Workshop on Structure-based ligand design: *Lecture 4: Chemometric analysis for COMBINE*

Rebecca C. Wade

European Media Laboratory
Heidelberg

rebecca.wade@eml.villa-bosch.de
http://www.eml.org/english/Research/MCM

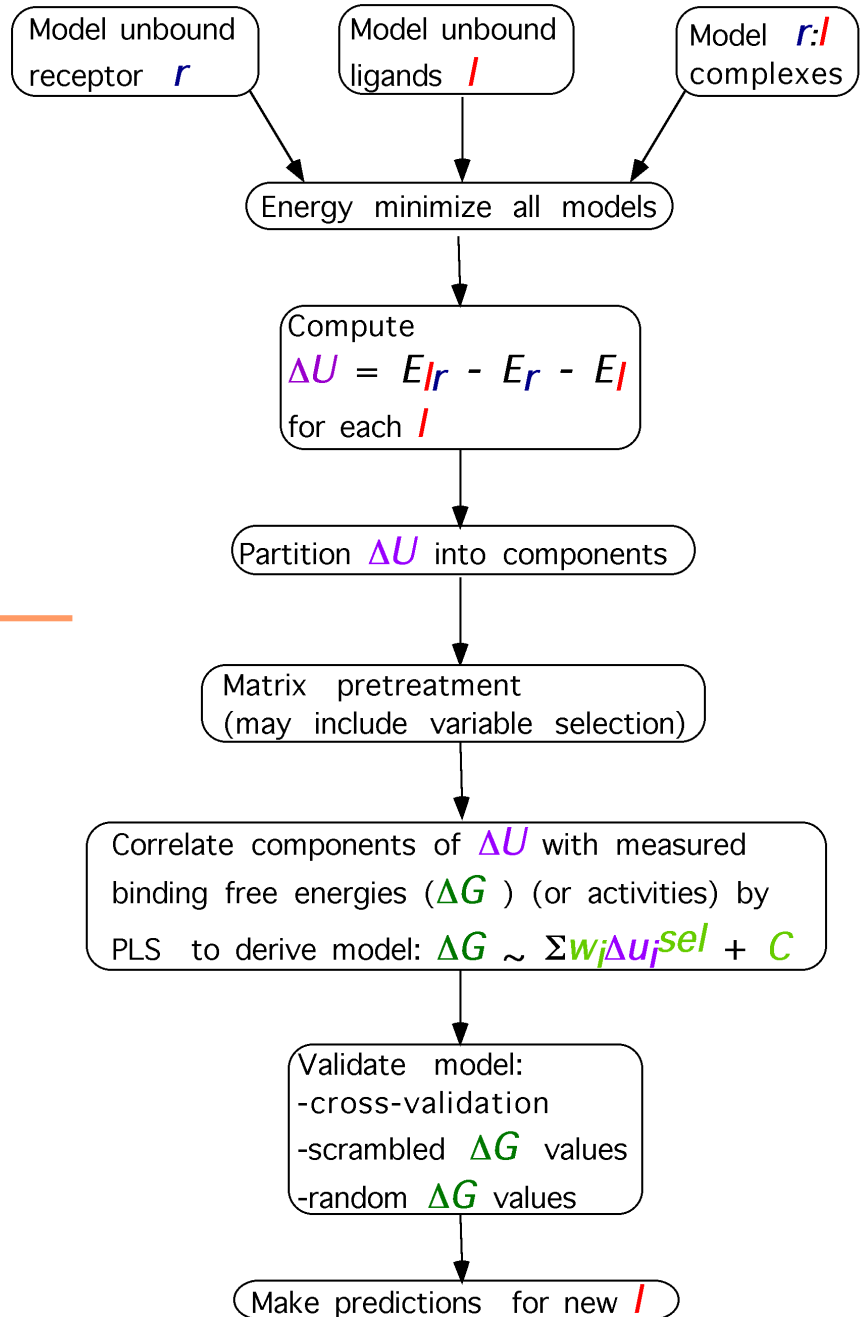
Workshop schedule

- **Lecture 1:** Introduction to Structure-based Drug Design, (GRID and 'flu)
- **Practical 1:** GRID
- **Lecture 2:** COMBINE Analysis overview
- **Lecture 3:** Molecular modeling for COMBINE
- **Practical 2:** COMBINE Analysis- molecular modeling
- **Lecture 4:** Chemometric analysis for COMBINE
- **Practical 3:** COMBINE Analysis- chemometrics
- **Lecture 5/Demo/Discussion/Practical 4**

