Thermodynamic profiling of protein-ligand binding energies

Application of machine learning methods in Bioinformatics

Chaitanya Koppisetty

Department of Computer Science and Engineering,
Chalmers University of Technology

and

Biognos AB

Overview

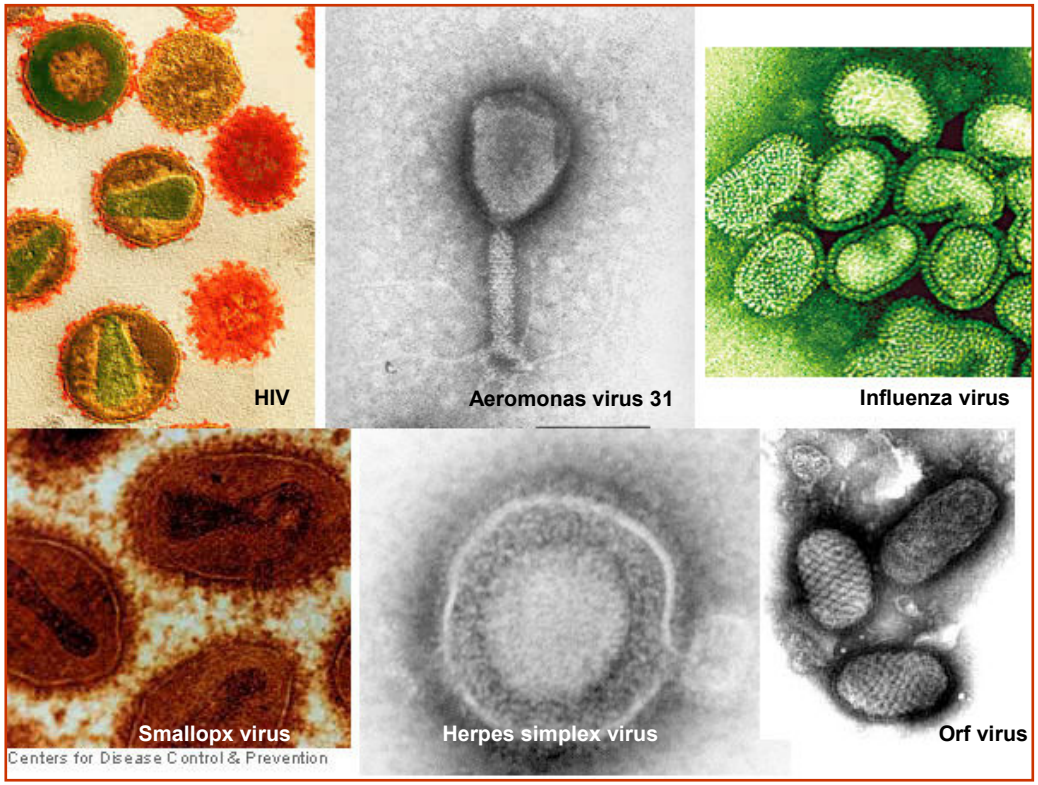
- Background
- Challenges
- Tools / methods
- Results / insights
- Summary

Background

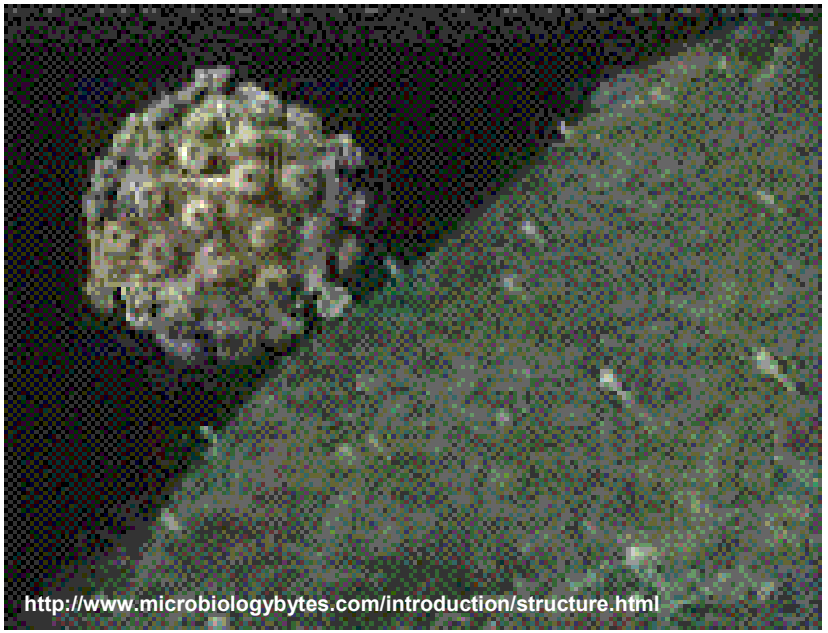
- Human body is constantly invaded by pathogens.

- ”Proteins” on the surface of pathogens are vital in adhesion and proliferation.

Background



Electron micrographs of viruses.

