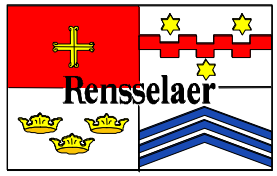
Prediction of protein affinity in HIC systems using state-of-the-art structure-property modeling techniques

Steven M. Cramer
Isermann Department of Chemical and Biological Engineering
Rensselaer Polytechnic Institute, Troy, NY

Presented at
Follow-on Biologics Workshop: Scientific Issues in Assessing the Similarity of Follow-on Protein Products

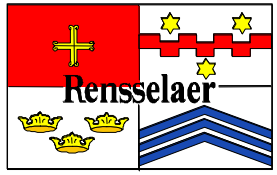
A New York Academy of Sciences Meeting

December 12 – 14, 2005
Brooklyn, New York

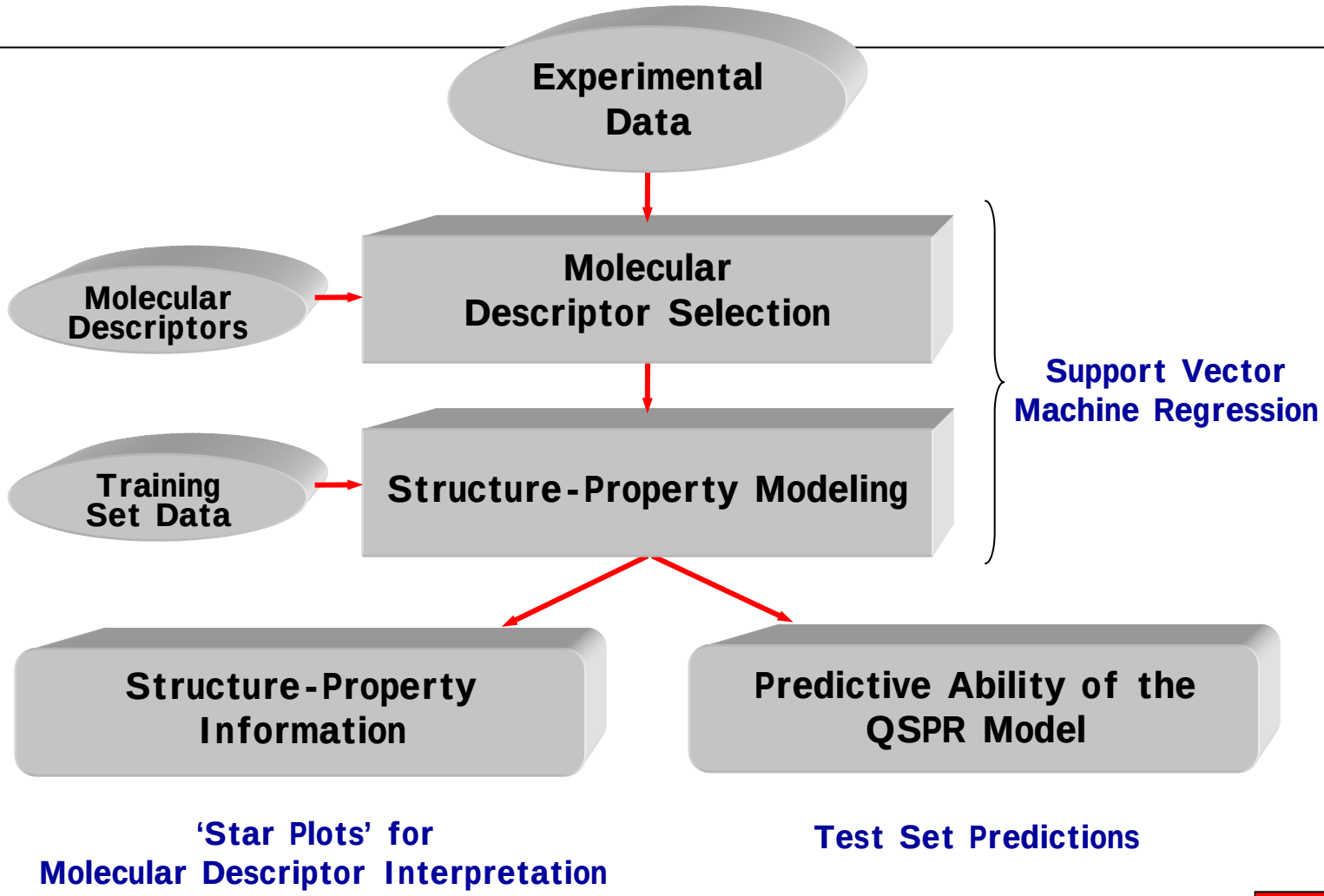


Motivation

- The *a priori* prediction of protein affinity and preparative chromatographic behavior has been a longstanding major goal in the bioseparations field. This work focuses on the development of novel Quantitative Structure-Property Relationship (QSPR) models for protein affinity in HIC systmes.
- In addition to providing *a priori* predictions, this work attempts to provide fundamental insights into the underlying mechanisms of chromatographic selectivity as well as a technique for predicting column chromatographic behavior directly from protein crystal structure data.
- Finally, this work attempts to establish a framework for evaluating the similarities of proteins.

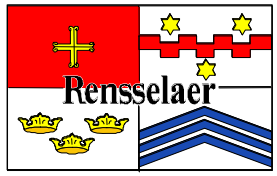


QSPR Modeling Flowchart

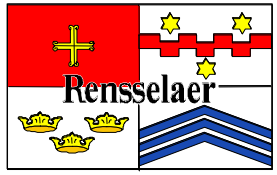


Description of the QSPR Modeling Approach

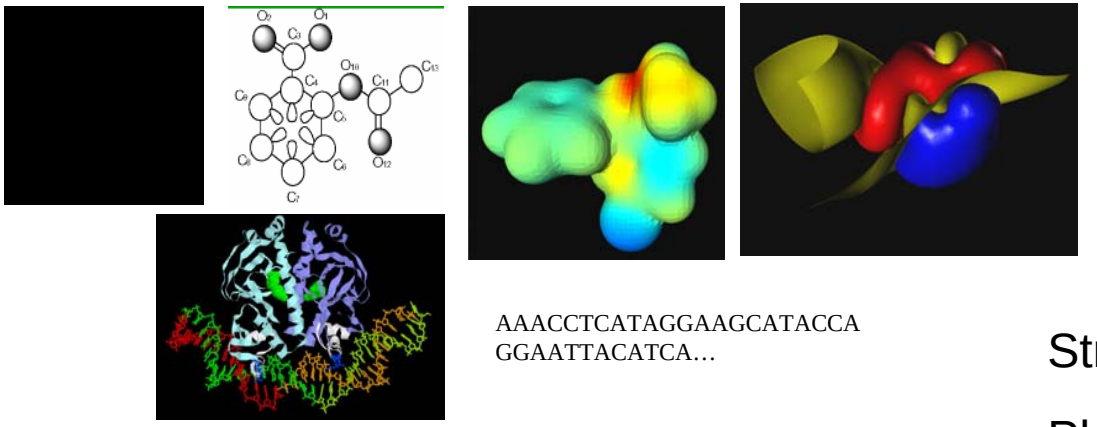
- **Obtain experimental data that will be used as the dependent variable (e.g., retention data, isotherm parameters, etc.).**
- **Calculate a relatively large number of molecular property descriptors for each protein used in the experimental data set.**
- **Carry out feature selection to determine the molecular descriptors most highly correlated with the experimental response.**
- **Construct a QSPR model from the experimental data and selected molecular descriptors for a training set of molecules.**
- **Test the predictive ability of this model using a test set of molecules that have not been used in the generation of the model.**
- **Examine a graphical depiction (star plot) showing the relative importance of the selected descriptors to gain insights into the underlying mechanisms.**



Molecular Descriptors



Encoding Structure : Descriptors



AAACCTCATAGGAAGCATACCA
GGAATTACATCA…

Structural Descriptors

Physiochemical Descriptors

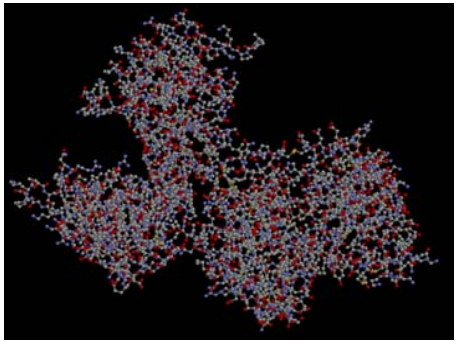
Topological Descriptors

Geometrical Descriptors

Molecular Structures → Descriptors ↔ Model ↔ Activity

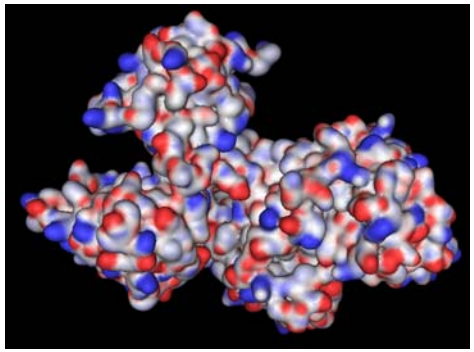
MOE Descriptors

- Classical physicochemical properties:
  - logP, molecular refractivity
- Pharmacophore features:
  - the number of H-bond donor/acceptor atom
  - polar or hydrophobic surface area
- Property-mapped subdivided surface area:



3D protein crystal geometry

map partial charge on molecular surface

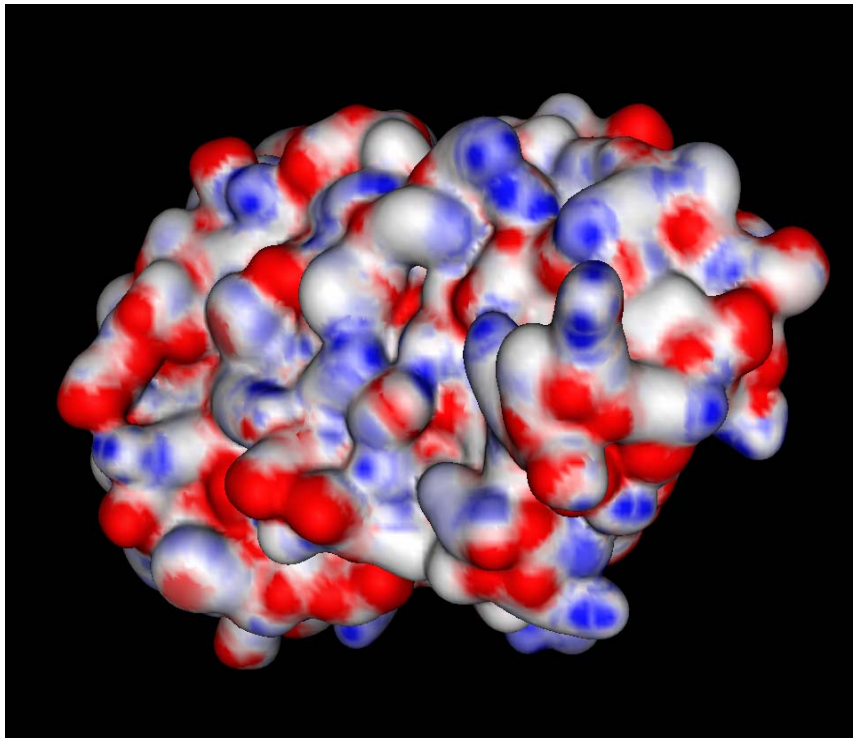


blue: positive; red: negative

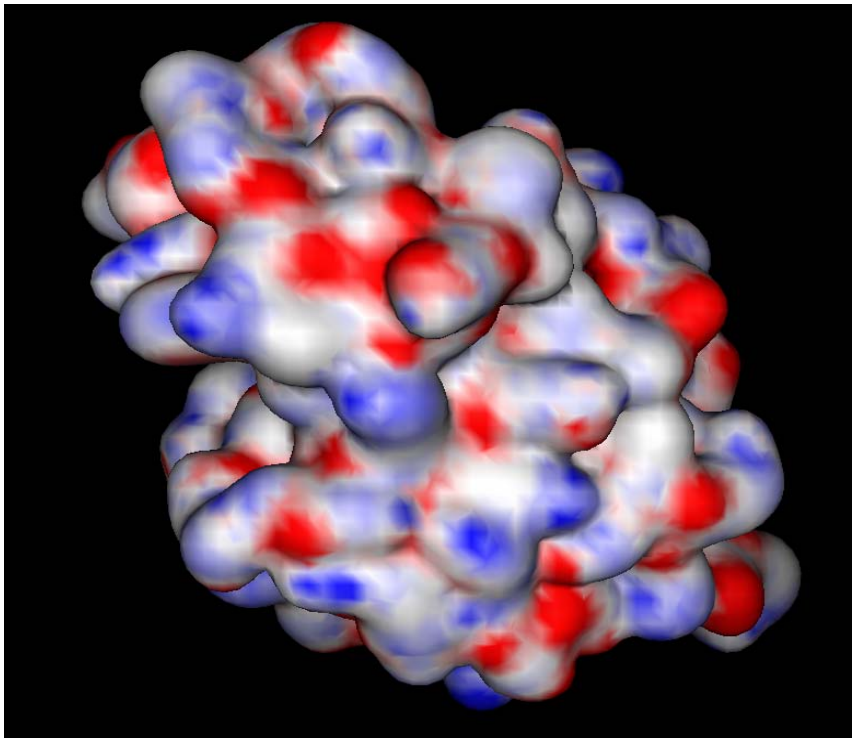
www.chemcomp.com

Rensselaer

Protein Surfaces (EP)