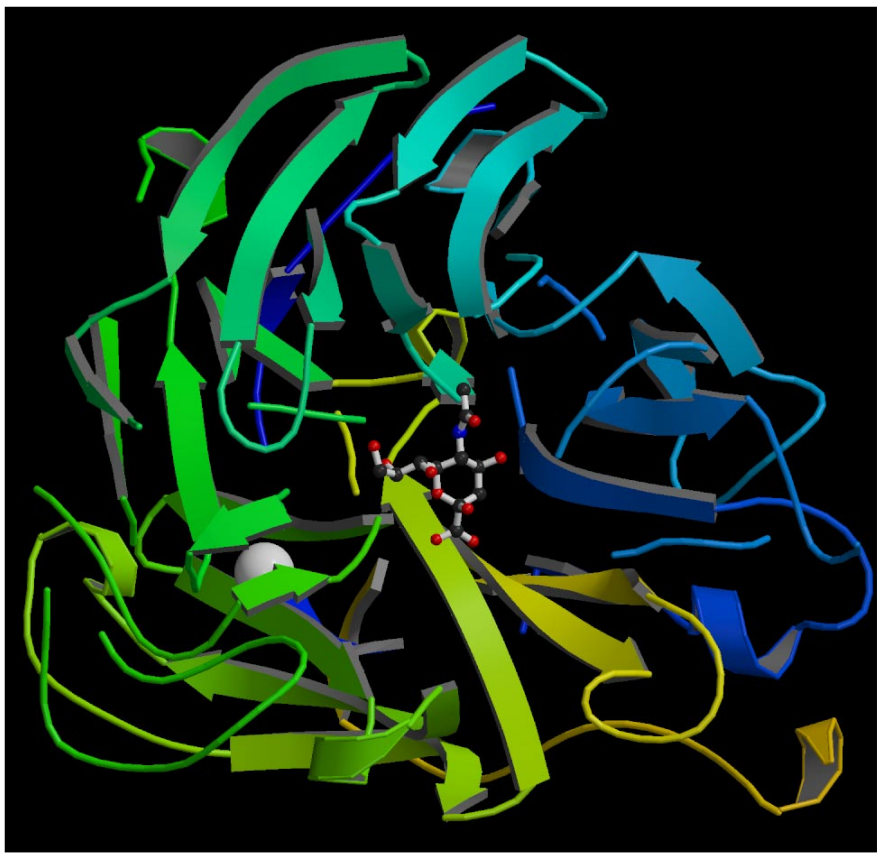
Lecture 4: Overview

- How can a COMBINE QSAR model be derived using the computed energy components?
  - Building an energy term/activity matrix
  - PCA: Principal Components Analysis
  - PLS: Projection to Latent Structures by means of Partial Least Squares
  - Variable selection
  - Model validation

Flowchart For Combine Analysis



Influenza neuraminidase inhibitors



- 43 complexes:
  - 29 inhibitors: sialic acid TS and benzoic acid derivatives
  - N9 + N2 subtypes + mutants
- 32 crystal structures
- 11 docked (comparative/AUTODOCK)
- Energy minimize: AMBER

Wang,T., Wade,R.C. *J. Med. Chem.* (2001) **44**, 961-971

Comparative Binding Energy (COMBINE) Analysis

- COMBINE: data + techniques
  - 3D macromolecular structure + experimental binding data
  - Empirical molecular mechanics energies + chemometric PLS

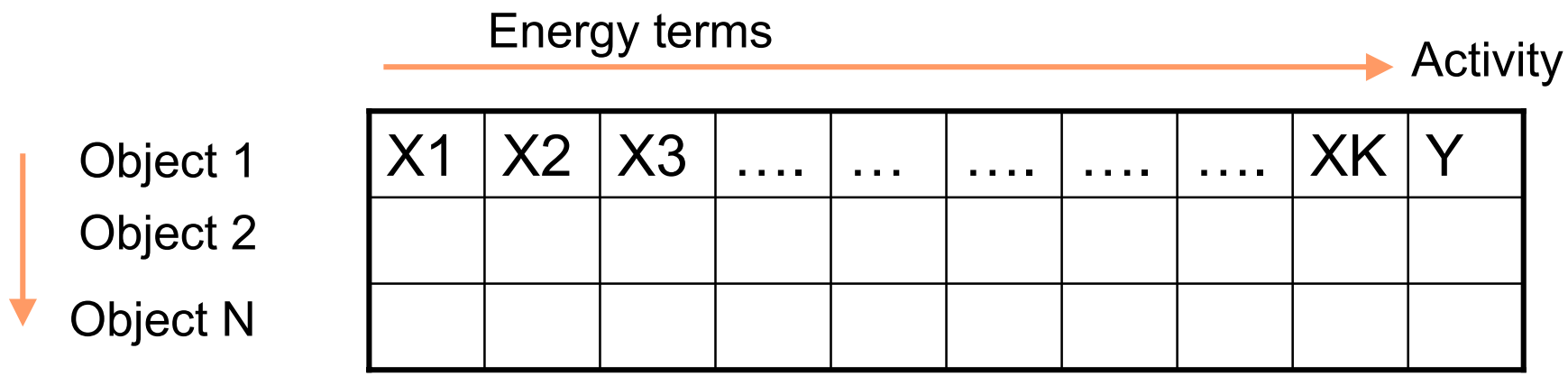
$$\Delta G = \sum_i w_i \Delta u_i + C$$

$$\Delta G = \sum_i w_i^{vdw} u_i^{vdw} + \sum_i w_i^{ele} u_i^{ele} + C$$

Ortiz,A.R., Pisabarro,M.T., Gago,F. Wade,R.C. *J. Med. Chem.* (1995) 38, 2681
Wade,R.C., Ortiz,A.R., Gago,F. *Persp. Drug. Disc. & Des.* (1998) 9, 19

