Electron density derived descriptors in ADME/Tox screening

Presented by

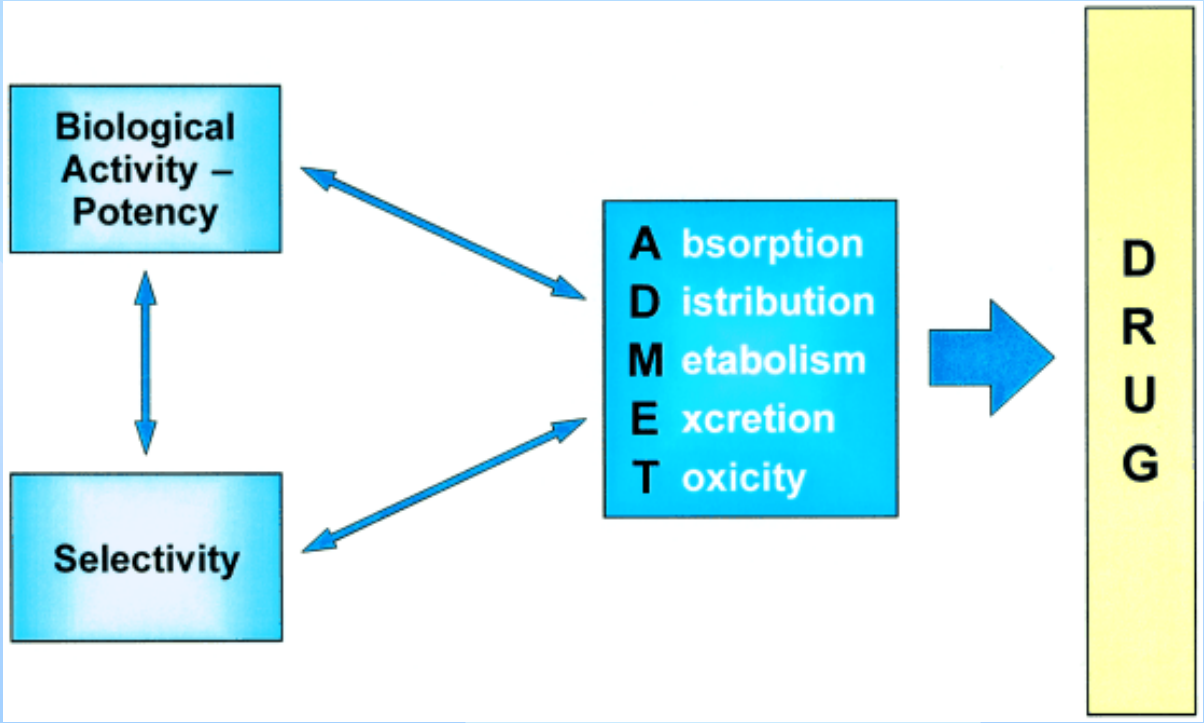
N. SUKUMAR

Curt M. Breneman, Mark J. Embrechts, Kristin P. Bennett and Dechuan Zhuang

http://www.drugmining.com/

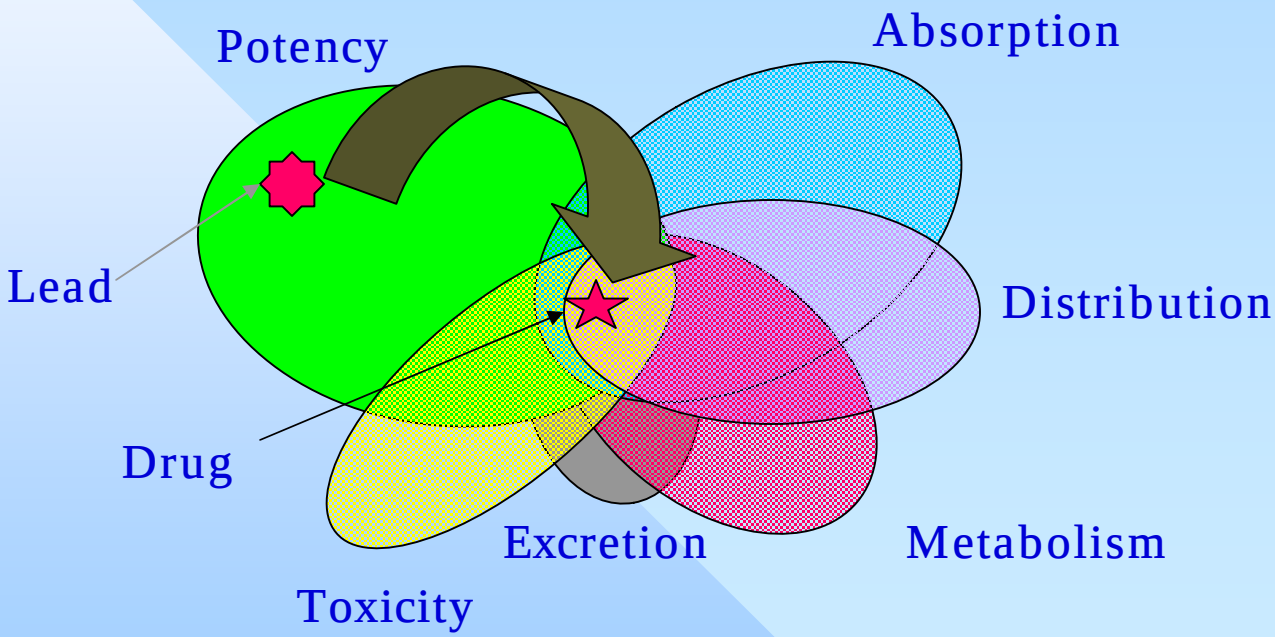
ADMET Property Prediction: Challenges in Medicinal Chemistry

Multiple-parameter optimization of lead structures

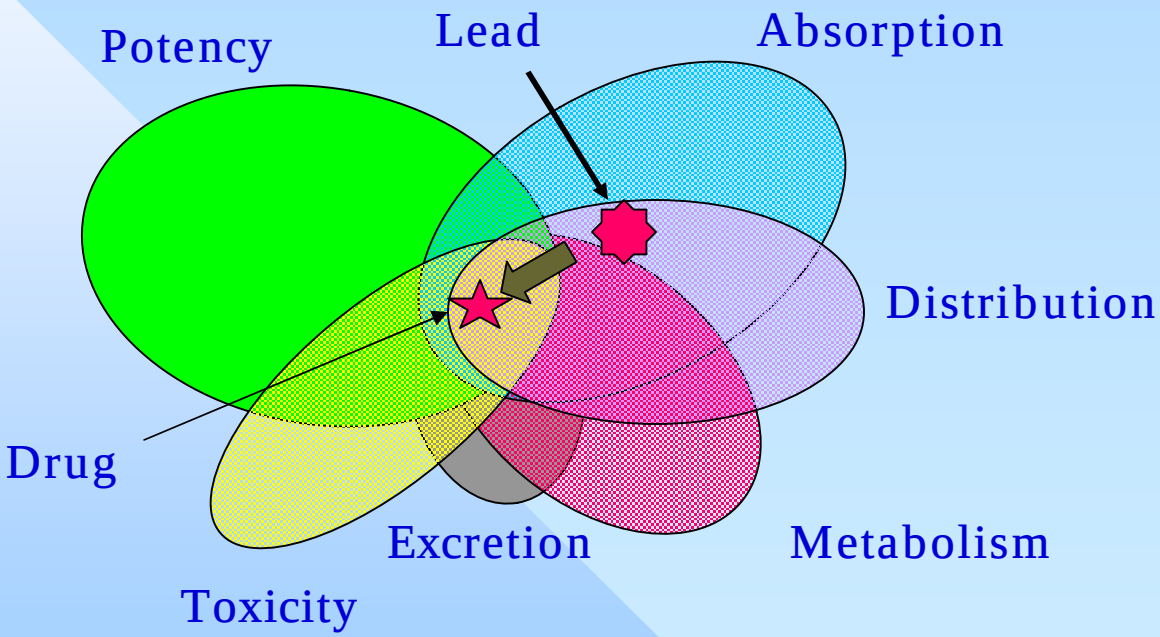


- Other parameters: patent position, chemical synthesis
- The greatest hurdle : ADMET properties.

Traditional Drug Discovery Scheme



In silico prediction of ADME properties



Electron Density Derived descriptors

- TAE (Transferable Atom Equivalent)
- WCD (Wavelet Coefficient Descriptors)
- PEST (Property-Encoded Surface Translator) Hybrid Shape-Property Descriptors
- Autocorrelation of Molecular Surface Properties

Molecular Surface Properties

Electronic Properties

- Electrostatic Potential

$$EP(r) = \sum_a \frac{Z_a}{|r - R_a|} - \int \frac{r(r')dr'}{|r - r'|}$$

- Electronic Kinetic Energy Density

$$K(r) = -(y^* \nabla^2 y + y \nabla^2 y^*)$$
$$G(r) = -\nabla y^* . \nabla y$$

- Electron Density Gradients

$\nabla\rho \bullet N$

- Laplacian of the Electron Density

$$L(r) = -\nabla^2 r(r) = K(r) - G(r)$$

- Local Average Ionization Potential

$$PIP(r) = \sum_i \frac{r_i(r)|e_i|}{r(r)}$$

- Bare Nuclear Potential (BNP)

- Fukui function

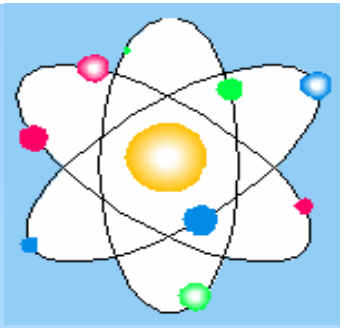
$F^+(r) = \rho_{HOMO}(r)$

The *TAE* Reconstruction Method

Transferable Atom Equivalents

- RECON algorithm for rapid reconstruction of molecular charge densities and molecular electronic properties
- based on Bader's quantum theory of Atoms In Molecules
- A library of atomic charge density fragments -- corresponding to structurally distinct atom types -- has been built in a form that allows for the rapid retrieval of the fragments and molecular assembly
- Associated with each atomic charge density fragment in the library is a data file which contains
  - information describing topological features of the atomic charge density used to orient the fragments into their proper molecular space orientations
  - atomic charge density-based descriptors encoding electronic and structural information relevant to the chemistry of intermolecular interactions.

Theory of Atoms in Molecules



Definition of an Atom in a Molecule:

- An atom is the union of an attractor and its basin
- Each atom contains one (*and only one*) nucleus, which is the attractor of its electron density distribution $\rho(\mathbf{r})$
- Every atom is bounded by an atomic surface of zero flux
- Atoms defined in this way satisfy the virial theorem
- They have properties that are approximately additive and transferable from one molecule to another.

