

True Positive Rate

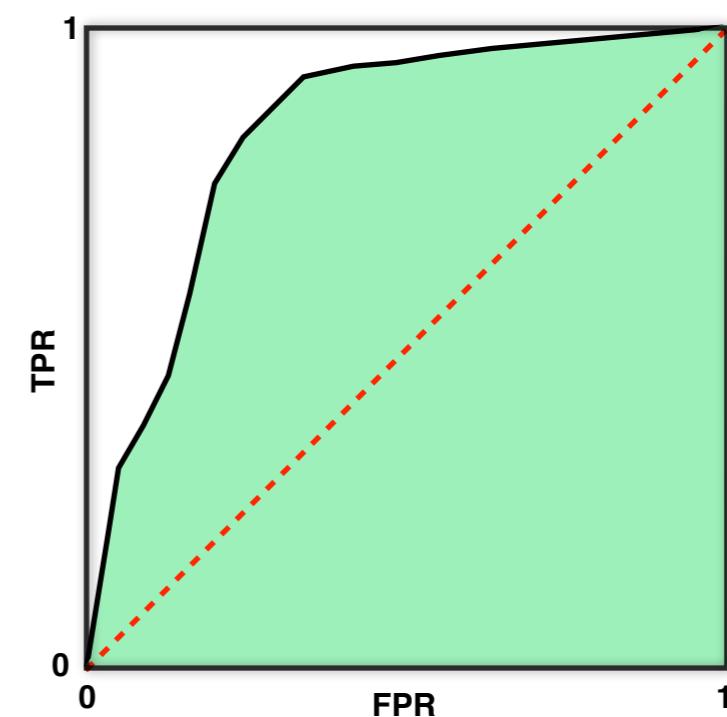
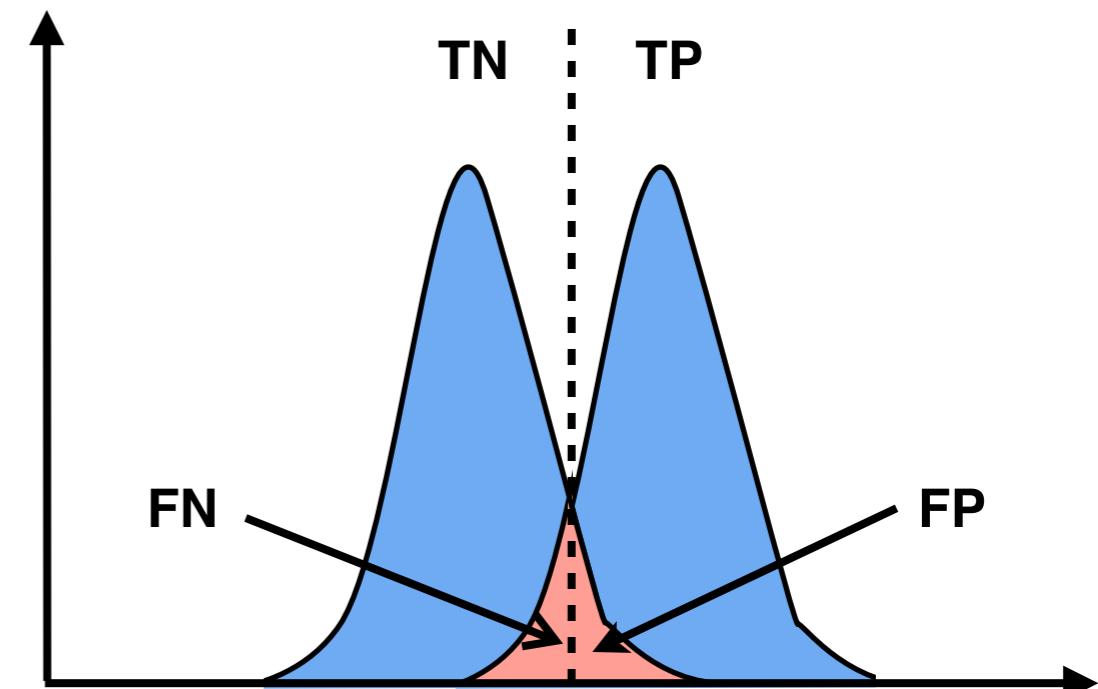
$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

The Area Under the Receiver operating characteristic (ROC) Curve (AUC) is a prediction evaluation measure that is 0.5 for completely random predictors and close to 1.0 for highly accurate predictors.

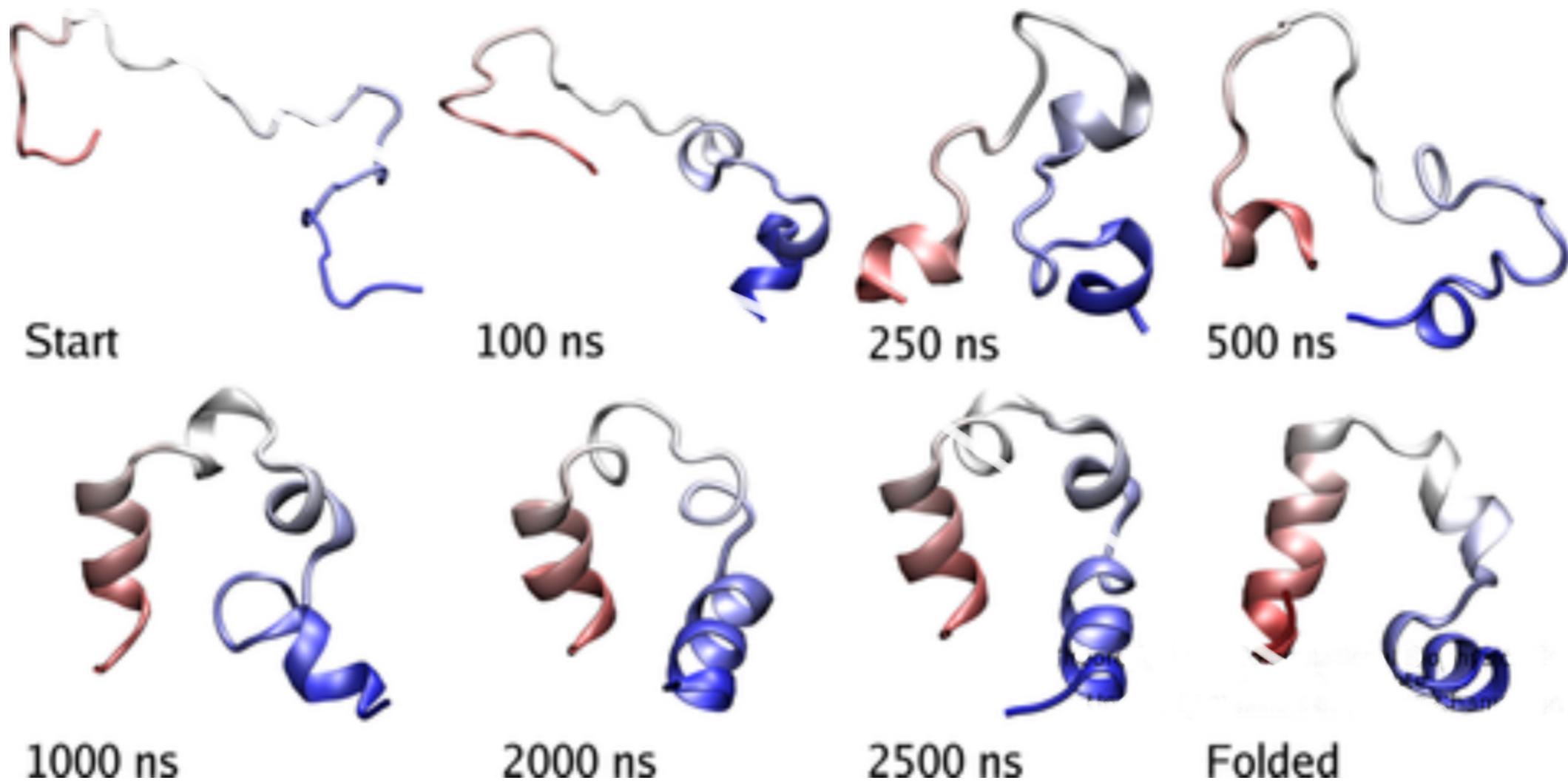
Baldi et al. (2000) Bioinformatics, 16:412-424



Mutation and Stability

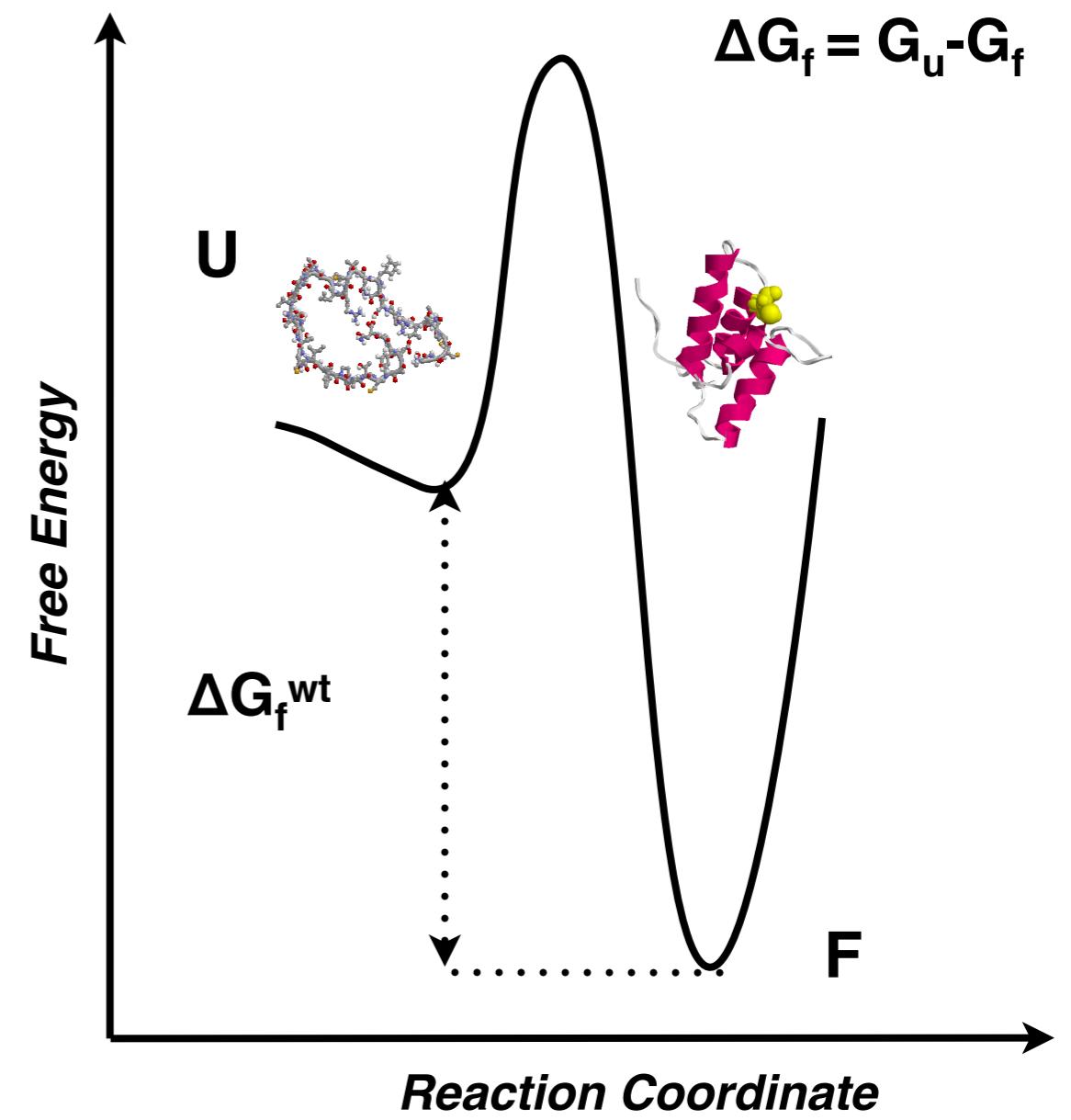
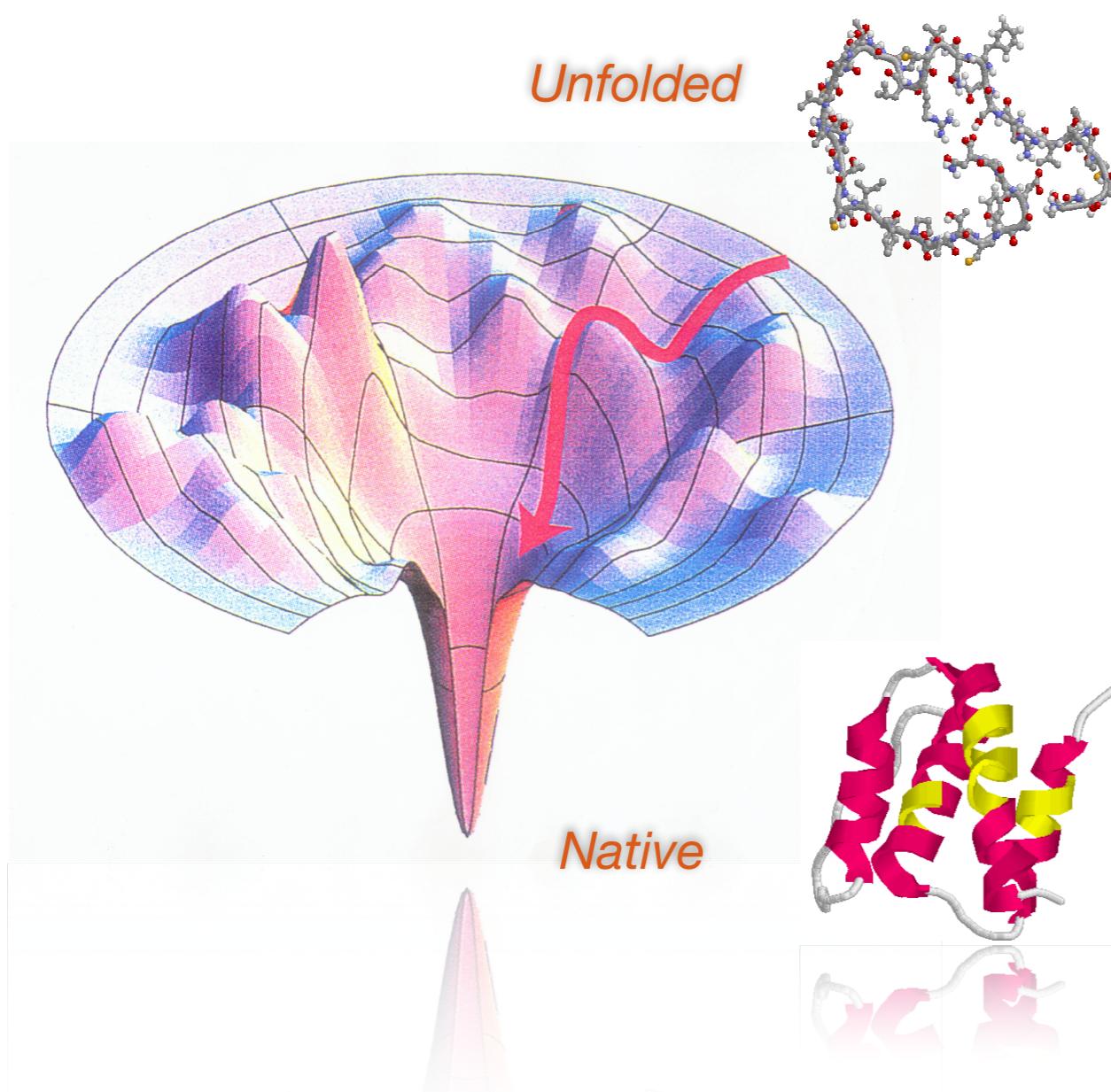
Protein folding

Protein folding is the **process by which a protein assumes its native structure** from the unfolded structure



Folding and stability

The folding free energy difference, ΔG_f , is typically small, of the order of -5 to -15 kcal/mol for a globular protein (compared to e.g. -30 to -100 kcal/mol for a covalent bond).