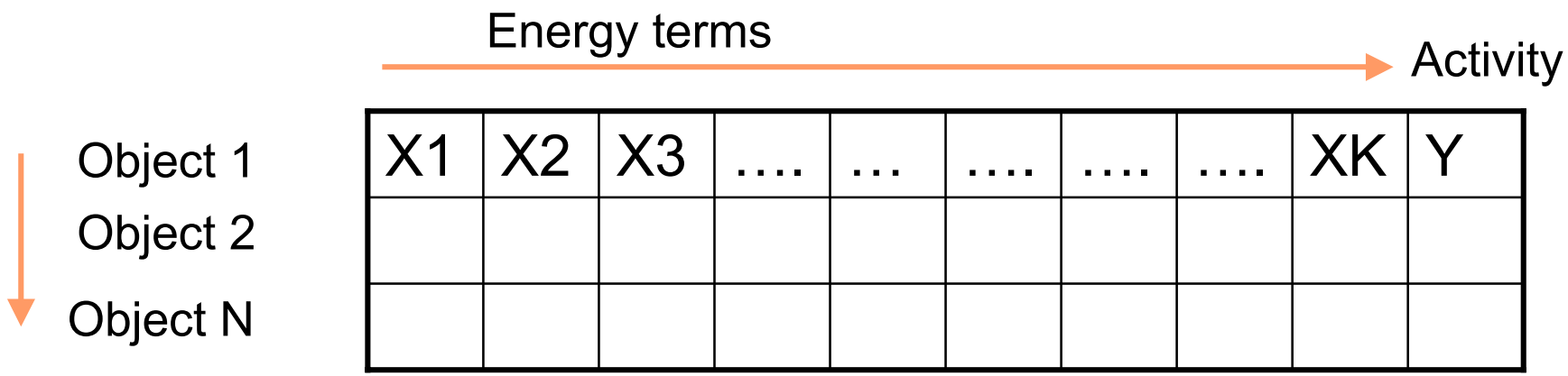
The data matrix

- For chemometric analysis, we need to construct a matrix of
  - X-variables: energy terms
  - Y-variable(s): Measured activity/binding affinity
- Our model will correlate selected, weighted X-variables with the Y-variable.



The data matrix: the problem

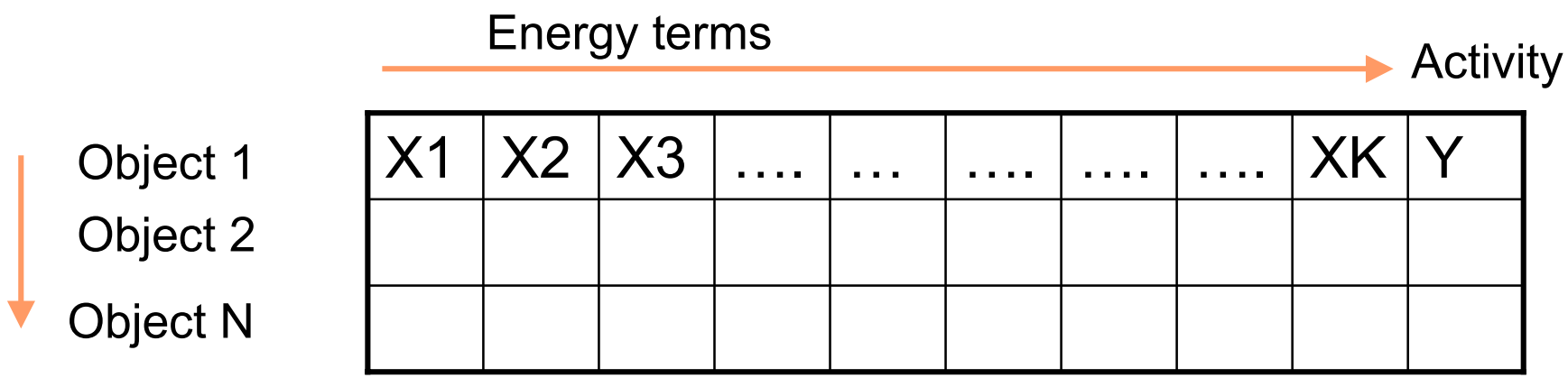
- **Typically:**
  - **100s of X-variables**
  - **1 Y-variable**
  - **10s of complexes**
- **Difficult to identify correlations by “looking”**



The data matrix: the problem

- **Classical regression techniques not suitable**
  - **Multiple Linear Regression (MLR) requires:**
    - **# objects N > 3x # variables K**
    - **Independent, uncorrelated X variables**