RECON

http://www.drugmining.com/

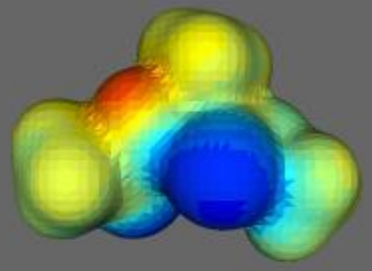
- Molecular input
  - PDB
  - Tripos MOL2
  - SIMS DRA
  - SMILES
  - SDF
- Determine atom types and assign closest match from atom type library according to the priority:
  - Perfect match
  - Ring size differs
  - Hybridization of nearest neighbor does not match
  - Atomic number of nearest neighbor does not match
  - Hybridization of atom does not match
  - For monovalent atom, hybridization of nearest neighbor differs
  - Atomic number of atom does not match
- Combine densities of atomic fragments
- Compute predicted molecular properties

RECON Timing performance

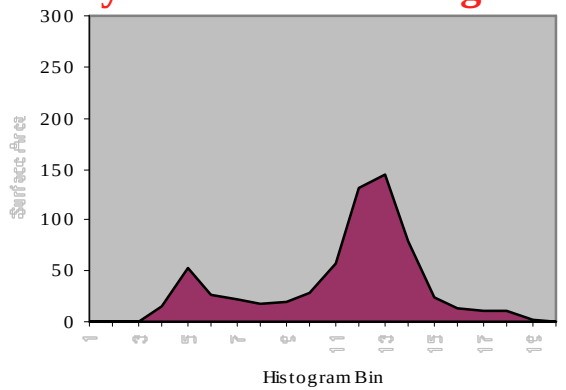
Hardware			SGI 300 MHz Octane		1.7 GHz Intel Pentium		
O/S			IRIX64		Windows 2000	Linux	
Test dataset	# of molecules	File Format	User CPU (sec.)	System CPU (sec.)	Total CPU (sec.)	User CPU (sec.)	System CPU (sec.)
Misc. small molecules	196	mixed	23.5	1.4			
	69				8.28		
MAO inhibitors	1650	SDF	102.7	44.5	650.7	15.3	3.5
	1641	SMILES	122.3	45.9	757.2	61.3	3.6
Proteins	25	PDB	186.8	194.5		65.1	17.6
NCI	42,689	SDF	2327.2	1131.0		391.0	67.5

TAE Electrostatic Surface and Histogram profile

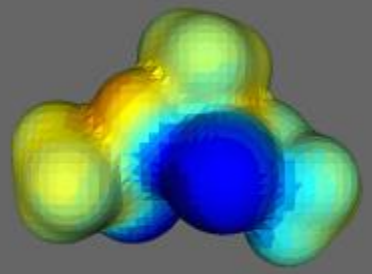
Analytical EP-encoded Surface



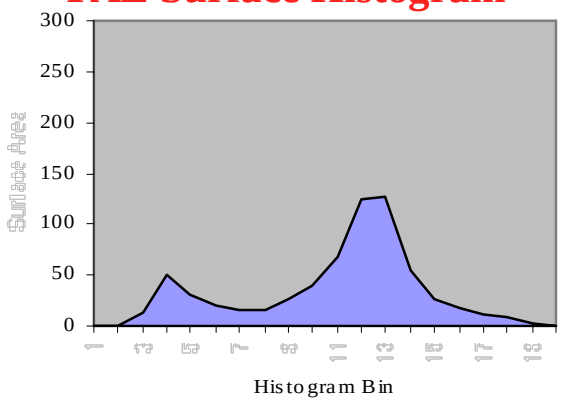
Analytical Surface Histogram



TAE EP-encoded Surface



TAE Surface Histogram

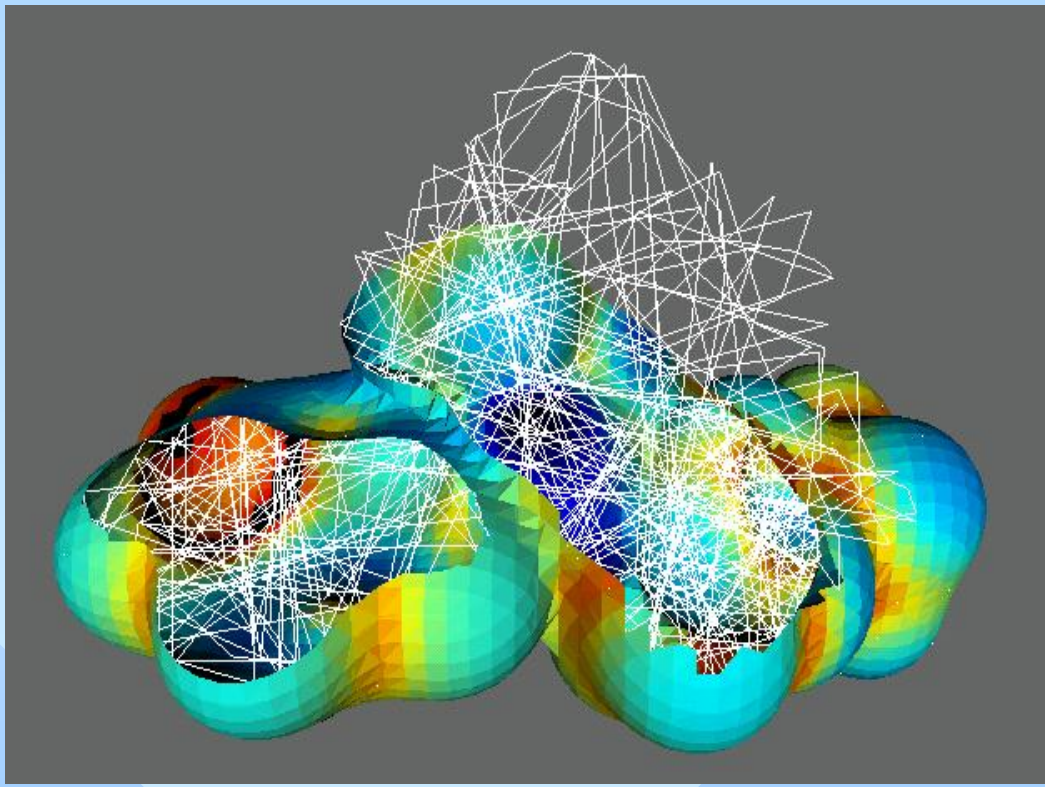


Christopher E. Whitehead, Curt M. Breneman, N. Sukumar and M. D. Ryan, "*Transferable Atom Equivalent Multi-Centered Multipole Expansion Method*" J. Comp. Chem. **24**, 512-529 (2003)

PEST Shape/Property Hybrid descriptors

- A TAE property-encoded surface is subjected to internal ray reflection analysis.
- A ray is initialized with a random location and direction within the molecular surface and reflected throughout inside the electron density isosurface until the molecular surface is adequately sampled.
- Molecular shape information is obtained by recording the ray-path information, including segment lengths, reflection angles and property values at each point of incidence.

Isosurface (portion removed) with 750 segments

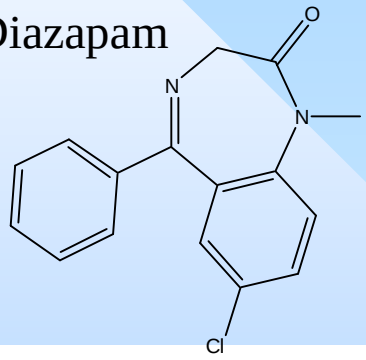


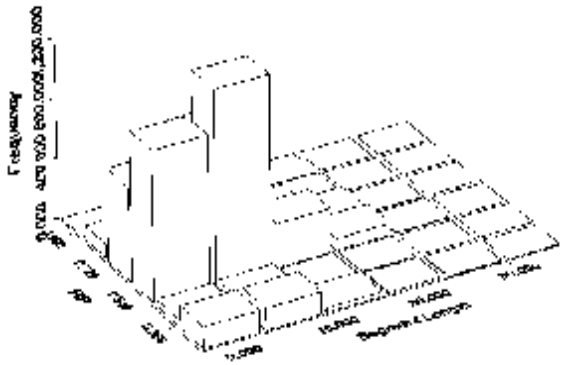
- Karthigeyan Nagarajan, Randy Zauhar, and William J. Welsh**, "***Enrichment of Ligands for the Serotonin Receptor Using the Shape Signatures Approach***"** *J. Chem. Inf. Model.***, 45**, 49-57 (2005)
- Curt M. Breneman, C. Matthew Sundling, N. Sukumar, Lingling Shen, William P. Katt and Mark J. Embrechts, "*New developments in PEST shape/property hybrid descriptors*" *J. Computer-Aided Mol. Design*, **17,** 231–240, (2003)

PEST Shape/Property Hybrid descriptors

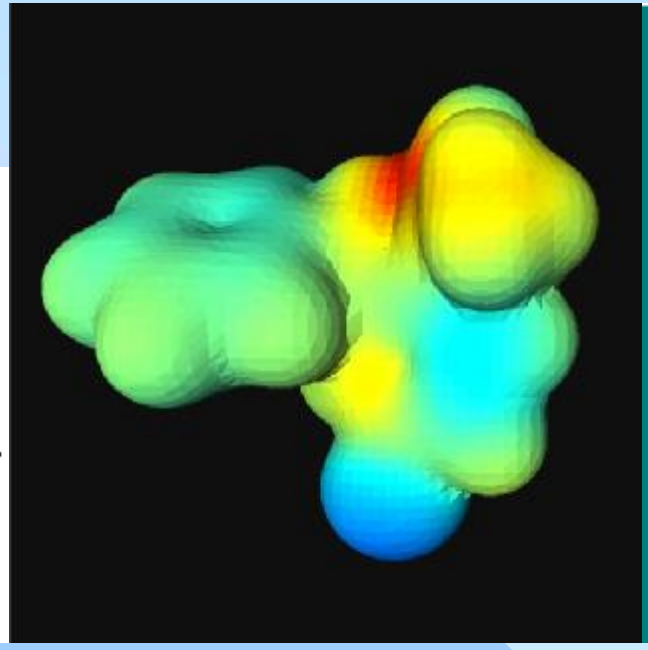
- Segment length and point-of-incidence value form 2D-histogram
- Each bin of 2D-histogram becomes a hybrid descriptor