HSA



lysozyme

TAE/RECON Descriptors

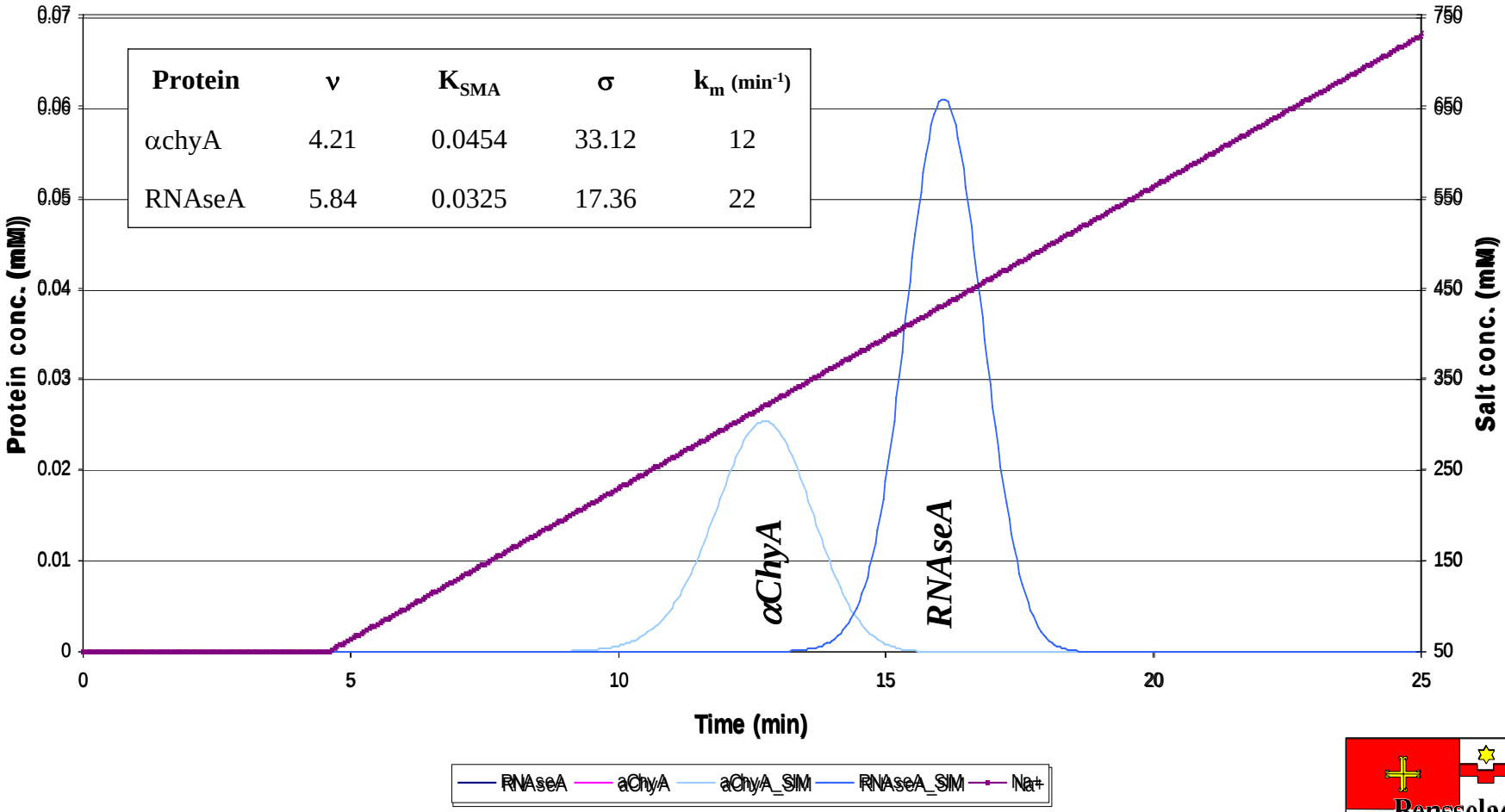
EP	Electrostatic Potential	$EP(r) = \sum_\alpha Z_\alpha / \lvert r - R_\alpha \rvert - \int \rho(r') d(r') / \lvert r - r' \rvert$
Del(Rho)•N	Electron Density Gradient normal to electron density iso-surface	
G	Electronic Kinetic Energy	$G = -(\eta / 4m) \int \{\nabla\psi^* \cdot \nabla\psi\} d\tau$
K	Electronic Kinetic Energy	$K = -(\eta / 4m) \int \{\psi^* \nabla^2 \psi + \psi \nabla^2 \psi^*\} d\tau$
Del(K)•N	Gradient of K Electronic Kinetic Energy normal to surface	
Del(G)•N	Gradient of G Electronic Kinetic Energy normal to surface	
Fuk	Fukui F^+ function scalar value	$F^+(r) = \rho_{HOMO}(r)$
Lapl	Laplacian of the electron density	$\nabla^2 \rho(r) = G(r) - K(r)$
BNP	Bare Nuclear Potential	$BNP_j = \sum_{i=1}^{n} Z_i / r_{ij}$
PIP	Local Average Ionization Potential	$PIP(r) = \sum_i \rho_i(r) \cdot \lvert \varepsilon_i \rvert / \rho(r)$

1. Bader, R.F.W. *Atoms in Molecules: A Quantum Theory*; Oxford Univ. Press, 1994.

2. Breneman, C.M.; Rhem, M. *J. Comp. Chem.* 18, 182-197, 1997.

Rensselaer

Prediction of Column Performance (Ladiwala et al, A priori prediction of adsorption isotherm parameters and chromatographic behavior in ion-exchange systems PNAS 2005 102: 11710-11715)



Rensselaer

Hydrophobicity Descriptors

$$H_{surface} = \sum (h_i r_i)$$

Hydrophobicity values assigned to residue i:

- Miyazawa-Jernigan (MJ)
- Cowan-Whittaker (CW)
- Milton Hearn (H1 & H2)

$$r_i = \frac{s_i}{\sum s_i} = \frac{\text{exposed s.a. of residue } i}{\text{total exposed s.a. of protein}}$$

AND

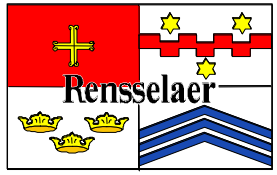
$$r_i = \frac{s_i}{asa_i} = \frac{\text{exposed s.a. of residue } i}{\left(\begin{array}{l}\text{s.a.s.a. of residue } i \text{ when it is} \\ \text{in a } Gly\text{-}X_i\text{-}Gly \text{ peptide}\end{array}\right)}$$

8 New Surface Hydrophobicity Descriptors

Rensselaer

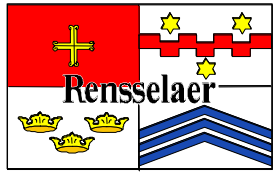
- Hydrophobicity values for amino acid residues based on the four hydrophobicity scales

	Cowan-Whittaker		Miyazawa-Jernigen		Hearn	
	Orig.	Norm.	Orig.	Norm.	1	2
Alanine	0.42	0.660	5.33	0.391	0.06	2.62
Arginine	-1.56	0.176	4.18	0.202	-0.85	1.26
Asparagine	-1.03	0.306	3.71	0.125	0.25	-1.27
Aspartic acid	-0.51	0.433	3.56	0.105	-0.20	-2.84
Cysteine	0.84	0.763	7.93	0.819	0.49	0.73
Glutamine	-0.96	0.323	3.87	0.151	0.31	-1.69
Glutamic acid	-0.37	0.467	3.65	0.115	-0.10	-0.45
Glycine	0.00	0.557	4.48	0.252	0.21	-1.15
Histidine	-2.28	0.000	5.10	0.354	-2.24	-0.74
Isoleucine	1.81	1.000	8.83	0.967	3.48	4.38
Leucine	1.80	0.998	8.47	0.908	3.50	6.57
Lysine	-2.03	0.061	2.95	0.000	-1.62	-2.78
Methionine	1.18	0.846	8.95	0.987	0.21	-3.12
Phenylalanine	1.74	0.983	9.03	1.000	4.80	9.24
Proline	0.86	0.768	3.87	0.151	0.71	-0.12
Serine	-0.64	0.401	4.09	0.188	-0.62	-1.39
Threonine	-0.26	0.494	4.49	0.253	0.65	1.81
Tryptophan	1.46	0.914	7.66	0.775	2.29	5.91
Tyrosine	0.51	0.682	5.89	0.484	1.89	1.39
Valine	1.34	0.885	7.63	0.770	1.59	2.30



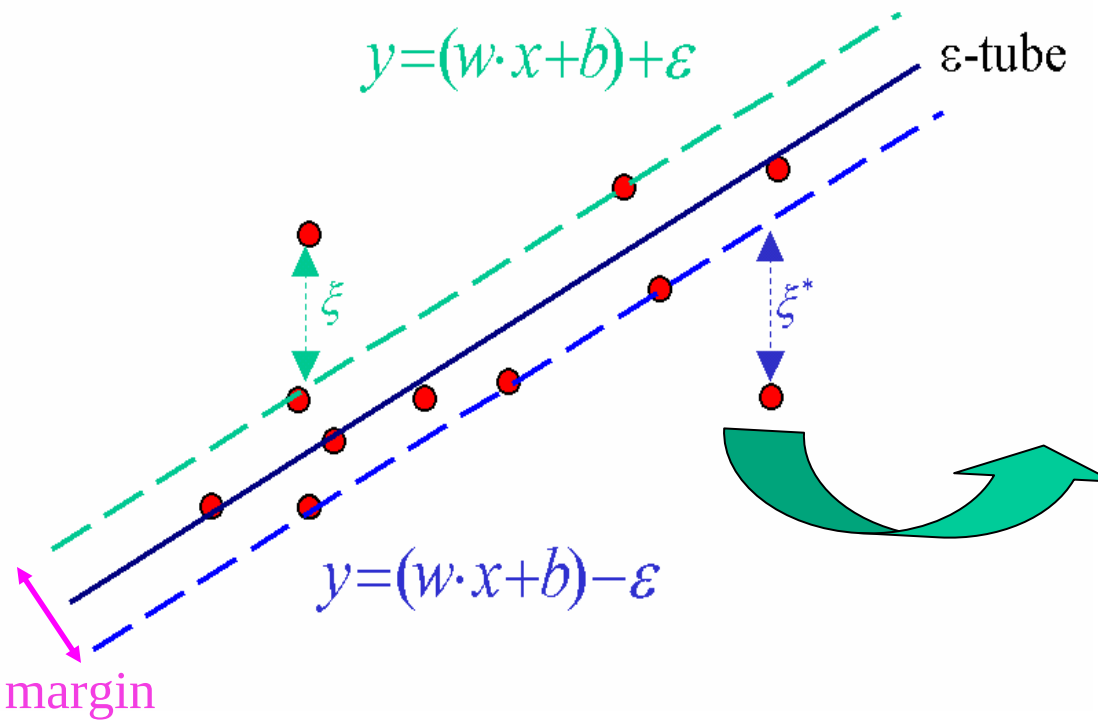
Machine Learning:

Support Vector Machines (SVM)



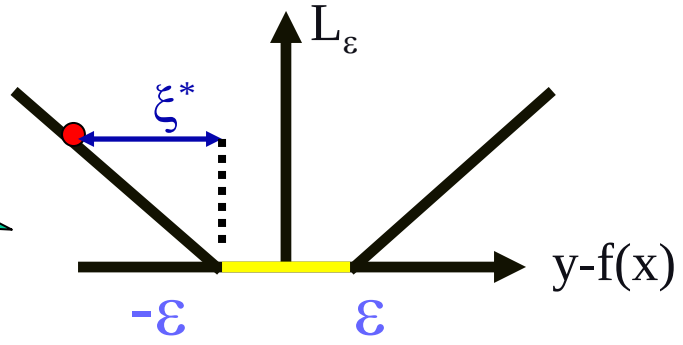
Support Vector Regression (SVR)

