Computer Aided Drug Design

——Quantitative Structure-Activity Relationships (QSAR)

Qin Xu
http://cbb.sjtu.edu.cn/~qinxu/CADD.htm

Structure-Activity Relationships (SAR)

- Rule of 5 (drug like)
  - donor<5, acceptor>10,mass<500Da,LogP<5
- Rule of 3 (lead like)
  - donor<=3, acceptor>=3, mas<300Da, LopP>=3, rotational bond<=3, polar SA<=60Å

Molecular Parameters

log P

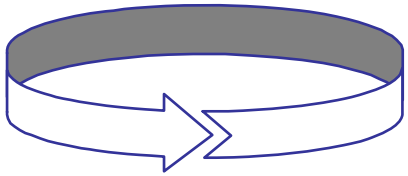
σ

MR

MLP

…

experiences



Biological activity

IC_{50}

K_i

MIC

Permeation

…

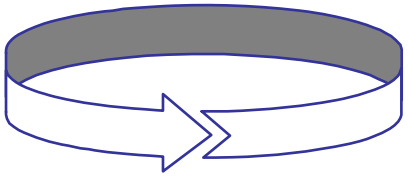
Quantitative Structure-Activity Relationships (QSAR)

- **Mathematical Models**

  Found a quantitative relationship

Molecular Parameters
- log P
- σ
- MR
- MLP
- …

Statistical tool



Biological activity
- IC_{50}
- K_i
- MIC
- Permeation
- …

Traditional methods - 2D(1964)

- parameters:
  - Lipophilicity (log P, π)
  - Electronics (σ ρ, Hammett correlation)
  - Sterics (E_s, molar refractivity)
- Statistical analysis
  - One parameter : Linear regression
  - Several Parameters : Multilinear regression
- Models of linear equations:

$$\text{Log } 1/C = k_1P_1 + k_2P_2 + k_3s + k_4Es + k_5$$

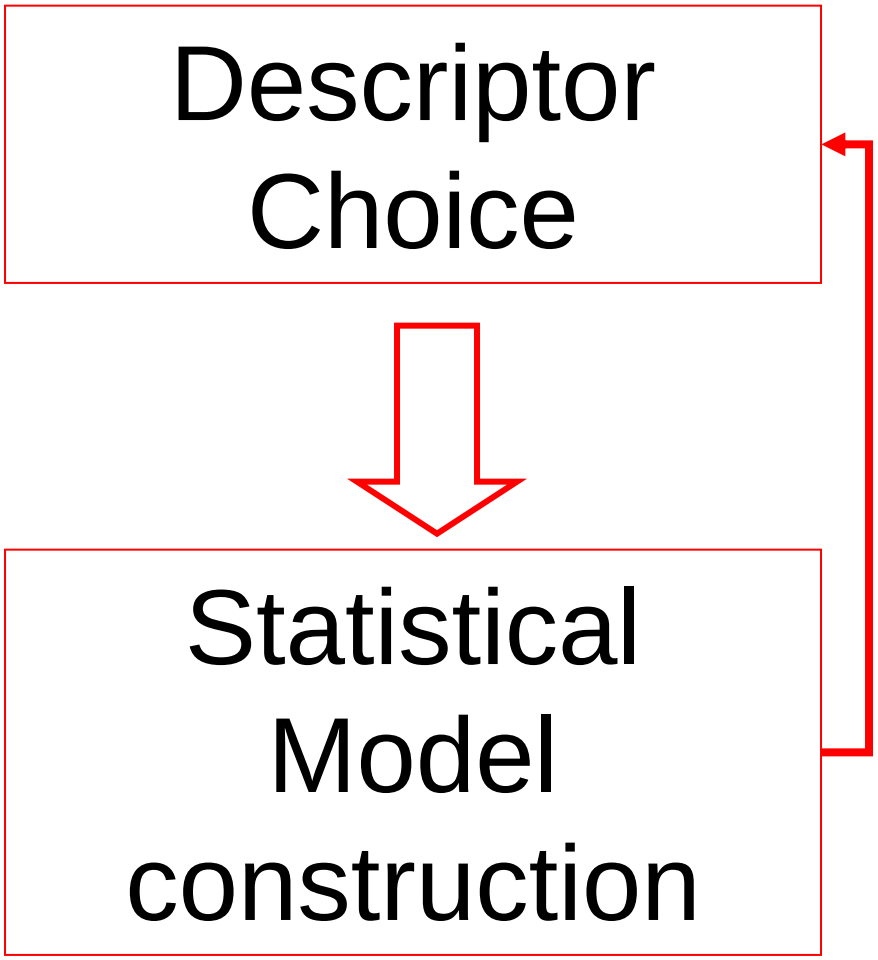
3D-QSAR(1988)

- Parameters
  - Molecular Field calculation (software)

    [f]Field contributions: Steric and electrostatic fields from CoMFA. Steric, Electrostatic, Hydrophobic, H-bond Donor and Acceptor fields
  - Interpretation of CoMFA results

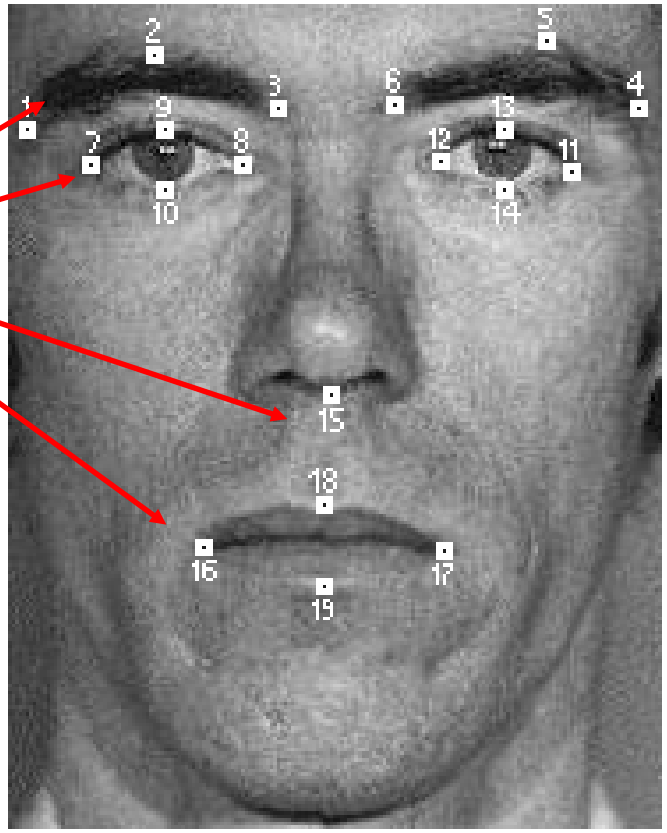
    Three dimensional structure
- Describe ligand-receptor intereactions
  - In vitro bioactivity

The key of QSAR

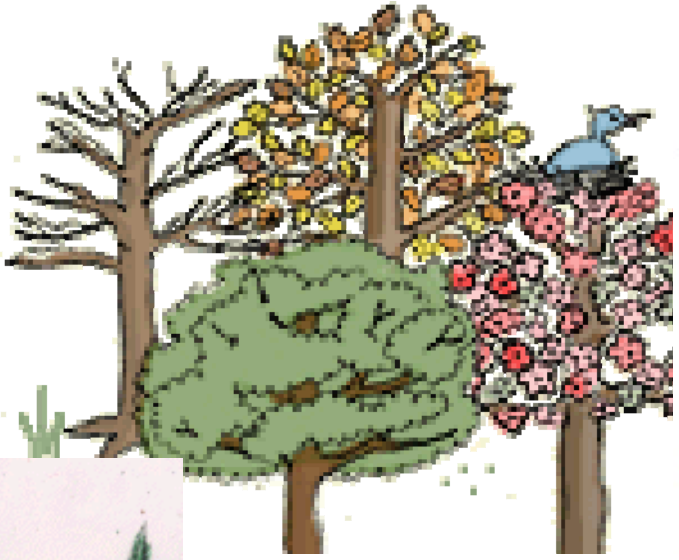


Feature Selection

- Face identification
- Key features.
- The same applies to molecules.



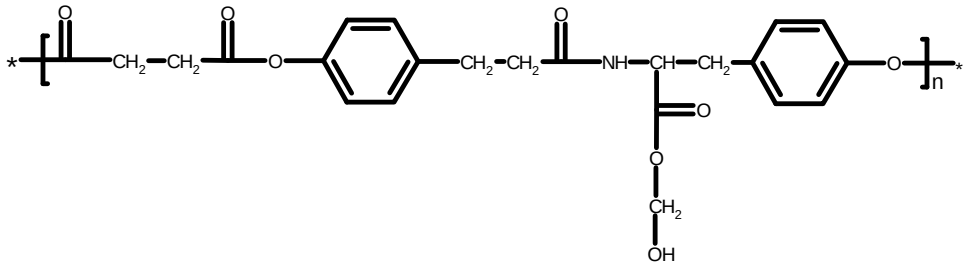
Descriptor Choice



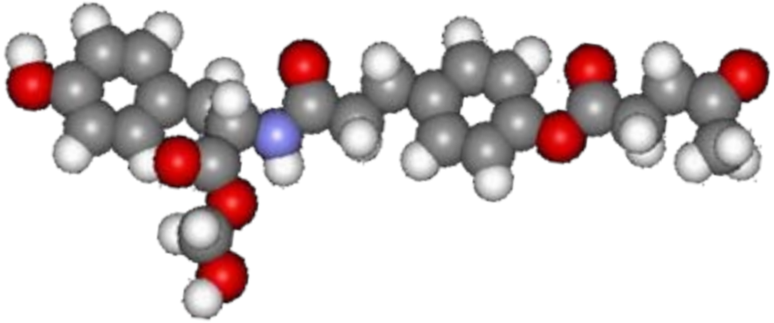
- 乔木、灌木
- 阔叶、针叶
- 落叶、常绿

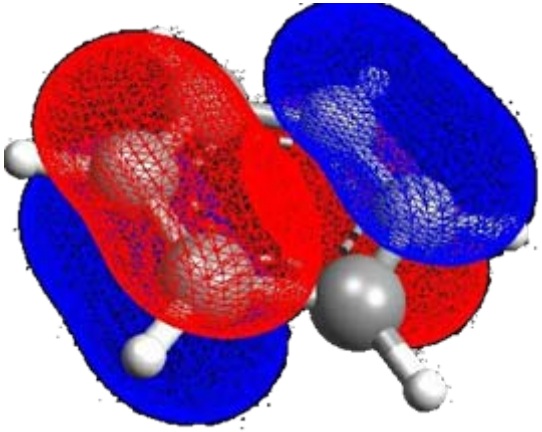
Types of descriptor for molecules

Topology



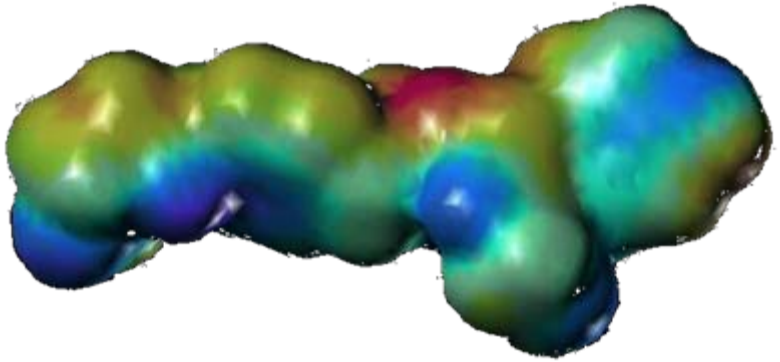
Geometry





Quantum

Electrostatic

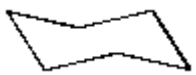


Fingerprint

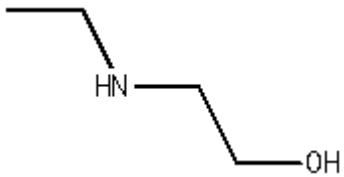
If a universe of features
U = { is-aromatic, has-ring, has-C, has-N, has-O, has-S, has-P, has-halogen }
there are 2^8=256 possible fingerprints

Tanimoto coefficient between fingerprints *X* and *Y* is defined to be:
features in intersect(A,B) / # features in union(A,B)