Diazapam

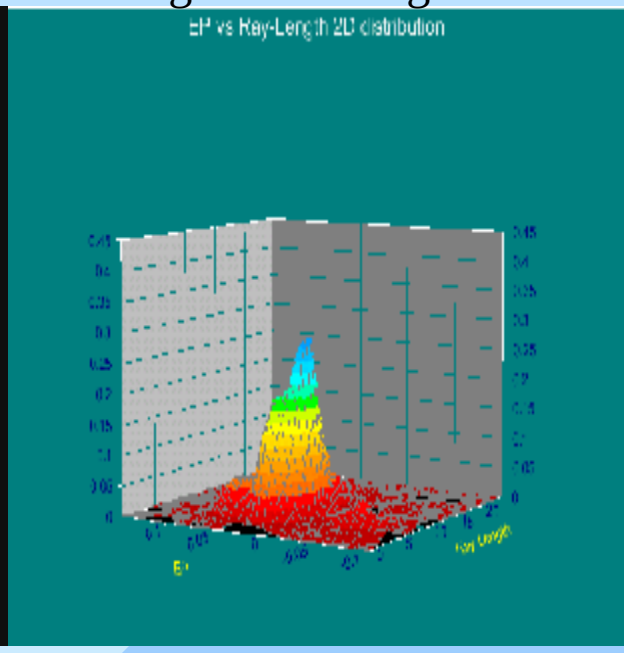




EP Surface

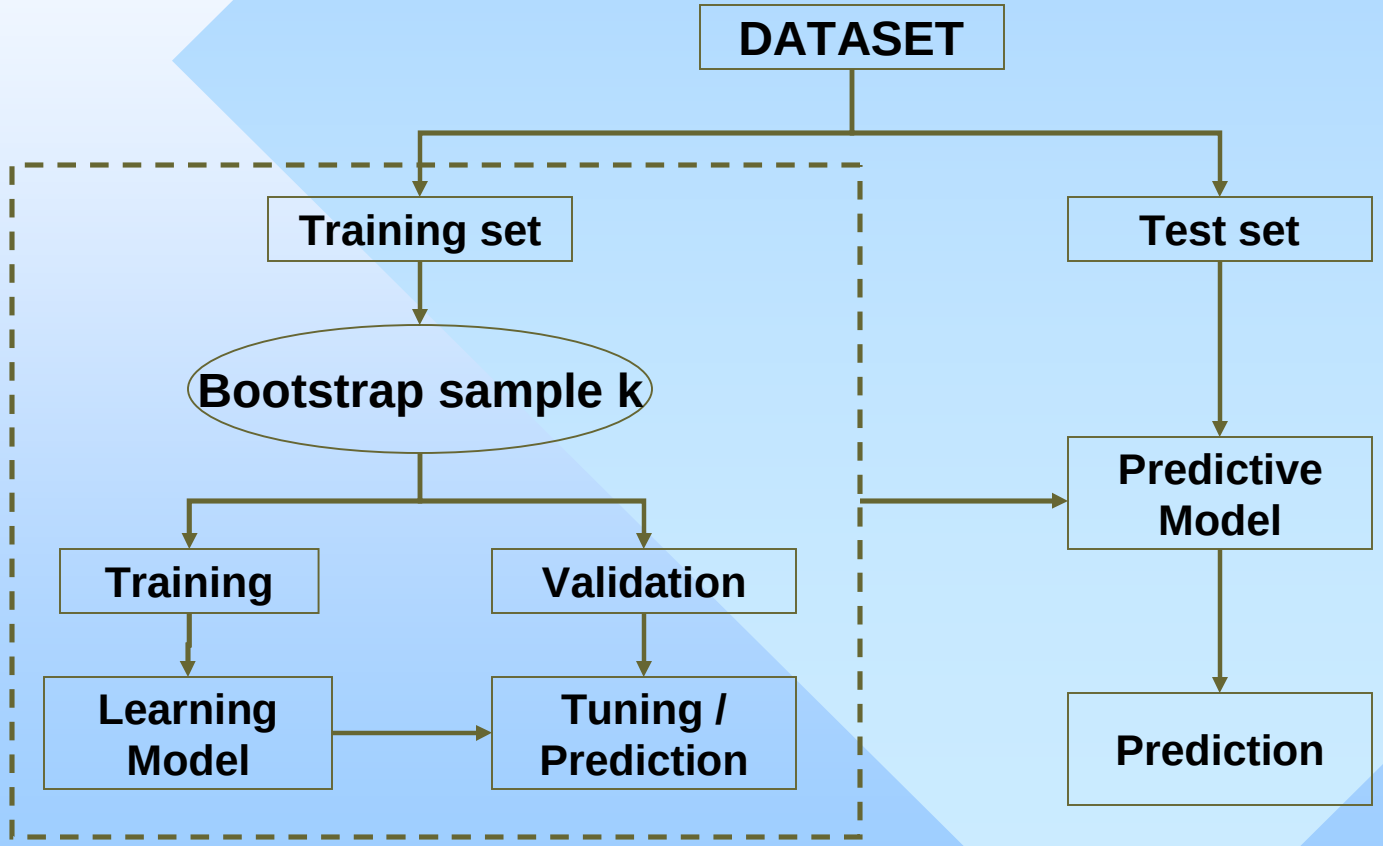


EP vs Length 2D histogram



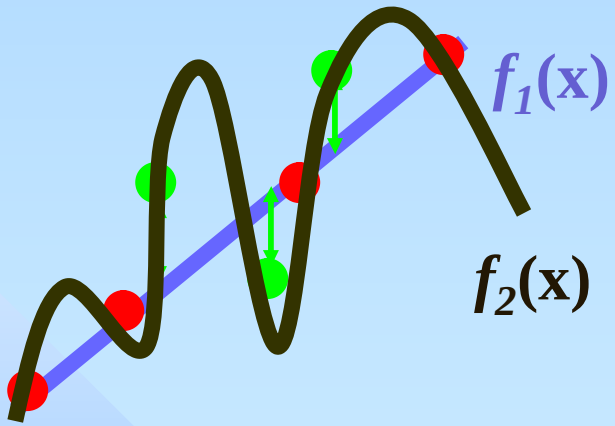
C. Matthew L. Sundling (COMP 345) *PEST vs. CoMFA: A comparative study of two 3D-QSAR technologies (4:00 PM, Thursday, March 17 -- Room 7B)*

General Modeling protocol



Traditional Regression approaches

- Minimize the training errors $\sum(y_i - f(x_i))$:



- Overfitting problem (HTS data with noise):
  - ø "perfect" prediction for training set
  - ø Poor prediction for unknown data

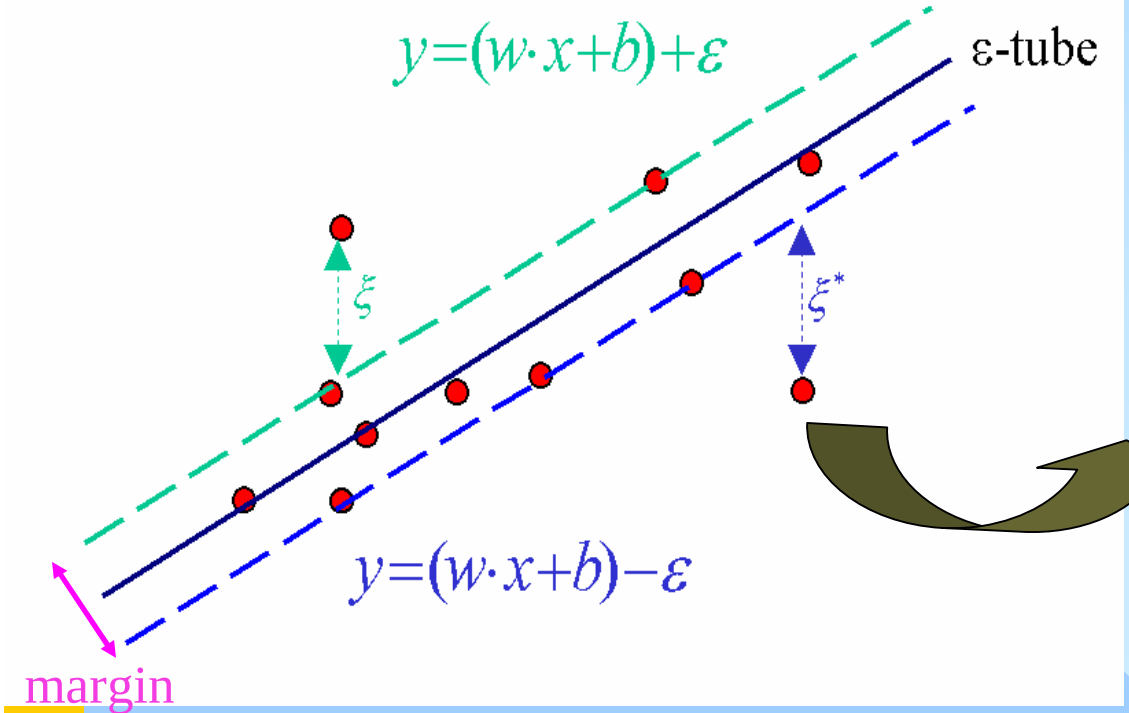
Support Vector Regression

Minimize the Generalization error:

- **ü** training error + model complexity

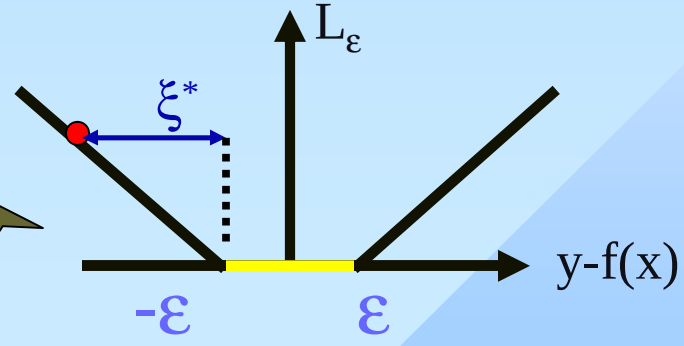
$$\min_{w,b,x_i,x_i^*} C\sum_{i=1}^{l}(x_i + x_i^*) + \frac{1}{2}\| w \|^2$$

- Parameter *C* controls tradeoff between error and capacity
- Minimizing ||*w*|| controls capacity of linear function



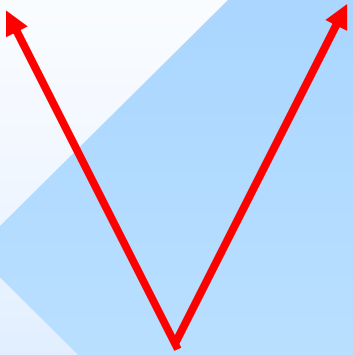
ε-insensitive loss function:

$$L_e(y - f(x)) \coloneqq \max(0, |y - f(x)| - e)$$

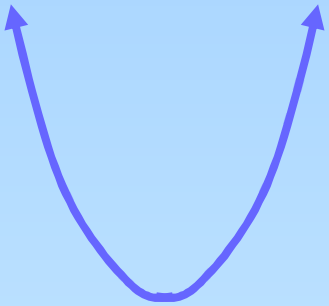


Avoid overfitting by controlling the model complexity

Variable Selection using SVM



1-norm



2-norm

The basic idea is very simple:

- Construct a series of sparse linear SVM that exhibit good generalization.
- Construct the subset of variables having nonzero weights in the linear models.
- Then use this subset of variables in nonlinear SVM to produce the final regression or classification function.

The method exploits the fact that linear SVM with 1-norm regularization inherently performs variable selection as a side-effect of minimizing capacity in the SVM model.

Nonlinear Support Vector Machine

- Construct linear function in higher-dimensional space.
- Mapping done using Kernel functions.
- Typically use RBF Kernel $K(u, x_i) = exp\left(\frac{\| u - x_i \|^2}{s}\right)$
- Regression function is

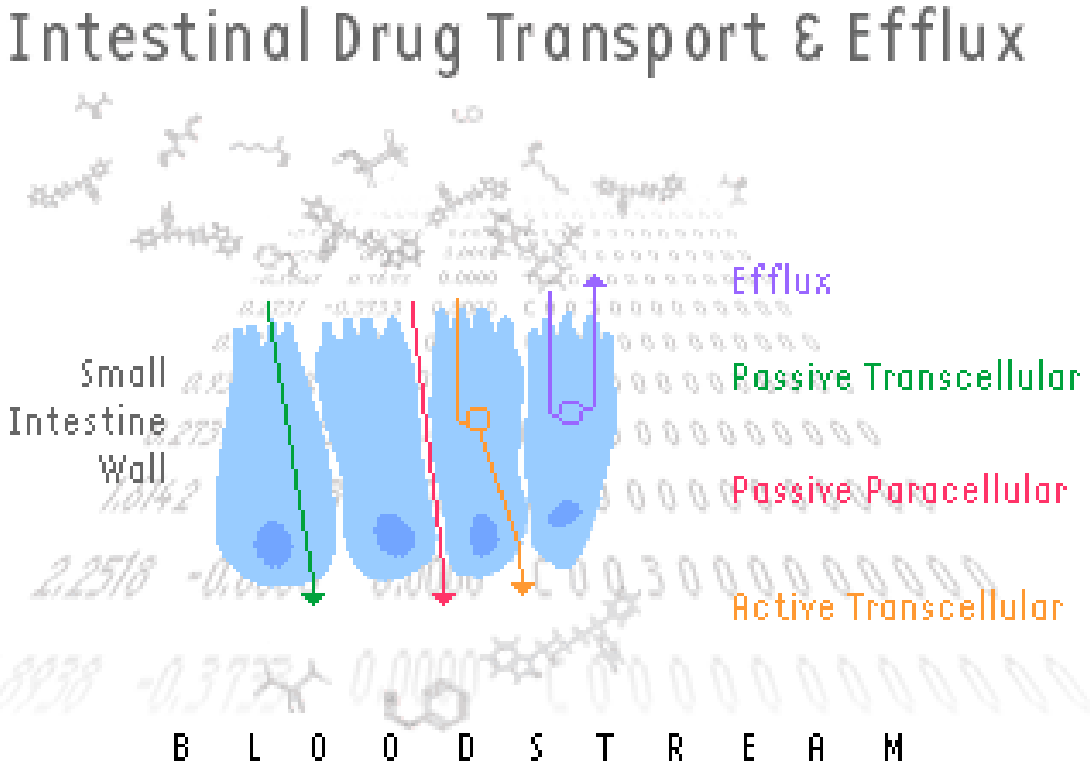
$$f(u) = \sum_{i=1}^{1} \left((a_i - a^*_i) K(u, x_i) \right) - b$$