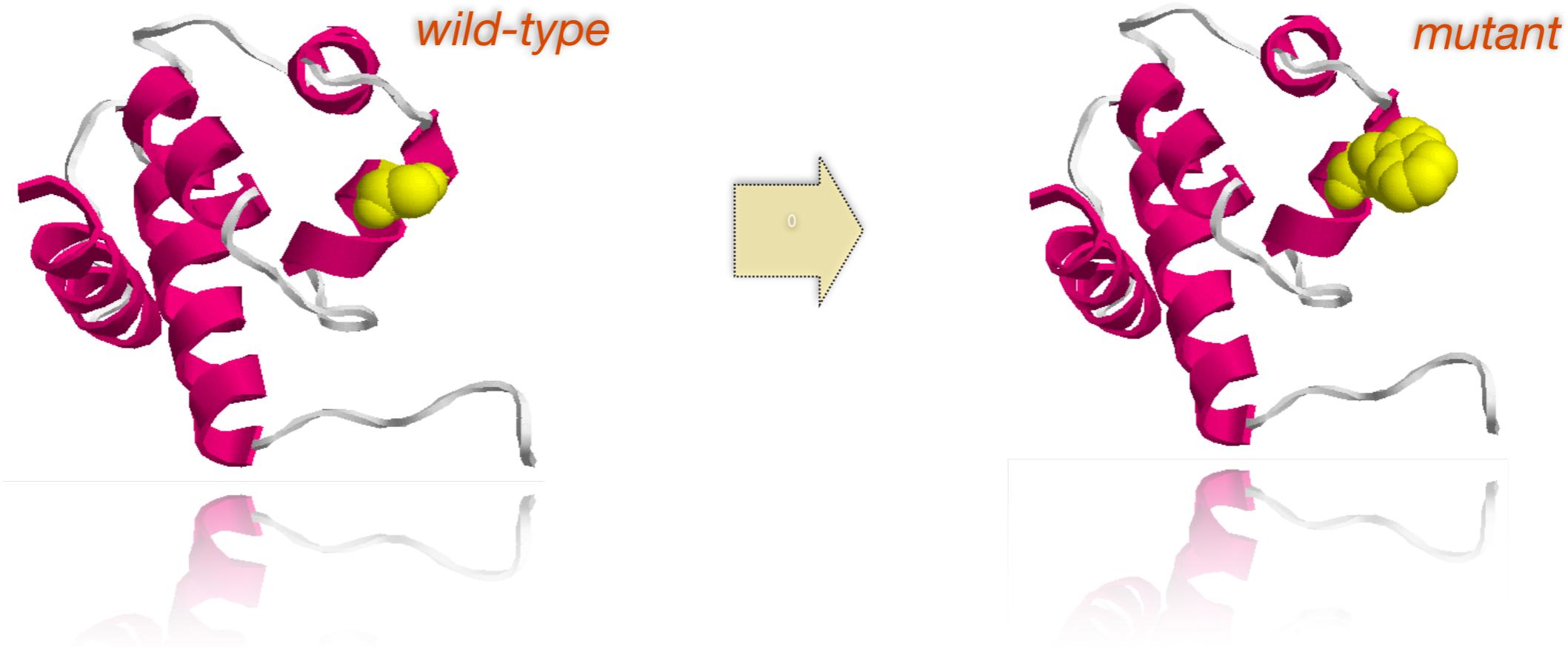


Folding and mutations

- Mutations of the protein sequence can affect the folding process changing the stability of the folded structure.
- Failure to folding process can produce inactive proteins with different properties even toxic. Protein misfolding is believed to be the main cause of neurodegenerative and other diseases.
- Web available databases are collecting large amount of thermodynamic data from mutagenesis experiments that can be used to develop methods for the prediction the protein stability change upon mutation.

Mutation and stability

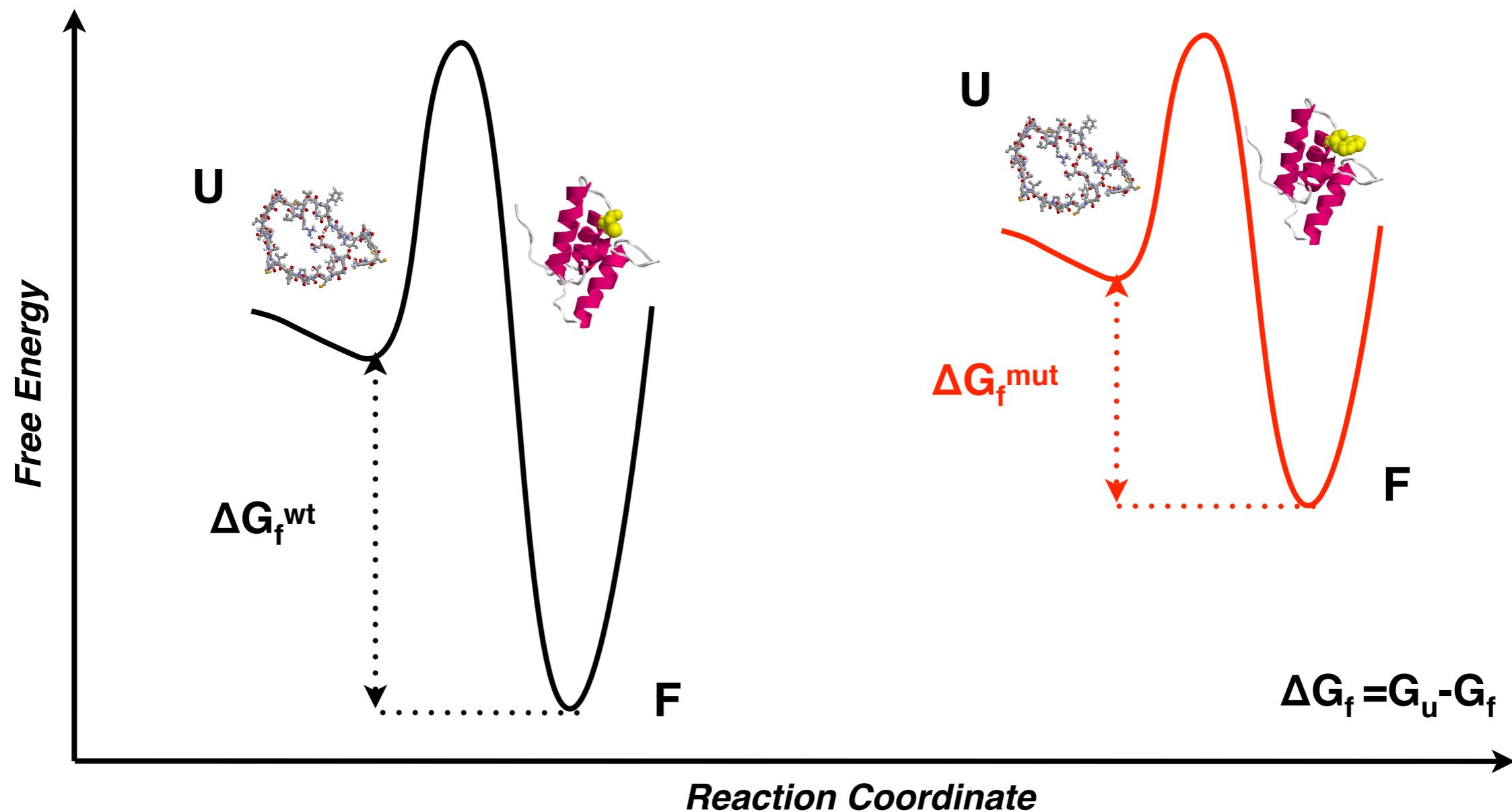
if a protein is mutated in a single site, what is the effect of the mutation on the stability of the protein?



Free energy change

If we mutate one residue in the protein sequence,
is the protein stability **increased or decreased?**

$$\Delta\Delta G_f = \Delta G_f^{\text{mut}} - \Delta G_f^{\text{wt}}$$



ProTherm database

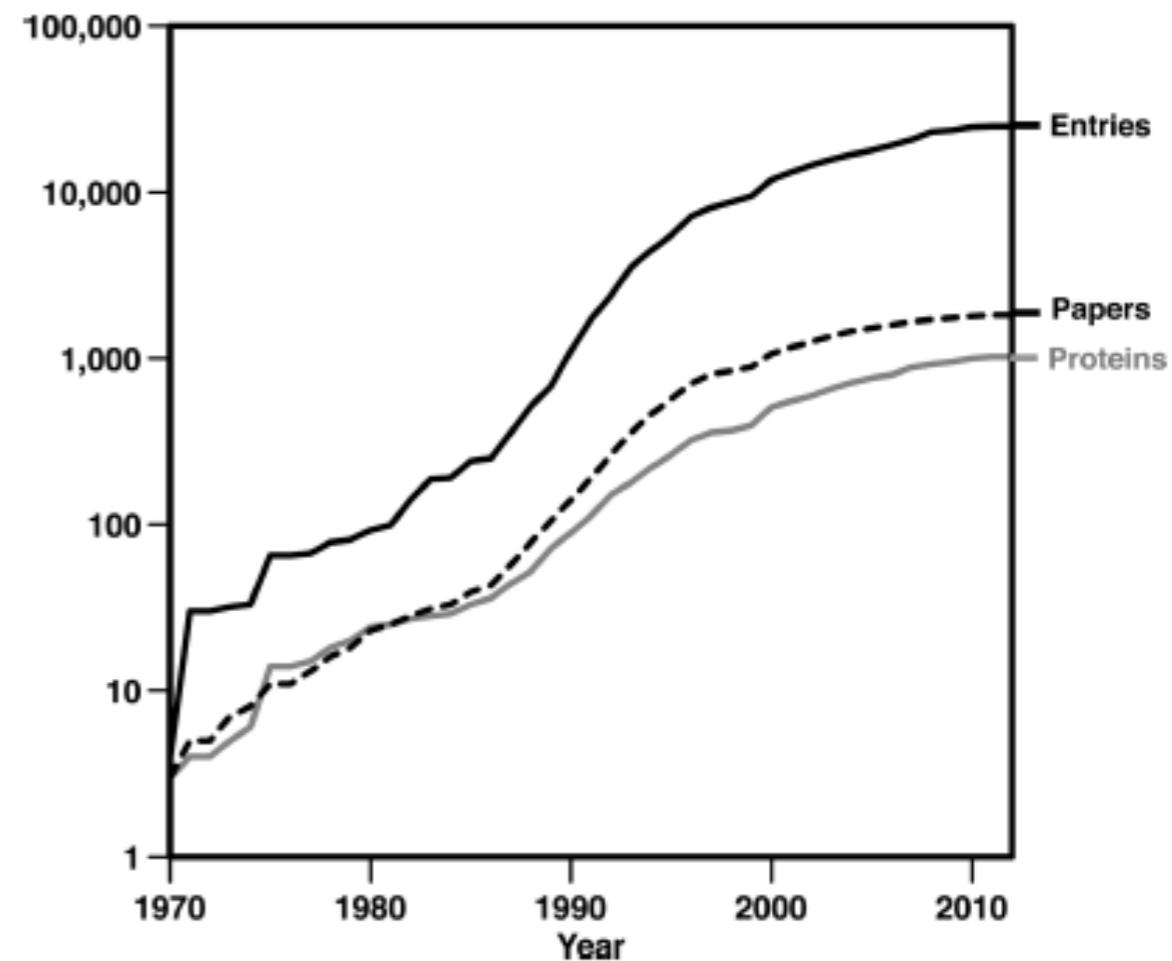
ProTherm is a collection of numerical data of thermodynamic parameters including **Gibbs free energy change, enthalpy change, heat capacity change, transition temperature** etc. for wild type and mutant proteins, that are important for understanding the structure and stability of proteins.

ProTherm is a collection of numerical data of thermodynamic parameters such as Gibbs free energy change, enthalpy change, heat capacity change, transition temperature etc. for wild type and mutant proteins, that are important for understanding the structure and stability of proteins. It also contains information about secondary structure and accessibility of wild type residues, experimental conditions (pH, temperature, buffer, ion and protein concentration), measurements and methods used for each data, and activity information (Km and Kcat).