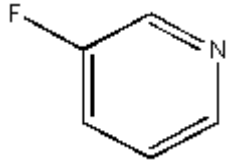
{ is-aromatic, has-ring, has-C }



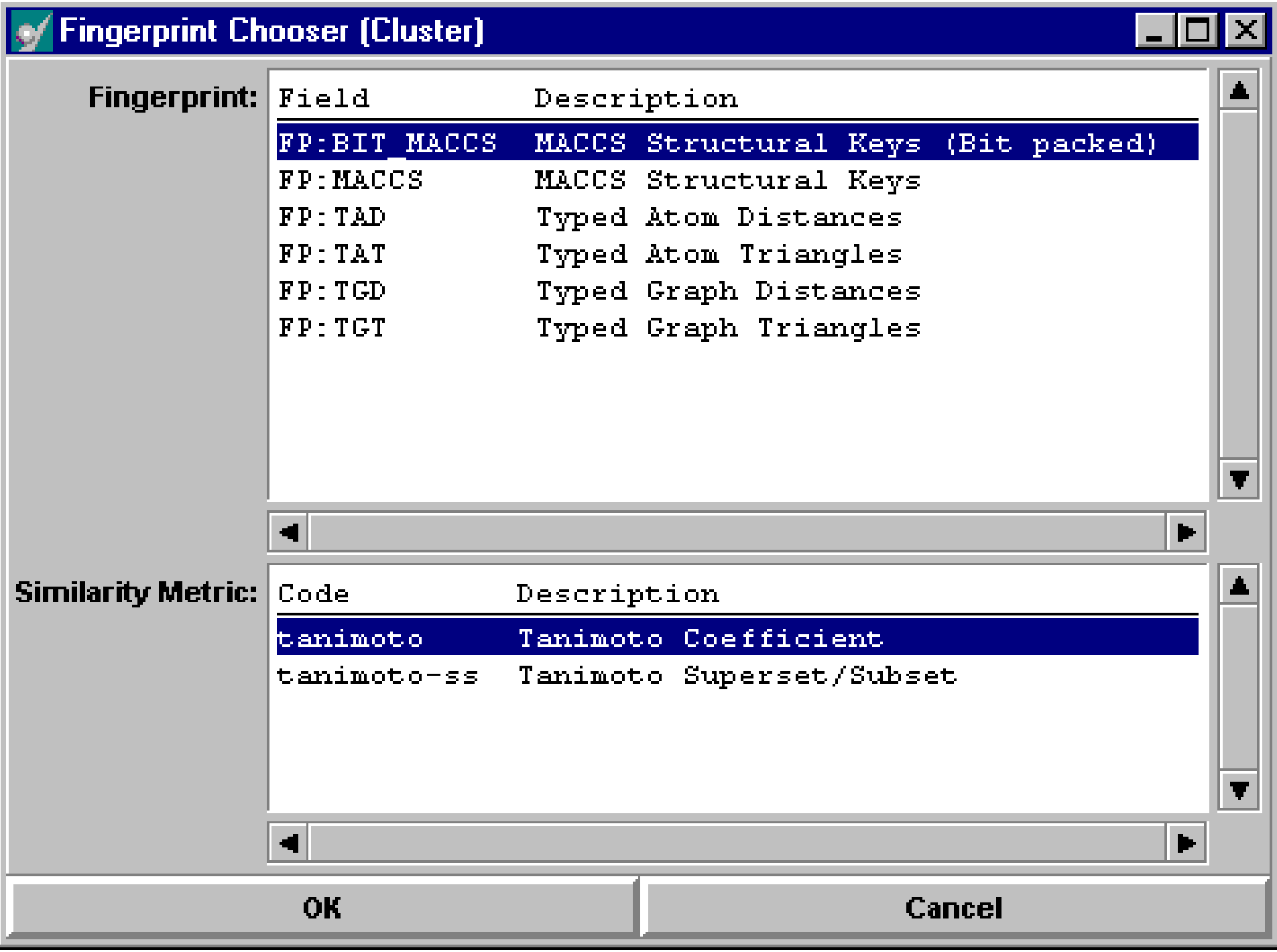
{ has-ring, has-C }

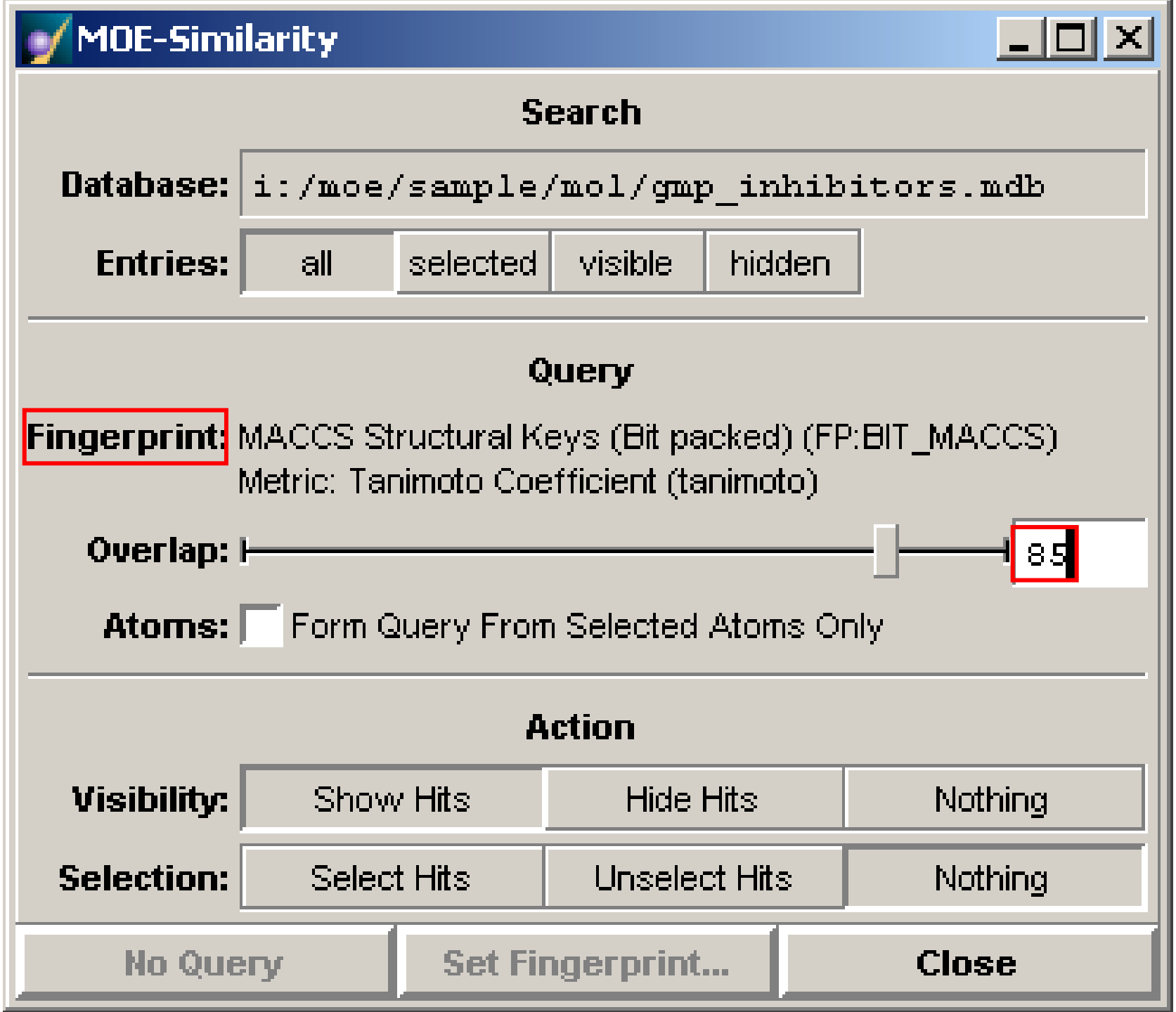


{ has-C, has-N, has-O }

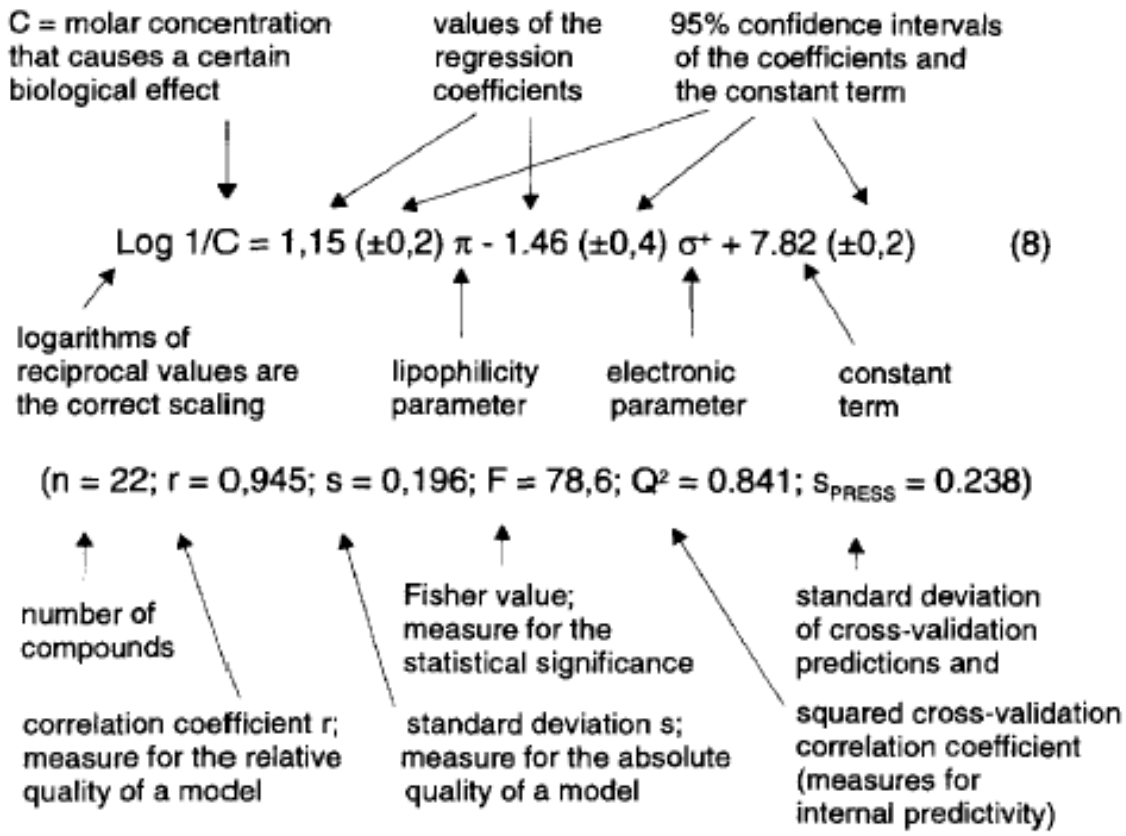


{ is-aromatic, has-ring, has-C, has-N, has-F }


Fingerprint Chooser (Cluster)
Fingerprint:
Field
Description
FP:BIT_MACCS
MACCS Structural Keys (Bit packed)
FP:MACCS
MACCS Structural Keys
FP:TAD
Typed Atom Distances
FP:TAT
Typed Atom Triangles
FP:TGD
Typed Graph Distances
FP:TGT
Typed Graph Triangles
Similarity Metric:
Code
Description
tanimoto
Tanimoto Coefficient
tanimoto-ss
Tanimoto Superset/Subset
OK
Cancel


MOE-Similarity
Search
Database:
i:/moe/sample/mol/gmp_inhibitors.mdb
Entries:
all
selected
visible
hidden
Query
Fingerprint:
MACCS Structural Keys (Bit packed) (FP:BIT_MACCS)
Metric: Tanimoto Coefficient (tanimoto)
Overlap:
85
Atoms:
Form Query From Selected Atoms Only
Action
Visibility:
Show Hits
Hide Hits
Nothing
Selection:
Select Hits
Unselect Hits
Nothing
No Query
Set Fingerprint...
Close

QSAR models


C = molar concentration that causes a certain biological effect
values of the regression coefficients
95% confidence intervals of the coefficients and the constant term
Log 1/C = 1,15 (±0,2) π - 1.46 (±0,4) σ+ + 7.82 (±0,2) (8)
logarithms of reciprocal values are the correct scaling
lipophilicity parameter
electronic parameter
constant term
(n = 22; r = 0,945; s = 0,196; F = 78,6; Q² = 0.841; sPRESS = 0.238)
number of compounds
correlation coefficient r; measure for the relative quality of a model
standard deviation s; measure for the absolute quality of a model
Fisher value; measure for the statistical significance
squared cross-validation correlation coefficient (measures for internal predictivity)
standard deviation of cross-validation predictions and