Dechuan Zhuang* (COMP 337) *Consensus feature selection for multi-objective SVM modeling of protein ion-exchange displacement chromatography (1:20 PM, Thursday, March 17 -- Room 7B)

Human Intestinal Absorption

- For most drugs, the preferred route of administration is by oral ingestion
- Oral drugs are absorbed across a series of separate barriers
- Diffusion mechanisms:
  - Passive diffusion
  - Active diffusion

Drugs Diffusion Mechanisms



Different barriers

Drugs

Mucus Gel Layer

Intestinal Epithelial Cells

Lamina Propria

Endothelium of Capillarics

Be absorbed

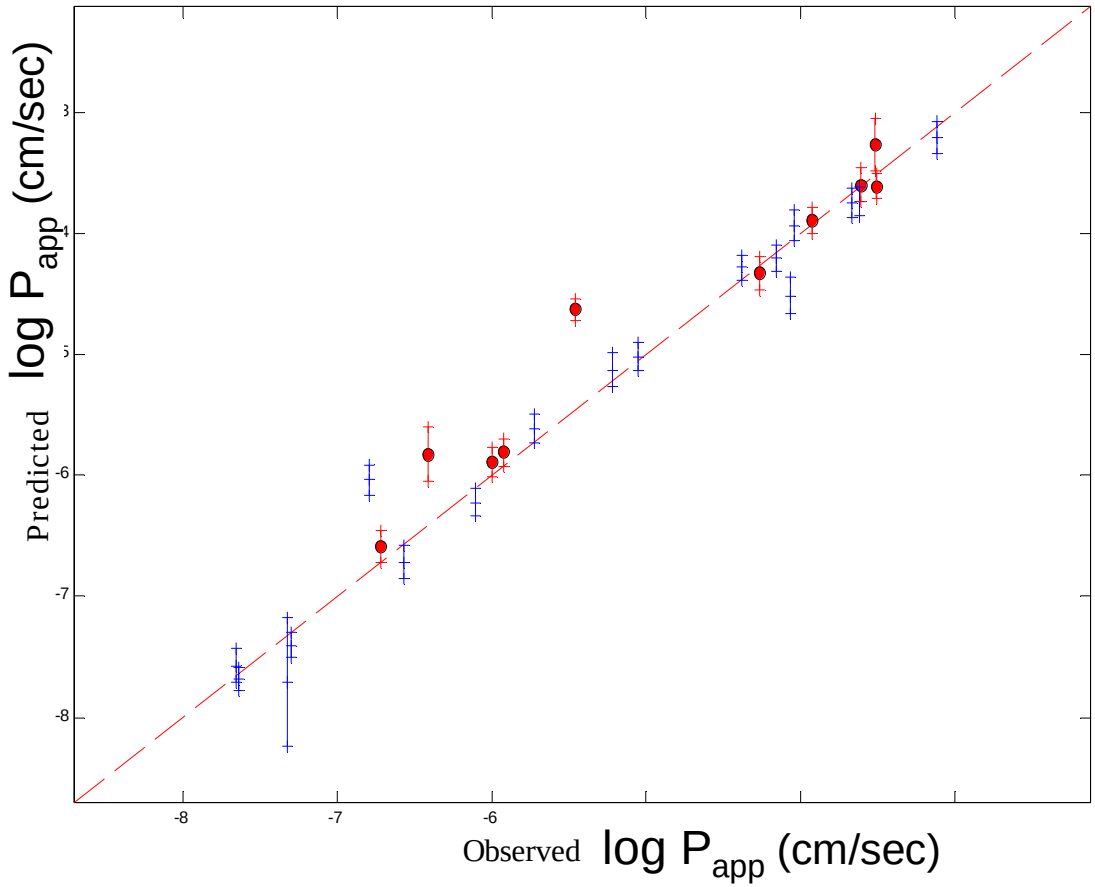
A series of separate barriers (epithelial layer is the most dominant barrier)

What is Caco-2 cell line?

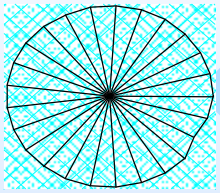
- Derived from a human colorectal carcinoma
- With remarkable morphological and biochemical similarity to the small intestinal columnar epithelium
- The apparent permeability coefficient (Papp) have been shown to correlate well with %HIA or %FA
- Extremely useful for mechanistic studies of drug absorption and as an absorption screening assay for pre-clinical drug selection.
- The absorption of a drug compound through the human intestinal cell line is an important property for potential drug candidates.
- Caco-2 cell line is a widely used model for drug absorption.
- Accurately measuring this property, however, can be very costly and time-consuming.
- The use of QSPRs is an attractive alternative to experimental measurement.

Caco-2 Bagged SVM Model
(RBF Kernel, 15 features selected by SVM/LP)

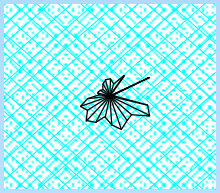
- Human intestinal cell line
- Predicts drug absorption
- 27 molecules with tested permeability
- 718 descriptors generated
  - Electronic TAE
  - Shape/Property (PEST)
  - MOE



Caco-2 – 14 Features



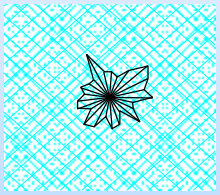
a.don



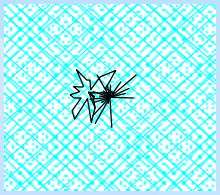
DRNB10

PEOE.VSA.FNEG

BNPB31



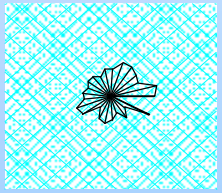
KB54



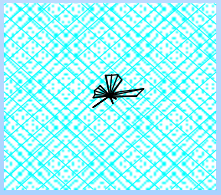
ABSDRN6

ABSKMIN

FUKB14

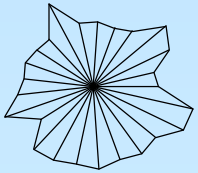


SMR.VSA2

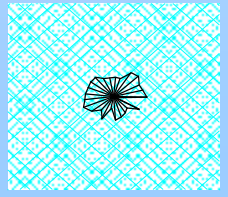


PEOE.VSA.FPPOS

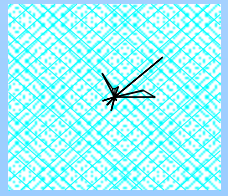
SIKIA



SlogP.VSA0



ANGLEB45



DRNB00

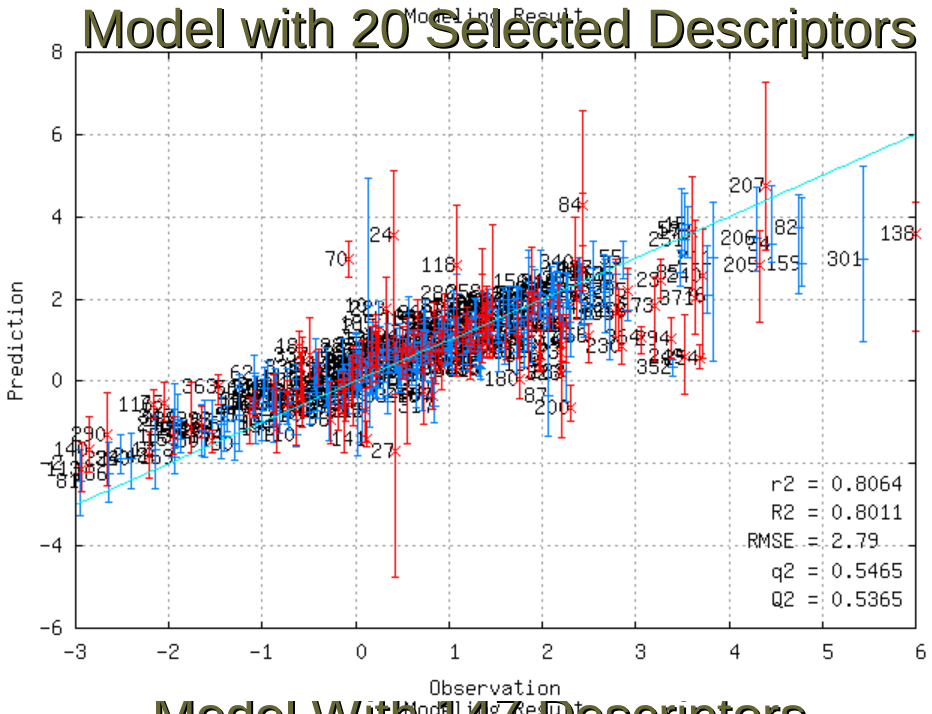
- Each star represents a descriptor
- Each ray is a separate bootstrap
- The area of a star represents the relative importance of that descriptor
- Descriptors shaded cyan have a negative effect
- Unshaded ones have a positive effect

- **Hydrophobicity** - a.don
- **Size and Shape** - ABSDRN6, SMR.VSA2, ANGLEB45 Large is bad. Flat is bad. Globular is good.
- **Polarity** – PEOE.VSA...: negative partial charge good.

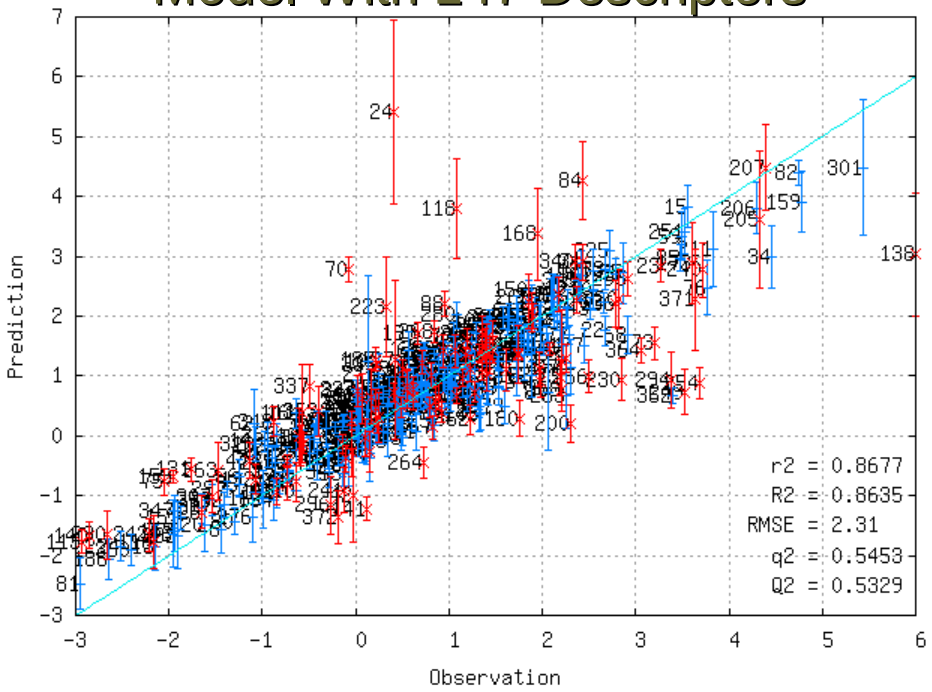
Modeling Fathead Minnow Toxicity

- 375 Diverse Organic Compounds
- Acute Toxicity in terms of $-\log LD_{50}$ (negative logarithm of mM concentration causing 50% fatality in fathead minnow population within 96 hours of exposure)
- 147 RECON 2D descriptors
- Randomly take out 125 molecules as the external test set
- Build Support Vector Regression models with all descriptors and with 20 selected descriptors
- Check model significance using Y-scrambling
  - Scramble the response values of the training set: the correlation coefficient measures the "messiness"
  - Build a model from the scrambled training data
  - Predict the external test set using the model
  - Repeat 50 times
- Compare the performance of the real model and the scrambled models