- Minimize the regularized empirical error:
  - training error + model complexity $\min_{w,b,\xi_i,\xi_i^*} C\sum_{i=1}^{l}(\xi_i + \xi_i^*) + \frac{1}{2}\|w\|^2$



ε-insensitive loss function:

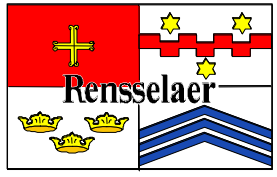
$L_\varepsilon(y - f(x)) := \max(0, |y - f(x)| - \varepsilon)$



- Avoid overfitting by controlling the model complexity

Rensselaer

Quantitative Structure-Retention Relationship Models for Protein Binding in HIC Systems



HIC: Protein Retention Data

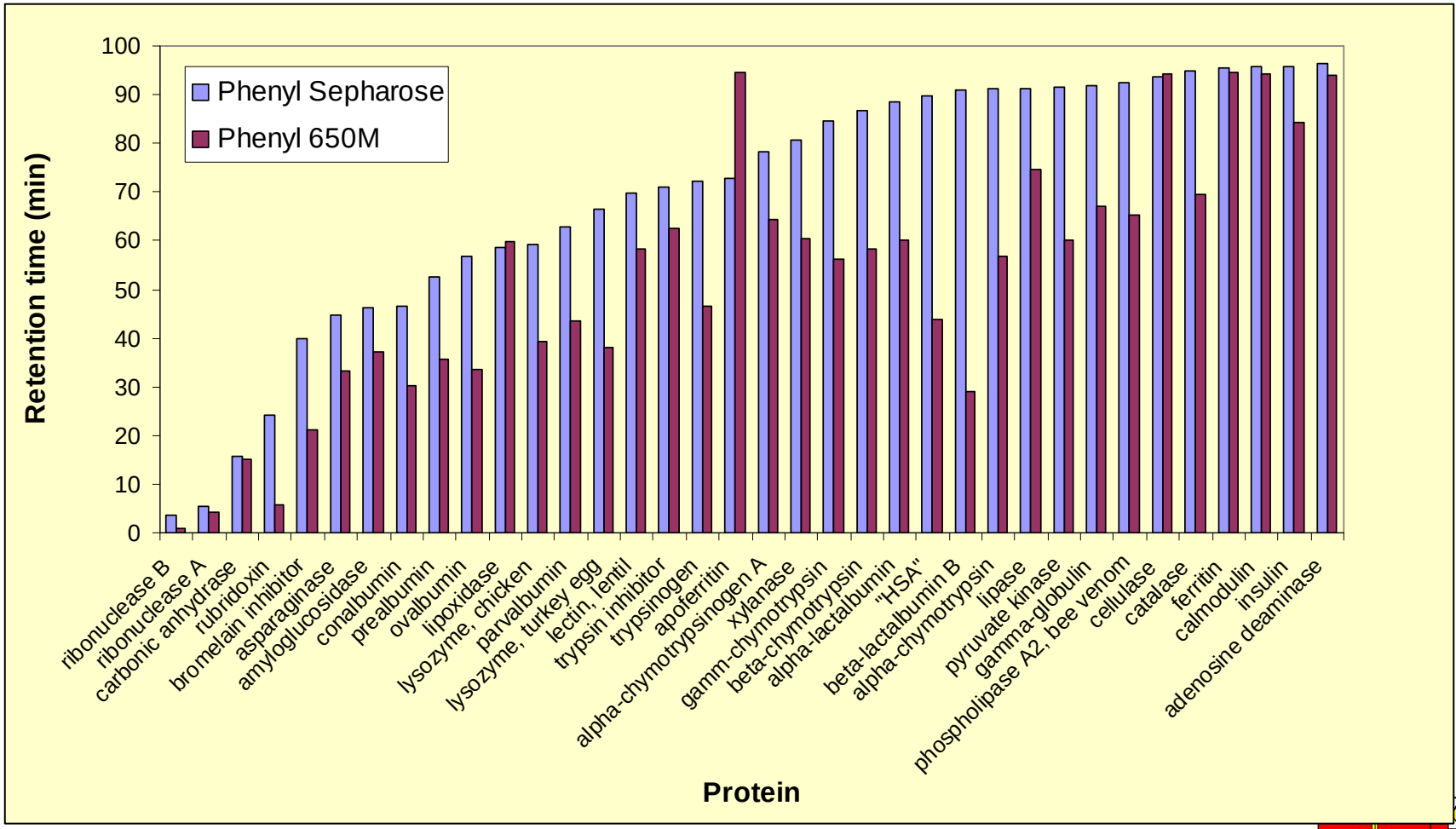
Retention Data on Butyl and Phenyl 650M Resins