Useful parameters

- **Log P** :(partition coefficient) Hydrophobicity
  P = [drug] in octanol / [drug] in water
- **p** :substituent hydrophobicity constant
  $\pi_x = \log P_x - \log P_H$

- **Hammett constant (1940)** σ :
  - Measure e-withdrawing or e-donating effects (compared to benzoic acid & how affected its ionization)
- **Taft's steric factor (Es)** (~1956)
  - an experimental value based on rate constants
- **Molar refractivity (MR)**
  - measure of the volume occupied by an atom or group
  - equation includes the MW, density, and the index of refraction
- **Verloop steric parameter-**
  - computer program uses bond angles, van der Waals radii, bond lengths

Sterimol parameter

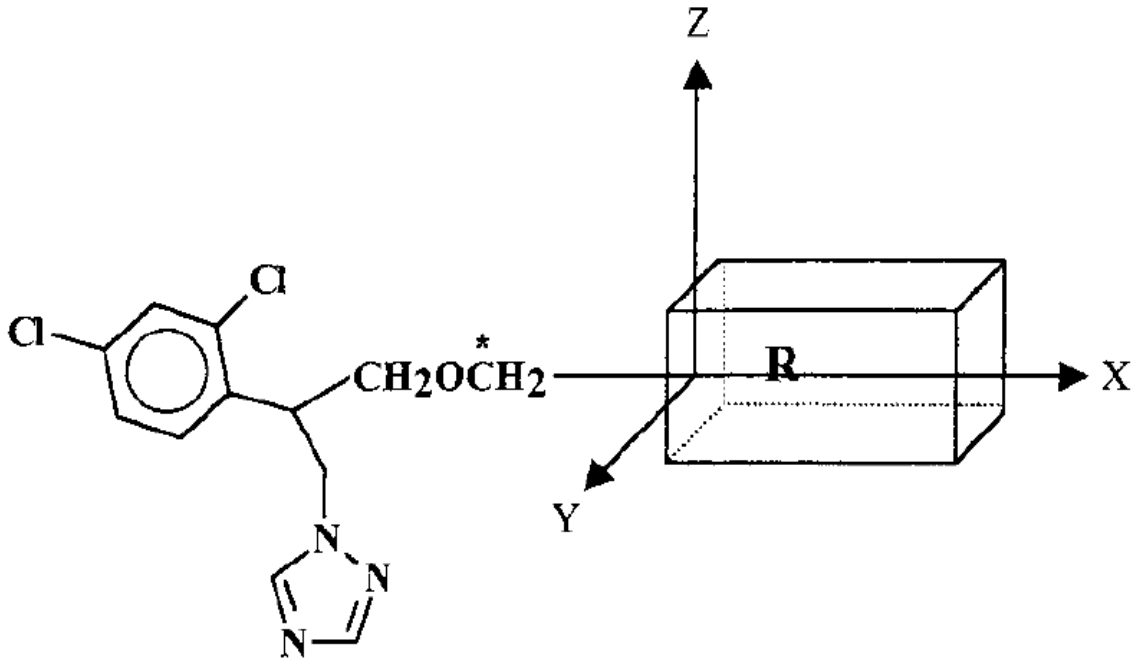


Figure 2. Definition of some of the parameters of Table 3. Geometrical descriptors: $|X|_{max} = X1$; $|Y|_{max} = Y1$; $|Y|_{min} = Y2$; $|Z|_{max} = Z1$; $|Z|_{min} = Z2$. Charge: * indicates the carbon used for the charge.

3D-QSAR Parameters

- Traditional CoMFA fields
  - Steric fields, Lennard-Jones potential

$$E_j = \sum_{k=1}^{natoms} \varepsilon \bullet \left[\left(\frac{r_{probe} + r_k}{r_{ij}} \right)^{12} - 2 \bullet \left(\frac{r_{probe} + r_k}{r_{jk}} \right)^{6} \right]$$

  - Electrostatic fields, Coulomb potential

$$E_j = \sum_{k=1}^{natoms} \frac{q_{probe} \bullet q_k}{\varepsilon \bullet r_{jk}}$$

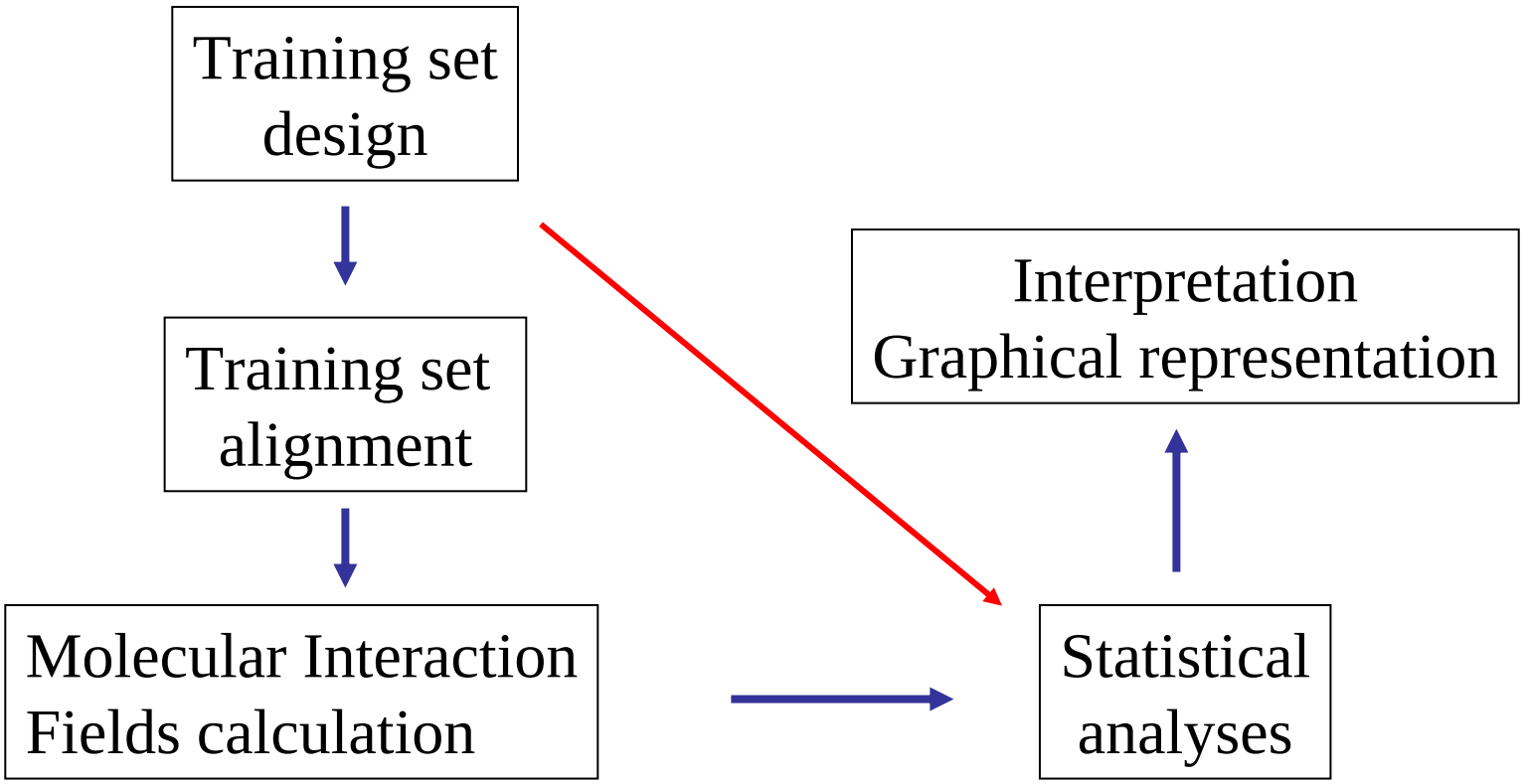
More 3D-QSAR Parameters

- Additional fields in CoMFA
  - Interaction energies with functional groups (GRID software)
  - hydrophobic field (HINT software)
  - Molecular Lipophilicity Field (CLIP software)

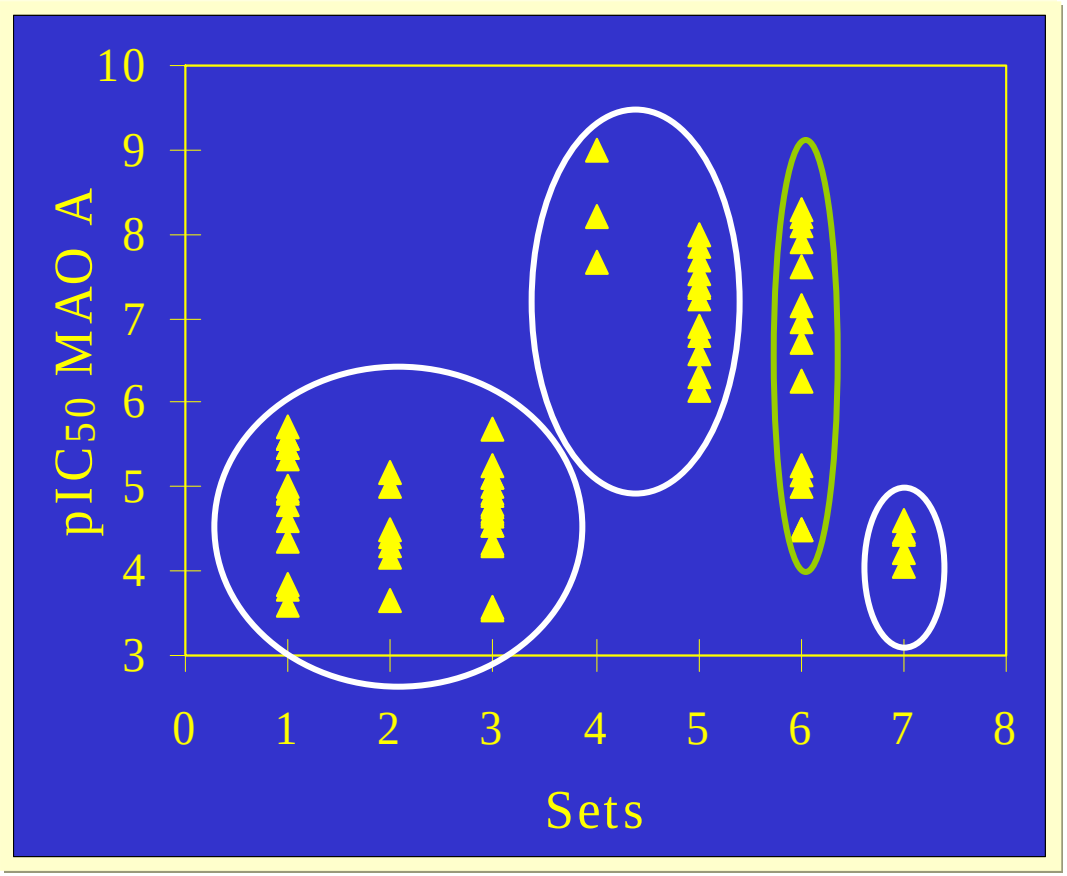
Biological activity parameters

– IC50 (50% inhibiting concentration),
– Ki (inhibitory constant),
– MIC (minimum inhibitory concentration),
– Permeation

Workflow


Training set design
Training set alignment
Molecular Interaction Fields calculation
Statistical analyses
Interpretation Graphical representation

Training set design



QSAR method

- Hansch method
- Free-Wilson method
- Craig Plots
- Topliss operational schemes

3D-QSAR methods

- Methods
  - Molecular-shape analysis [1]
  - Hypothetical Active Site Lattice (HASL) [2]
  - RECEPS Programs [3]
  - Crippen's distance geometry [4]
  - Voronoi binding sites [5]
  - Comparative Molecular Field Analysis (CoMFA) [6]

- [1] Hopfinger A.J. *J. Am. Chem. Soc*. **1980**, 102, 7196-7206
- [2] Doweyko A.M. *J. Med. Chem.* **1988**, 31, 1396-1406
- [3] Kato Y.; Itai A.; Iiatake Y. *Tetrahedron* **1987**, 43, 5229-5236
- [4] Ghose A.K.; Crippen G.M. *J. Med. Chem.* **1985**, 28, 333-346
- [5] Boulu L.D.; Crippen G.M. *J. Comp. Chem.* **1989**, 10, 673-682
- [6] Cramer R.D. III; Patterson D.E.; Bunce J.D. *J. Am. Chem. Soc.* **1988**, 110, 5959-5967

