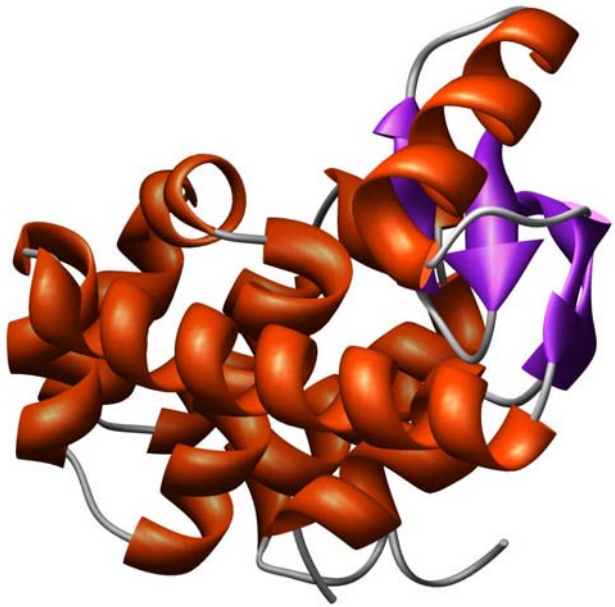
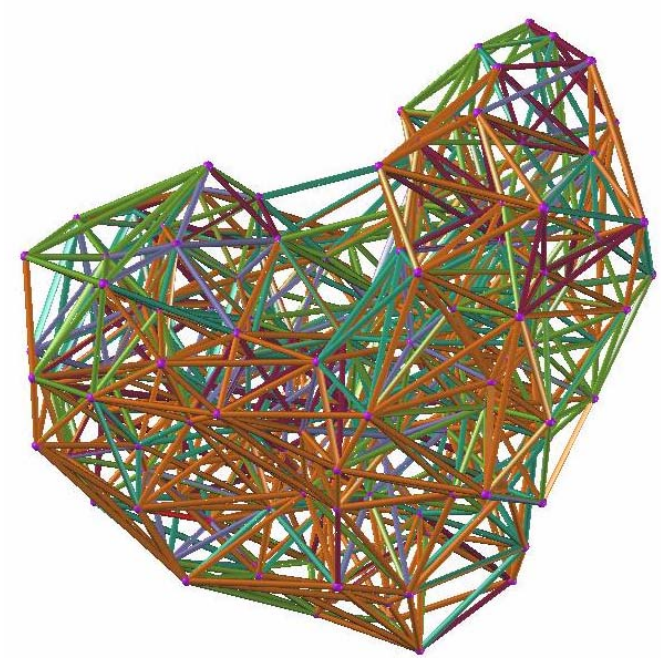
Applications of Statistical Geometry to the Functional Analysis of Protein Mutants



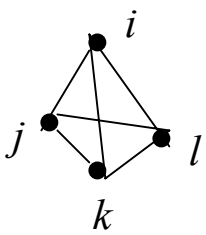


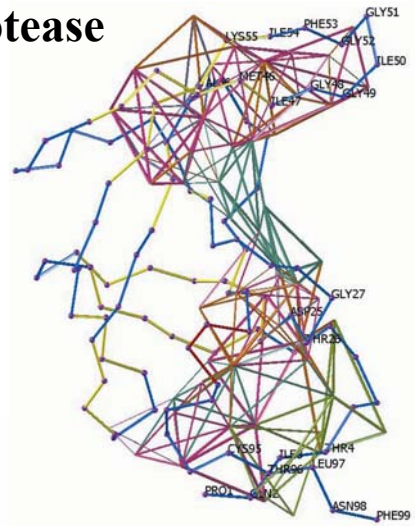
Majid Masso

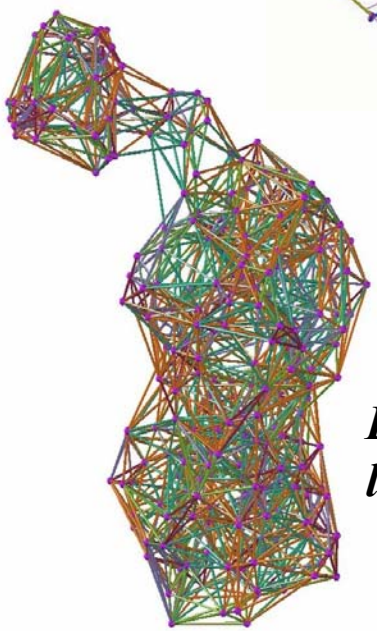
Ph.D. Dissertation Defense

Four-Body Statistical Potential

HIV-1 protease





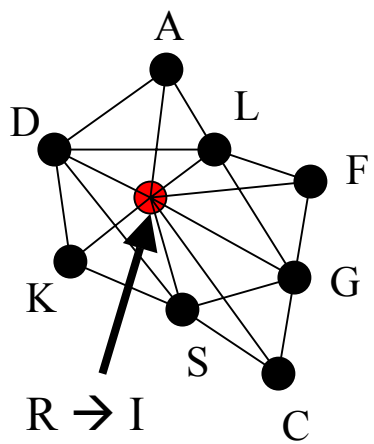


***E. coli* *lac* repressor**

- Protein structures are represented as discrete sets of points in 3D, each corresponding to an amino acid (aa)
- Delaunay tessellation of a protein structure yields an aggregate of space-filling, non-overlapping, irregular tetrahedra (simplices) that each define a quadruplet of nearest-neighbor aa's
- A four-body statistical potential function is derived via tessellation of a training set of structures, assigning a log-likelihood score to all possible quadruplets of aa's

Computational Mutagenesis Methodology

- *Total Potential* or *Topological Score* of a protein structure, a global measure of sequence-structure compatibility, is obtained by summing the scores of all the simplices in the tessellation
- *Individual Residue Potential* or *Residue Environment Score* of each aa in a protein structure is obtained by locally summing the scores of only the simplices that use the aa's point representation as a vertex; the scores of all the aa's form a *Potential Profile* vector



- Assumption: minor structural differences and similar tessellations between each mutant and the wild-type (wt) protein
- Approach: the total potential and potential profile of every mutant can be derived from the tessellation of the wt structure

Computational Mutagenesis Methodology

Based on the methodology, each mutant is characterized by a scalar *Residual Score* and a vector *Residual Profile*:

- *Residual Score* – difference between mutant and wt total potentials
  - Measures the relative change in mutant sequence-structure compatibility from wt
- *Residual Profile* – difference between mutant and wt potential profiles
  - Quantifies environmental perturbations from wt at every aa position
  - Each component in the profile is referred to as an *environmental change (EC) score* for the corresponding aa position

Comprehensive Mutational Profile (CMP)

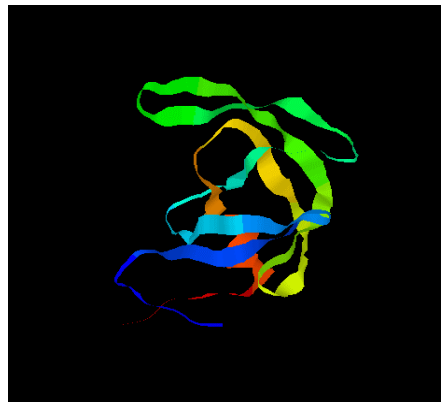
- At each residue position in a protein structure, a CMP score is obtained by calculating the mean of the 20 residual scores associated with all possible aa replacements (including the degenerate mutant obtained by substituting the wt aa with itself, with residual score 0)
- Mathematically,

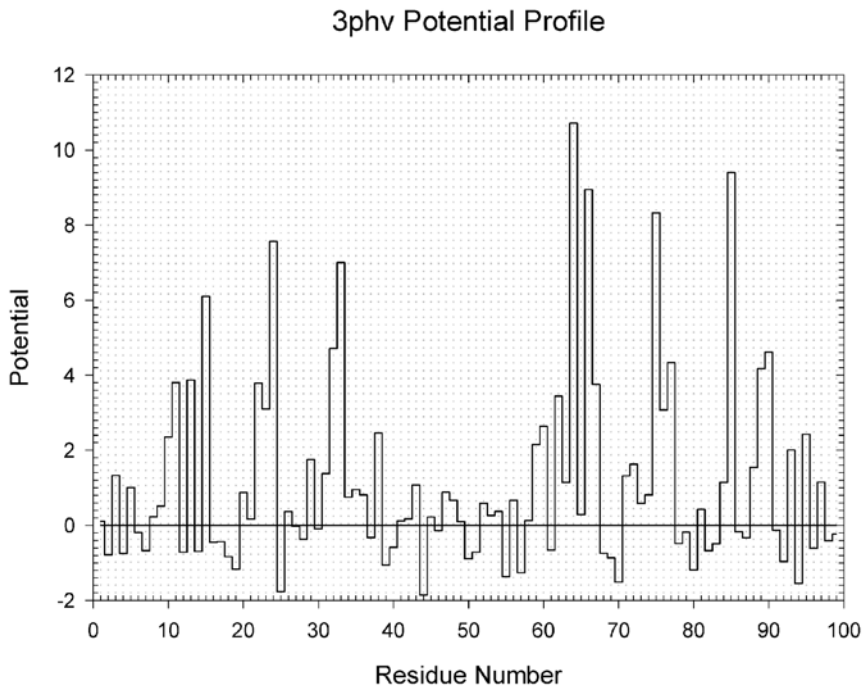
$$\begin{aligned} \mathrm{CMP}_j &= \tfrac{1}{20}\sum_{i=1}^{20} \left[(\text{mutant topological score})_{ij} - (\text{wt topological score})\right] \\ &= \tfrac{1}{20}\sum_{i=1}^{20} (\text{mutant residual score})_{ij} \\ &= \left\{\text{mean residual score}\right\}_j \end{aligned}$$

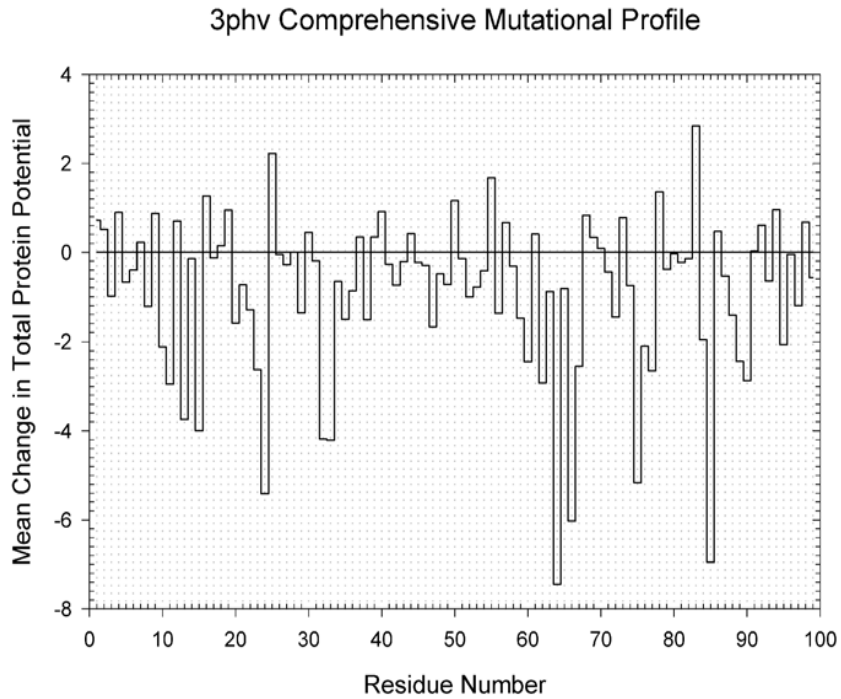
where index *i* refers to the 20 aa's, and index *j* refers to the position in the 1° sequence of the protein

CMP Example: HIV-1 Protease

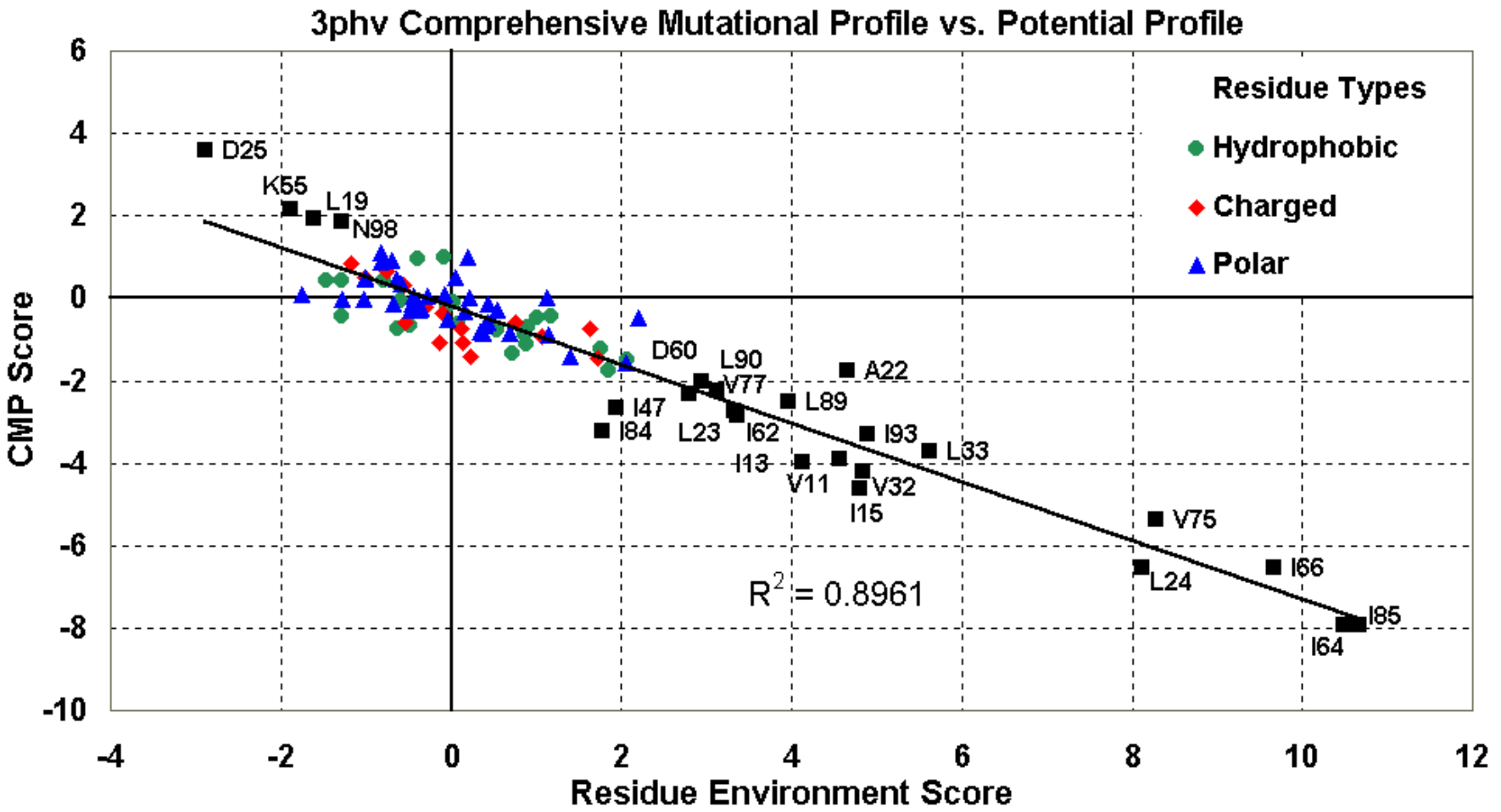
- PDB ID: 3phv (monomer, 99 aa's)
- Functional as a homodimer
  - Interface: P1-T4 and C95-F99
  - Catalytic triad: D25-T26-G27
  - Flap region: M46-V56