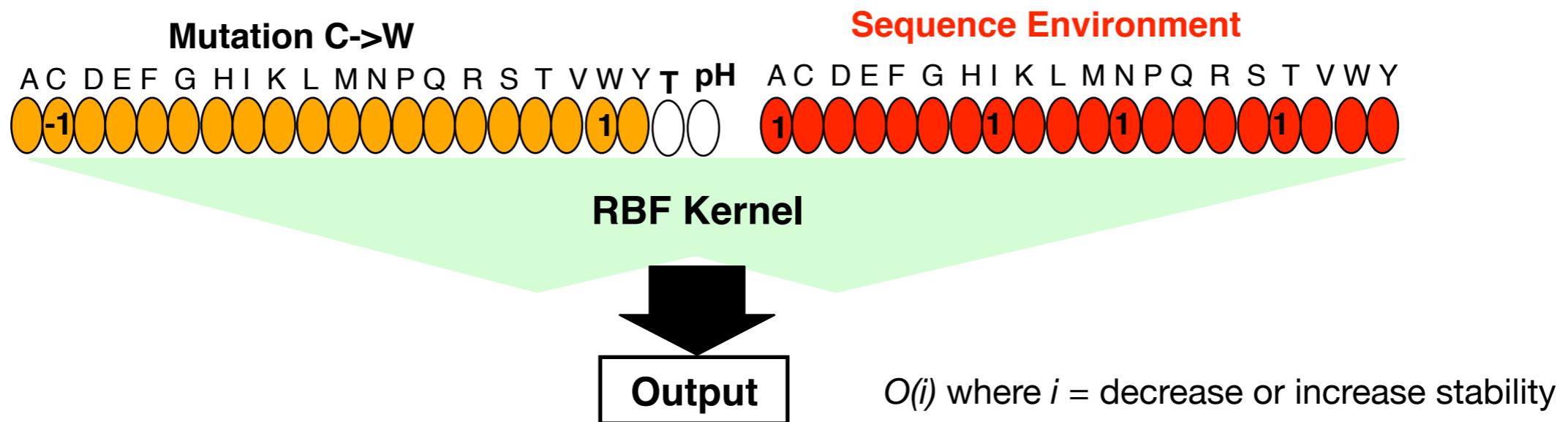
ProTherm is cross-linked with sequence databases (PIR and SWISS-PROT), structural database (Protein Data Bank), functional database (Protein Mutant Database), and literature database (PubMed). Moreover, the thermodynamic information is integrated with structural and functional information through the relational database, [3DinSight](#). The WWW interface enables users to search data based on various terms with different sorting options for outputs, and view three dimensional structures with automatically mapped mutation sites and surrounding amino acids. For more detail about ProTherm, please see [here](#).

Please note that this database is under constant development. There will be changes without prior notice. We welcome your comments and suggestions to improve this database.

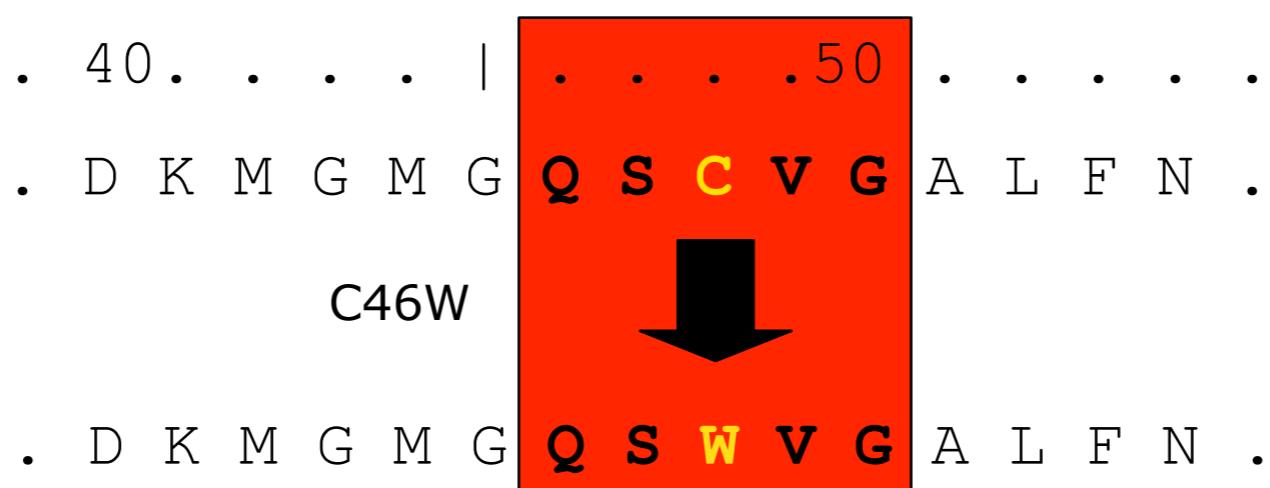
Total number of entries	25820
Number of unique proteins	740
Total number of all proteins	1045
Number of Proteins with mutants	311
Number of Single Mutations	12561
Number of Double Mutations	1744
Number of Multiple Mutations	1132
Number of Wild Type	10383



Sequence-based predictor



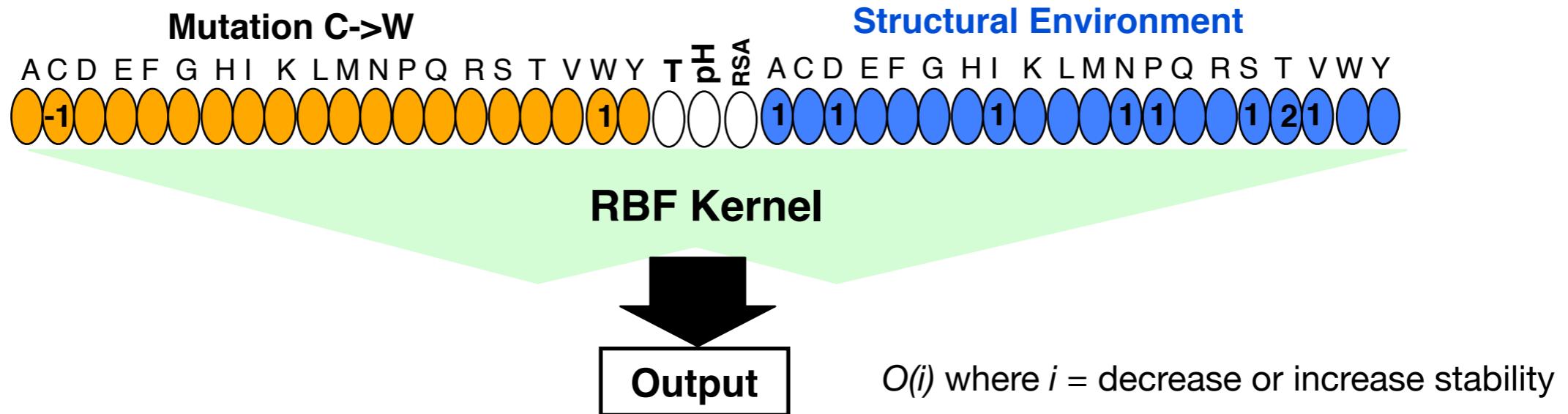
SVM-SEQUENCE: 20 element vector that describes the amino acid mutation,
 2 element pH and T (experimental conditions)
 20 more input features (40 in total) encoding the sequence residue environment



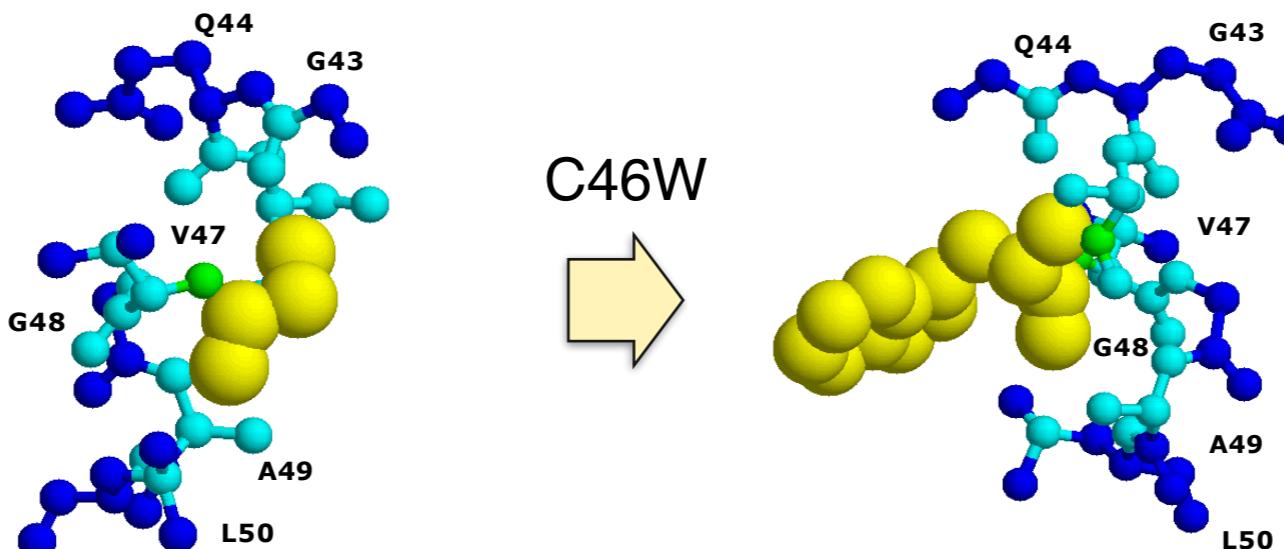
■ Mutated Aminoacid

■ Sequence Window

Structure-based predictor



SVM-STRUCTURE: 20 element vector that describes the amino acid mutation,
3 element pH, T and relative solvent accessible area
20 more input features (43 in total) encoding the structure residue environment



■ Mutated Aminoacid

■ $0 < R < 2\text{\AA}$

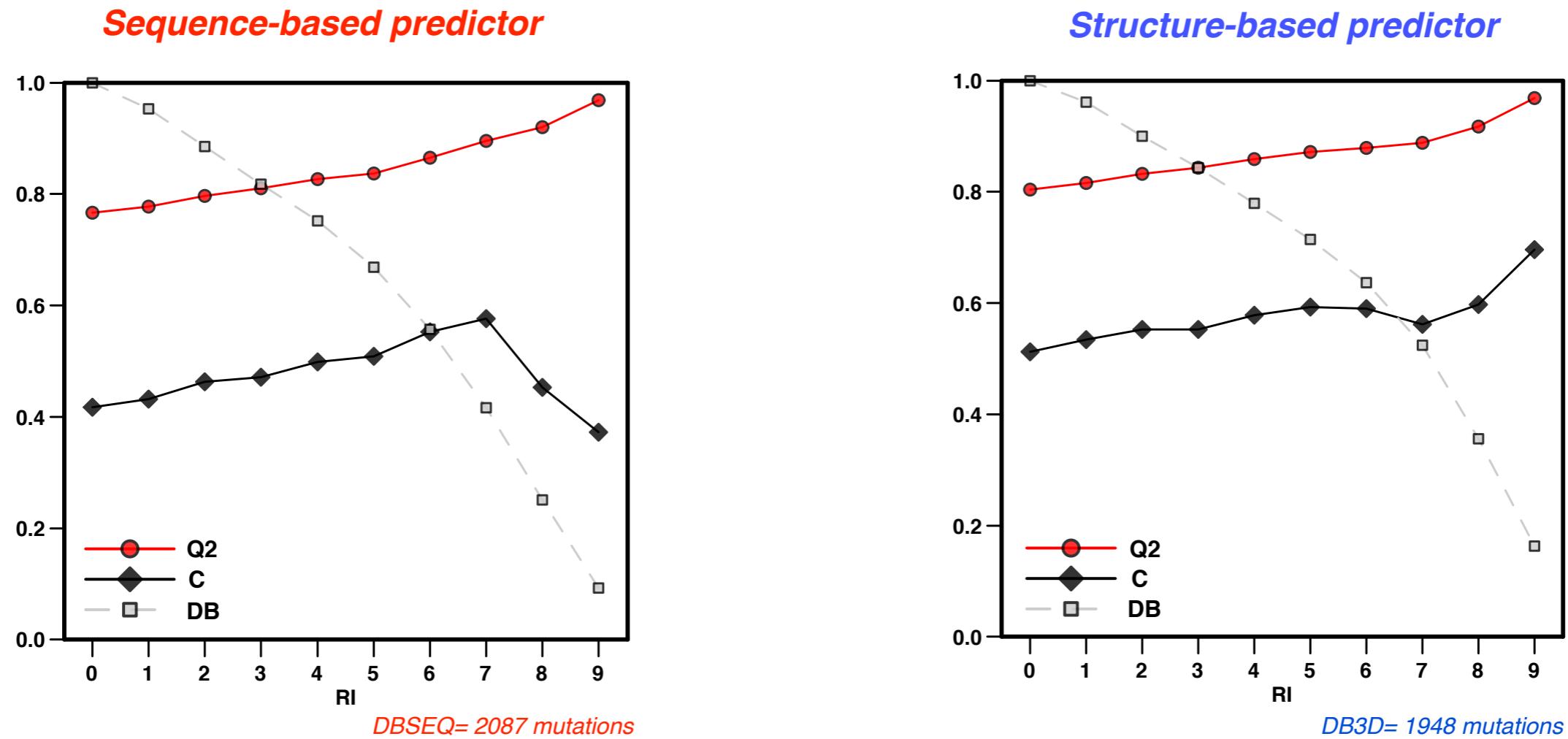
■ $2 < R < 4\text{\AA}$

■ $4 < R < 6\text{\AA}$

Classification results

	Q2	P[-]	S[-]	P[+]	S[+]	C
SVM-Sequence	0.77	0.79	0.91	0.69	0.46	0.42
SVM-Structure	0.80	0.83	0.91	0.73	0.56	0.51

