Generation of QSAR Sets with a Self Organizing Map

Rajarshi Guha, Jon Serra and Peter C. Jurs
Department of Chemistry
The Pennsylvania State University

1

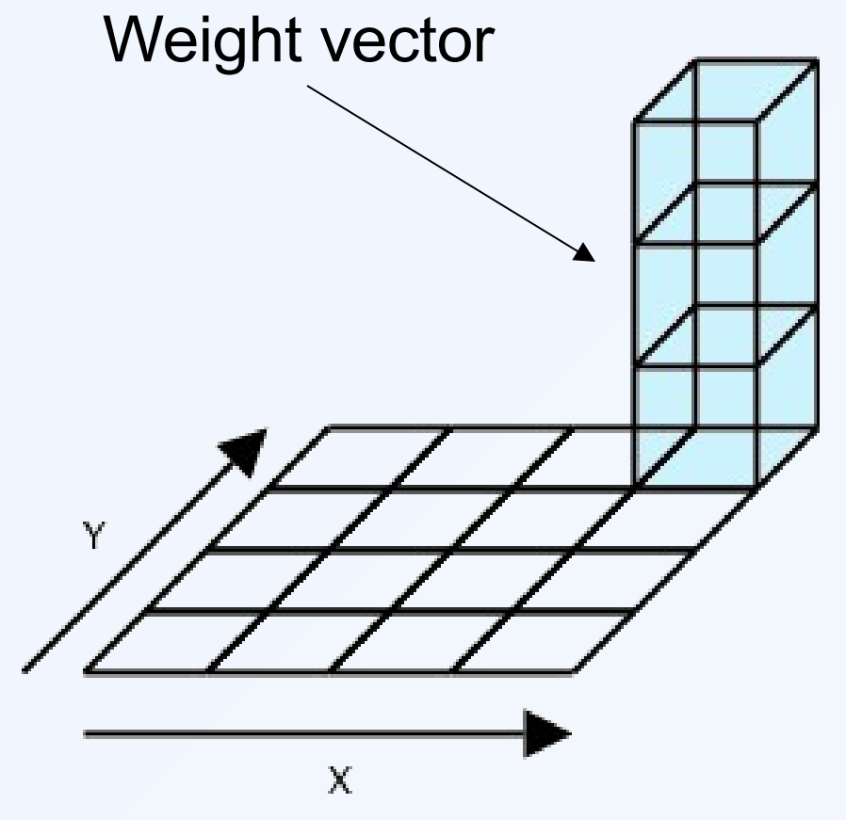
Introduction

- QSAR modelling requires the creation of a training, cross validation and prediction set, collectively termed QSAR sets. There are several methods available to create QSAR sets including random selection, activity binning and sphere selection.
- However, QSAR models are trained and validated using different sets of molecules. Since a QSAR models tries to capture features of the dataset and use those features to make predictions, it is important that the training and validation sets be representative of the dataset as a whole.
- Thus the goal of QSAR set creation should be to generate representative sets. That is, similar groups of molecules in the overall dataset should be represented in the QSAR sets in similar proportions.
- Thus we need to detect ***similar*** sets of molecules. This can be achieved with the help of s Self Organizing Map (SOM)

2

What is a SOM?

- Unsupervised neural network
- Transforms an N-dimensional dataset into a 2-D grid
- Maintains topology
- Competitive training
- Does not require the dependant variable



Kohonen, T., *Self Organizing Maps*, Springer; 1994

3

How Is The SOM Used?

- Construct the map.
- Train the map.
  - Allow neurons to compete.
  - Modify winning neurons.
- Use the map to detect classes.
- Use the SOM classes to generate QSAR sets.