CMP Example: HIV-1 Protease



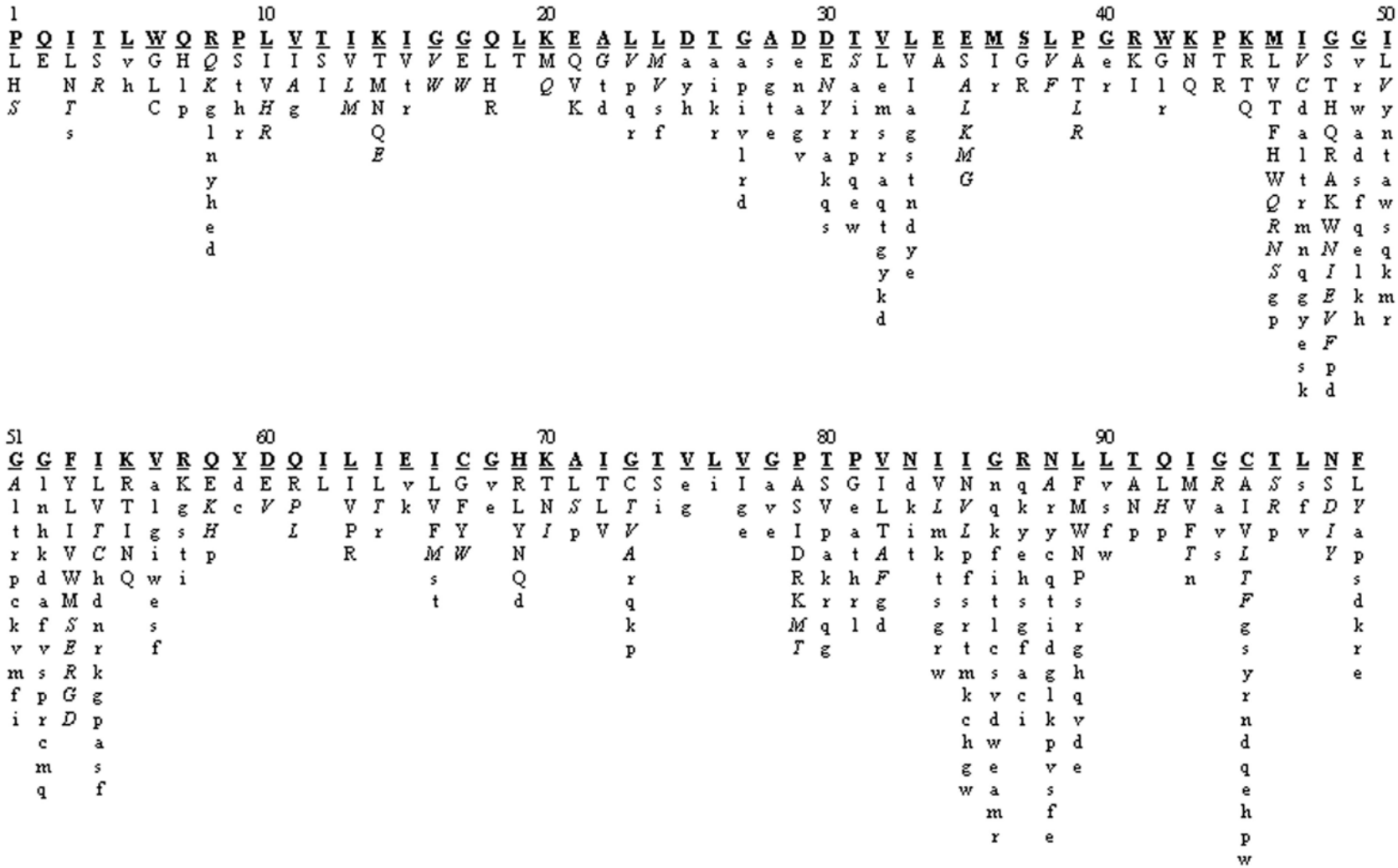
HIV-1 Protease Experimental Data

- Synthesis and analysis of 536 single site missense mutants
  - 336 published mutants: Loeb, D.D., Swanstrom R., Everitt, L., Manchester, M., Stamper, S.E. & Hutchison III, C.A. (1989) Complete mutagenesis of the HIV-1 protease. *Nature* **340**, 397-400.
  - 200 mutants provided by R. Swanstrom (UNC)

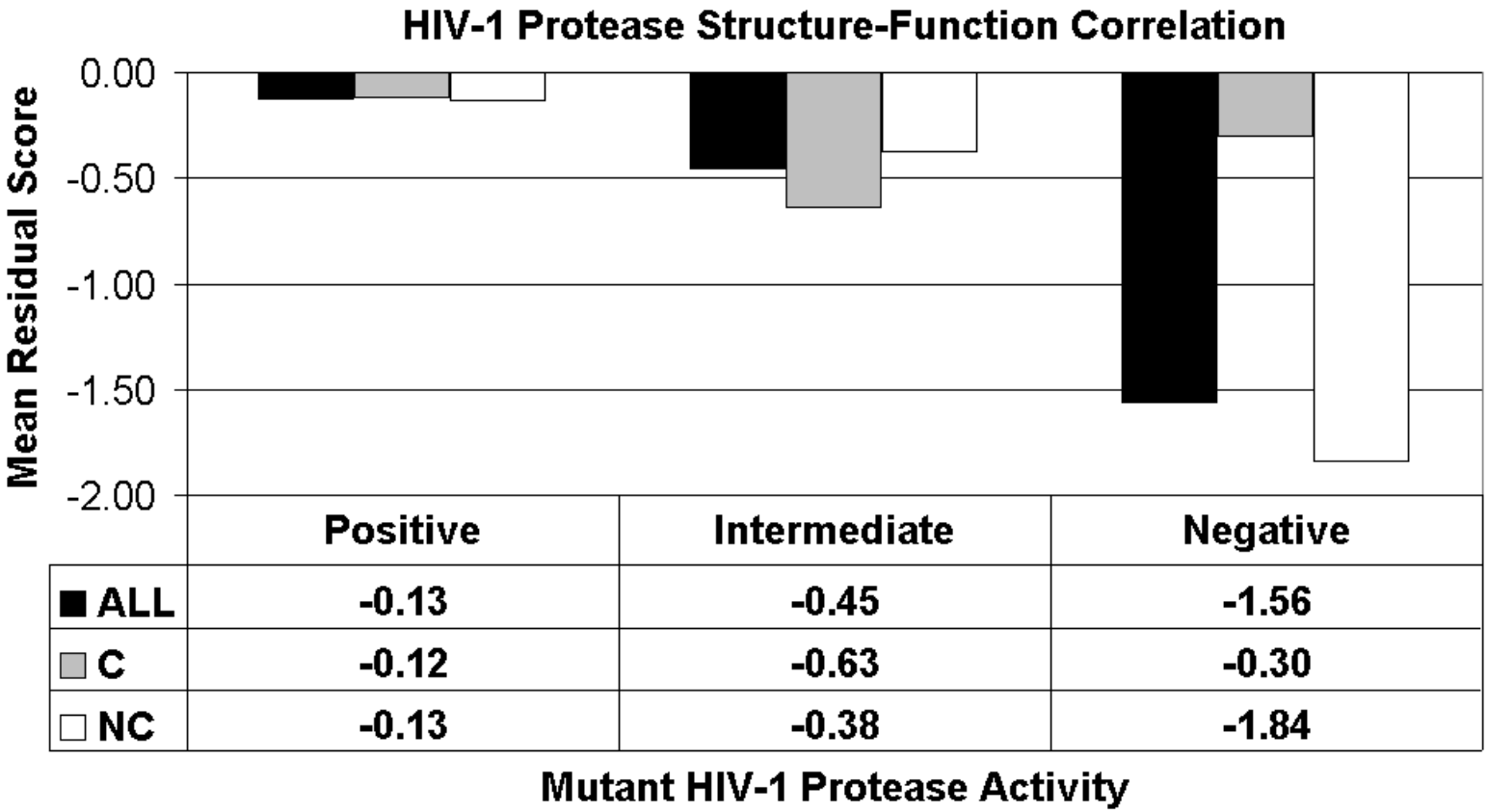
- Each mutant placed in one of 3 phenotypic categories, positive, negative, or intermediate, based on activity (ability to process the Pol polyprotein)

- Residual scores of the mutants can be used to elucidate the structure-function relationship in HIV-1 protease

HIV-1 Protease Experimental Data



Structure-Function Correlations Based on Residual Scores: **HIV-1 Protease**

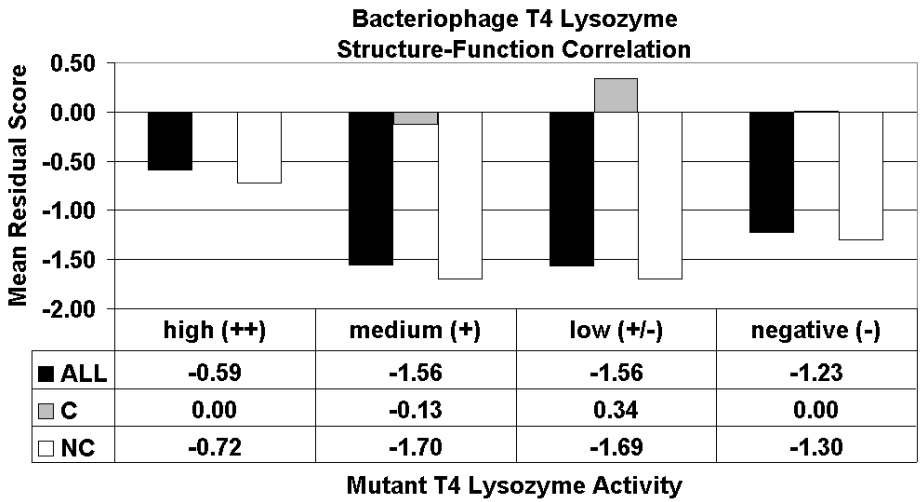


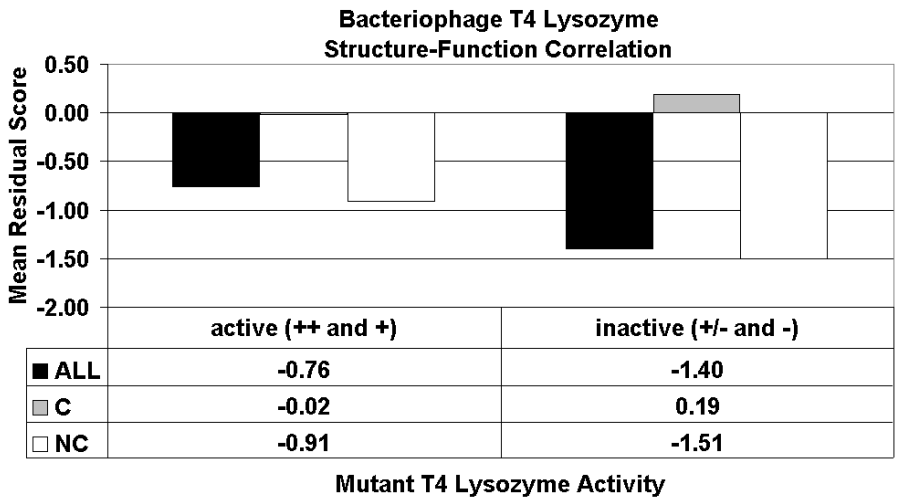
How significant are the differences in class-pair means?

Pos-Neg: $p = 1.65 \times 10^{-11}$; Int-Neg: $p = 9.90 \times 10^{-6}$; and Pos-Int: $p = 0.086$.