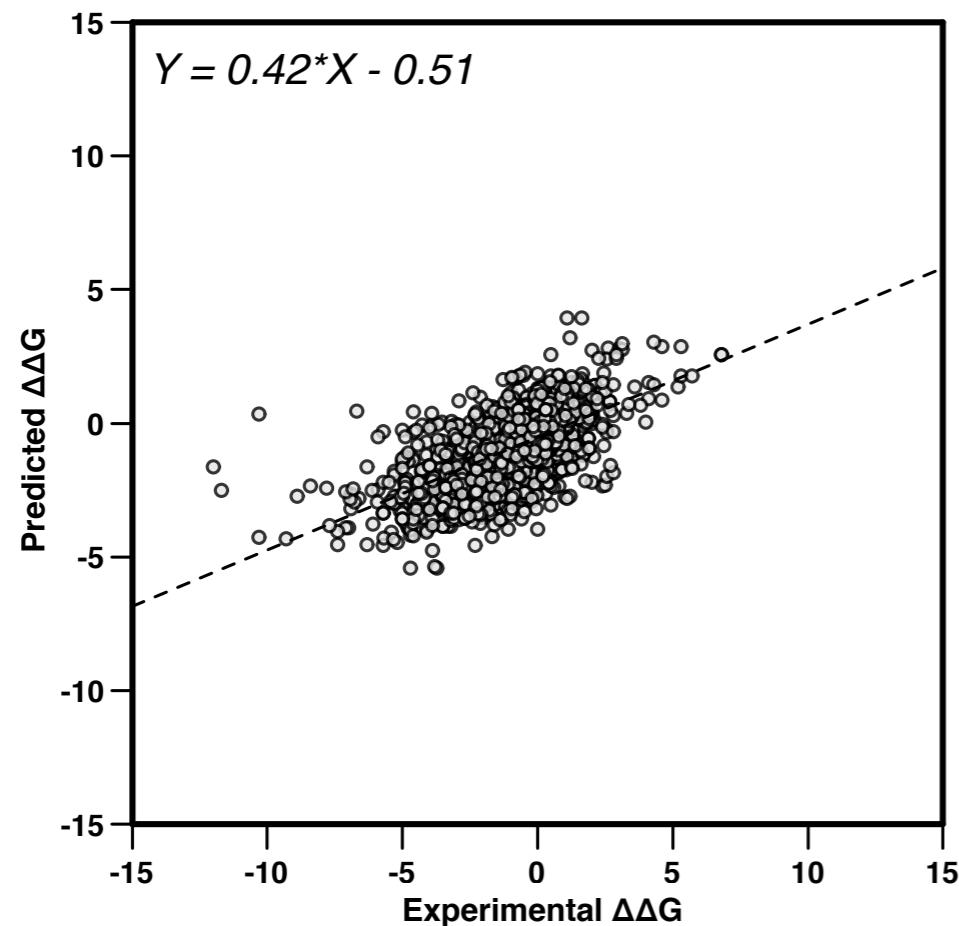
+ Increase stability – Decrease stability



Q2: Overall Accuracy **C:** Mean Correlation Coefficient **DB:** Fraction of database that are predicted with a reliability \geq the given threshold

Regression results

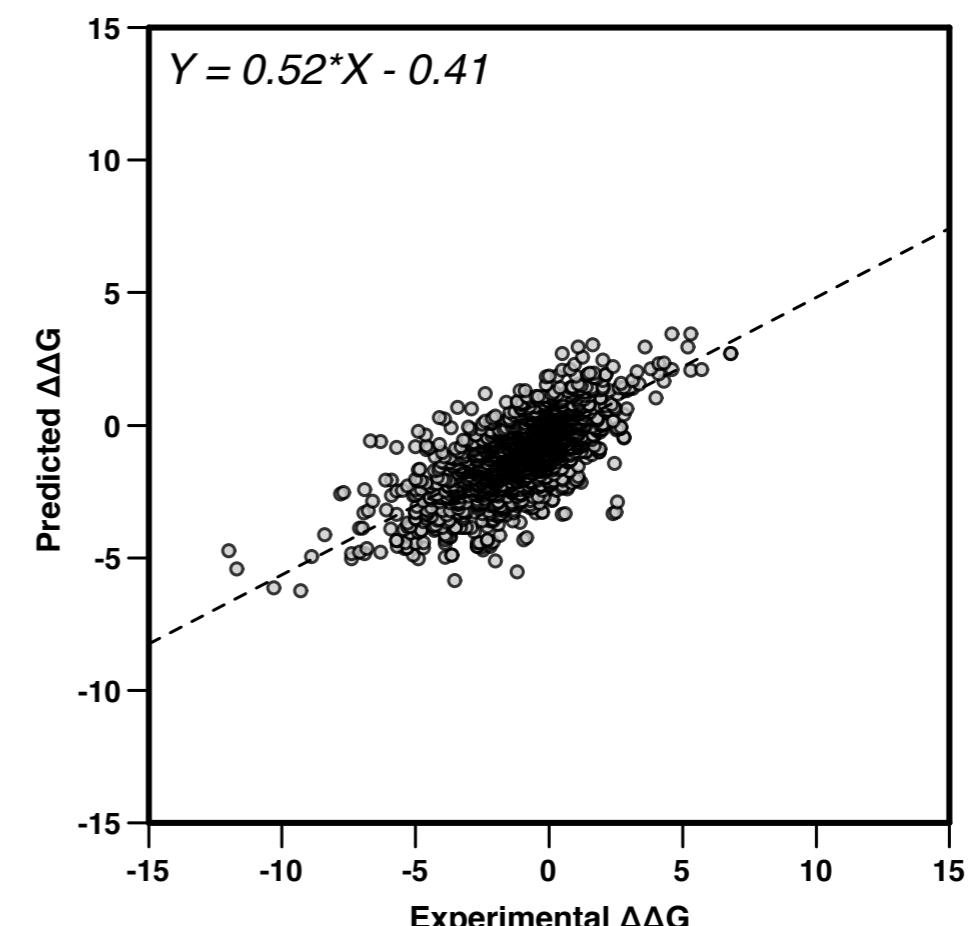
Sequence-based predictor



DBSEQ= 2087 mutations

C= 0.62 (RMSE= 1.45 kcal/mole)

Structure-based predictor



DB3D= 1948 mutations

C= 0.71 (RMSE= 1.30 kcal/mole)

<http://folding.biofold.org/i-mutant>

Capriotti et al. (2005) Nucleic Acids Research 33, W306-W310.

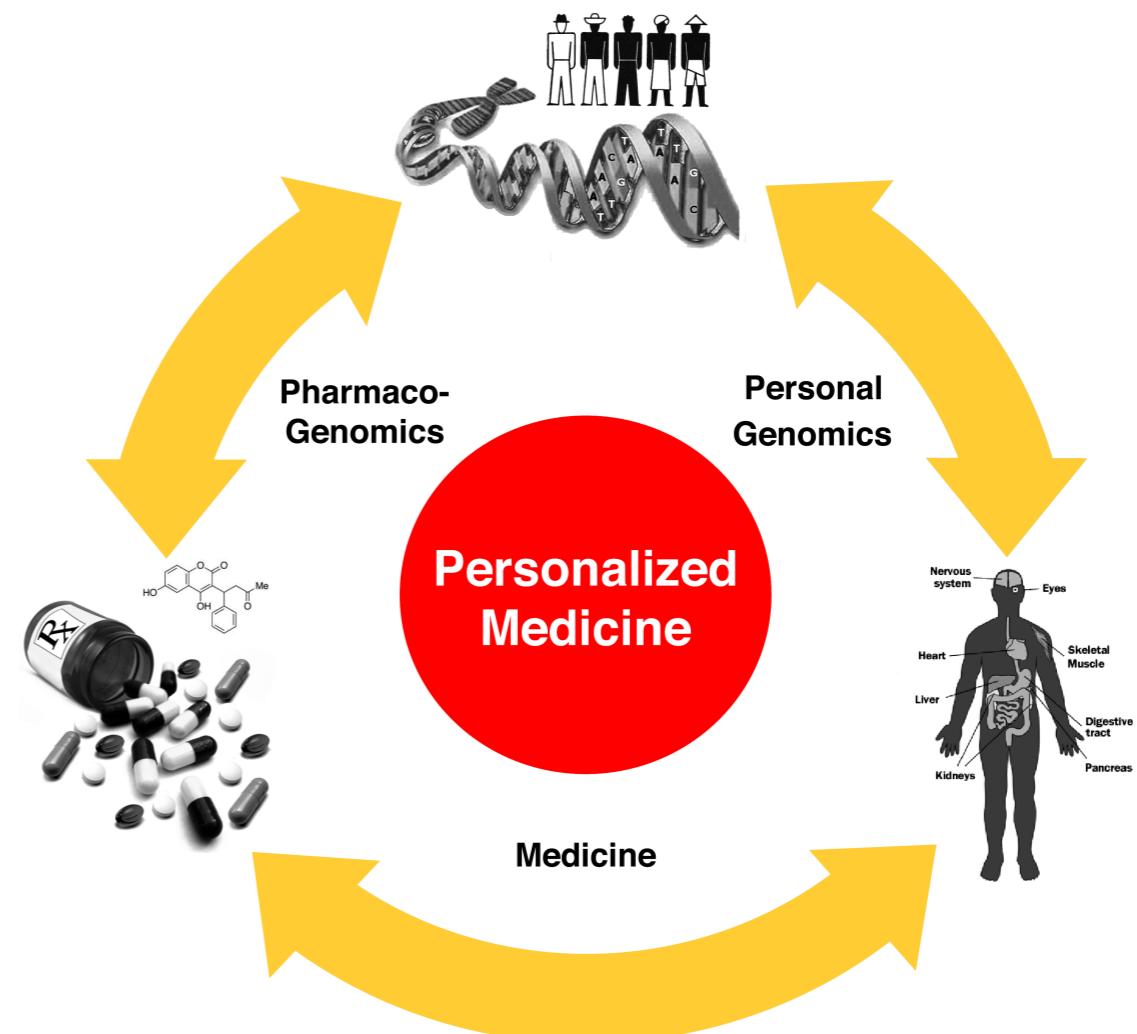
Mutation and Disease

Personalized medicine

Currently direct to consumers company are performing **genotype test** on **markers** associated to **genetic traits**, and soon **full genome sequencing** will cost about \$1000.

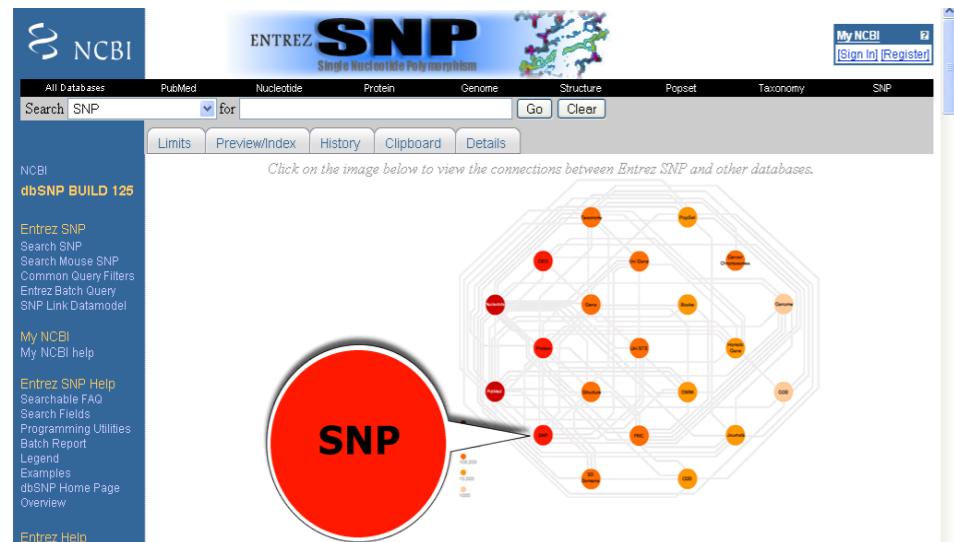
The future bioinformatics challenges for personalized medicine will be:

1. Processing Large-Scale **Robust Genomic Data**
2. **Interpretation** of the Functional Effect and the Impact of Genomic Variation
3. Integrating Systems and Data to **Capture Complexity**
4. Making it all **clinically relevant**



SNVs and SAVs databases

dbSNP (2016/2017) @ NCBI



<http://www.ncbi.nlm.nih.gov/>

SwissVar (Jun 2017) @ ExPASy



<http://www.expasy.ch/swissvar/>

Single Nucleotide Variants

<i>Homo sapiens</i>	135,967,291
<i>Bos taurus</i>	39,722,628
<i>Mus musculus</i>	16,396,141

Single Amino acid Variants

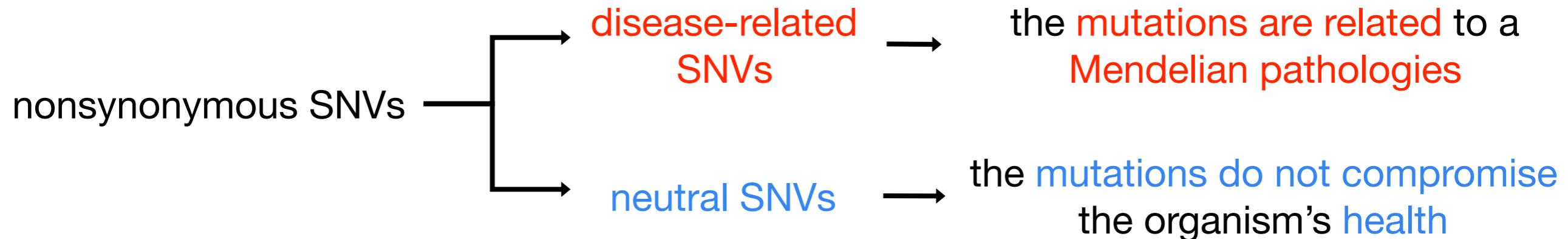
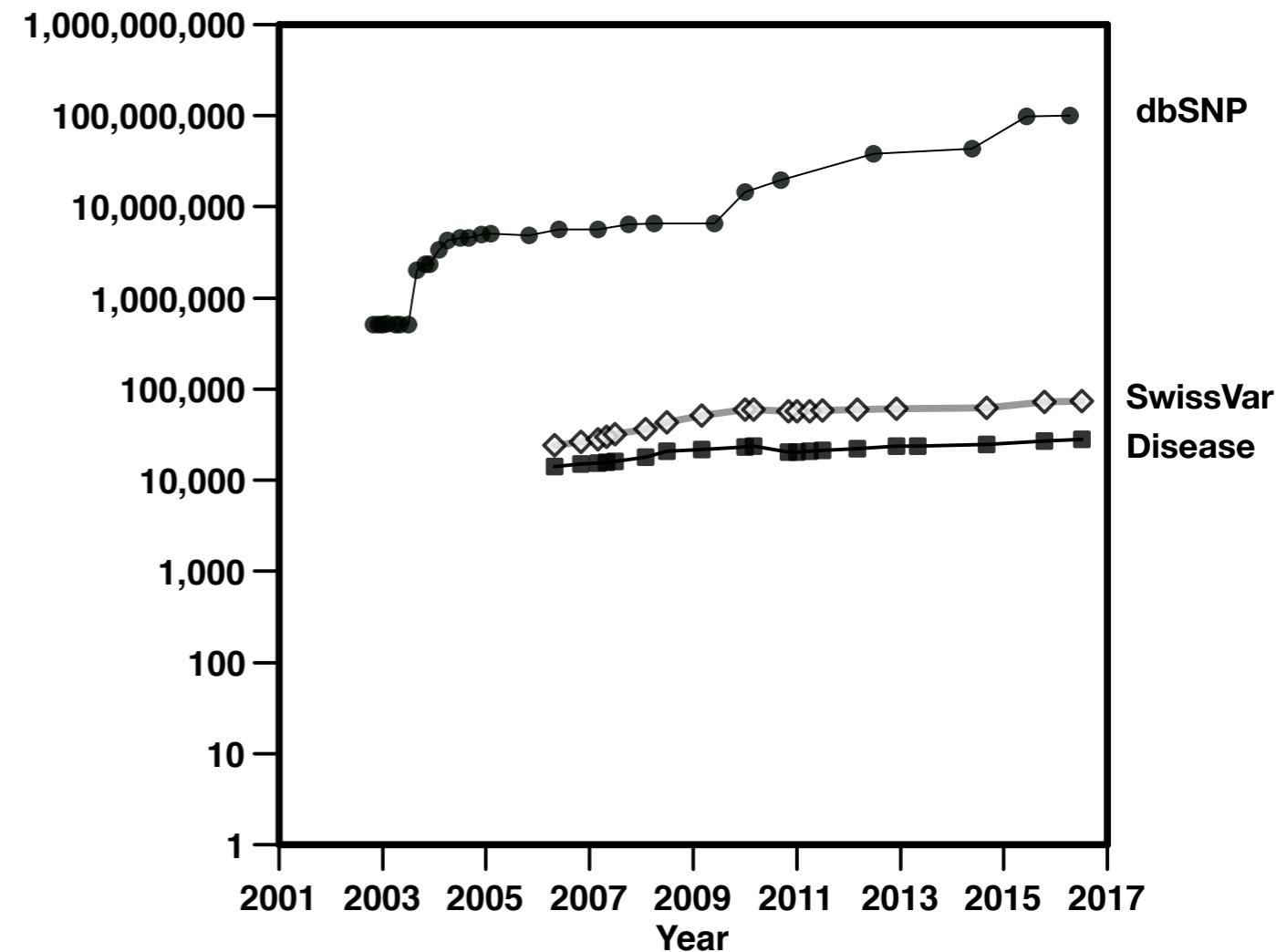
<i>Homo sapiens</i>	76,608
<i>Disease</i>	29,529
<i>Polymorphisms</i>	39,779