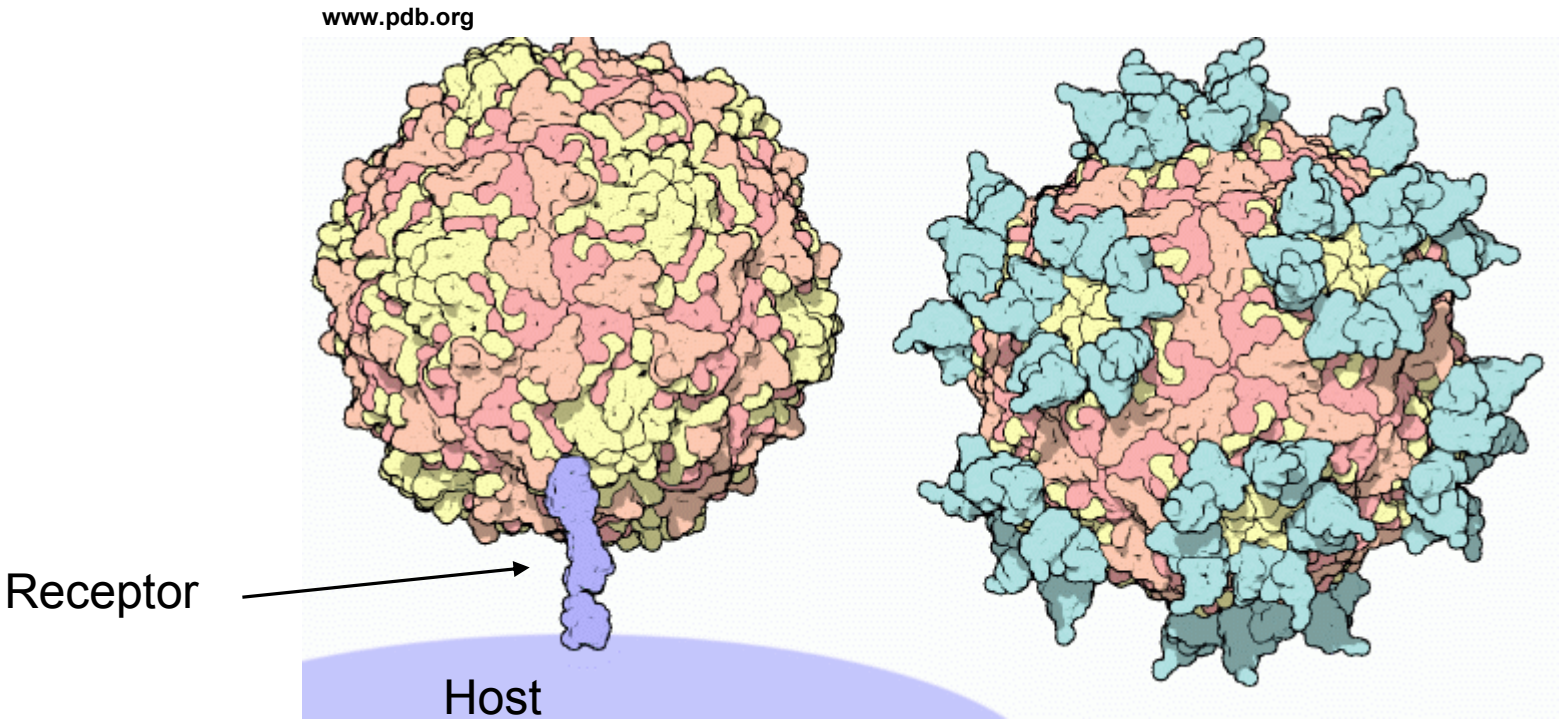


Background

- Drugs / Inhibitors / ligands
  - Small molecules that prevent the adhesion or proliferation of pathogens.
    - ”Relenza” is the trade name for infuenza virus inhibitors (ligand) that binds to a surface protein of influenza virus that stops proliferation.

Background

Inhibitors



Two important properties of a drug

- Affinity
- Specificity

Background

Affinity

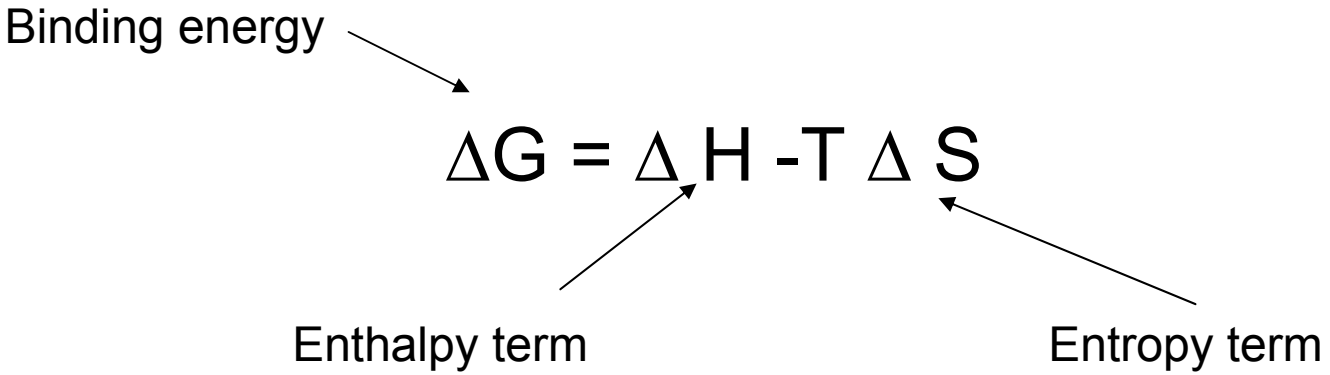
- How strong does a drug bind to the target.

Specificity

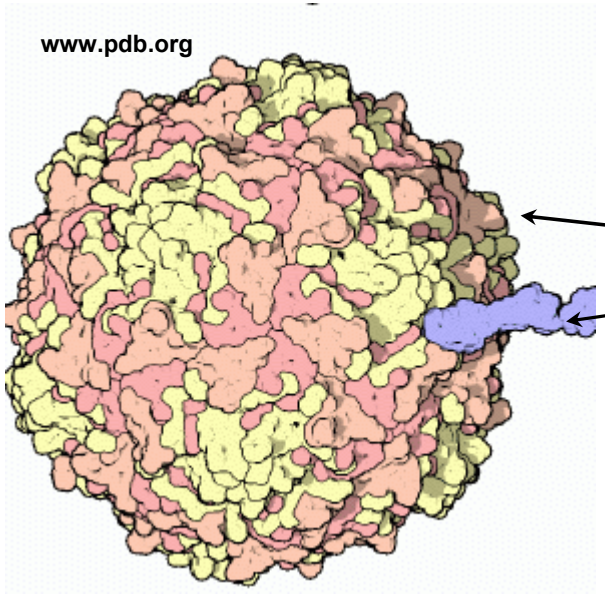
- How specific are the drug's interactions to that particular target
  - Is it binding to other proteins in the human body ?
  - Main cause of side-effects

Background

- Binding energy
  - Strength of interaction between the protein and ligand (negative value indicates binding)



