The effect of variations on protein stability: problems and issues

Piero Fariselli
Dept. Medical Sciences, University of Torino
piero.fariselli@unito.it

UNIVERSITÀ
DEGLI STUDI
DI TORINO

TESE

Traguardi di Eccellenza nelle Scienze mediche Esplorando le Omiche

Outline

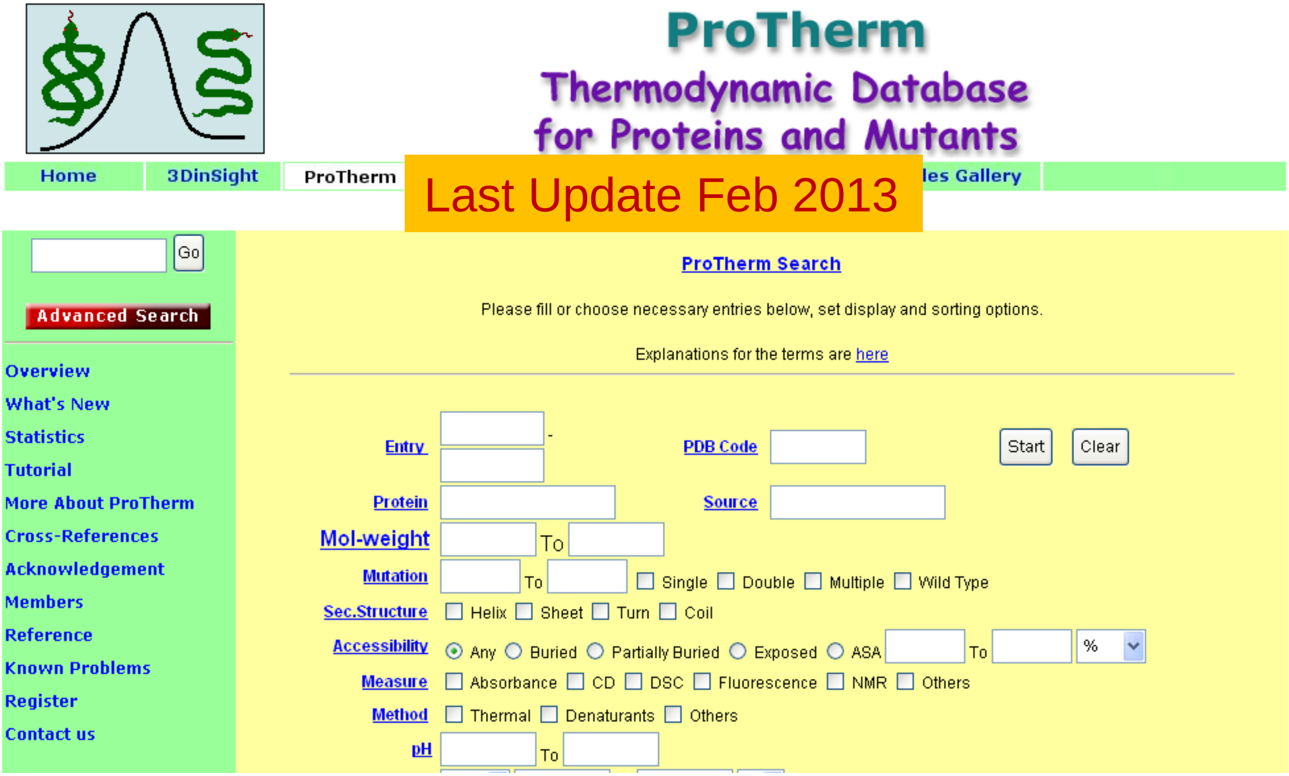
1. Dataset influence (*intrinsic to the data*)

2. Biases in the $\Delta\Delta G$ predictions (*dependent on the method design*)

3. Proper evaluation of the performance (*human behaviour*)

Datasets

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



Last Update Feb 2013

ProTherm Search

Please fill or choose necessary entries below, set display and sorting options.

Explanations for the terms are here

Entry - PDB Code Start Clear

Protein Source

Mol-weight To

Mutation To □ Single □ Double □ Multiple □ Wild Type

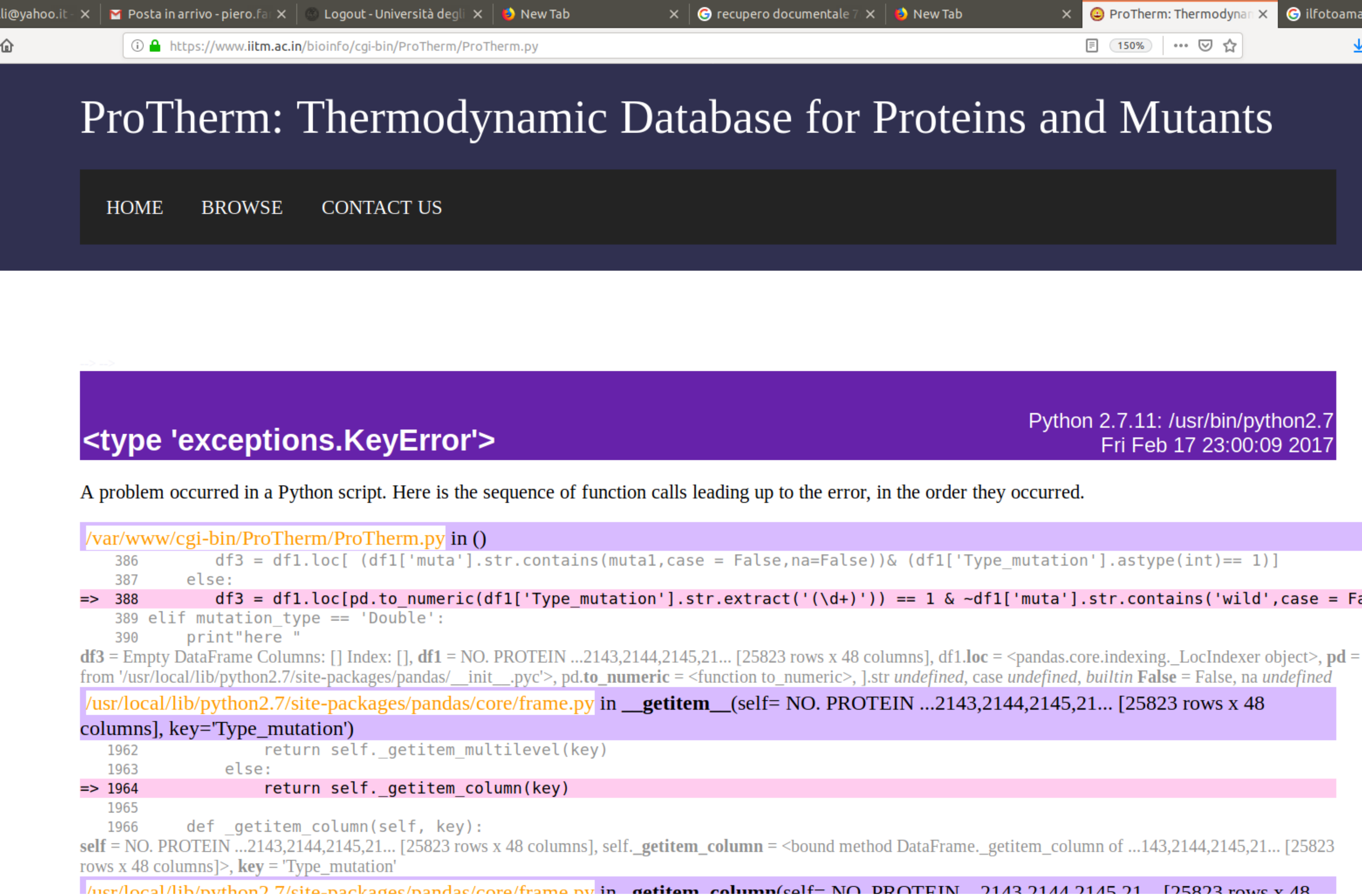
Sec.Structure □ Helix □ Sheet □ Turn □ Coil

Accessibility ● Any ○ Buried ○ Partially Buried ○ Exposed ○ ASA To %

Measure □ Absorbance □ CD □ DSC □ Fluorescence □ NMR □ Others

Method □ Thermal □ Denaturants □ Others

pH To


li@yahoo.it -
Posta in arrivo - piero.fa
Logout - Università degli
New Tab
recupero documentale 7
New Tab
ProTherm: Thermodynam
ilfotoama
https://www.iitm.ac.in/bioinfo/cgi-bin/ProTherm/ProTherm.py
150%

ProTherm: Thermodynamic Database for Proteins and Mutants

HOME BROWSE CONTACT US

--> -->

<type 'exceptions.KeyError'>

Python 2.7.11: /usr/bin/python2.7
Fri Feb 17 23:00:09 2017

A problem occurred in a Python script. Here is the sequence of function calls leading up to the error, in the order they occurred.

/var/www/cgi-bin/ProTherm/ProTherm.py in ()

```
    386         df3 = df1.loc[ (df1['muta'].str.contains(muta1,case = False,na=False))& (df1['Type_mutation'].astype(int)== 1)]
    387     else:
=>  388         df3 = df1.loc[pd.to_numeric(df1['Type_mutation'].str.extract('(\d+)')) == 1 & ~df1['muta'].str.contains('wild',case = Fa
    389 elif mutation_type == 'Double':
    390     print"here "
```

df3 = Empty DataFrame Columns: [] Index: [], **df1** = NO. PROTEIN ...2143,2144,2145,21... [25823 rows x 48 columns], df1.**loc** = <pandas.core.indexing._LocIndexer object>, **pd** = from '/usr/local/lib/python2.7/site-packages/pandas/__init__.pyc'>, pd.**to_numeric** = <function to_numeric>,].str *undefined*, case *undefined*, *builtin* **False** = False, na *undefined*

/usr/local/lib/python2.7/site-packages/pandas/core/frame.py in **__getitem__**(self= NO. PROTEIN ...2143,2144,2145,21... [25823 rows x 48 columns], key='Type_mutation')

```
    1962             return self._getitem_multilevel(key)
    1963         else:
=>  1964             return self._getitem_column(key)
    1965 
    1966     def _getitem_column(self, key):
```

self = NO. PROTEIN ...2143,2144,2145,21... [25823 rows x 48 columns], self.**_getitem_column** = <bound method DataFrame._getitem_column of ...143,2144,2145,21... [25823 rows x 48 columns]>, **key** = 'Type_mutation'

/usr/local/lib/python2.7/site-packages/pandas/core/frame.py in **_getitem_column**(self= NO. PROTEIN ...2143,2144,2145,21... [25823 rows x 48

1. **Dataset influence (*intrinsic to the data*)**

2. Biases in the ΔΔG predictions (*dependent on the method design*)

3. Proper evaluation of the performance (*human behaviour*)

May the *available* experimental measures affect the prediction performance?

Estimation of a predictor upper bound

An experimental measure of ΔΔG depends on several factors

- $\Delta\Delta G = f\,(\text{pH, T, salts, C\_i,C\_j,}\ldots)$

- In many cases we usually talk only of ΔΔG which is an average:
  $\Delta\Delta G = \sum_i \sum_j \sum_k \ldots \sum_n \Delta\Delta G(\text{i,j,k,}\ldots\text{n})$

- Sometimes we consider the dependence of pH and T:
  $\Delta\Delta G = f\,(\text{pH,T})$

Problem when we compare different measures of the same variation

Examples

1. In Keeler et al. (2009) the variation H180A in the human prolactin (pdb code 2Q98) measured at T=25°C, but different pH,
   $\Delta\Delta G$ = 1.39 kcal/mol at pH=5.8;
   $\Delta\Delta G$ = -0.04 at pH=7.8
   .
2. In Gribenko and Makhatadze 2007, the variation E3R in protein 1CSP, 6 different $\Delta\Delta G$ values ranging from 1.4 kcal/mol to 2.4 kcal/mol were measured at the same temperature (55°C) and pH (7.5) as function of different salt concentrations.

3. In Ferguson and Shaw (2002) the variant L3S of the calcium-binding protein S100B (1UWO) measured in two different starting conditions and techniques, but at the same temperature (25°C) and pH (7.2) yielded two
   $\Delta\Delta G$ = 1.91kcal/mol and
   $\Delta\Delta G$ = -2.77kcal/mol

Given the dataset (σ_{DB}) and the measure uncertainty (σ) is there an upper bound to the prediction performance?

Theoretical of estimation of an upper bound: a “*Gedankenexperiment*”

Given N protein variations, we may think to perform a set of N pairs of experiments ($\{x_i\}$, $\{y_i\}$), two for each variation.

Then we use one set of ΔΔG measures as “predictor” and the other as a set of experimental measures.