Jun 2017

SNVs and Disease

Single Nucleotide Variants (SNVs) are the most common type of genetic variations in human accounting for more than **90% of sequence differences** (1000 Genome Project Consortium, 2012).

SNVs can also be responsible of genetic diseases (Ng and Henikoff, 2002; Bell, 2004).



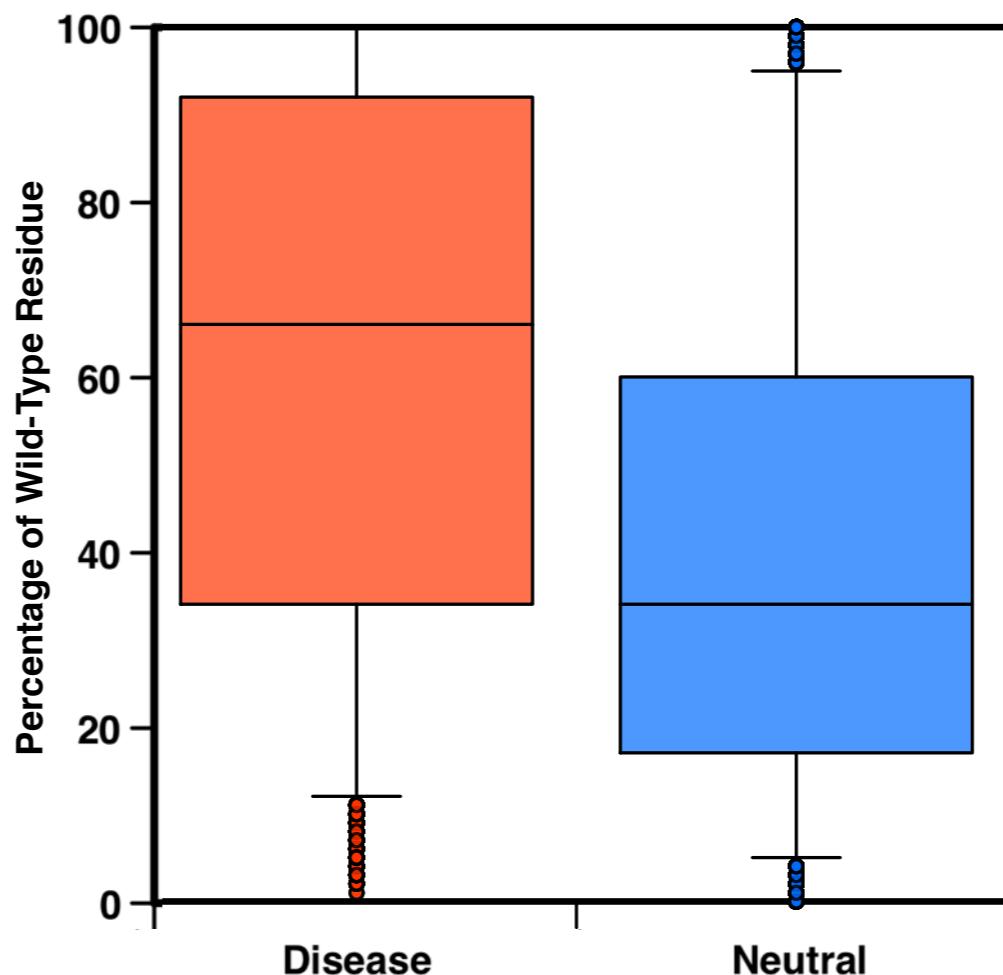
Conserved or not?

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

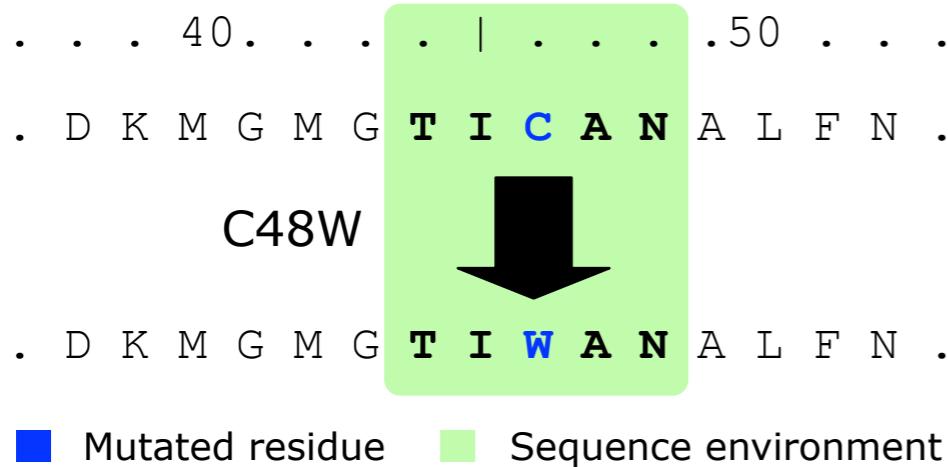
Sequence profile

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

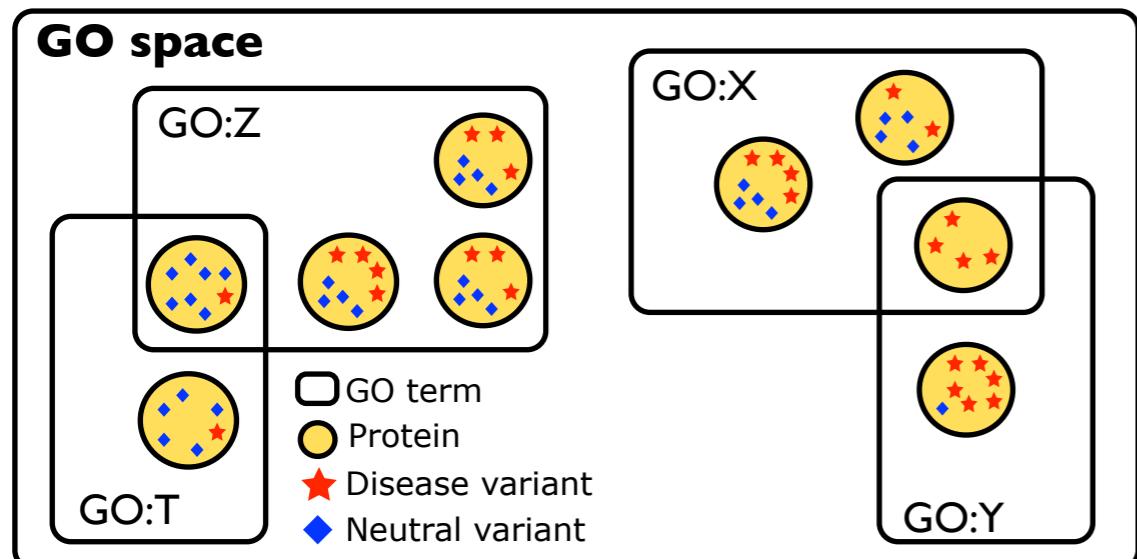
The **frequency distributions of the wild-type residues** for disease-related and neutral variants are significantly different (KS p-value=0).



SNPs&GO input features



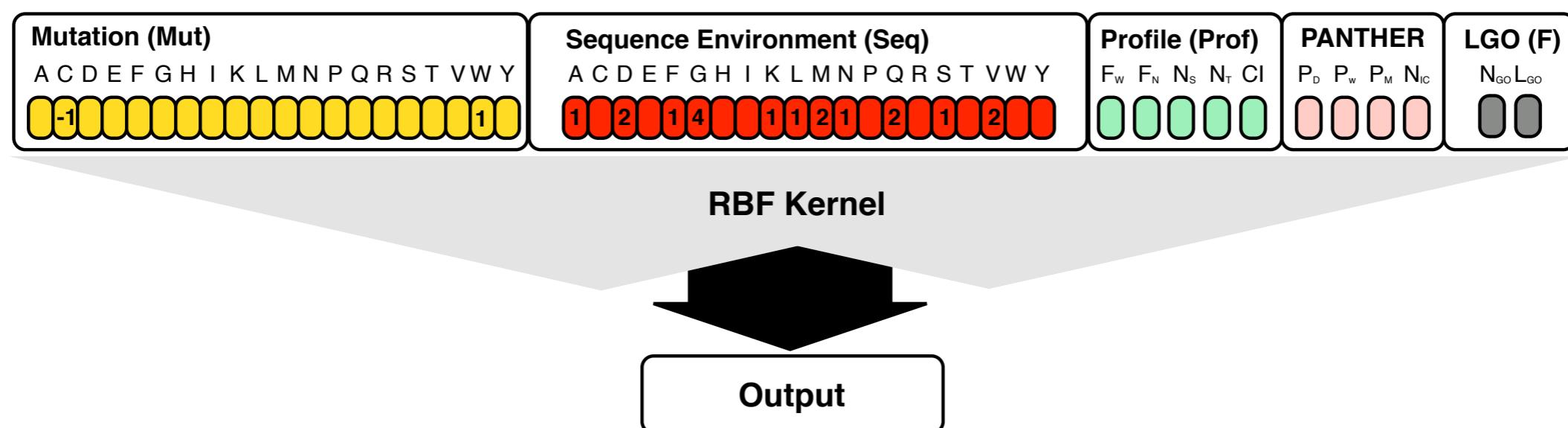
Protein sequence profile information derived from a multiple sequence alignment. It is encoded in a 5 elements vector corresponding to different features general and local features



The GO information are encoded in a 2 elements vector corresponding to the number unique of GO terms associated to the protein sequences and the sum of the logarithm of the total number of disease-related and neutral variants for each GO term.

SNPs&GO performance

SNPs&GO results in better performance with respect to previously developed methods.



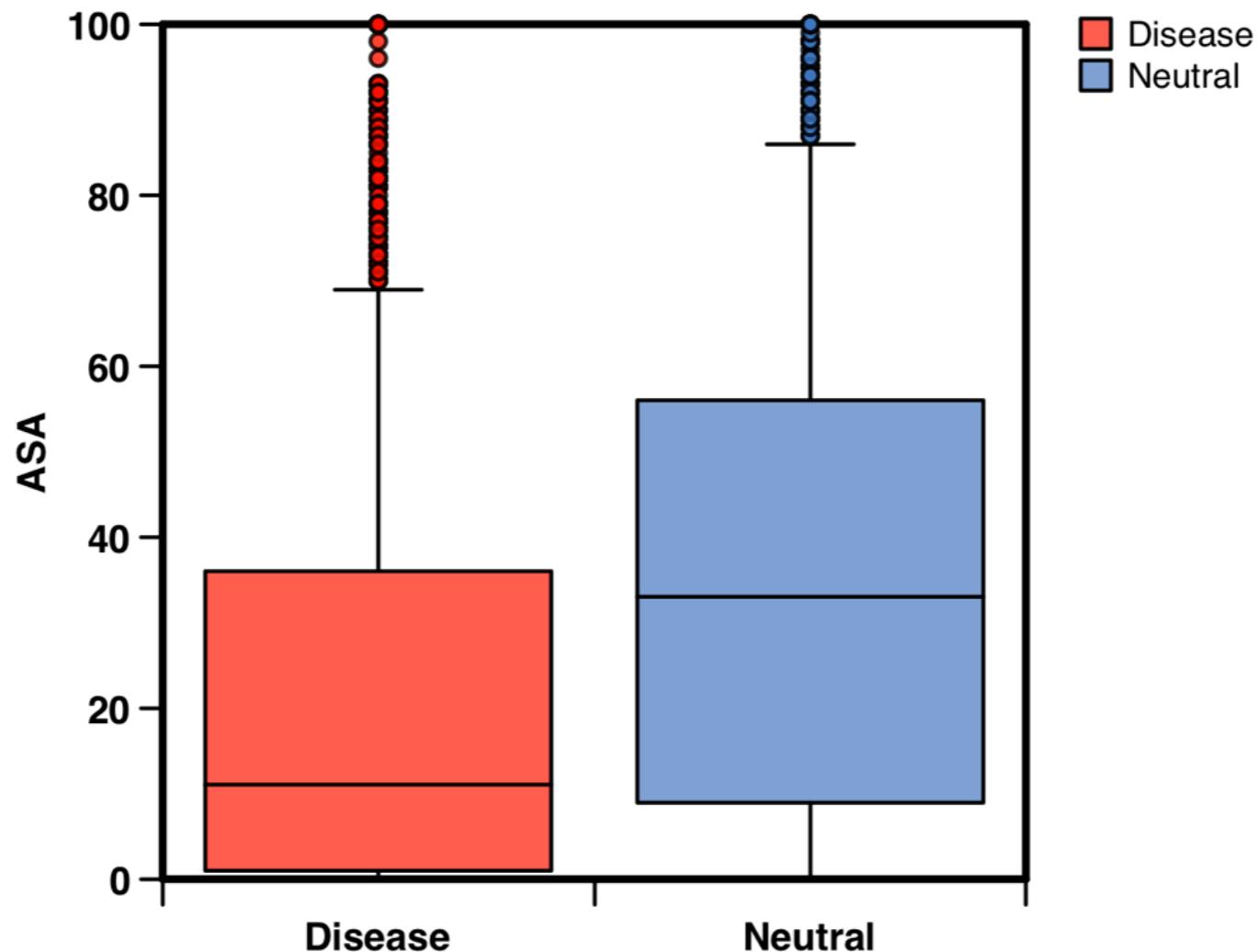
Method	Q2	P[D]	Q[D]	P[N]	Q[N]	C	PM
PolyPhen	0,71	0,76	0,75	0,63	0,64	0,39	58
SIFT	0,76	0,75	0,76	0,77	0,75	0,52	93
PANTHER	0,74	0,77	0,73	0,71	0,76	0,48	76
SNPs&GO	0,82	0,83	0,78	0,80	0,85	0,63	100