Challenges



Scoring functions for estimating ΔG

Features / Descriptors based on physical and chemical properties

$$\Delta G_{Estimated} = \sum_{i=1}^{n} G_i f(Descriptor_i) \qquad eq\text{-}1$$

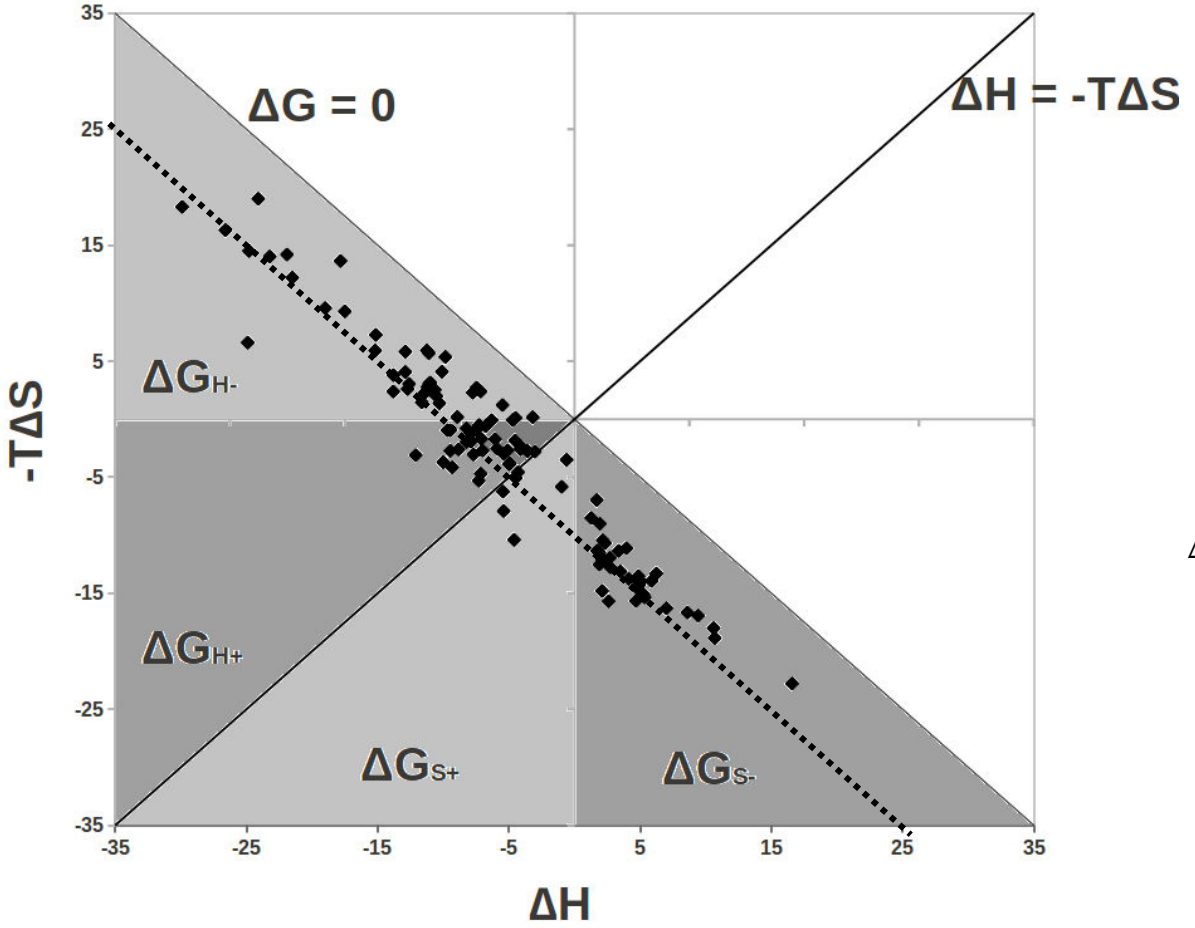
$$\Delta G_{Estimated} = \Delta H_{Estimated} - T\Delta S_{Estimated} \qquad eq\text{-}2$$

where

$$\Delta H_{Estimated} = \sum_{i=1}^{n} H_i f(Descriptor_i) \text{ and } T\Delta S_{Estimated} = \sum_{i=1}^{n} S_i f(Descriptor_i)$$

Challenges

$\Delta G = \Delta H - T\Delta S$



.....

$\Delta G = -10$ kcal/mol

Accurate estimation of ΔH and $T\Delta S$ is necessary for precise placement of ΔG case

Tools

Qualitative classification

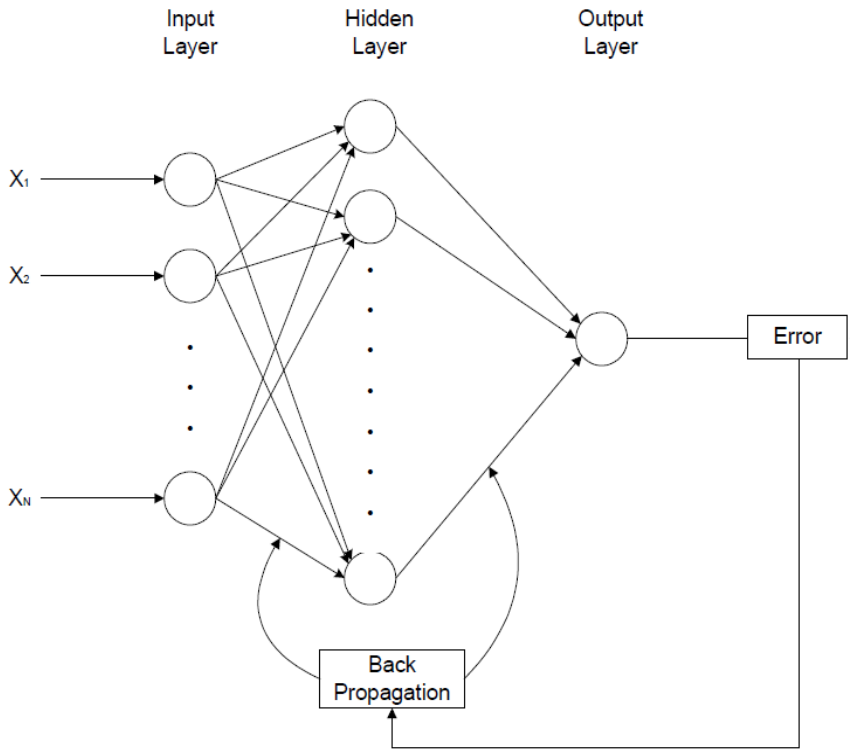
- Neural networks
- Support vector machines

Quantitative estimation

- Support vector machine regression

Tools

Neural networks (Multilayered perceptron)