Proteins exhibit differences in affinity for different ligand types

Rensselaer

HIC: Protein Retention Data

Retention Data on Phenyl Sepharose and Phenyl 650M Resins

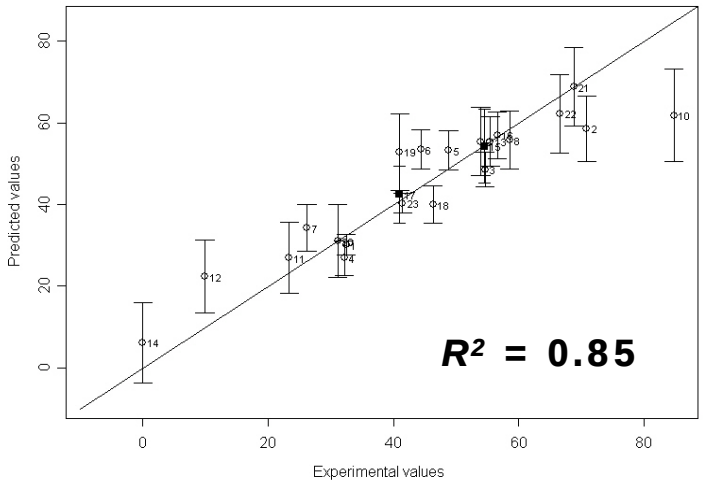


Differences in retention for different resin backbone chemistry

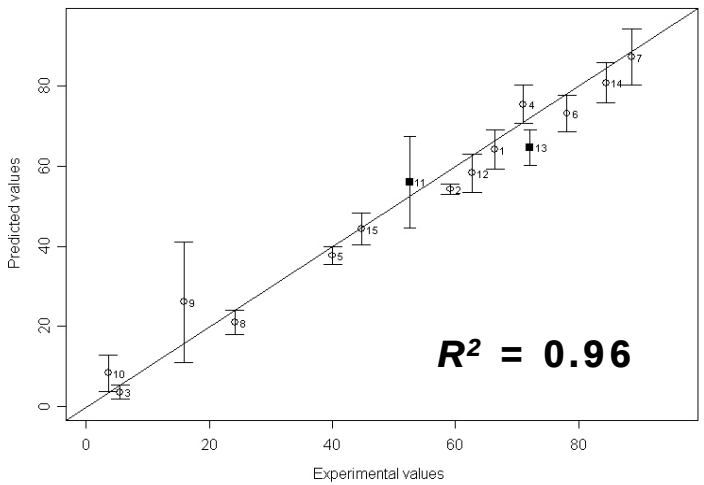
Rensselaer

QSRR: Training

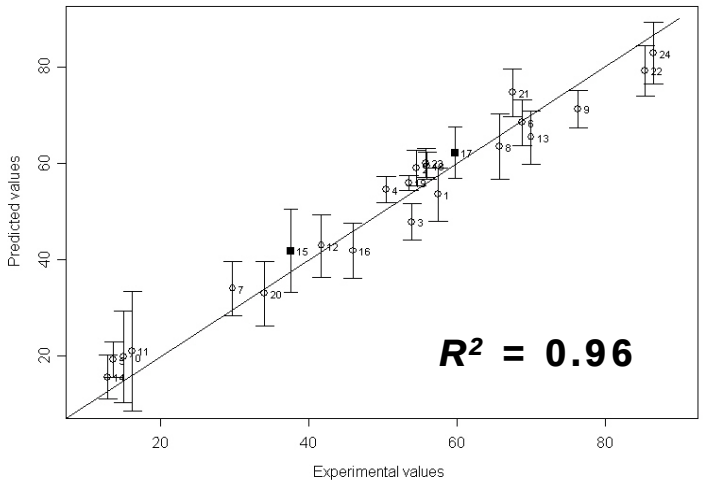
Butyl Sepharose



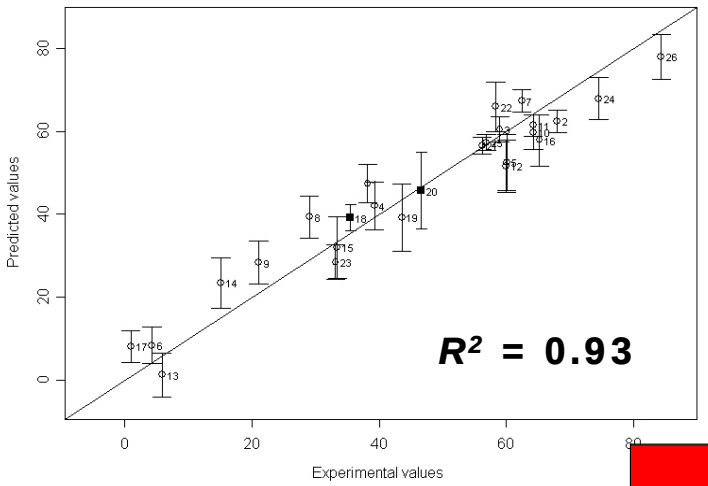
Phenyl Sepharose



Butyl 650M



Phenyl 650M

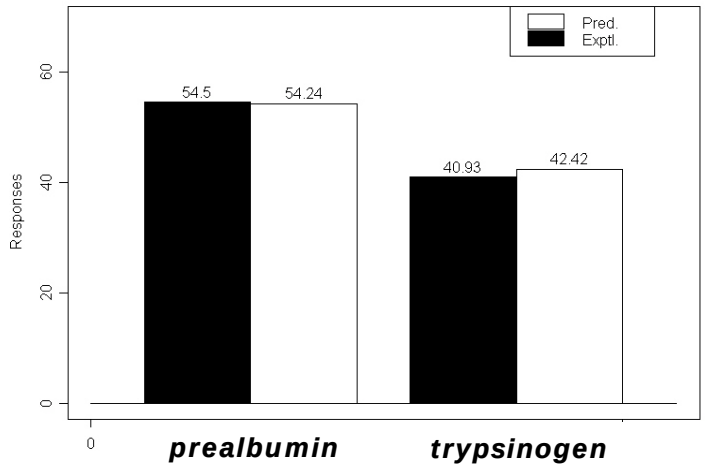


QSRR models can capture the differences in binding affinity

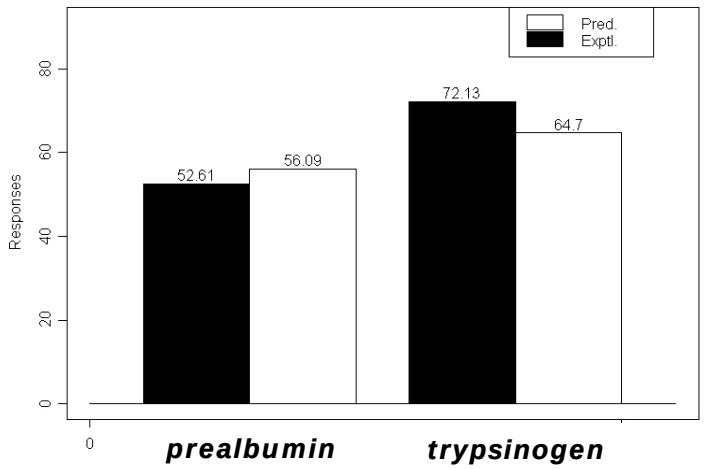
Rensselaer

QSRR: Test Set Predictions

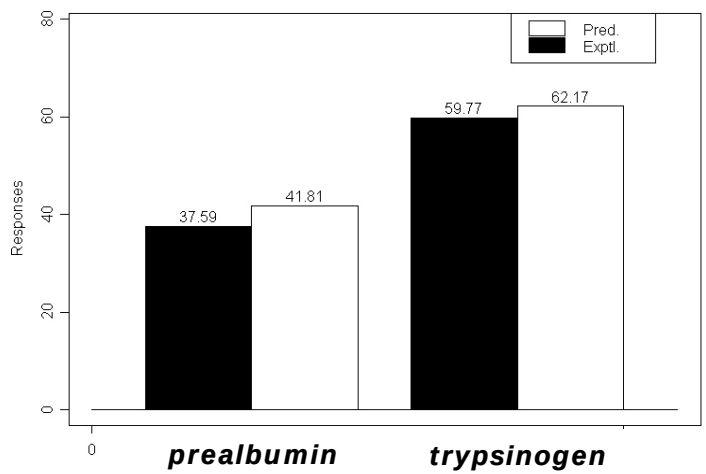
Butyl Sepharose



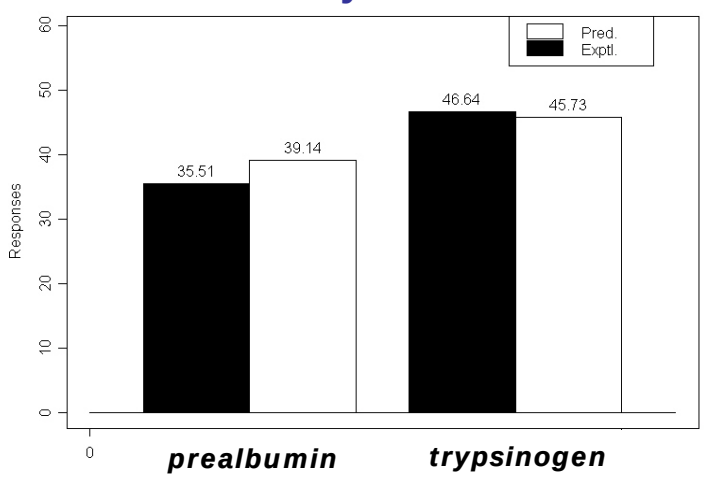
Phenyl Sepharose



Butyl 650M



Phenyl 650M



QSRR models can predict t_r for test set proteins

Rensselaer

QSRR: Model Validation

- *Y*-Scrambling Analysis
  - Test of the modeling algorithm

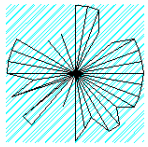
Model	"Real" model		"Scrambled" models		Probability $P(R^2_s \geq R^2_r)$
	R^2_r	Q^2_r	Avg. R^2_s	Avg. Q^2_s	
Butyl Sepharose	0.84	0.98	0.38	-2.36	0.53 %
Phenyl Sepharose	0.96	0.65	0.46	-4.84	2.24 %
Butyl 650M	0.96	0.90	0.42	-0.46	0.18 %
Phenyl 650M	0.93	0.77	0.35	-2.27	0.18 %

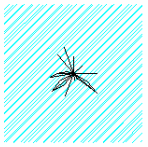
  - Extremely low *P* values indicate that the non-linear SVR algorithm cannot fit scrambled data

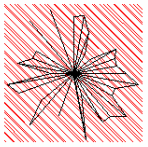
Can't fit random data using the SVM modeling approach

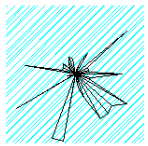
Rensselaer

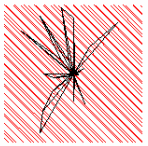
QSRR: Star Plots

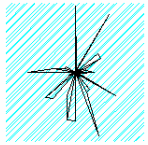

"STD.DIM2 + 20.8%"

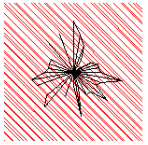

"DEL.RHO.NIA + 9.8%"


"REACTIVE - 17.8%"

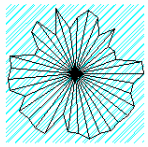

"PEOE.VSA.FHYD + 12.7%"

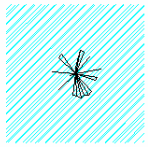

"Q.VSA.FPPOS - 12.4%"

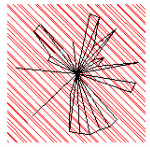

"E.STB + 12.5%"

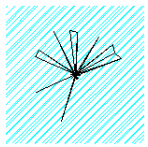

"B.ROTR - 14.1%"

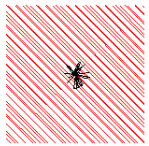
Butyl Sepharose

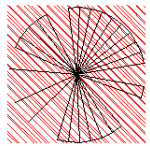

"PEOE.VSA.FHYD + 21.7%"

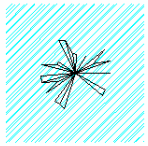

"DASA + 7%"

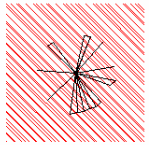

"Q.VSA.FPOS - 15.3%"


"SIEPIA + 10.2%"

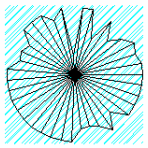

"E.OOP - 5.8%"

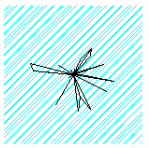

"B.ROTR - 22.6%"

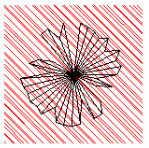

"DEL.RHO.NIA + 8.3%"

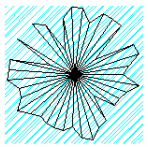

"DENSITY - 9.2%"

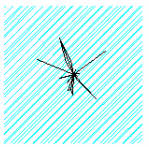
Phenyl Sepharose

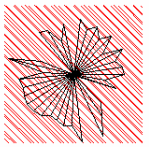

"HYDROSURFDE.H1 + 18.5%"

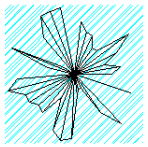

"SIEPIA + 6%"

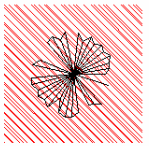

"DEL.RHO.NMIN - 13.8%"


"HYDROSURF.MJ + 16.7%"

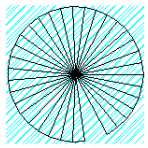

"A.ICM + 5.4%"

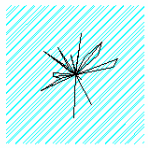

"B.1ROTR - 14.1%"

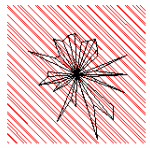

"DEL.K.IA + 14%"

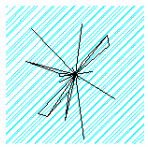

"REACTIVE - 11.5%"

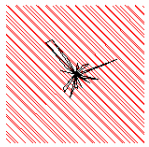
Butyl 650M

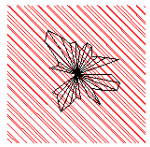

"HYDROSURF.MJ + 34.5%"

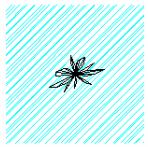

"FASA + 5.7%"

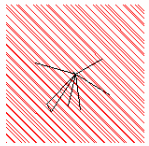

"PEOE.VSA.FPNEG - 13.7%"


"PEOE.VSA.FHYD + 12.1%"


"B.ROTR - 6.4%"


"REACTIVE - 15.1%"

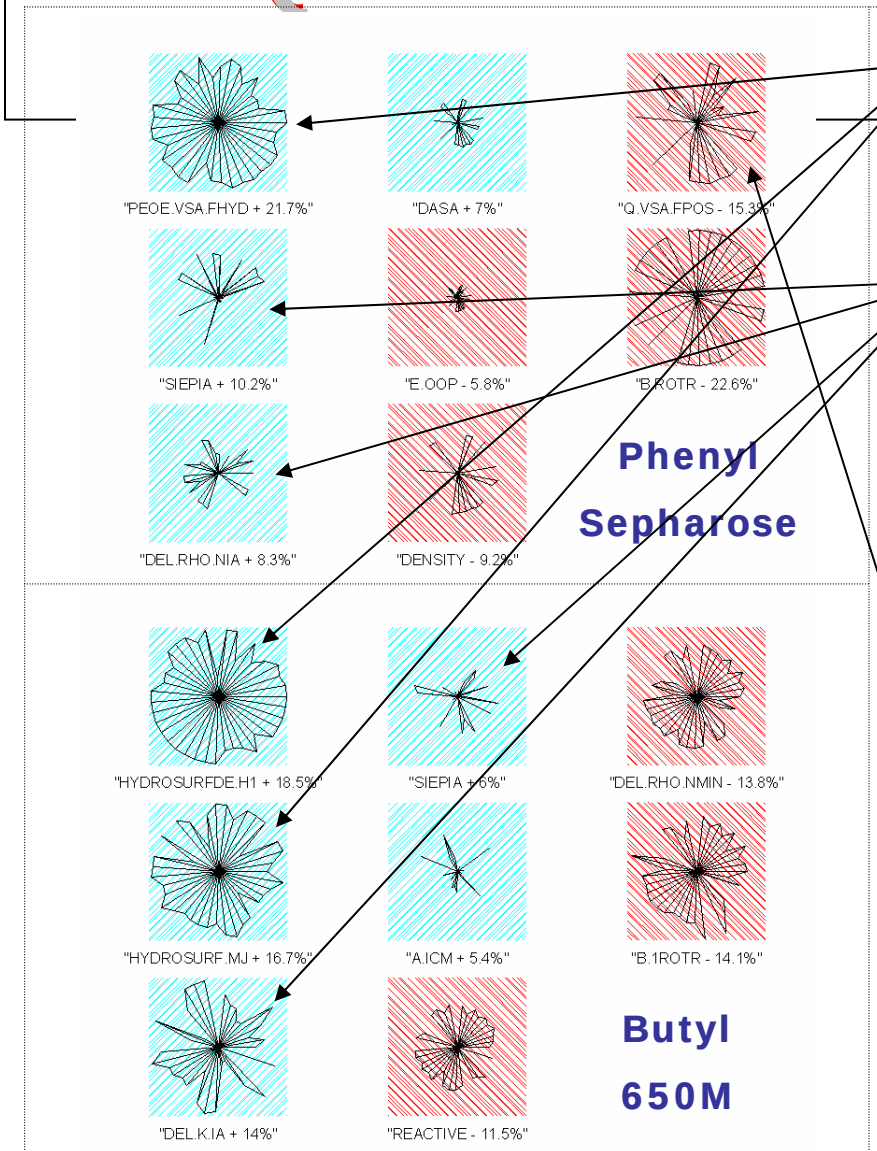

"HYDROSURFDE.H1 + 6%"


"HYDROSURF.CW - 6.5%"

Phenyl 650M

Rensselaer

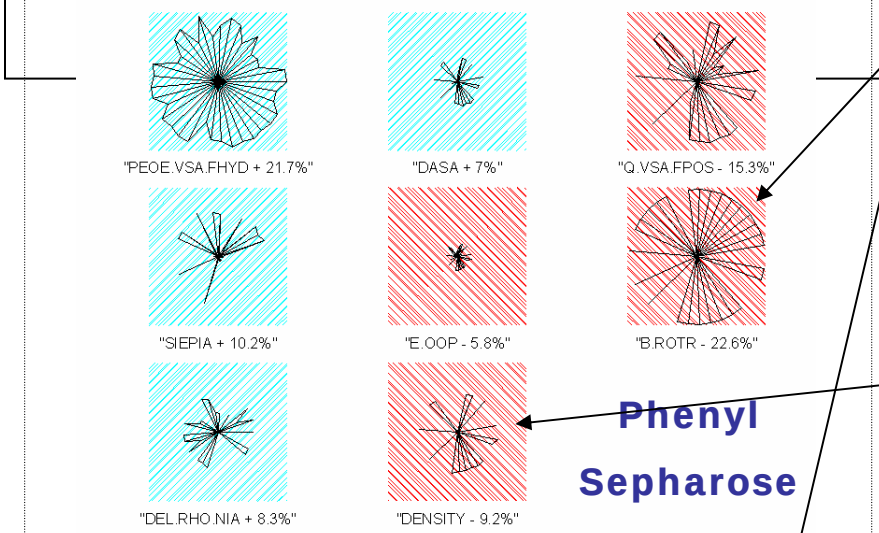
QSRR: Model Interpretation



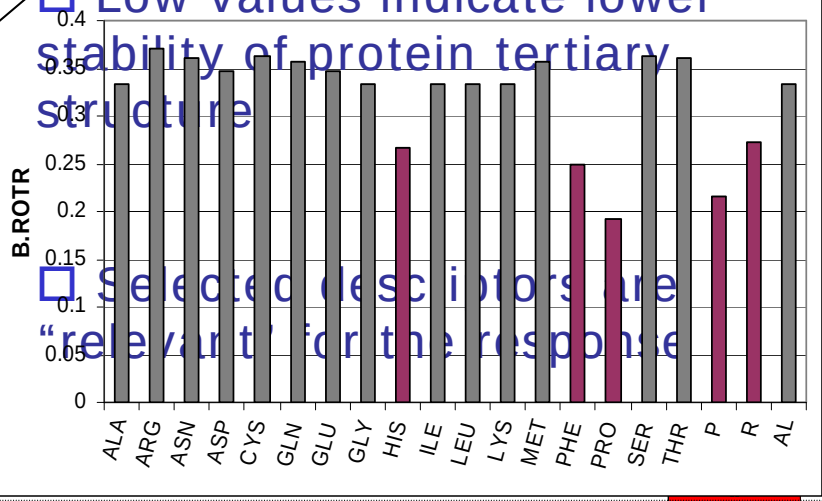
- Exposed hydrophobic (hϕ) surface area of the protein
- Associated with regions of low polarizability
  - High (less negative) values indicate the presence of more non-polar/hϕ residues
- Polar/charged surface area