Statistical models

- Linear (quantitative model)
  - LR (linear regression)
  - MLR (multi-linear regression)
  - PCR (principle component regression)
  - PLS (partial least square)
  - SVR (support vector regression)
- Nonlinear (qualitative model)
  - SVM (support vector machine)
  - Bayesian statistics

Evaluation of QSAR models

拟合效果

- 相关系数**R**，

$$R = \sqrt{1 - \frac{\sum (y_{calc} - y_{\exp})^2}{\sum (y_{calc} - y_{mean})^2}}$$

- 均方根偏差**RMSE**

$$RMSE = \sqrt{1 - \frac{\sum (y_{calc} - y_{\exp})^2}{n - k - 1}}$$

- **Fisher检验值F**

$$F = \sqrt{\frac{R^2 (n - k - 1)}{k (1 - R)^2}}$$

y_{calc} 为计算活性；　$y_{\exp}$ 为实验活性；　y_{mean} 为计算活性平均值；　n 为样本个数；　k 为变量个数

Descriptor Choice

编号	描述子名称	物理意义	类别
1	**Pmi**	主惯性矩	空间参数
2	**E_vdw**	范德华构成的势能	电子参数
3	**PmiY**	以Y做主惯性矩	空间参数
4	**PmiZ**	以Z做主惯性矩	空间参数
5	**Zagreb**	Zagreb指数	拓扑参数
6	**PEOE- VSA+1**	在[0.05，0.1）之间q_i的v_i之和	空间参数

描述子及其所代表的物理量及类型

Descriptor Combinations	**RMSE**	**R**	**F 值**
1	**0.40642**	**0.37643**	**2.698**
1+2	**0.26828**	**0.79120**	**11.671**
1+2+3	**0.23969**	**0.70148**	**12.369**
1+2+3+4	**0.21909**	**0.86641**	**13.366**
1+2+3+4+5	**0.16442**	**0.92729**	**22.814**
1+2+3+4+5+6	**0.09823**	**0.97461**	**30.763**

根据**QSAR**模型的拟合能力评价指标确定描述子的组合

Evaluation

- 交叉验证
  - 相关系数 $R_{cv} = \sqrt{1 - \frac{\sum (y_{pred} - y_{exp})^2}{\sum (y_{calc} - y_{mean})^2}}$
  - 预测误差平方

$$PRESS = \sum (y_{pred} - y_{exp})^2$$

- 预测集的预测

R_{CV}→1 or PRESS→0

Better prediction

Statistical value for 3D-QSAR

- For the step of crossvalidation:
  - Cross-validated correlation coefficient, q^2.
  - Optimal number of components, N.

- For the final model:
  - Squared of correlation coefficient, r^2.
  - Standard error of estimate, s.
  - Residuals.
  - F values.

Crossvalidated r^2_{cv} (q^2)

$$q^2 = 1.0 - \frac{\sum (y_{predicted} - y_{experimental})^2}{\sum (y_{predicted} - \overline{y})^2}$$

1.00 = Perfect prediction

Significant statistical results

0.50

Use results only with care when $q^2 > 0.4$)

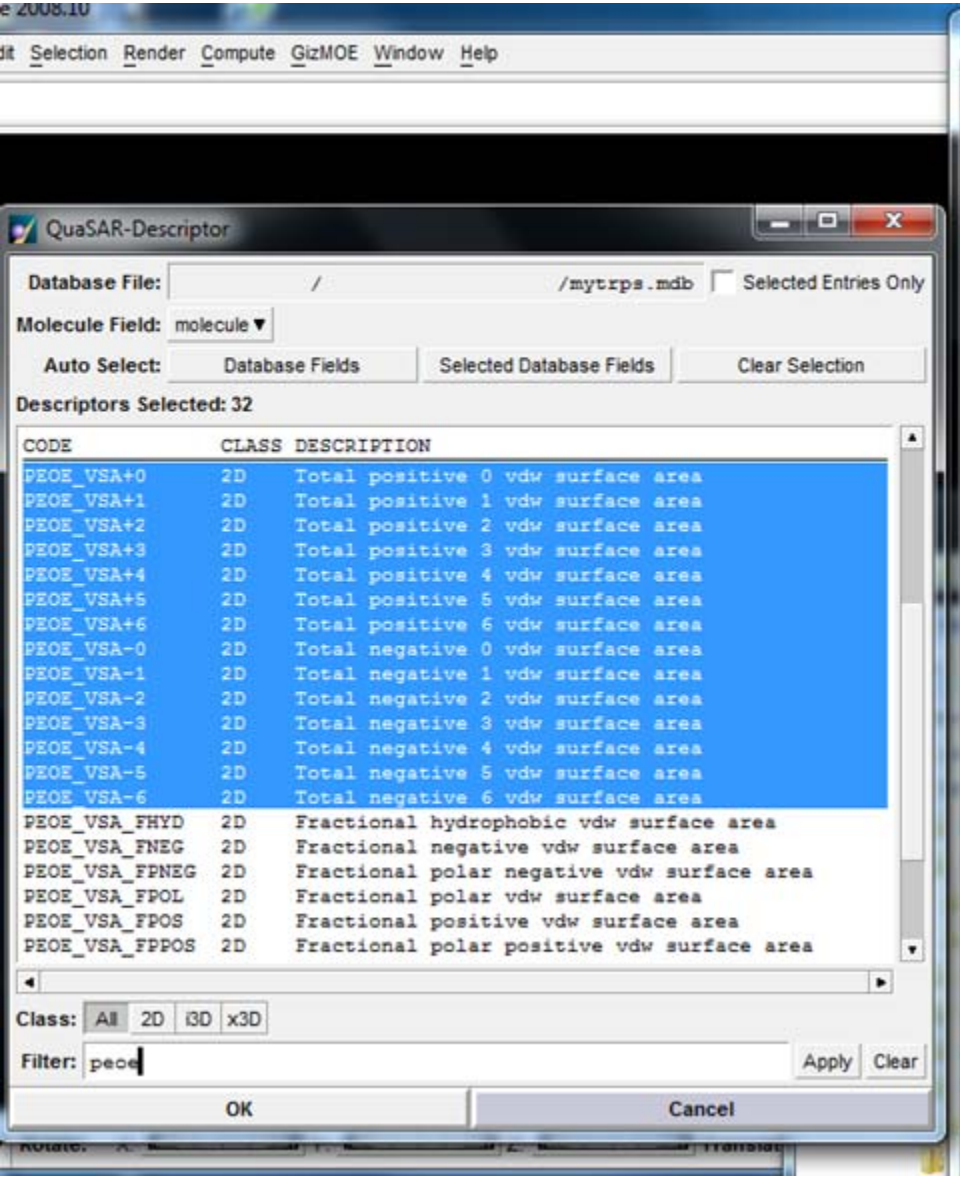
0.00 = **No Model!**

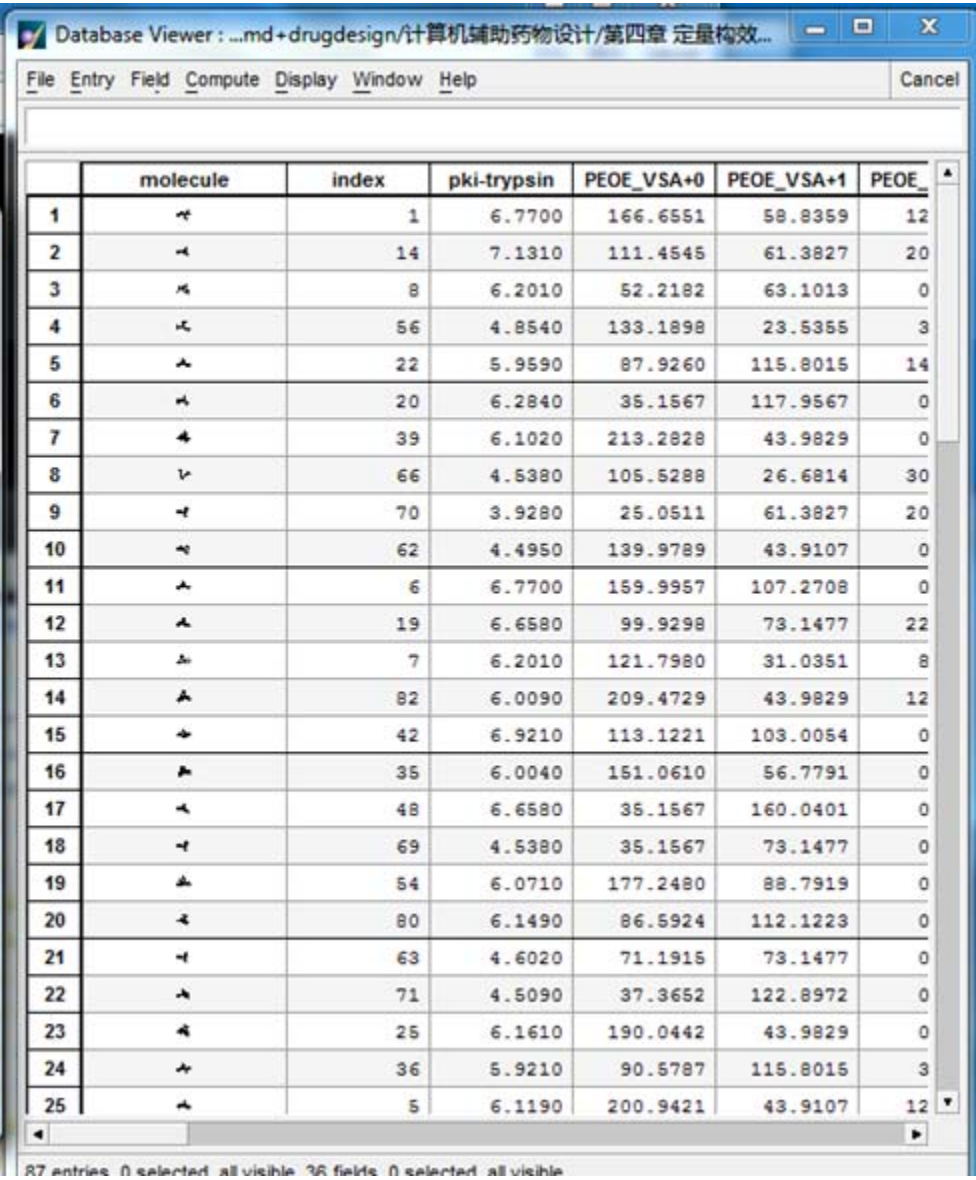
Negative values = prediction worse than those based on the mean over all compounds !