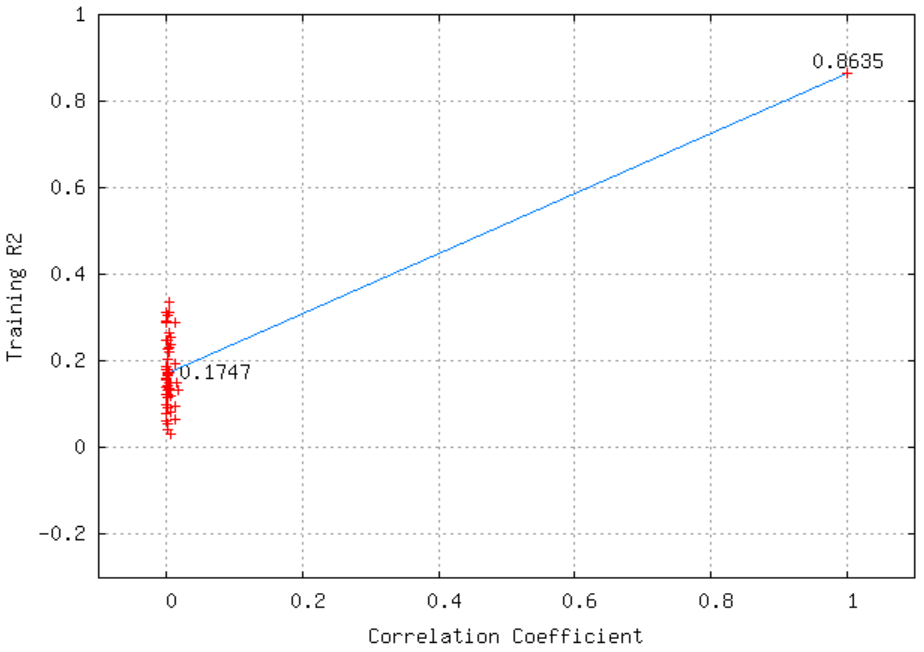
Model with 20 Selected Descriptors

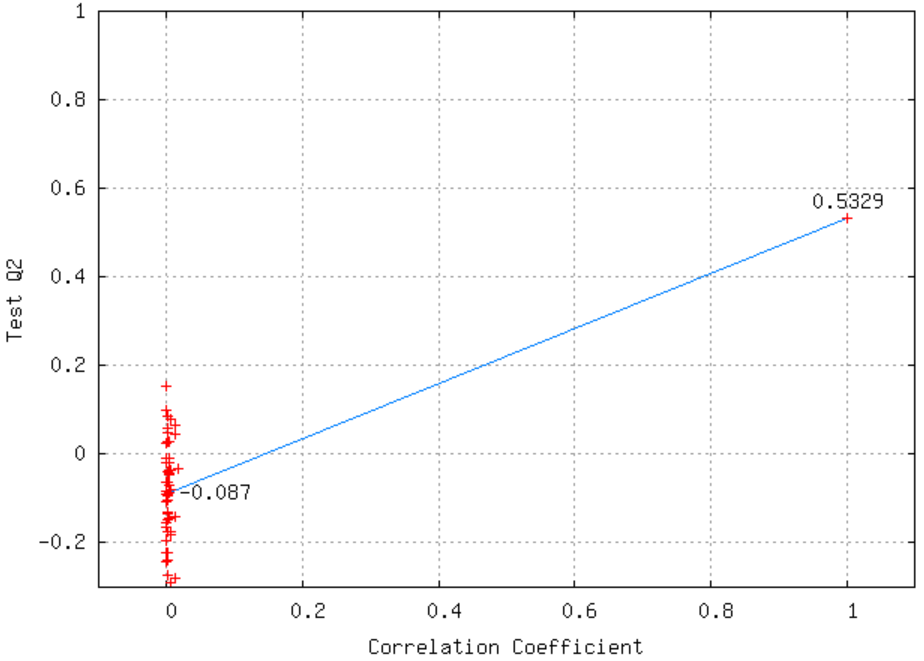


Model With 147 Descriptors

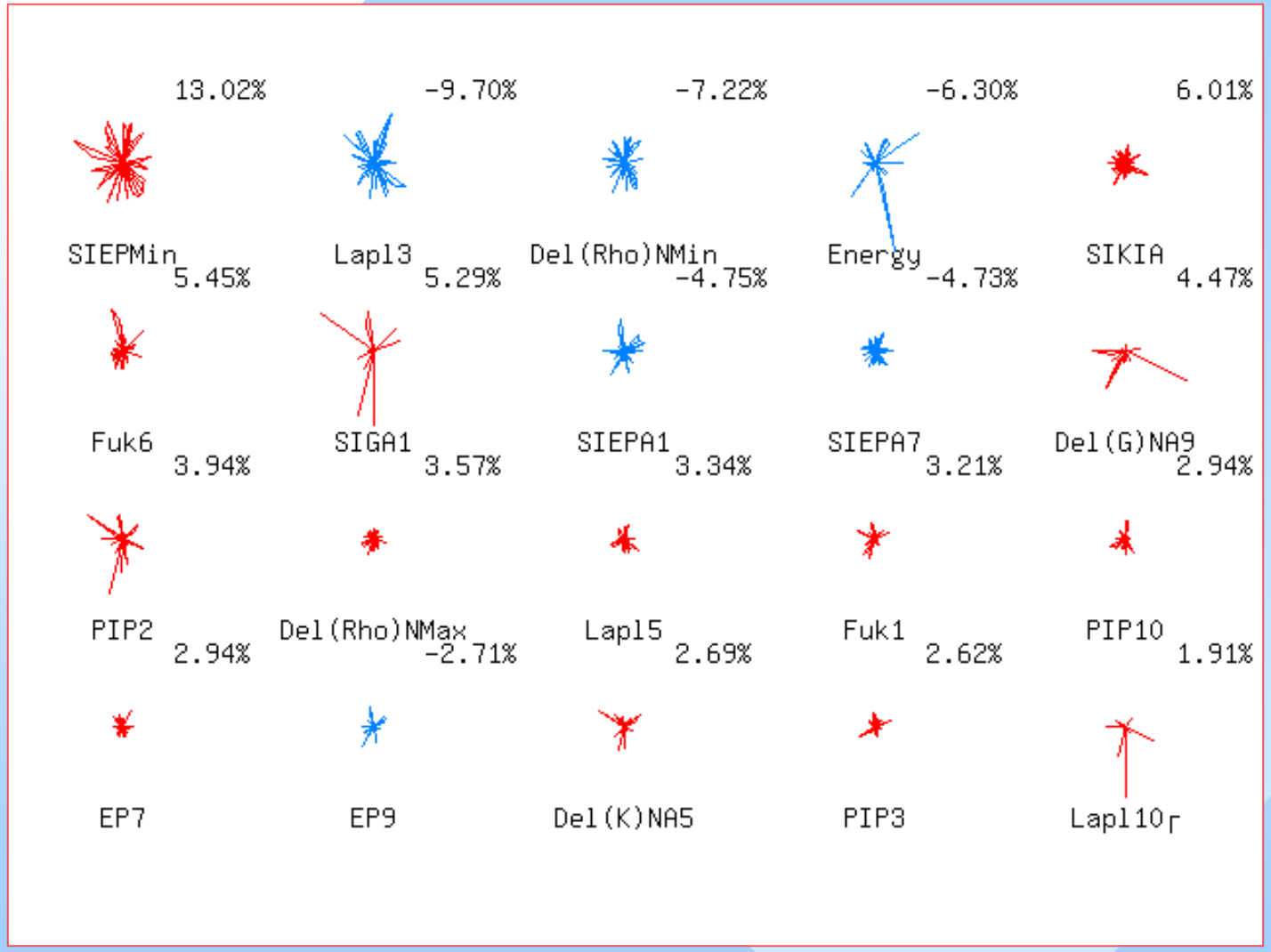


Models with Y-Scrambling



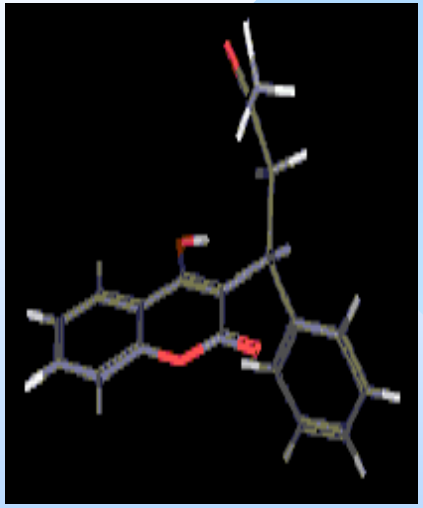


Starplots of 20 Selected Descriptors

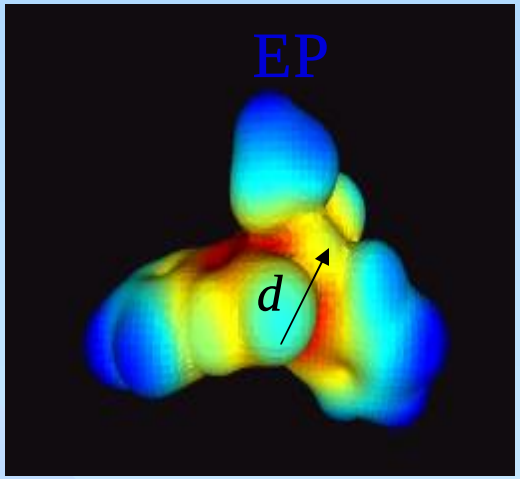


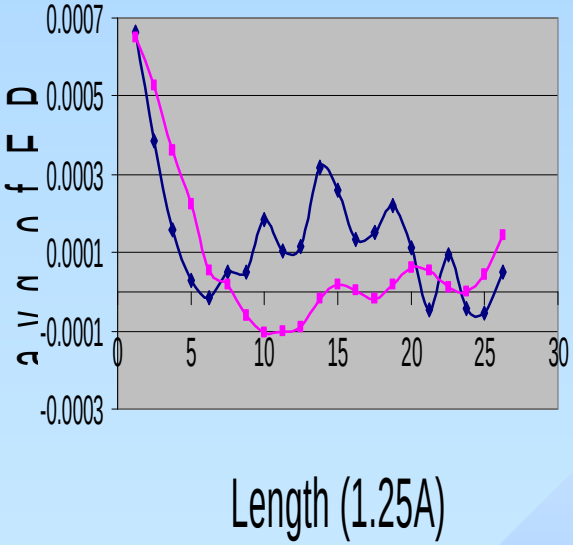
- **Solubility, surface area important**
- **Del(G)N, Del(Rho)N describe polarizability/hydrophobicity**
- **EP correlated with solvation behavior**