- **Therefore: Partial Least Squares (PLS)**



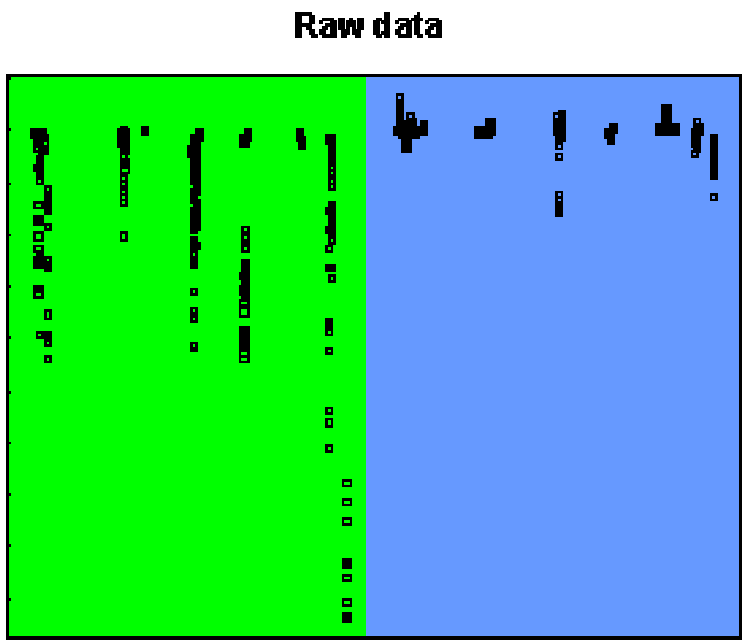
Data Preparation

- **Zeroing: Some X variables hardly vary across the dataset, so:**
  - **If their SD is very low, they are zeroed**
  - **If their magnitude is very low, they are zeroed**

Data Pretreatment

- **Scaling:**
  - **None – try it!**
  - **Autoscaling (full) – be careful!**
  - **Block unscaled weights (BUW) – usually most suitable!**
    - **E.g. give same importance to LJ block of X-variables as to electrostatic block of X-variables**

$$x_{ij} = \frac{x_{ij} - < x_j >}{\sigma_j}$$



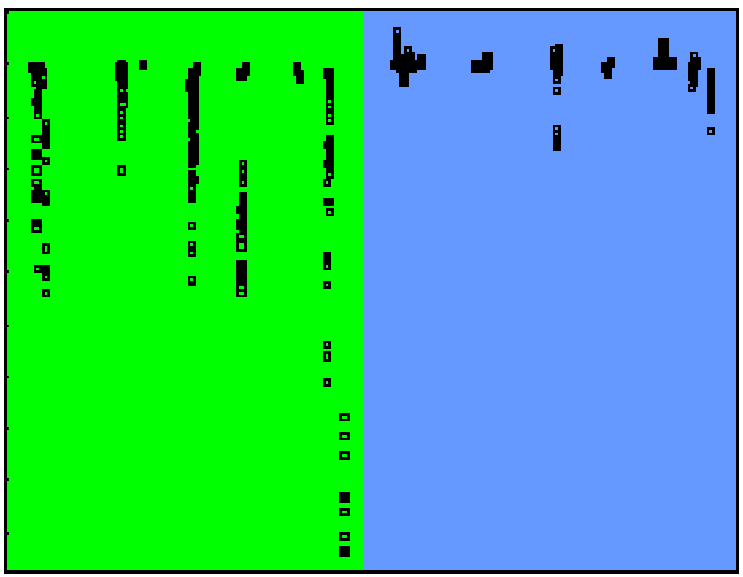
Golpe manual

Data Pretreatment: BUW

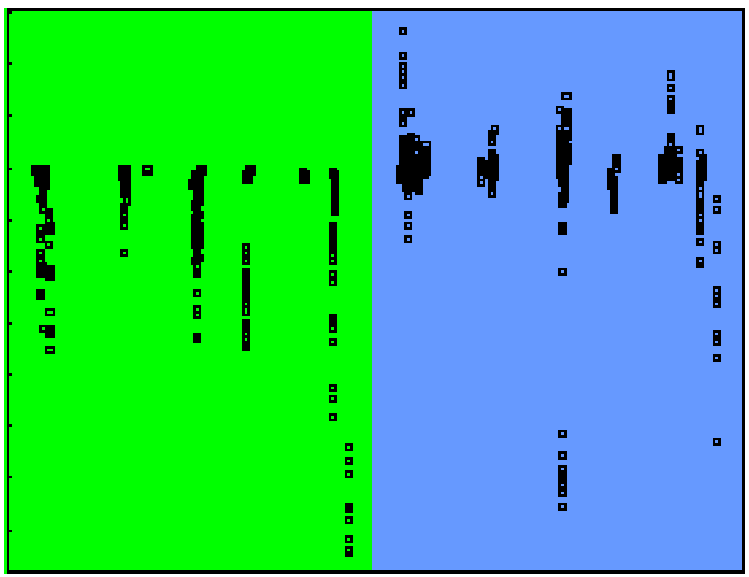
- **X-variables are scaled such that relative importance within a block, *k*, is maintained but overall the blocks have equal importance**

$$x_{ij,k} = \frac{x_{ij,k}.\sigma_X}{\sigma_k}$$

Raw data



Block Unscaled Weights (BUW)