Structure-Function Correlations Based on Residual Scores: **Bacteriophage T4 Lysozyme**

- Experimental data: 2015 single site mutants generated by introducing the same 13 aa replacements at 163/164 positions - all but M1 (PDB ID: 3lzm)
  - Rennell, D., Bouvier, S.E., Hardy, L.W. & Poteete, A.R. (1991) Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222**, 67-88.
- Four mutant activity classes: high, medium, low, negative
- Investigators recommend data analysis using only two classes (active = high + med, inactive = low + neg): $p = 0.0003$

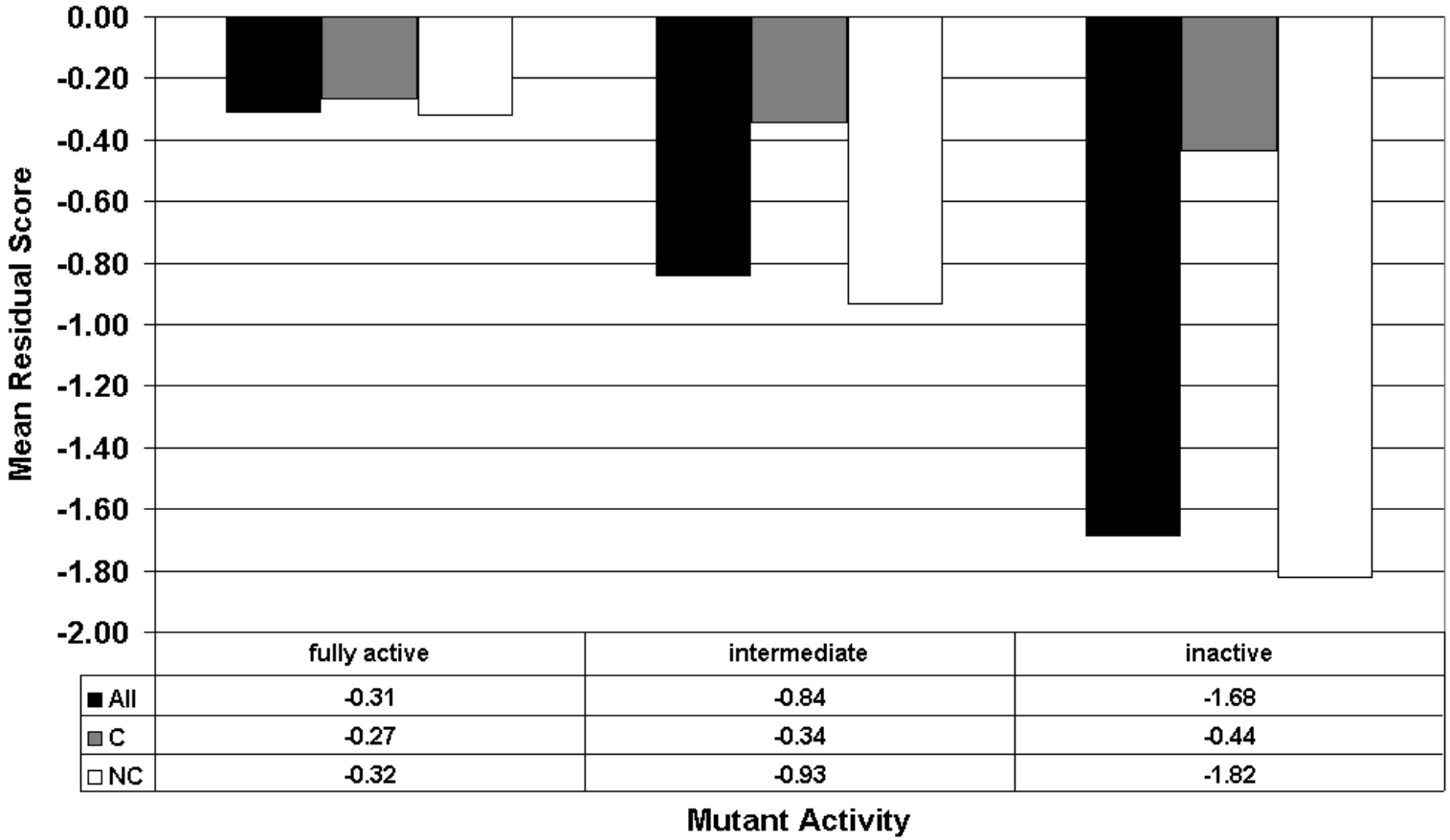




Structure-Function Correlations Based on Residual Scores: *E. coli Lac* **Repressor**

- Experimental data: 4041 single site mutants generated by introducing the same 13 aa replacements at positions 2-329 (PDB ID: 1efaB)
  - Markiewicz, P., Kleina, L.G., Cruz, C., Ehret, S. & Miller, J.H. (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J. Mol. Biol.* **240**, 421-433.
- Four mutant activity classes based on degree of repression of β-galactosidase: fully active (greater than 200-fold), moderate (20 to 200-fold), low (4 to 20-fold), inactive (less than 4-fold)
- Investigators suggest combining moderate + low = intermediate
- Recent computational studies using this data set define two classes: unaffected (fully active) and affected (all other classes combined)
- All 328 *lac* repressor residue positions were annotated and clustered into 15 groups based on their structural locations, functional roles, and level of tolerance to mutations

Structure-Function Correlations Based on Residual Scores: *E. coli Lac* **Repressor**

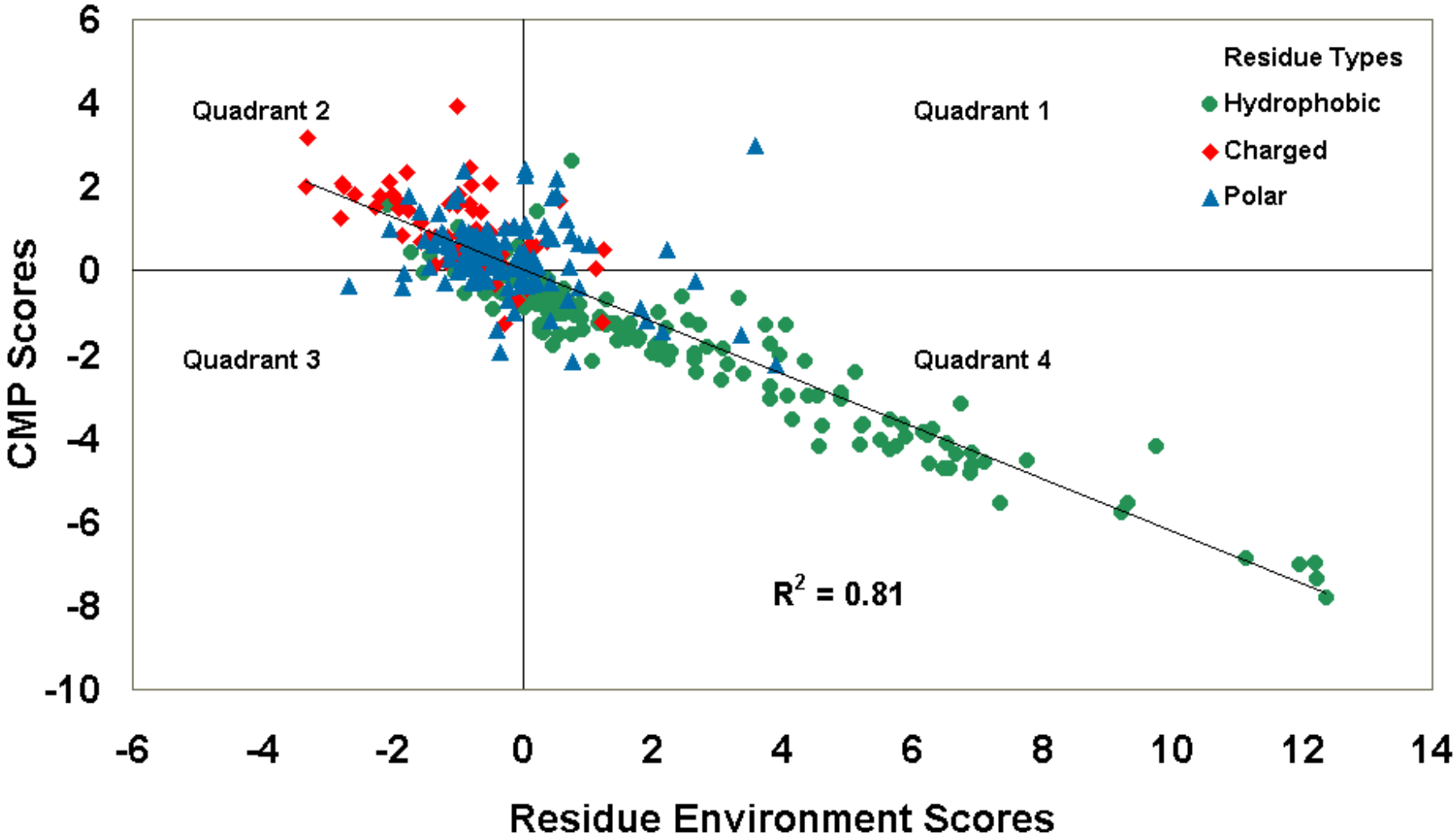


	fully active	intermediate	inactive
■ All	-0.31	-0.84	-1.68
■ C	-0.27	-0.34	-0.44
□ NC	-0.32	-0.93	-1.82

How significant are the differences in class-pair means?

full-inter: $p = 4.64 \times 10^{-7}$; full-inactive: $p = 1.95 \times 10^{-36}$; and inter-inactive: $p = 6.57 \times 10^{-10}$.

Lac Repressor: CMP vs. Potential Profile

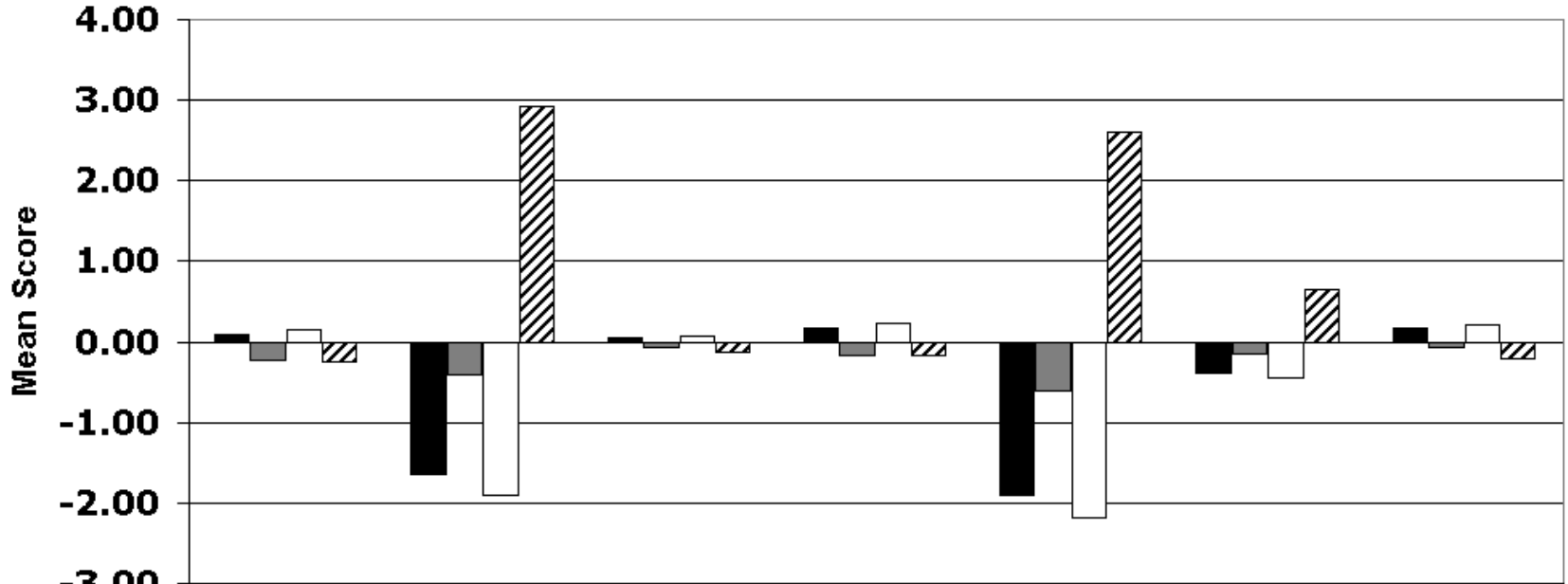


Distribution of *Lac* Repressor Residue Positions

Graph Quadrants	Residue Groups: Surface	Buried	DNA binding	IPTG binding	Stability	Interface	Spacers	Total
Q1	8	10	0	2	1	6	4	31
Q2	49	12	9	9	8	15	20	122
Q3	13	5	4	2	2	5	6	37
Q4	31	46	5	4	25	17	10	138
Total	101	73	18	17	36	43	40	328

Apply chi-square test with 18 df: $\chi^2 = 51.11$, so reject null hypothesis that no association exists between structural/functional groups and quadrant locations, with $p < 0.0001$

Characterizing Structural or Functional Roles of *Lac* Repressor Residues Based on Residual Scores and Residue Environment Scores



	Surface	Buried	DNA binding	IPTG binding	Stability	Interface	Spacers
All	0.08	-1.65	0.04	0.17	-1.91	-0.39	0.16
C	-0.23	-0.42	-0.07	-0.17	-0.62	-0.15	-0.08
NC	0.14	-1.91	0.07	0.24	-2.17	-0.44	0.21
M.R.E.S.	-0.26	2.91	-0.12	-0.17	2.61	0.65	-0.20

Mutant Residual Profiles: Motivation

- Residual profile vectors encode much more sequence and structure information about the mutants than residual scores; hence, they may prove to be more useful for classification and inference for mutants belonging to different activity classes
- Nonzero components (EC scores) of a mutant residual profile identify the mutated position(s) as well as all of their topological nearest-neighbors based on tessellation (i.e., all positions that participate in simplices with the mutated positions)
- For any single site mutant, the EC score at the residual profile component corresponding to the mutated position is precisely the residual score of the mutant
- A consequence of the above is that all 19 single site mutants at a particular position have residual profiles w/ identical arrangements of zero and nonzero components (only the EC scores at any given nonzero component differ among the 19 residual profiles)