A example by MOE

- Compound database
  - mytrps.mdb
- Statistical Model
  - PLS

Descriptor Choice



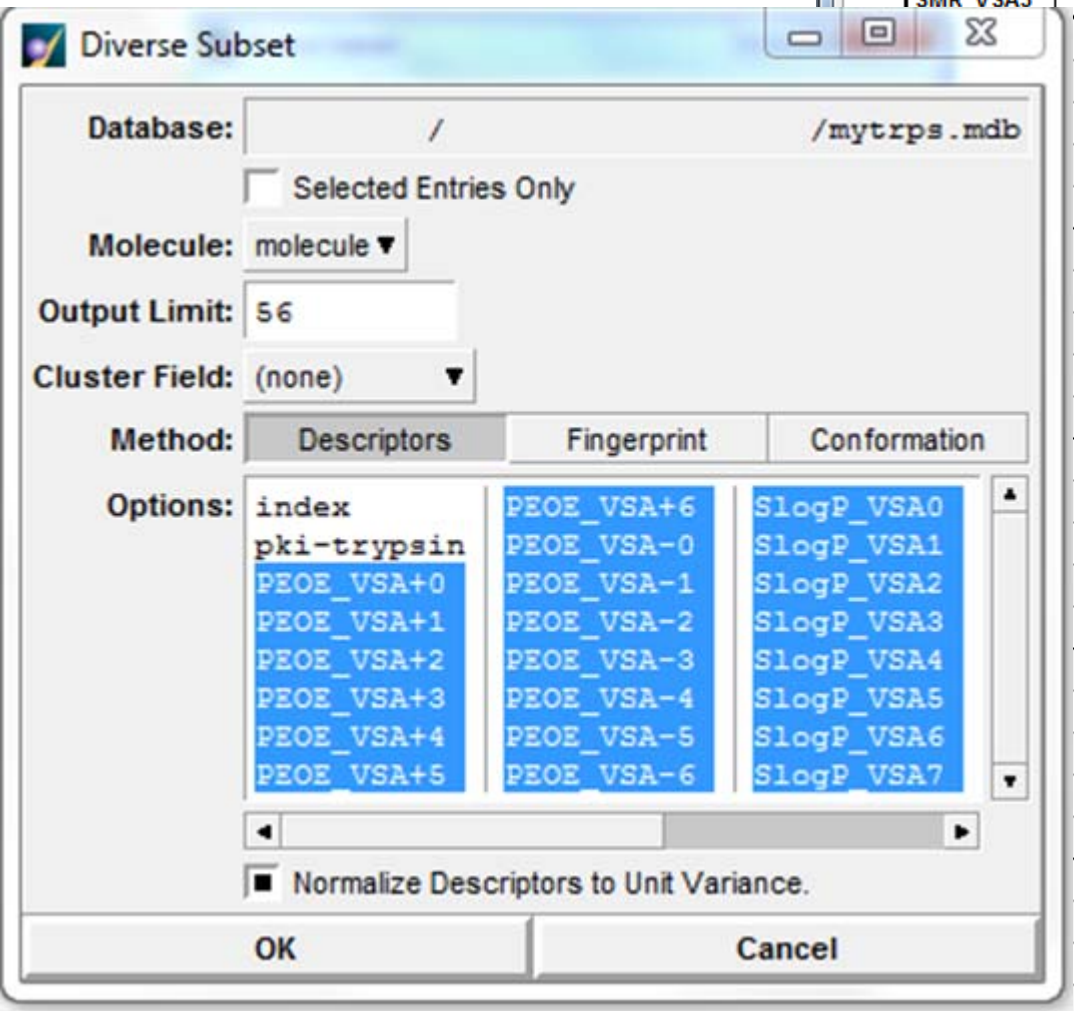


DBV|Compute|Descriptor

87 molecules altogether

Subsets

Popup|Sort up



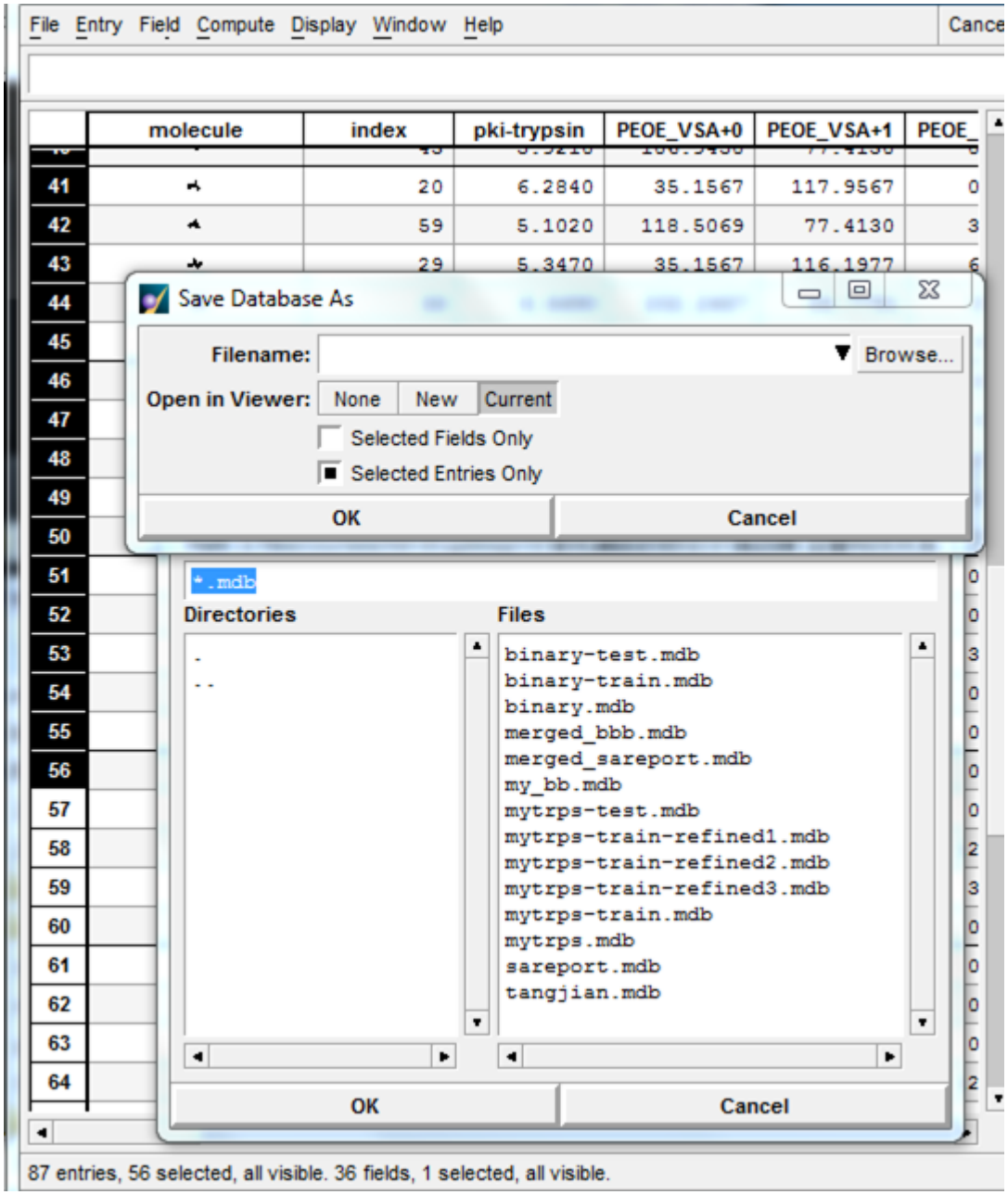
DBV|Compute|Diverse Subset

File Entry Field Compute Display Window Help Cancel

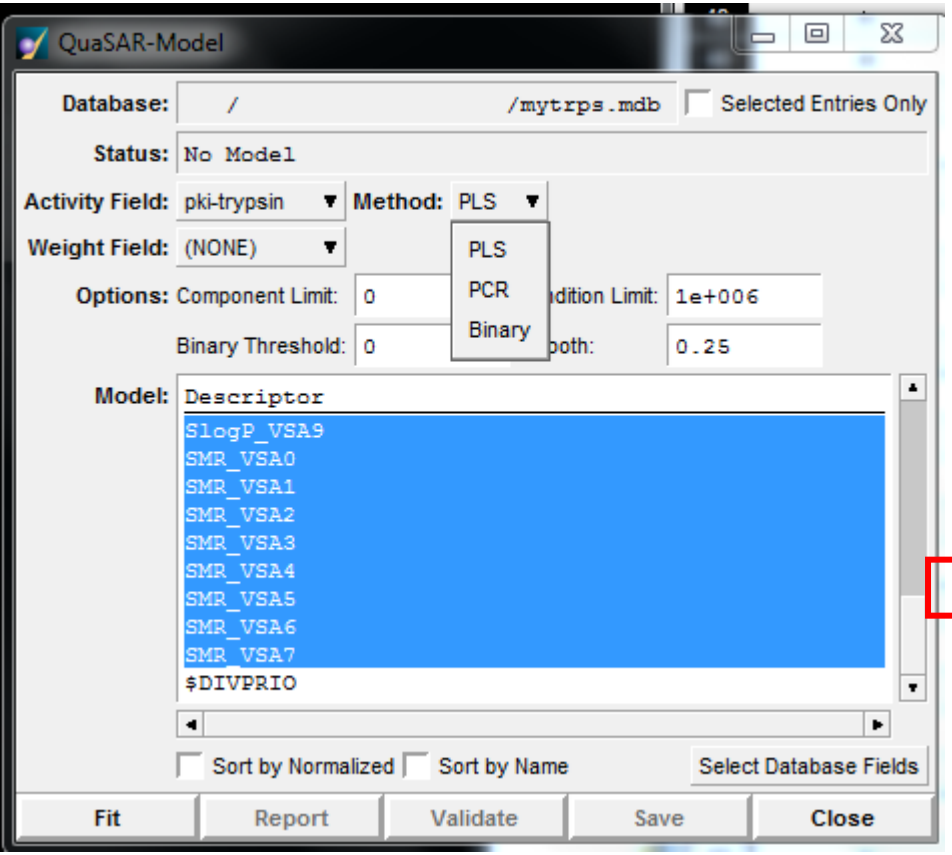
	SMR_VSA3	SMR_VSA4	SMR_VSA5	SMR_VSA6	SMR_VSA7	$DIVPRIO
		40.7999	195.3002	81.3450	110.7670	1
		32.8972	250.6802	55.8411	39.9668	2
		38.0431	250.6802	44.4659	39.9668	3
		32.8972	127.1780	51.7825	133.3041	4
		38.0431	212.9434	107.5301	6.6407	5
		38.0431	322.0856	10.0734	6.6407	41
		32.8972	162.4643	49.7246	239.4940	30
		65.7944	89.4411	81.3450	38.3319	6
		38.0431	212.9434	15.7610	6.6407	13
		32.8972	127.1780	7.5867	206.1680	88
		38.0431	212.9434	86.1750	72.4351	26
		38.0431	250.6802	44.4659	26.7860	9
		32.8972	146.0464	67.3870	39.9668	15
		32.8972	124.7275	81.3450	244.9287	11
		38.0431	357.3358	49.7246	6.6407	12
		32.8972	162.4643	52.2113	142.0027	21
		38.0431	212.9434	81.3450	6.6407	31
		38.0431	212.9434	15.7610	6.6407	17
		32.8972	271.6065	52.2113	206.5968	7
		32.8972	160.0139	86.3184	39.9668	36
		38.0431	212.9434	7.5867	39.5379	27
		40.7999	283.5162	51.4149	6.6407	29
		32.8972	162.4643	44.4659	239.9229	34
		38.0431	320.8242	15.3322	42.0246	24
25	28.3388	32.8972	127.1780	81.3450	211.6027	16

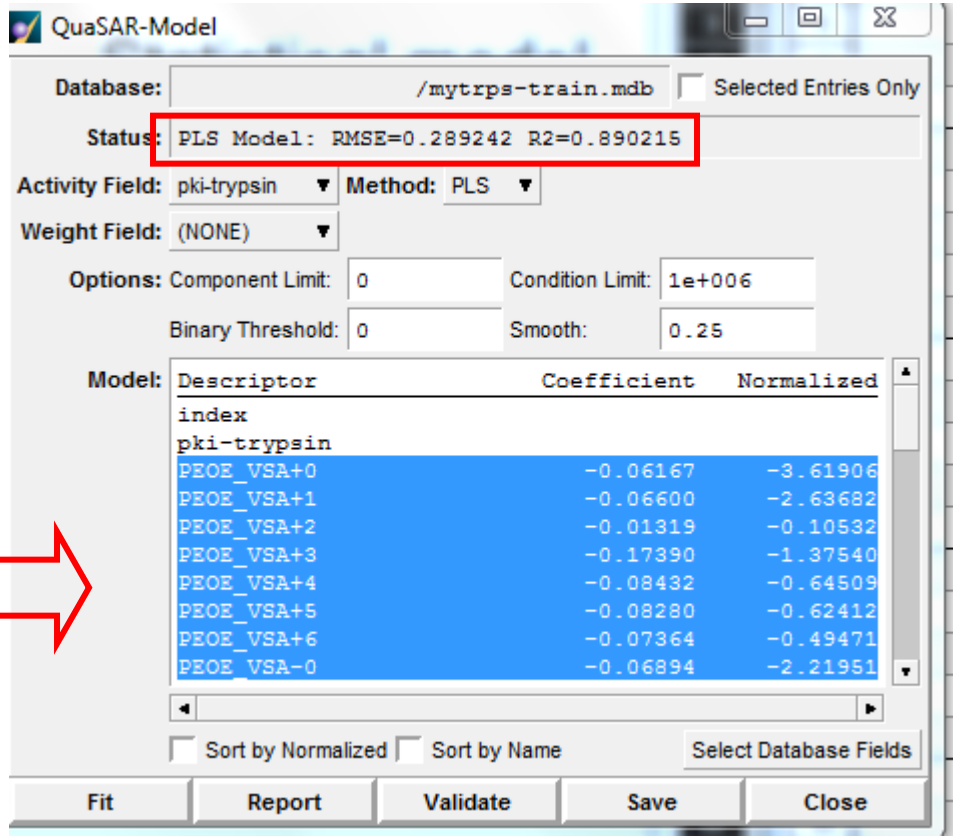
87 entries, 0 selected, all visible. 36 fields, 0 selected, all visible.