Golpe manual

PCA: Principal Components Analysis

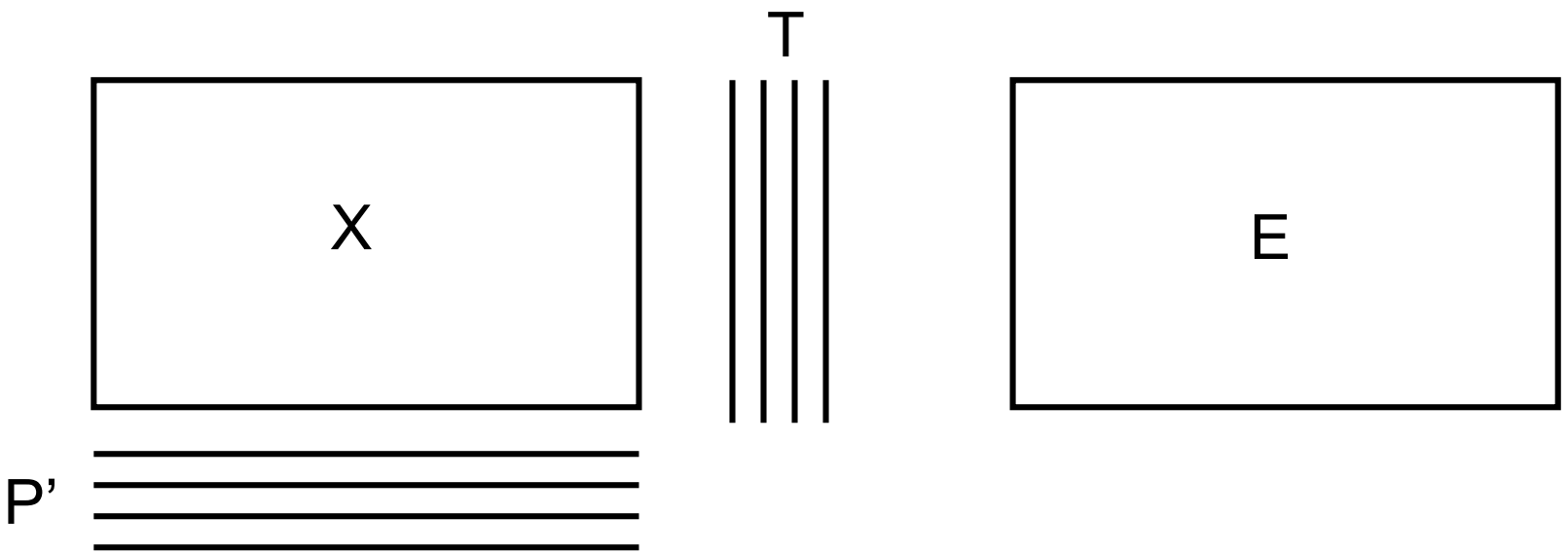
- How are my objects (complexes) distributed in X-variable (energy terms) space?
  - Are there clusters or outliers?
  - Is there good coverage?
  - Are there obvious patterns?
- Analysis of X-matrix only
  - Describe data in terms of a few new vectors:
    - principal components (PC)
    - PCs are linear combinations of original X-variables
    - PCs are orthogonal (uncorrelated)
    - PC1 is more important than PC2 etc

PCA: Principal Components Analysis

- Decompose X-matrix into 2 smaller matrices:
  - X = TP' + E
  - Loading matrix P
    - contains PCs (describing variables)
    - P’ is the transpose of P
  - Score matrix T
    - contains projections onto PCs (describing objects)
  - Residual matrix E
    - contains unexplained X-variance
- Choose how many PCs to extract (e.g. by CV)
  - Typically ca. 4 PCs explain ca. 90% of X-variance