HIV-1 Protease Dataset: Residual Profiles of the Experimental Mutants

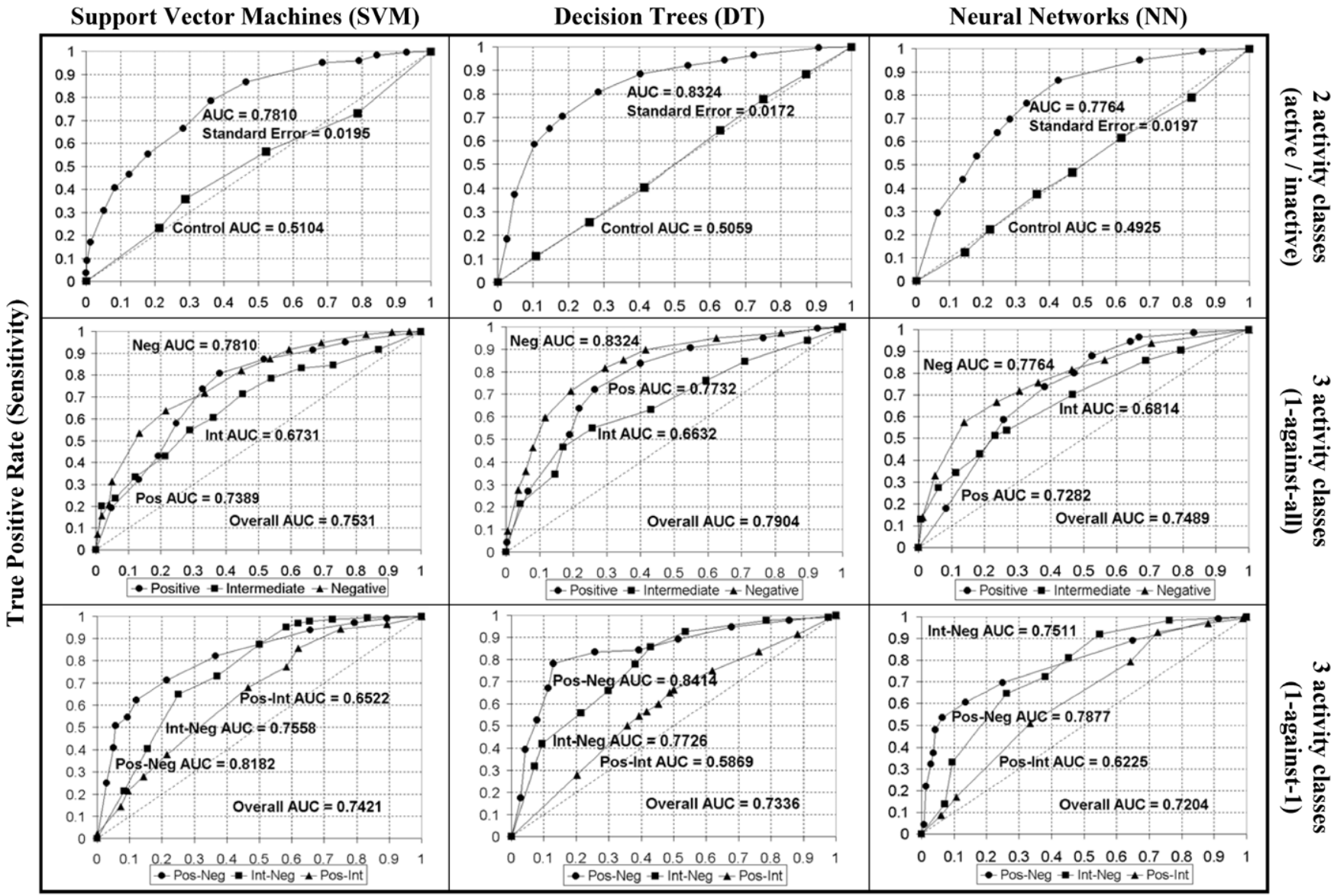
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	...	CW	CX	CY
1	PRO	1	HIS	1.89369	0.12473	0.2462	-0.01137	0	0	0	0	0	0	0		0.2482	0	pos
2	PRO	1	LEU	1.61399	-0.21225	1.51021	0.14456	0	0	0	0	0	0	0		-0.7566	0	pos
3	PRO	1	SER	0.80073	0.19565	0.14197	0.15969	0	0	0	0	0	0	0		0.30934	0	int
4	GLN	2	GLU	-0.6395	-1.55273	-0.24116	-1.33969	-0.4477	-0.41718	0	0	0	0	0		-0.29306	-0.31513	pos
5	ILE	3	ASN	-0.32949	0.76726	-2.46203	0.5757	-1.49592	0	0.31665	0	-0.93573	-0.49091	-1.47315		0.46809	0	pos
6	ILE	3	LEU	0.35974	0.41178	1.5984	0.10011	0.37716	0	0.2498	0	0.42616	0.2479	0.19533		0.50297	0	pos
7	ILE	3	SER	0.35207	0.88747	-1.14271	0.53599	-1.30293	0	0.40746	0	-0.52978	-0.29686	-1.07501		0.38893	0	neg
8	ILE	3	THR	0.28471	0.89302	-0.3196	0.72597	-1.06583	0	0.60907	0	-0.17343	-0.1048	-0.43737		0.29873	0	int
9	THR	4	ARG	-0.36146	-0.33689	-0.18267	-0.34217	-0.43148	0.00263	0.25453	0	-0.16441	0	0		-0.18464	-0.18971	int
10	THR	4	SER	0.03021	-0.26497	-0.21622	-0.33293	-0.23951	0.0838	-0.11714	0	-0.11618	0	0		-0.08467	0.06375	pos
11	LEU	5	HIS	0	0.06901	-1.55951	0.05785	-0.9789	0.1661	0.55983	0.86038	0.44361	0	0		-0.09357	-0.48623	neg
12	LEU	5	VAL	0	0.00037	-0.2512	0.07167	-0.33375	-0.05122	-0.07882	-0.14561	-0.02276	0	0		0.09464	-0.01646	neg
13	TRP	6	CYS	0	-0.24419	0	-0.521	-0.58979	-1.12732	-0.66335	-0.45596	0	0	0		0	-0.26395	pos
14	TRP	6	GLY	0	-0.18178	0	-0.63535	-0.90704	-1.28979	-0.33159	-0.17572	0	0	0		0	-0.62764	pos
15	TRP	6	LEU	0	-0.03694	0	-0.00334	0.26617	0.26431	-0.04368	0.14435	0	0	0		0	0.08937	pos
16	GLN	7	HIS	0	0	0.22456	0.14707	-0.05542	0.16744	0.24723	-0.08248	-0.0548	0.17104	0.14183		0	0	pos
17	GLN	7	LEU	0	0	1.13621	0.28754	0.24948	0.54479	1.00782	-0.41464	0.37055	1.21177	0.94688		0	0	neg
18	GLN	7	PRO	0	0	0.20172	-0.12112	0.03098	-0.03136	0.00232	0.20147	0.33796	0.19486	0.06676		0	0	neg
19	ARG	8	ASN	0	0	0	0	-0.38913	0.18631	-0.63722	-2.26973	-0.61127	-0.75384	0		0	0	neg
20	ARG	8	ASP	0	0	0	0	-0.94424	-0.29427	-1.15565	-4.07861	-0.73567	-1.05439	0		0	0	neg
21	ARG	8	GLN	0	0	0	0	0.02021	0.48854	0.52975	-0.80067	0.15343	-0.06552	0		0	0	int
22	ARG	8	GLU	0	0	0	0	-0.95011	-0.35115	-0.5433	-3.12437	-0.62964	-0.65032	0		0	0	neg
23	ARG	8	GLY	0	0	0	0	-0.42784	6.00E-05	-1.3967	-3.00439	-0.60337	-0.61053	0		0	0	neg
24	ARG	8	HIS	0	0	0	0	0.18617	0.41218	-0.14344	-0.53493	0.01364	-0.13521	0		0	0	neg
25	ARG	8	LEU	0	0	0	0	0.69068	0.95149	-0.60797	0.0926	0.18717	0.90623	0		0	0	int
26	ARG	8	LYS	0	0	0	0	-0.61972	-0.26158	-0.45997	-1.35066	-0.56148	-0.48045	0		0	0	neg
27	ARG	8	TYR	0	0	0	0	0.46293	0.69359	-0.68478	-0.51269	0.08071	0.13992	0		0	0	neg
28	PRO	9	ARG	0	0	-0.53754	-0.11854	0.08246	0	0.06947	0.34747	0.05305	-0.37048	-0.40188		0	0	neg
29	PRO	9	HIS	0	0	-0.03502	0.01097	0.29562	0	0.07942	0.04235	0.37048	-0.05895	-0.01009		0	0	pos
30	PRO	9	SER	0	0	-0.04716	-0.02244	-0.10916	0	-0.12694	-0.17164	-0.2665	-0.17832	-0.00902		0	0	neg
31	PRO	9	THR	0	0	0.75899	0.25165	0.35114	0	0.19194	-0.08298	1.00916	0.38118	0.43864		0	0	int
32	LEU	10	ARG	0	0	-0.4688	0	0	0	0.6295	0.43639	-0.74725	-1.8945	-1.58816		0	0	int

In each of the 536 rows, the initial three components identify the mutant. This is followed by the 99-dimensional residual profile. The final component is the mutant activity class.

Supervised Classification

- Algorithms: Neural Network (NN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF)
- Implementations available with the Weka suite of machine learning tools: http://www.cs.waikato.ac.nz/ml/weka/
- Training set: Residual profile vectors for the mutants of a protein that have been studied experimentally, along with the activity class of each mutant (i.e., supervised)
- Each mutant (represented as a residual profile + activity class) is referred to as an ***instance***; each component of the residual profiles is referred to as an ***attribute***

Model Performance: HIV-1 Protease Mutants



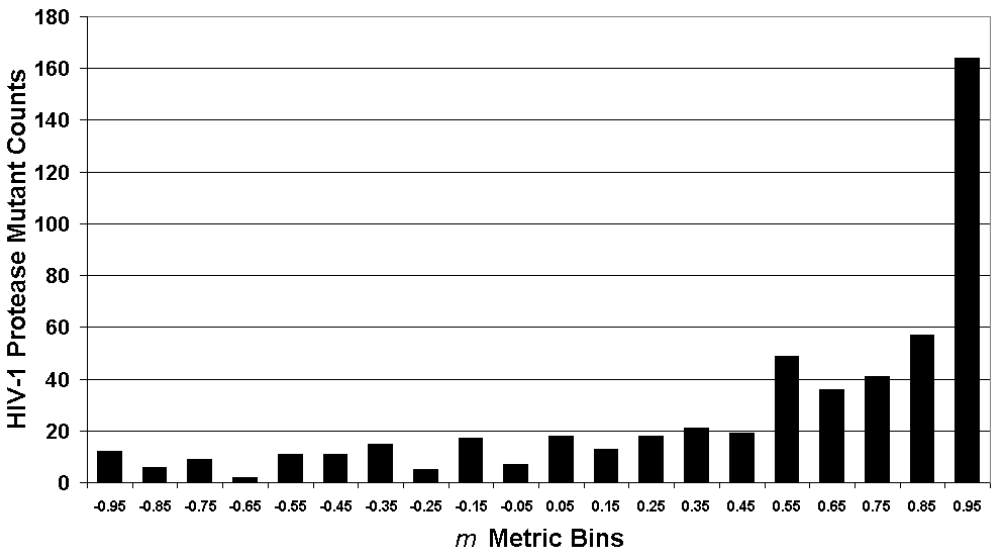
AUC Summary for HIV-1 Protease ROC Curves

	Pos (1-against-1)	Int (1-against-1)	Neg (1-against-1)	Others Combined (1-against-all)
Pos	---	**0.6522 (SVM)** **0.5869 (DT)** **0.6225 (NN)**	**0.8182 (SVM)** **0.8414 (DT)** **0.7877 (NN)**	**0.7389 (SVM)** **0.7732 (DT)** **0.7282 (NN)**
Int		---	**0.7558 (SVM)** **0.7726 (DT)** **0.7511 (NN)**	**0.6731 (SVM)** **0.6632 (DT)** **0.6814 (NN)**
Neg			---	**0.7810 (SVM)** **0.8324 (DT)** **0.7764 (NN)**

- Most disparate signals stem from residual profiles of the positive and negative mutants, followed closely by the intermediate and negative mutants
  - consistent with biological notion that fully active and inactive mutants display the greatest differences in structural and functional properties, while partially active and inactive mutants display significant, albeit less dramatic differences
- Residual profiles of mutants in the positive and intermediate classes display the least divergent signals
  - reflects the fact that both classes contain mutants that are more or less functionally active and display at most minimal structural changes from wt

Reliability/Reproducibility of Model Predictions

- 60/40 split test option => model learned with 60% of the mutants is used to predict the activity classes of the remaining 40%; 60 runs => expect approx. 24 predictions/mutant
- Apply two-class decision tree learning (default costs)
- For each mutant, n_c (n_i) = total # of correct (incorrect) predictions
- Mutant reliability metric: $m = (n_c - n_i) / (n_c + n_i)$
- $m = 0$ => equal # of correct and incorrect predictions; $m = 1$ => all predictions correct; $m = -1$ => all predictions incorrect



	536 HIV-1 Protease Mutants				40 Mutants Satisfying $m < -0.5$				
residue	**wt**	**sub.**	**sum**	**proportion**	**wt**	**sub.**	**sum**	**proportion**	**ratio**
TYR	2	15	17	0.0159	0	0	0	0	0.00
LYS	20	27	47	0.0438	0	1	1	0.0125	0.29
PHE	20	21	41	0.0382	1	0	1	0.0125	0.33
HIS	6	20	26	0.0243	1	0	1	0.0125	0.52
ASP	18	22	40	0.0373	0	2	2	0.025	0.67
ILE	92	28	120	0.1119	6	0	6	0.075	0.67
VAL	35	40	75	0.0700	0	4	4	0.05	0.71
ARG	27	45	72	0.0672	1	3	4	0.05	0.74
CYS	21	11	32	0.0299	1	1	2	0.025	0.84
MET	14	18	32	0.0299	2	0	2	0.025	0.84
ASN	24	22	46	0.0429	1	2	3	0.0375	0.87
LEU	52	37	89	0.0830	2	4	6	0.075	0.90
GLU	12	29	41	0.0382	2	1	3	0.0375	0.98
GLY	98	29	127	0.1185	7	4	11	0.1375	**1.16**
SER	2	40	42	0.0392	0	4	4	0.05	**1.28**
THR	32	35	67	0.0625	4	3	7	0.0875	**1.40**
GLN	17	25	42	0.0392	3	2	5	0.0625	**1.60**
PRO	28	23	51	0.0476	2	5	7	0.0875	**1.84**
ALA	10	33	43	0.0401	3	3	6	0.075	**1.87**
TRP	6	16	22	0.0205	4	1	5	0.0625	**3.05**
SUM	536	536	1072	1.00	40	40	80	1.00	

Assessment of the Statistical Significance for the Number of Correctly Classified Instances

- Random split: 436 HIV-1 protease mutants used as a training set for **decision tree** learning; remaining 100 mutants form a test set
- Training: 121 pos, 66 int, 249 neg; Testing: 19 pos, 18 int, 63 neg
- Result below based on two classes (similar method for 3 classes):

```
Correctly Classified Instances          74    74%
Incorrectly Classified Instances        26    26%
Total Number of Instances              100

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  Class
  0.73      0.254      0.628     0.73   active
  0.746     0.27       0.825     0.746  inactive

=== Confusion Matrix ===

  a  b   <-- classified as
 27 10 |  a = active
 16 47 |  b = inactive
```

Assessment of the Statistical Significance for the Number of Correctly Classified Instances

- Training: 187 active, 249 inactive; Testing: 37 active, 63 inactive
- Let $X = X_1+X_2+ \ldots+X_{100}$, where each X_i is a Bernoulli random variable representing the outcome of a test set instance prediction.
- $\mu = E(X) = 37\cdot(187/436) + 63\cdot(249/436) = 52$
- $\sigma^2 = Var(X) = 100\cdot(187/436)\cdot(249/436) = 24.5$
- So $\sigma = 4.95$, and *p*-value is