D = Disease related N = Neutral

DB= 33672 nsSNVs

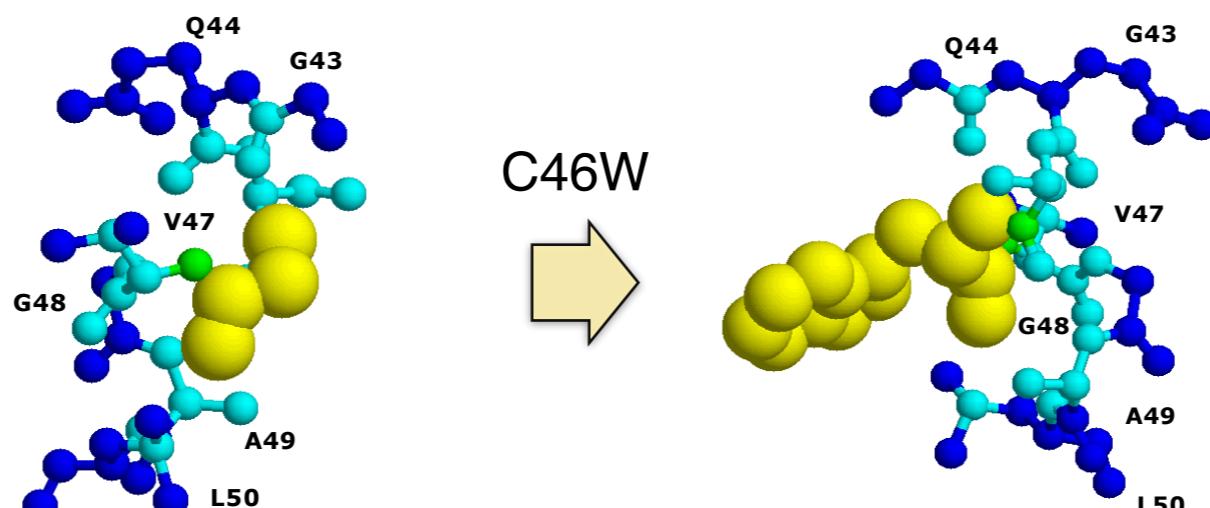
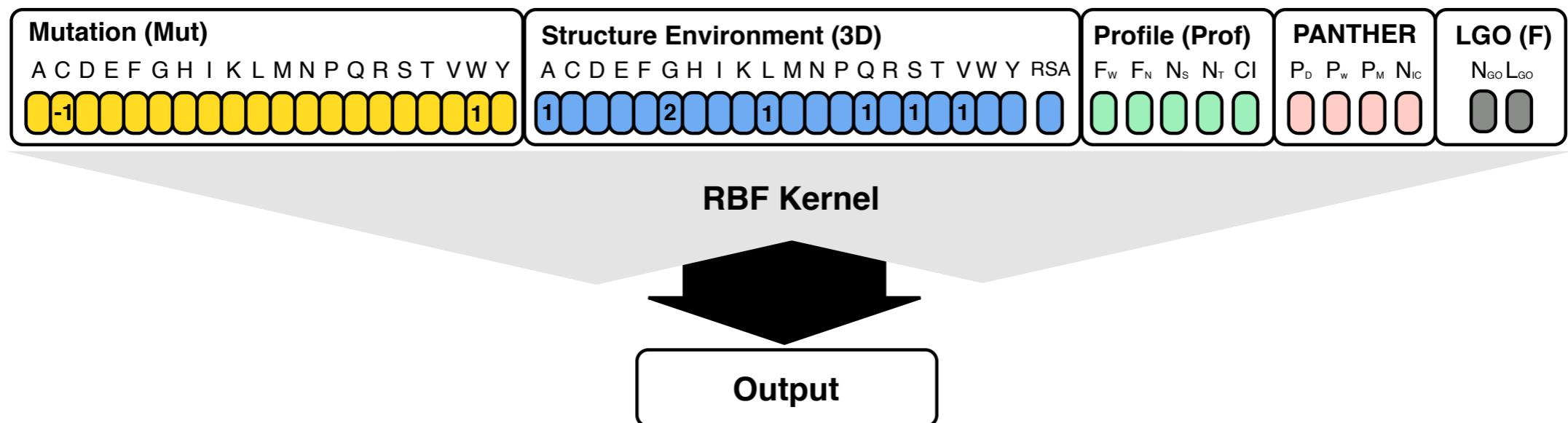
Structure environment

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.



The structure-based method

The method takes in to input 4 types of information encoded in a 48 elements vector. The input features are: mutation data; structure environment, sequence profile and functional score based on GO terms.



Mutated Aminoacid

0 < R < 2 Å

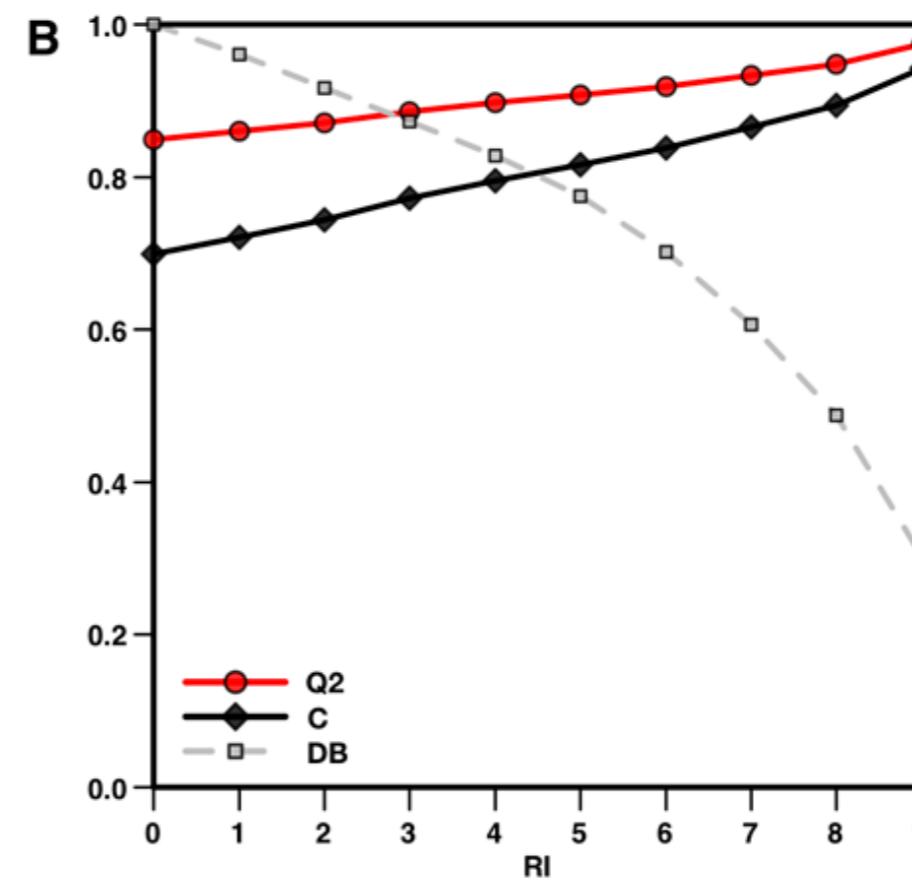
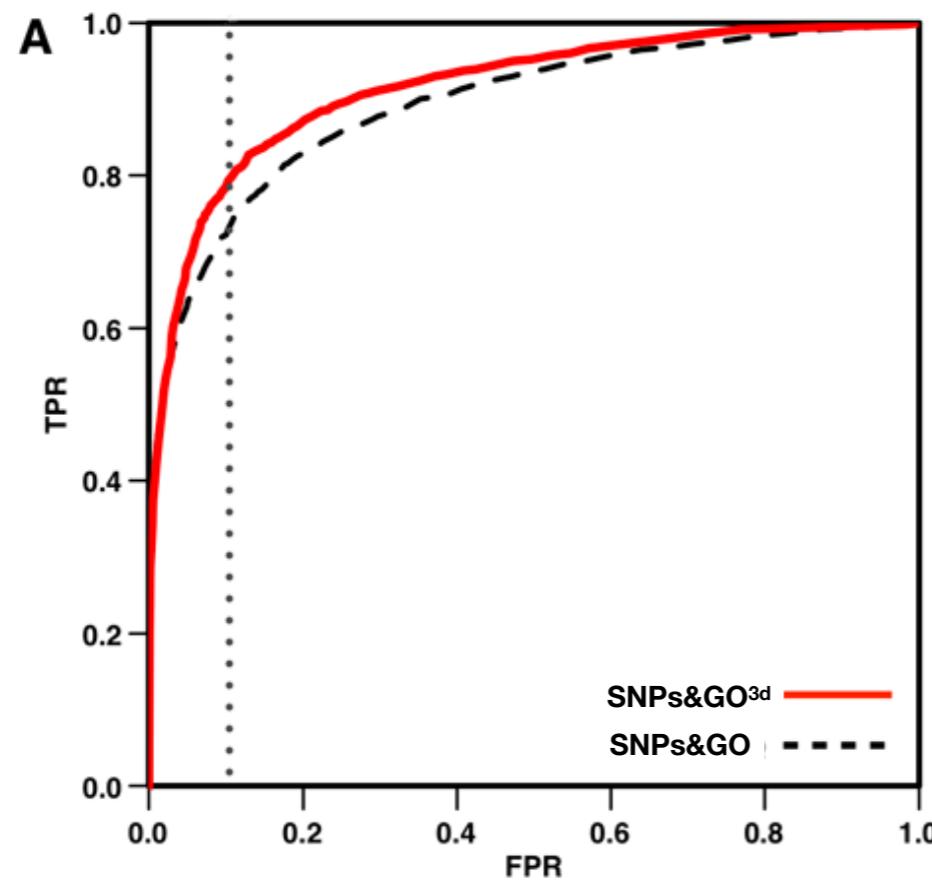
2 < R < 4 Å

4 < R < 6 Å

Sequence vs Structure

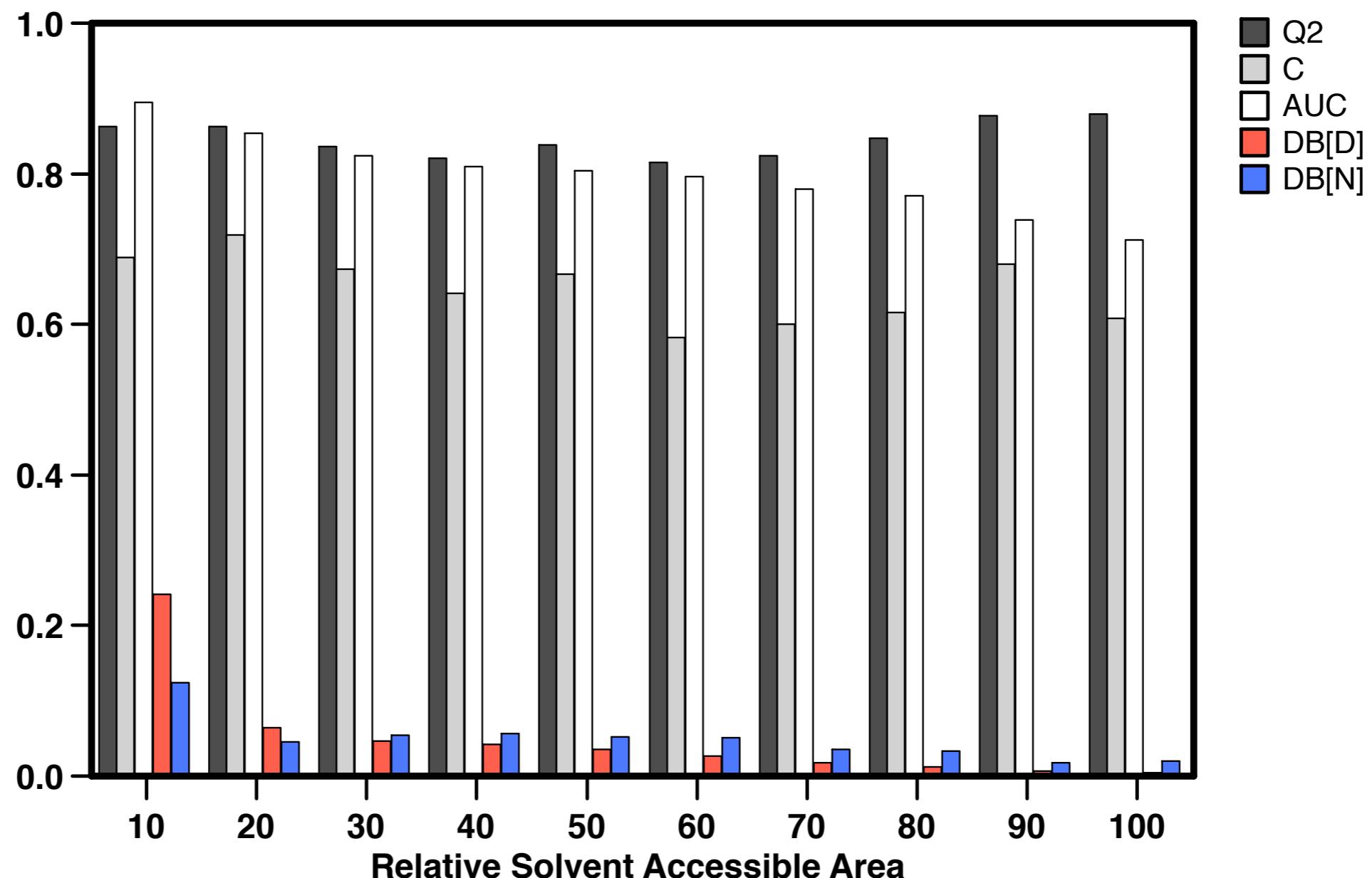
The structure-based method results in better accuracy with respect to the sequence-based one. **Structure based prediction are 3% more accurate** and **correlation coefficient increases of 0.06**. If 10% of FP are accepted the TPR increases of 7%.

