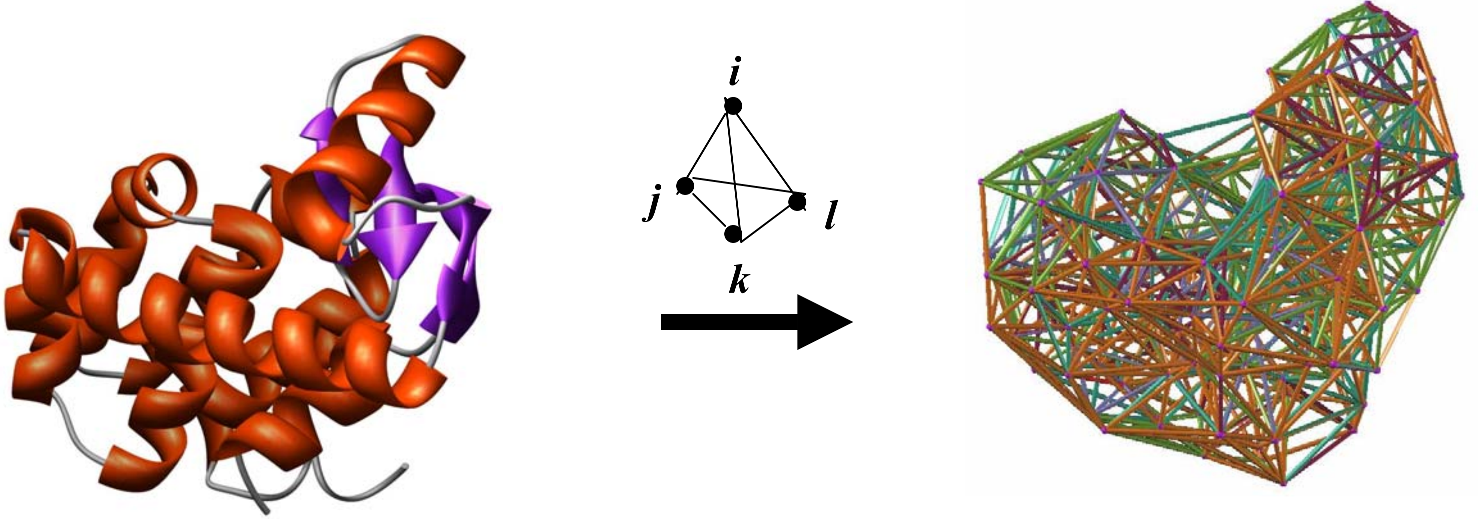
Computational Mutagenesis for Predicting Functional Consequences of Amino Acid Replacements in Proteins



Majid Masso, Ph.D.

Laboratory for Structural Bioinformatics

George Mason University

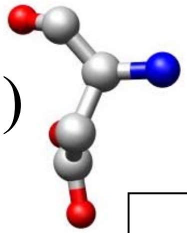
http://binf.gmu.edu/mmasso   mmasso@gmu.edu

What Constitutes a "Functional Consequence" Due to Amino Acid Substitutions?

- Change in protein stability:
  - Effect on melting temperature: $\Delta Tm = Tm\ (mutant) – Tm\ (wt)$
  - Effect on thermal denaturation: $\Delta\Delta G = \Delta G\ (mutant) – \Delta G\ (wt)$
  - Effect on denaturant denaturation: $\Delta\Delta G^{H_2O} = \Delta G^{H_2O}\ (mutant) – \Delta G^{H_2O}\ (wt)$
- Change in protein activity:
  - Mutant enzymatic activity relative to wt
  - Mutant strength of DNA binding relative to wt
- Disease potential of human coding nsSNPs
  - Neutral polymorphism or disease-associated mutation?
- For protein (human, bacterial, viral) targets of inhibitor drugs:
  - Continued sensitivity or (degree of) resistance that patients with the mutant protein have to the inhibitor
  - Inhibitor binding energy to mutant target relative to wt

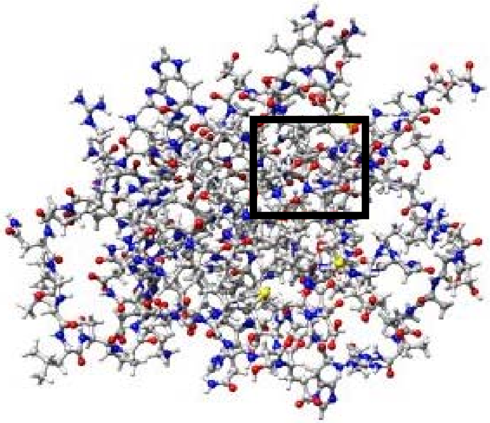
Delaunay Tessellation of Protein Structure

D (Asp)

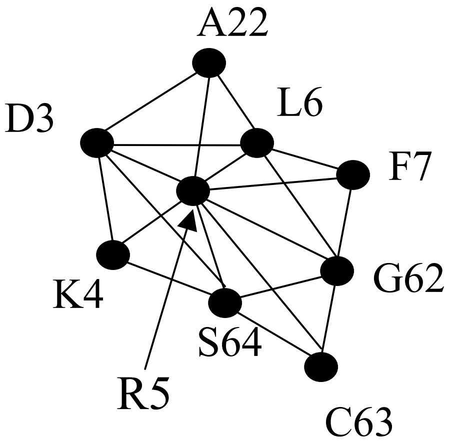


→ ● C_α or center of mass

Abstract each amino acid to a point
Atomic coordinates – Protein Data Bank (PDB)



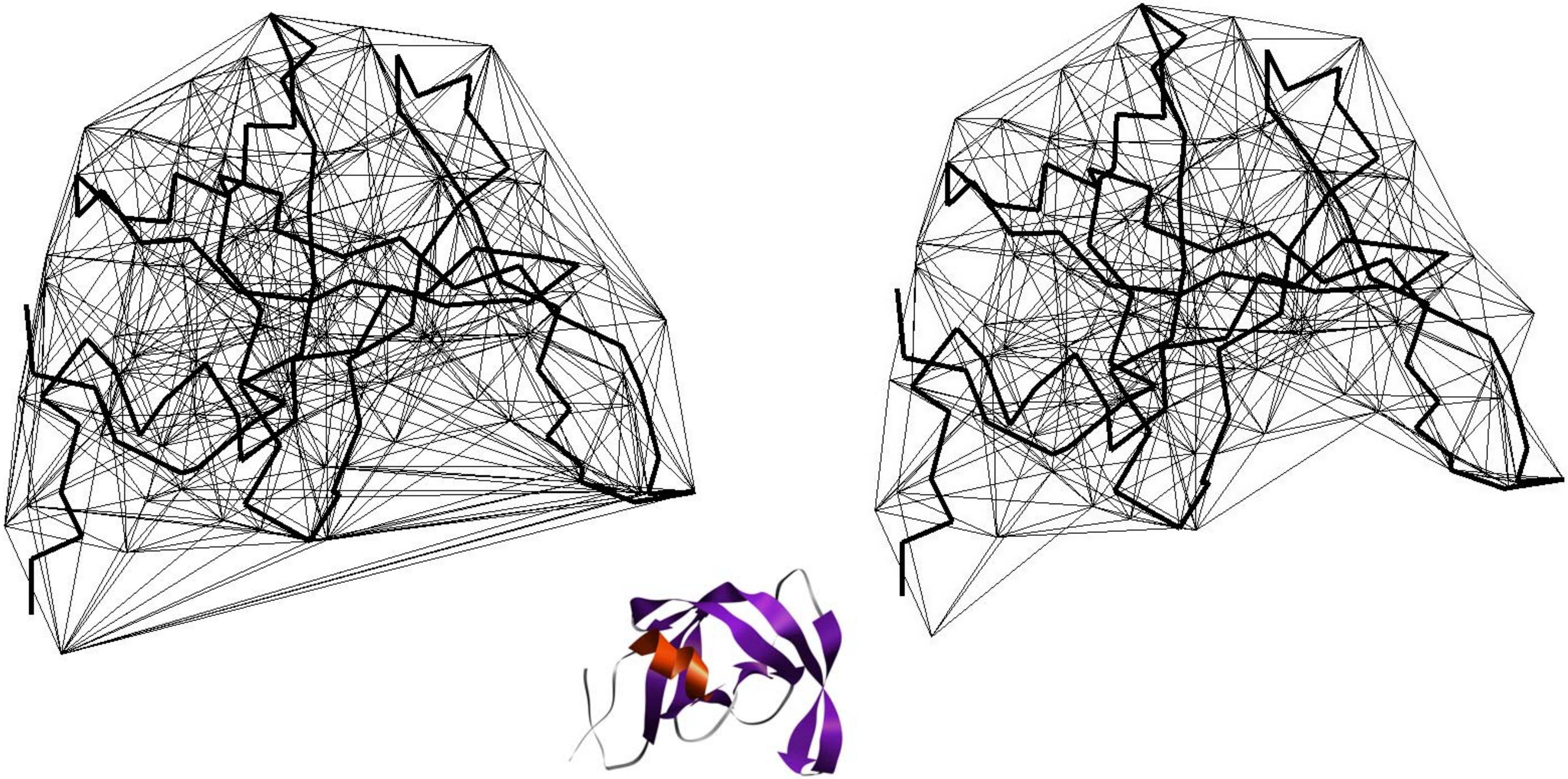
→



Delaunay tessellation: 3D “tiling” of space into non-overlapping, irregular tetrahedral simplices. Each simplex objectively defines a quadruplet of nearest-neighbor amino acids at its vertices.

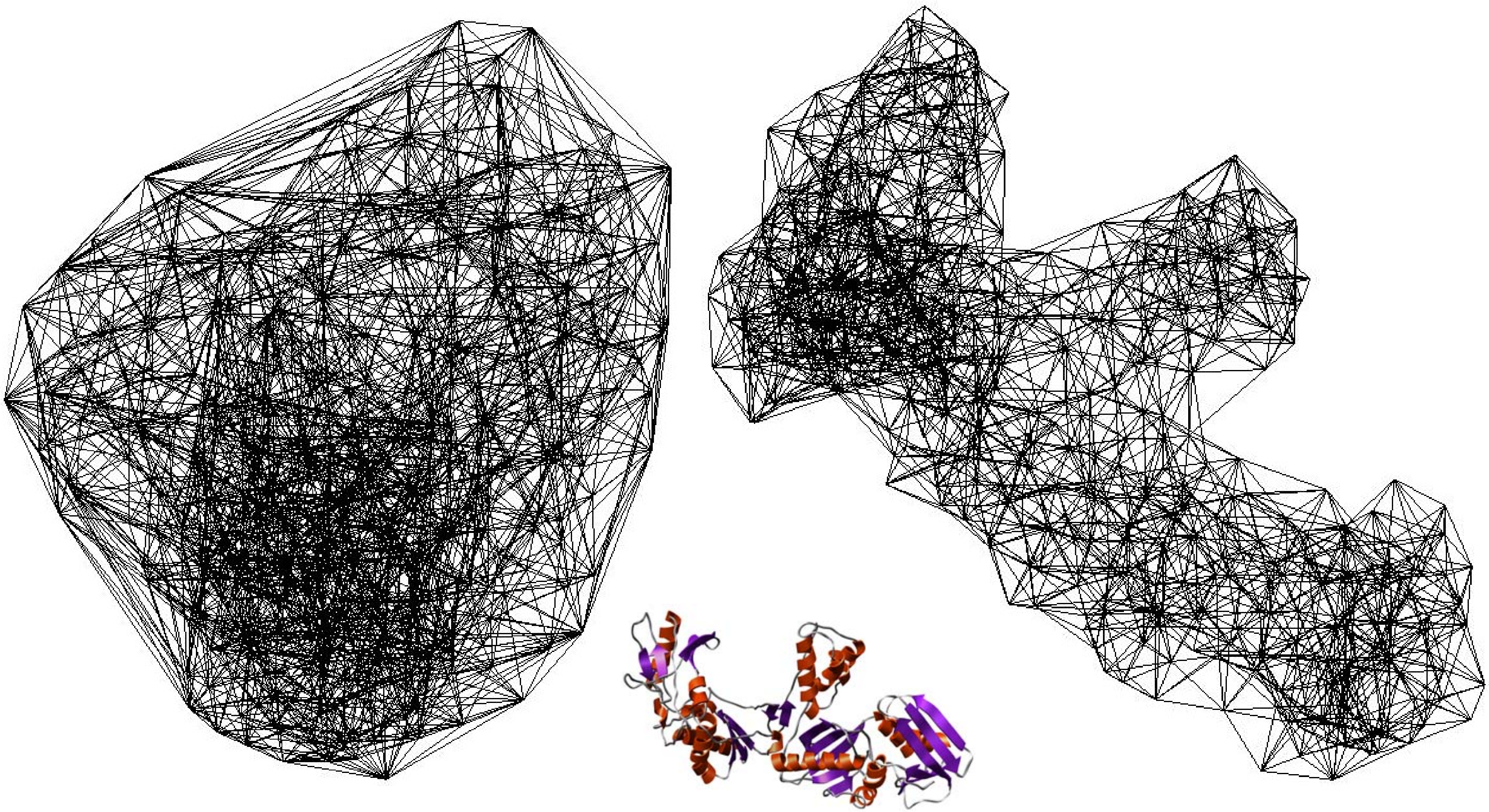
Example 1: HIV-1 Protease (3phv)

Vertices: Weighted side chain center of mass (CM) points for 99 aa’s
Dark line: C-alpha backbone trace (coincides with a vertex for Gly)
Left: complete tessellation; Right: partial (12A filter), “true” neighbors



Example 2: HIV-1 Reverse Transcriptase (1rtjA)

CM vertices; Left – full tessellation; Right – 12A filter on edges



Counting Amino Acid Quadruplets

Ordered quadruplets: $20^4 = 160{,}000$ (too many)

Order-independent quadruplets (our approach):

$$\underbrace{\mathrm{C}}\ \underbrace{\mathrm{D}}\ \underbrace{\mathrm{E}}\ \underbrace{\mathrm{F}} \qquad \binom{20}{4}$$

$$\mathrm{C}\ \ \mathrm{C}\ \ \underbrace{\mathrm{D}}\ \underbrace{\mathrm{E}} \qquad 20 \cdot \binom{19}{2}$$