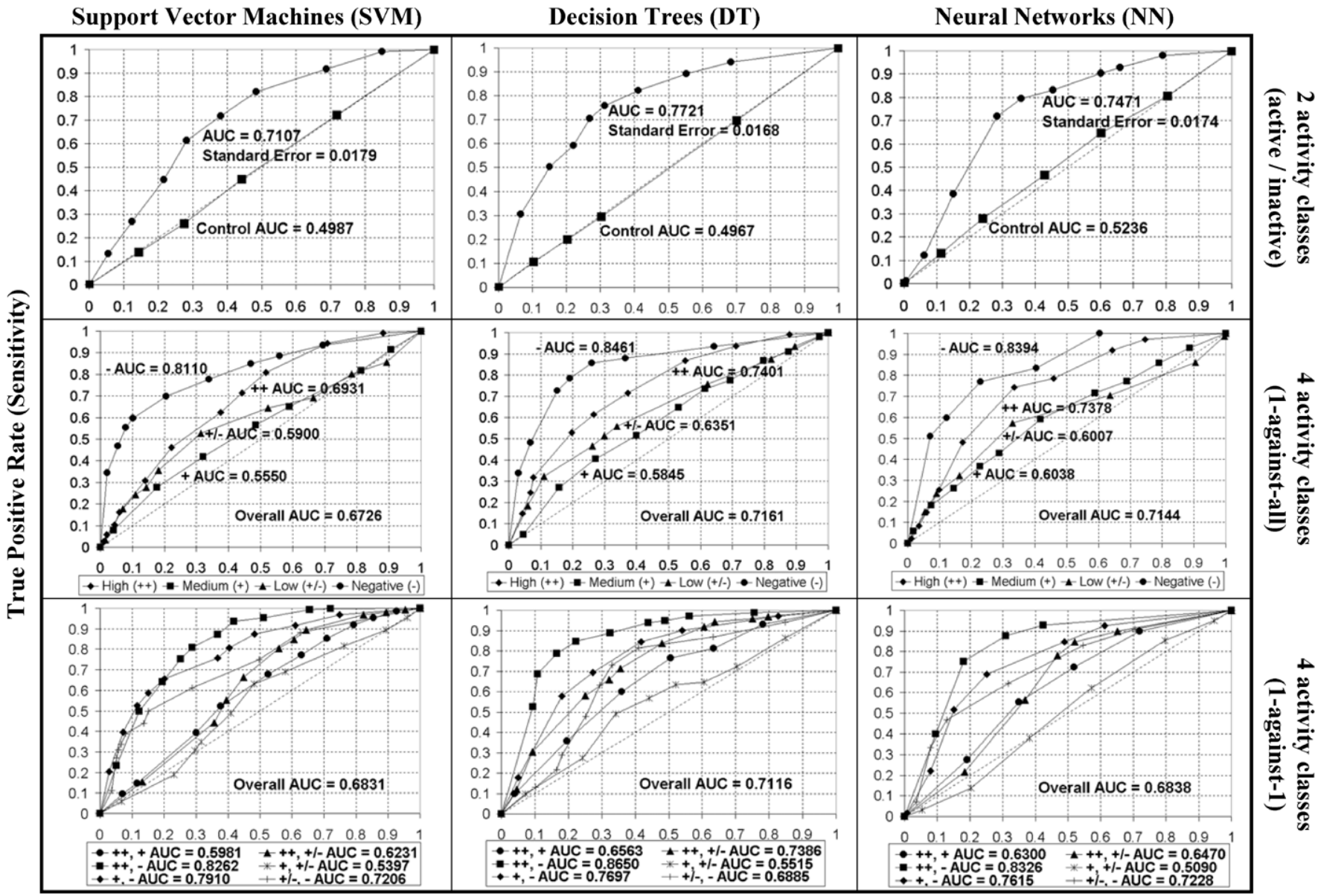
$$P(X > 74; \mu = 52) = P\left(\frac{X-\mu}{\sigma} > \frac{74-52}{4.95}\right) = P(z > 4.44) \approx 1-\Phi(4.44) = 4.42\times10^{-6}$$

where Φ is the cumulative dist. fn. for a standardized normal var.

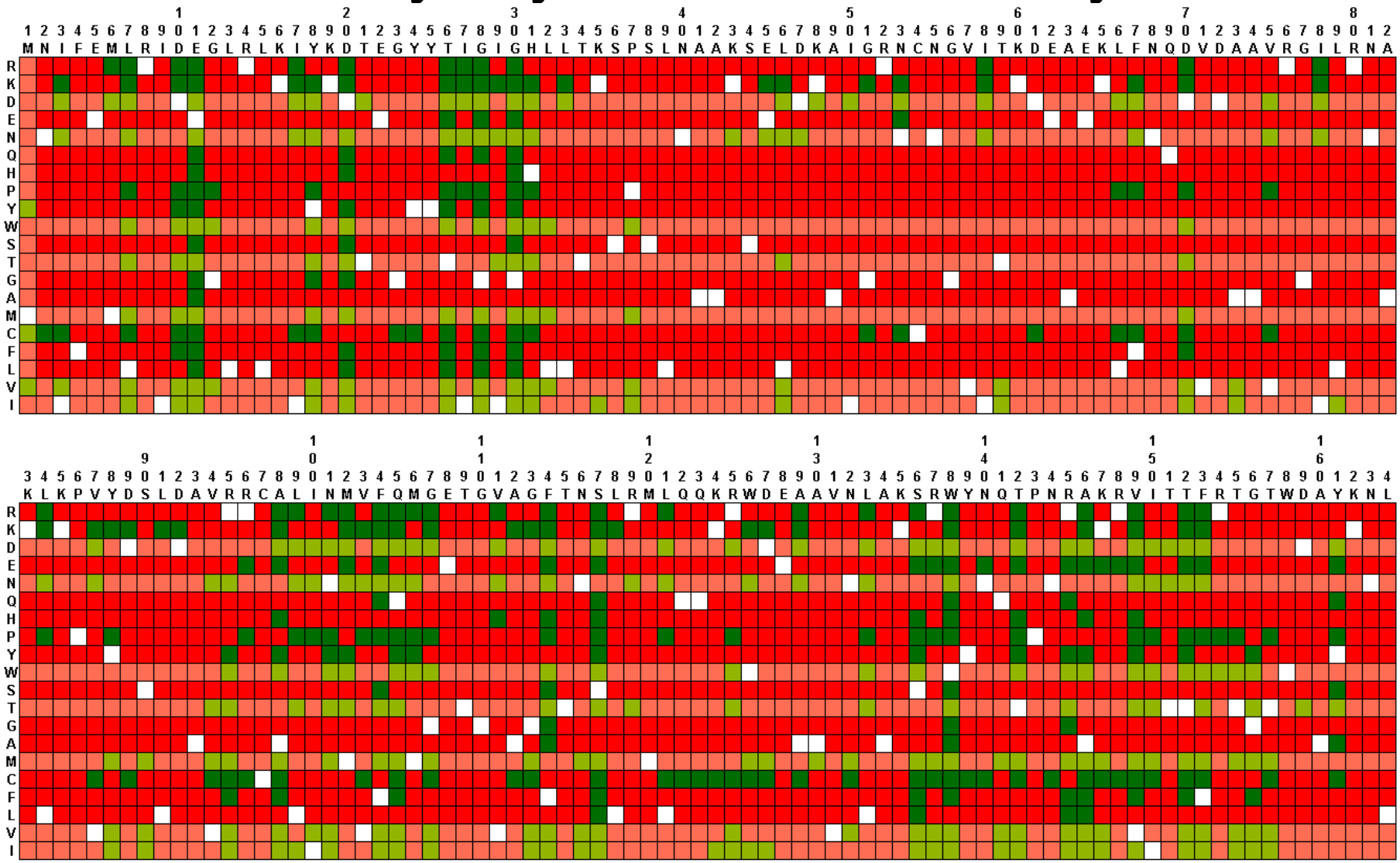
Summary of Results Based on Two and Three Classes

HIV-1 protease mutants (436 for training, 100 for testing)	2 classes	3 classes (1-against-all)	3 classes (1-against-1)
expected number of correctly classified test mutants (std. dev.)	52 (4.95)	44 (4.64)	44 (4.64)
actual number of correctly classified test mutants	74 $p = 4.42\times10^{-6}$	69 $p = 3.57\times10^{-8}$	66 $p = 1.06\times10^{-6}$

Model Performance: T4 Lysozyme Mutants



T4 Lysozyme Mutational Array



Training set mutants (n = 2015)

Active Inactive

Predicted test set mutants (n = 1101)

Active Inactive

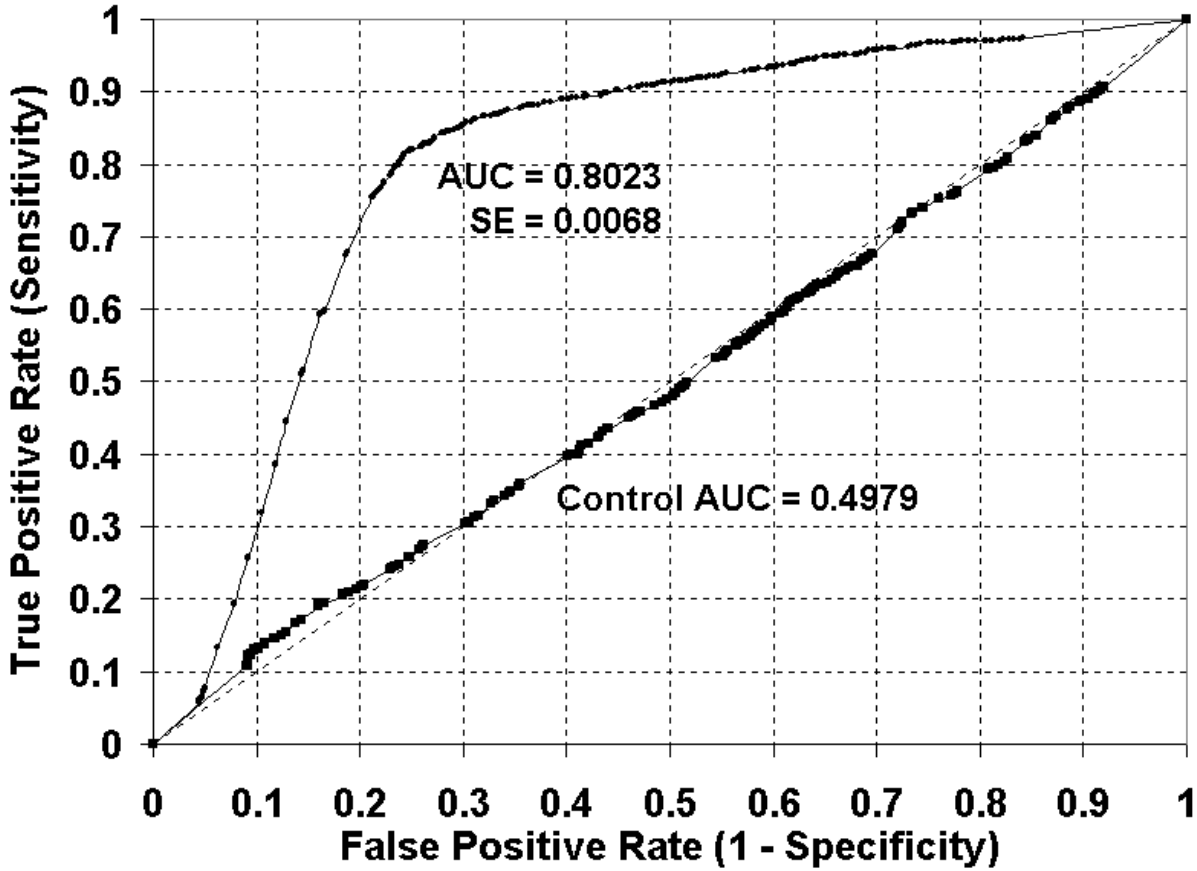
T4 Lysozyme Prediction Results

- Predicted activities compared with exp. activity from 8 labs
- Exp. data obtained from ProTherm database
- Exp. activity ≤ 5 inactive, and values > 5 active
- Result: 30/35 correct predictions, ~86%

#	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	

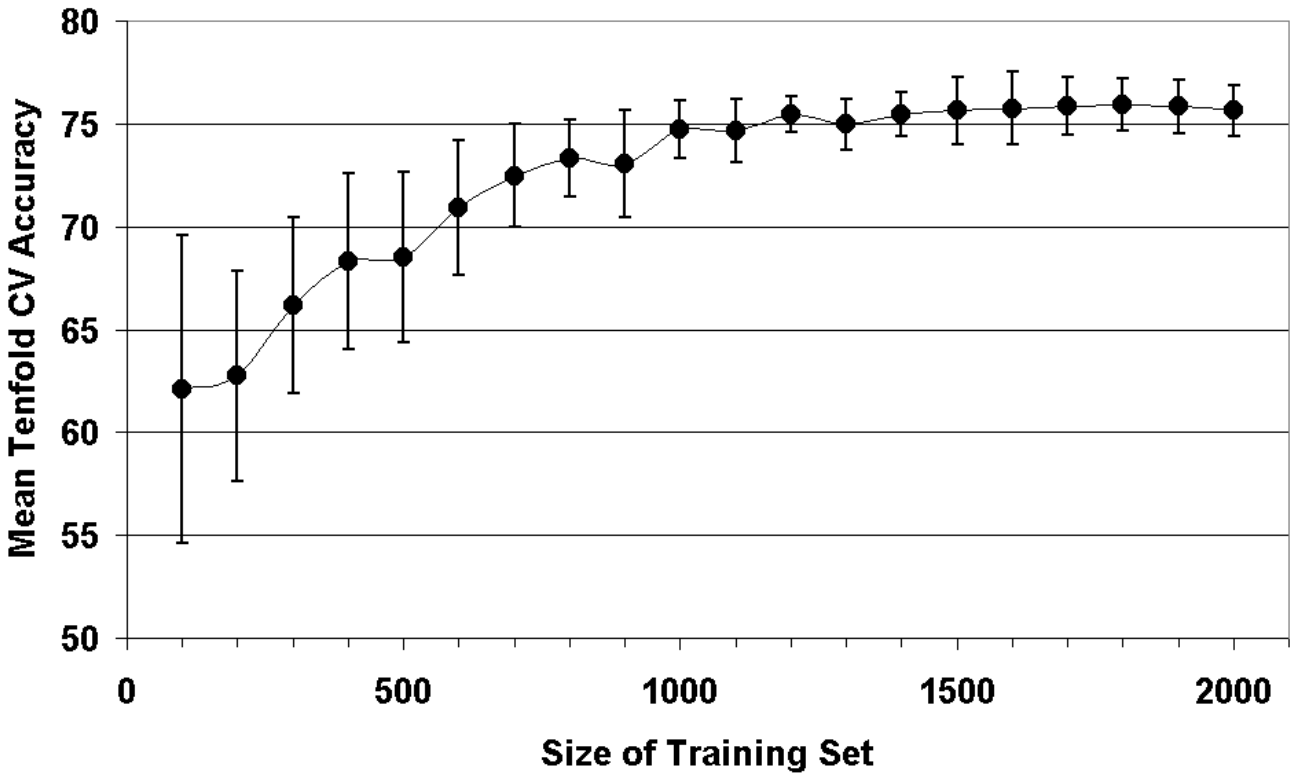
Lac Repressor Decision Tree Model Performance: Two Activity Classes (Unaffected/Affected)