The idea is that, given the experimental condition, the best possible predictor is another set of experimental data (considering the experimental uncertainty)

Theoretical estimation

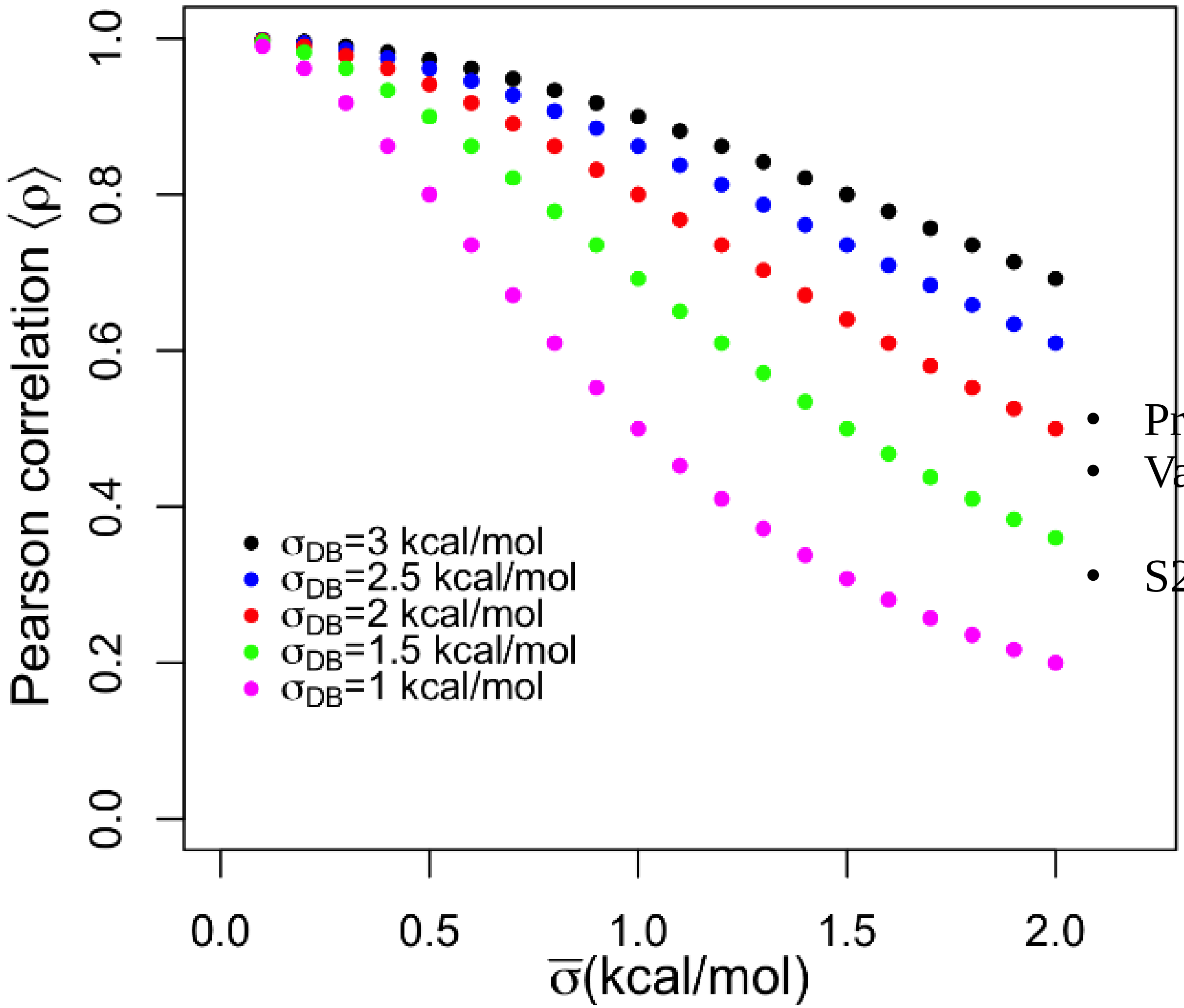
The Pearson’s correlation:

$$\langle \rho \rangle \cong \frac{\langle \sigma_{xy} \rangle}{\langle {\sigma_x}^2 \rangle} \frac{{\sigma_{DB}}^2}{\overline{\sigma^2} + {\sigma_{DB}}^2} = \frac{1}{1 + \left(\frac{\overline{\sigma^2}}{\sigma_{DB}^2} \right)}$$

The upper bound of the Coefficient of determination (*R*^{2}) is even lower than the Pearson with

$$R^2_{ub} = \langle R^2 \rangle = 1 - \langle S_e \rangle / \langle St \rangle \approx \frac{\sigma^2_{DB} - \overline{\sigma}^2}{\sigma^2_{DB} + \overline{\sigma}^2} = \frac{1 - \overline{\sigma}^2 / \sigma^2_{DB}}{1 + \overline{\sigma}^2 / \sigma^2_{DB}}$$

Expected Pearson correlation <ρ> vs. data average uncertainty ($\bar{\sigma}$) for different values of dataset standard deviation σ_{DB}



- ProTherm 2.06 kcal/mol
- Varibench 1.91 kcal/mol

- S2648 1.47 kcal/mol

Experimental Datasets

From

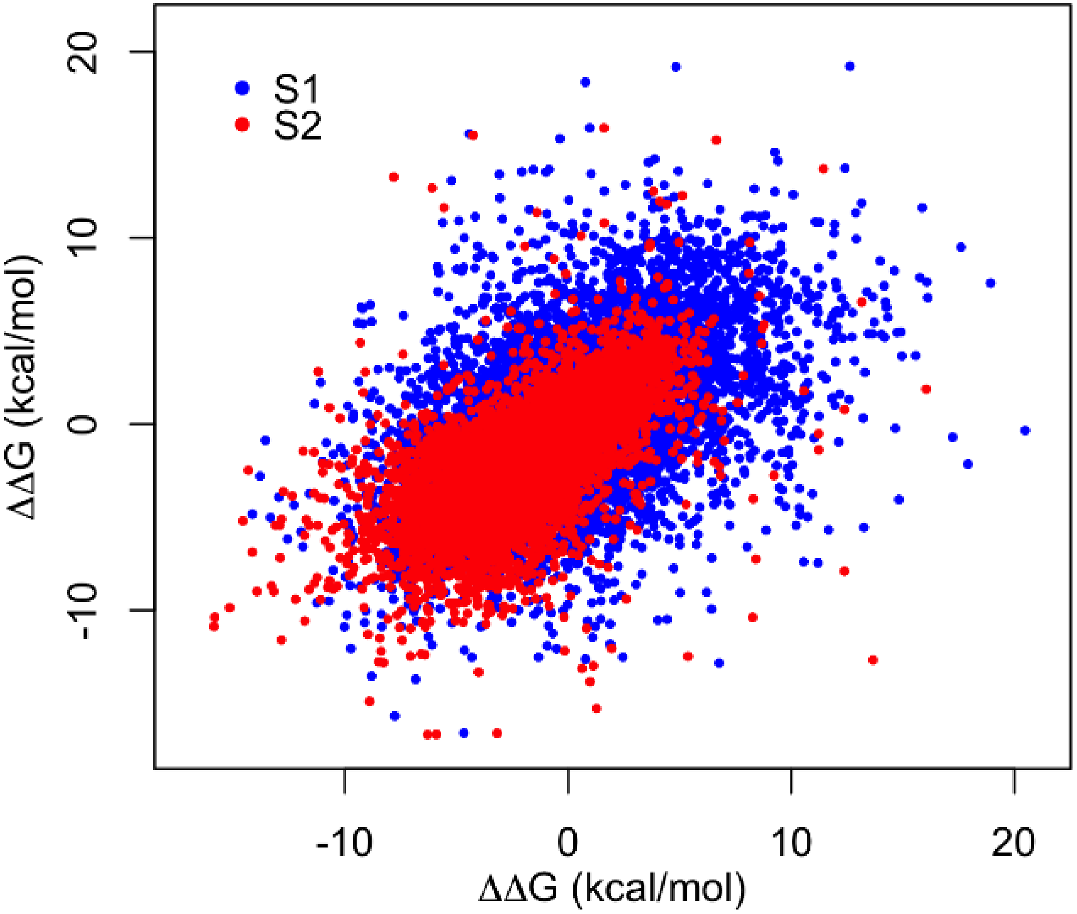
$$\langle \rho \rangle = \frac{1}{1 + \left(\frac{\overline{\sigma^2}}{\sigma_{DB}^2} \right)}$$

and using the experimental data we have

- S1: Theoretical estimation with $\bar{\sigma} = 1.04$ and σ_{DB}=1.72
  => R= 0.73

- S2: Theoretical estimation with $\bar{\sigma} = 0.72$ and σ_{DB}=1.57
  => R= 0.83

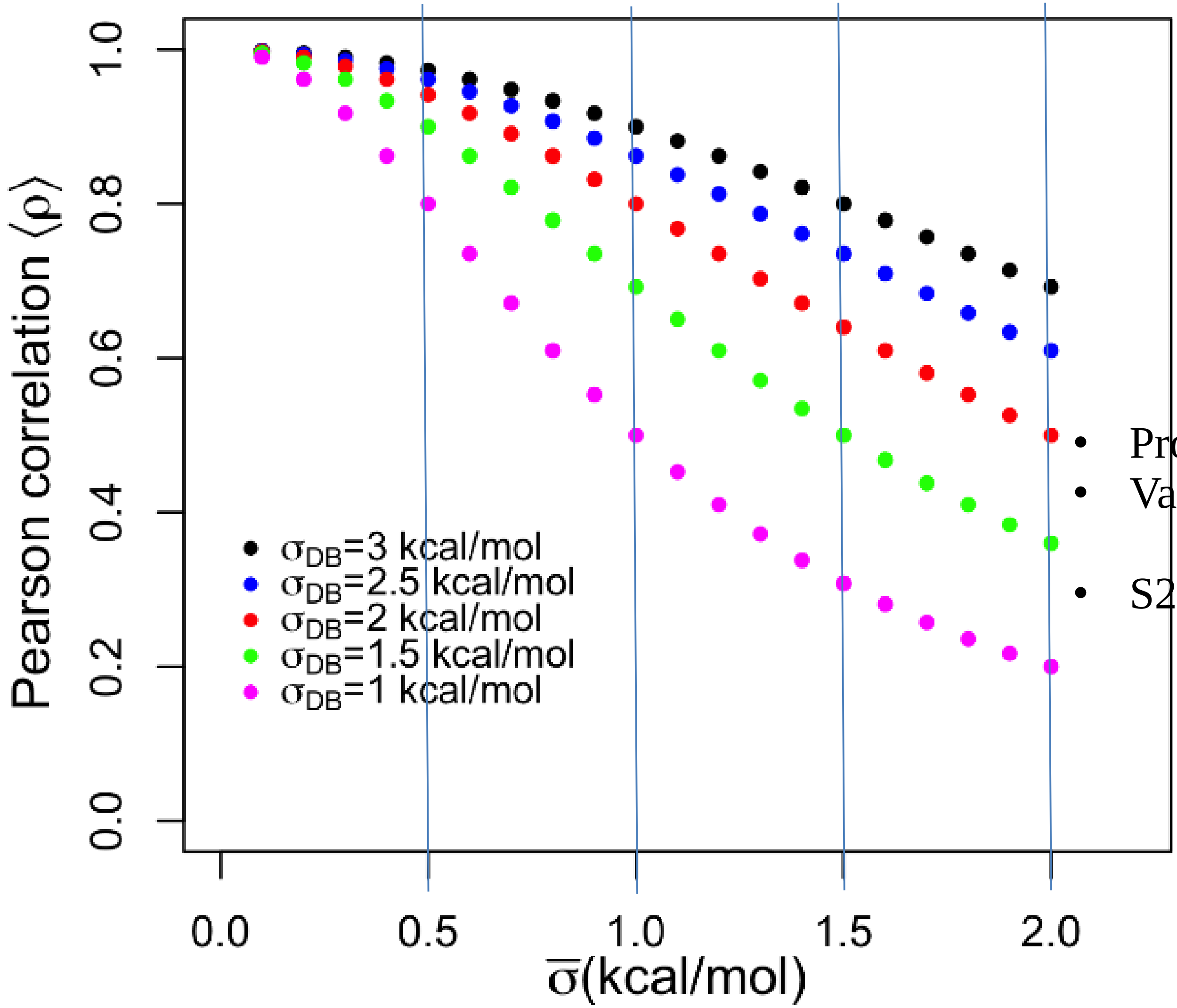
Simulation with the experimental Datasets



Scatterplot of two randomly generated observations for a given variation.

After 100 runs the Pearson correlation are
S1-> 0.74 ± 0.02
S2 -> 0.84 ± 0.02

Multiple Mutations ?