PCA: Principal Components Analysis

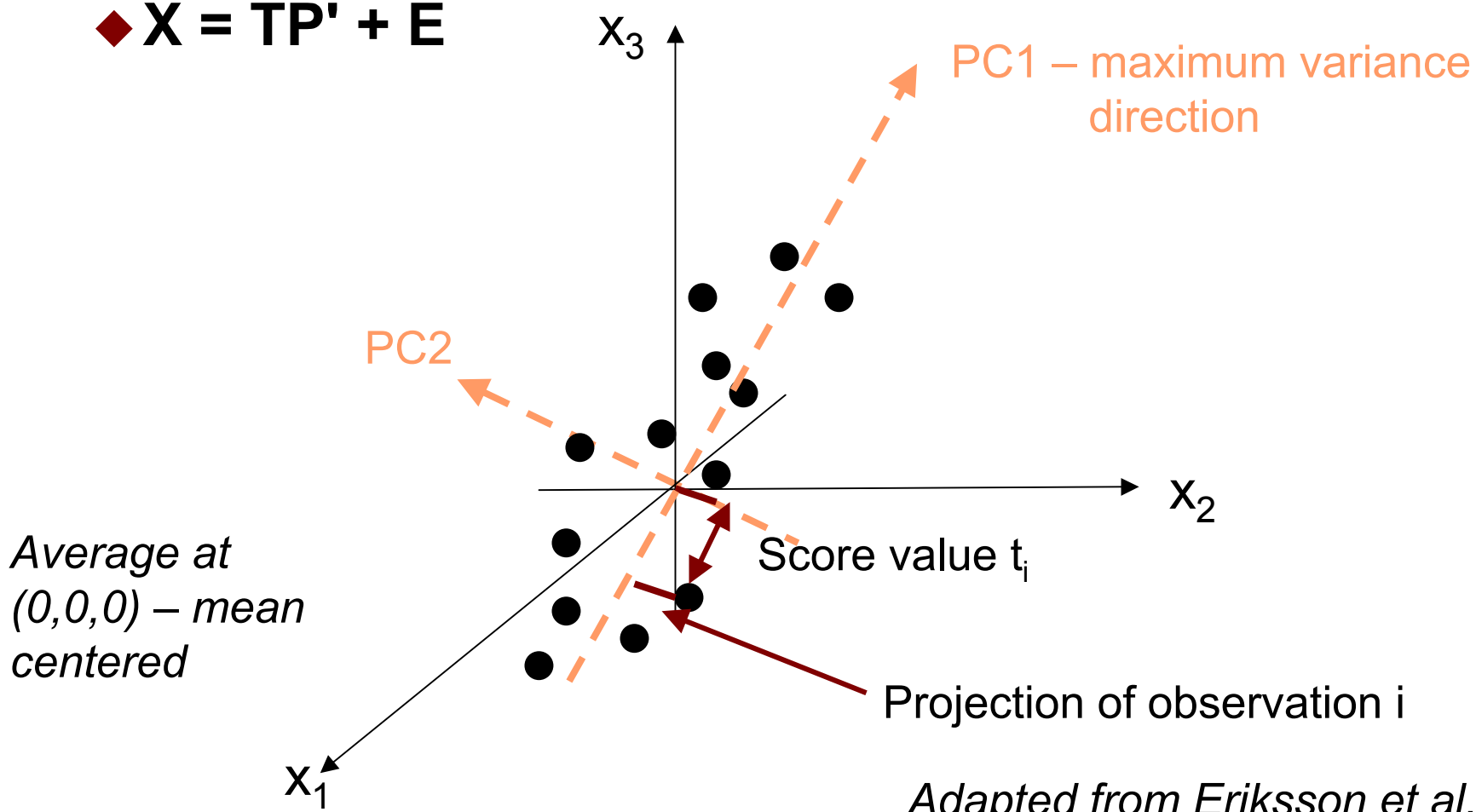
- **Matrices:**
  - **X = TP' + E**



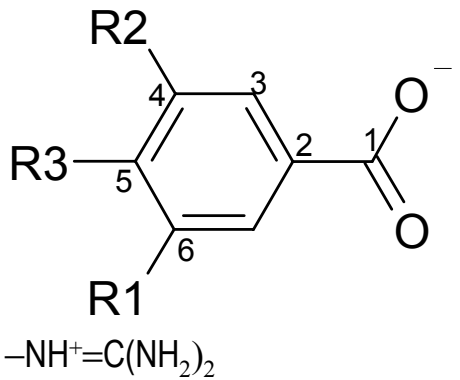
- **PCA = singular value decomposition (SVD)**
- **PCA = Eigenvector analysis**