Each node/neuron in hidden layer is a non-linear activation function

$$y_i = \frac{1}{1+e^{-S_i}}$$

Where,

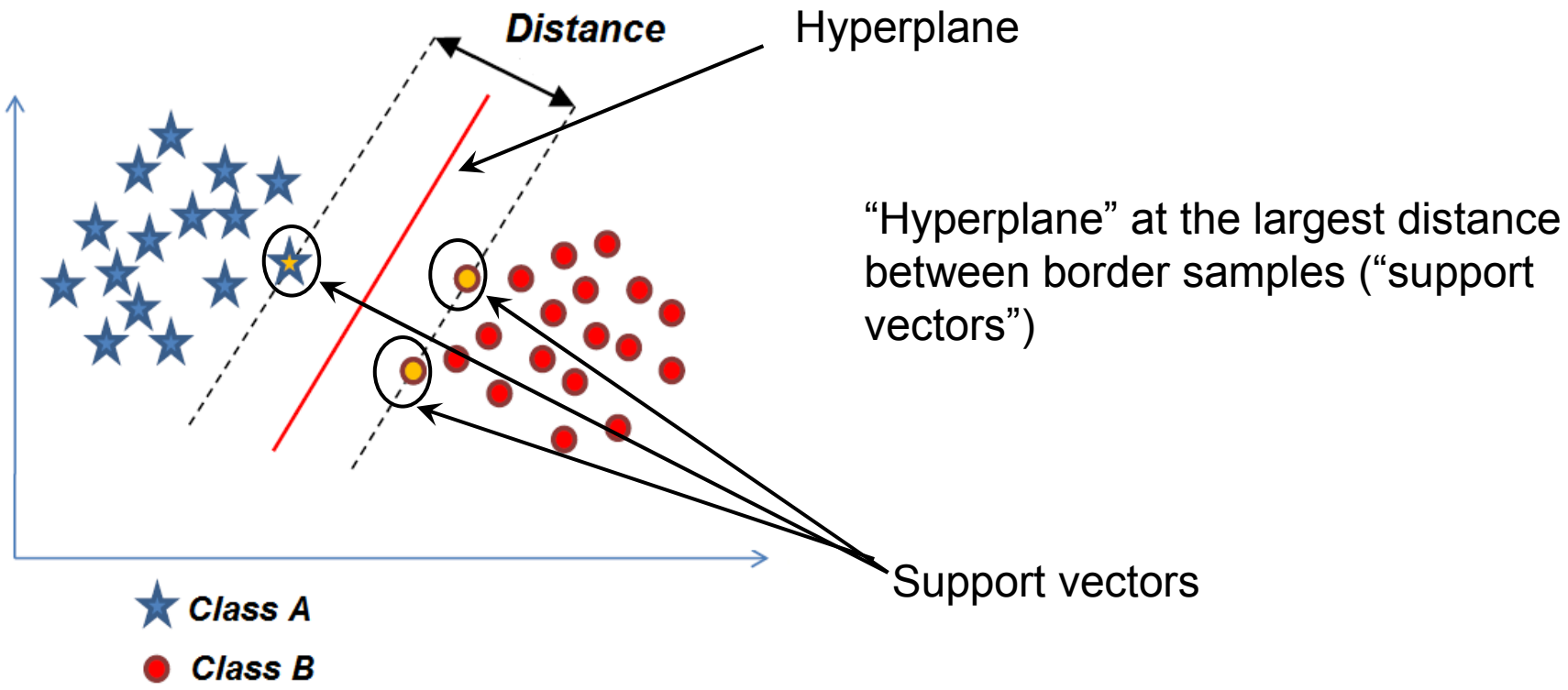
y_i is output of neuron i,

S_i is the weighted sum of all inputs and bias to neuron i

Error back-propagation algorithms

Tools

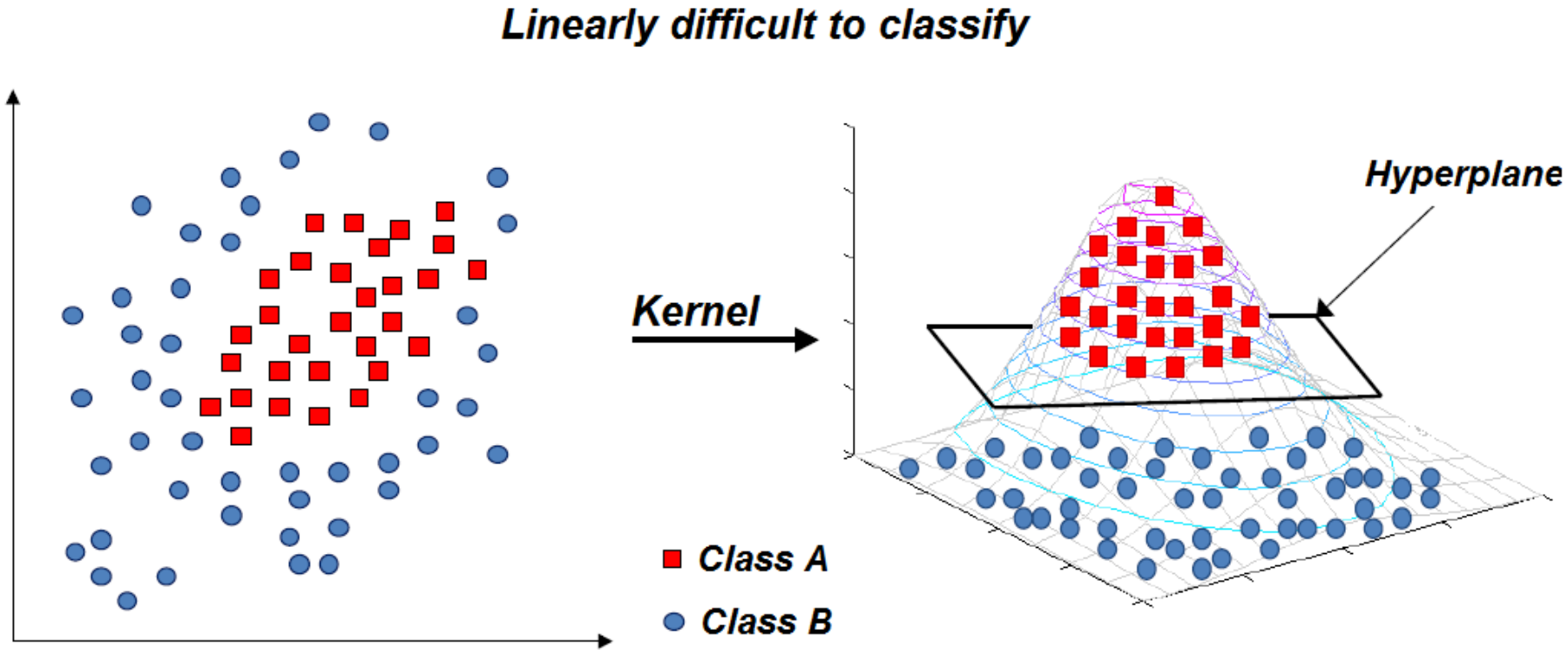
Support vector machines classification



"Hyperplane" at the largest distance between border samples ("support vectors")