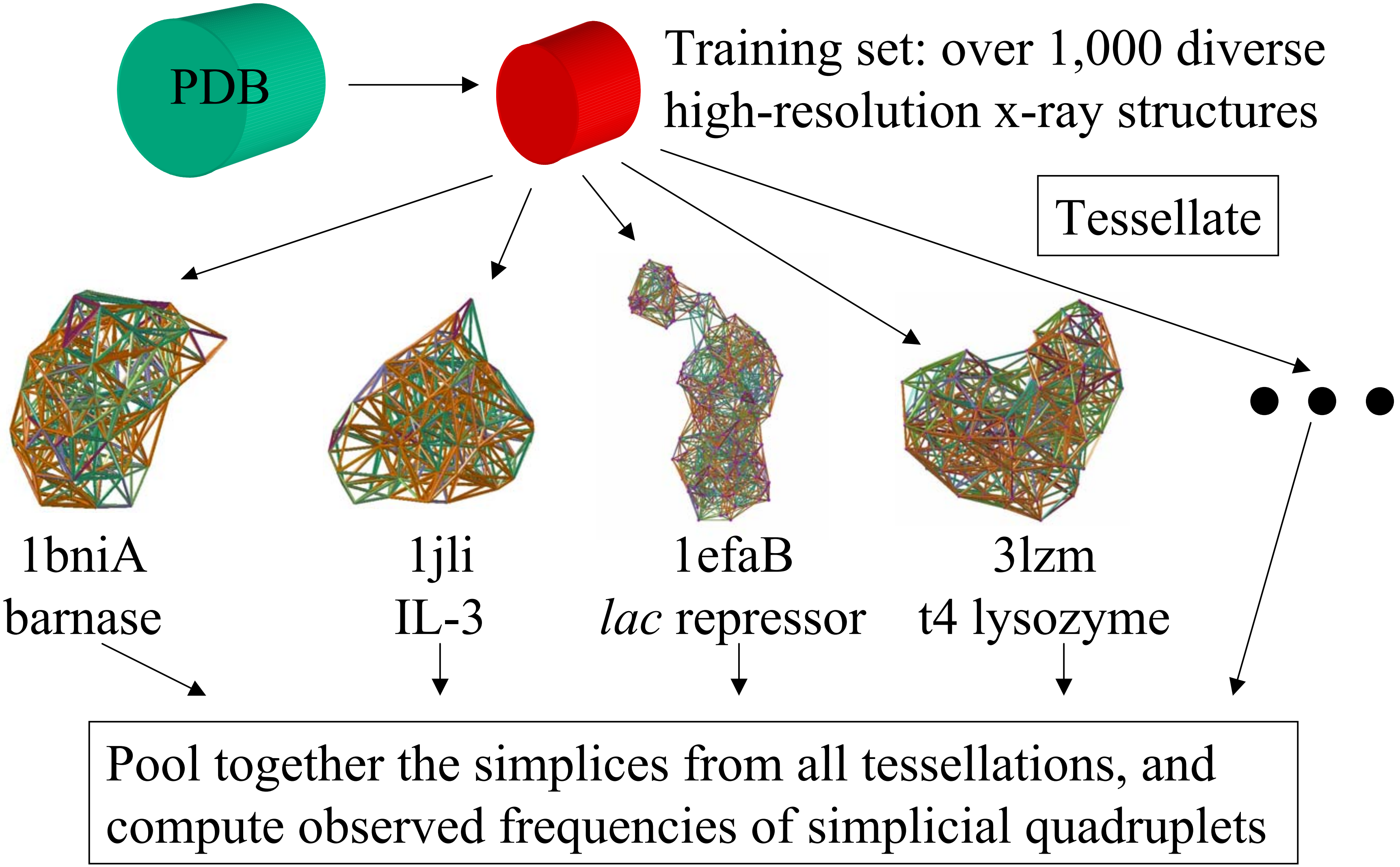
$$\underbrace{\mathrm{C}\ \ \mathrm{C}}\ \underbrace{\mathrm{D}\ \ \mathrm{D}} \qquad \binom{20}{2}$$

$$\mathrm{C}\ \ \mathrm{C}\ \ \mathrm{C}\ \ \mathrm{D} \qquad 20 \cdot 19$$

$$\mathrm{C}\ \ \mathrm{C}\ \ \mathrm{C}\ \ \mathrm{C} \qquad 20$$

Total: 8,855 distinct unordered quadruplets

Four-Body Statistical Potential



PDB

Training set: over 1,000 diverse high-resolution x-ray structures

Tessellate

1bniA
barnase

1jli
IL-3

1efaB
lac repressor

3lzm
t4 lysozyme

Pool together the simplices from all tessellations, and compute observed frequencies of simplicial quadruplets

Four-Body Statistical Potential

- Knowledge-based, modeled after inverse Boltzmann law:
  p_i = Frequency (feature i) ∂ $\mathrm{e}^{-\,\text{Energy (feature } i\text{)} / KT}$, i.e., $E_i \,\partial\, -KT \ln p_i$;
  and Potential (feature i) = $E_i - E_{ref} = \Delta E_i = -KT \ln(p_i / p_{ref})$

- For amino acid quadruplet (i,j,k,l), a log-likelihood score (interaction "pseudo-energy") is given by $s(i,j,k,l) = \log(f_{ijkl} / p_{ijkl})$

- f_{ijkl} = observed proportion of training set simplices whose four vertex residues are i,j,k,l

- p_{ijkl} = rate expected by chance (multinomial distribution, based on training set proportions of residues i,j,k,l)

- Four-body statistical potential: the collection of 8855 quadruplet (or simplex) types and their respective log-likelihood scores

Reference (Multinomial) Distribution

- Empirical potential of quadruplet interaction:

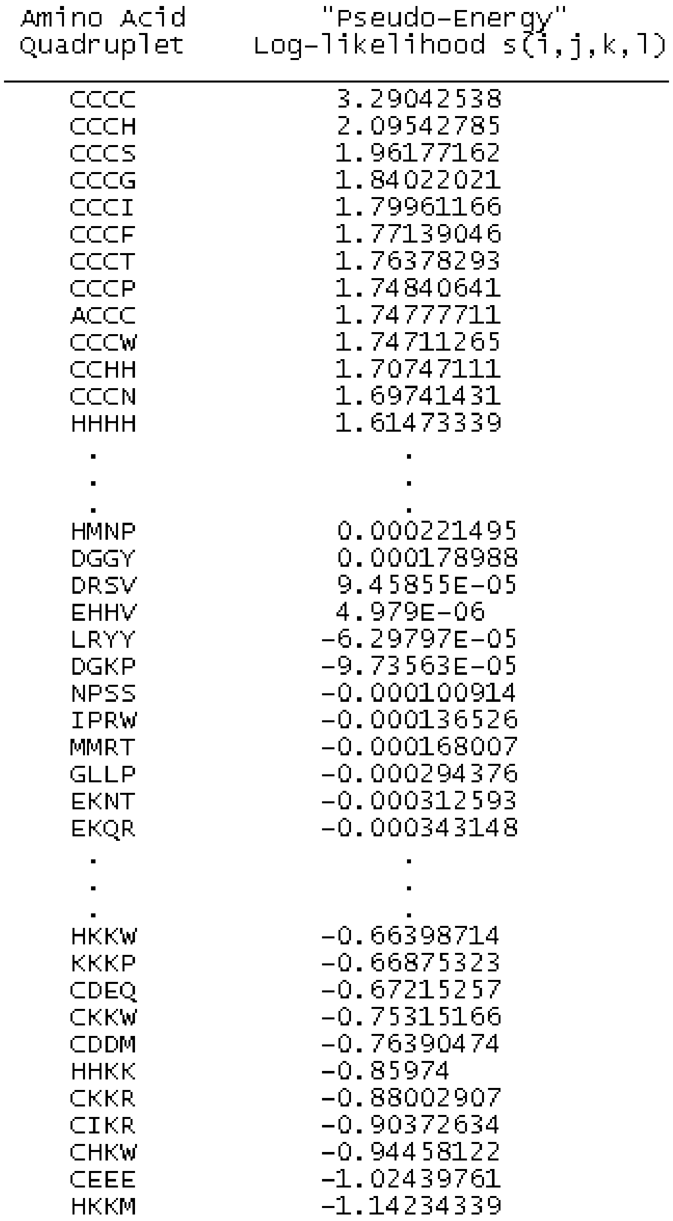
$$s(i,j,k,l) = \log(f_{ijkl} / p_{ijkl})$$

- Multinomial distribution:

$$p_{ijkl} = c a_i a_j a_k a_l$$

- a_i = total number of occurrences of residue i divided by total number of residues, in the entire training set of protein structures
- $c = \frac{4!}{\prod_i^n (t_i!)}$, where n = number of distinct residue types in the quadruplet, and t_i is the number of residues of type i.
- Potential problem: The collection of all amino acids exist in hundreds of separate training set structures
- Potential solution: Weighted average of separate multinomials for each structure, where weight = proportion of residues in structure

Four-Body Statistical Potential

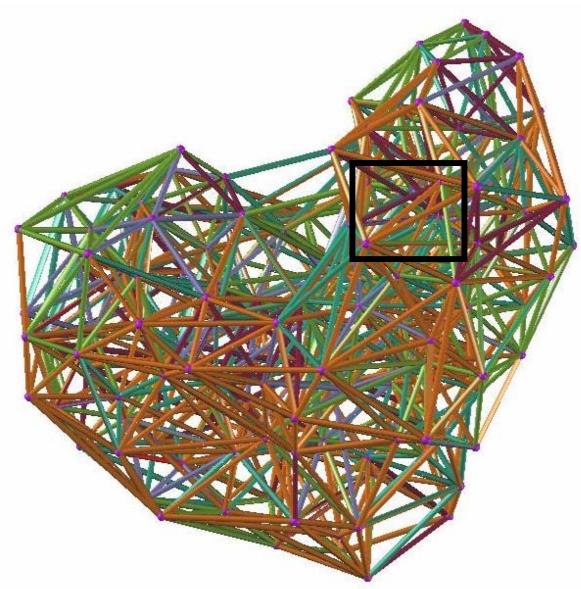


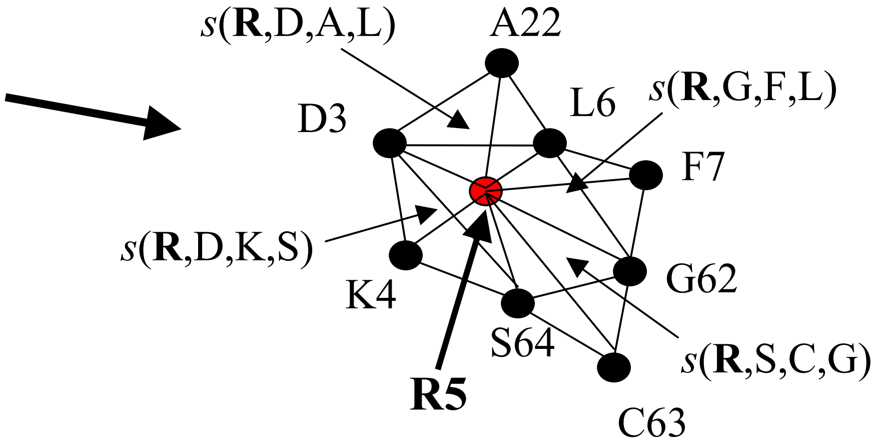
Amino Acid Quadruplet	"Pseudo-Energy" Log-likelihood s(i,j,k,l)
CCCC	3.29042538
CCCH	2.09542785
CCCS	1.96177162
CCCG	1.84022021
CCCI	1.79961166
CCCF	1.77139046
CCCT	1.76378293
CCCP	1.74840641
ACCC	1.74777711
CCCW	1.74711265
CCHH	1.70747111
CCCN	1.69741431
HHHH	1.61473339
.	.
.	.
.	.
HMNP	0.000221495
DGGY	0.000178988
DRSV	9.45855E-05
EHHV	4.979E-06
LRYY	-6.29797E-05
DGKP	-9.73563E-05
NPSS	-0.000100914
IPRW	-0.000136526
MMRT	-0.000168007
GLLP	-0.000294376
EKNT	-0.000312593
EKQR	-0.000343148
.	.
.	.
.	.
HKKW	-0.66398714
KKKP	-0.66875323
CDEQ	-0.67215257
CKKW	-0.75315166
CDDM	-0.76390474
HHKK	-0.85974
CKKR	-0.88002907
CIKR	-0.90372634
CHKW	-0.94458122
CEEE	-1.02439761
HKKM	-1.14234339

Application 1:Topological Score of a Protein

- Global measure of sequence-structure compatibility, also referred to as the “total (empirical or statistical) potential of the protein”
- Obtained by summing the log-likelihood scores of **all** simplicial quadruplets defined by the tessellation

TS = $\sum_{\hat{\imath}} s(\mathbf{x})$, sum taken over **all** simplex quadruplets **x** in the entire tessellation.

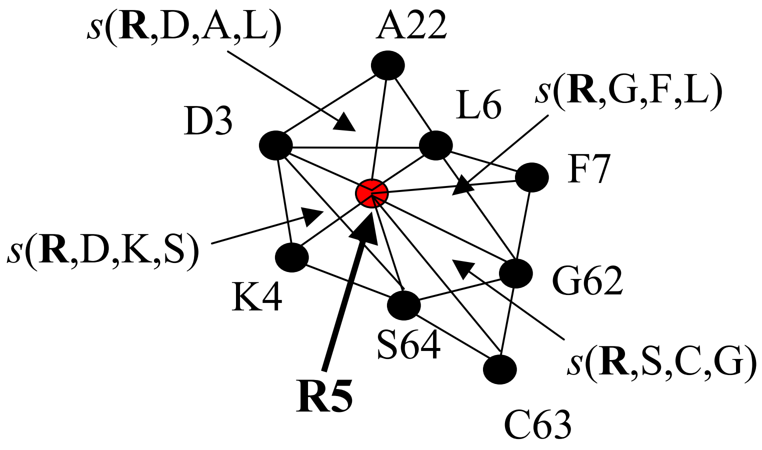




Close-up view of **only** the four simplices that use R at position 5 as a vertex (hypothetical)

Application 2: Residue Environment Scores

- For each amino acid position, locally sum log-likelihood scores $s(i,j,k,l)$ of only simplices that use the amino acid point as a vertex



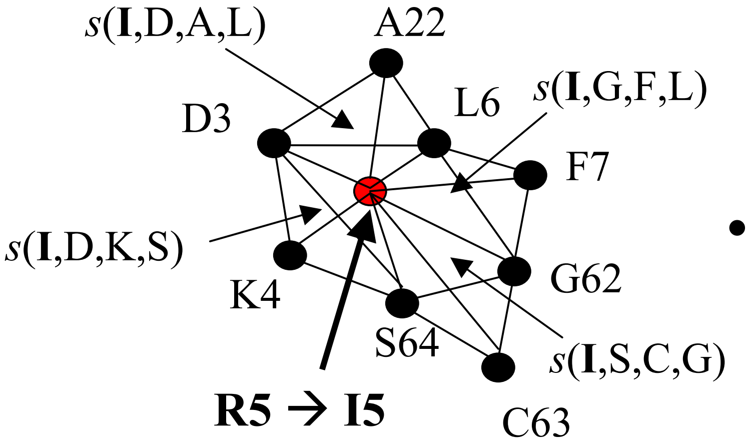
Example: $q_5 = q(\text{R5}) = \sum_{(i,j,k,l)} s(i,j,k,l)$, sum is taken over all simplex quads (i,j,k,l) that contain amino acid R5

- The scores of all the amino acid positions in the protein structure form a **Potential Profile** vector $\mathbf{Q} = < q_1,\ldots,q_N >$ (N = length of primary sequence in the solved structure)

Computational Mutagenesis Methodology

- Observation: mutant and wild type (wt) protein structure tessellations are very similar or identical

- Approach: obtain topological score and potential profile of mutant from wt structure tessellation, by changing residue labels at points