- ProTherm 2.06 kcal/mo
- Varibench 1.91 kcal/mo
- S2648 1.47 kcal/mol

- Comparison among methods on different datasets
- Performance of a method on different datasets
- Evaluate method over-fitting
- Effect on multiple mutations

1. Dataset influence (*intrinsic to the data*)

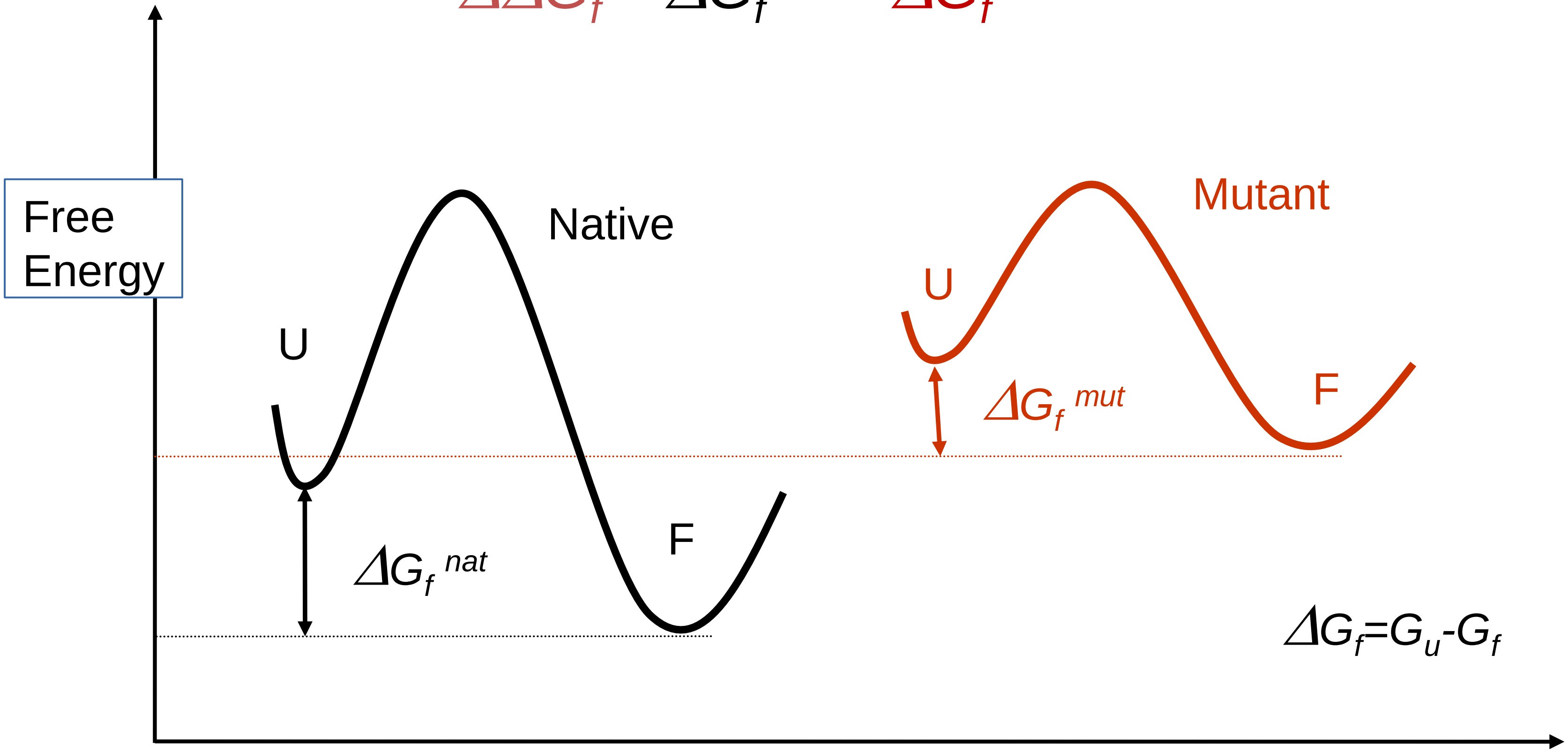
2. **Biases in the ΔΔG predictions (*dependent on the method design*)**

3. Proper evaluation of the performance (*human behaviour*)

Biases in ΔΔG predictions

If we change Alanine 35 with a Leucine,
is the protein stability *Increased or Decreased*?

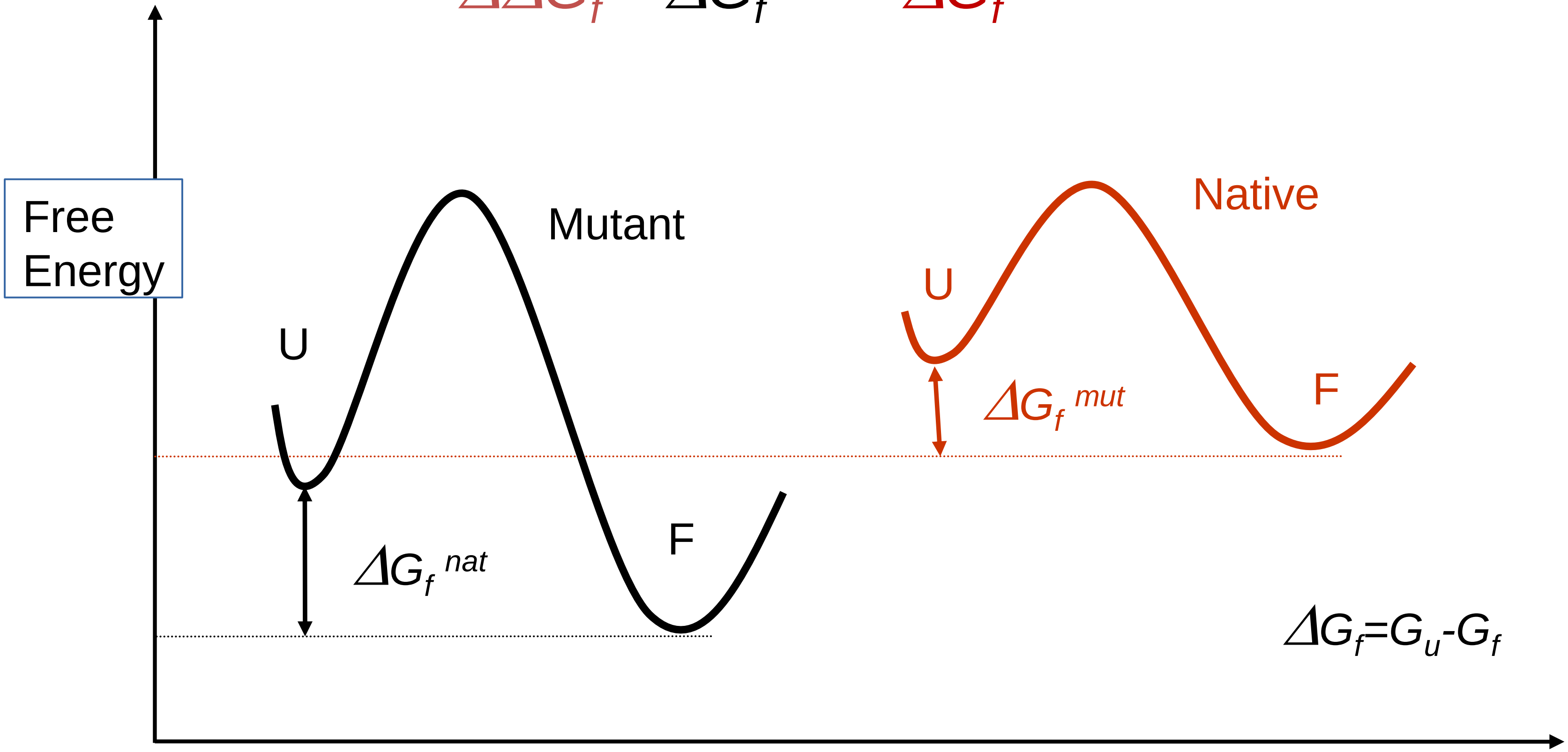
$$\Delta\Delta G_f = \Delta G_f^{\,nat} - \Delta G_f^{\,mut}$$



$$\Delta G_f = G_u - G_f$$

If we change Alanine 35 with a Leucine,
is the protein stability *Increased or Decreased*?

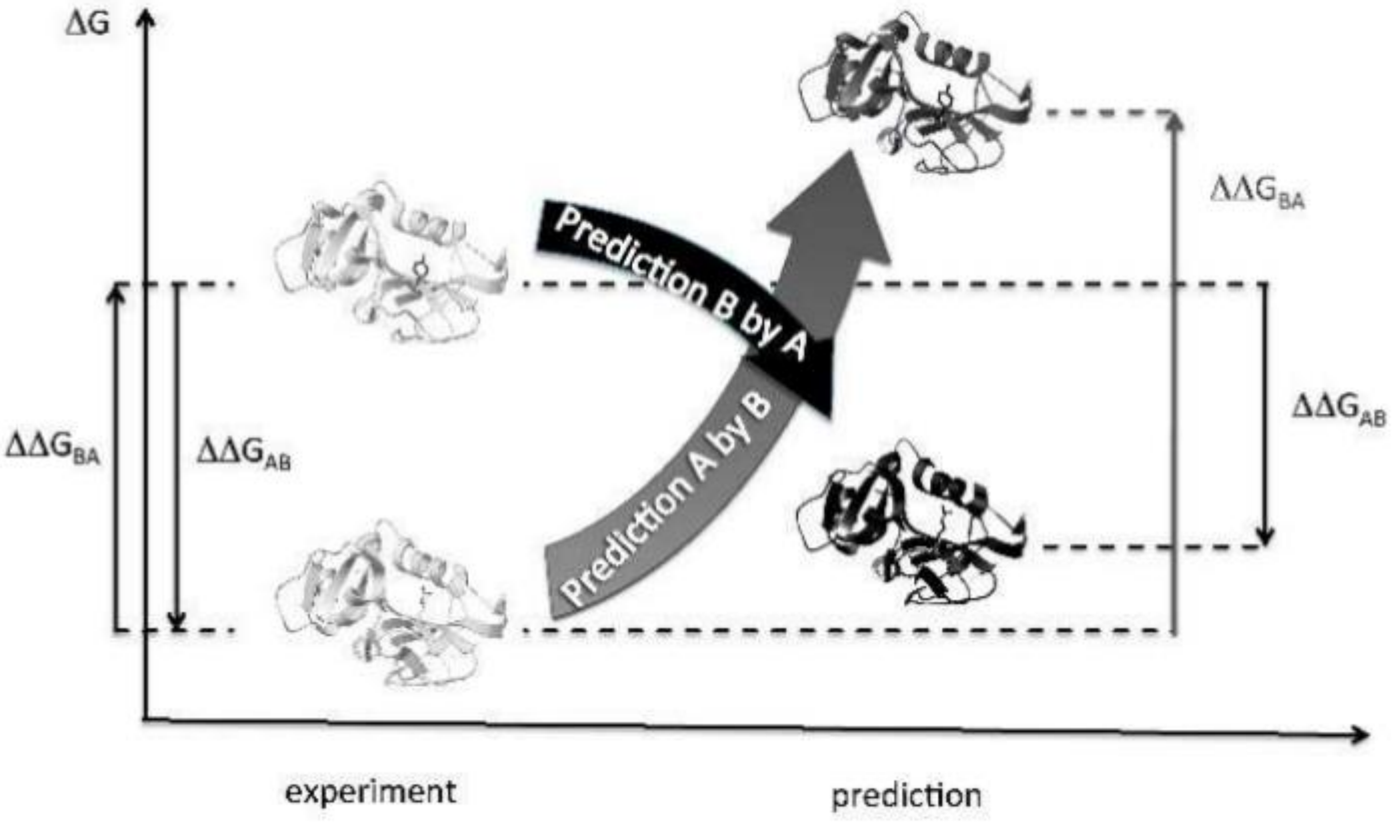
$$\Delta\Delta G_f = \Delta G_f^{\ mut} - \Delta G_f^{\ nat}$$



$$\Delta G_f = G_u - G_f$$

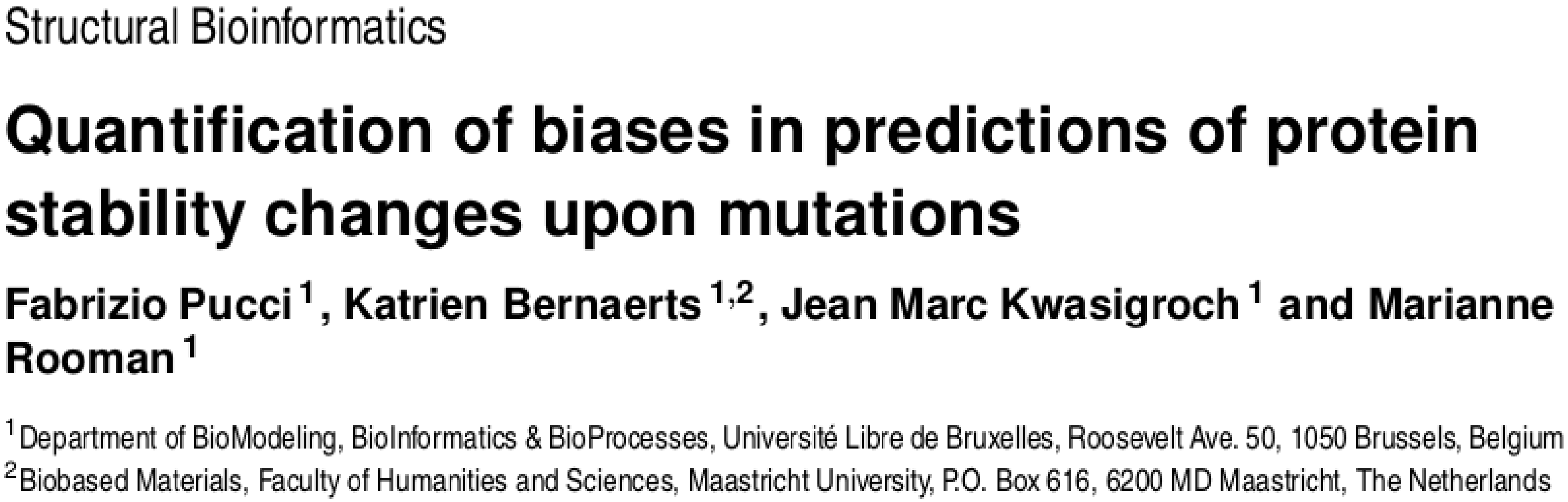
Protein A = Protein B with variation Y25R
Protein B = Protein A with variation R25Y

$$\Delta\Delta G_{AB} = - \Delta\Delta G_{BA}$$



Usmanova,D.R., *et al*. (2018)