4

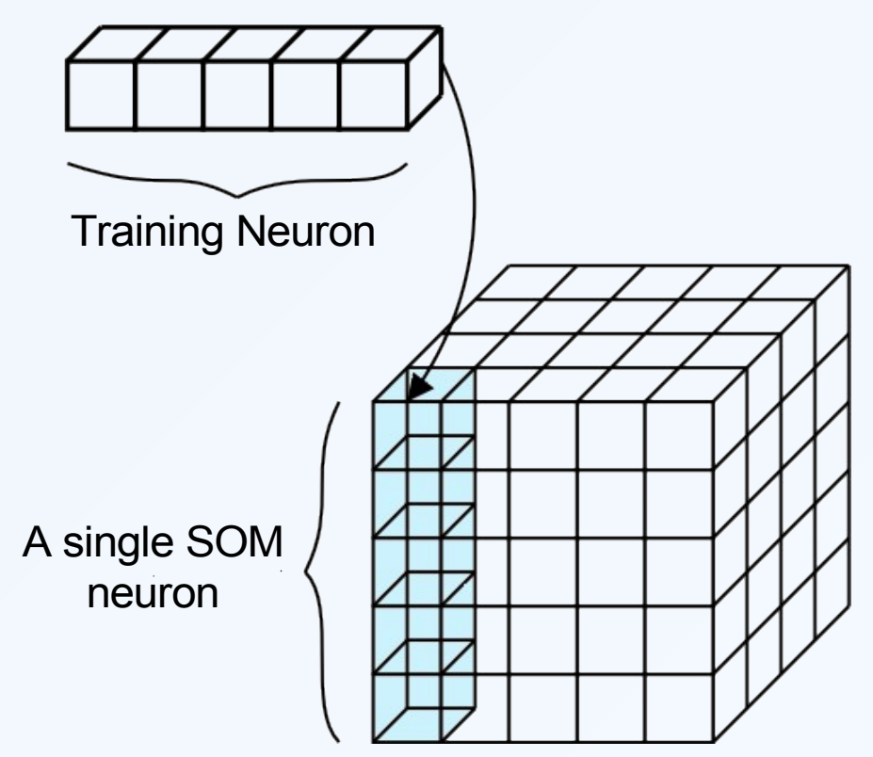
Preparing the SOM

- ***Requirements***
  - A training set. Since we only want to divide the dataset, the whole dataset is used as the training set
  - A set of molecular descriptors which characterize whole molecules rather than specific features
- ***Constructing the SOM***
  - 13x13 square grid wrapped round the edges
  - Each point is a neuron represented by a weight vector
  - The length of the vector equals the number of descriptors used
  - Neurons are initialized with random weights

5

Training the Map

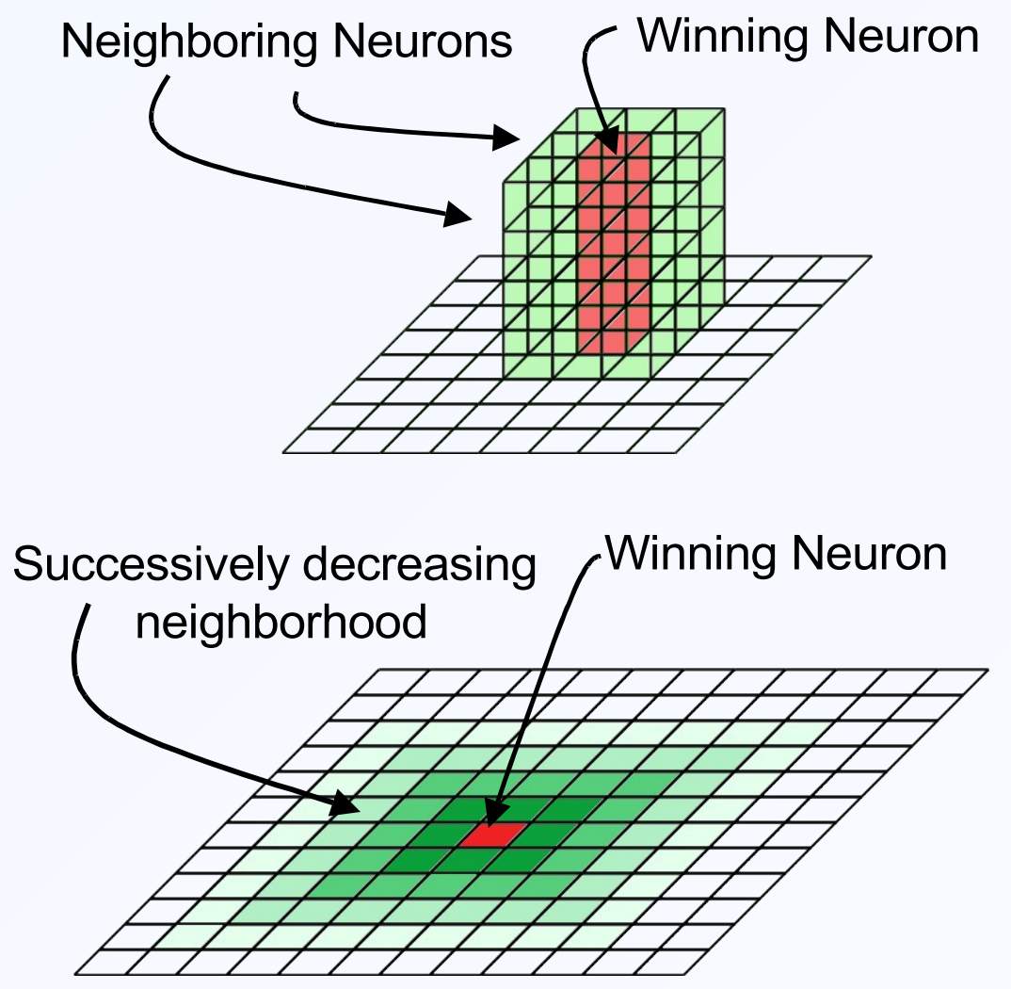
- Training examples are presented to all units
- Units compete for selection
- The winner is unit which is closest to the example in an Euclidean sense.
- The winner is modified and this process is repeats for all the training examples multiple times.