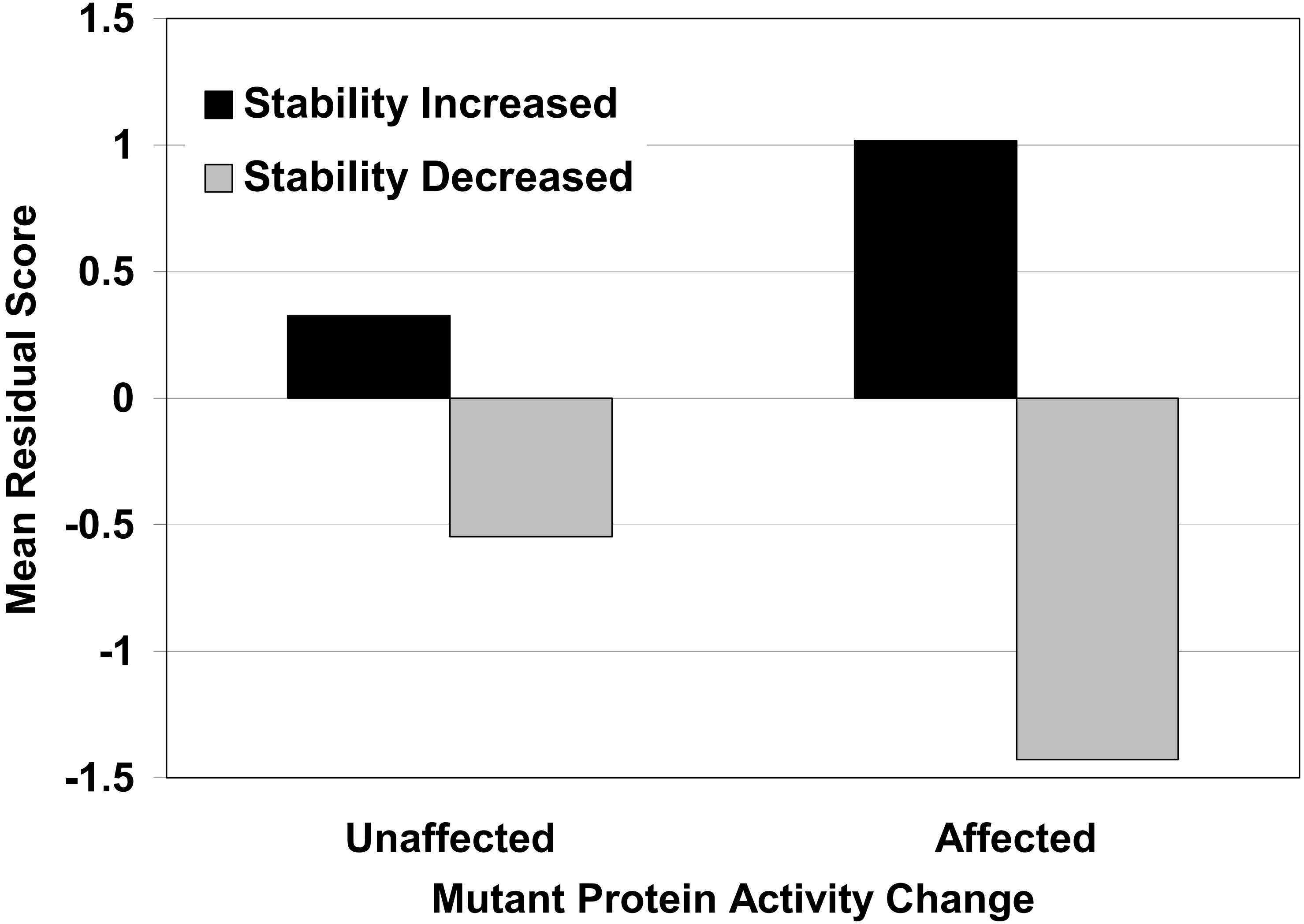


- **Scalar "Residual Score"**: mutant – wt topological scores = $TS_{mut} - TS_{wt}$ (empirical measure of overall relative structural impact due to mutation)

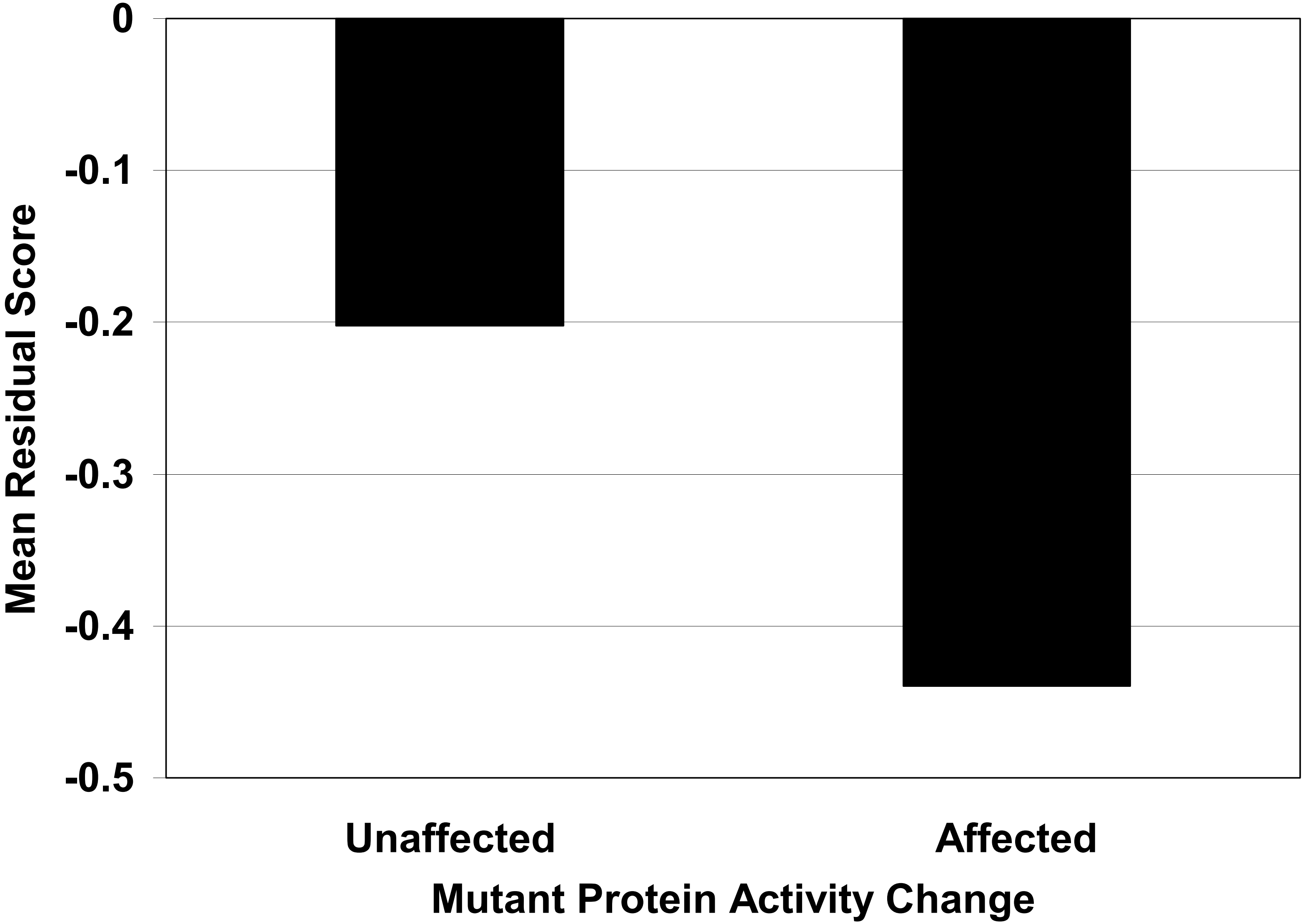
- **Vector "Residual Profile":**

  $\mathbf{R} = \mathbf{Q}_{mut} - \mathbf{Q}_{wt}$ = difference between mutant and wt potential profile vectors (environmental perturbation score for every amino acid position in structure)

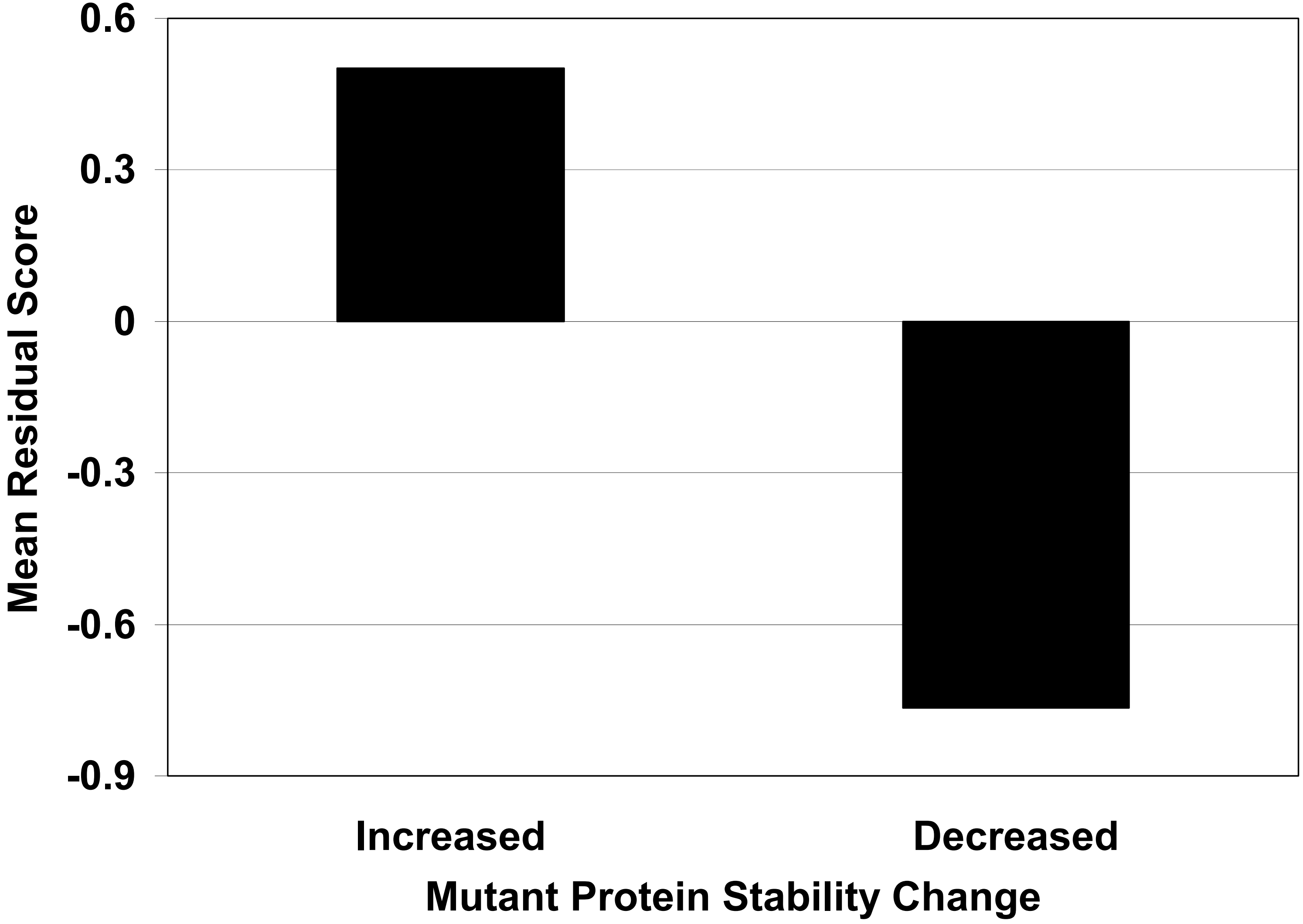
Residual Scores Example: 980 Distinct Single-Point Mutants in 20 Proteins



Residual Score Example (Continued)

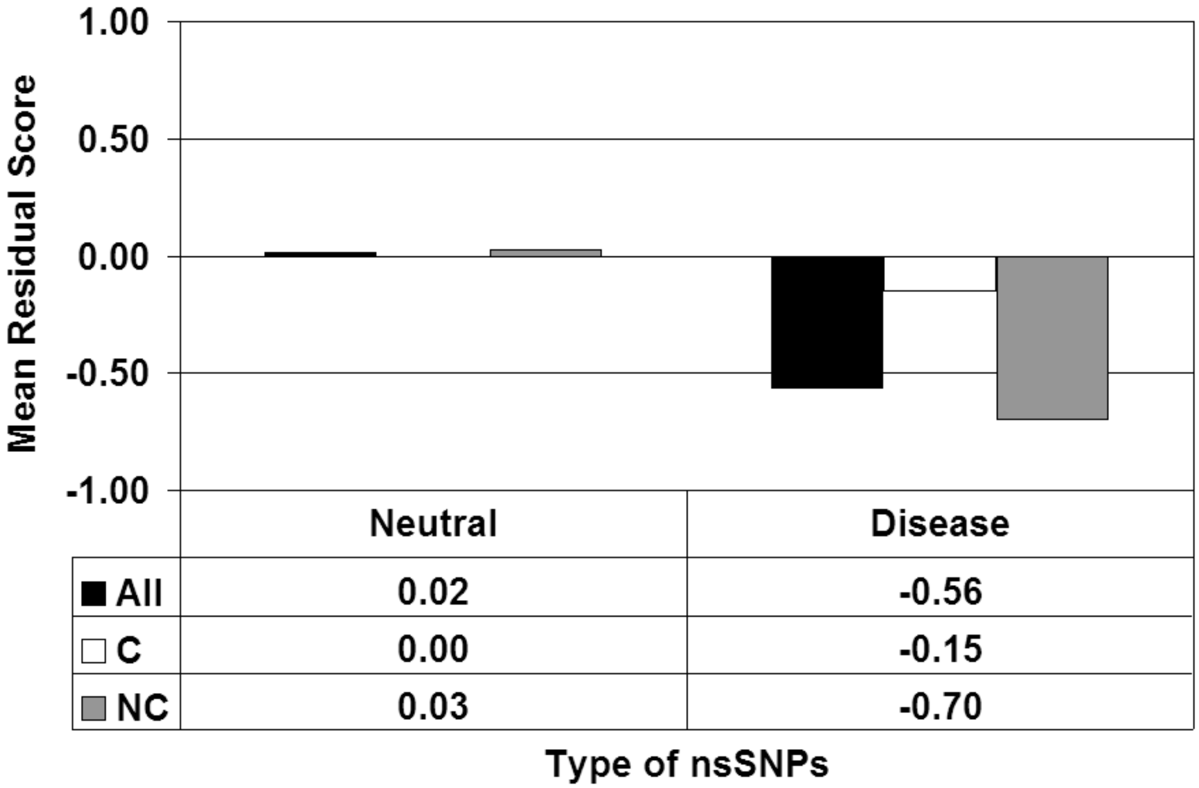


Residual Score Example (Continued)