Biases in ΔΔG predictions

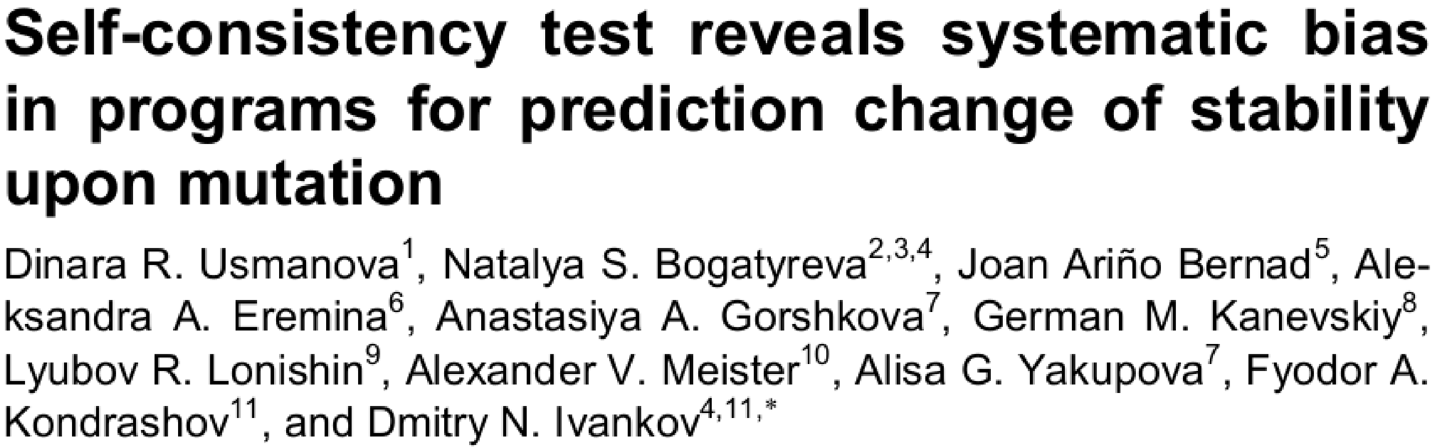


Structural Bioinformatics

Quantification of biases in predictions of protein stability changes upon mutations

Fabrizio Pucci[1], Katrien Bernaerts[1,2], Jean Marc Kwasigroch[1] and Marianne Rooman[1]

[1] Department of BioModeling, BioInformatics & BioProcesses, Université Libre de Bruxelles, Roosevelt Ave. 50, 1050 Brussels, Belgium
[2] Biobased Materials, Faculty of Humanities and Sciences, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands



Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation

Dinara R. Usmanova[1], Natalya S. Bogatyreva[2,3,4], Joan Ariño Bernad[5], Aleksandra A. Eremina[6], Anastasiya A. Gorshkova[7], German M. Kanevskiy[8], Lyubov R. Lonishin[9], Alexander V. Meister[10], Alisa G. Yakupova[7], Fyodor A. Kondrashov[11], and Dmitry N. Ivankov[4,11,*]

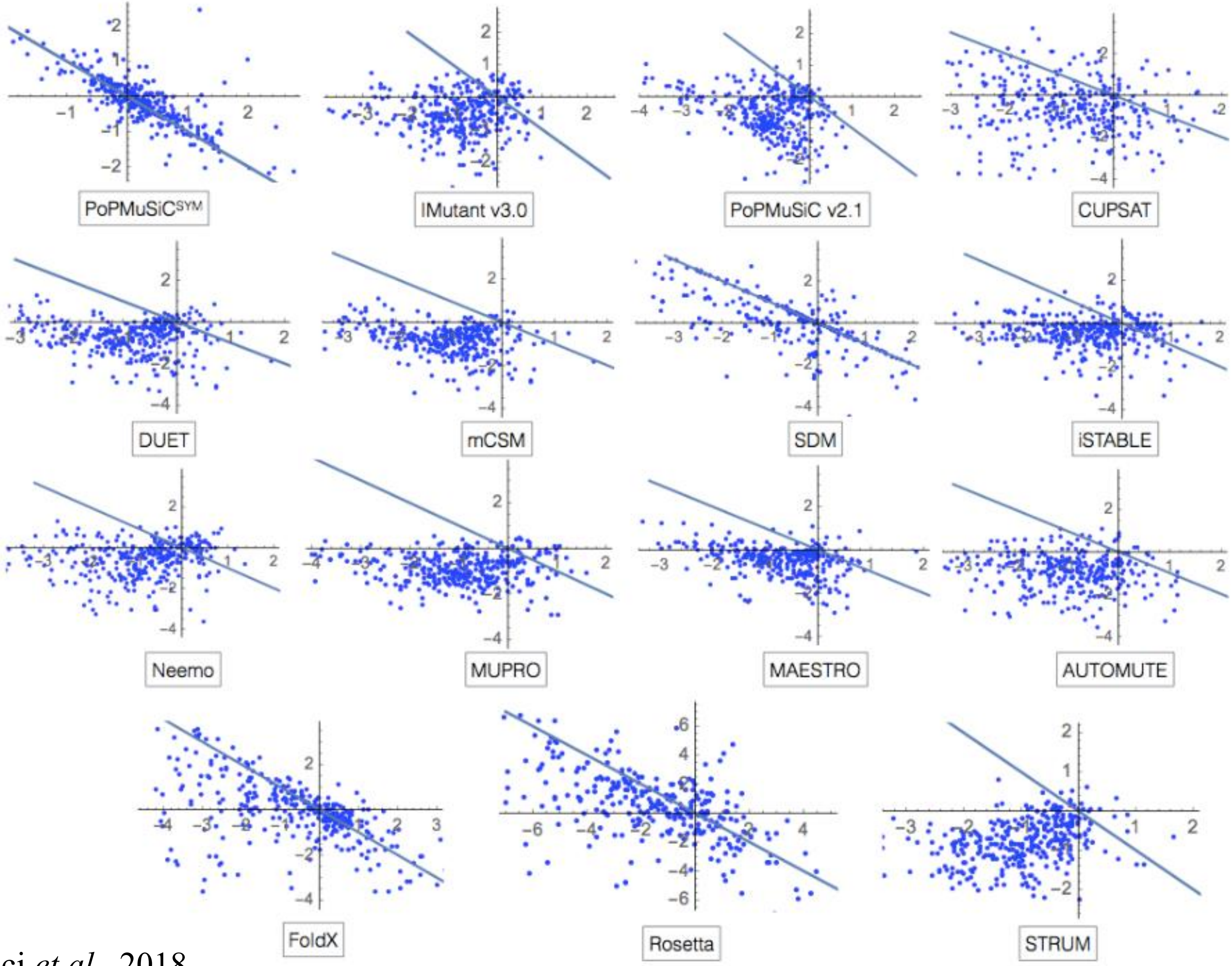
Biases in ΔΔG predictions

Structural Bioinformatics

Quantification of biases in predictions of protein stability changes upon mutations

Fabrizio Pucci[1], Katrien Bernaerts[1,2], Jean Marc Kwasigroch[1] and Marianne Rooman[1]

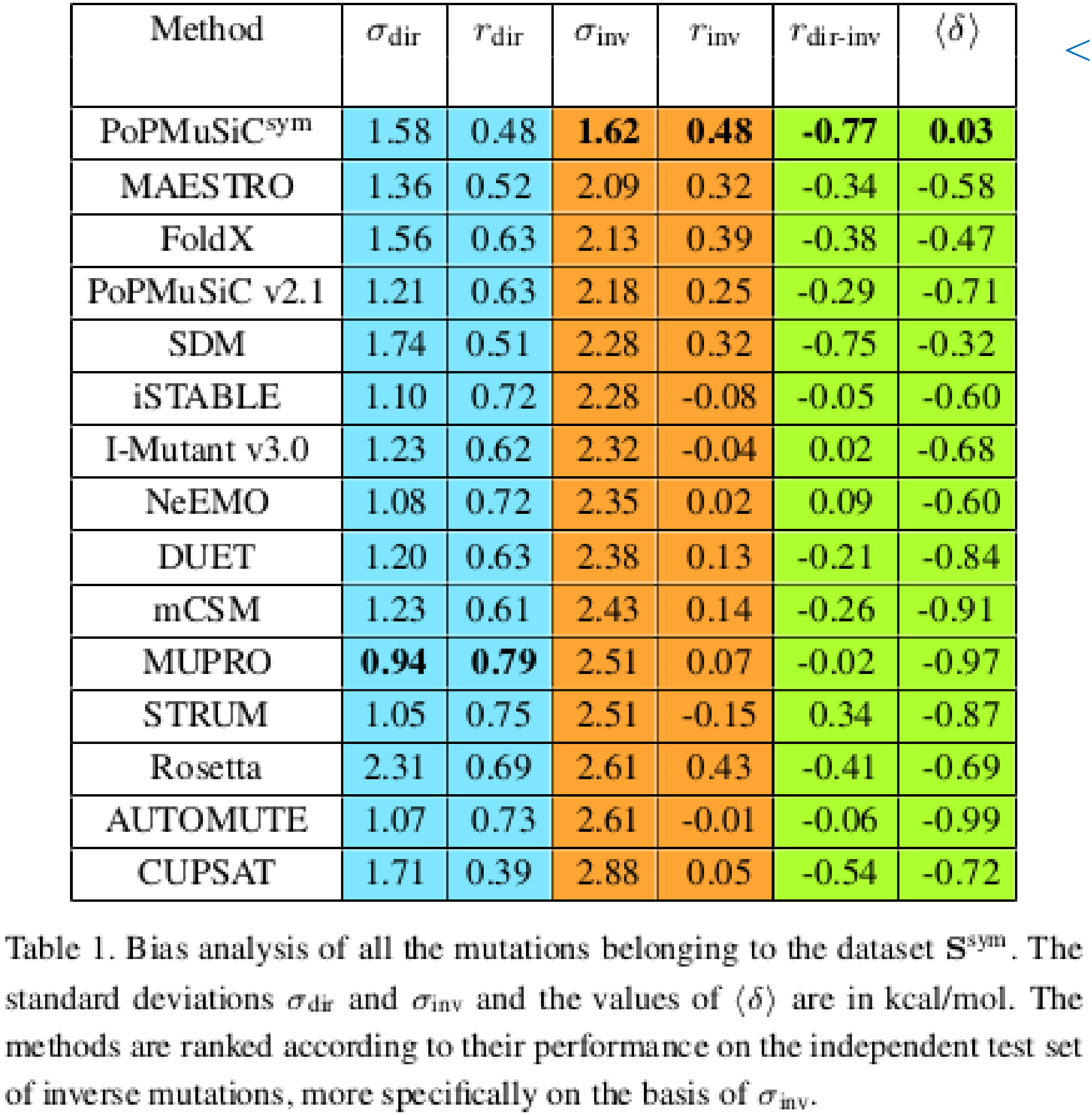
- The **Ssym** dataset is a manually curated selection of variations from the ProTherm database.
- It contains mutations with experimental ΔΔG values for which the 3D structures of both the wild-type and variant proteins were solved by X-ray crystallography.
- **Ssym** consists of 684 variations, 342 are direct (reported in the literature) and 342 are obtained by anti-symmetry, and associated to the variant PDB structure



Pucci *et al.*, 2018

Quantification of biases in predictions of protein stability changes upon mutations

Fabrizio Pucci[1], Katrien Bernaerts[1,2], Jean Marc Kwasigroch[1] and Marianne Rooman[1]