	Q2	P[D]	S[D]	P[N]	S[N]	C	AUC
SNPs&GO	0.82	0.81	0.83	0.82	0.81	0.64	0.89
SNPs&GO^{3d}	0.85	0.84	0.87	0.86	0.83	0.70	0.92



Accuracy vs Accessibility

The predictions are more accurate for mutations occurring in buried region (0-30%). Mutations of exposed residues results in lower accuracy.

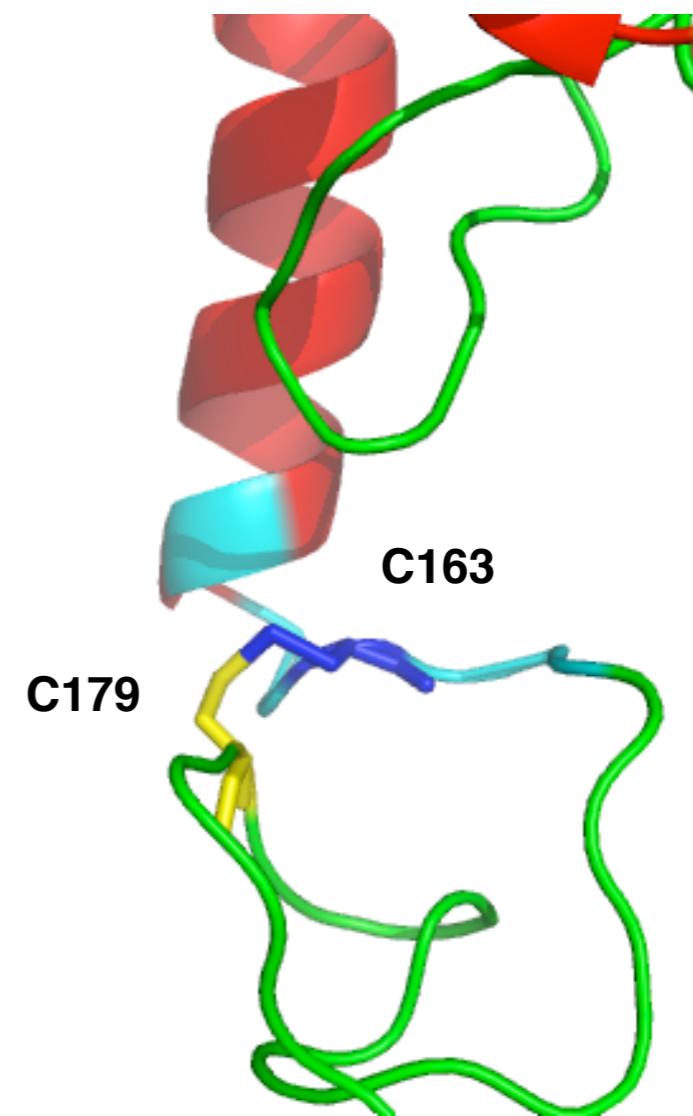


Prediction example

Damaging missing Cys-Cys interaction in the Glycosylasparaginase. The mutation p.Cys163Ser results in the loss of the disulfide bridge between Cys163 and Cys179. This SAP is responsible for Aspartylglucosaminuria.



1APY: Chain A, Res: 2.0 Å



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](#)



[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>

The P16 challenge

CDKN2A is the most common, high penetrance, susceptibility gene identified to date in **familial malignant melanoma**. **p16^{INK4A}** is one of the two **oncosuppressor** which promotes cell cycle arrest by inhibiting cyclin dependent kinase (CDK4/6).

Challenge: Evaluate how different variants of p16 protein impact its ability to block cell proliferation.

Provide a number between **50%** that represent the normal **proliferation rate of control cells** and **100%** the maximum proliferation rate in case cells.

SNPs&GO prediction

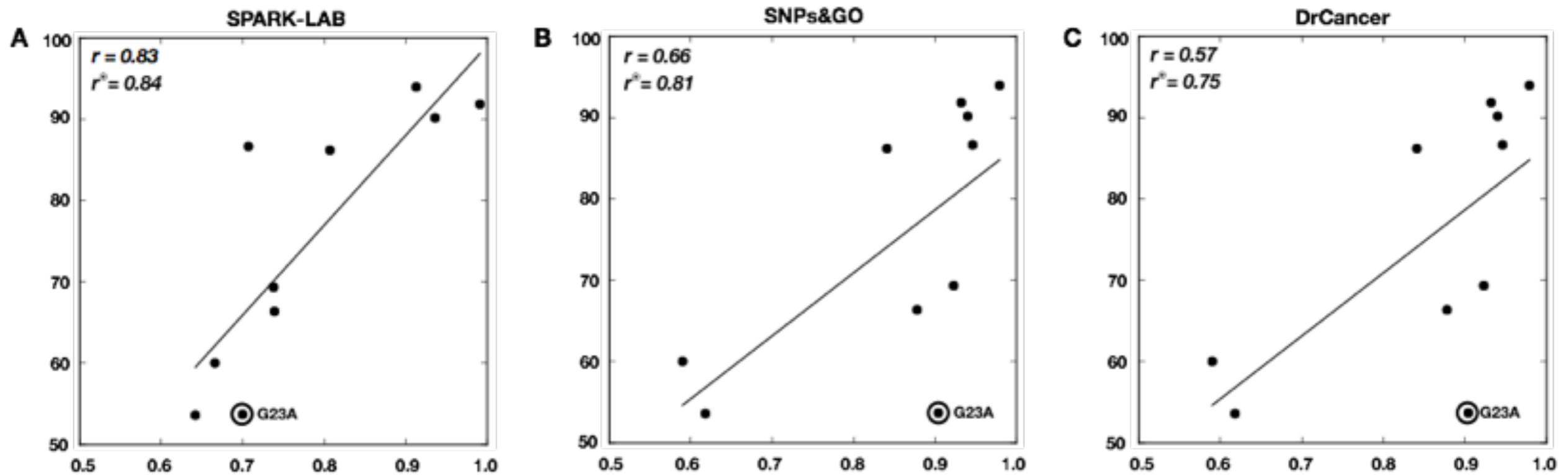
Proliferation rates predicted using the **output of SNPs&GO** without any optimization.

Variant	Prediction	Real	Δ	%WT	%MUT
G23R	0,932	0,918	0,014	84	0
G23S	0,923	0,693	0,230	84	1
G23V	0,940	0,901	0,039	84	0
G23A	0,904	0,537	0,367	84	2
G23C	0,946	0,866	0,080	84	0
G35E	0,590	0,600	0,010	12	14
G35W	0,841	0,862	0,021	12	0
G35R	0,618	0,537	0,081	12	4
L65P	0,878	0,664	0,214	15	1
L94P	0,979	0,939	0,040	56	0

P16 predictions

SNPs&GO resulted among the best methods for predicting the impact of P16INK4A variants on cell proliferation.

Method	Q2	AUC	MC	RMSE	rPearson	rSpearman	rKendallTau
SPARK-LAB	0.900	0.920	0.816	0.30	0.595	0.619	0.443
SNPs&GO	0.700	0.880	0.500	0.33	0.575	0.616	0.445
DrCancer	0.600	0.840	0.333	0.46	0.477	0.495	0.409



The NAGLU challenge

NAGLU is a lysosomal glycohydrolase which deficiency causes a rare disorder referred as Sanfilippo B disease

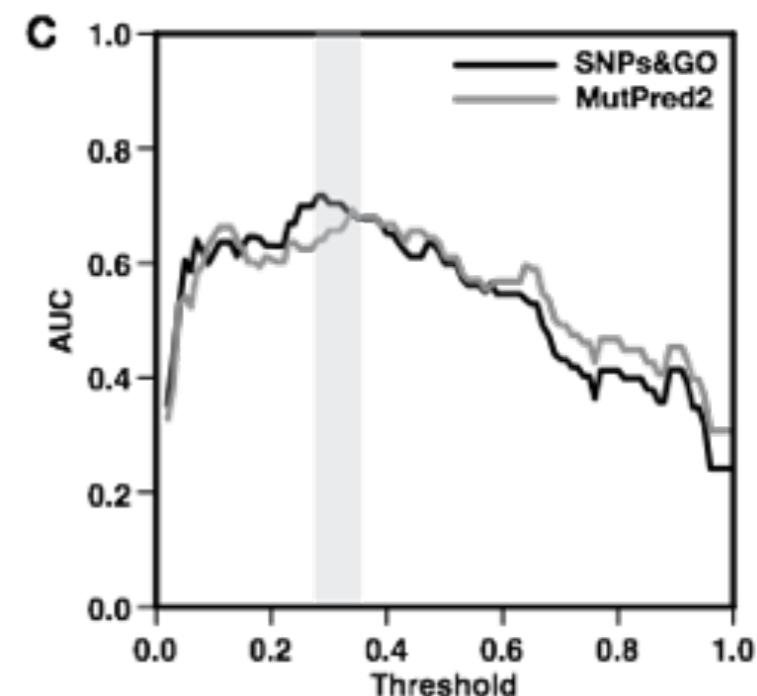
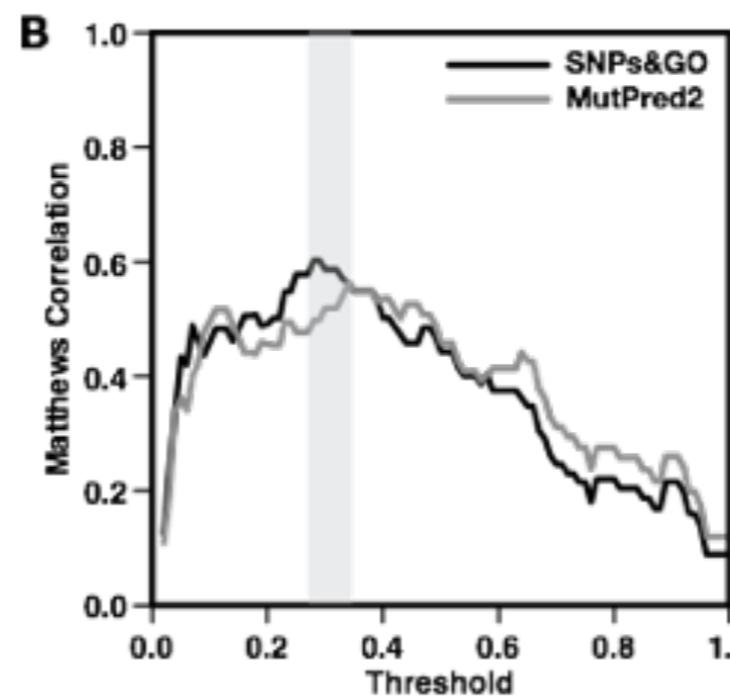
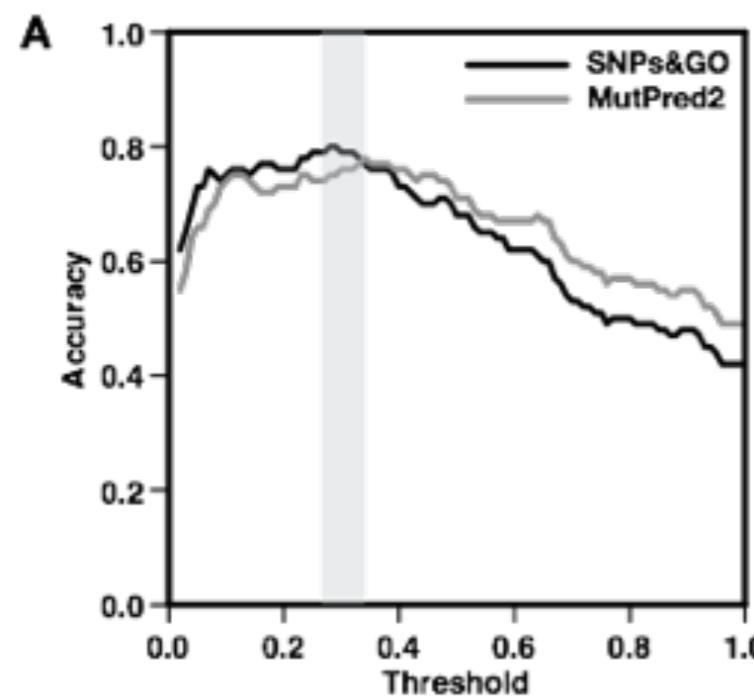
Challenge: Predict the effect of the 165 variants on NAGLU enzymatic activity.

The submitted prediction should be a numeric value ranging from 0 (no activity) to 1 (wild-type level of activity).