Training set and Test set



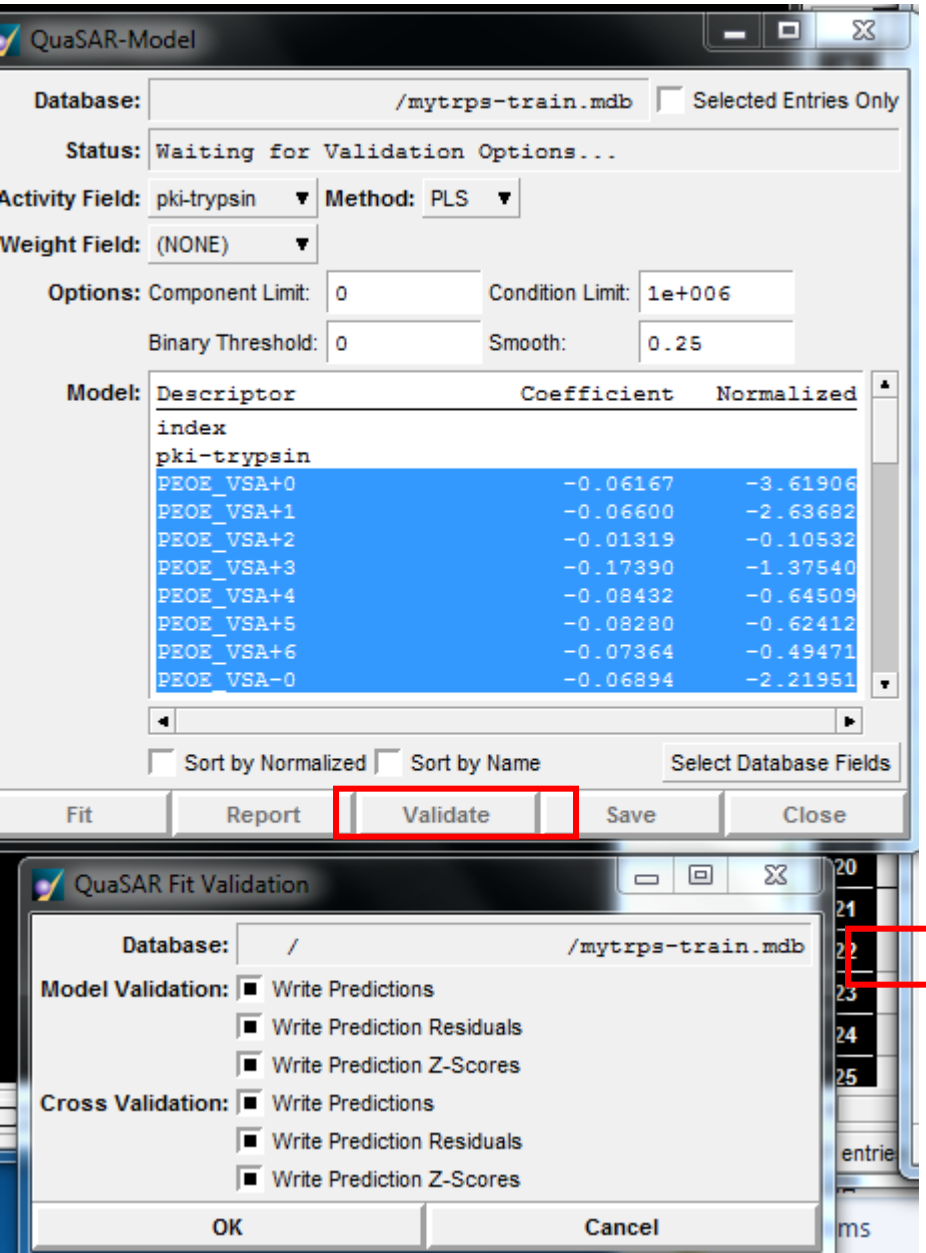
Statistical model

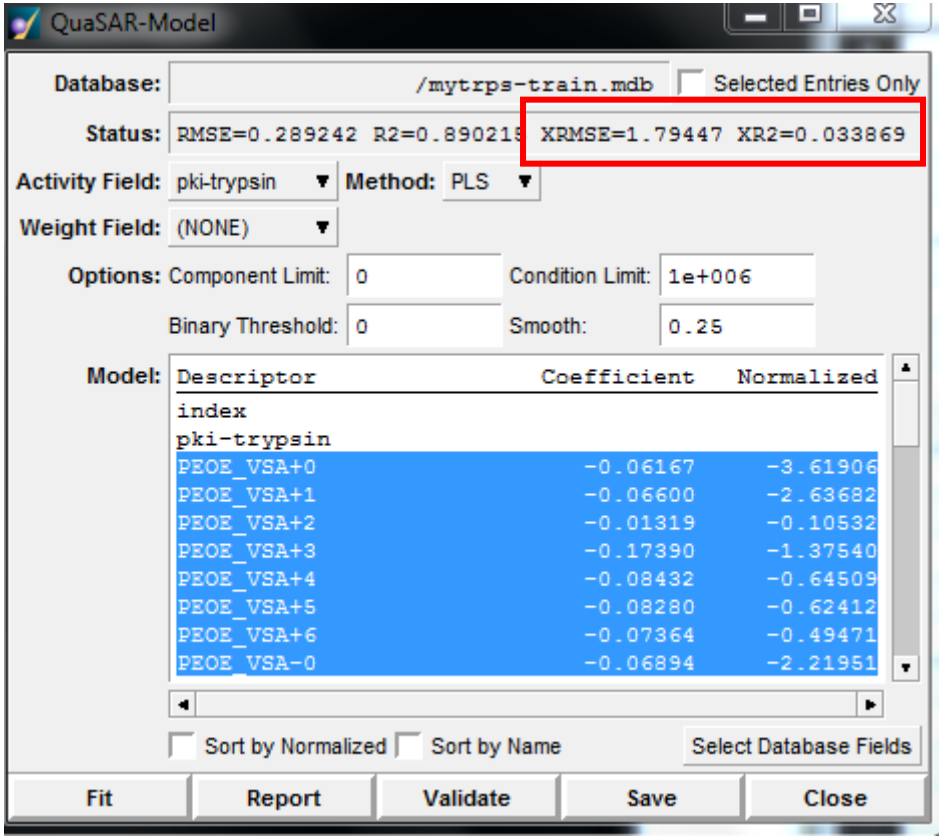




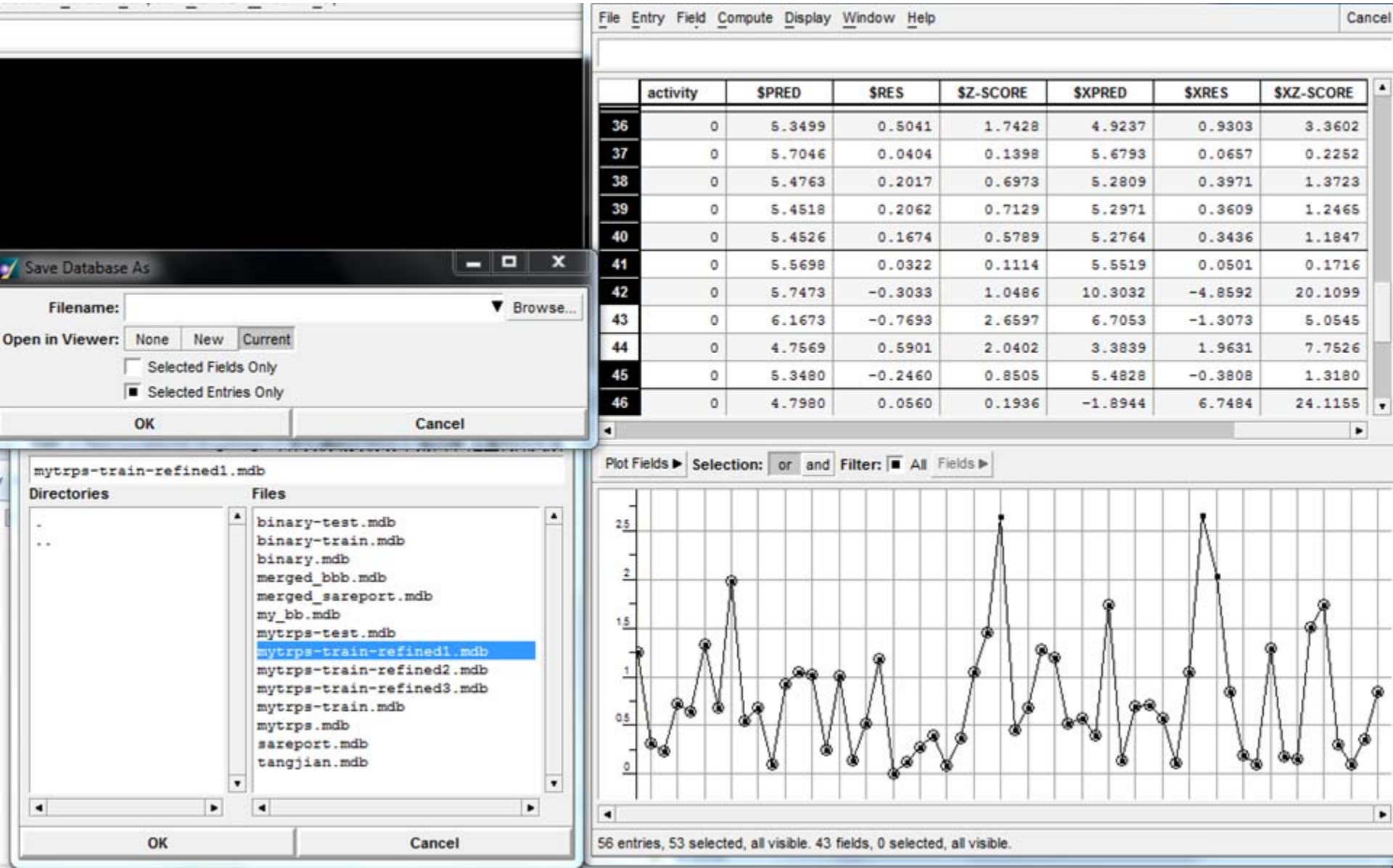
DBV|Compute|QuaSAR-Model

Validation

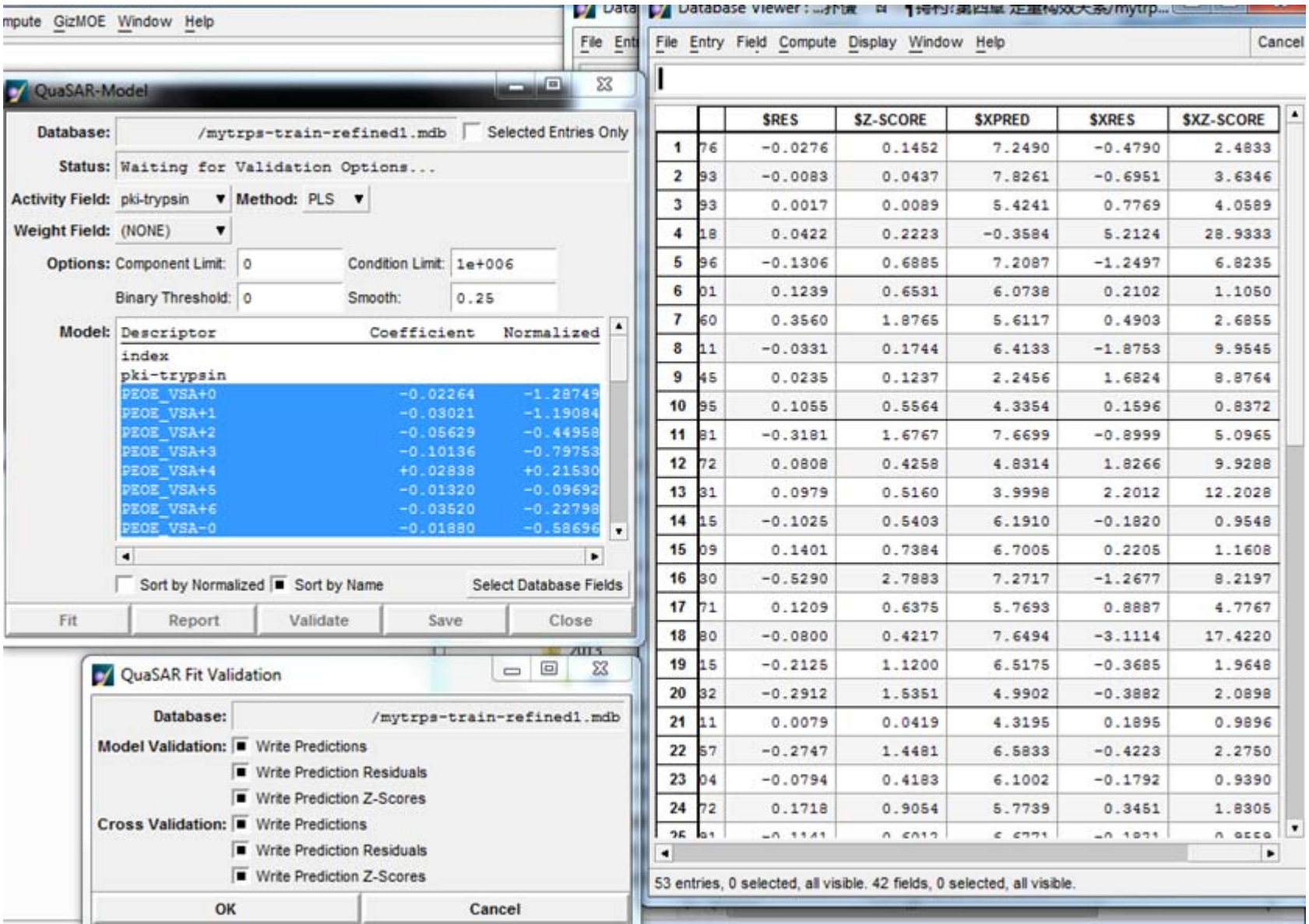

QuaSAR-Model
Database: /mytrps-train.mdb
Selected Entries Only
Status: Waiting for Validation Options...
Activity Field: pki-trypsin
Method: PLS
Weight Field: (NONE)
Options: Component Limit: 0
Condition Limit: 1e+006
Binary Threshold: 0
Smooth: 0.25
Model:
Descriptor Coefficient Normalized
index
pki-trypsin
PEOE_VSA+0 -0.06167 -3.61906
PEOE_VSA+1 -0.06600 -2.63682
PEOE_VSA+2 -0.01319 -0.10532
PEOE_VSA+3 -0.17390 -1.37540
PEOE_VSA+4 -0.08432 -0.64509
PEOE_VSA+5 -0.08280 -0.62412
PEOE_VSA+6 -0.07364 -0.49471
PEOE_VSA-0 -0.06894 -2.21951
Sort by Normalized
Sort by Name
Select Database Fields
Fit
Report
Validate
Save
Close
QuaSAR Fit Validation
Database: / /mytrps-train.mdb
Model Validation:
Write Predictions
Write Prediction Residuals
Write Prediction Z-Scores
Cross Validation:
Write Predictions
Write Prediction Residuals
Write Prediction Z-Scores
OK
Cancel


QuaSAR-Model
Database: /mytrps-train.mdb
Selected Entries Only
Status: RMSE=0.289242 R2=0.890215 XRMSE=1.79447 XR2=0.033869
Activity Field: pki-trypsin
Method: PLS
Weight Field: (NONE)
Options: Component Limit: 0
Condition Limit: 1e+006
Binary Threshold: 0
Smooth: 0.25
Model:
Descriptor Coefficient Normalized
index
pki-trypsin
PEOE_VSA+0 -0.06167 -3.61906
PEOE_VSA+1 -0.06600 -2.63682