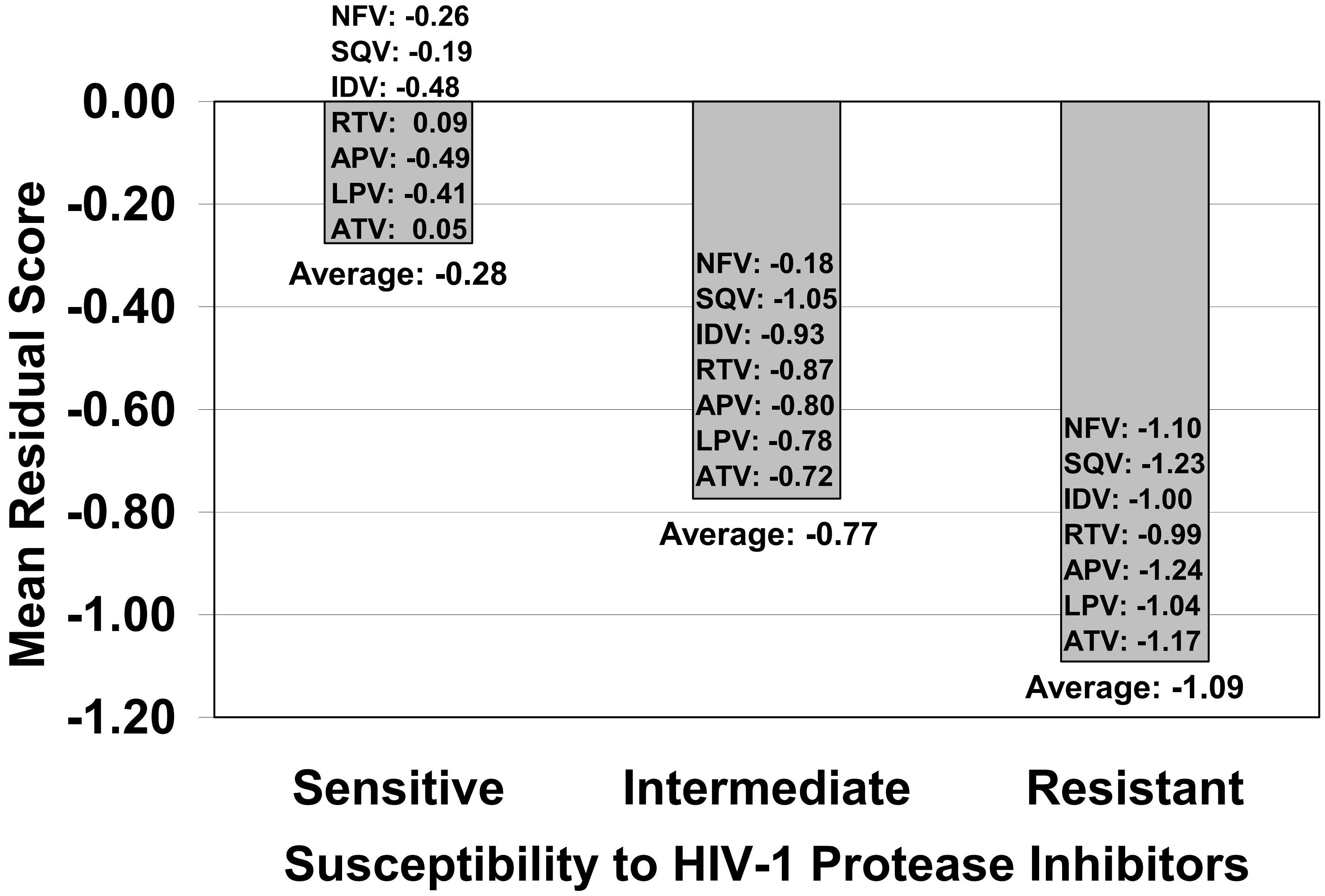
Structure-Function Correlations Based on Residual Scores: nsSNPs

- 1790 nsSNPs corresponding to single amino acid substitutions in several hundred proteins with tessellatable structures
- Function: 1332 nsSNPs associated with disease; 458 neutral
- Data obtained from Swiss-Prot and HPI



	Neutral	Disease
■ All	0.02	-0.56
□ C	0.00	-0.15
▨ NC	0.03	-0.70

Structure-Function Correlations Based on Residual Scores: Drug Susceptibility



Mutant Residual Profiles: Motivation

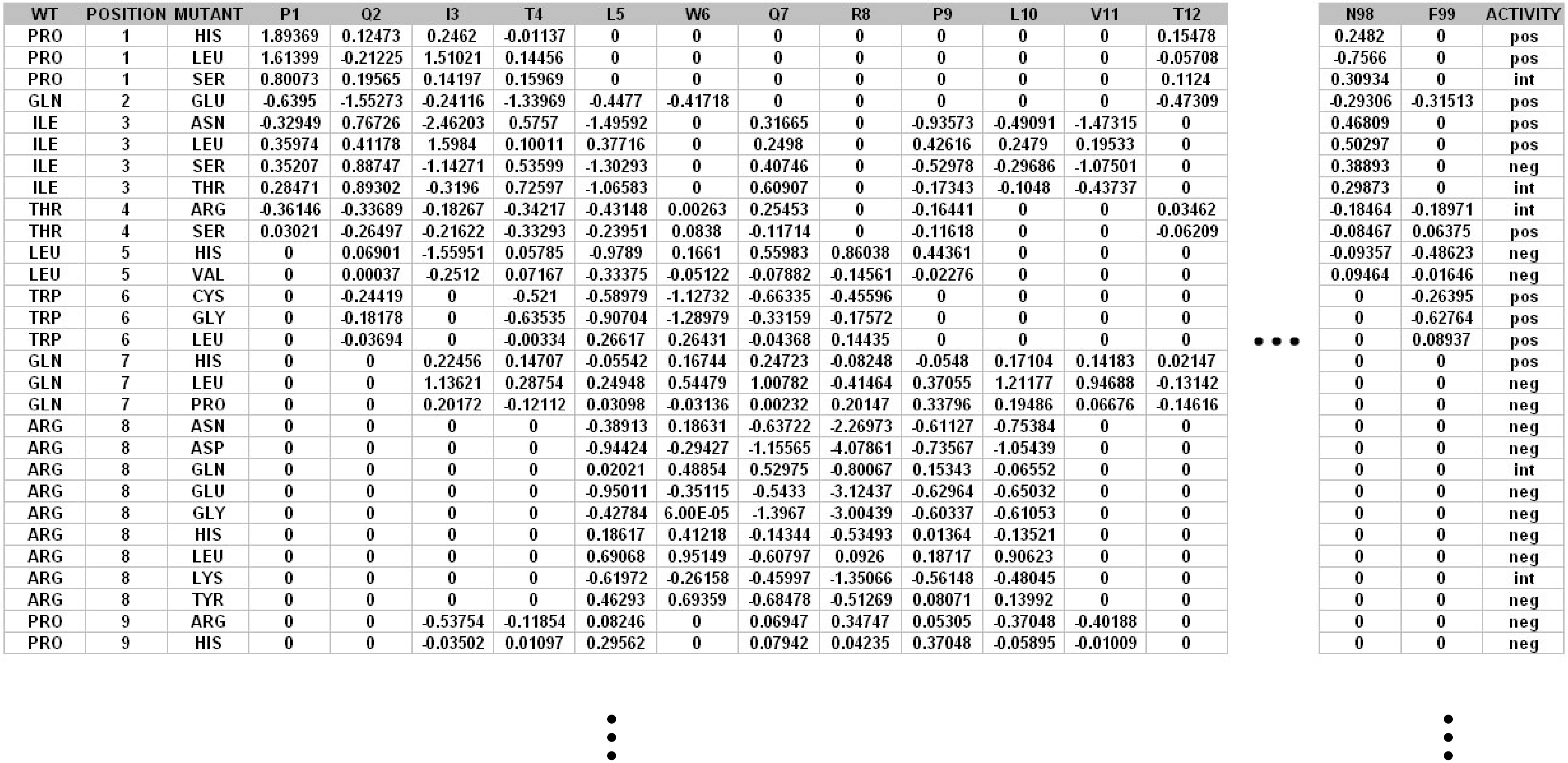
- Residual profile vectors encode much more sequence and structure information about mutants than scalar residual scores; Denote **R** = $< EC_1,\ldots,EC_N >$, where EC_i = *Environmental Change* at position *i* relative to wt

- $EC_i = 0$ unless either position *i* has been mutated, or position *i* is involved in a simplex with a mutated position (structure info)

- For the special case of single point mutants, residual scores are explicitly incorporated into the residual profiles (EC score at mutated position = residual score of mutant protein)

- Residual profiles of all 19 single point mutants at one position have identical arrangements of zero and nonzero components; only the values of the nonzero components differ (sequence info)

HIV-1 PR Dataset Example:
Residual Profiles of 536 Experimental Mutants



WT	POSITION	MUTANT	P1	Q2	I3	T4	L5	W6	Q7	R8	P9	L10	V11	T12	…	N98	F99	ACTIVITY
PRO	1	HIS	1.89369	0.12473	0.2462	-0.01137	0	0	0	0	0	0	0	0.15478		0.2482	0	pos
PRO	1	LEU	1.61399	-0.21225	1.51021	0.14456	0	0	0	0	0	0	0	-0.05708		-0.7566	0	pos
PRO	1	SER	0.80073	0.19565	0.14197	0.15969	0	0	0	0	0	0	0	0.1124		0.30934	0	int
GLN	2	GLU	-0.6395	-1.55273	-0.24116	-1.33969	-0.4477	-0.41718	0	0	0	0	0	-0.47309		-0.29306	-0.31513	pos
ILE	3	ASN	-0.32949	0.76726	-2.46203	0.5757	-1.49592	0	0.31665	0	-0.93573	-0.49091	-1.47315	0		0.46809	0	pos
ILE	3	LEU	0.35974	0.41178	1.5984	0.10011	0.37716	0	0.2498	0	0.42616	0.2479	0.19533	0		0.50297	0	pos
ILE	3	SER	0.35207	0.88747	-1.14271	0.53599	-1.30293	0	0.40746	0	-0.52978	-0.29686	-1.07501	0		0.38893	0	neg
ILE	3	THR	0.28471	0.89302	-0.3196	0.72597	-1.06583	0	0.60907	0	-0.17343	-0.1048	-0.43737	0		0.29873	0	int
THR	4	ARG	-0.36146	-0.33689	-0.18267	-0.34217	-0.43148	0.00263	0.25453	0	-0.16441	0	0	0.03462		-0.18464	-0.18971	int
THR	4	SER	0.03021	-0.26497	-0.21622	-0.33293	-0.23951	0.0838	-0.11714	0	-0.11618	0	0	-0.06209		-0.08467	0.06375	pos
LEU	5	HIS	0	0.06901	-1.55951	0.05785	-0.9789	0.1661	0.55983	0.86038	0.44361	0	0	0		-0.09357	-0.48623	neg
LEU	5	VAL	0	0.00037	-0.2512	0.07167	-0.33375	-0.05122	-0.07882	-0.14561	-0.02276	0	0	0		0.09464	-0.01646	neg
TRP	6	CYS	0	-0.24419	0	-0.521	-0.58979	-1.12732	-0.66335	-0.45596	0	0	0	0		0	-0.26395	pos
TRP	6	GLY	0	-0.18178	0	-0.63535	-0.90704	-1.28979	-0.33159	-0.17572	0	0	0	0		0	-0.62764	pos
TRP	6	LEU	0	-0.03694	0	-0.00334	0.26617	0.26431	-0.04368	0.14435	0	0	0	0	• • •	0	0.08937	pos
GLN	7	HIS	0	0	0.22456	0.14707	-0.05542	0.16744	0.24723	-0.08248	-0.0548	0.17104	0.14183	0.02147		0	0	pos
GLN	7	LEU	0	0	1.13621	0.28754	0.24948	0.54479	1.00782	-0.41464	0.37055	1.21177	0.94688	-0.13142		0	0	neg
GLN	7	PRO	0	0	0.20172	-0.12112	0.03098	-0.03136	0.00232	0.20147	0.33796	0.19486	0.06676	-0.14616		0	0	neg
ARG	8	ASN	0	0	0	0	-0.38913	0.18631	-0.63722	-2.26973	-0.61127	-0.75384	0	0		0	0	neg
ARG	8	ASP	0	0	0	0	-0.94424	-0.29427	-1.15565	-4.07861	-0.73567	-1.05439	0	0		0	0	neg
ARG	8	GLN	0	0	0	0	0.02021	0.48854	0.52975	-0.80067	0.15343	-0.06552	0	0		0	0	int
ARG	8	GLU	0	0	0	0	-0.95011	-0.35115	-0.5433	-3.12437	-0.62964	-0.65032	0	0		0	0	neg
ARG	8	GLY	0	0	0	0	-0.42784	6.00E-05	-1.3967	-3.00439	-0.60337	-0.61053	0	0		0	0	neg
ARG	8	HIS	0	0	0	0	0.18617	0.41218	-0.14344	-0.53493	0.01364	-0.13521	0	0		0	0	neg
ARG	8	LEU	0	0	0	0	0.69068	0.95149	-0.60797	0.0926	0.18717	0.90623	0	0		0	0	neg
ARG	8	LYS	0	0	0	0	-0.61972	-0.26158	-0.45997	-1.35066	-0.56148	-0.48045	0	0		0	0	int
ARG	8	TYR	0	0	0	0	0.46293	0.69359	-0.68478	-0.51269	0.08071	0.13992	0	0		0	0	neg
PRO	9	ARG	0	0	-0.53754	-0.11854	0.08246	0	0.06947	0.34747	0.05305	-0.37048	-0.40188	0		0	0	neg
PRO	9	HIS	0	0	-0.03502	0.01097	0.29562	0	0.07942	0.04235	0.37048	-0.05895	-0.01009	0		0	0	neg

⋮

⋮

Machine Learning Algorithms

- Supervised Classification: Neural Network (NN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF); Regression: Tree regression, Support Vector Regression
- Training set: residual profiles ("attribute" or "feature" vectors) of protein mutants ("instances" or "examples") with experimentally measured function (categorical "class" or a numerical "value")
- Common approach among algorithms: train a model capable of accurately classifying or determining the value of each example, based on the values of the attribute set
- Learned model: a consistent set of relationships or rules (complex nonlinear function) between the attributes of the examples and their classes (or values), used for predicting the class memberships (or values) of new, unstudied instances