A posteriori evaluation

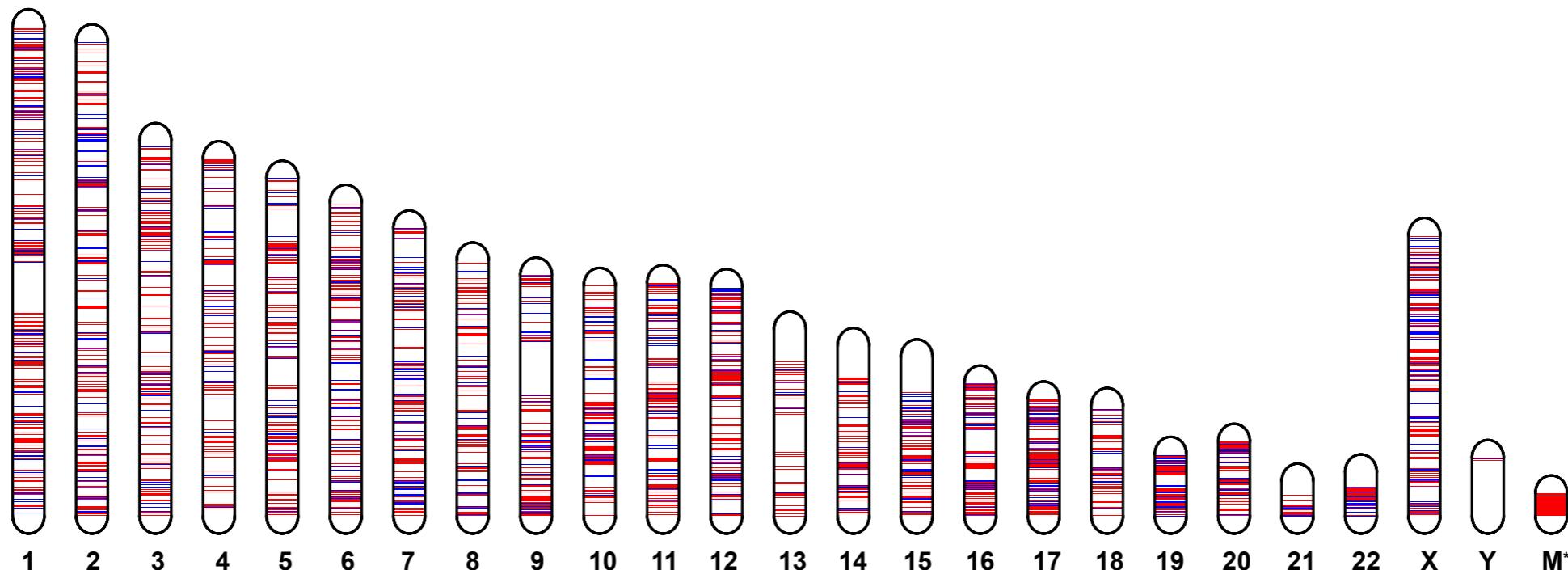
I performed a posteriori evaluation of the performance based on my version of the predictor and found that **SNPs&GO reaches similar accuracy than the best method (MutPred2)**

Method	Q2	AUC	MC	RMSE	rPearson	rSpearman	rKendallTau
MutPred2	0.780	0.850	0.565	0.30	0.595	0.619	0.443
SNPs&GO	0.800	0.854	0.603	0.33	0.575	0.616	0.445
SNPs&GO ⁰⁹	0.750	0.749	0.499	0.46	0.477	0.495	0.409



Whole-genome predictions

Most of the genetic variants occur in non-coding region that represents >98% of the whole genome.

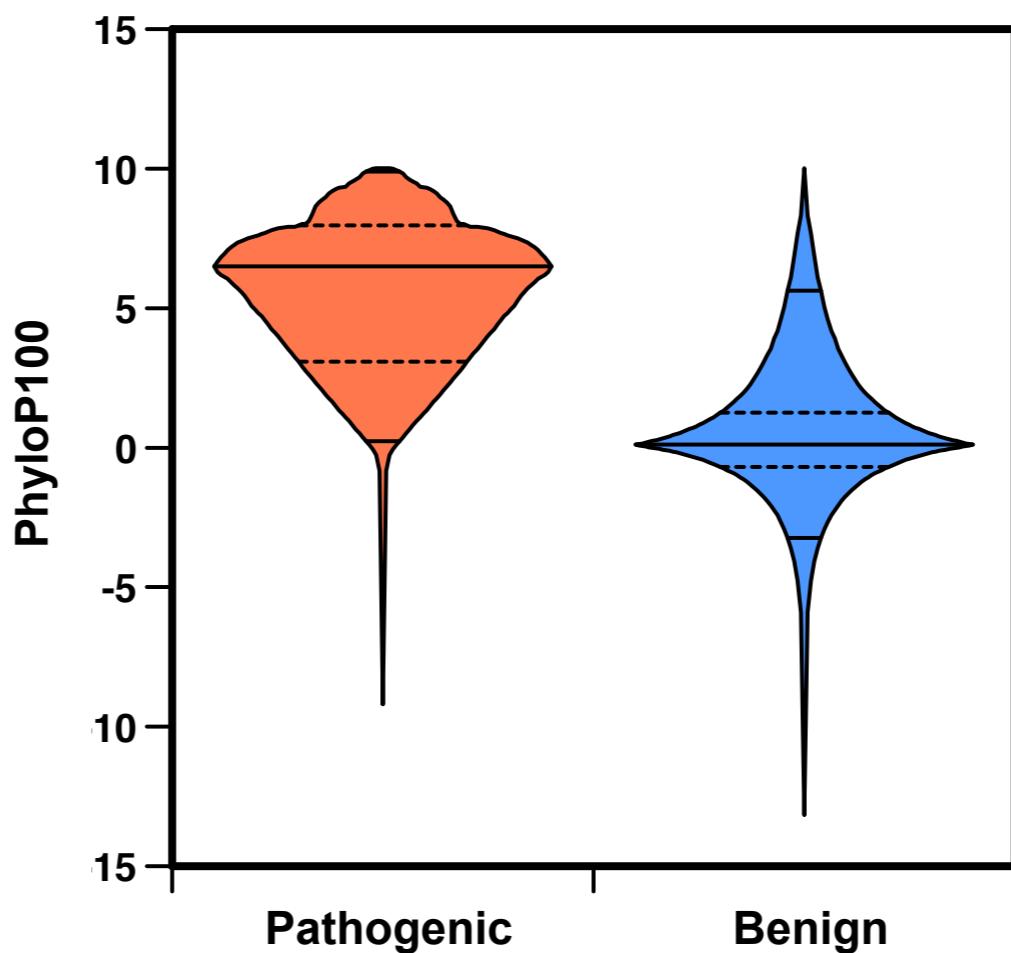


Predict the effect of SNVs in non-coding region is a challenging task because conservation is more difficult to estimate.

Sequence alignment is more complicated for sequences from non-coding regions.

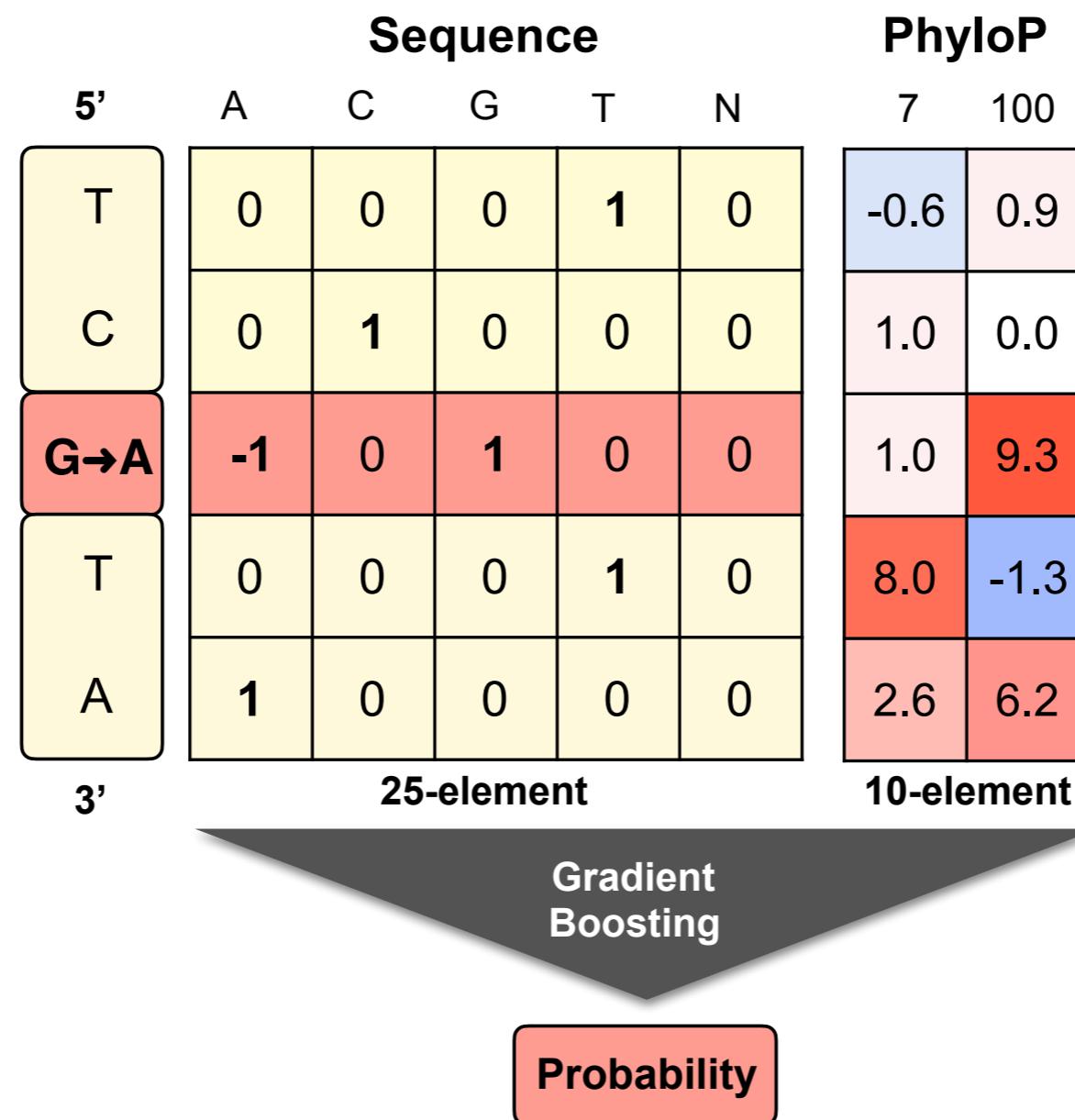
PhyloP100 score

Conservation analysis based on the pre-calculated score available at the UCSC revealed a **significant difference between the distribution of the PhyloP100 scores in Pathogenic and Benign SNVs.**



PhD-SNPG

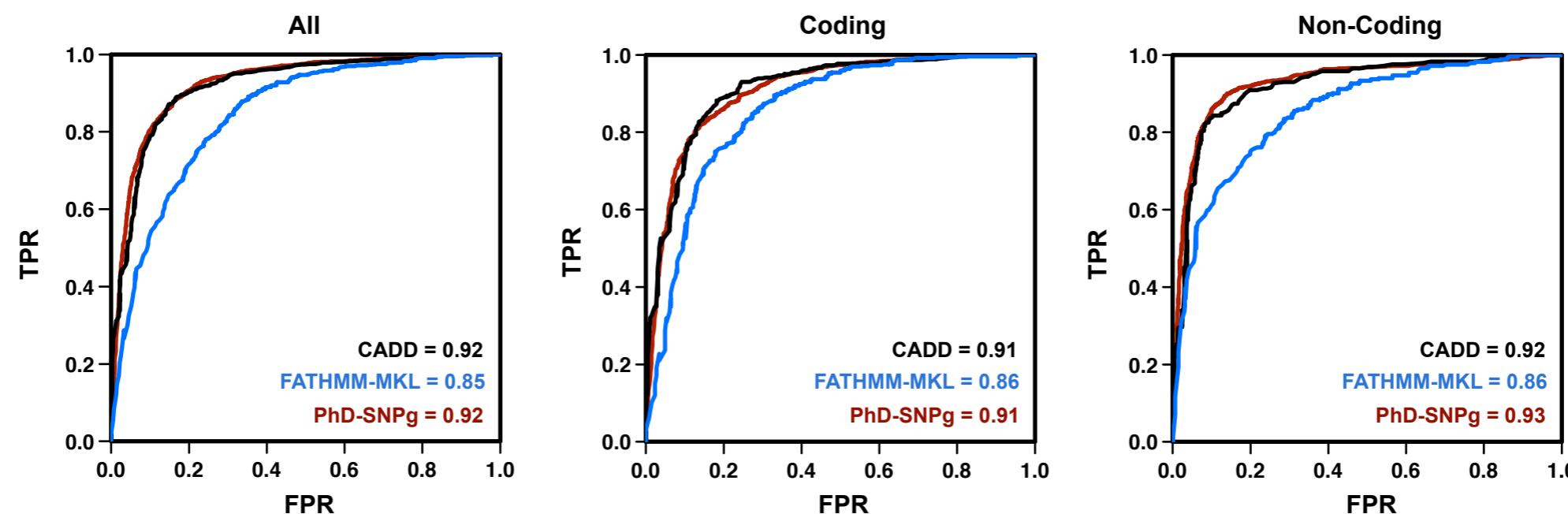
PhD-SNPG is a simple method that takes in input **35 sequence-based features** from a window of 5 nucleotides around the mutated position.



Benchmarking

PhD-SNP^g has been tested in cross-validation on a set of 35,802 SNVs and on a blind set of 1,408 variants recently annotated.

	Q2	TNR	NPV	TPR	PPV	MCC	F1	AUC
PhD-SNP^g	0.861	0.774	0.884	0.925	0.847	0.715	0.884	0.924
Coding	0.849	0.671	0.845	0.938	0.850	0.651	0.892	0.908
Non-Coding	0.876	0.855	0.911	0.901	0.839	0.753	0.869	0.930



Conclusions

- The machine learning methods based on sequence and structural information, trained to predict the sign and the value of $\Delta\Delta G$, reach a good level of accuracy.
- Evolutionary information are important for predicting deleterious variants. Wild-type residues in disease-related sites are more conserved than in neutral sites.
- Protein structure information improves performance of machine learning methods to discriminate between disease-causing and neutral variants.
- Nucleotide conservation is an important feature to predict the impact of SNVs in non coding regions

Acknowledgments

Structural Genomics @CNAG

Marc A. Marti-Renom

Davide Bau

David Dufour

Francois Serra

Bologna Biocomputing Group

Rita Casadio

Pier Luigi Martelli

University of Padova

Piero Fariselli

University of Camerino

Mario Compiani

Computational Biology and Bioinformatics Research Group (UIB)

Jairo Rocha

Division of Informatics at UAB

Rui Tian

Shivani Viradia

Malay Basu

Diego E. Penha

Helix Group (Stanford University)

Russ B. Altman

Jennifer Lahti

Tianyun Liu

Grace Tang

Other Collaborations

Yana Bromberg, Rutger University, NJ

Francisco Melo, Universidad Catolica, Chile

Sean Mooney, Buck Institute, Novato

Cedric Notredame, CRG Barcelona

Gustavo Parisi, Univesidad de Quilmes

Frederic Rousseau, KU Leuven

Joost Schymkowitz, KU Leuven

FUNDING

Startup funding Dept. of Pathology UAB

NIH:3R00HL111322-04S1 Co-Investigator

EMBO Short Term Fellowship

Marie Curie International Outgoing Grant

Marie Curie Reintegration Grant

Marco Polo Research Project

BIOSAPIENS Network of Excellence

SPINNER Consortium

Biomolecules, Folding and Disease

<http://biofold.org/>