Rensselaer

QSRR: Model Interpretation

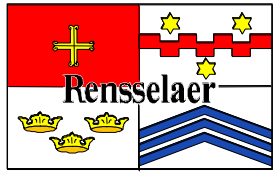


- Number of rotatable bonds on the amino acid side chains
  - Low values associated mainly with hϕ residues
- Molecular density. Low values imply a "floppy" protein
- Low values indicate lower stability of protein tertiary structure
- Selected descriptors are "relevant" for the response



Rensselaer

Investigation of Protein Binding in HIC Systems under Low Salt Conditions



Motivation

- **Industrial HIC processes which employ low salt binding conditions are desirable for the following reasons:**
  - *reduce protein denaturation*
  - *improve protein recovery*
  - *reduce the expense associated with high salt buffer preparation*
  - *minimize the time and cost related to desalting*

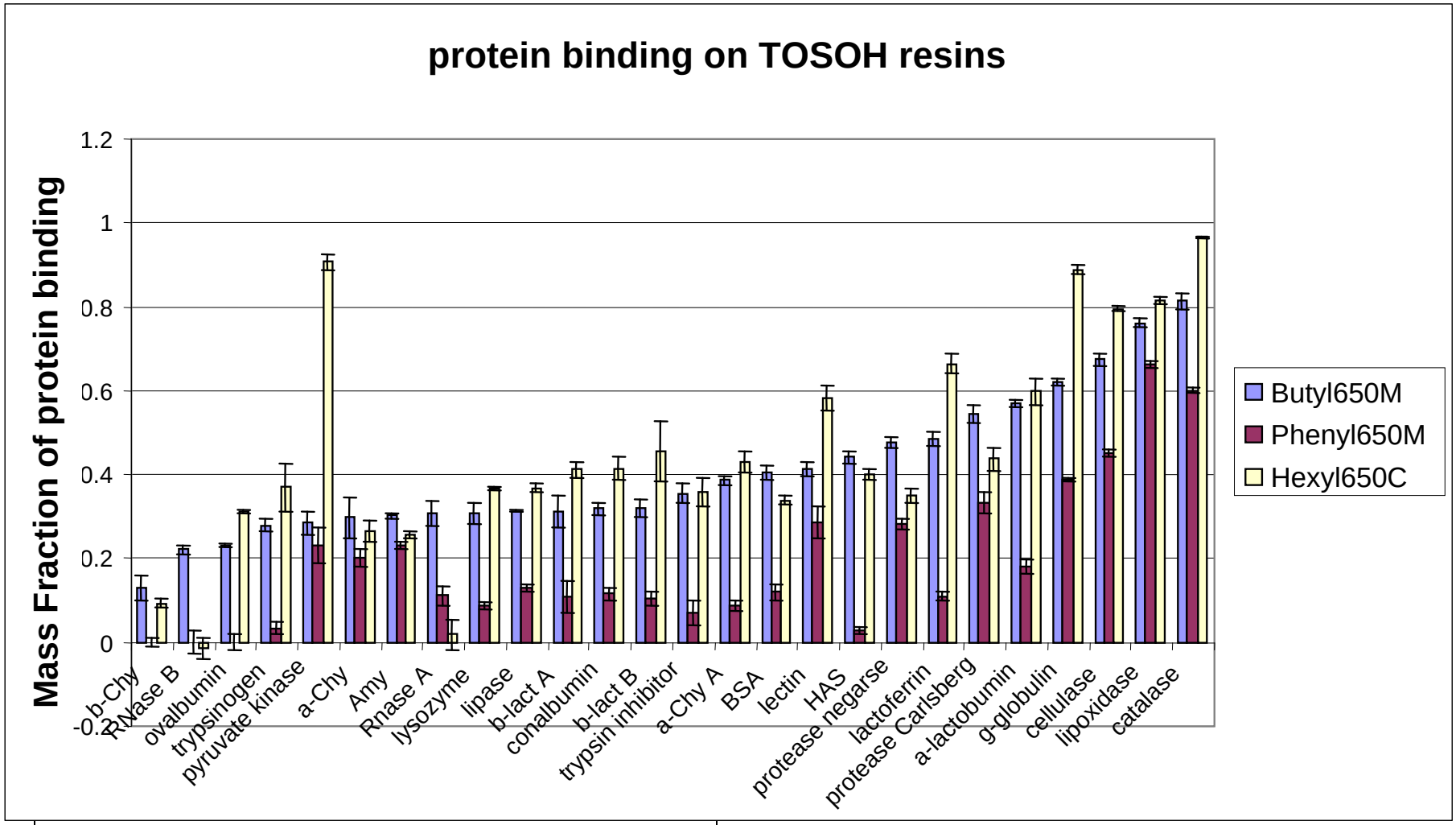
Experiment Conditions:

- 26 proteins
- Batch mode by applying High Throughput Screening (HTS) technique
- Binding at 0.5 M $(NH_4)_2SO_4$, 25 mM phosphate pH7.0 and elution at 25 mM phosphate pH7.0 buffer.
- 5 resins:

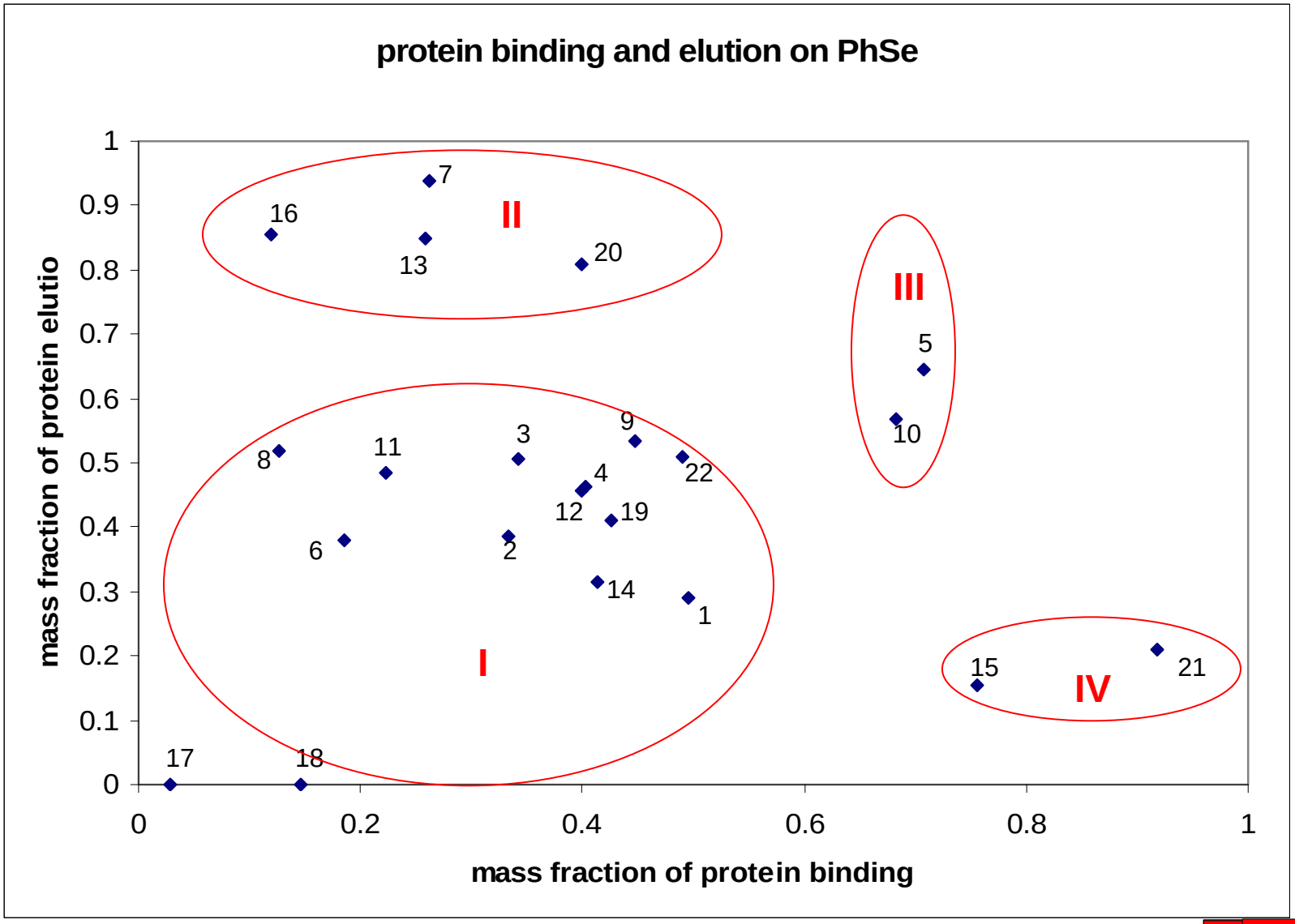
  GE Healthcare Resin: Butyl Sepharose, Phenyl Sepharose (high sub)

  Tosoh Resin: Butyl 650M, Phenyl 650M and Hexyl 650C.

Comparison of Protein Binding on Different Resins



Protein Binding and Elution (on Phenyl Sepharose_high sub)



Summary of Protein Classification on Different Resins

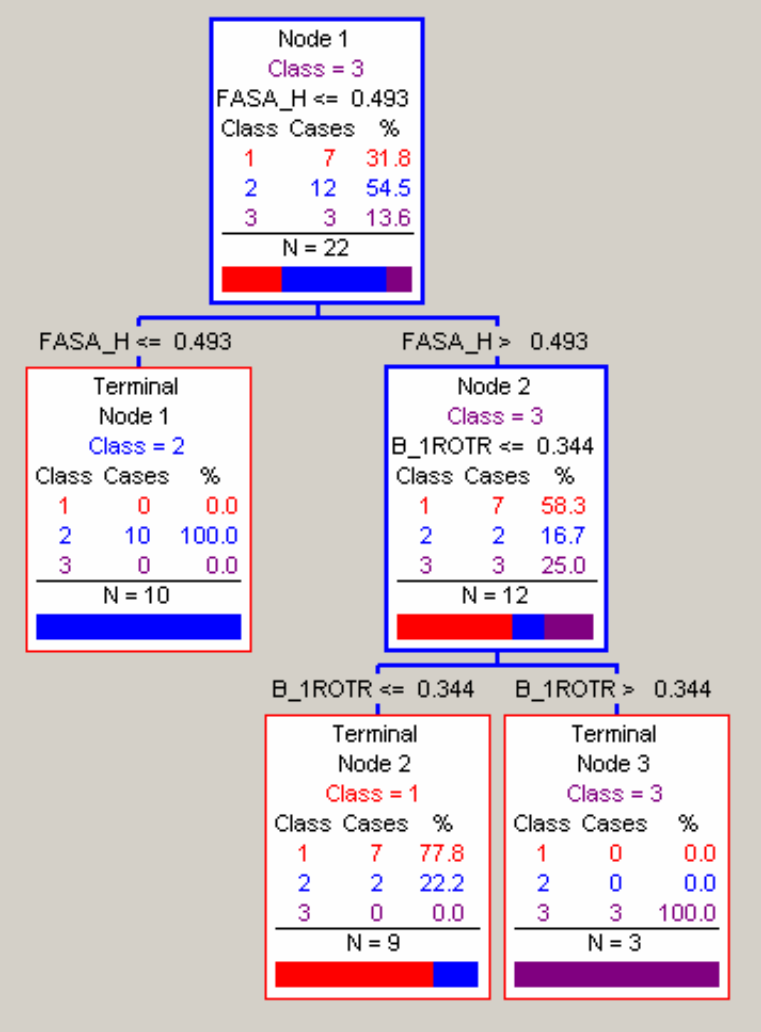
	BuSe	PhSe	Bu650	Ph650	Hx650
Class I: low binding/ low elution	1,3,5,6,7,8, 11,12,13,161 7,18,19,2022, 23,25	1,2,3,6,8,9,111 2,14,1718, 19,22,23,26	1,2,3,5,6,8,9, 11,12,14,16, 17,18,19,20, 21,22,2325,26	1,2,3,5,6,7,811 ,12,13,14,16,1 7,18,19, 22,23,24,25,26	1,2,3,7,8, 11,12,13, 16,17,18, 20,21,22, 23,26
Class II: low binding/ high elution	2,4,9,10 14,24,26	7,13,16,20,25	7,13	9,18,20	6,25
Class III: High binding/ high elution	21	5,10,24	10,24	10	5,19
Class IV: high binding/ low elution	15	15,21	15,21	15,21	9,10,14 15,21, 24

Decision Tree Learning for Protein Binding

- Recursive Partitioning (RP) : discover logical patterns within datasets
- Given data characterized by descriptors and belonging to different categories, derive rules based on the descriptors which correctly categorizes as many observations as possible.
- Method identifying the best splitting rule at each step is important. (e.g. Gini Impurity score minimize the impurity of the resultant nodes.)
- Output in the form of a tree diagram
- CART (Classification And Regression Trees)
  - Developed by Stanford University and UC Berkeley
  - Automatic Self-Validation Procedures
- Data: 22 proteins categorized according to binding percentage (high, medium, low) on 5 different resins.