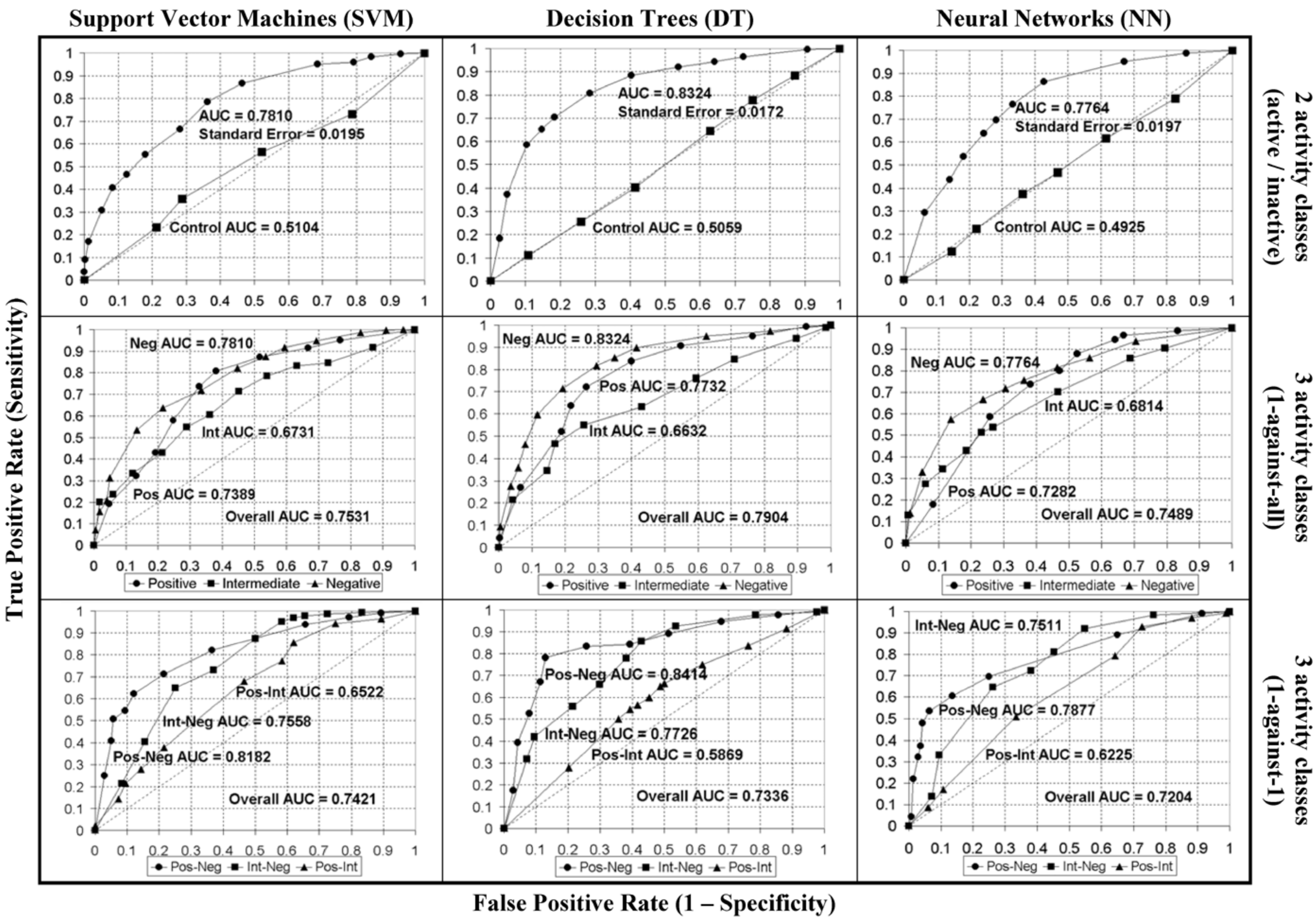
Evaluating Algorithm Performance

- Overall goal: Develop model with known examples to accurately predict class (or value) of examples that have not yet been assayed experimentally (potentially great savings of time and money)

- Approaches: Tenfold cross-validation (CV);
  leave-one-out (i.e., jackknife or N-fold CV, N = dataset size);
  % split (e.g., use only 2/3 for training, 1/3 held out for testing)

- Classification performance measures:
  accuracy = (TP+TN) / (TP+FP+TN+FN); sensitivity = TP / (TP+FN);
  specificity = TN / (TN+FP); precision = TP / (TP+FP);
  BER = 0.5 × [FP / (FP+TN) + FN / (FN+TP)];
  MCC = (TP×TN – FP×FN) / ◊(TP+FN)(TP+FP)(TN+FN)(TN+FP);
  AUC = area under ROC curve (plot of sensitivity vs. 1 – specificity)
  For regression models: correlation coefficient, standard error

Algorithm Performance: HIV-1 PR Mutants



Real-World Application: HIV-1 PR

- Model: two-class decision tree, trained with the 536 HIV-1 PR mutants
- Test set: experimental activity for 47 additional mutants discovered while searching the literature (12 different studies)
- Residual profiles of the mutants fed into model for predictions
- Result: 37/47 (79%) of the mutant activity predictions match experimental activity

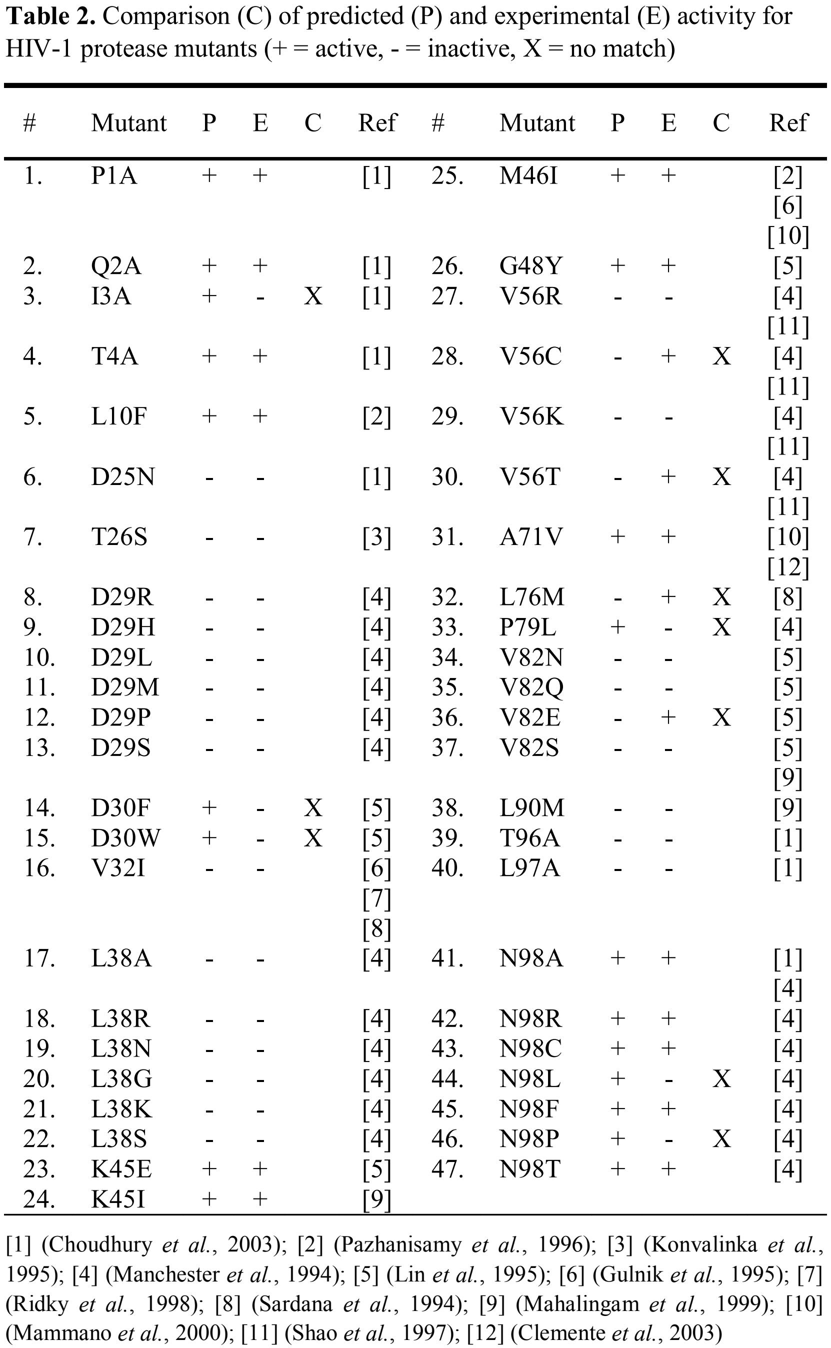
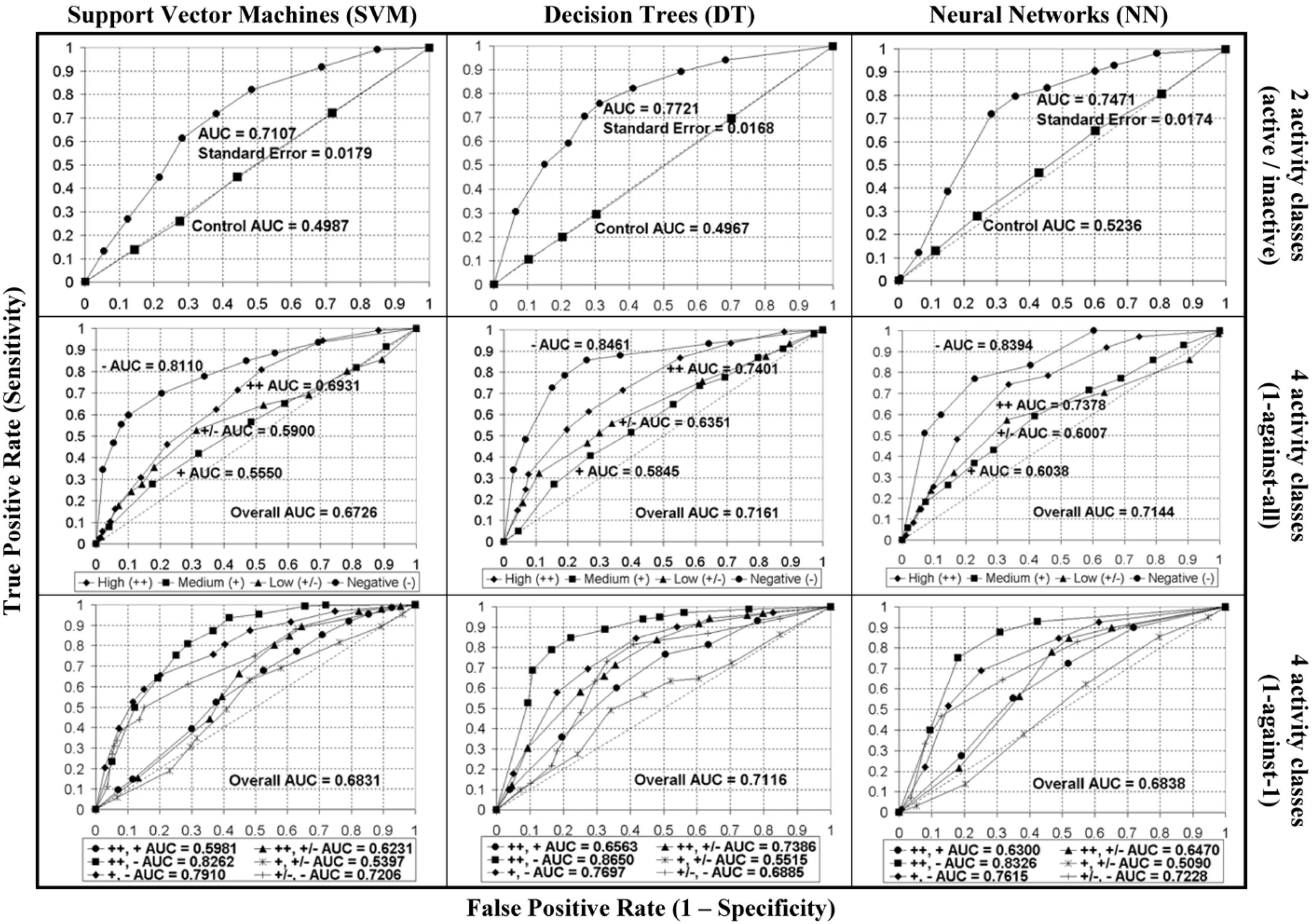


Table 2. Comparison (C) of predicted (P) and experimental (E) activity for HIV-1 protease mutants (+ = active, - = inactive, X = no match)

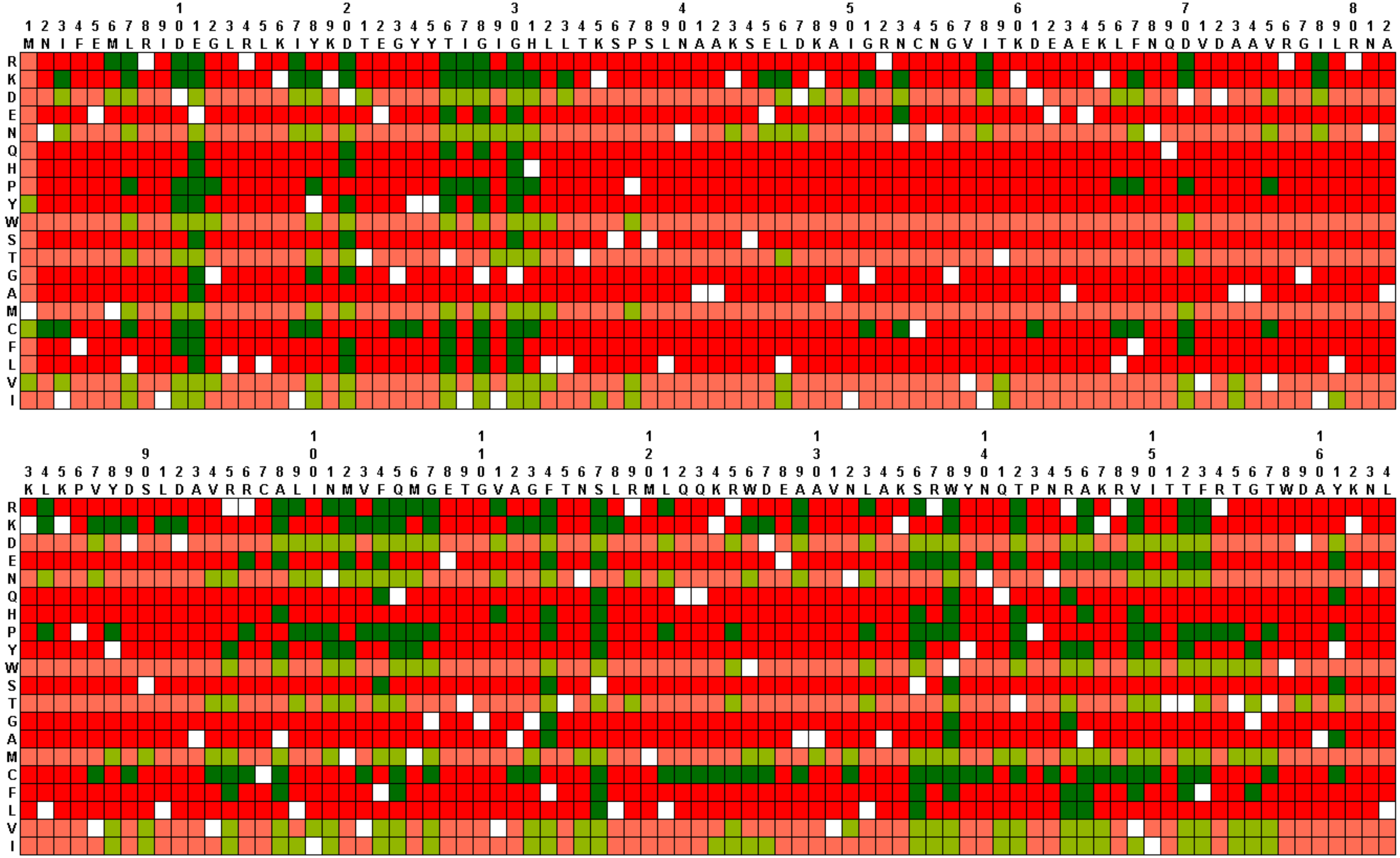
#	Mutant	P	E	C	Ref	#	Mutant	P	E	C	Ref
1.	P1A	+	+		[1]	25.	M46I	+	+		[2] [6] [10]
2.	Q2A	+	+		[1]	26.	G48Y	+	+		[5]
3.	I3A	+	-	X	[1]	27.	V56R	-	-		[4] [11]
4.	T4A	+	+		[1]	28.	V56C	-	+	X	[4] [11]
5.	L10F	+	+		[2]	29.	V56K	-	-		[4] [11]
6.	D25N	-	-		[1]	30.	V56T	-	+	X	[4] [11]
7.	T26S	-	-		[3]	31.	A71V	+	+		[10] [12]
8.	D29R	-	-		[4]	32.	L76M	-	+	X	[8]
9.	D29H	-	-		[4]	33.	P79L	+	-	X	[4]
10.	D29L	-	-		[4]	34.	V82N	-	-		[5]
11.	D29M	-	-		[4]	35.	V82Q	-	-		[5]
12.	D29P	-	-		[4]	36.	V82E	-	+	X	[5]
13.	D29S	-	-		[4]	37.	V82S	-	-		[5] [9]
14.	D30F	+	-	X	[5]	38.	L90M	-	-		[9]
15.	D30W	+	-	X	[5]	39.	T96A	-	-		[1]
16.	V32I	-	-		[6] [7] [8]	40.	L97A	-	-		[1]
17.	L38A	-	-		[4]	41.	N98A	+	+		[1] [4]
18.	L38R	-	-		[4]	42.	N98R	+	+		[4]
19.	L38N	-	-		[4]	43.	N98C	+	+		[4]
20.	L38G	-	-		[4]	44.	N98L	+	-	X	[4]
21.	L38K	-	-		[4]	45.	N98F	+	+		[4]
22.	L38S	-	-		[4]	46.	N98P	+	-	X	[4]
23.	K45E	+	+		[5]	47.	N98T	+	+		[4]
24.	K45I	+	+		[9]						