6

Modification of the Map

- Winning neurons and ***topologically*** close neurons are modified
- This leads to global ordering
- The extent of modification decreases in successive training cycles
- The number of neighbor hood neurons is controlled by the neighborhood function
- The size of the neighborhood is decreased as training progresses



7

The Mathematics of the Map

- Neurons are modified according to

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)]$$

- The neighborhood function is Gaussian

$$h_{ci}(t) = \alpha(t)\exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right)$$

- $\alpha(t)$ is the learning factor and controls the extent of modification of a neuron

$$\alpha(t+1) = \alpha(t) - 0.01$$

8

Detecting Classes in the Map

- Assign an arbitrary class to the first neuron.
- For each neuron calculate distances to the nearest neighbors.
- If the distance to a neighbor is less than a user-specified threshold, then the neighbor is in the same class as the grid point.
- After class assignments of the grid neurons, use these assignments to divide the dataset into two classes.
- Using the SOM classes assign classes to the dataset.

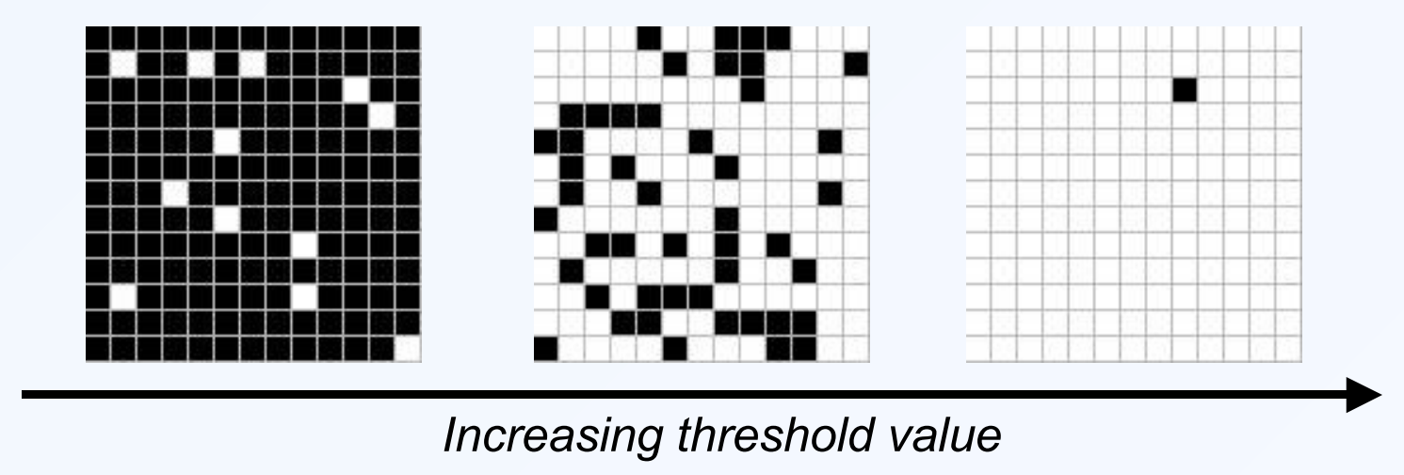
9

Selecting a Threshold Value

- Run the SOM once to get an idea about the distances between each neuron
- Next run the SOM multiple times with threshold values ranging from 10% to 90% of the maximum distance observed
- Look for class breakups which are approximately 80-20