- Accuracy: 78.67%
- AUC ± SE: 0.8023 ± 0.0068
- Control: activity labels randomly shuffled among the 4041 mutant residual profile vectors in the training set prior to applying decision tree learning

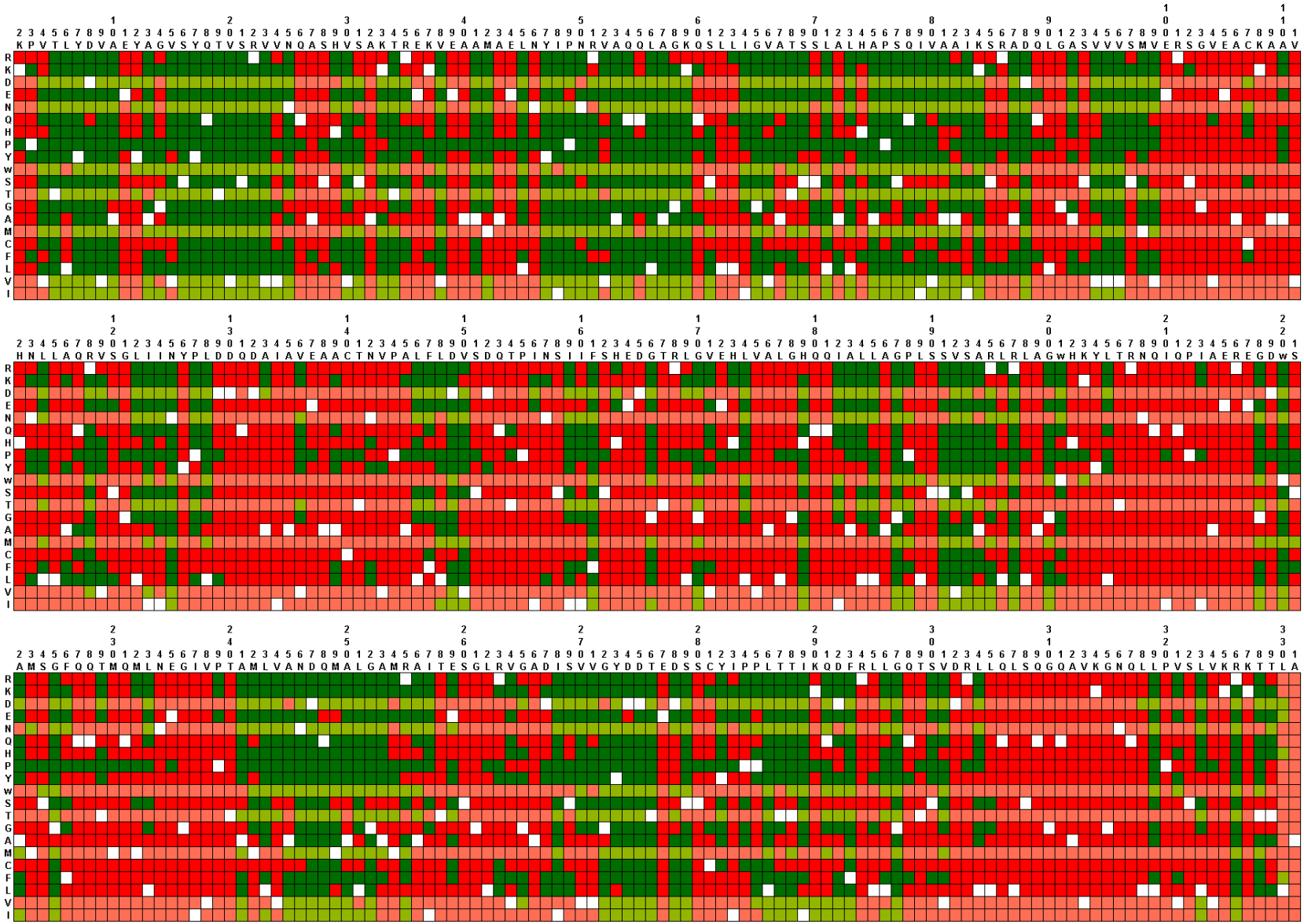


Learning Curve Example: *Lac* Repressor

- Stratified training sets randomly chosen with replacement in increments of 100 mutants
- At each training set size, mean 10 CV accuracy based on average of 10 runs using two-class decision tree supervised learning
- Error bars represent ±1 std. dev. from the mean

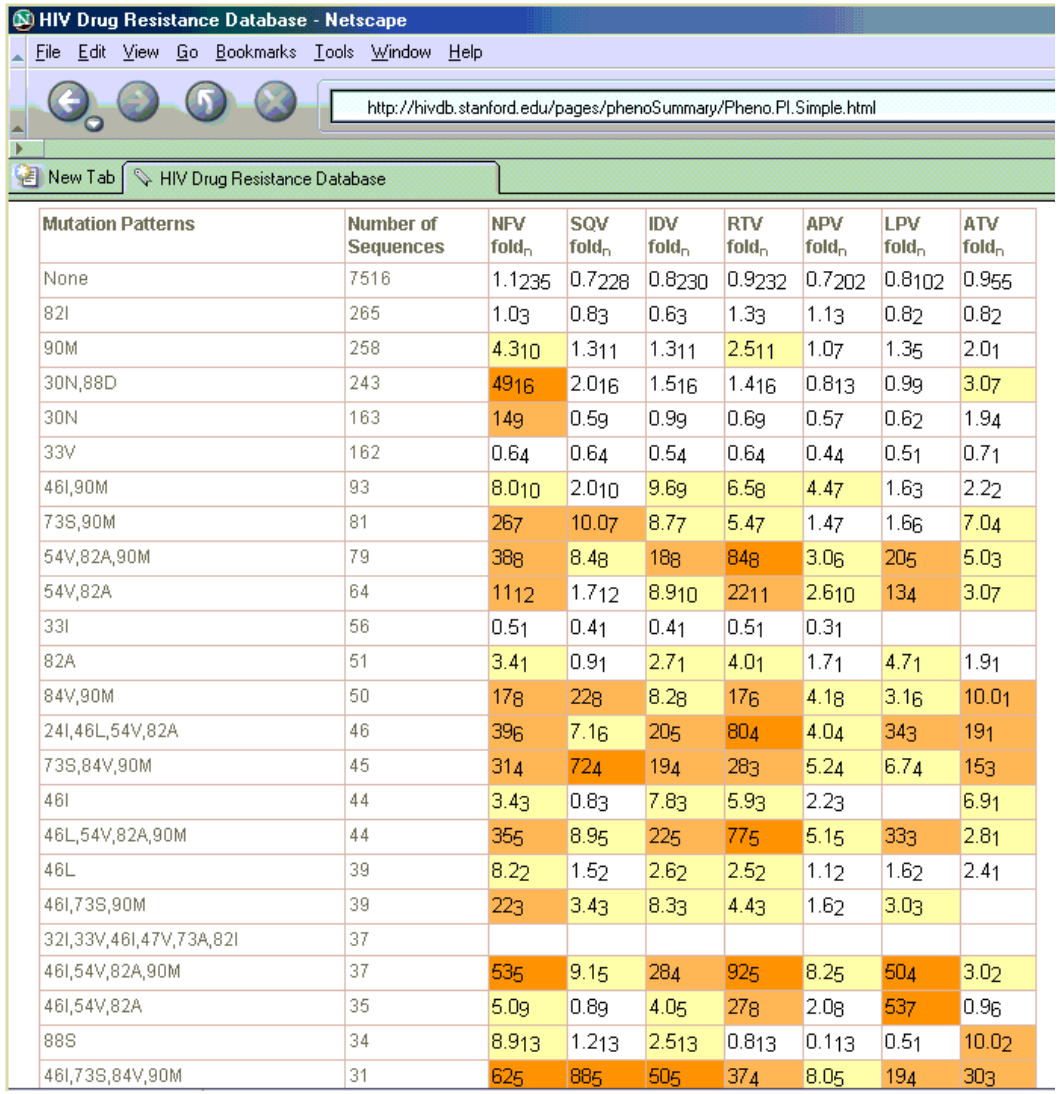


Lac Repressor Mutational Array



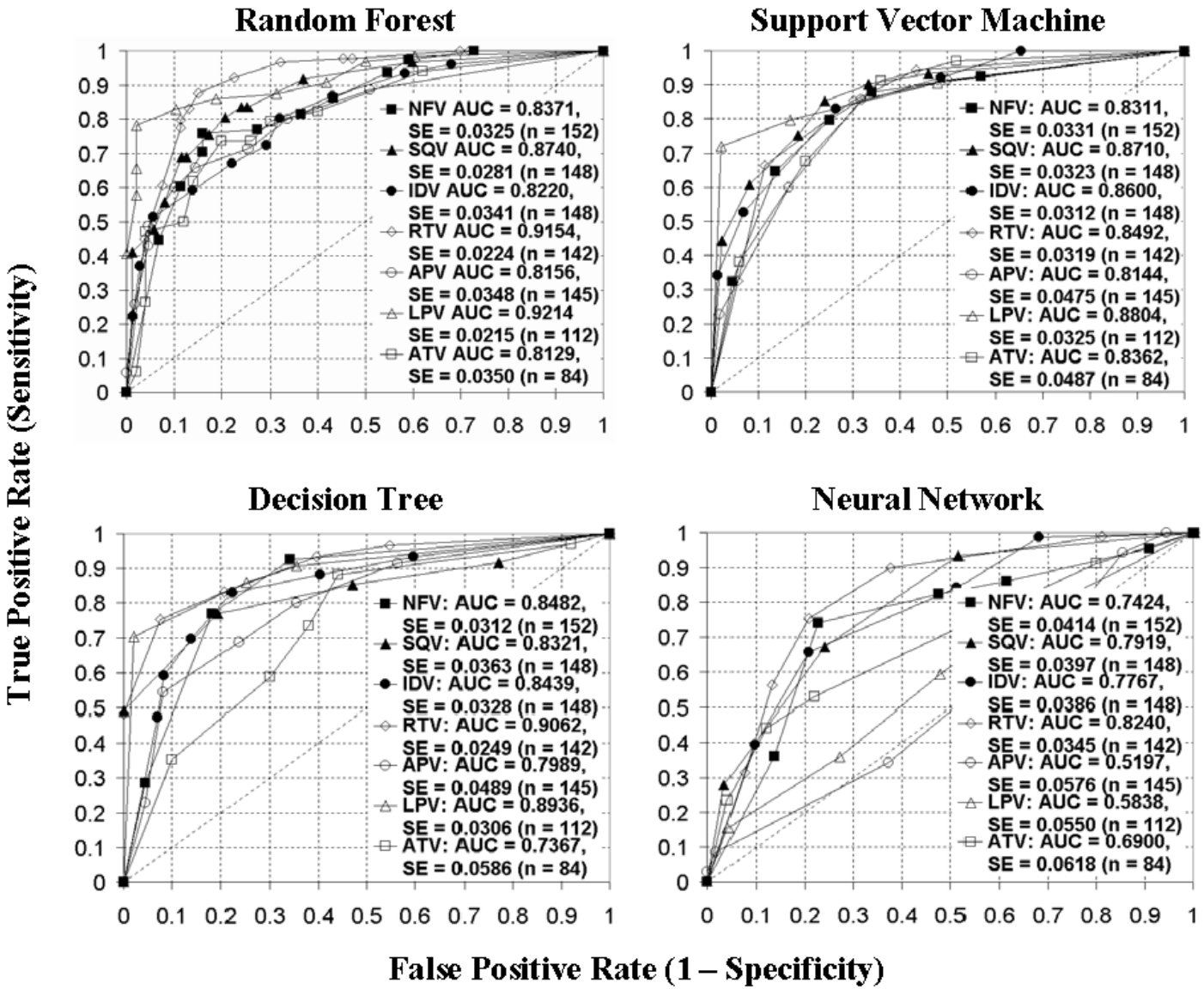
Clinical Application: Prediction of Drug Resistance Protein Mutational Patterns

- Nearly 400 (single and multiple) mutants of HIV-1 protease, isolated and sequenced from over 4000 patients
- Monogram Biosciences PhenoSense assay:
  - High: 152 distinct mutational patterns assayed for NFV
  - Low: 84 patterns assayed for ATV
- Mutant fold change = IC_{50}(mutant) / IC_{50}(wt)
- Subscripts in table = no. of assayed mutants; fold change value in table = median value
- Individual fold changes all show small abs. dev. from median, reflecting assay consistency
- Clinical cutoffs (based on latest data, studies still underway):
  - 2 classes: Sensitive ≤ 10, Resistant > 10
  - 3 classes: S ≤ 2.5, $2.5 <$ I ≤ 10, R > 10
- Each of the 7 inhibitors uses a distinct training set; separate models are trained and their performance is evaluated for each drug
- For each inhibitor, the learned models are used to predict the susceptibility of the unassayed mutational patterns for the given drug