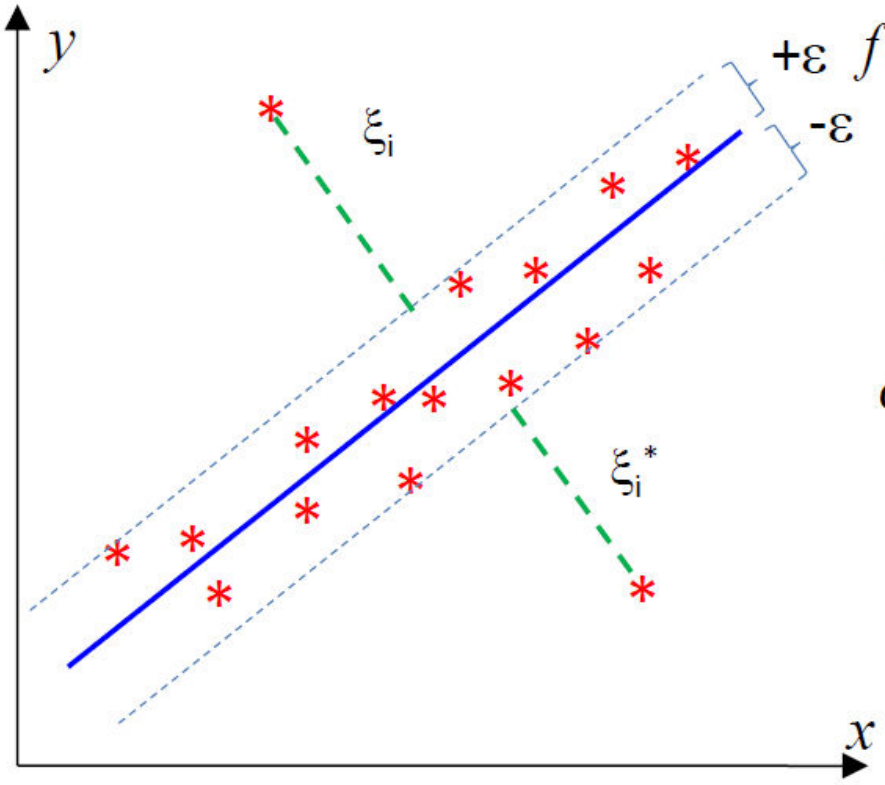
Tools

Support vector machines classification



Tools

SVM regression



$$f(\vec{x}) = \vec{w} \cdot \vec{x} + b$$

Minimize $\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right)$

Constraints

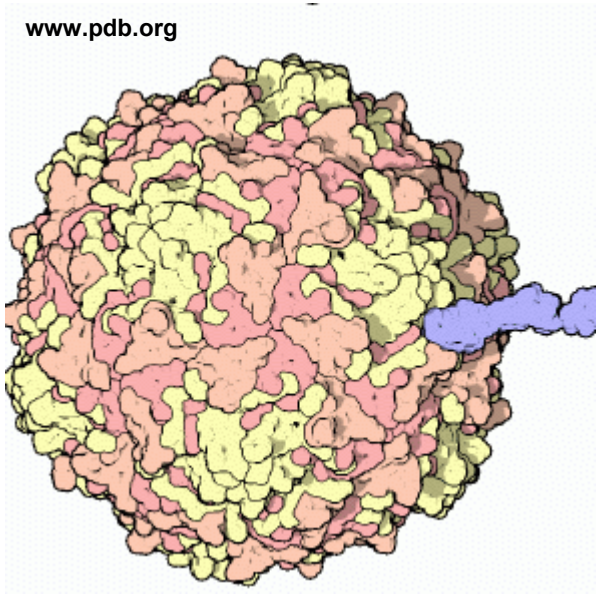
$$y_i - (\vec{w} \cdot \vec{x} + b) \le \varepsilon + \xi_i$$

$$y_i - (\vec{w} \cdot \vec{x} + b) \ge -\varepsilon - \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$

$$i = 1,\ldots,N.$$

Feature Generation



Features / descriptors based on structure based physico chemical properties of protein-ligand structures.

Training dataset of 120 protein-ligand structures.

Validation dataset of 1300 protein-ligand structures independent of training dataset.

220 Features / descriptors.

Descriptors grouped into 8 classes based on method of generation.