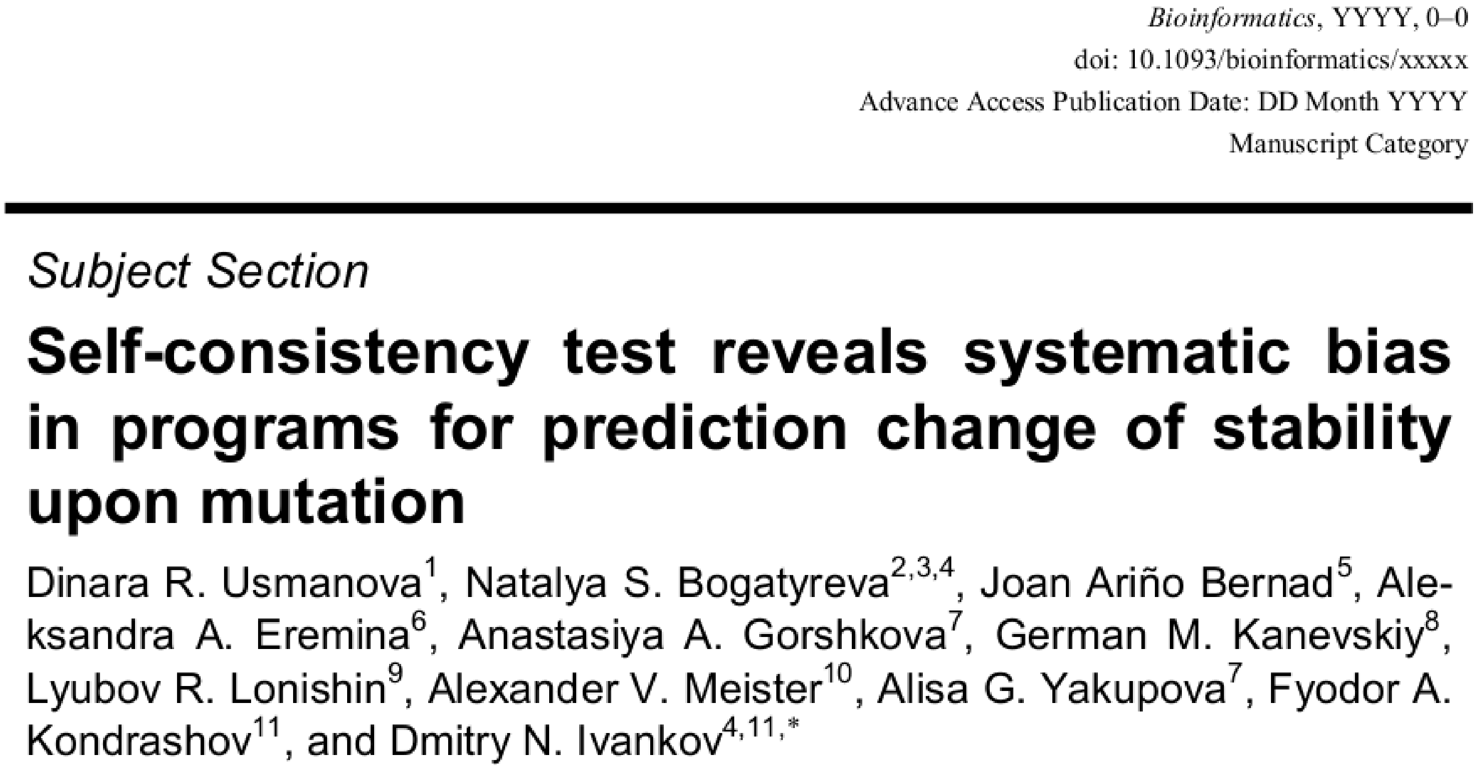


Method	σ_{dir}	r_{dir}	σ_{inv}	r_{inv}	$r_{\text{dir-inv}}$	$\langle\delta\rangle$
PoPMuSiC$^{\text{sym}}$	1.58	0.48	**1.62**	**0.48**	**-0.77**	**0.03**
MAESTRO	1.36	0.52	2.09	0.32	-0.34	-0.58
FoldX	1.56	0.63	2.13	0.39	-0.38	-0.47
PoPMuSiC v2.1	1.21	0.63	2.18	0.25	-0.29	-0.71
SDM	1.74	0.51	2.28	0.32	-0.75	-0.32
iSTABLE	1.10	0.72	2.28	-0.08	-0.05	-0.60
I-Mutant v3.0	1.23	0.62	2.32	-0.04	0.02	-0.68
NeEMO	1.08	0.72	2.35	0.02	0.09	-0.60
DUET	1.20	0.63	2.38	0.13	-0.21	-0.84
mCSM	1.23	0.61	2.43	0.14	-0.26	-0.91
MUPRO	**0.94**	**0.79**	2.51	0.07	-0.02	-0.97
STRUM	1.05	0.75	2.51	-0.15	0.34	-0.87
Rosetta	2.31	0.69	2.61	0.43	-0.41	-0.69
AUTOMUTE	1.07	0.73	2.61	-0.01	-0.06	-0.99
CUPSAT	1.71	0.39	2.88	0.05	-0.54	-0.72

Table 1. Bias analysis of all the mutations belonging to the dataset S^{sym}. The standard deviations σ_{dir} and σ_{inv} and the values of $\langle\delta\rangle$ are in kcal/mol. The methods are ranked according to their performance on the independent test set of inverse mutations, more specifically on the basis of σ_{inv}.

$<\delta> = \Delta\Delta G_{AB} + \Delta\Delta G_{BA}$

Biases in ΔΔG predictions



Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

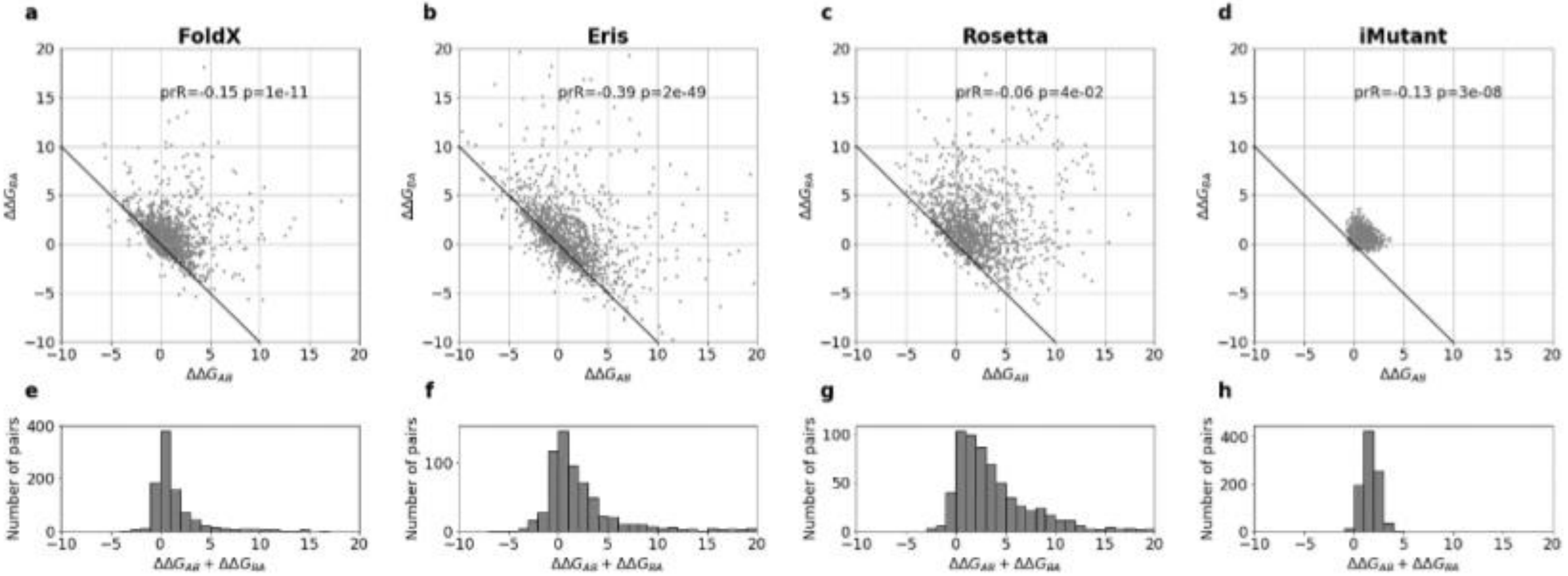
Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation

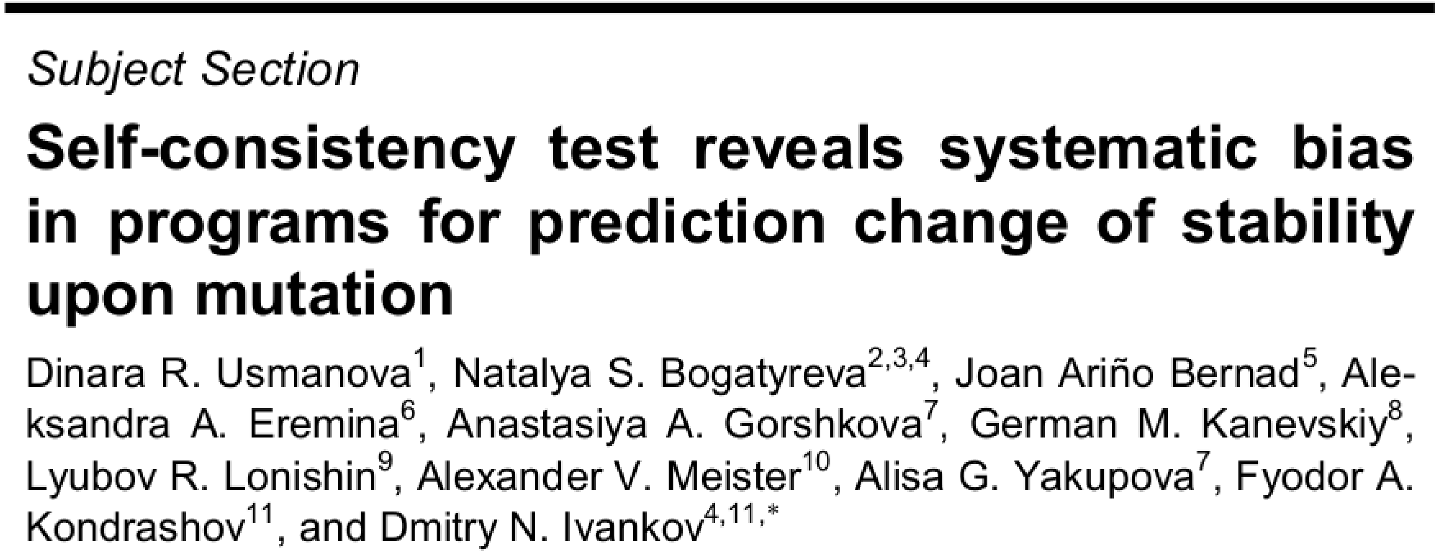
Dinara R. Usmanova[1], Natalya S. Bogatyreva[2,3,4], Joan Ariño Bernad[5], Aleksandra A. Eremina[6], Anastasiya A. Gorshkova[7], German M. Kanevskiy[8], Lyubov R. Lonishin[9], Alexander V. Meister[10], Alisa G. Yakupova[7], Fyodor A. Kondrashov[11], and Dmitry N. Ivankov[4,11,*]

Biases in ΔΔG predictions

- dataset was built by Usmanova *et al.* 2018, by extracting high-resolution pairs of proteins from the Protein Data Bank (PDB) differing by one to ten amino acids.

- Large datasets, with 1000 pairs of protein structures differing by one residue


a
FoldX
prR=-0.15 p=1e-11
b
Eris
prR=-0.39 p=2e-49
c
Rosetta
prR=-0.06 p=4e-02
d
iMutant
prR=-0.13 p=3e-08
$\Delta\Delta G_{BA}$
$\Delta\Delta G_{AB}$
20
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20
e
f
g
h
Number of pairs
400
200
0
100
50
$\Delta\Delta G_{AB} + \Delta\Delta G_{BA}$



Subject Section

Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation

Dinara R. Usmanova[1], Natalya S. Bogatyreva[2,3,4], Joan Ariño Bernad[5], Aleksandra A. Eremina[6], Anastasiya A. Gorshkova[7], German M. Kanevskiy[8], Lyubov R. Lonishin[9], Alexander V. Meister[10], Alisa G. Yakupova[7], Fyodor A. Kondrashov[11], and Dmitry N. Ivankov[4,11,*]

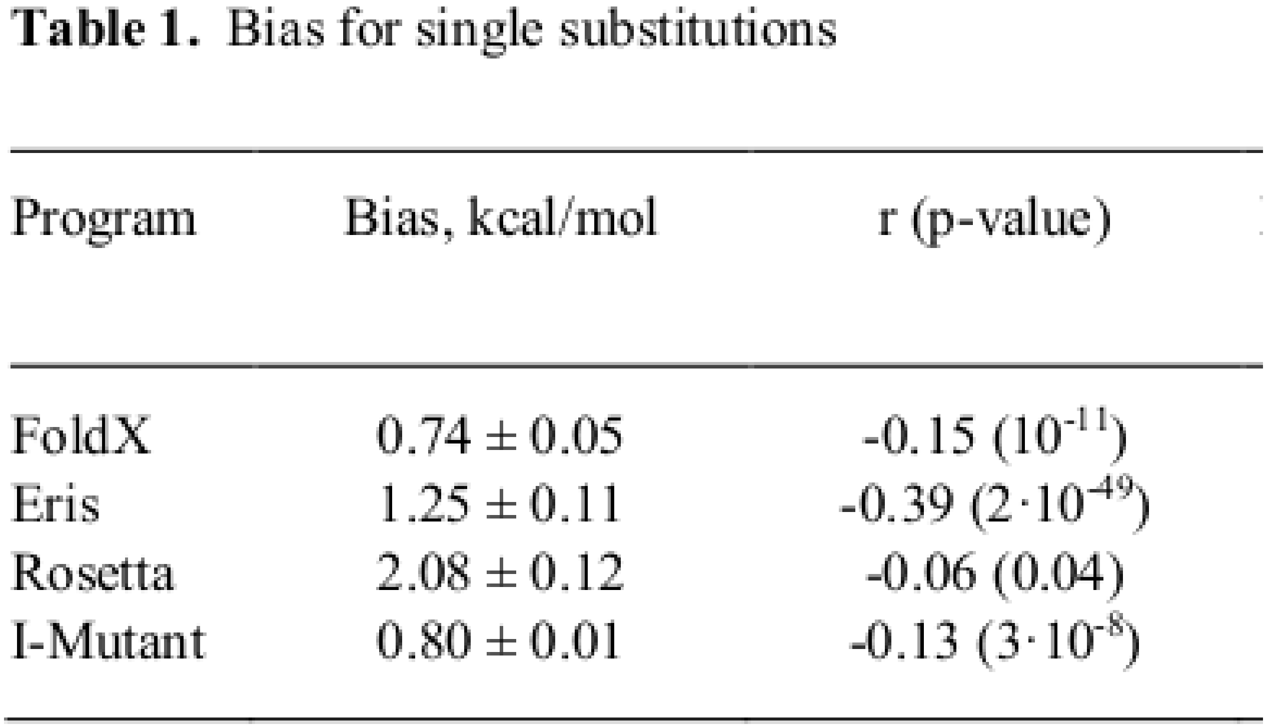


Table 1. Bias for single substitutions

Program	Bias, kcal/mol	r (p-value)
FoldX	0.74 ± 0.05	-0.15 (10^{-11})
Eris	1.25 ± 0.11	-0.39 ($2 \cdot 10^{-49}$)
Rosetta	2.08 ± 0.12	-0.06 (0.04)
I-Mutant	0.80 ± 0.01	-0.13 ($3 \cdot 10^{-8}$)

$$\text{Bias} = (\Delta\Delta G_{AB} + \Delta\Delta G_{BA})/2$$

An important property that a predictor has to fulfil is the “variation” anti-symmetry :

$$\Delta\Delta G_{AB} = - \Delta\Delta G_{BA}$$

DDGun: DDG Untrained baseline method

A way to implement the predictor anti-symmetry is to provide in input to it only anti-symmetric features