Mutation Patterns	Number of Sequences	NFV $fold_n$	SQV $fold_n$	IDV $fold_n$	RTV $fold_n$	APV $fold_n$	LPV $fold_n$	ATV $fold_n$
None	7516	1.1_{235}	0.7_{228}	0.8_{230}	0.9_{232}	0.7_{202}	0.8_{102}	0.9_{55}
82I	265	1.0_{3}	0.8_{3}	0.6_{3}	1.3_{3}	1.1_{3}	0.8_{2}	0.8_{2}
90M	258	4.3_{10}	1.3_{11}	1.3_{11}	2.5_{11}	1.0_{7}	1.3_{5}	2.0_{1}
30N,88D	243	49_{16}	2.0_{16}	1.5_{16}	1.4_{16}	0.8_{13}	0.9_{9}	3.0_{7}
30N	163	14_{9}	0.5_{9}	0.9_{9}	0.6_{9}	0.5_{7}	0.6_{2}	1.9_{4}
33V	162	0.6_{4}	0.6_{4}	0.5_{4}	0.6_{4}	0.4_{4}	0.5_{1}	0.7_{1}
46I,90M	93	8.0_{10}	2.0_{10}	9.6_{9}	6.5_{8}	4.4_{7}	1.6_{3}	2.2_{2}
73S,90M	81	26_{7}	10.0_{7}	8.7_{7}	5.4_{7}	1.4_{7}	1.6_{6}	7.0_{4}
54V,82A,90M	79	38_{8}	8.4_{8}	18_{8}	84_{8}	3.0_{6}	20_{5}	5.0_{3}
54V,82A	64	111_{2}	1.7_{12}	8.9_{10}	221_{1}	2.6_{10}	13_{4}	3.0_{7}
33I	56	0.5_{1}	0.4_{1}	0.4_{1}	0.5_{1}	0.3_{1}		
82A	51	3.4_{1}	0.9_{1}	2.7_{1}	4.0_{1}	1.7_{1}	4.7_{1}	1.9_{1}
84V,90M	50	17_{8}	22_{8}	8.2_{8}	17_{6}	4.1_{8}	3.1_{6}	10.0_{1}
24I,46L,54V,82A	46	39_{6}	7.1_{6}	20_{5}	80_{4}	4.0_{4}	34_{3}	19_{1}
73S,84V,90M	45	31_{4}	72_{4}	19_{4}	28_{3}	5.2_{4}	6.7_{4}	15_{3}
46I	44	3.4_{3}	0.8_{3}	7.8_{3}	5.9_{3}	2.2_{3}		6.9_{1}
46L,54V,82A,90M	44	35_{5}	8.9_{5}	22_{5}	77_{5}	5.1_{5}	33_{3}	2.8_{1}
46L	39	8.2_{2}	1.5_{2}	2.6_{2}	2.5_{2}	1.1_{2}	1.6_{2}	2.4_{1}
46I,73S,90M	39	22_{3}	3.4_{3}	8.3_{3}	4.4_{3}	1.6_{2}	3.0_{3}	
32I,33V,46I,47V,73A,82I	37							
46I,54V,82A,90M	37	53_{5}	9.1_{5}	28_{4}	92_{5}	8.2_{5}	50_{4}	3.0_{2}
46I,54V,82A	35	5.0_{9}	0.8_{9}	4.0_{5}	27_{8}	2.0_{8}	53_{7}	0.9_{6}
88S	34	8.9_{13}	1.2_{13}	2.5_{13}	0.8_{13}	0.1_{13}	0.5_{1}	10.0_{2}
46I,73S,84V,90M	31	62_{5}	88_{5}	50_{5}	37_{4}	8.0_{5}	19_{4}	30_{3}

ROC Curves Based on Two-Class Training Sets



Factors Contributing to Classification Capability

Factors

F1: values (magnitude and sign) of the non-zero components in each vector

F2: location of the non-zero components in each vector

F3: number of non-zero components in each vector

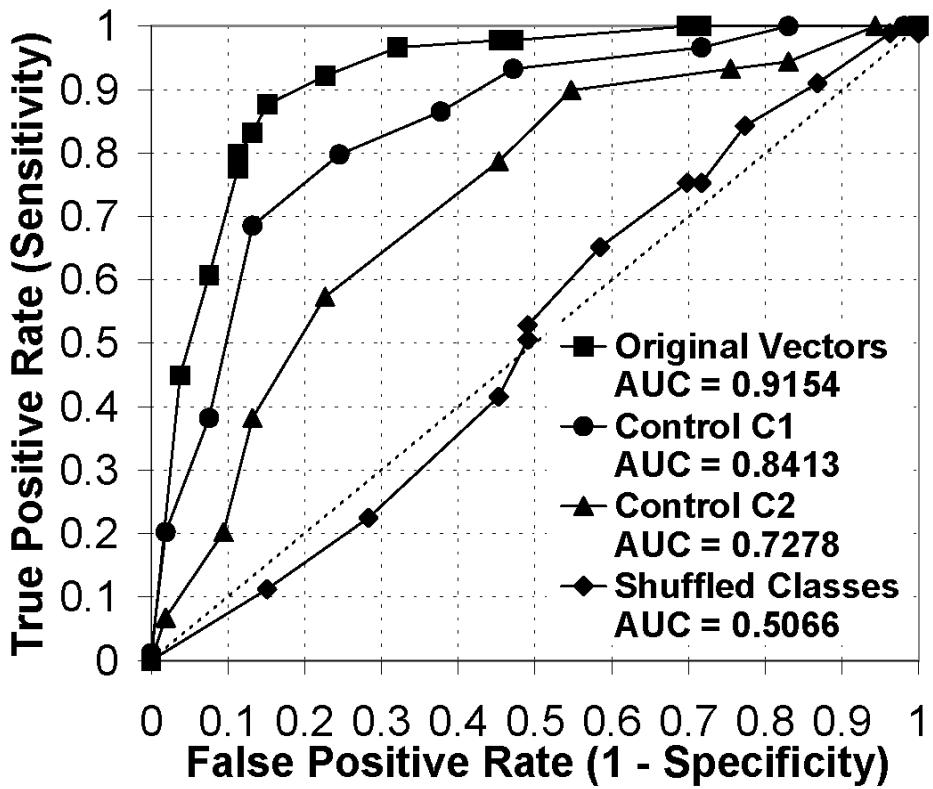
Controls

C1: multiply each non-zero vector component by a random number generated from the interval [-2, 2] (removes influence of F1, measures contributions of F2 and F3)

C2: randomly shuffle the components of each vector in C1 independently (removes influences of F1 and F2, measures contribution of F3)

Graphed ROC Example: RTV

- Apply Random Forest (RF) supervised classification
- Shuffled classes control: S, R class labels randomly shuffled among mutant vectors prior to RF learning



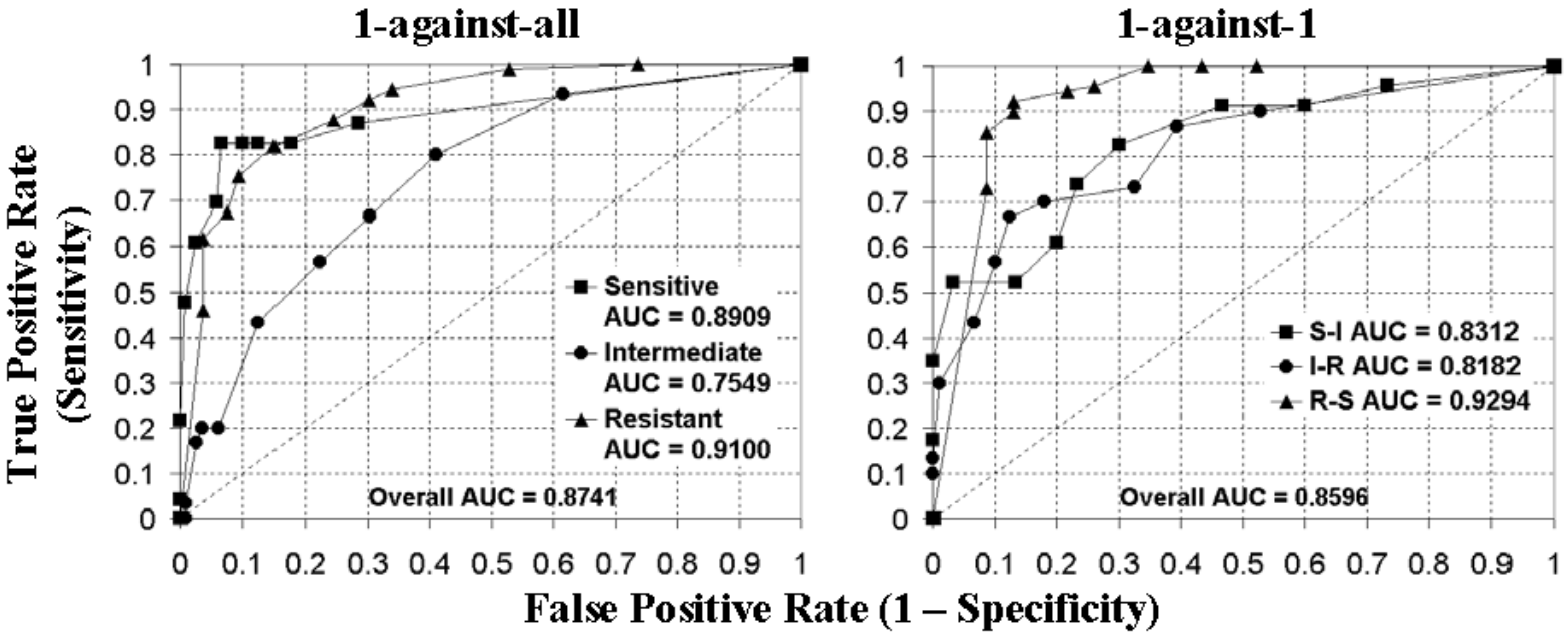
RF Results For All Inhibitors

	AUC Values			
Inhibitor	Original Vectors	Control C1	Control C2	Shuffled Classes
NFV	0.8371	0.7796	0.7404	0.4762
SQV	0.8740	0.7455	0.6629	0.4957
IDV	0.8220	0.7166	0.6552	0.5757
RTV	0.9154	0.8413	0.7278	0.5066
APV	0.8156	0.6382	0.5349	0.4764
LPV	0.9214	0.8818	0.7163	0.4277
ATV	0.8129	0.7671	0.6826	0.5185

RF AUCs Based on Three Susceptibility Classes

Inhibitor	Ref. Class	No. of Mutants	Ref. Class AUC	Overall AUC (1-against-all)	Class Pairs	Class Pair AUC	Overall AUC (1-against-1)
NFV	S	13	0.7939	**0.8197**	S-I	0.6923	**0.7816**
	I	31	0.7457		I-R	0.8202	
	R	108	0.8441		R-S	0.8323	
SQV	S	53	0.8892	**0.8064**	S-I	0.7142	**0.7918**
	I	34	0.5938		I-R	0.6941	
	R	61	0.8530		R-S	0.9671	
IDV	S	29	0.8864	**0.7845**	S-I	0.7791	**0.8299**
	I	43	0.6729		I-R	0.7463	
	R	76	0.8088		R-S	0.9644	
RTV	S	23	0.8909	**0.8741**	S-I	0.8312	**0.8596**
	I	30	0.7549		I-R	0.8182	
	R	89	0.9100		R-S	0.9294	
APV	S	54	0.7939	**0.7562**	S-I	0.7002	**0.7959**
	I	56	0.6640		I-R	0.7773	
	R	35	0.8455		R-S	0.9101	
LPV	S	22	0.8495	**0.8883**	S-I	0.7762	**0.8711**
	I	26	0.7744		I-R	0.8684	
	R	64	0.9479		R-S	0.9688	
ATV	S	21	0.7192	**0.7306**	S-I	0.5263	**0.6918**
	I	29	0.6207		I-R	0.7277	
	R	34	0.8315		R-S	0.8214	

Graphed RF ROC Example: RTV



Publications

- Masso, M. & Vaisman, I.I. (2003) Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach. *Biochem. Biophys. Res. Comm.* **305**, 322-326.
- Masso, M. (2003) DC-SIGN points the way to a novel mechanism for HIV-1 transmission. *Medscape General Medicine* **5** (2). Available at: http://www.medscape.com/viewarticle/455538.
- Masso, M. & Jagota, A. Computational Methods in Phylogenetic Analysis. Sunnyvale: Bioinformatics by the Bay Press, 2005. ISBN: 0970029764.
- Masso, M. & Vaisman, I.I. Computational mutagenesis studies of protein structure-function correlations. *Proteins* (accepted).
- Masso, M. & Vaisman, I.I. Accurate prediction of enzyme mutant activity based on a multibody statistical potential. *BMC Bioinform.* (to be submitted).
- Masso, M. & Vaisman, I.I. Functional inference of enzyme mutants using a four-body statistical potential. *J. Mol. Biol.* (to be submitted).
- Masso, M. & Vaisman, I.I. Computational mutagenesis of *lac* repressor: insights into structure-function correlations and accurate inferential models of mutant activity. *J. Proteome Res.* (to be submitted).
- Masso, M. & Vaisman, I.I. Inferential models of susceptibility to HIV-1 protease inhibitors: a combined sequence-structure approach to predicting resistance. *AIDS* (to be submitted).

Selected Conference Presentations

- Masso M. and Vaisman I.I. Functional Prediction of Protein Mutants Using a Four-Body Potential, Intelligent Systems for Molecular Biology (ISMB), Detroit, MI, June 25-29, 2005.
- Masso M. and Vaisman I. Automated Functional Inference of Enzyme Mutants Utilizing a Four-Body Statistical Potential, The Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB), Cambridge, MA, May 14-18, 2005.
- Masso M. and Vaisman I. Structure-Function Correlation in HIV-1 Protease Using a Four-Body Statistical Potential, International Conference on Structural Genomics (ICSG), Washington Hilton Hotel, Washington, DC, November 17-21, 2004.
- Functional Analysis of HIV-1 Protease Using a Four-Body Statistical Potential, Annual Meeting of the Society for Mathematical Biology (SMB), University of Michigan, Ann Arbor, MI, July 25-28, 2004.
- Masso M. and Vaisman I. Comprehensive Mutagenesis of HIV-1 Protease: A Statistical Geometry Approach, European Conference on Computational Biology (ECCB), Centre de Conférences de la Villette, Paris, France, September 27-30, 2003.
- A Statistical Geometry Approach to the Study of Protein Structure, Annual Summer Meeting of the Mathematical Association of America (MAA), University of Colorado, Boulder, CO, July 29-August 2, 2003.
- Masso M. and Vaisman I. Analyzing Protein Structure-Function Correlations Using Statistical Geometry, Intelligent Systems for Molecular Biology (ISMB), Brisbane Convention & Exhibition Centre, Brisbane, Australia, June 29-July 3, 2003.