[1] (Choudhury *et al.*, 2003); [2] (Pazhanisamy *et al.*, 1996); [3] (Konvalinka *et al.*, 1995); [4] (Manchester *et al.*, 1994); [5] (Lin *et al.*, 1995); [6] (Gulnik *et al.*, 1995); [7] (Ridky *et al.*, 1998); [8] (Sardana *et al.*, 1994); [9] (Mahalingam *et al.*, 1999); [10] (Mammano *et al.*, 2000); [11] (Shao *et al.*, 1997); [12] (Clemente *et al.*, 2003)

Performance: 2015 T4 Lysozyme Mutants



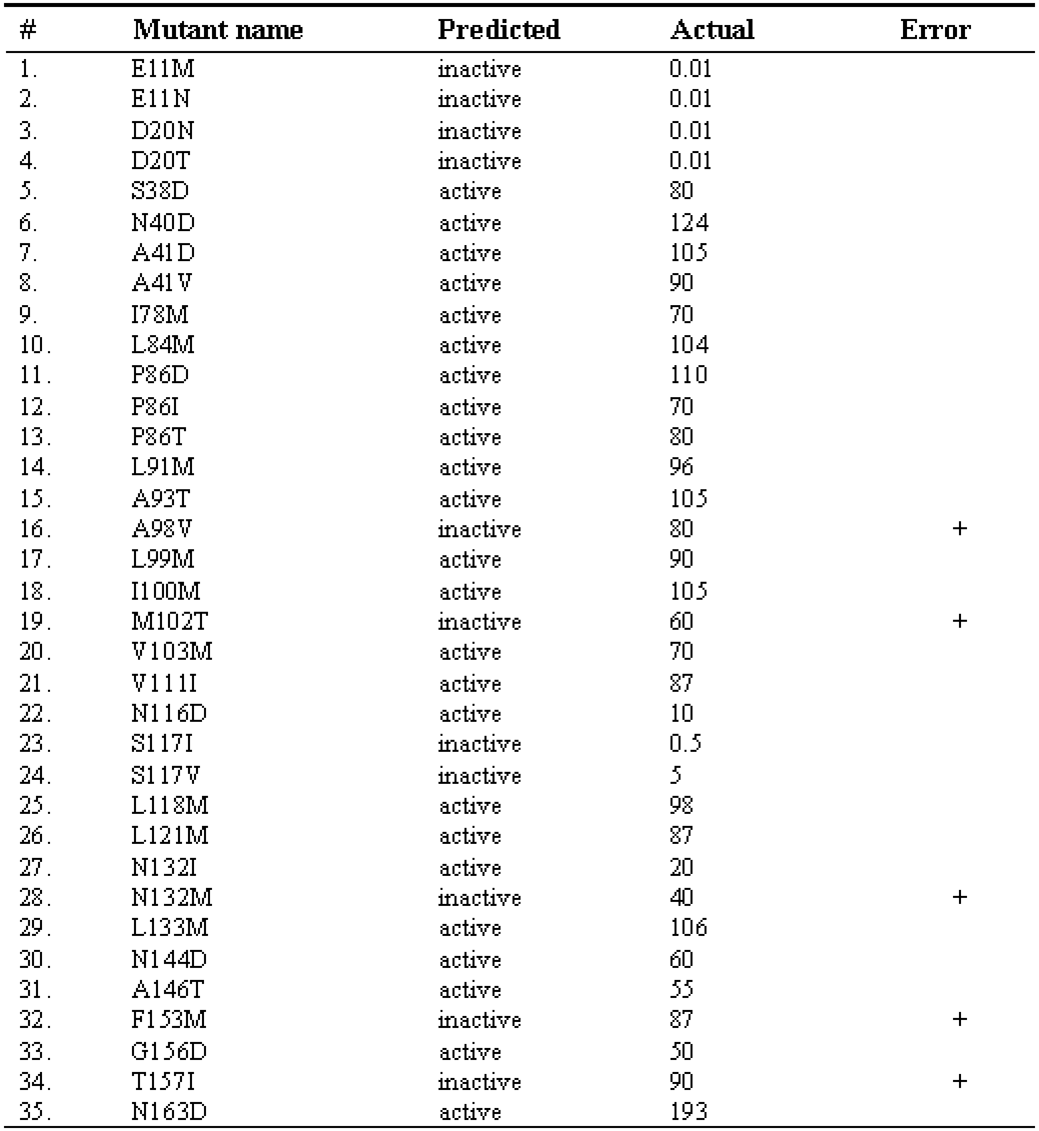
T4 Lysozyme Mutational Array



Training set mutants (n = 2015) Predicted test set mutants (n = 1101)

🟥 Active 🟩 Inactive 🟥 Active 🟩 Inactive

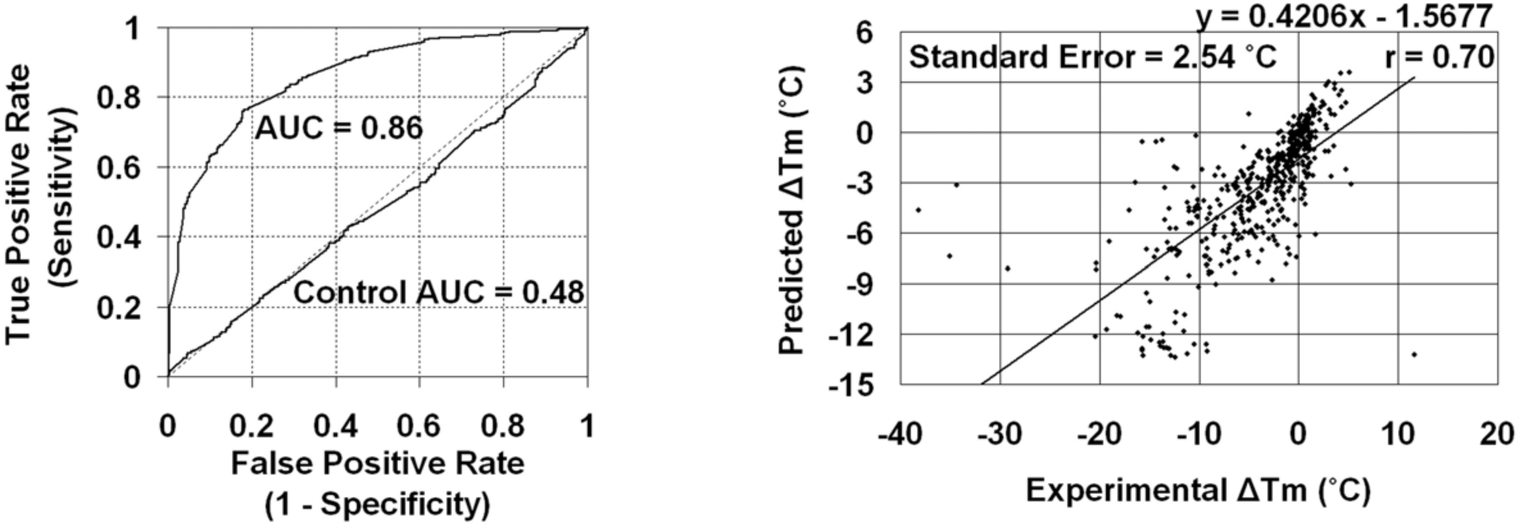
Real-World T4 Lysozyme Prediction Results



#	Mutant name	Predicted	Actual	Error
1.	E11M	inactive	0.01	
2.	E11N	inactive	0.01	
3.	D20N	inactive	0.01	
4.	D20T	inactive	0.01	
5.	S38D	active	80	
6.	N40D	active	124	
7.	A41D	active	105	
8.	A41V	active	90	
9.	I78M	active	70	
10.	L84M	active	104	
11.	P86D	active	110	
12.	P86I	active	70	
13.	P86T	active	80	
14.	L91M	active	96	
15.	A93T	active	105	
16.	A98V	inactive	80	+
17.	L99M	active	90	
18.	I100M	active	105	
19.	M102T	inactive	60	+
20.	V103M	active	70	
21.	V111I	active	87	
22.	N116D	active	10	
23.	S117I	inactive	0.5	
24.	S117V	inactive	5	
25.	L118M	active	98	
26.	L121M	active	87	
27.	N132I	active	20	
28.	N132M	inactive	40	+
29.	L133M	active	106	
30.	N144D	active	60	
31.	A146T	active	55	
32.	F153M	inactive	87	+
33.	G156D	active	50	
34.	T157I	inactive	90	+
35.	N163D	active	193	

- Experimental data (not part of training set) obtained from ProTherm database
- **Result:** predictions match experiments for 30/35 (~86%) of the mutants

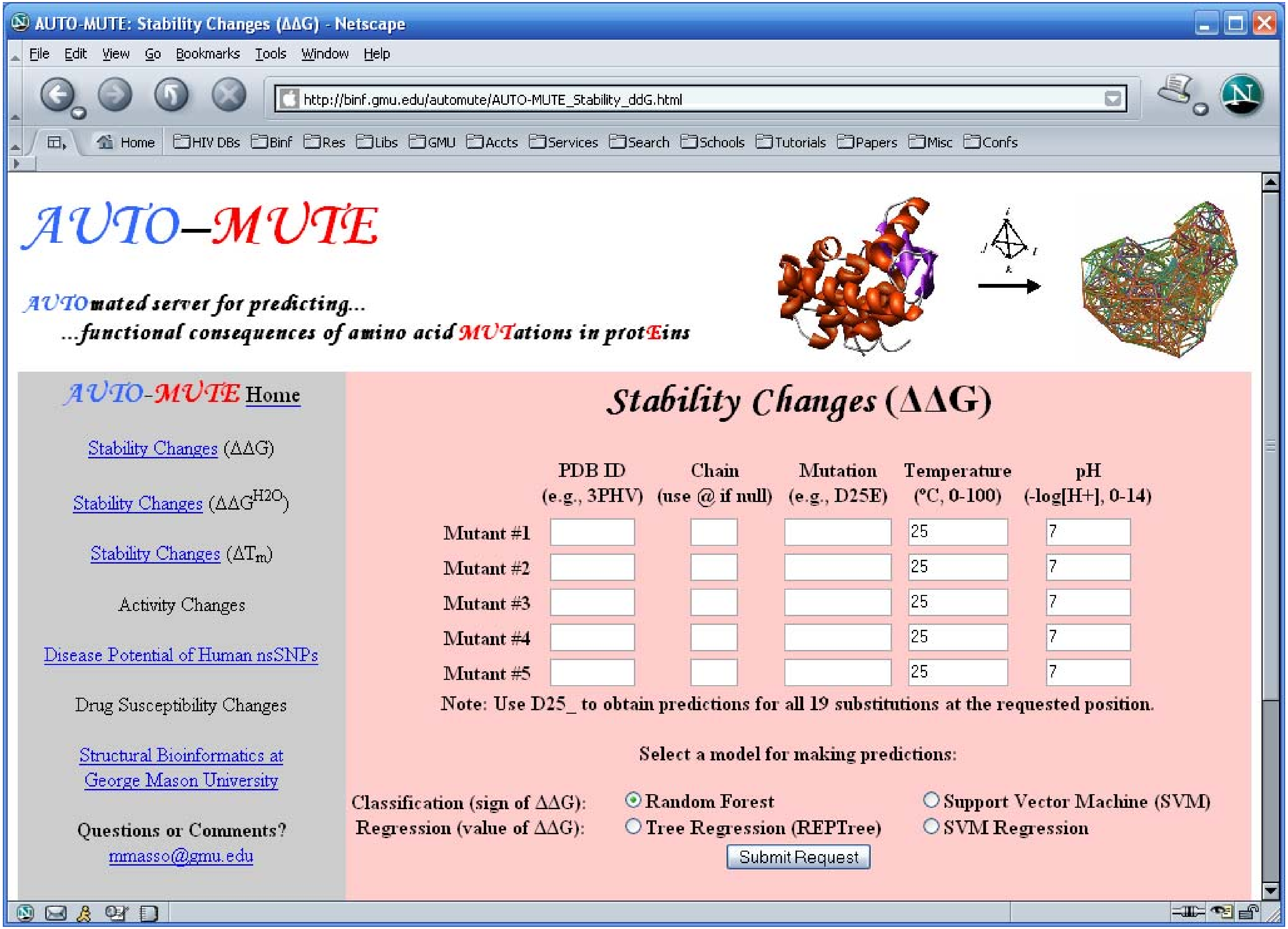
Algorithm Performance: T4 Lysozyme Activity and Stability Mutants



Left: Random forest algorithm, tenfold cross-validation, 2015 single-point activity mutants (1724 active and 291 inactive), overall accuracy is 80.4% (81.9% active class, 71.8% inactive class).
Right: Support vector regression algorithm, tenfold cross-validation, 507 single-point stability mutants.