10

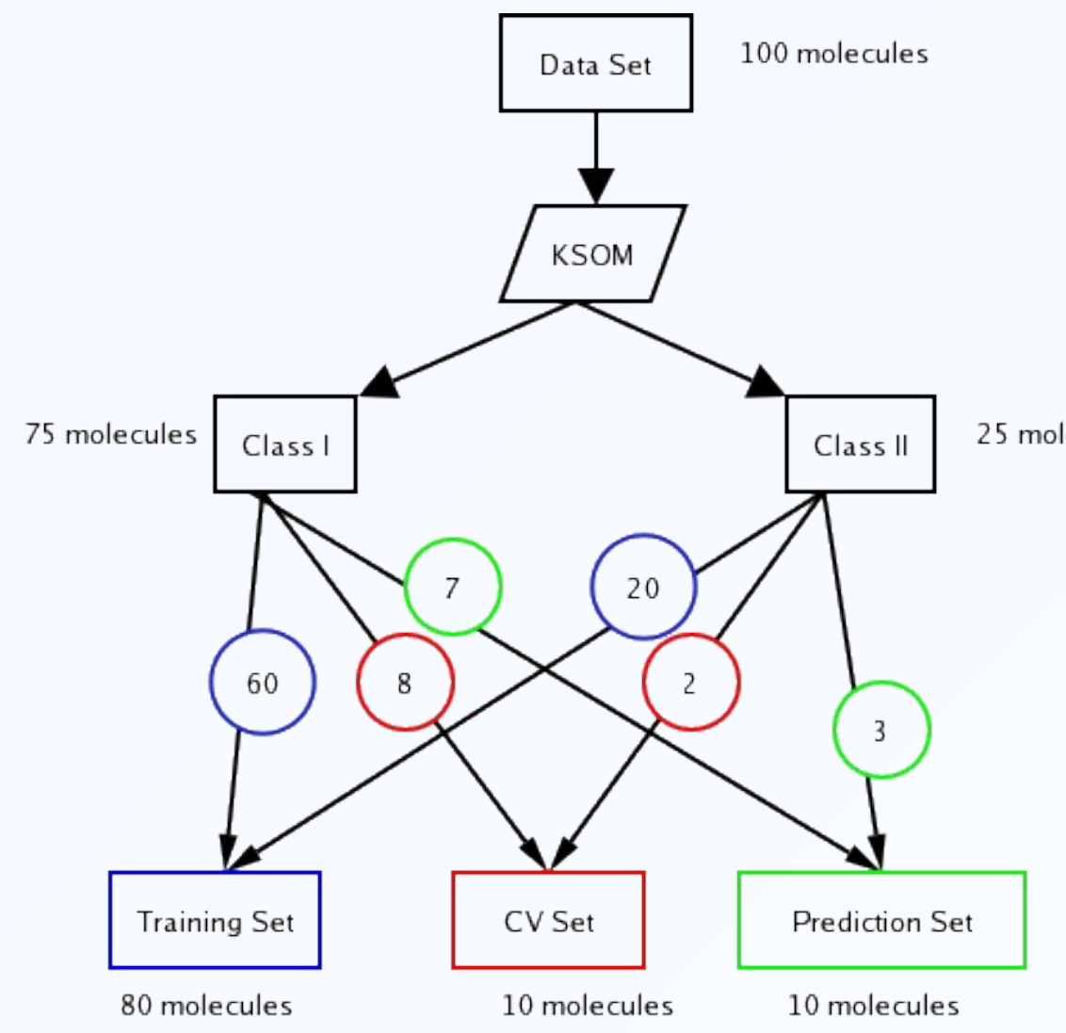
The Clustered SOM



- The above diagrams represent the clusters in a SOM trained with the BCUT – 2D Autocorrelation Dragon descriptor
- The differences arise due to the different threshold values used in the cluster detection algorithm