Autocorrelation of Molecular Surface Properties



Warfarin





$p(x)$ property at point x

N # of products in distance interval

d distance interval

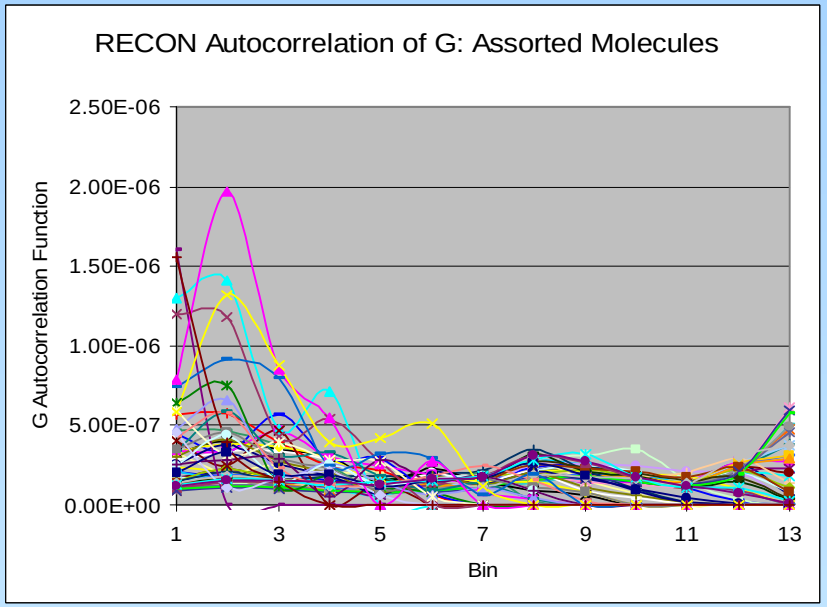
M. Wagener, J. Sadowski, J. Gasteiger, *J. Am. Chem. Soc*. **1995**, *117*, 7769

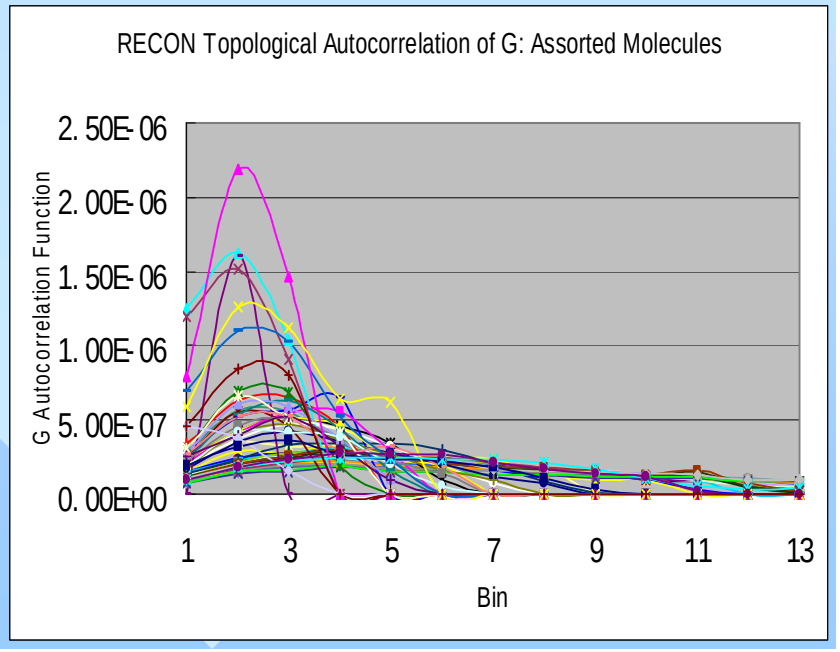
RECON Autocorrelation Descriptors

- Derived from RECON binned histogram descriptors
- Based on Gasteiger's Autocorrelation Formula

$$A(R_{xy}) = 1/n \times P_y \sum_{x,y} P_x$$

- Binned by distance between atoms x and y (RAD) or by topological distance between these atoms (TRAD)
- Reflect a 3-dimensional information content but can be calculated almost as rapidly as 2-dimensional descriptors
- Calculated for each type of standard RECON descriptor
- Higher bin number => greater distance between atom





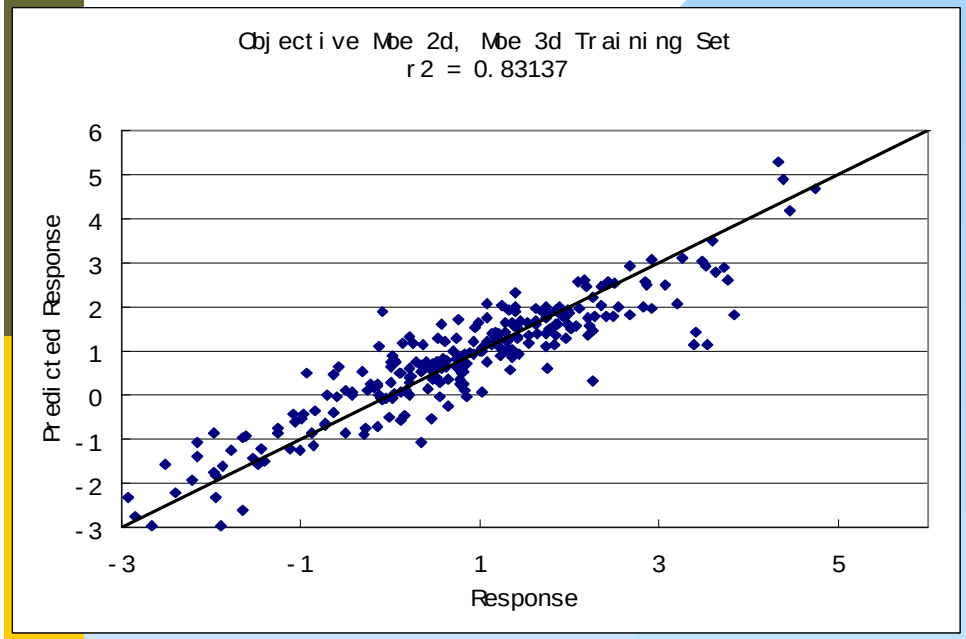
Modeling Fathead Minnow Toxicity

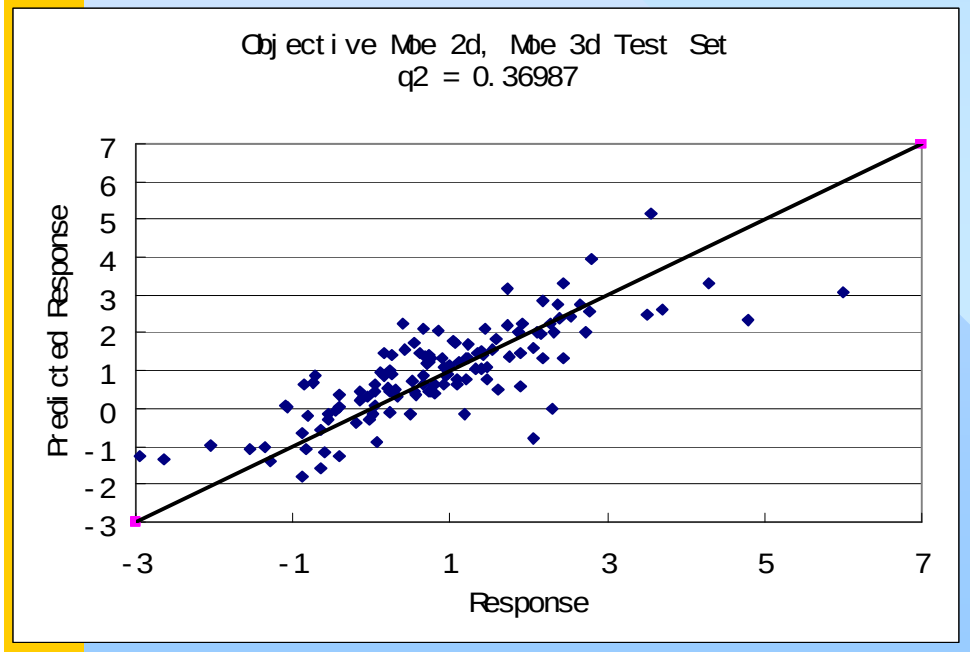
Kernel PLS with			
Subjective Feature Selection			
Descriptors	r2 (training)	q2 (test)	# Features
RECON	0.821	0.611	100
Moe2d, RECON	0.851	0.61	100
RECON, RAD	0.836	0.613	100
RECON, TRAD	0.839	0.632	100
RAD, TRAD	0.766	0.541	100
Moe2d, Moe3d, RECON	0.859	0.625	100
Moe2d, RECON, TRAD	0.864	0.63	100
RECON, RAD, TRAD	0.857	0.587	100
All 5 sets	0.875	0.622	100