PEOE_VSA+2 -0.01319 -0.10532
PEOE_VSA+3 -0.17390 -1.37540
PEOE_VSA+4 -0.08432 -0.64509
PEOE_VSA+5 -0.08280 -0.62412
PEOE_VSA+6 -0.07364 -0.49471
PEOE_VSA-0 -0.06894 -2.21951
Sort by Normalized
Sort by Name
Select Database Fields
Fit
Report
Validate
Save
Close

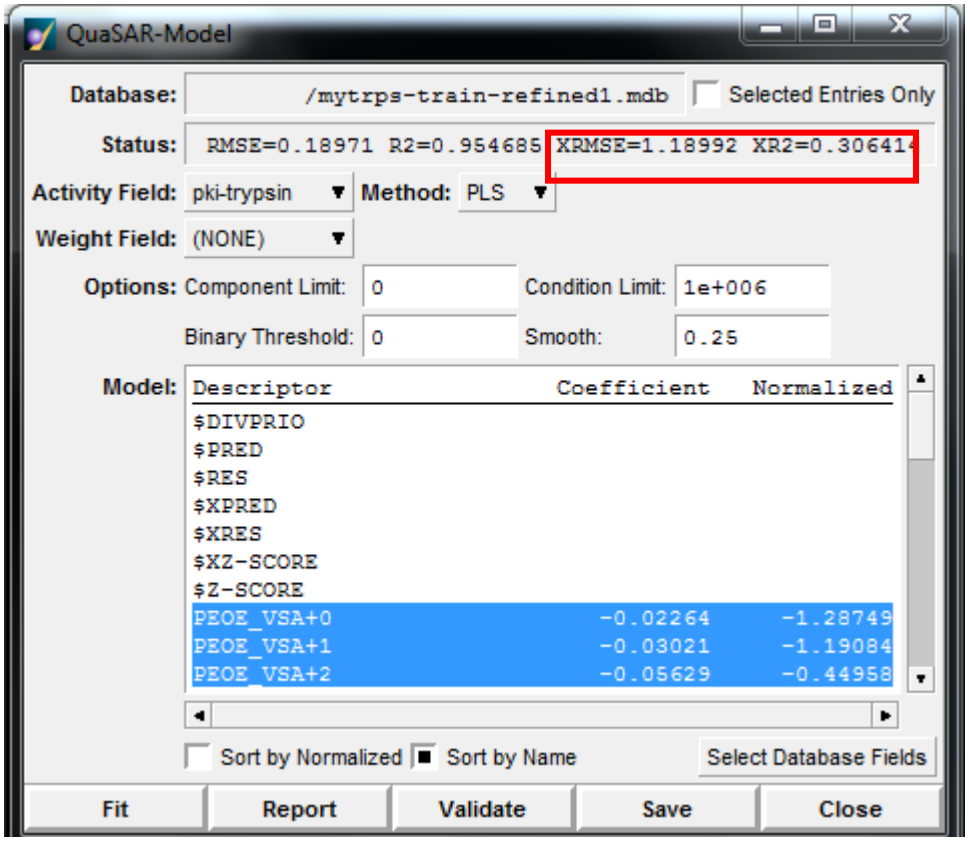
Outliers?


File Entry Field Compute Display Window Help
Cancel
activity $PRED $RES $Z-SCORE $XPRED $XRES $XZ-SCORE
36 0 5.3499 0.5041 1.7428 4.9237 0.9303 3.3602
37 0 5.7046 0.0404 0.1398 5.6793 0.0657 0.2252
38 0 5.4763 0.2017 0.6973 5.2809 0.3971 1.3723
39 0 5.4518 0.2062 0.7129 5.2971 0.3609 1.2465
40 0 5.4526 0.1674 0.5789 5.2764 0.3436 1.1847
41 0 5.5698 0.0322 0.1114 5.5519 0.0501 0.1716
42 0 5.7473 -0.3033 1.0486 10.3032 -4.8592 20.1099
43 0 6.1673 -0.7693 2.6597 6.7053 -1.3073 5.0545
44 0 4.7569 0.5901 2.0402 3.3839 1.9631 7.7526
45 0 5.3480 -0.2460 0.8505 5.4828 -0.3808 1.3180
46 0 4.7980 0.0560 0.1936 -1.8944 6.7484 24.1155
Plot Fields
Selection: or and
Filter: All Fields
2.5
2
1.5
1
0.5
0
56 entries, 53 selected, all visible. 43 fields, 0 selected, all visible.
Save Database As
Filename:
Browse...
Open in Viewer: None New Current
Selected Fields Only
Selected Entries Only
OK
Cancel
mytrps-train-refined1.mdb
Directories
Files
.
..
binary-test.mdb
binary-train.mdb
binary.mdb
merged_bbb.mdb
merged_sareport.mdb
my_bb.mdb
mytrps-test.mdb
mytrps-train-refined1.mdb
mytrps-train-refined2.mdb
mytrps-train-refined3.mdb
mytrps-train.mdb
mytrps.mdb
sareport.mdb
tangjian.mdb
OK
Cancel