Rensselaer

BuSe (CART analysis)



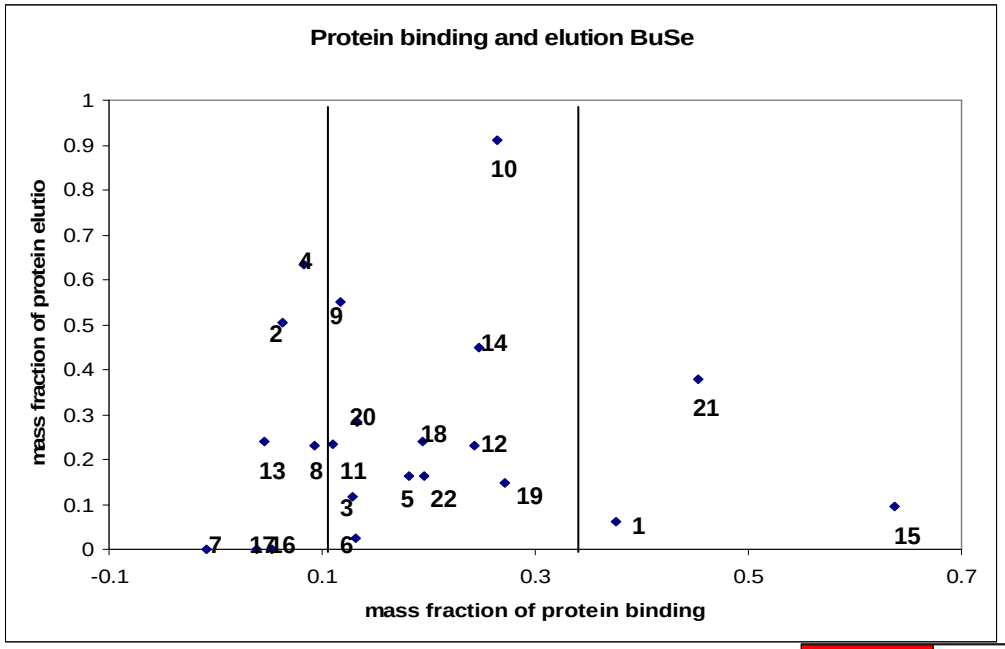
Terminal Node 1: 3,5,6,9,10,12,14,18,19,22.

Terminal Node 2: 2,4,7,8,11,13,16,17,20.

Terminal Node 3: 1,15,21.

FASA_H: fractional water accessible surface area of all hydrophobic atoms.

B_1ROTR: fraction of rotatable single bonds.



Rensselaer

Protein Similarity using PEST: Shape-Aware Molecular Descriptors from Property/Segment-Length Distributions

PEST (Property-Encoded Surface Translation) adds shape information that encodes the spatial relationships of surface properties.

A property-encoded surface is subjected to internal ray reflection analysis.

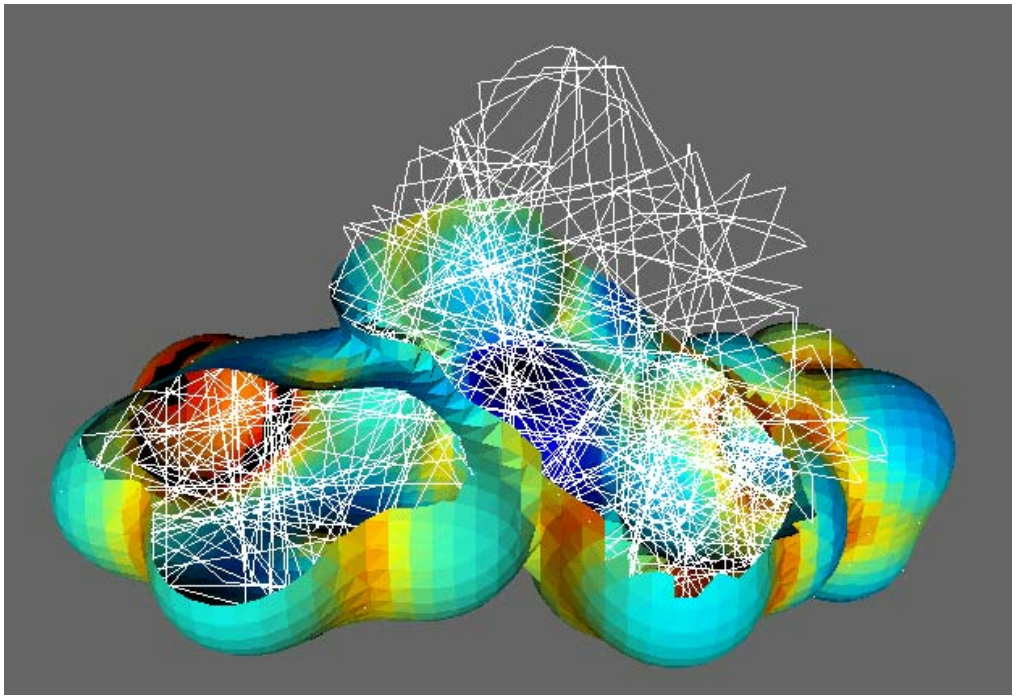
Molecular shape information is obtained by recording the ray-path information, including segment lengths, reflection angles and property values at each point of incidence.

Breneman et al., "*New developments in PEST shape/property hybrid descriptors*" *J. Computer-Aided Mol. Design*, **17,** 231–240, (2003)

Rensselaer

PEST Descriptors

- TAE Internal Ray Reflection - low resolution scan



Isosurface (portion removed) with 750 segments

Protein EP & Hydrophobic Mapping

- **EP:** $\sum \frac{Q_i}{\varepsilon r_{ij}}$
- **Hydrophobicity**
  - **MLP (Molecular Lipophilic Potential):**

$$MLP(t) = \sum_{i=1}^{n} f_i / (1 + d_{it})$$

  - **MHM(Molecular Hydrophobic Mapping)**

$$MLP\ 2 = \frac{\sum_i f_i g(d_i)}{\sum_i g(d_i)}$$

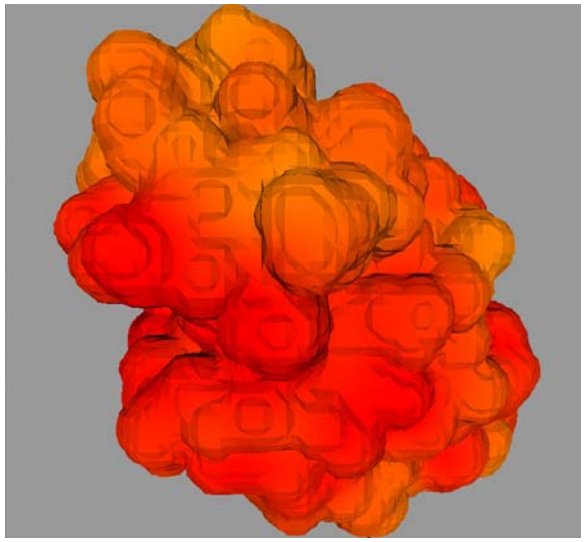
$$g(d_i) = \frac{\exp\lfloor -a \cdot d_{cut-off} \rfloor + 1}{\exp\left[a\left(d_i - d_{cut-off}\right)\right] + 1} \qquad d_{cut\text{-}off} = 4\text{Å},\ a = 1.5$$

  - **HINT**

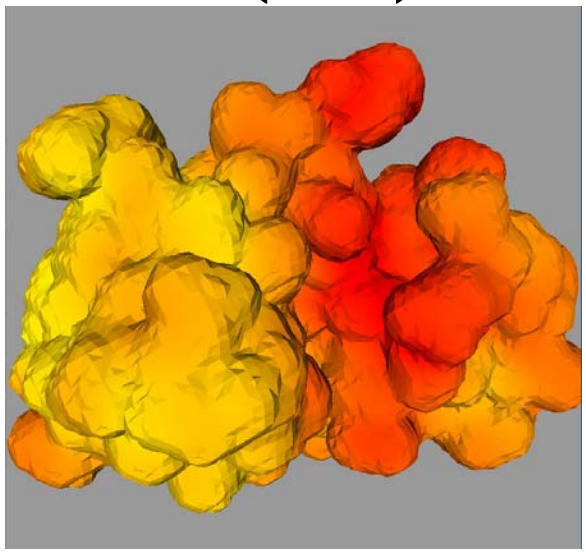
$$A_t = \sum_{i=1}^{n} S_i a_i R_{it}(r) \qquad R_{it}(r) = e^{-r}$$

Rensselaer

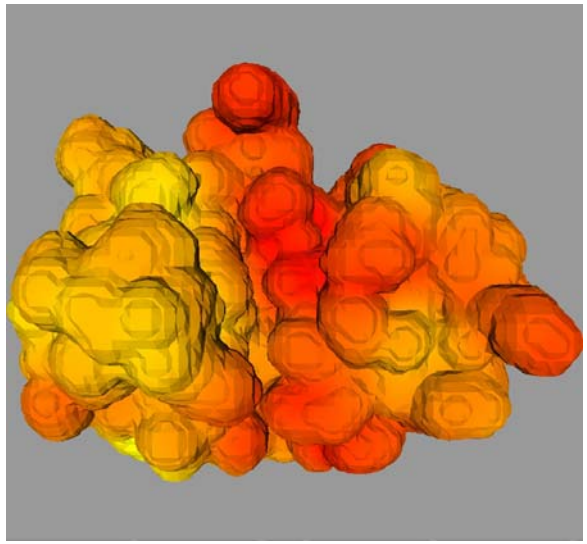
EP(GL)