DDGun: DDG Untrained baseline method

Assuming that the profile *p* does not change for the mutant and the wild type protein sequence, we can compute some feature scores such as

Evolutionary
(*B*=Blosum62)

$$s_{Bl} = \sum_{i=1}^{20} p(a_i)\big(B(a_i, m) - B(a_i, w)\big)$$

Skolnick (P_{Sk})
Local potential

$$s_{Sk} = \sum_{j=-2\ j\neq 0}^{j+2} \sum_{i=1}^{20} p(a_j)\big(P_{Sk}(w, a_i) - P_{Sk}(m, a_i)\big)$$

Hydrophobicity (*K*)

$$s_{Hp} = p(m)K(m) - p(w)K(w)$$

3D contact potential (P_{BV})

$$s_{BV} = \sum_{j\in I} \sum_{i=1}^{20} p(a_{ij})\big(P_{BV}(w, a_i) - P_{BV}(m, a_i)\big)$$

DDGun: DDG Untrained baseline method

Anti-symmetry performances of DDGun on the Ssym data set (Pucci et al, 2018, PopMusicSym)

	Performances		Anti-symmetry	
Method	Direct variations Pearson r, RMSE	Inverse variations Pearson r, RMSE	$r_{dir-inv}$	Bias <δ> (kcal/mol)
DDGun	0.48, 1.47	0.48, 1.50	-0.99	-0.007
DDGun3D	0.56, 1.42	0.53, 1.46	-0.99	-0.02
PopMusicSym	0.48, 1.58	0.48, 1.62	-0.77	0.03
SDM	0.51, 1.74	0.32, 2.28	-0.75	-0.32
Maestro	0.52, 1.36	0.32, 2.09	-0.34	-0.58
FoldX	0.63, 1.56	0.39, 2.13	-0.38,	-0.47

DDGun: DDG Untrained baseline method

Performances on the 914 multiple site variation from Protherm.

	Performances			Anti-symmetry	
Method	Direct and Inverse Pearson r, RMSE	Direct variations Pearson r, RMSE	Inverse variations Pearson r, RMSE	$r_{dir-inv}$	Bias <δ> (kcal/mol)
DDGun	0.44, 2.23	0.37, 2.23	0.37, 2.23	-1.00	0.00
DDGun3D	0.45, 2.27	0.39, 2.24	0.38, 2.25	-0.99	-0.007
Maestro	0.30, 2.59	0.55, 1.96	0.08, 3.10	-0.20	-0.92
FoldX	0.44, 3.10	0.41, 2.95	0.33, 3.24	-0.71	-0.21

1. Dataset influence (*intrinsic to the data*)

2. Biases in the ΔΔG predictions (*dependent on the method design*)

3. Proper evaluation of the performance (*human behaviour*)

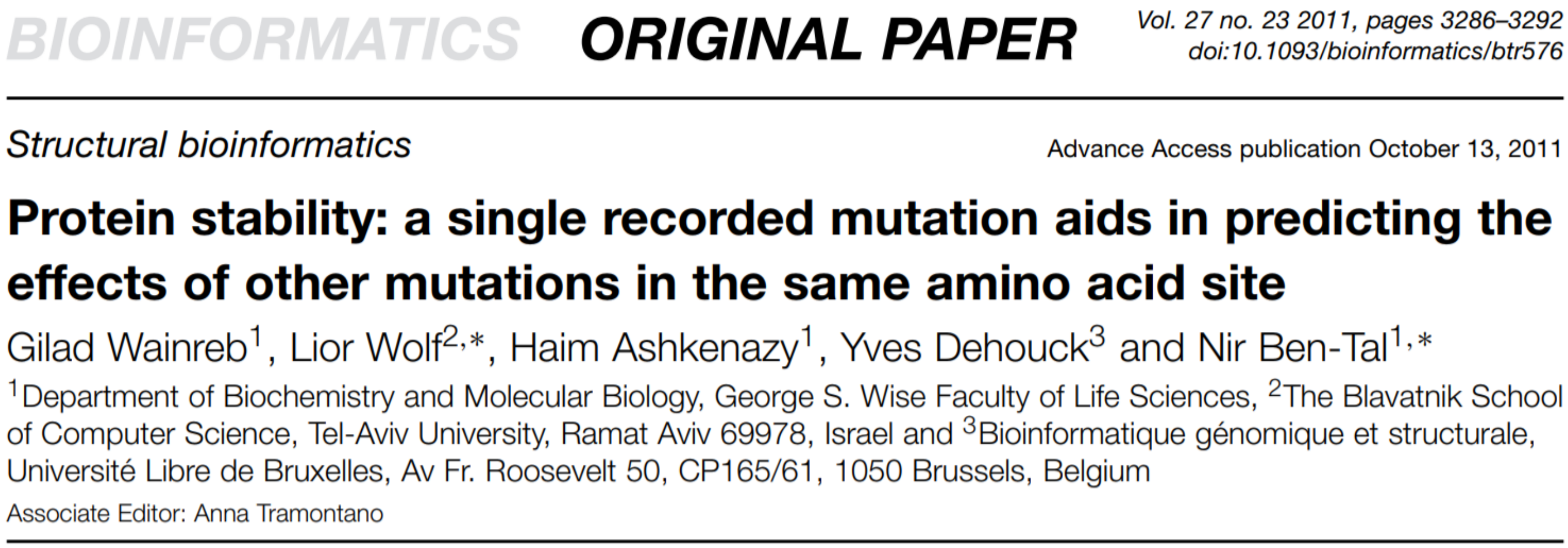
Evaluation problems:

1. many Variations in the same (similar) protein
2. many Variations in the same protein position

Classical mistake: random partition of training and testing sets to fit the parameters or train models

I

Problem of similarity between training and testing sets



BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 23 2011, pages 3286–3292
doi:10.1093/bioinformatics/btr576

Structural bioinformatics

Advance Access publication October 13, 2011

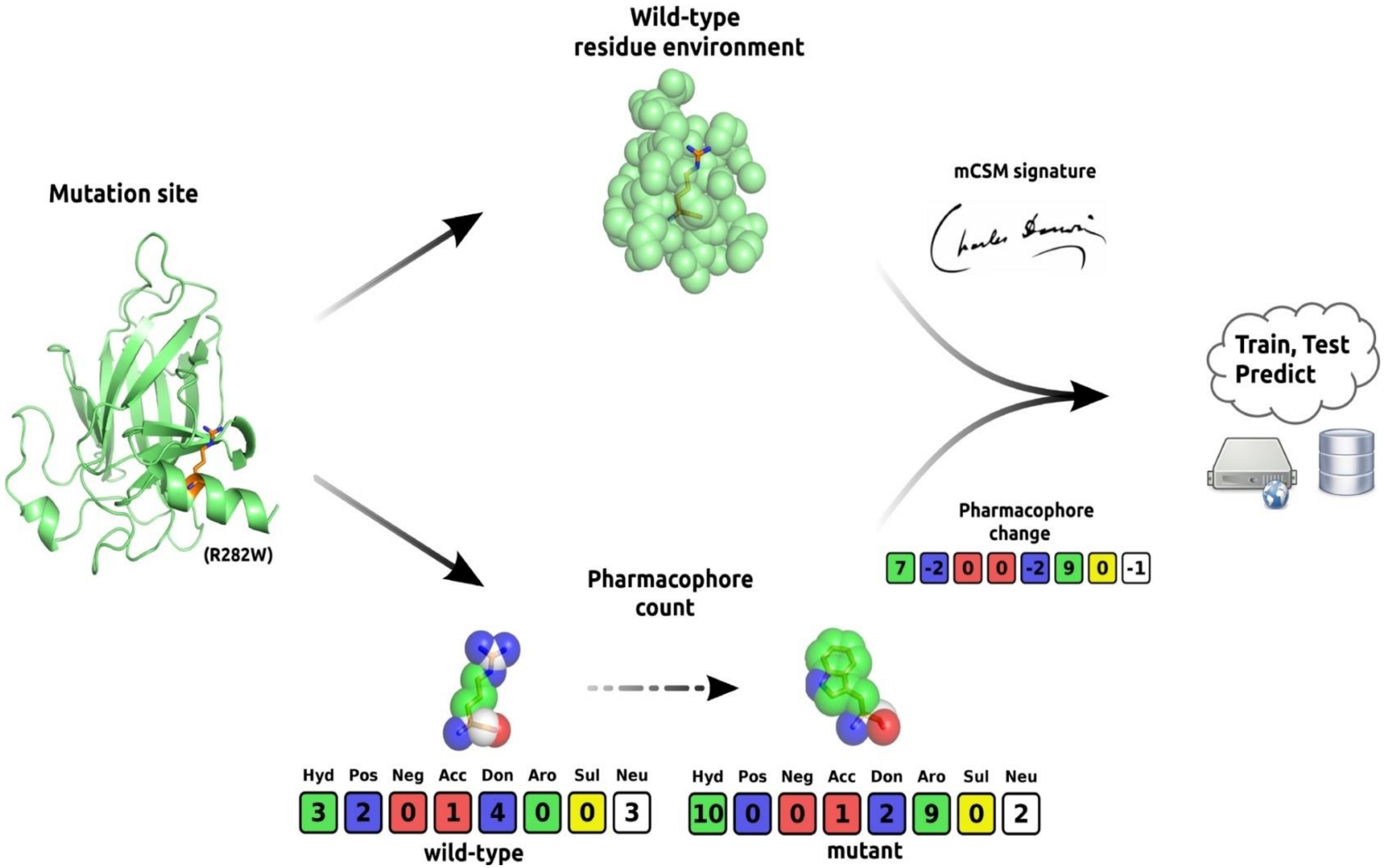
Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site

Gilad Wainreb[1], Lior Wolf[2,*], Haim Ashkenazy[1], Yves Dehouck[3] and Nir Ben-Tal[1,*]

[1]Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, [2]The Blavatnik School of Computer Science, Tel-Aviv University, Ramat Aviv 69978, Israel and [3]Bioinformatique génomique et structurale, Université Libre de Bruxelles, Av Fr. Roosevelt 50, CP165/61, 1050 Brussels, Belgium

Associate Editor: Anna Tramontano

The case of mCSM



Pharmacophore change

7	-2	0	0	-2	9	0	-1

wild-type

Hyd	Pos	Neg	Acc	Don	Aro	Sul	Neu
3	2	0	1	4	0	0	3

mutant

Hyd	Pos	Neg	Acc	Don	Aro	Sul	Neu
10	0	0	1	2	9	0	2

(*) mCSM: predicting the effect of mutations in proteins using graph-based signatures