Training Set: 250 molecules

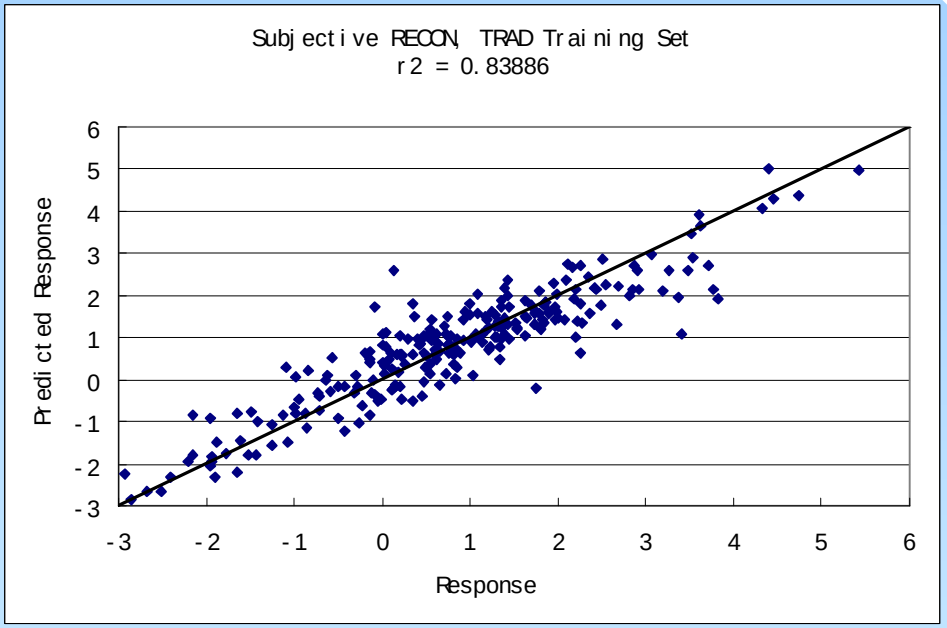
Test Set: 125 molecules

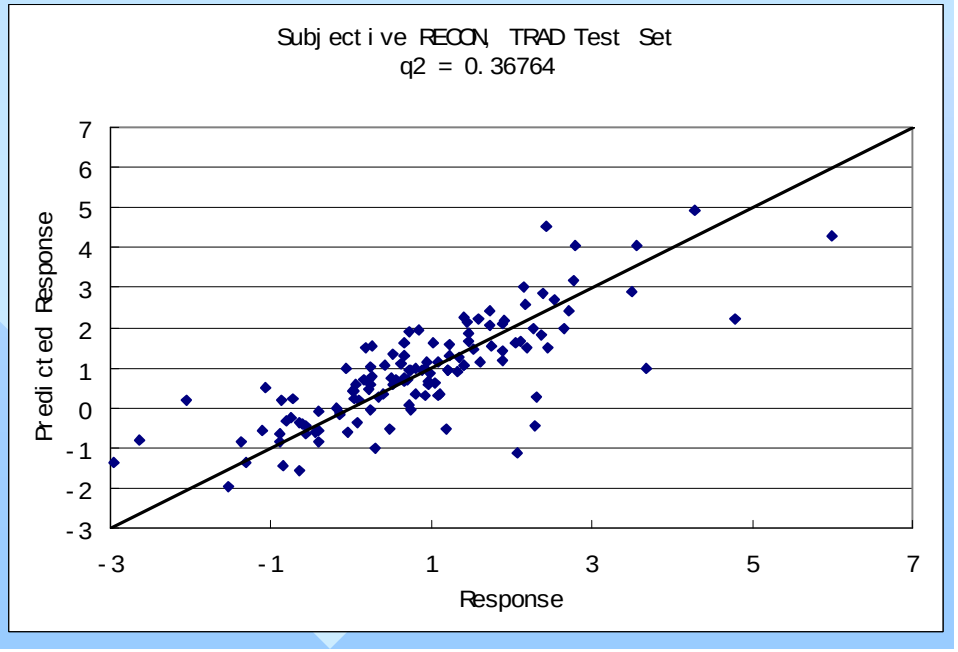
MOE 2-D & 3-D descriptors





RECON & TRAD descriptors

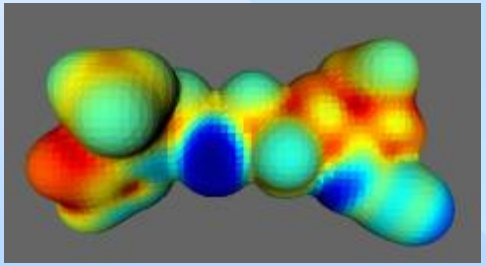




Electron Density Derived Surface Descriptors

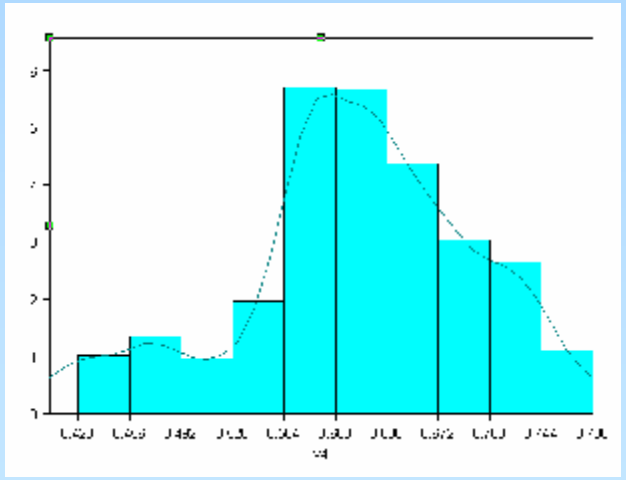
Histograms or wavelet encoding of surface properties gives TAE property descriptors

- Surface histograms can represent property distributions with 70-75% accuracy when 10-20 histogram bins are used

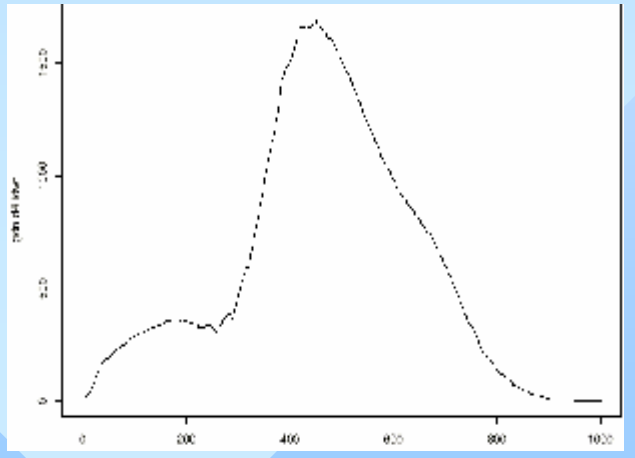


PIP (Local Ionization Potential)

Histograms



Wavelet Coefficients



- >95% accuracy with 16 wavelet coefficients

The $\rho(\mathbf{r}) = 0.002$ e/au^3 isosurface corresponds roughly to the molecular van der Waals surface.

Wavelet Decomposition

Photographic images are routinely compressed using Discrete Cosine Transforms: Joint Photographic Experts Group (JPEG)

Original image	High Resolution JPEG	Medium Resolution JPEG	Low Resolution JPEG
514 KB	**26 KB**	**11 KB**	**6.6 KB**

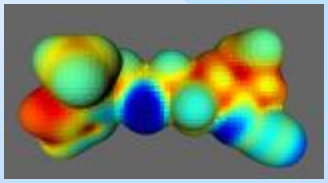
In this example, nearly 90-fold data compression is achieved by retaining only the highest amplitude coefficients.

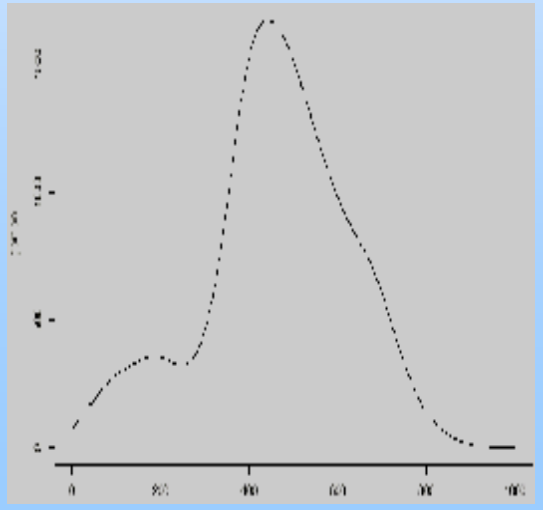
We use Discrete Wavelet Transforms (***DWT***) to encode quantum chemical information content

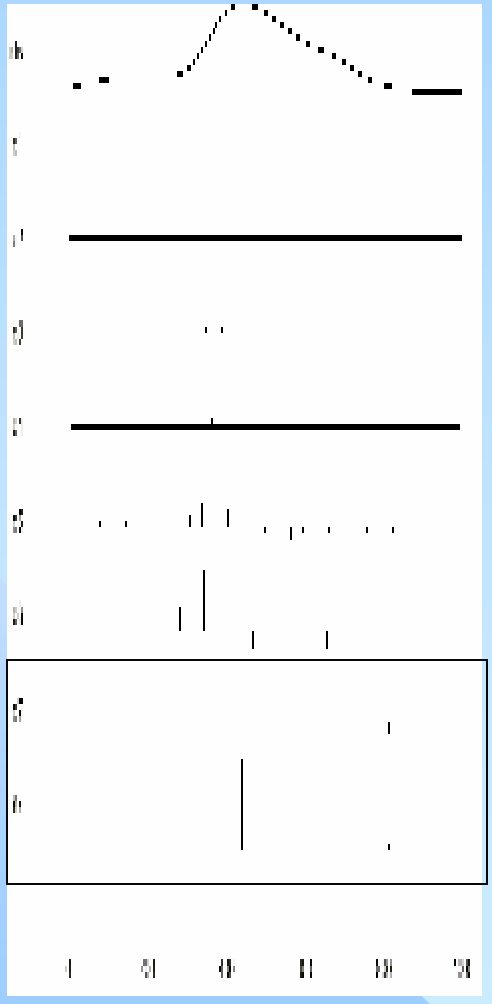
Electron Density-Derived Molecular Properties: Wavelet Coefficient Descriptors (*WCD*)

Wavelet Decomposition:

- Creates a set of coefficients that represent a waveform.
- Small coefficients may be omitted to compress data.



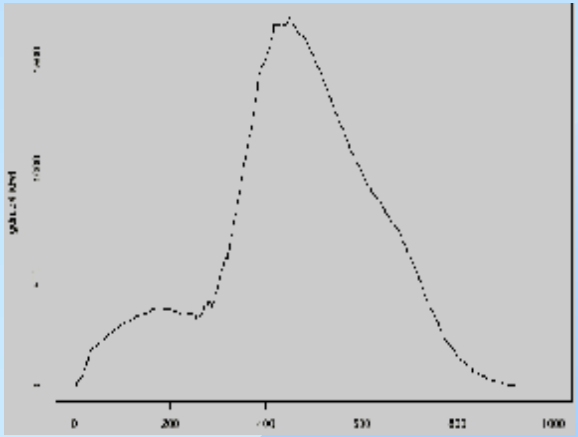




1024 raw wavelet coefficients capture PIP distribution on molecular surface.

Wavelet Surface Property Density Reconstruction:

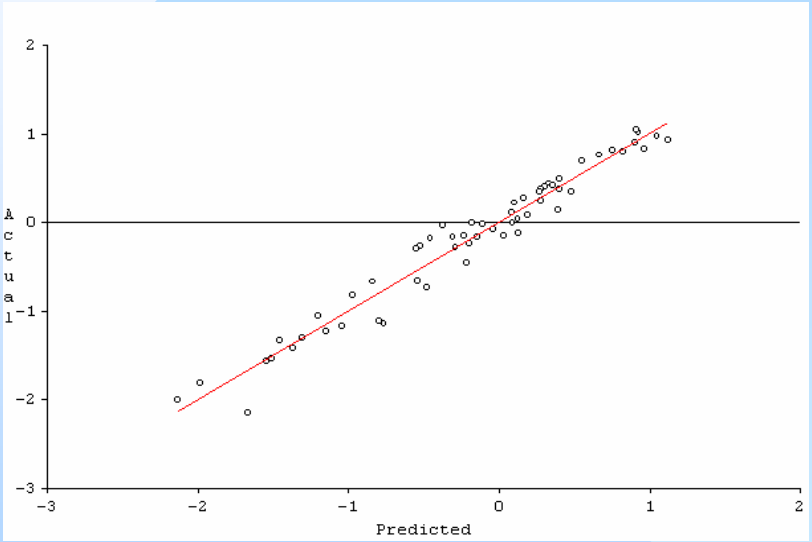
16 coefficients from S7 and D7 portions of the WCD vector represent surface property densities with >95% accuracy.



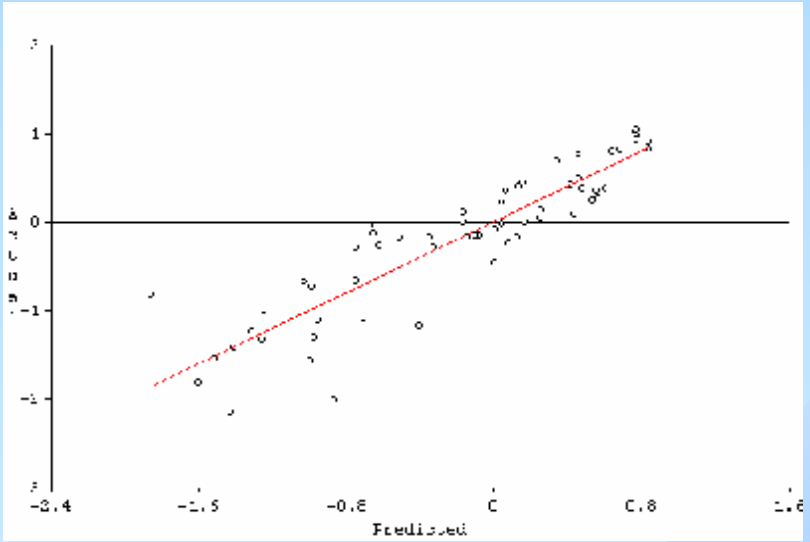
Increased property accuracy of descriptors provides more predictive models.