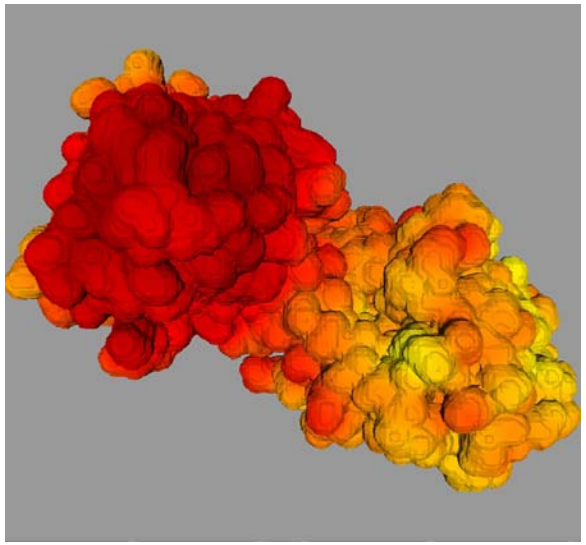
lysozyme



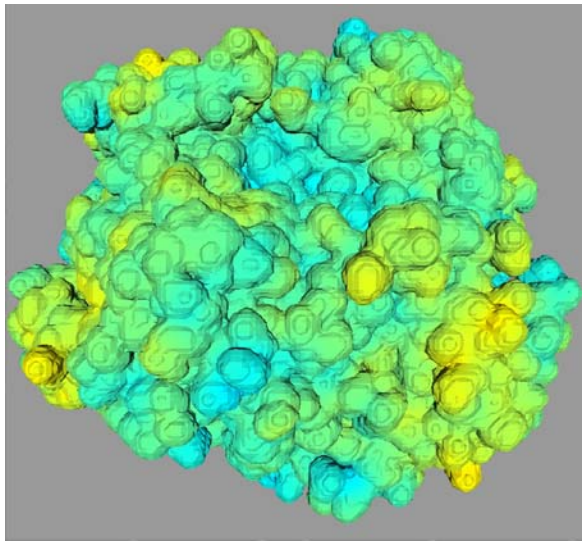
RnaseA



RnaseB



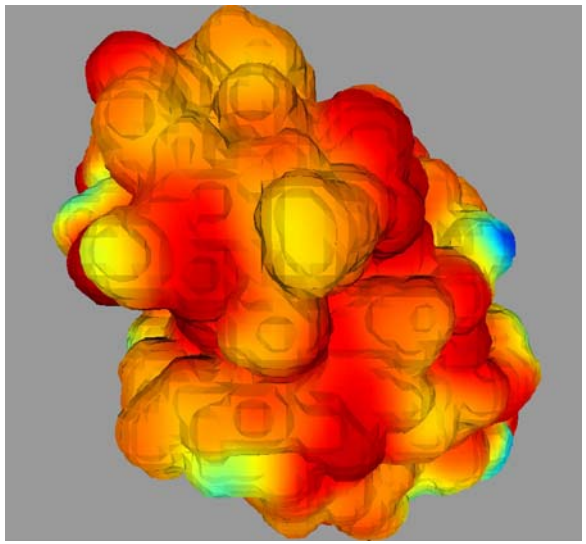
Lactoferrin



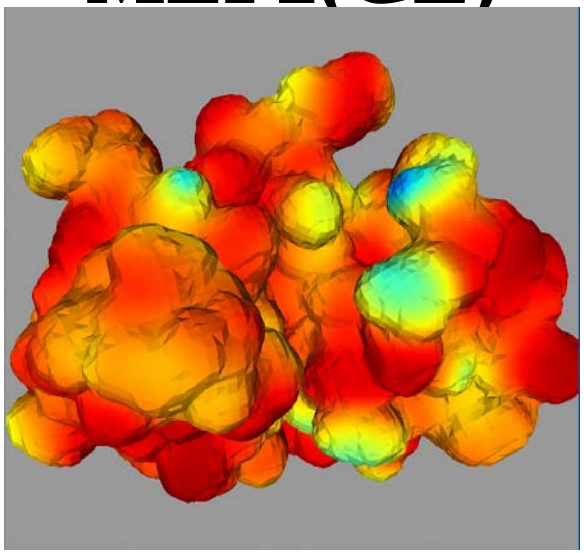
Catalase

Rensselaer

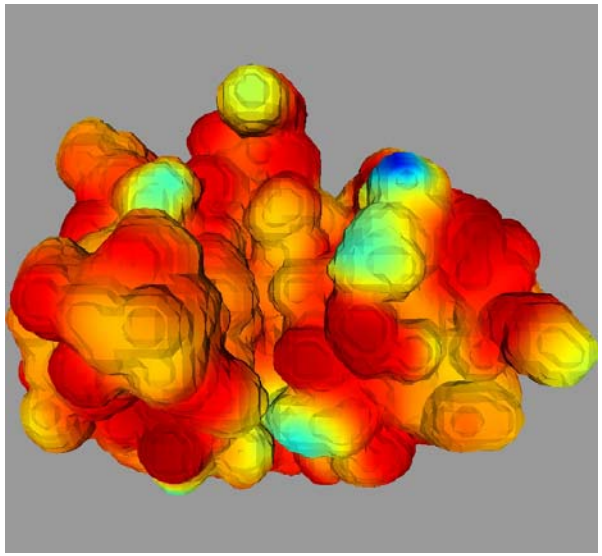
MLP2(GL)



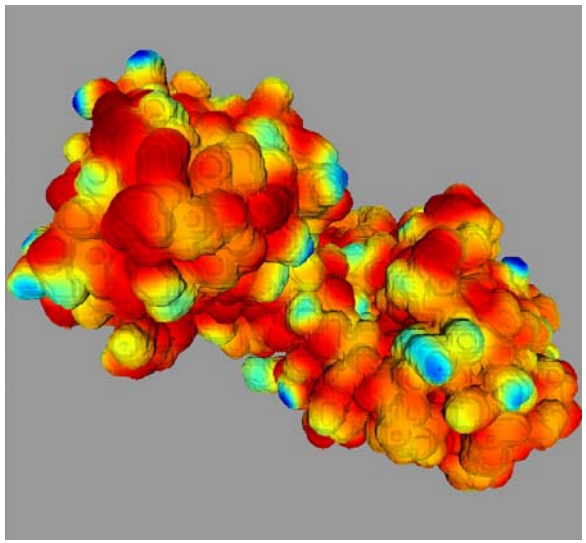
lysozyme



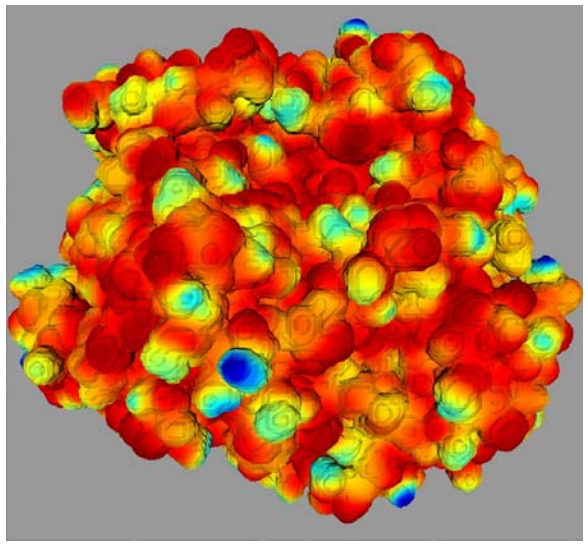
Rnase A



Rnase B



Lactoferrin

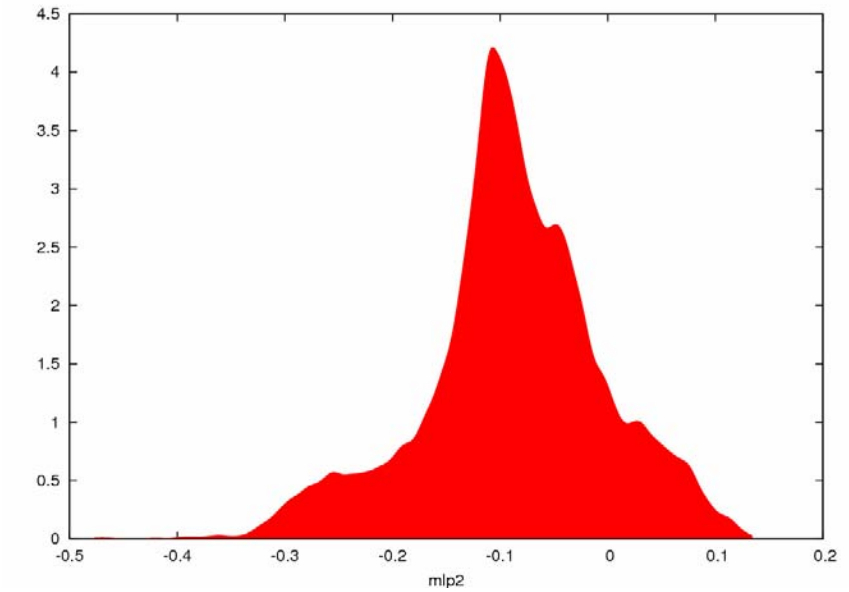


Catalase

Rensselaer

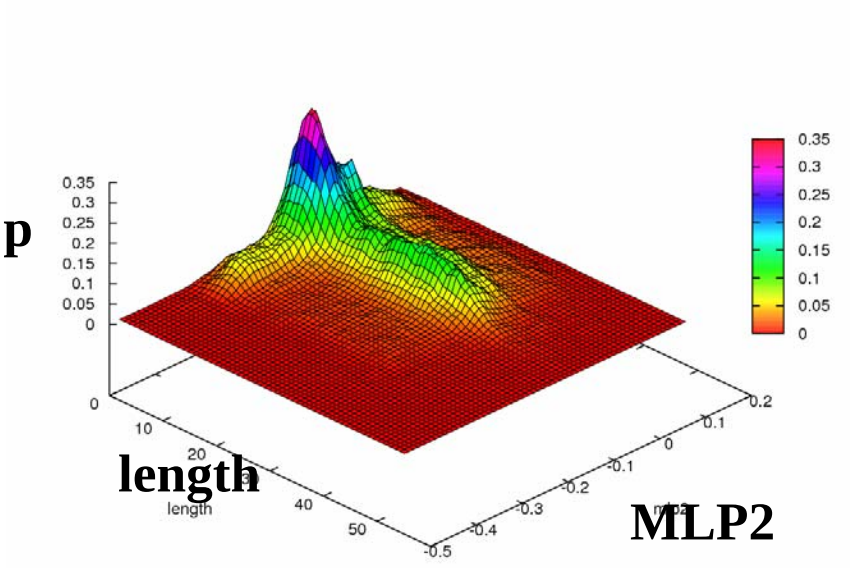
PPEST lysozyme mlp2

P



MLP2

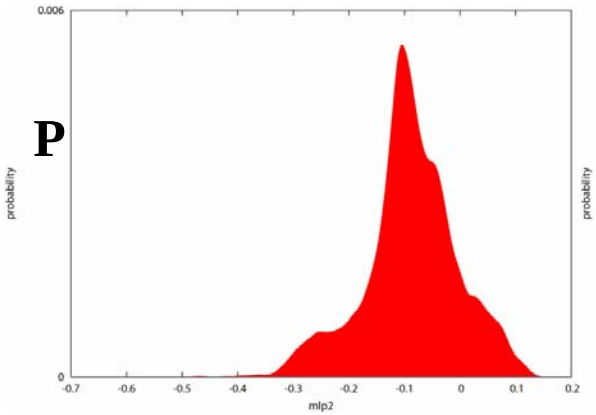
p



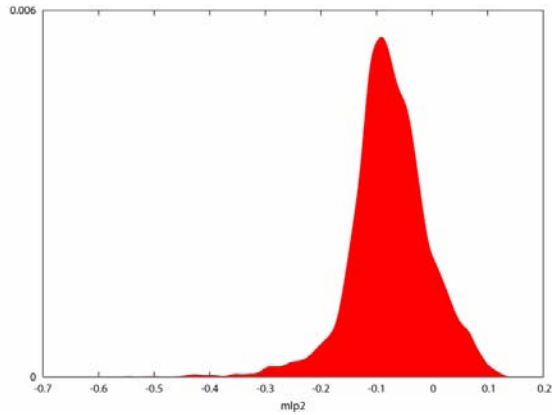
Rensselaer

MLP2(GL)