Douglas E. V. Pires, David B. Ascher, Tom L. Blundell, 2014

mCSM

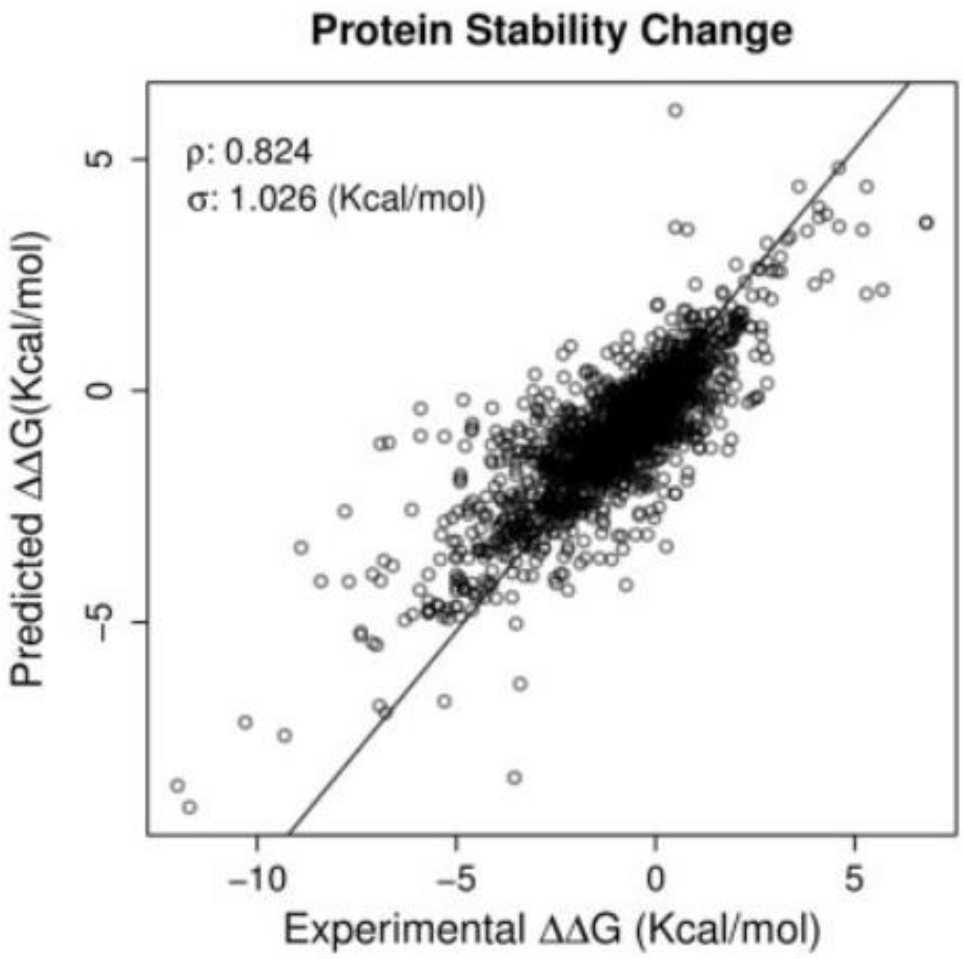


Fig. 2. Regression results for mCSM signature pred

main paper results

(*) mCSM: predicting the effect of mutations in proteins using graph-based signatures
Douglas E. V. Pires, David B. Ascher, Tom L. Blundell, 2014

mCSM

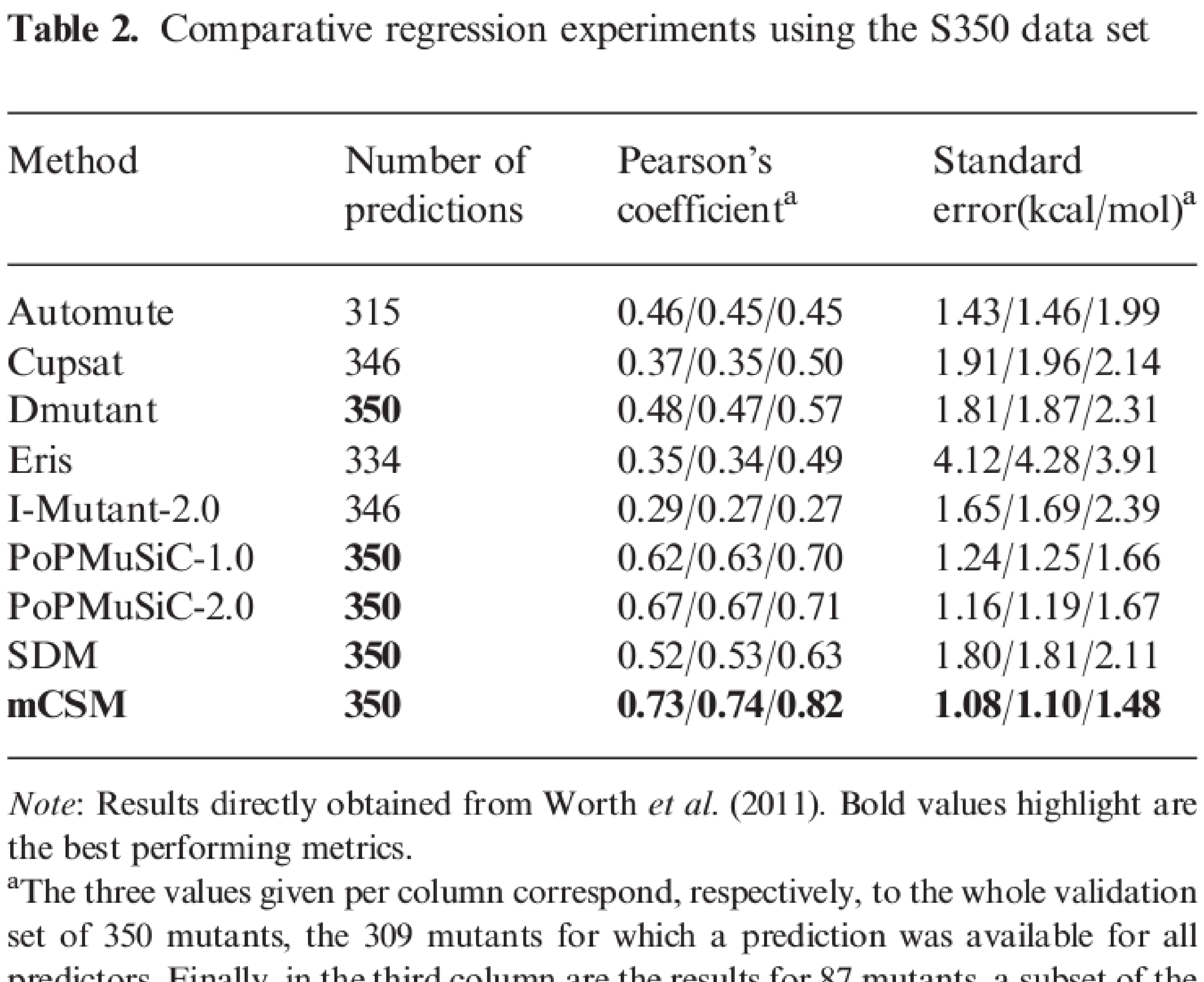


Table 2. Comparative regression experiments using the S350 data set

Method	Number of predictions	Pearson's coefficient[a]	Standard error(kcal/mol)[a]
Automute	315	0.46/0.45/0.45	1.43/1.46/1.99
Cupsat	346	0.37/0.35/0.50	1.91/1.96/2.14
Dmutant	**350**	0.48/0.47/0.57	1.81/1.87/2.31
Eris	334	0.35/0.34/0.49	4.12/4.28/3.91
I-Mutant-2.0	346	0.29/0.27/0.27	1.65/1.69/2.39
PoPMuSiC-1.0	**350**	0.62/0.63/0.70	1.24/1.25/1.66
PoPMuSiC-2.0	**350**	0.67/0.67/0.71	1.16/1.19/1.67
SDM	**350**	0.52/0.53/0.63	1.80/1.81/2.11
mCSM	**350**	**0.73/0.74/0.82**	**1.08/1.10/1.48**

Note: Results directly obtained from Worth *et al.* (2011). Bold values highlight are the best performing metrics.

[a] The three values given per column correspond, respectively, to the whole validation set of 350 mutants, the 309 mutants for which a prediction was available for all predictors. Finally, in the third column are the results for 87 mutants, a subset of the

main paper results

(*) mCSM: predicting the effect of mutations in proteins using graph-based signatures
Douglas E. V. Pires, David B. Ascher, Tom L. Blundell, 2014

mCSM

Supplementary material

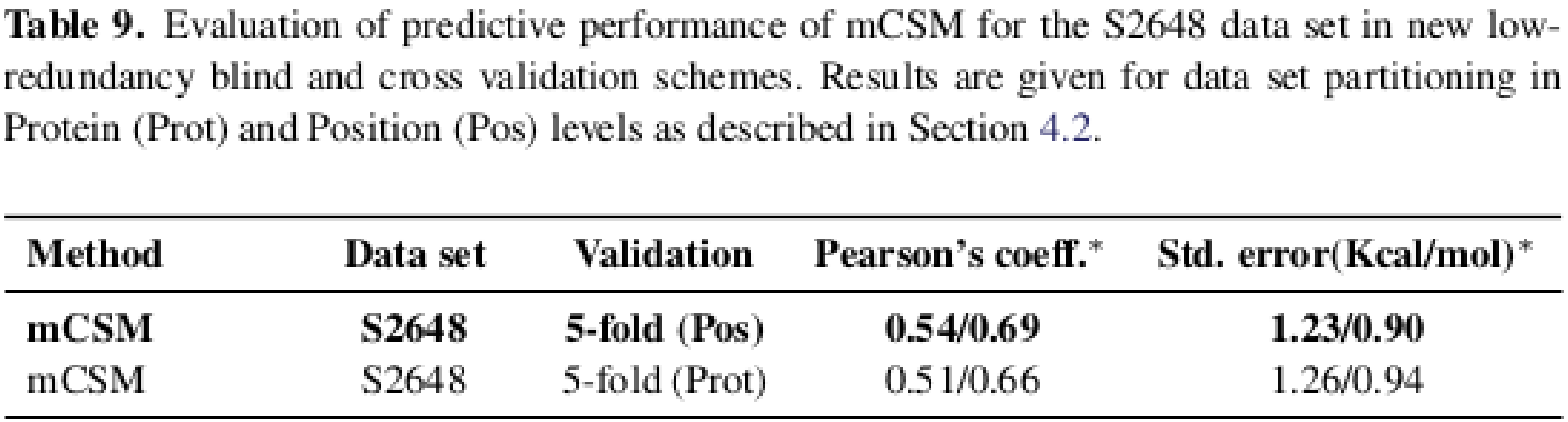


Table 9. Evaluation of predictive performance of mCSM for the S2648 data set in new low-redundancy blind and cross validation schemes. Results are given for data set partitioning in Protein (Prot) and Position (Pos) levels as described in Section 4.2.

Method	Data set	Validation	Pearson's coeff.*	Std. error(Kcal/mol)*
mCSM	**S2648**	**5-fold (Pos)**	**0.54/0.69**	**1.23/0.90**
mCSM	S2648	5-fold (Prot)	0.51/0.66	1.26/0.94

(*) mCSM: predicting the effect of mutations in proteins using graph-based signatures
Douglas E. V. Pires, David B. Ascher, Tom L. Blundell, 2014

mCSM

Summary:

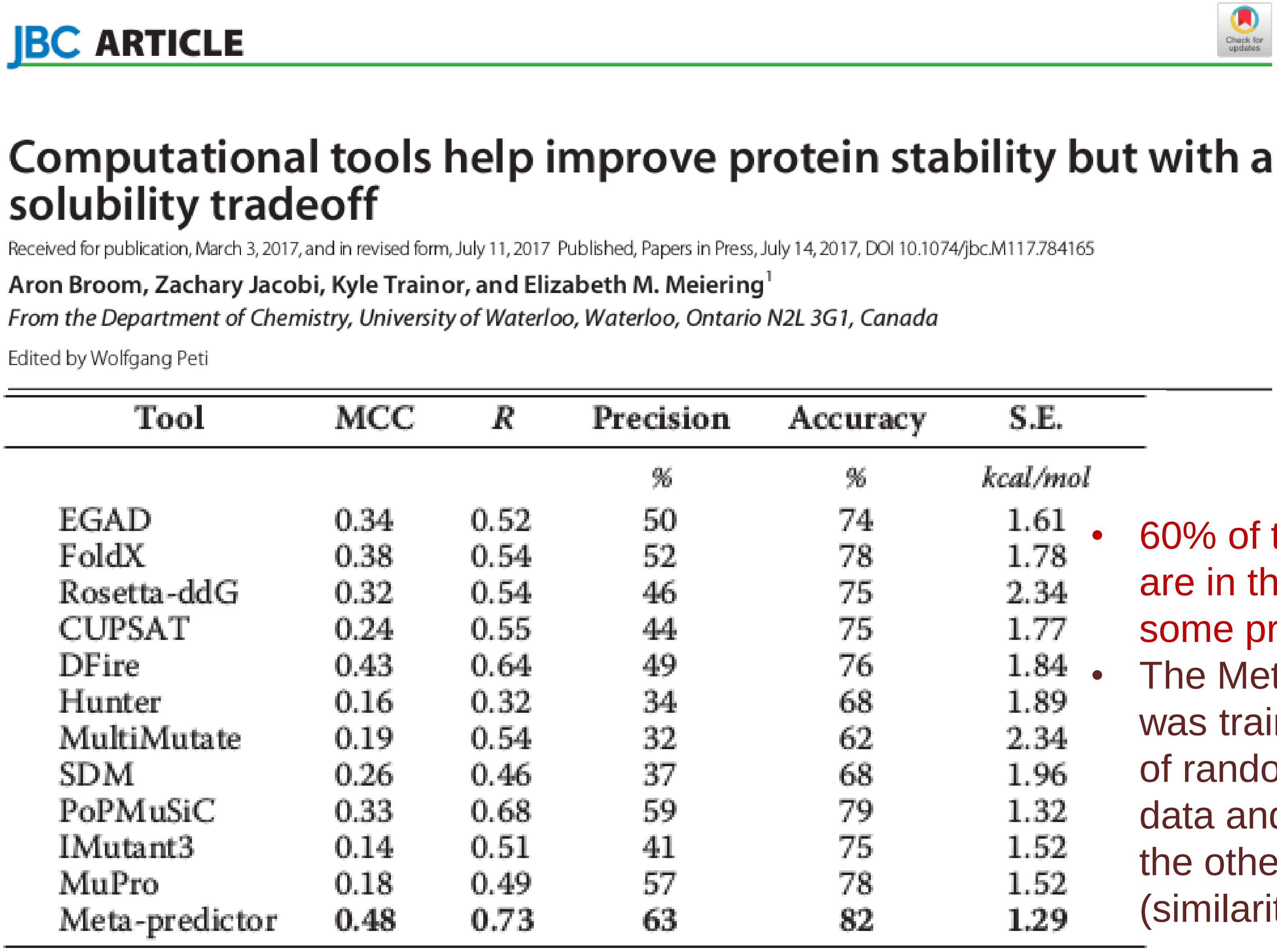
- Training-> r = 0.82
- Random split -> r = 0.73
- CV for positions -> r = 0.54
- CV for proteins -> r=0.51