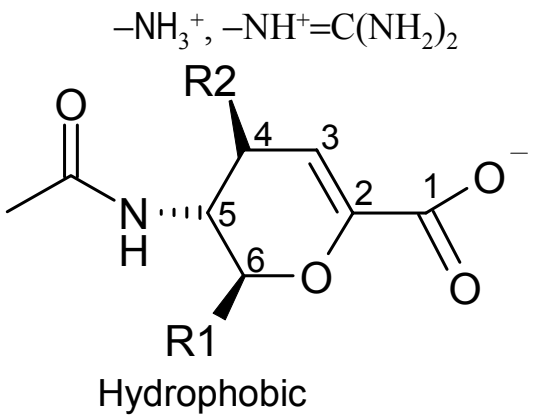
PCA: Principal Components Analysis

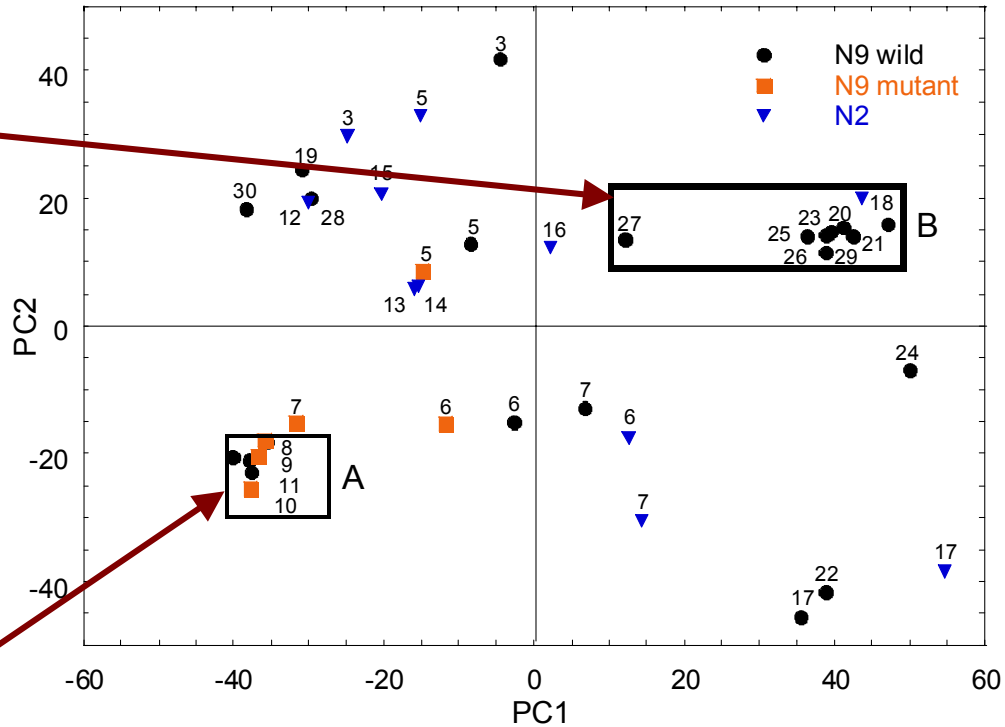
- **Matrices:**
  - **X = TP' + E**



PCA score plot: discrimination of influenza neuraminidase inhibitors







-BUW pretreated variables
-Compound classes distinguished
-No obvious distinction between N2 and N9

PLS: Partial Least Squares

- *Aim*: To correlate X-variables with Y-activity and produce QSAR by a linear multivariate model
- Y = f(X) + E
- Many different models may fit data
  - Need predictive models
  - Need to avoid over-fitting to Y-activity data that contain errors
  - "No regression model is better than the series it was obtained from"
    - Interpolate better than extrapolate

PLS: Partial Least Squares

- Decompose X-matrix into 2 smaller matrices (as for PCA):
  - X = TP' + E
  - Loading matrix P
    - contains Latent Variables, LVs (like PCs, describing variables)
    - LVs are orthogonal and numbered in order of decreasing importance
    - P' is the transpose of P
  - Score matrix T
    - contains projections onto LVs (describing objects)
  - Residual matrix E
    - contains unexplained X-variance