Feature Selection

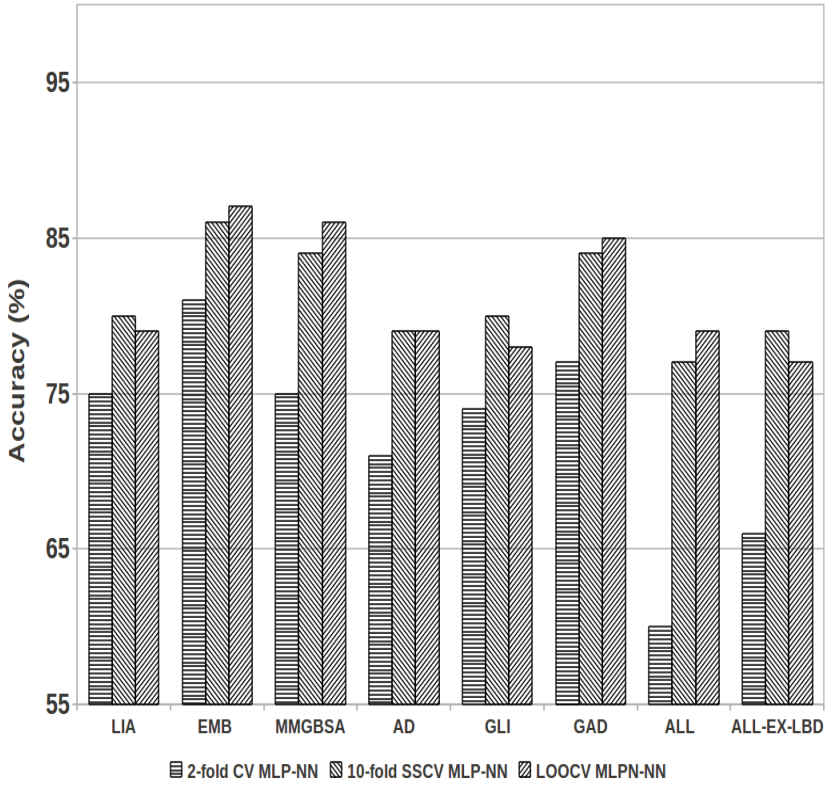
Feature selection / elimination

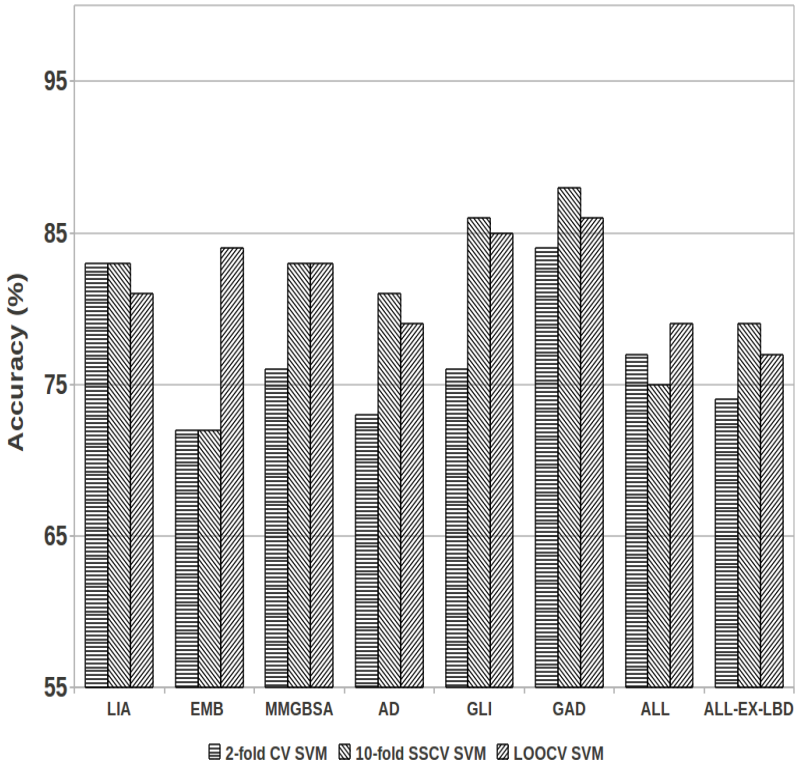
- Feature reduction using Principal Component Analysis (PCA)
  - Reduce the dimensionality of the data by fewer samples but still preserving the variance
- Backward feature elimination (BFE)

Cross validation

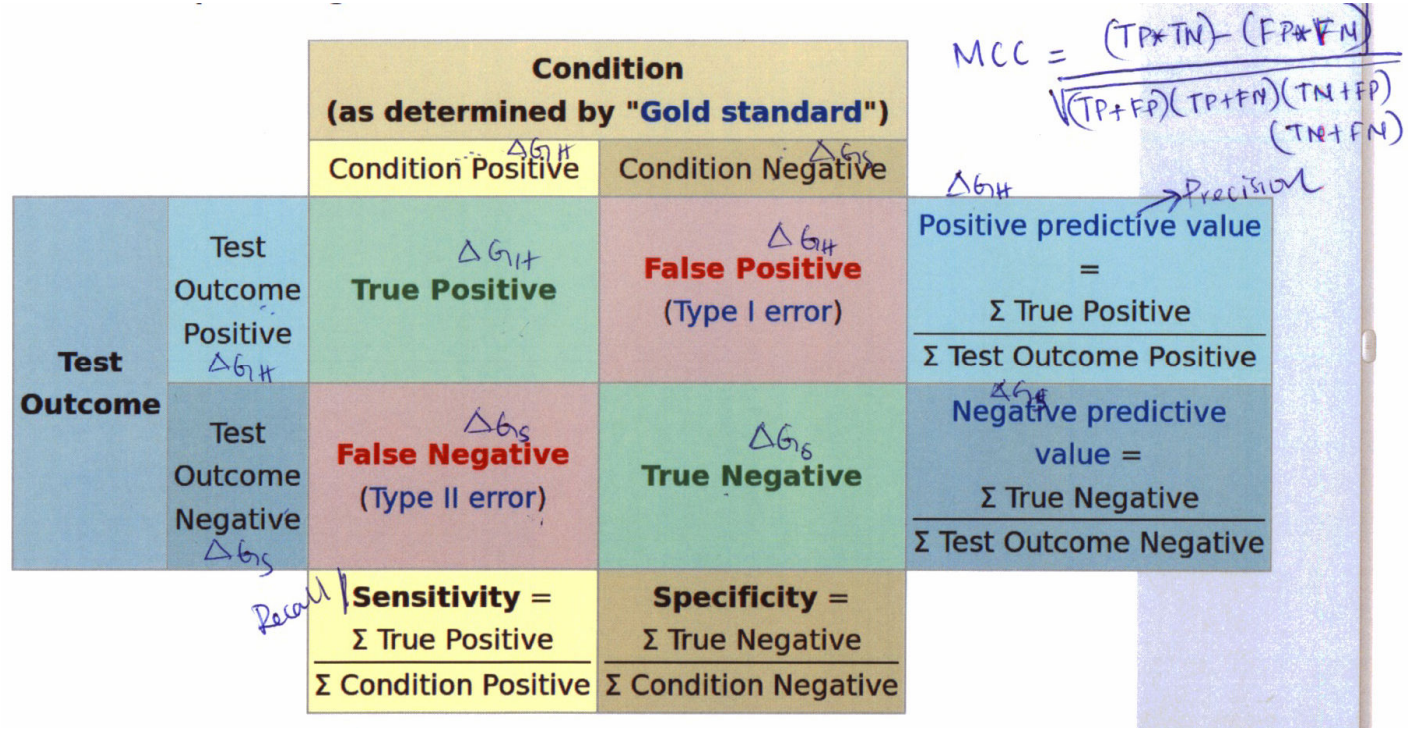
- 2 fold Cross validation
- Leave-one-out Cross validation
- N-fold Stratified sampling cross validation