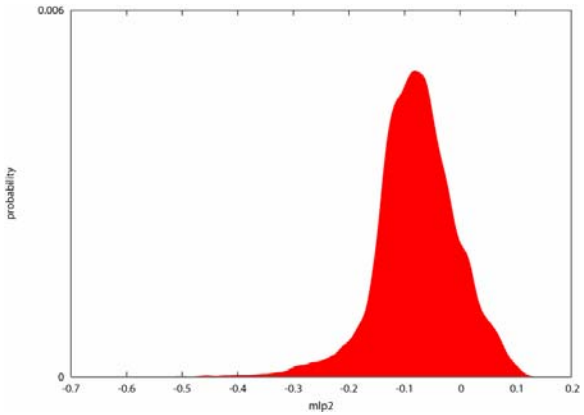
P



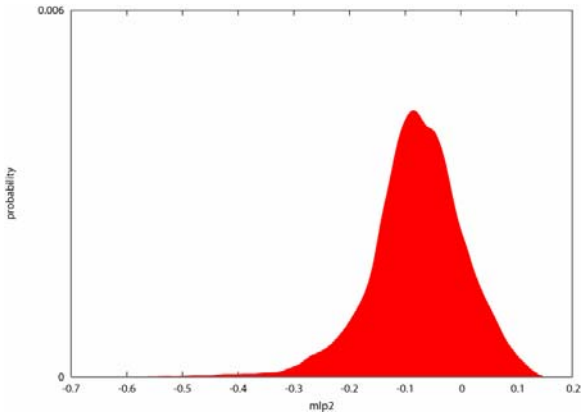
lysozyme



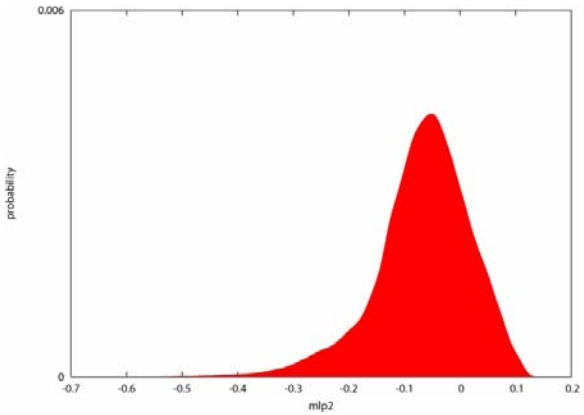
RnaseA



Rnase B



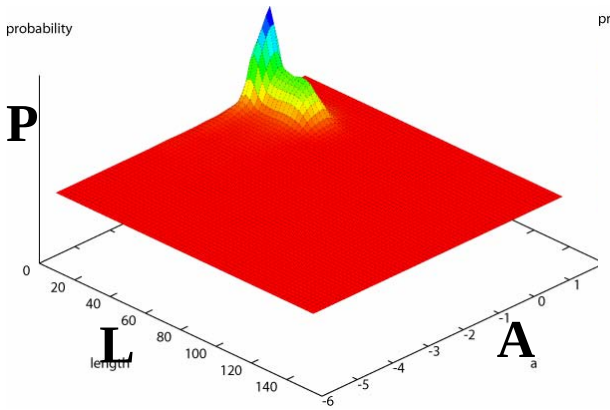
Lactoferrin



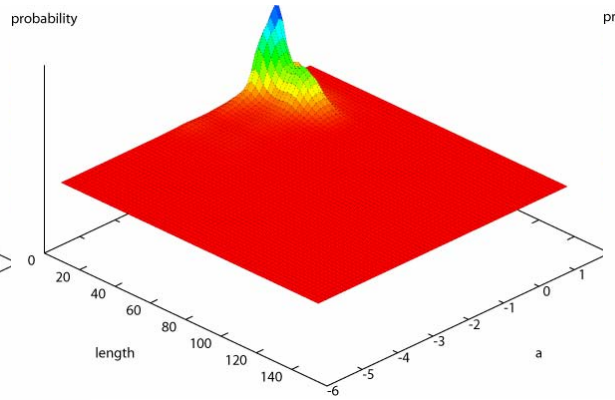
Catalase

Rensselaer

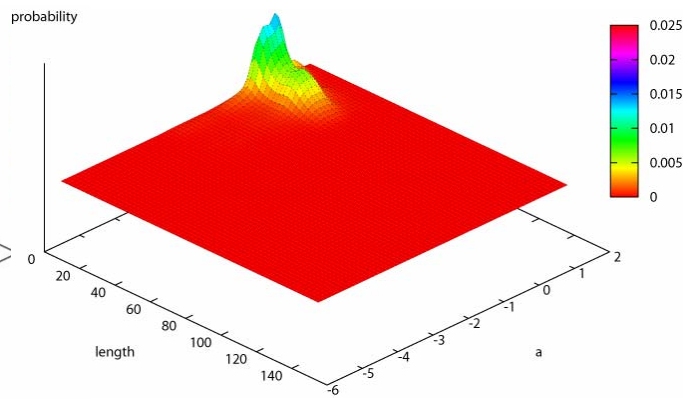
A(GL)



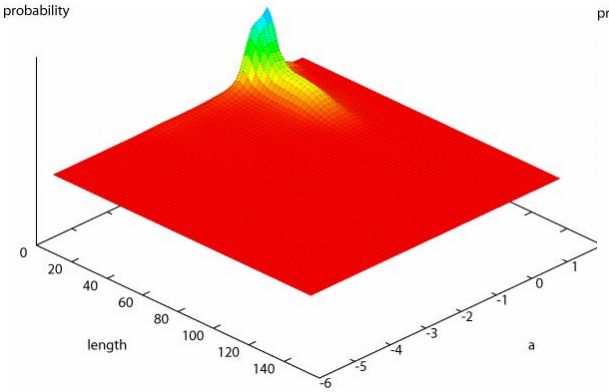
lysozyme



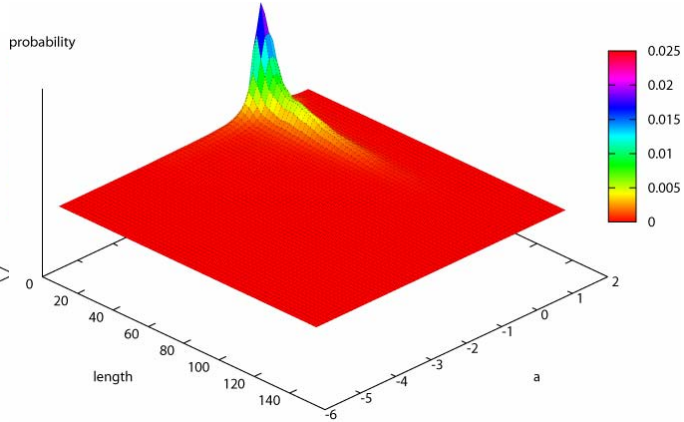
Rnase A



Rnase B

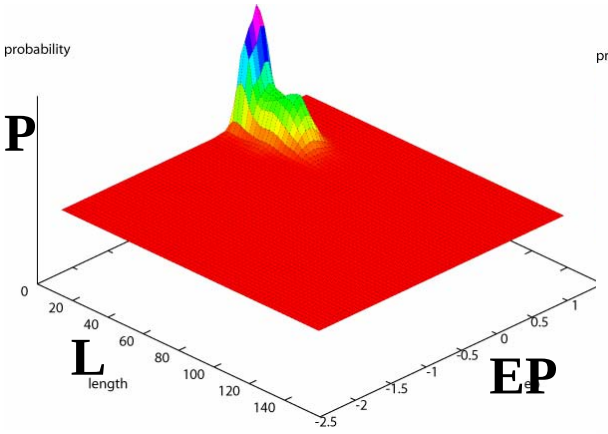


Lactoferrin

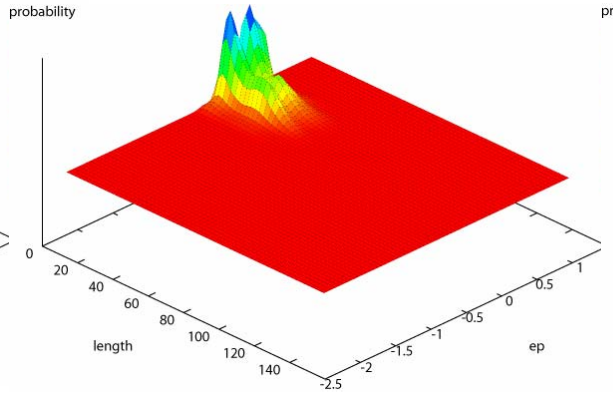


Catalase

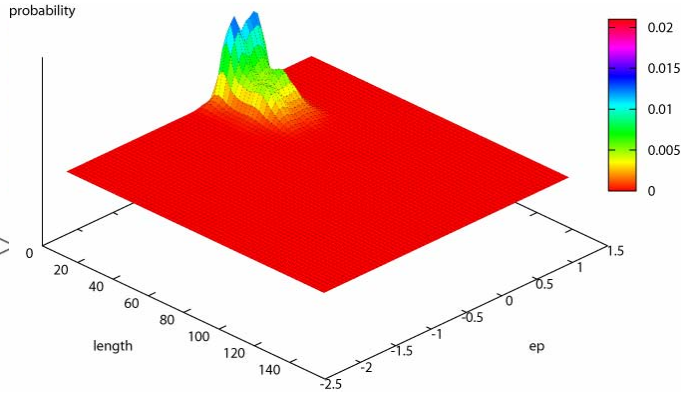
EP(GL)



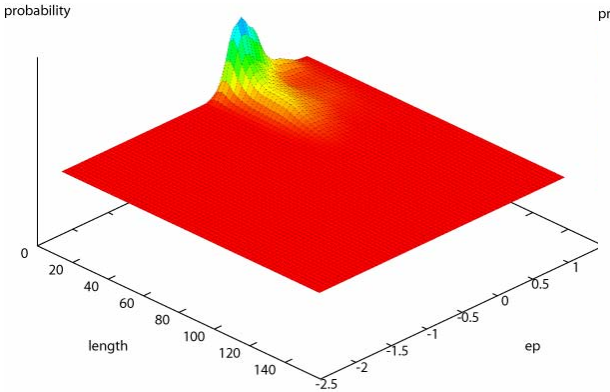
lysozyme



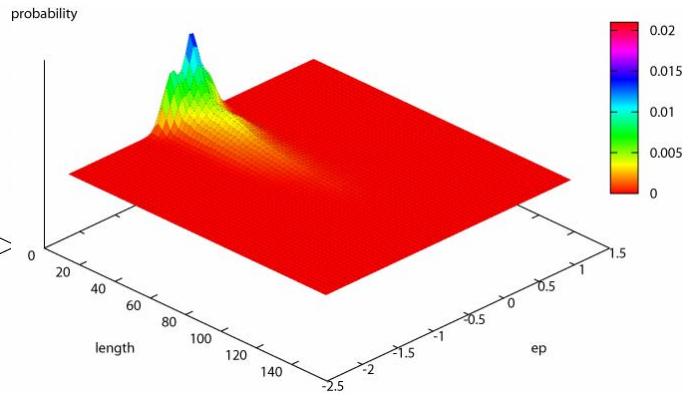
RnaseA



Rnase B



lactoferrin



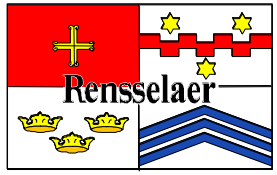
catalase

Rensselaer

Similarity Measurement

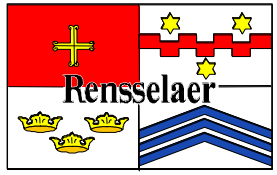
$$d_{ij} = 1 - \frac{2 \cdot \sum_{k=1}^{K} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{K} x_{ik} + \sum_{k=1}^{K} x_{jk}}$$

- x_{ik}: value of the kth descriptor for the ith protein
- range from 0 to 1.
  - 0: complete identity
  - 1: have nothing in common



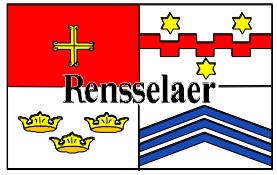
MLP2

mlp2(d1)	lys	RnaseA	RnaseB	lactoferrin	catalase
Lys	0	0.120	0.105	0.130	0.229
RnaseA	0.120	0	0.022	0.139	0.205
RnaseB	0.105	0.022	0	0.132	0.194
Lactoferrin	0.130	0.139	0.132	0	0.112
Catalase	0.229	0.205	0.194	0.112	0



EP & MLP2

ep&mlp2(d2)	lys	RnaseA	RnaseB	lactoferrin	catalase
Lys	0	0.449	0.451	0.351	0.761
RnaseA	0.449	0	0.043	0.366	0.693
RnaseB	0.451	0.043	0	0.375	0.690
Lactoferrin	0.351	0.366	0.375	0	0.707
catalase	0.761	0.693	0.690	0.707	0

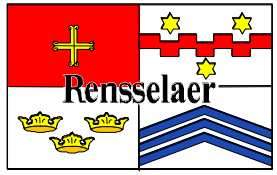


Potential uses of these approaches for follow-on biologics

- After identifying key variants by mass spec, use QSRR to design appropriate analytical chromatographic steps for quantitation and/or process chromatographic steps for variant removal.
- Carry out detailed similarity measurements using a range of property-shape hybrid molecular descriptors to examine the “similarity” of follow on protein products with respect to various properties.

Summary

- QSPR models were successfully generated for predicting protein retention in HIC systems from protein sequence and crystal structure.
- Proteins can be classified based on their low salt binding and subsequent elution and CART can be employed as a classification tool.
- The ability to quantitatively relate shape, surface EP, and surface MLP differences between proteins without alignment provides new information for studying protein surface hydrophobicity and for evaluating protein similarities.
- The synergy of these methods provides a unique opportunity to develop powerful predictive tools and methods for gaining significant insight into the fundamental physics of the protein chromatographic processes.



Acknowledgements

- Students: Asif Ladiwala, Jie Chen, Fang Xia, Matt Sundling and Qiong Luo.
- Professors: Curt Breneman, Kristen Bennett.
- Funding: NIH, NSF (PHAT), GE Healthcare,

Rensselaer