Universal Models for Single-Point Mutants

- Current models are protein-specific since residual profile vectors of mutants from different proteins have different sizes
- New approach: use a **subset** of seven components (EC scores) extracted from the residual profile vector, corresponding to
  - the mutated position (residual score of the mutant protein)
  - the six nearest neighbors that participate in simplices with the mutated position, ordered by Euclidean distance away
- Include native and new amino acids at the mutated position, ordered amino acids at the six neighbors, and ordered primary sequence distance of the six neighbors from the mutated position
- Include location (surface, undersurface, or buried) and secondary structure (helix, strand, coil, turn) of the mutated position
- Include temperature as well as pH of experimental conditions



AUTO-MUTE: Stability Changes (ΔΔG) - Netscape
File Edit View Go Bookmarks Tools Window Help
http://binf.gmu.edu/automute/AUTO-MUTE_Stability_ddG.html
Home HIV DBs Binf Res Libs GMU Accts Services Search Schools Tutorials Papers Misc Confs

AUTO–MUTE

AUTOmated server for predicting...
...functional consequences of amino acid MUTations in protEins

AUTO-MUTE Home

Stability Changes (ΔΔG)

Stability Changes ($\Delta\Delta G^{H2O}$)

Stability Changes (ΔT_m)

Activity Changes

Disease Potential of Human nsSNPs

Drug Susceptibility Changes

Structural Bioinformatics at George Mason University

Questions or Comments?
mmasso@gmu.edu

Stability Changes (ΔΔG)

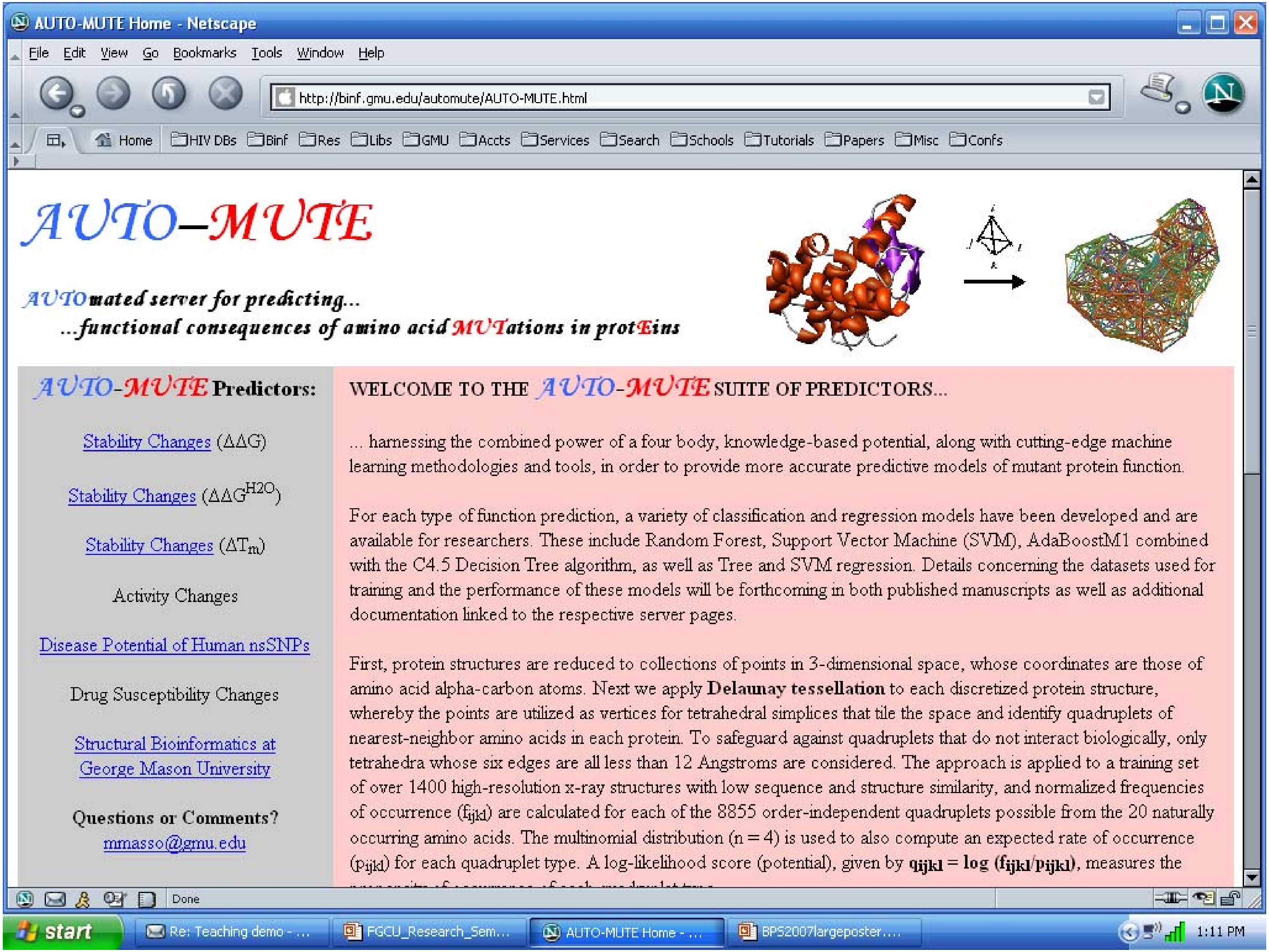
	PDB ID (e.g., 3PHV)	Chain (use @ if null)	Mutation (e.g., D25E)	Temperature (°C, 0-100)	pH (-log[H+], 0-14)
Mutant #1				25	7
Mutant #2				25	7
Mutant #3				25	7
Mutant #4				25	7
Mutant #5				25	7

Note: Use D25_ to obtain predictions for all 19 substitutions at the requested position.

Select a model for making predictions:

Classification (sign of ΔΔG): ● Random Forest ○ Support Vector Machine (SVM)
Regression (value of ΔΔG): ○ Tree Regression (REPTree) ○ SVM Regression

Submit Request


AUTO-MUTE Home - Netscape
File Edit View Go Bookmarks Tools Window Help
http://binf.gmu.edu/automute/AUTO-MUTE.html
Home HIV DBs Binf Res Libs GMU Accts Services Search Schools Tutorials Papers Misc Confs
Done
start
Re: Teaching demo - ...
FGCU_Research_Sem...
AUTO-MUTE Home - ...
BPS2007largeposter....
1:11 PM

AUTO–MUTE

AUTOmated server for predicting...
...functional consequences of amino acid MUTations in protEins

AUTO-MUTE Predictors:

Stability Changes ($\Delta\Delta G$)

Stability Changes ($\Delta\Delta G^{H2O}$)

Stability Changes (ΔT_m)

Activity Changes

Disease Potential of Human nsSNPs

Drug Susceptibility Changes

Structural Bioinformatics at George Mason University

Questions or Comments?
mmasso@gmu.edu

WELCOME TO THE AUTO-MUTE SUITE OF PREDICTORS...

... harnessing the combined power of a four body, knowledge-based potential, along with cutting-edge machine learning methodologies and tools, in order to provide more accurate predictive models of mutant protein function.

For each type of function prediction, a variety of classification and regression models have been developed and are available for researchers. These include Random Forest, Support Vector Machine (SVM), AdaBoostM1 combined with the C4.5 Decision Tree algorithm, as well as Tree and SVM regression. Details concerning the datasets used for training and the performance of these models will be forthcoming in both published manuscripts as well as additional documentation linked to the respective server pages.

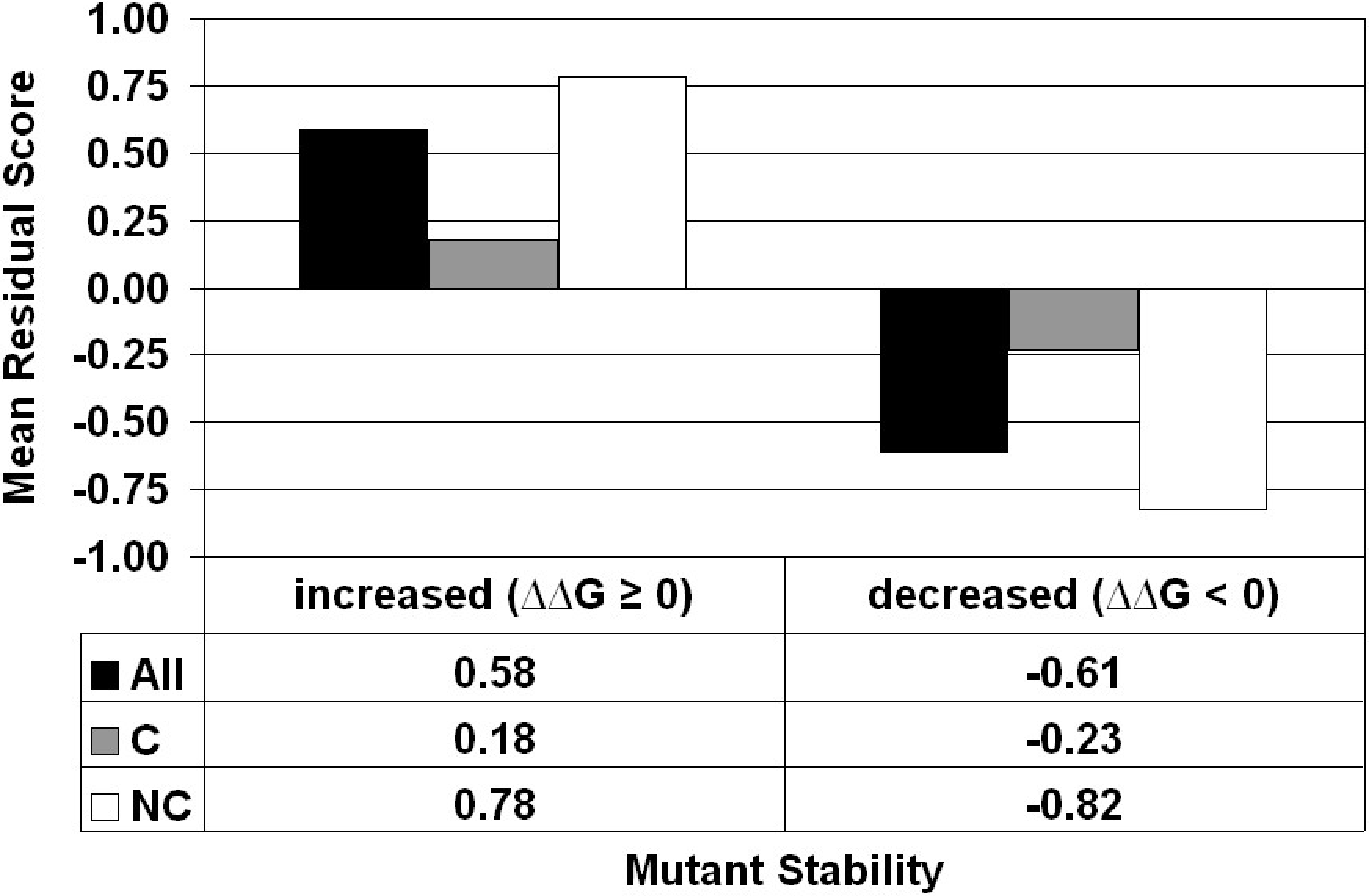
First, protein structures are reduced to collections of points in 3-dimensional space, whose coordinates are those of amino acid alpha-carbon atoms. Next we apply **Delaunay tessellation** to each discretized protein structure, whereby the points are utilized as vertices for tetrahedral simplices that tile the space and identify quadruplets of nearest-neighbor amino acids in each protein. To safeguard against quadruplets that do not interact biologically, only tetrahedra whose six edges are all less than 12 Angstroms are considered. The approach is applied to a training set of over 1400 high-resolution x-ray structures with low sequence and structure similarity, and normalized frequencies of occurrence (f_{ijkl}) are calculated for each of the 8855 order-independent quadruplets possible from the 20 naturally occurring amino acids. The multinomial distribution ($n = 4$) is used to also compute an expected rate of occurrence (p_{ijkl}) for each quadruplet type. A log-likelihood score (potential), given by $\mathbf{q_{ijkl} = \log(f_{ijkl}/p_{ijkl})}$, measures the propensity of occurrence of each quadruplet type.

ΔΔG Dataset Used to Train Models

- Over 1900 single-site mutants derived from 53 proteins with low sequence and structure homology
- All protein structures are tessellatable
- Experimental stability of each mutant reported as the free energy of unfolding ($\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wt}$) in kcal/mol
- Data collected from the ProTherm database by Capriotti *et al.*

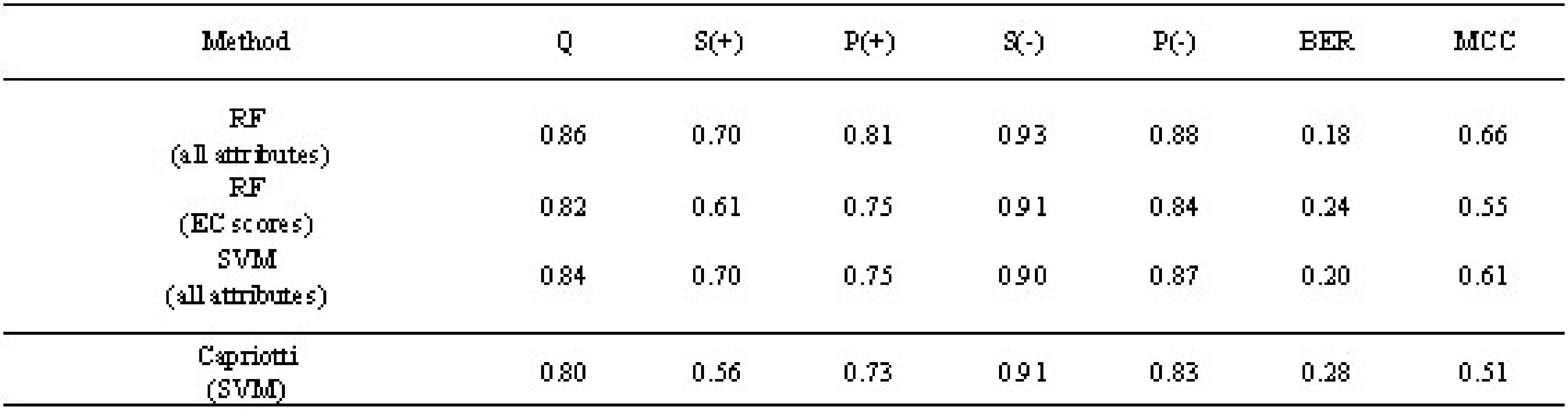
  Bava, K.A., Gromiha, M.M., Uedaira, H., Kitajima, K. and Sarai, A. (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.,* **32**, D120–D121.