11

SOM Classes to QSAR Sets



12

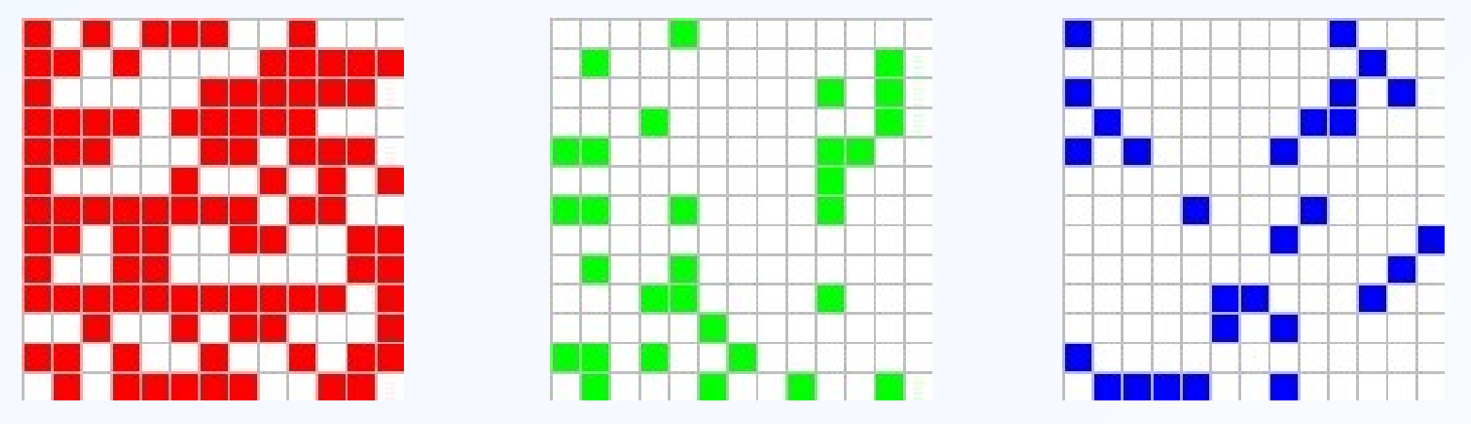
Testing the Technique

- The technique was applied to a dataset of DHFR inhibitors studied by Mattioni et al.
- The original work used activity binning to select QSAR sets.
- The dataset contained 333 molecules.
- 6 combinations of Dragon descriptors were submitted to the SOM to create 6 QSAR sets.
- Each QSAR set was used to generate models using the ADAPT methodology.

Mattioni, B., et al., *J. Mol. Graph. And Modell.* (**2003**) *21*, 391

13

Distribution of QSAR Sets Over the SOM



Training Set CV Set Prediction Set

The distributions were obtained by using the BCUT & 2D Autocorrelation Dragon descriptors.

14

Computational Neural Network Models

Dragon Descriptor Set	Architecture	RMS Error		
		Training	Cross Validation	Prediction
BCUT & 2D Auto.	5-3-1	0.63	0.68	0.79
BCUT & Galvez	5-3-1	0.62	0.62	0.71
GETAWAY	5-2-1	0.73	0.73	0.65
MoRSE & 2D Auto.	5-3-1	0.63	0.63	0.68
MoRSE & GETAWAY	9-5-1	0.49	0.59	0.76
MoRSE & WHIM	6-5-1	0.60	0.61	0.65
Published	10-6-1	0.45	0.49	0.66

Mattioni, B., et al., *J. Mol. Graph. And Modell.* (**2003**) *21*, 391

15

Random Sets vs. SOM Sets

	Random Sets			MoRSE – WHIM Set		
	Mean RMSE	Std Dev	R^2	Mean RMSE	Std Dev	R^2
TSET	0.57	0.02	0.75	0.58	0.005	0.74
CSET	0.59	0.03	0.73	0.57	0.001	0.76
PSET	0.80	0.13	0.56	0.63	0.63	0.63