Acknowledgements

Committee

Dr. Vaisman – Ph.D. Director

Dr. Grefenstette

Dr. Jamison

Dr. Royt

Structural Bioinformatics Group

Vadim Ravich

Ewy Mathe

Todd Taylor

Andrew Carr

Tariq Alsheddi

Greg Reck

Summer '05 Interns

Dr. Saleet Jafri – Organizer

Kahkeshan Hijazi, Nida Parvez

Support Staff

Glenda Wilson, Chris Ryan,

Susan Beale

Software

Qhull – tessellation (Barber)

Glisten – tessellation visualization (Carr)

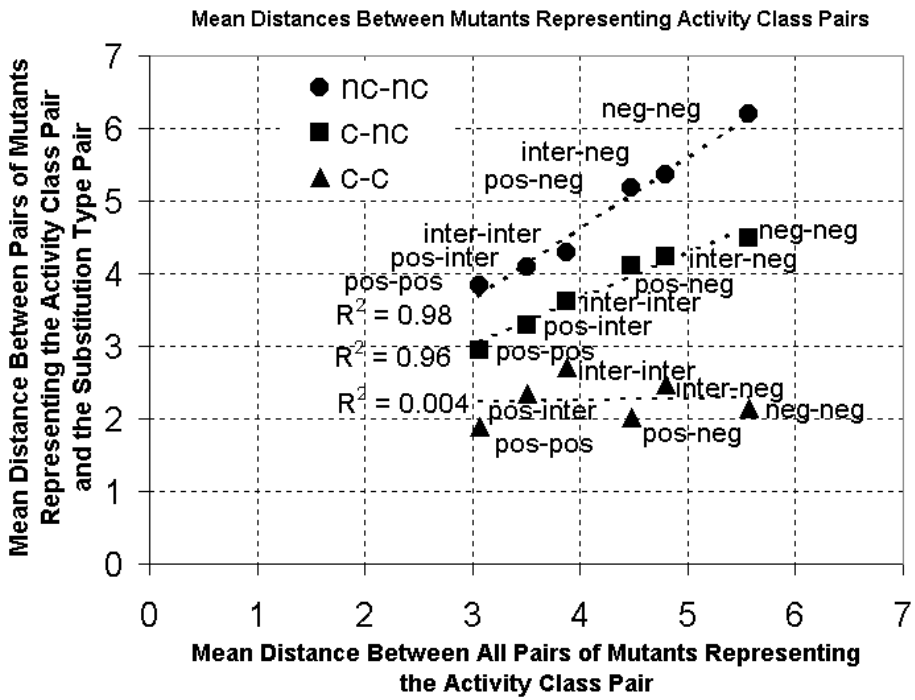
Chimera – ribbon diagrams (Ferrin)

Base Java programs – to generate raw data based on tessellation (Lu)

Weka – machine learning (Witten, Frank)

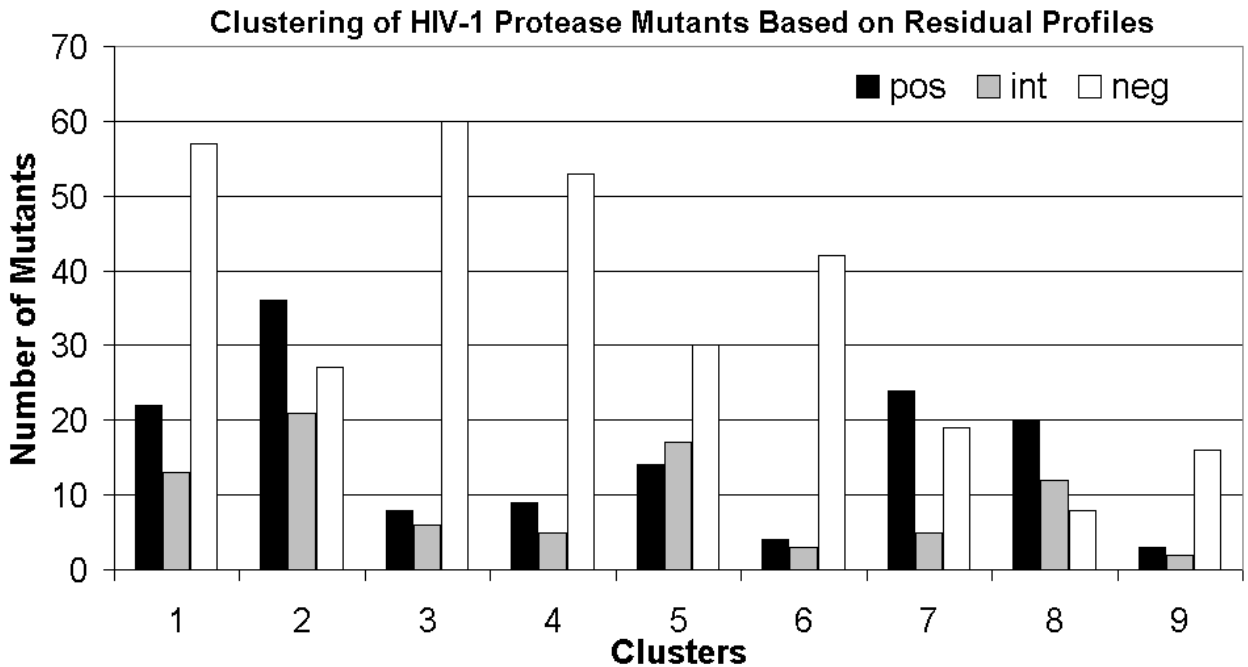
Mutant Activity Class Distribution in $\mathbb{R}^{99}$

- Given d(A,B) = mean Euclidean distance between all possible pairs of mutants (one from class A, and the other from class B),
  d(pos,pos) < d(pos,inter) < d(inter,inter) < d(pos,neg) < d(inter,neg) < d(neg,neg)
- order agrees with biological notions on impact of mutations
- mutant pairs for which at least one of the mutants represents a NC substitution drive the order of the mean distances



Clustering Example: HIV-1 Protease

- Ron Shamir's Expander software: http://www.cs.tau.ac.il/~rshamir/expander/expander.html
- Similar to k-means, but no a priori value of k needed; algorithm derives optimal number of clusters
- Leaves open the question of how well the residual profiles can be used to classify mutants with differing levels of activity



Test Options

- Use (partitioned) training set only—for assessing performance
  - Tenfold cross-validation (10 CV): Stratified partitioning of the instances into ten equally sized subsets
    1. One subset is held out, while the other nine subsets (90% of the original instances) are combined to form a modified training set
    2. The supervised classification algorithm is used to learn a model with the modified training set; the learned model is used to predict the activity classes of the instances in the hold-out subset (the test set)
    3. The process is repeated ten times, whereby each subset serves once as a hold-out for prediction; hence, a single activity prediction is made for each instance
  - Leave-one-out (or N CV, where N = size of full training set): Each subset consists of one instance; no stratification by definition; deterministic
  - % split: Stratified partitioning of the instances into two (not necessarily equal) subsets; larger subset serves as a training set for model building, and smaller subset is a test set
- Use the full training set (for model building) and an independent test set (for example, to predict the activity classes of mutants that have not been studied experimentally, if performance as described above is acceptable)

Evaluation of Model Performance (Two Classes: P, N)

- Confusion matrix: tabulated number of test predictions (shown)
- Sensitivity = TP / (TP + FN), Specificity = TN / (TN + FP), and 1-Specificity = FP / (FP + TN)
- Sensitivity = True Positive Rate (TPR)
  1-Specificity = False Positive Rate (FPR)

		Predicted as Pos	Predicted as Neg
Actual class	Pos	TP	FN
	Neg	FP	TN

- Accuracy = (TP + TN) / (TP + FP + TN + FN); simple measure, but highly sensitive to class skew in test sets
- Default costs assigned prior to model building are 0 (TP, TN) and 1 (FP, FN); ↑ FP cost only → ↓ no. of FP's → ↑ specificity; ↑ FN cost only → ↓ no. of FN's → ↑ sensitivity
- ROC (Receiver-Operating Characteristic) Curve: Plot of TPR vs. FPR in unit square using 10 CV for a range of FN/FP cost ratios
- Area under ROC curve (AUC): performance measure that is insensitive to unequal class distributions in test sets
  - Perfect classifier: Piecewise linear ROC joining (0,0) to (0,1) and (0,1) to (1,1); AUC = 1.0
  - Random guessing model: Diagonal line ROC joining (0,0) to (1,1); AUC = 0.5

Application to Multiple (n > 2) Classes

- One-against-all approach (use all training set instances)
  1. Choose one class as a reference (class 1); combine all other classes together by re-labeling as non-reference (class 2)
  2. Apply ROC analysis to this two-class system
  3. Repeat n times so that each class serves as a reference once
  4. Overall AUC for the multi-class system is a weighted average of the two-class AUCs (each two-class AUC weight is the proportion of mutants belonging to the respective reference class in the training set); this method is sensitive to class skew in theory, but performs well in practice
- One-against-one approach (truncate the original training set)
  1. Choose one pair of classes; form a truncated training set consisting of only instances that belong to either of these two classes
  2. Apply ROC analysis to this two-class system
  3. Repeat n(n-1)/2 times, so that every pair of classes is considered
  4. Overall AUC for the multi-class system is a simple average of the two-class AUCs; this method remains insensitive to class skew in test sets

Factors Contributing to Classification Capability

Factors

F1: no. of non-zero components in each vector

F2: value (magnitude and sign) of the non-zero components in each vector

F3: location of the non-zero components in each vector

F4: no. of non-zero columns in each group of vectors (submatrix of the training set) representing all mutants generated by amino acid substitutions at the same position

F5: location of the non-zero columns in each group

Controls

C1: multiply each non-zero vector component by a different random no. generated from the interval [-2, 2] (removes influence of F2, measures contributions of F1 and F3)

C2: randomly shuffle the components of each vector independently (removes influence of F3, measures contributions of F1 and F2)

C3: composite of C1 followed by C2 (removes influences of F2 and F3, measures contribution of F1)

C4: randomly shuffle the columns within each group independently (removes influence of F5, measures contributions of F2 and F4)

C5: composite of C1 followed by C4 (removes influences of F2 and F5, measures contribution of F4)

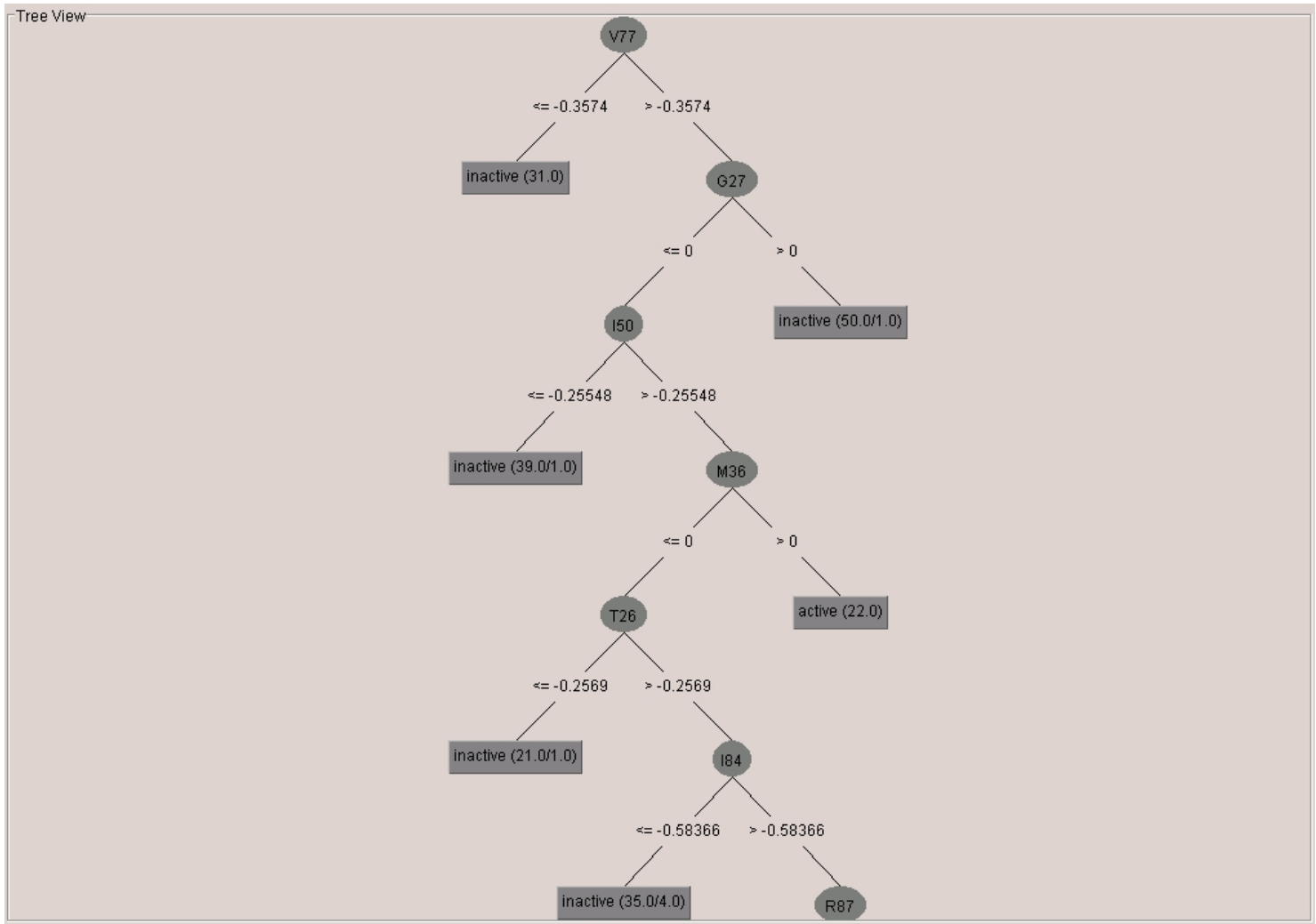
Ten independent versions of each control training set were prepared, and two-class decision tree learning (default costs) was applied

testing method	original vectors	C1	C2	C3	C4	C5
10 CV (10 runs)	76.8	72.6 (0.89)	54.6 (1.36)	53.5 (1.14)	73.0 (1.09)	65.8 (1.69)
60/40 split (100 runs)	75.3	71.0 (0.67)	54.8 (0.55)	53.4 (0.86)	70.3 (0.78)	63.6 (1.35)
536 CV	75.4	73.8 (3.24)	53.9 (3.74)	51.8 (3.86)	72.4 (2.12)	67.0 (2.32)

mean accuracy as a percentage (std. dev.)

Decision Tree

- Default cost model learned from the training set of 536 experimental HIV-1 protease mutants (active/inactive)



Alternative Testing Approaches and Learning Curves

- Apply RF supervised learning to the 142 RTV mutants
- 100 stratified 66/34 random splits: accuracy = 83.2%, std. dev. = 4.7%
- 100 iterations of 10 CV: accuracy over 1000 folds = 84.3%, std. dev. = 9.5%
- Leave-one-out (142 CV): accuracy = 85.9%

- Learning Curves using the 142 RTV mutants and DT, SVM, and RF supervised learning
- Stratified training sets randomly chosen with replacement in increments of 20 mutants
- Mean 10 CV accuracy based on average of 10 runs
- Error bars = ±1 std. dev.