Results Classification





Confusion matrix

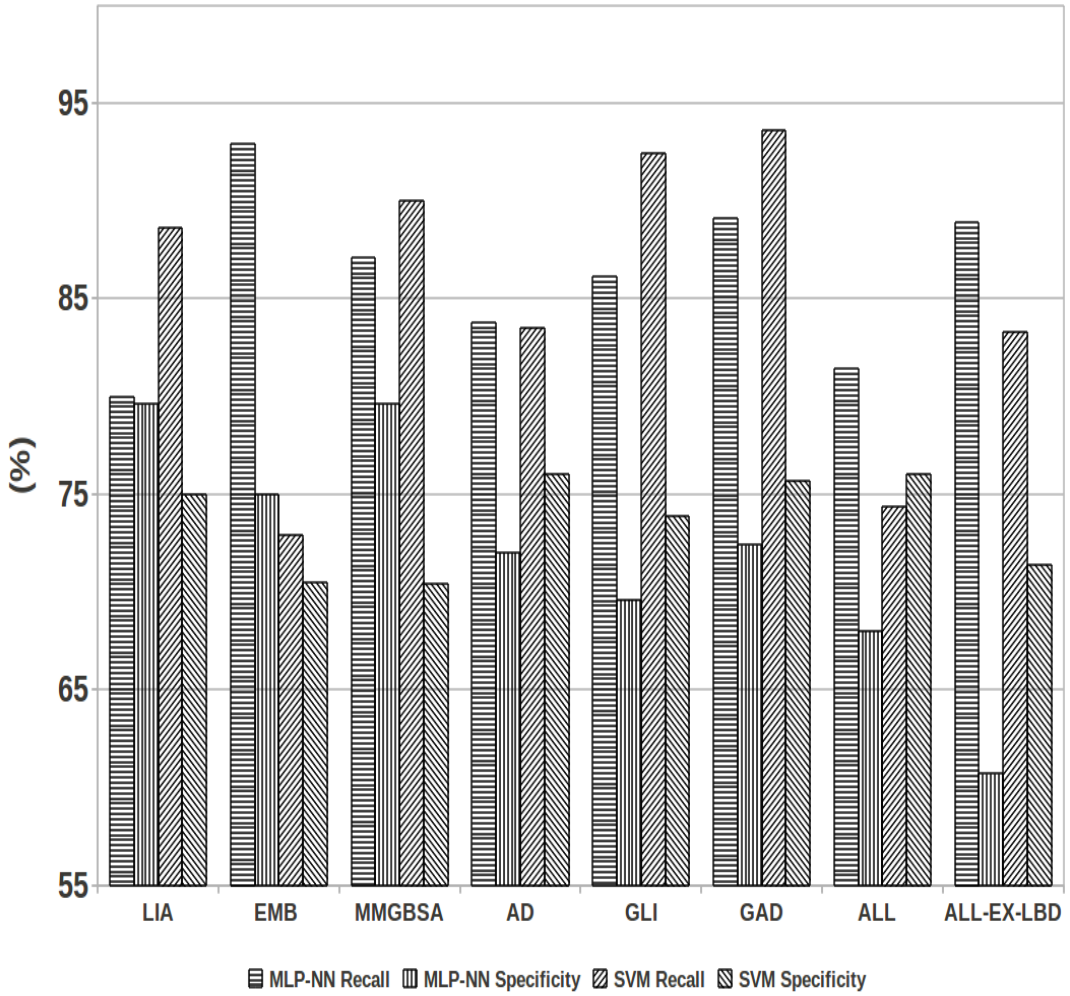


		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = Σ True Positive / Σ Test Outcome Positive
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = Σ True Negative / Σ Test Outcome Negative
		Sensitivity = Σ True Positive / Σ Condition Positive	**Specificity** = Σ True Negative / Σ Condition Negative	

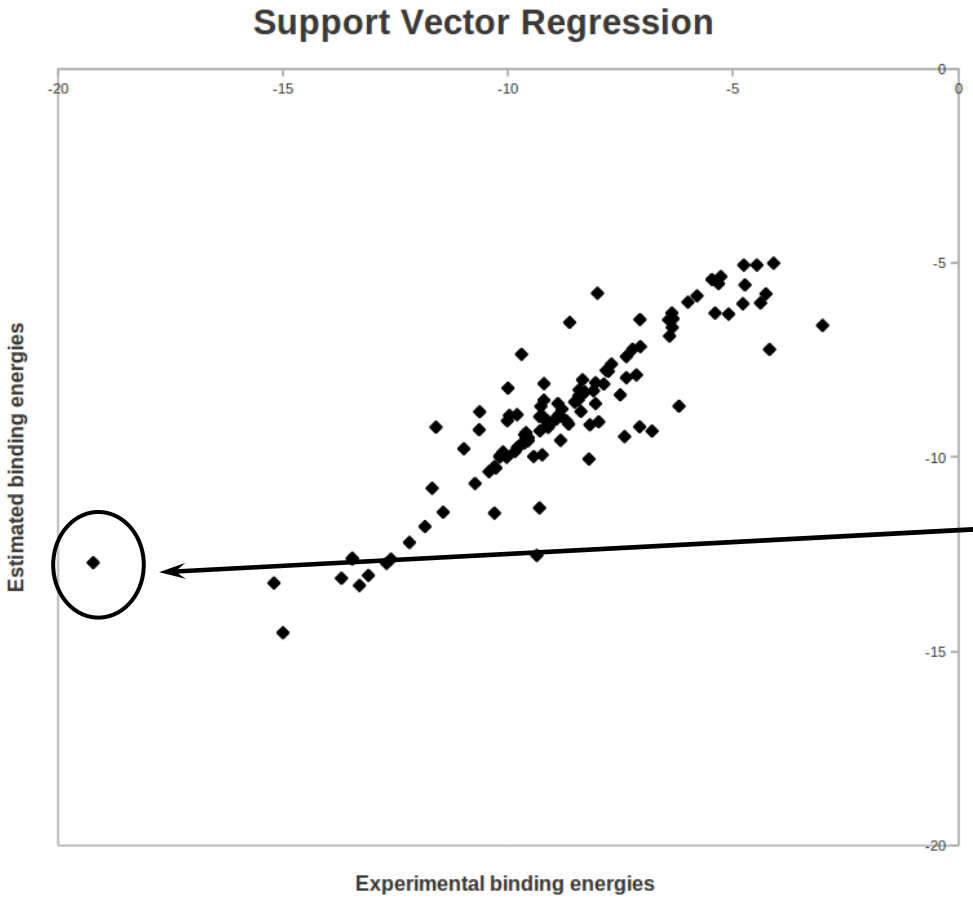
$$MCC = \frac{(TP \star TN) - (FP \star FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Source : Wikipedia