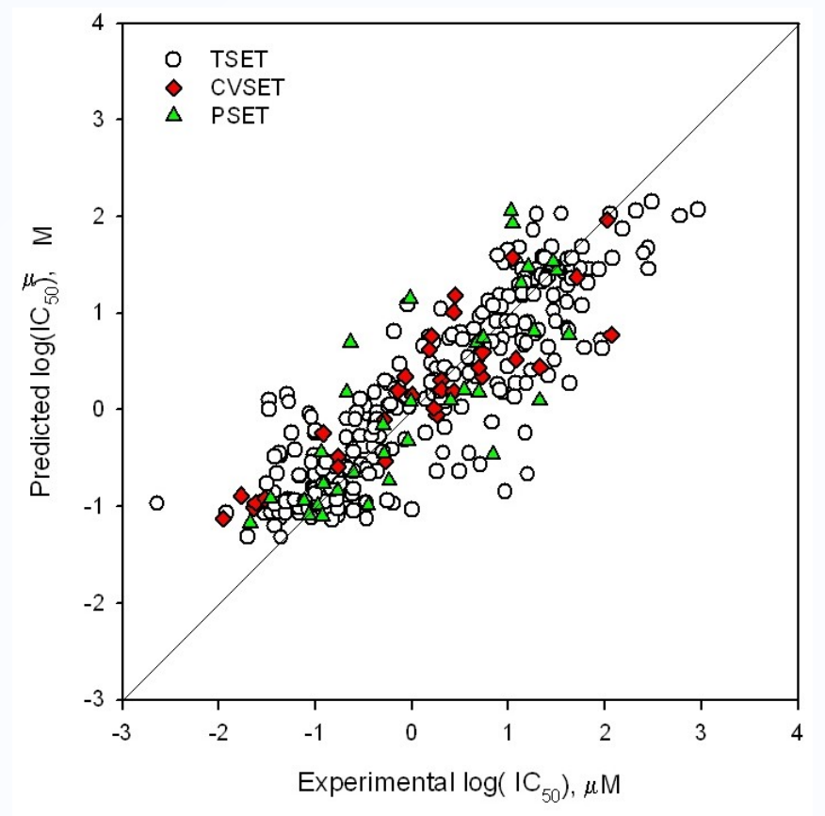
16

Scrambled Dependent Variable

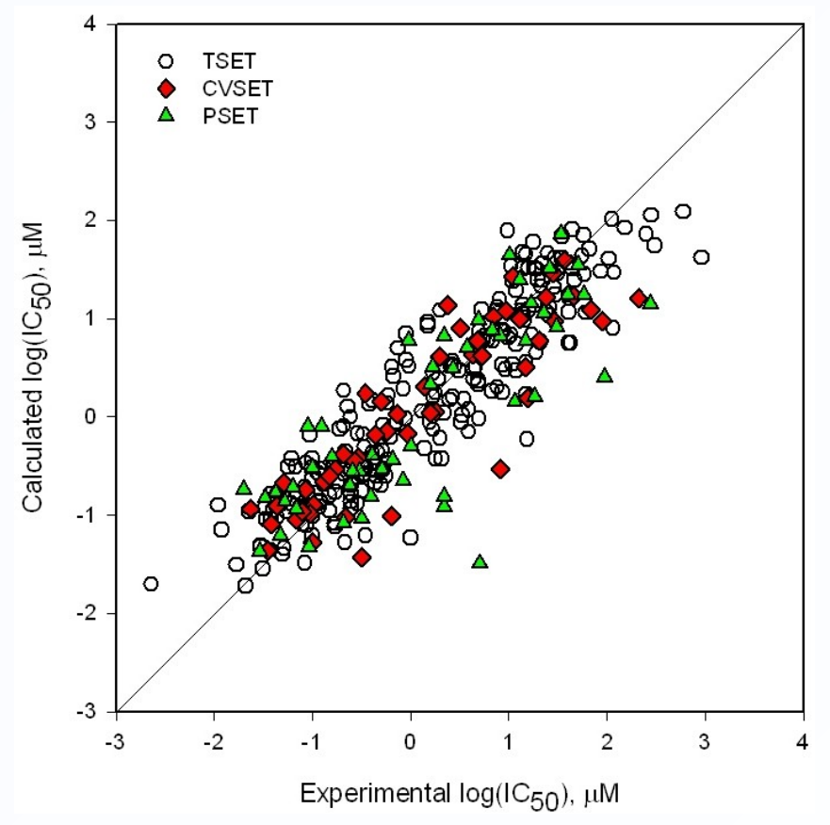
	Scrambled		MoRSE - WHIM	
	RMSE	R^2	RMSE	R^2
TSET	0.74	0.62	0.60	0.75
CSET	0.85	0.48	0.61	0.78
PSET	0.90	0.39	0.65	0.64

17

CNN Model Plots



Best SOM Based Model



Original Results

18

Conclusions

- The SOM appears to generate representative sets.
- This is evidenced by
  - More consistent statistics.
  - Smaller & simpler QSAR models.
- The SOM technique represents a more rational method QSAR set selection.