Fit model II → Vailidation II



After refinement



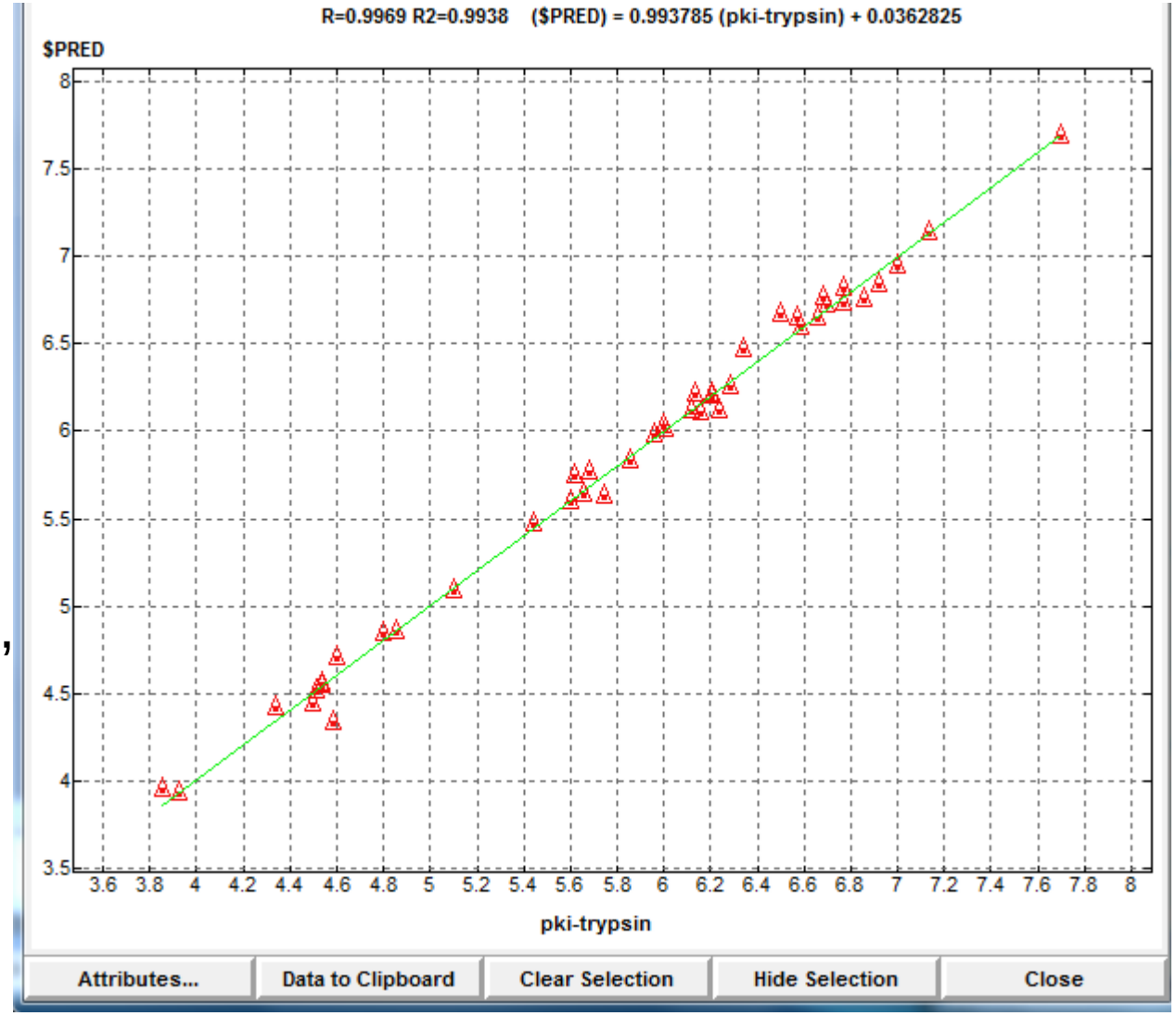
Training dataset III

Again, delete outliers with Z-SCORE>1.5,

→44 entries

RMSE=0.0739945,
R2=0.993785,
XRMSE=0.48619,
XR2=0.770392

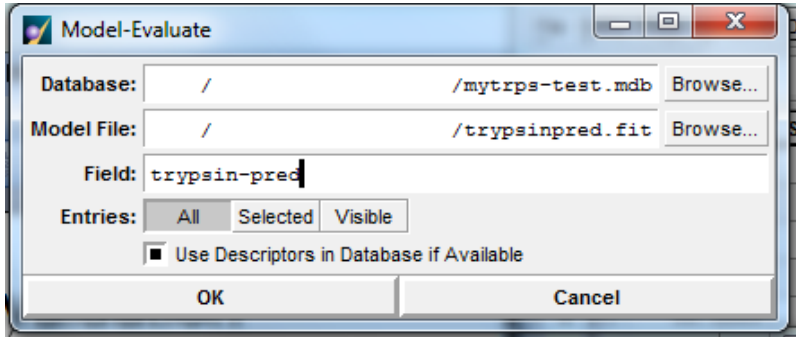
→trypsinpred.fit

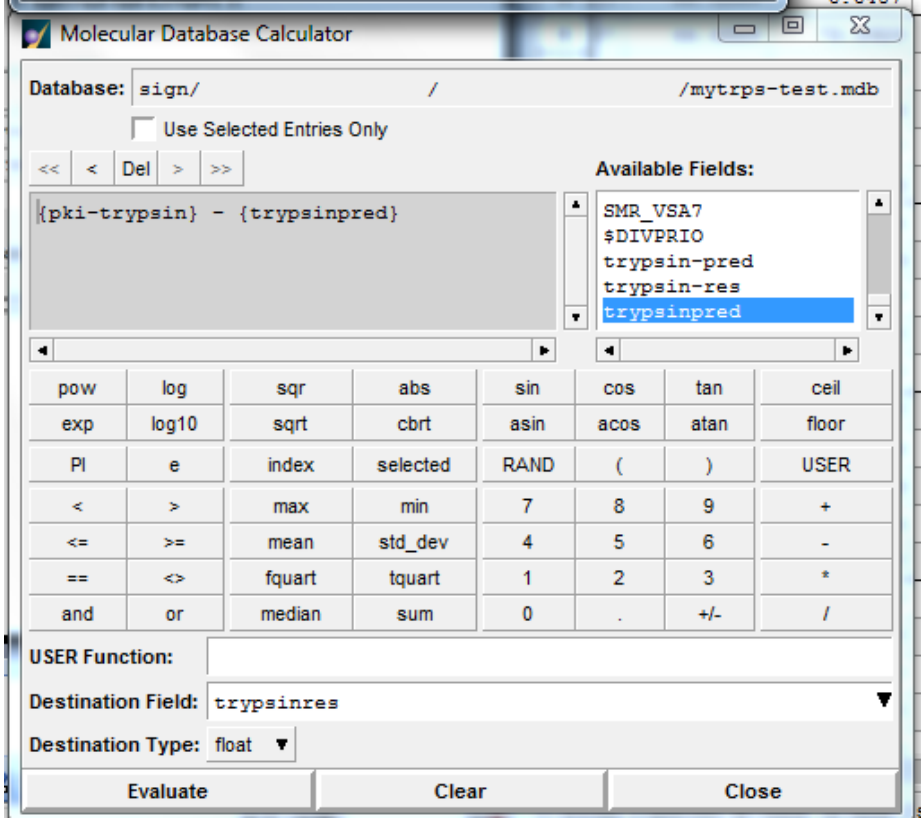


DBV|Compute|Analysis|Correlation plot

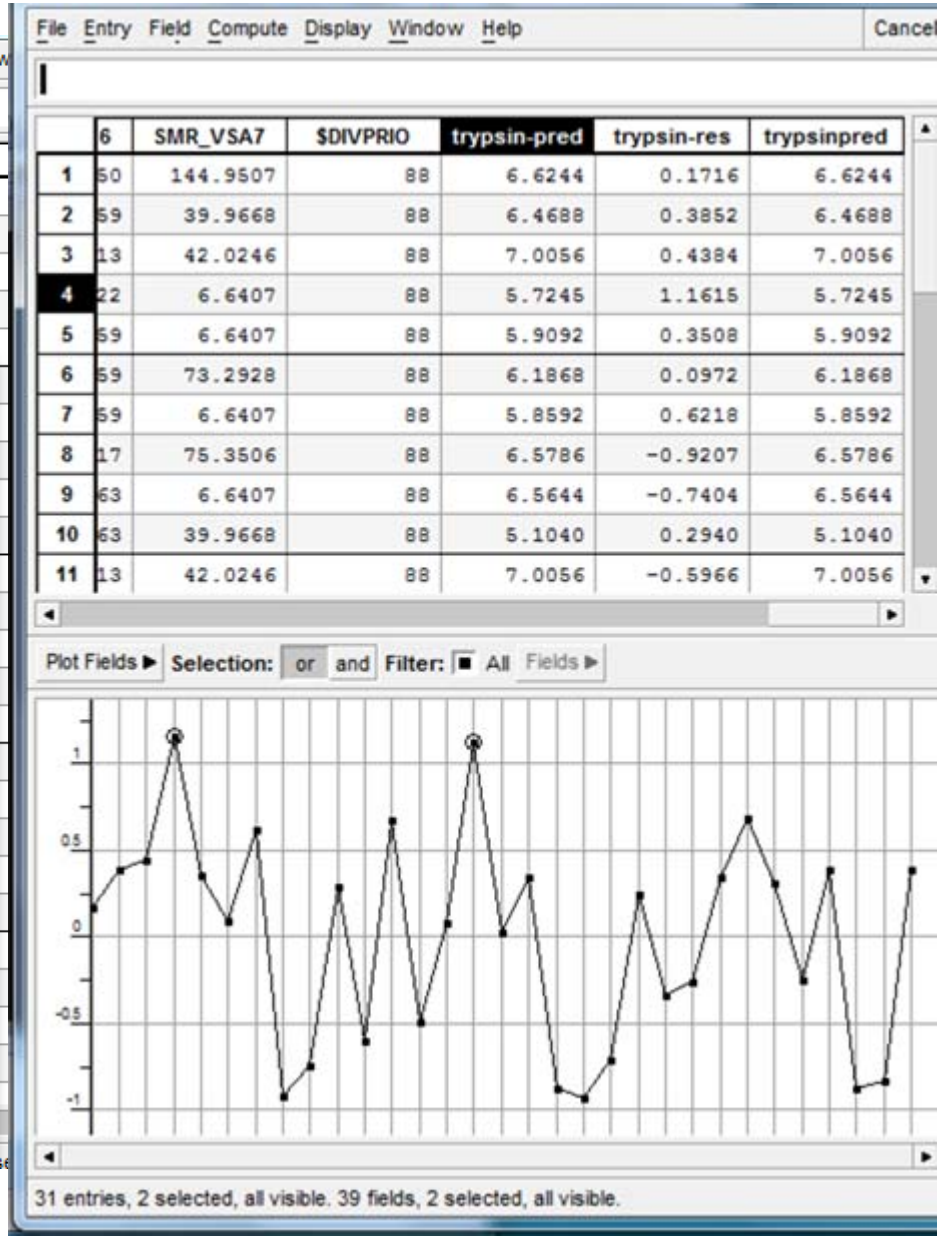
Model Evaluation

DBV|Compute|Model-Evaluate



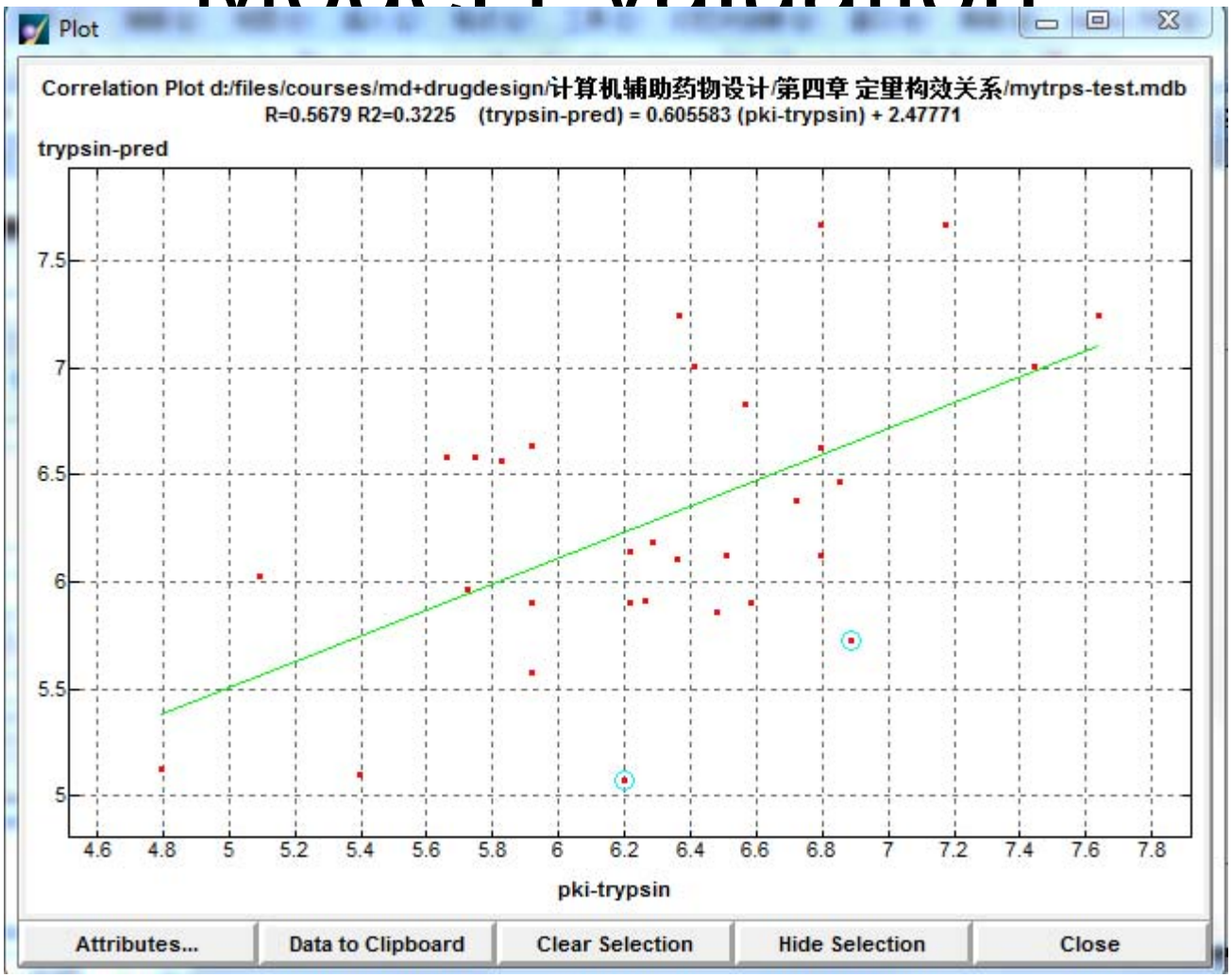


DBV|Compute|Calculator



	6	SMR_VSA7	$DIVPRIO	trypsin-pred	trypsin-res	trypsinpred
1	50	144.9507	88	6.6244	0.1716	6.6244
2	59	39.9668	88	6.4688	0.3852	6.4688
3	13	42.0246	88	7.0056	0.4384	7.0056
4	22	6.6407	88	5.7245	1.1615	5.7245
5	59	6.6407	88	5.9092	0.3508	5.9092
6	59	73.2928	88	6.1868	0.0972	6.1868
7	59	6.6407	88	5.8592	0.6218	5.8592
8	17	75.3506	88	6.5786	-0.9207	6.5786
9	63	6.6407	88	6.5644	-0.7404	6.5644
10	63	39.9668	88	5.1040	0.2940	5.1040
11	13	42.0246	88	7.0056	-0.5966	7.0056

Model Evaluation



Course Outline

- Introduction and Case Study
- Drug Targets
  - Sequence analysis
  - Protein structure prediction
  - Molecular simulation
- Molecular Docking
- Drug Design
  - QSAR
  - Pharmacophore
  - De novo Drug Design
  - Combinatorial library