  Capriotti, E., Fariselli, P. and Casadio, R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.,* **33**, W306–W310.
- Additional experimental data in ProTherm for each mutant includes temp. (˚C) and pH; relative accessibility (RSA) for each mutant computed with the DSSP program by Capriotti *et al.*

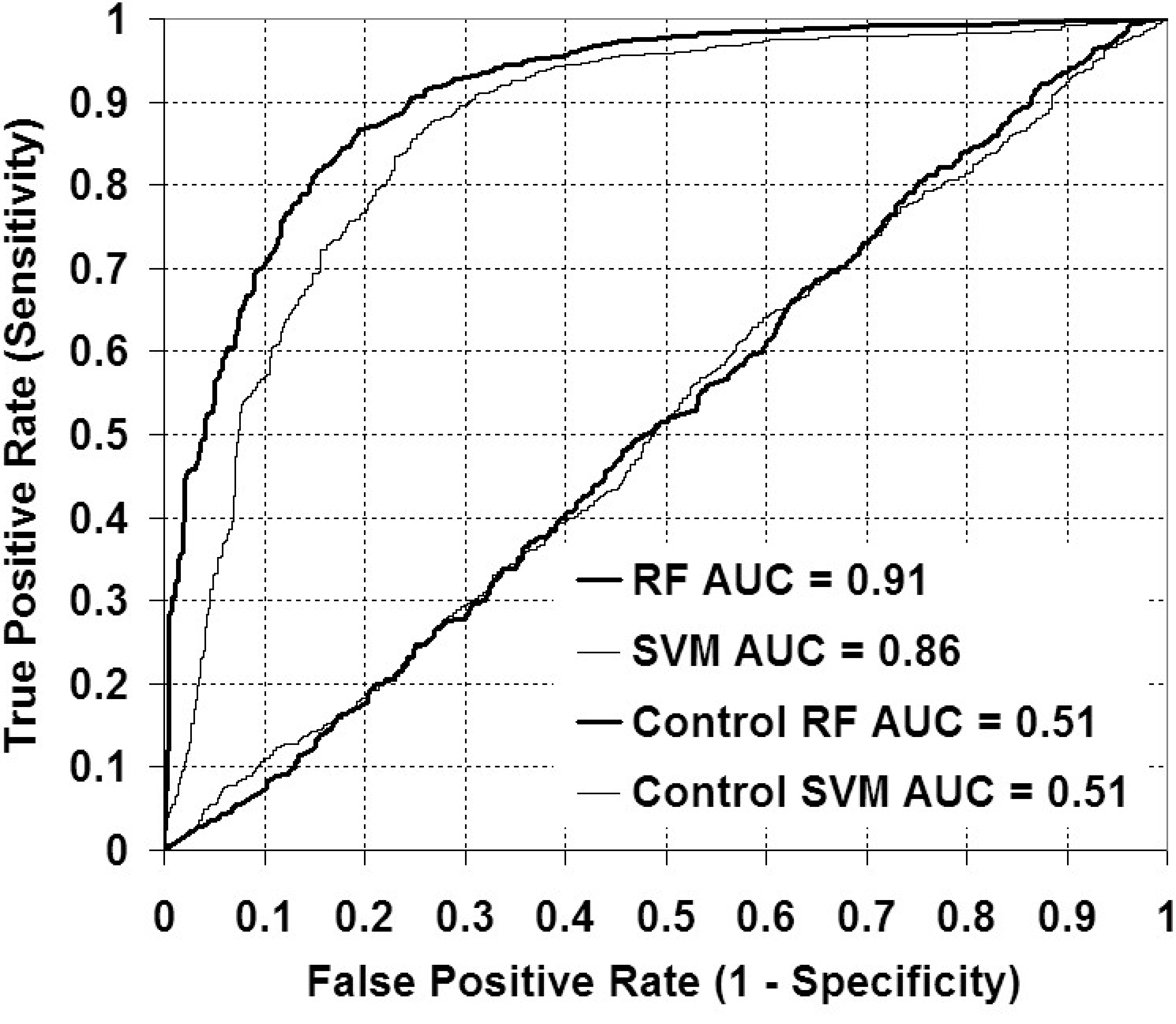


	increased (ΔΔG ≥ 0)	decreased (ΔΔG < 0)
■ All	0.58	-0.61
▨ C	0.18	-0.23
□ NC	0.78	-0.82

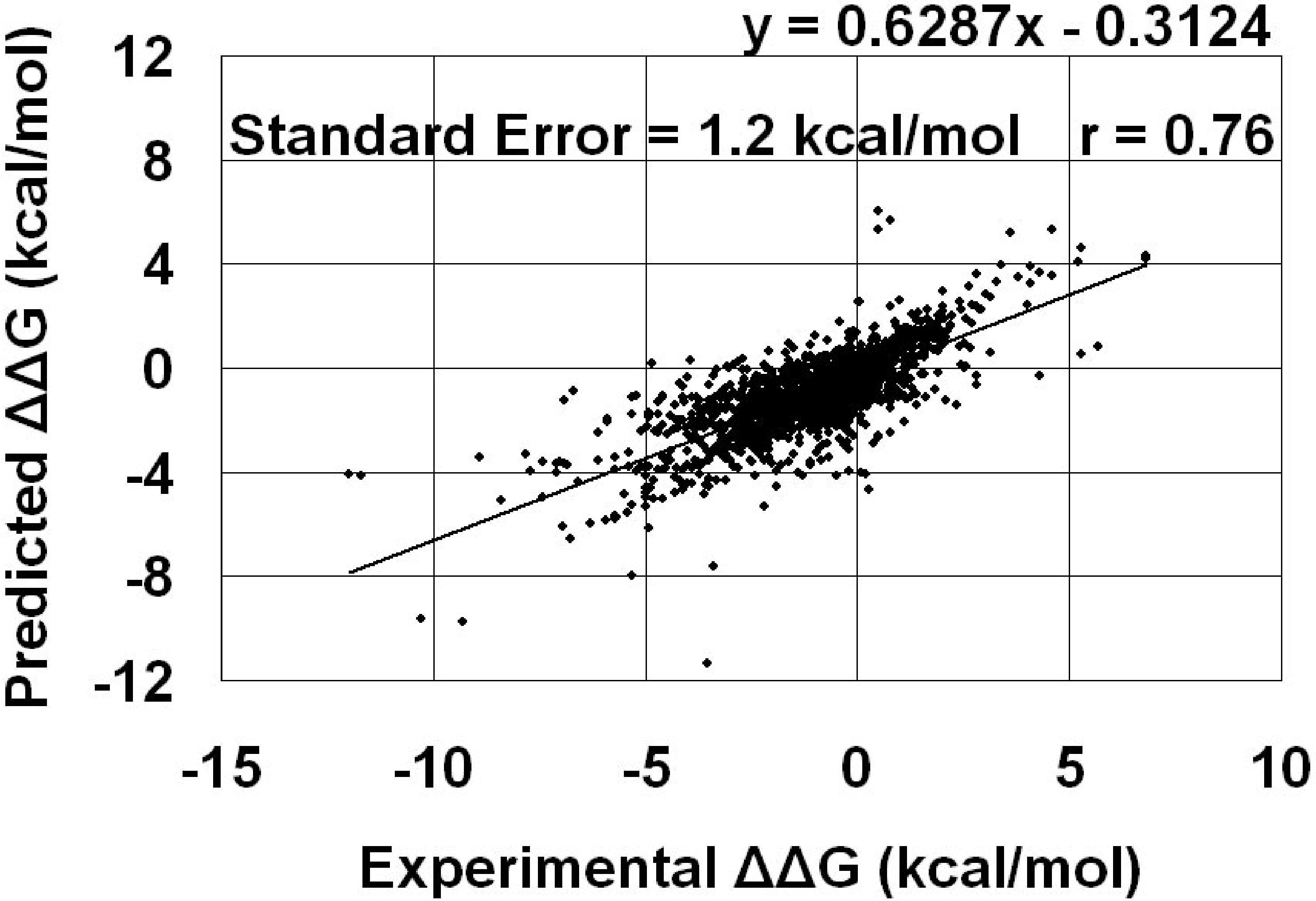
Supervised Classification Performance Measures



Method	Q	S(+)	P(+)	S(-)	P(-)	BER	MCC
RF (all attributes)	0.86	0.70	0.81	0.93	0.88	0.18	0.66
RF (EC scores)	0.82	0.61	0.75	0.91	0.84	0.24	0.55
SVM (all attributes)	0.84	0.70	0.75	0.90	0.87	0.20	0.61
Capriotti (SVM)	0.80	0.56	0.73	0.91	0.83	0.28	0.51

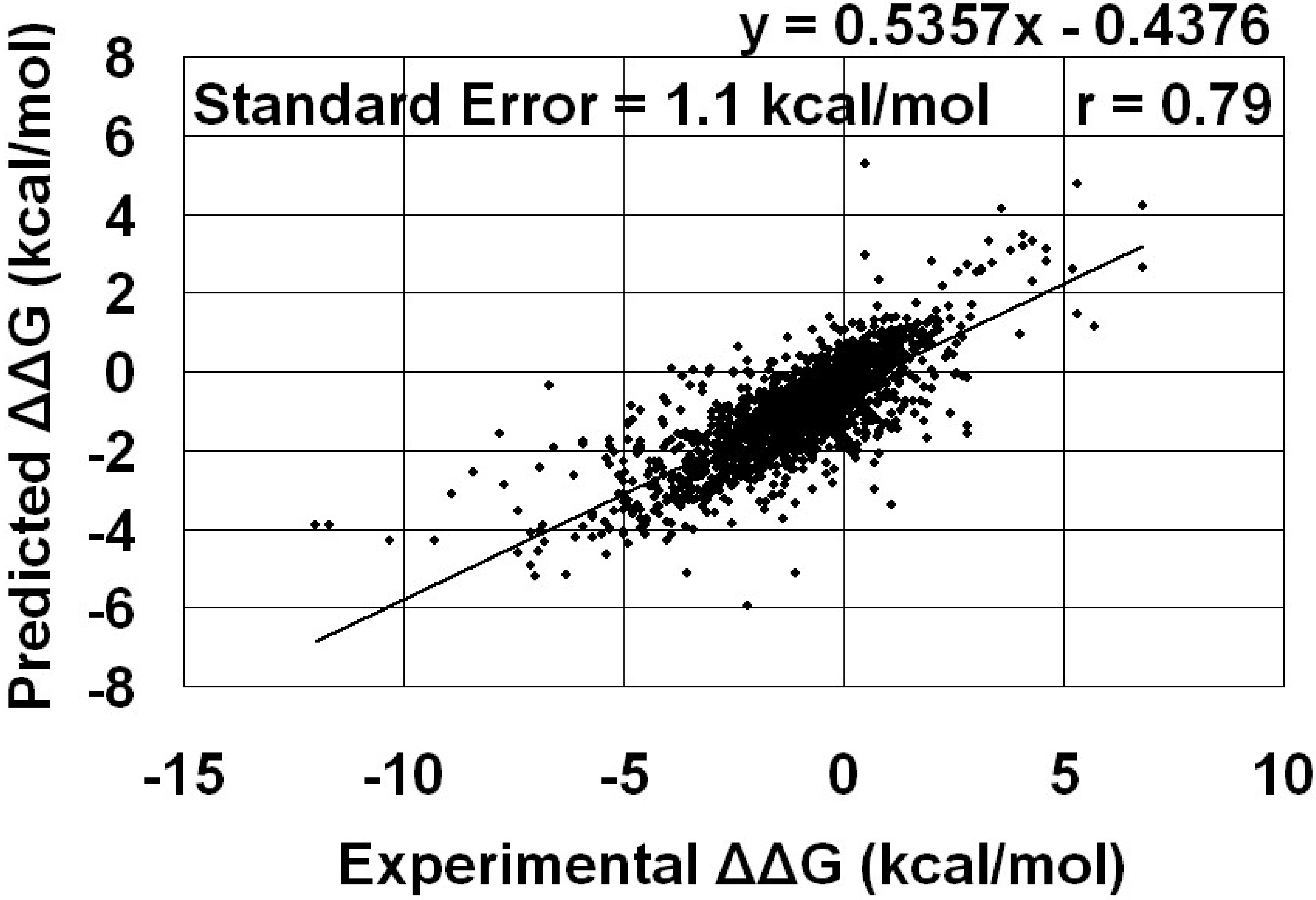

True Positive Rate (Sensitivity)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
False Positive Rate (1 - Specificity)
RF AUC = 0.91
SVM AUC = 0.86
Control RF AUC = 0.51
Control SVM AUC = 0.51

Support Vector Regression



Capriotti *et al.* SVM regression (for comparison):
r = 0.71, Standard Error = 1.3 kcal/mol, $y = 0.5223x - 0.4705$

Tree Regression (REPTree)



Comparison of classification algorithms on S388

Method	Q	S(+)	P(+)	S(-)	P(-)	BER	MCC
RF (all attributes)	0.87	0.36	0.42	0.94	0.92	0.35	0.31
Cheng (SVM/ST)	0.86	0.31	0.40	0.93	0.91	0.38	0.27
Capriotti (NN)	0.87	0.21	0.44	0.96	0.90	0.42	0.25
PoPMuSiC[a]	0.85	0.25	0.33	0.93	0.90	0.41	0.20
DFIRE[b]	0.68	0.44	0.18	0.71	0.90	0.43	0.11
FOLDX[c]	0.75	0.56	0.26	0.78	0.93	0.33	0.25

[a] http://babylone.ulb.ac.be (Gilis and Rooman, 1997; Kwasigroch *et al.*, 2002)

[b] http://sparks.informatics.iupui.edu (Zhou and Zhou, 2002)

[c] http://fold-x.embl-heidelberg.de (Guerois *et al.*, 2002)

Conclusions and Future Directions

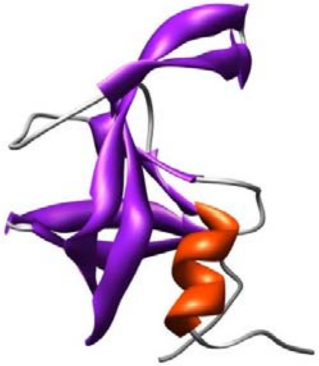
- A novel computational mutagenesis arising from a four-body, knowledge-based statistical potential uniquely characterizes each protein mutant using properties of sequence and structure
- Descriptors correlate well with mutant function and are valuable for developing accurate predictive models by combining with machine learning tools (novel approach not described in literature)
- **Future work:**
  - Develop atomic-level four-body statistical potentials
    - How to define alphabet?
    - Distinguish between protein and ligand atoms?
  - Apply to the development of predictive models
    - protein-protein interactions
    - protein-ligand binding energies

Acknowledgements and References

People

Iosif Vaisman Todd Taylor

Andrew Carr Vadim Ravich



Software

Qhull – tessellation (Barber)

Glisten – tessellation visualization (Carr); also Matlab

Chimera – ribbon diagram (Ferrin)

Ad hoc Java programs – potential (Taylor), residual profiles (Lu)

Weka – machine learning (Witten, Frank)

Publications

1. Masso M. and Vaisman I. Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach. *Biochem Biophys Res Comm* 2003; **305**: 322-326.
2. Masso M., Lu Z., and Vaisman I. Computational studies of protein structure-function correlations. *Proteins* 2006; **64**: 234-245.
3. Masso M. and Vaisman I. A novel sequence-structure approach for accurate prediction of resistance to HIV-1 protease inhibitors. *IEEE Proc BIBE* 2007; 952-958.
4. Masso M. and Vaisman I. Accurate prediction of enzyme mutant activity based on a multibody statistical potential. *Bioinformatics* (**in press**).
5. Barenboim M., Masso M., Vaisman I., and Jamison C. Statistical geometry based prediction of non-synonymous SNP functional effects using random forest and neuro-fuzzy classifiers. *Proteins* (**in press**).