Results Classification



Results – SVR prediction models

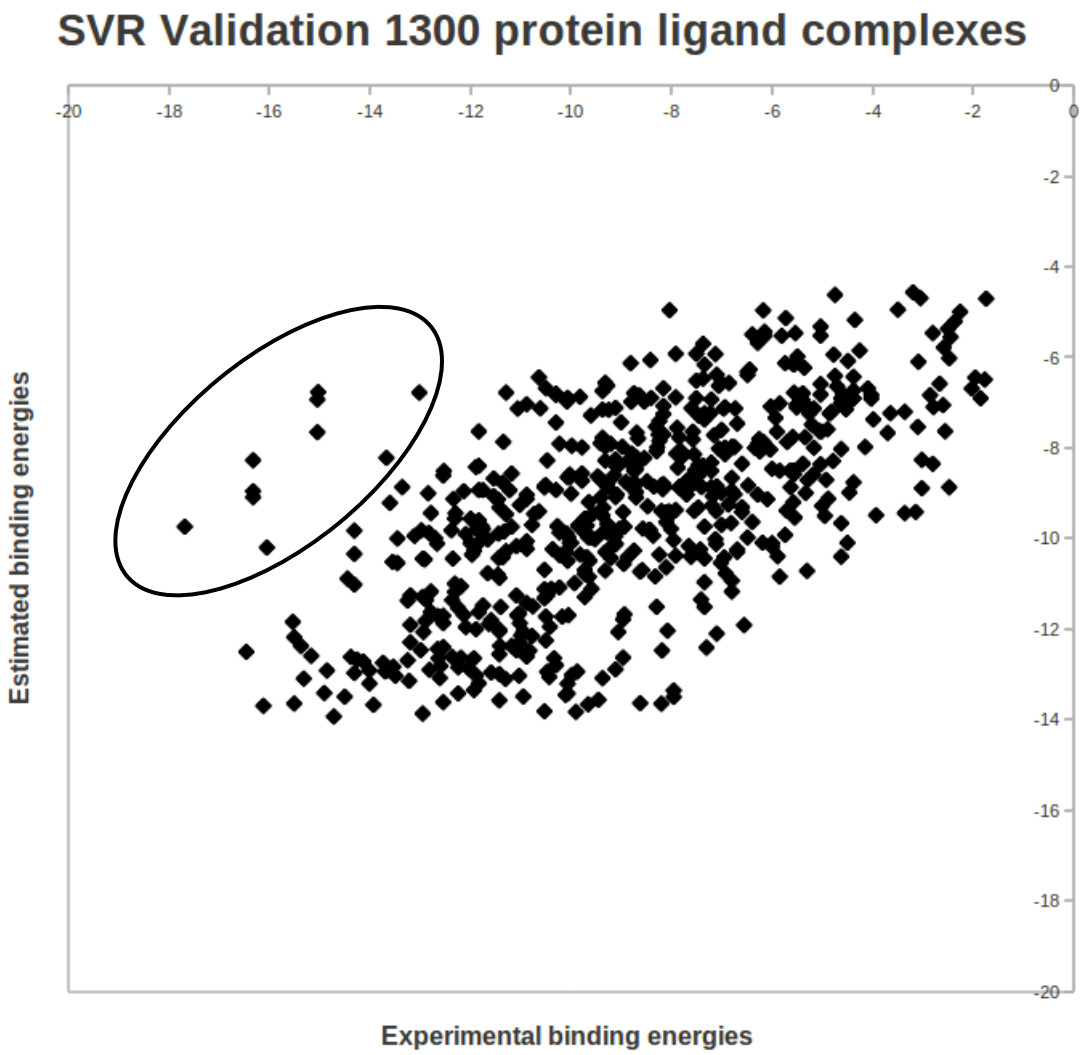


Training dataset

Things to consider

- Kernel and Hyper -parameters
- Cross validation methods
- Outlier analysis

Results – SVR prediction models



Validation dataset

- Outlier analysis
- Correlation coefficient and
- Standard error

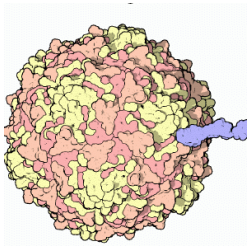
Application of SVR to real data

- Kernel choice and parameters
    - Linear
    - Polynomial (degree)
    - Gaussian (width parameter)
- Hyper-parameters
    - Parameter C
    - Parameter ε

Parameter optimization by cross-validation and grid search

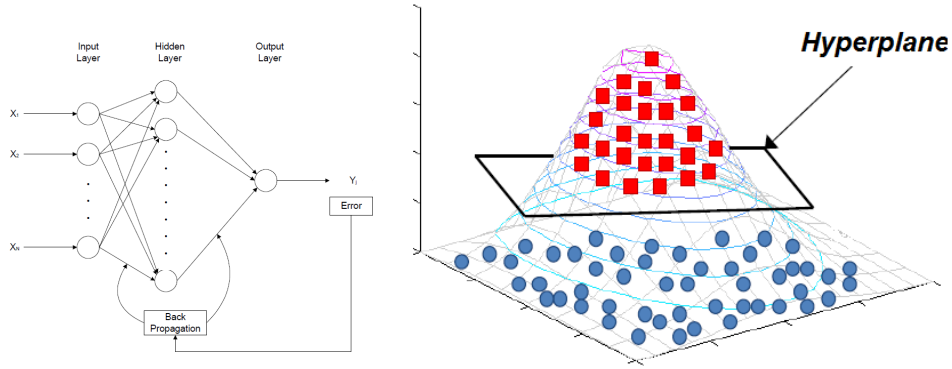
- Data Normalization (?)

Summary

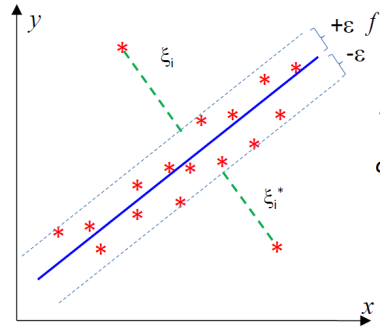


Training dataset of 120 protein-ligand structures, 220 Features / descriptors.

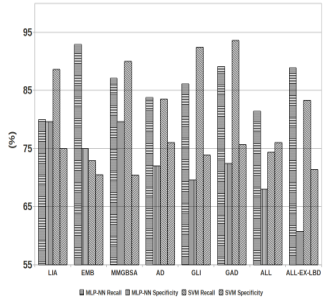
Validation dataset of 1300 protein-ligand structures.

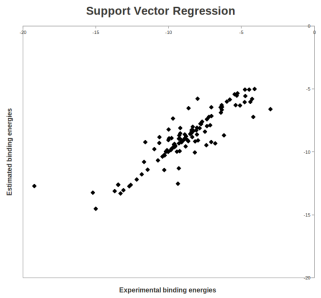


MLP-NN and SVM Classifiers



SVM regression models





Thanks for listening !