(*) mCSM: predicting the effect of mutations in proteins using graph-based signatures
Douglas E. V. Pires, David B. Ascher, Tom L. Blundell, 2014

mCSM

Other examples with meta-predictors

Broom et al. 2017



JBC ARTICLE

Computational tools help improve protein stability but with a solubility tradeoff

Received for publication, March 3, 2017, and in revised form, July 11, 2017 Published, Papers in Press, July 14, 2017, DOI 10.1074/jbc.M117.784165

Aron Broom, Zachary Jacobi, Kyle Trainor, and Elizabeth M. Meiering[1]

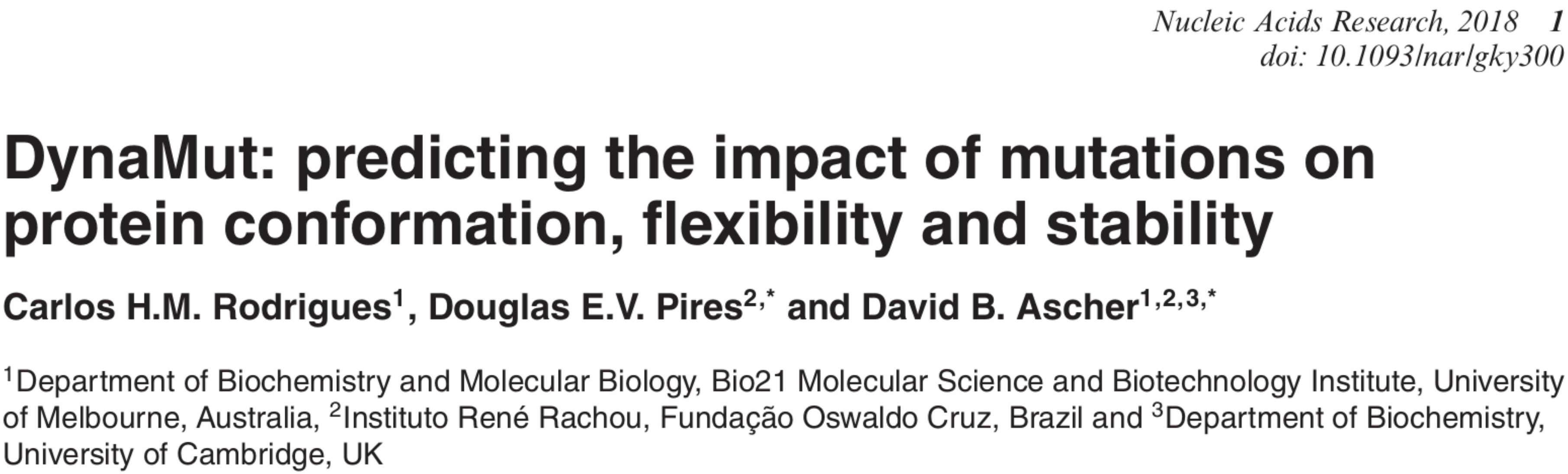
From the Department of Chemistry, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

Edited by Wolfgang Peti

Tool	MCC	*R*	Precision	Accuracy	S.E.
			%	%	*kcal/mol*
EGAD	0.34	0.52	50	74	1.61
FoldX	0.38	0.54	52	78	1.78
Rosetta-ddG	0.32	0.54	46	75	2.34
CUPSAT	0.24	0.55	44	75	1.77
DFire	0.43	0.64	49	76	1.84
Hunter	0.16	0.32	34	68	1.89
MultiMutate	0.19	0.54	32	62	2.34
SDM	0.26	0.46	37	68	1.96
PoPMuSiC	0.33	0.68	59	79	1.32
IMutant3	0.14	0.51	41	75	1.52
MuPro	0.18	0.49	57	78	1.52
Meta-predictor	**0.48**	**0.73**	**63**	**82**	**1.29**

- 60% of the proteins are in the training of some predictors
- The Meta-predictor was trained on 50% of randomly selected data and tested on the other 50% (similarity issue)

Rodrigues et al. 2018



Nucleic Acids Research, 2018 ***1***
doi: 10.1093/nar/gky300

DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability

Carlos H.M. Rodrigues[1], Douglas E.V. Pires[2,*] and David B. Ascher[1,2,3,*]

[1]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Australia, [2]Instituto René Rachou, Fundação Oswaldo Cruz, Brazil and [3]Department of Biochemistry, University of Cambridge, UK

- Dataset S2648
- Data presented in:
  - training on the data
  - cross-validation with random split
  - random generation on a “blind” set of 351 variations from S2648, and trained on the remainder 2297 variants.
- Based on the output of DUET, SDM2, mCSM that were TOTALLY trained on S2648

Please: *test your model using data (predictors) that have no sequence similarity (trained on proteins similar) to those of your test set!*

Contributors

Rita Casadio

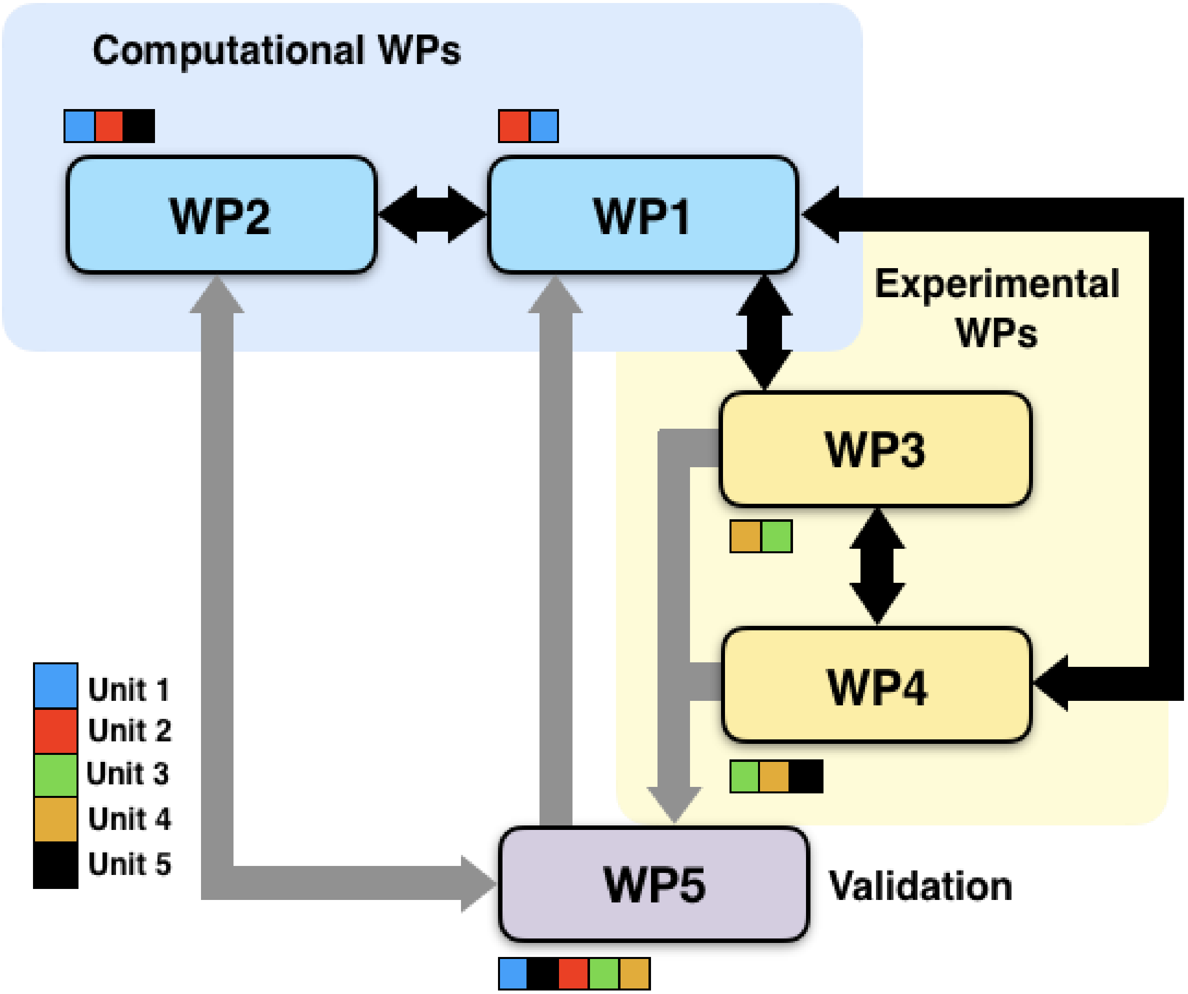
Emidio Capriotti

Pier Luigi Martelli

Ludovica Montanucci

Castrense Savojardo

Project:



Computational WPs

WP2

WP1

Experimental WPs

WP3

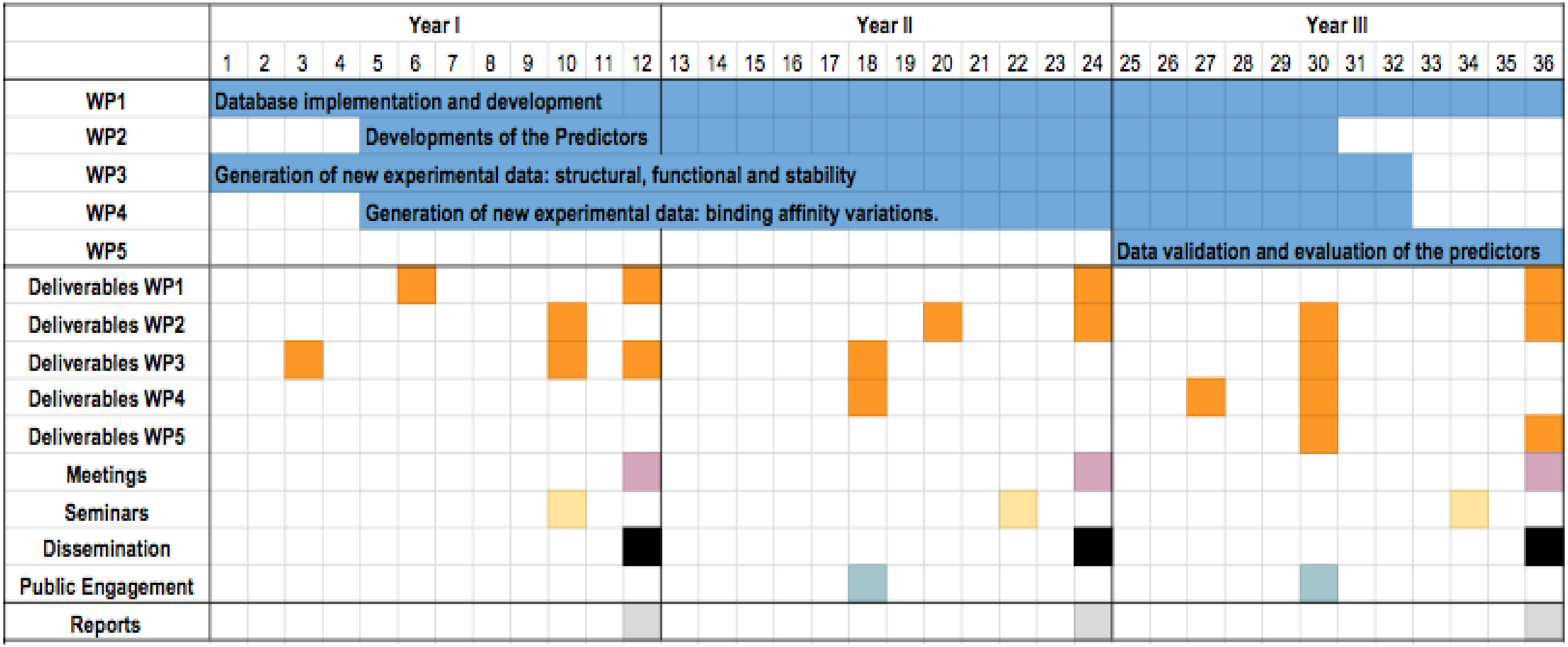
WP4

WP5

Validation

🟦 Unit 1
🟥 Unit 2
🟩 Unit 3
🟧 Unit 4
⬛ Unit 5

Project:



	Year I												Year II												Year III											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
WP1	Database implementation and development	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
WP2					Developments of the Predictors	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦						
WP3	Generation of new experimental data: structural, functional and stability	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦				
WP4					Generation of new experimental data: binding affinity variations.	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦				
WP5																									Data validation and evaluation of the predictors	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Deliverables WP1						🟧						🟧												🟧												🟧
Deliverables WP2										🟧										🟧				🟧						🟧						🟧
Deliverables WP3			🟧							🟧		🟧						🟧												🟧						
Deliverables WP4																		🟧									🟧			🟧						
Deliverables WP5																														🟧						🟧
Meetings												🟪												🟪												🟪
Seminars										🟨												🟨												🟨		
Dissemination												⬛												⬛												⬛
Public Engagement																		🟩												🟩						
Reports												▨												▨												▨

Project:

Definition of the objectives and deliverables in the light of the 38% cut