Blood-Brain Barrier Modeling
WCD vs Traditional Descriptors

16 TAE Wavelet Coefficient Descriptors

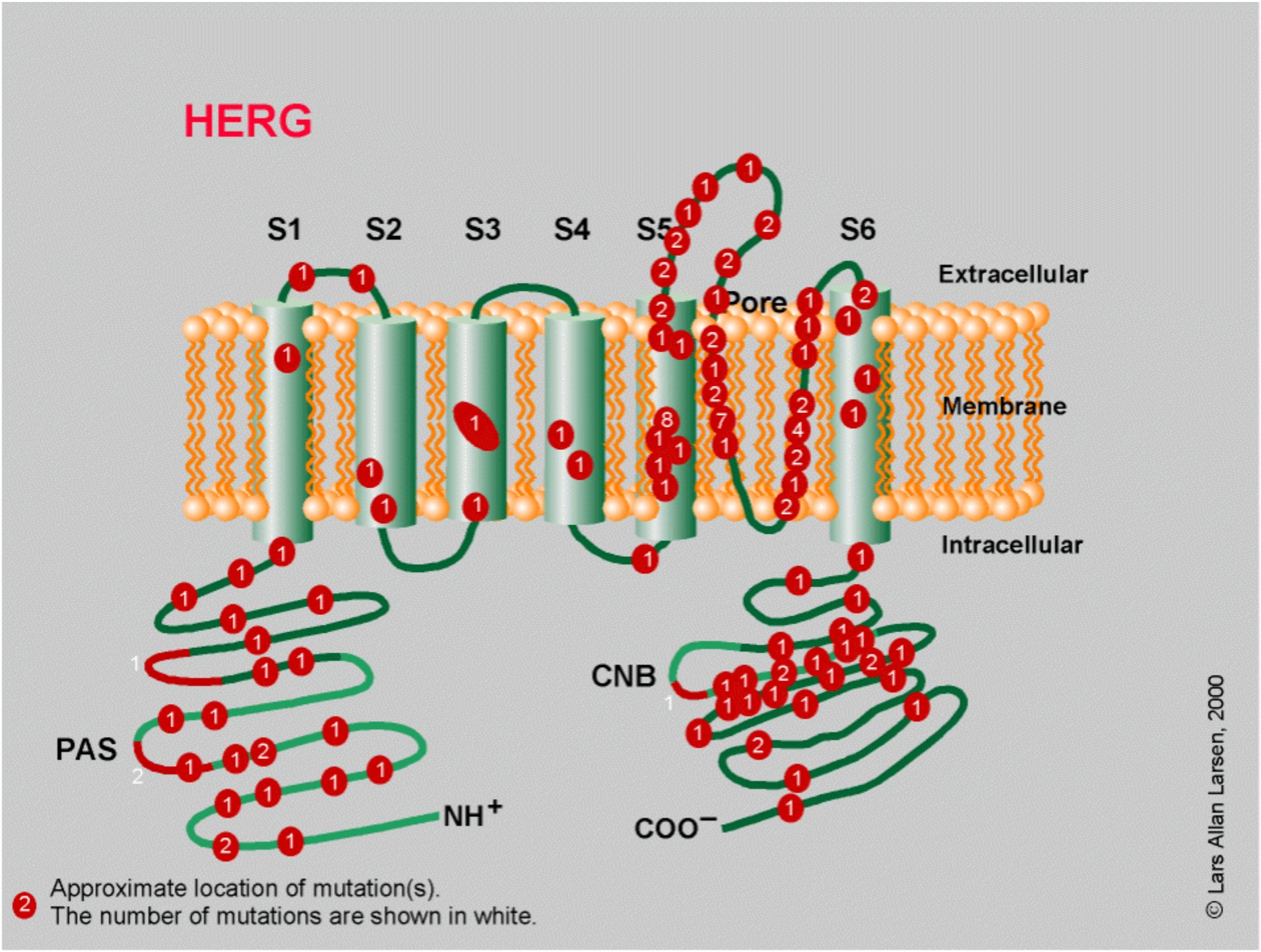


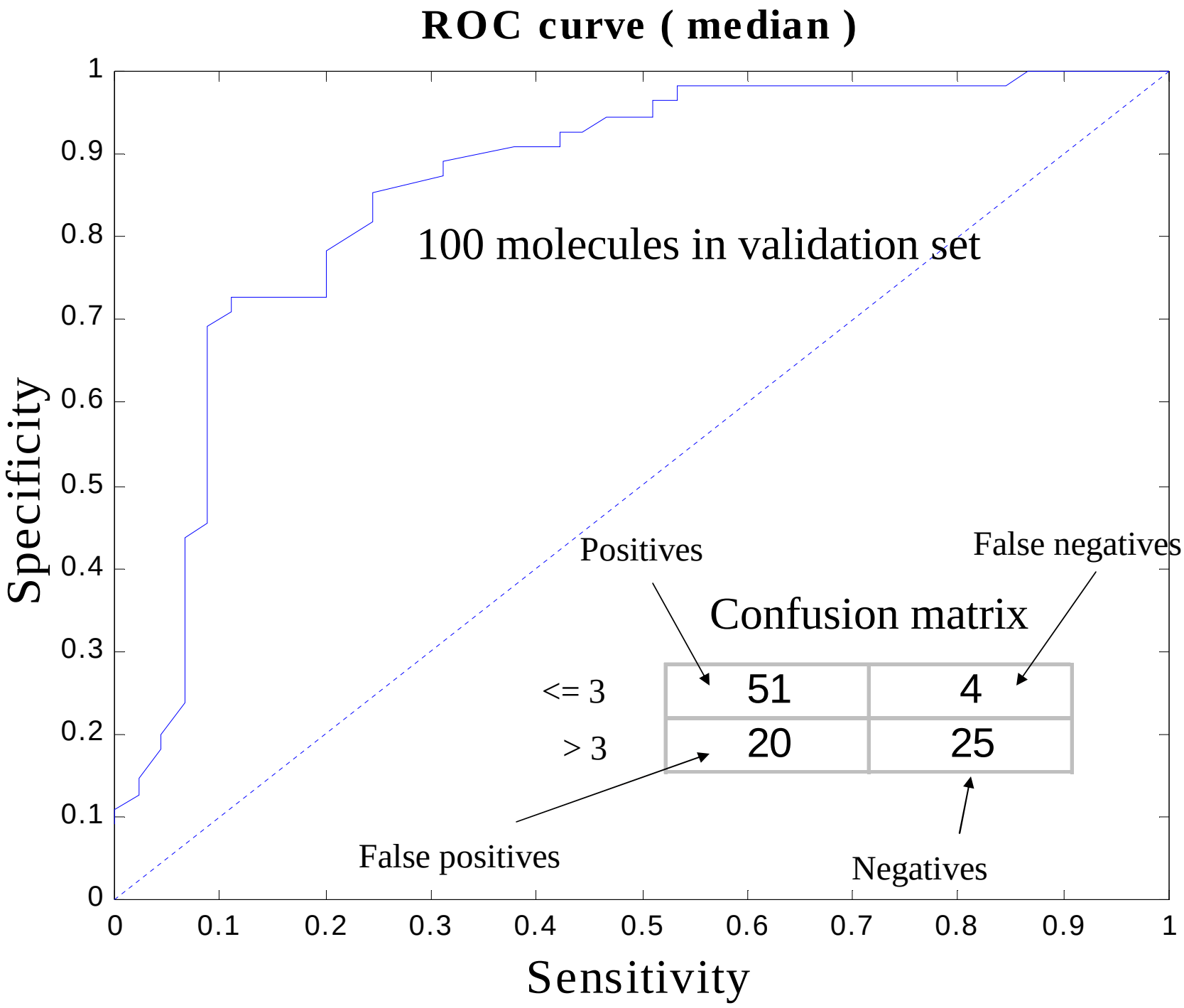
16 Traditional 2D & 3D Descriptors

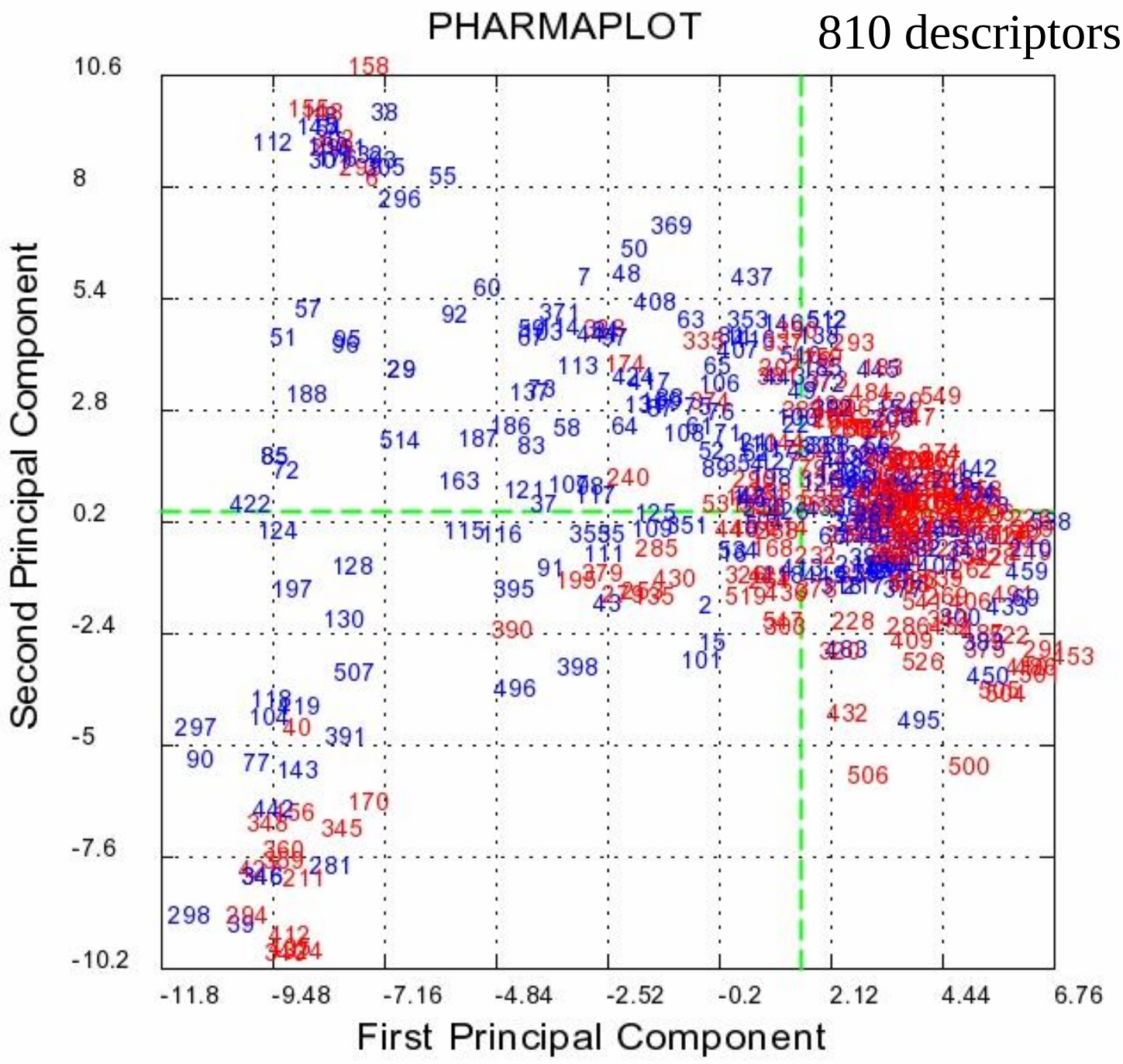


Both cases used GAPLS feature selection and three latent variables. Cross-validation was performed using bootstrapping techniques. All data values shown were predicted as unknowns using 80/20 training/test groupings.

Blood-Brain Barrier dataset: Lombardo, F; Blake, J.; Curatolo, W. *J. Med. Chem.* **39**, no. 24 (1996): 4750-4755


HERG
S1
S2
S3
S4
S5
S6
Pore
Extracellular
Membrane
Intracellular
PAS
CNB
NH+
COO−
Approximate location of mutation(s).
The number of mutations are shown in white.
© Lars Allan Larsen, 2000


ROC curve (median)
100 molecules in validation set
Specificity
Sensitivity
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Positives
False negatives
Confusion matrix
<= 3
> 3
51
4
20
25
False positives
Negatives


PHARMAPLOT
810 descriptors
Second Principal Component
First Principal Component
10.6
8
5.4
2.8
0.2
-2.4
-5
-7.6
-10.2
-11.8
-9.48
-7.16
-4.84
-2.52
-0.2
2.12
4.44
6.76

Acknowledgements

- Breneman Research Group :
  - Dechuan Zhuang
  - Matt Sundling
  - Bill Katt
  - Minghu Song
  - Lingling Shen
  - Bo Jiang
  - Larry Lockwood
- Embrechts Research Group
- Bennett Research Group

See: www.drugmining.com for more details