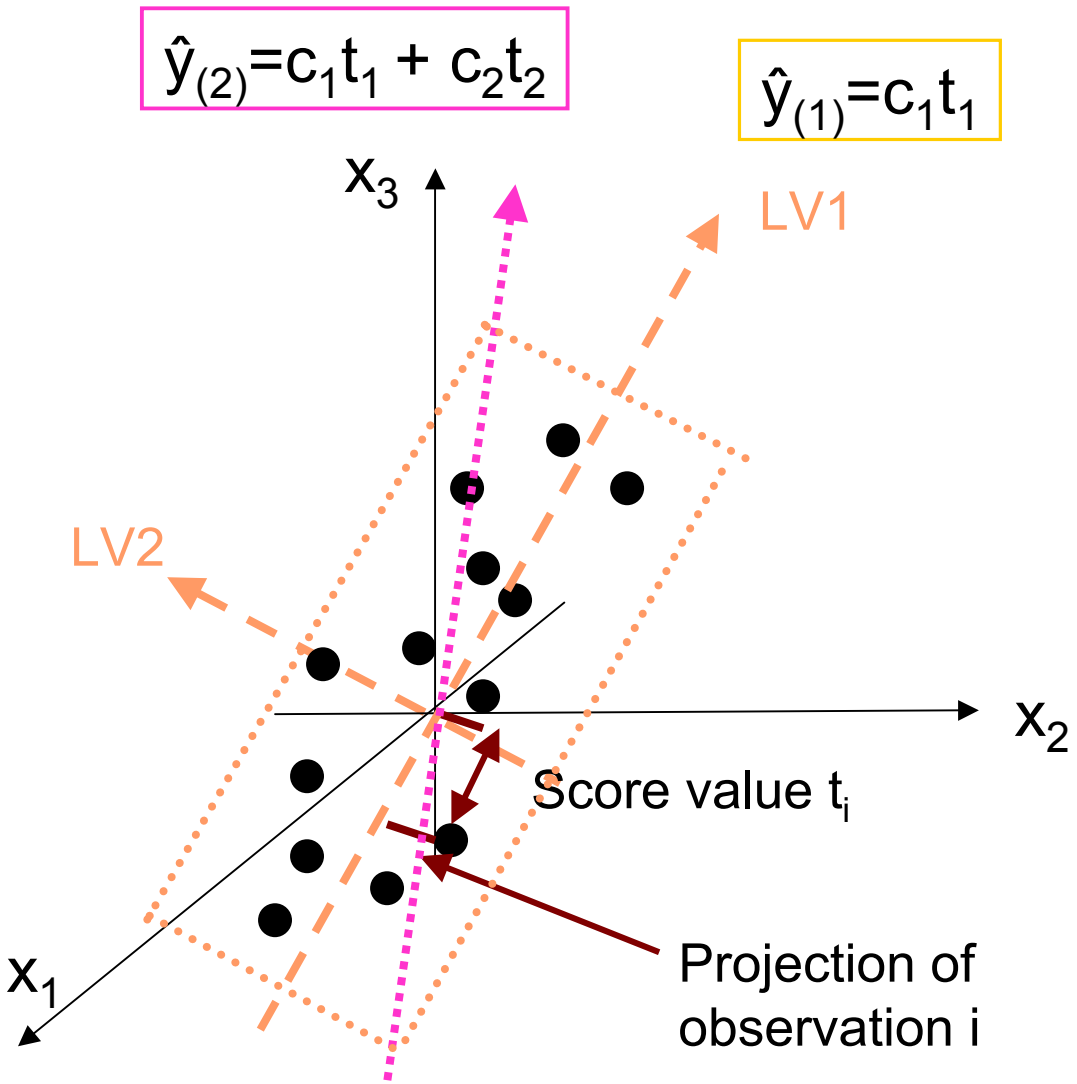
PLS: Partial Least Squares

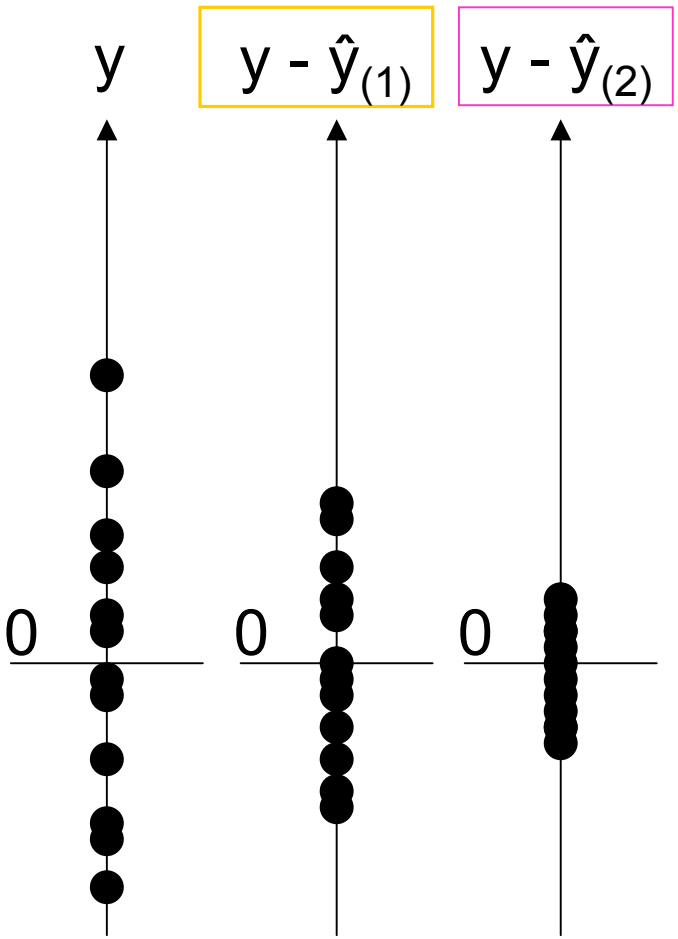
- **PLS vs PCA:**
  - **PCs in PCA represent structure of X-matrix only**
  - **LVs in PLS are obtained under 2 constraints via an iterative process:**
    - **Must represent structure of X- and Y-matrices**
    - **Must maximize fitting between X's and Y's**

PLS: Partial Least Squares

- **Matrices of final PLS model:**
  - **X = TW' + E**
    - **Xs decomposed to X-scores T and X-weights W**
  - **Y = UC' + F**
    - **Ys decomposed to Y-scores U and Y-weights C**
  - **U = T + H**
    - **X-scores correlate with Y-scores (inner relation)**

PLS: Partial Least Squares





Adapted from Eriksson et al.

Variable Selection

- **Aim:**
  - **eliminate noisy X-variables**
  - **simplify PLS models**

- **GOLPE: Generating Optimal Linear PLS Estimations**
  - **Build initial PLS model, select optimal dimensionality**
  - **Optional : D-optimal preselection**
  - **Build design matrix. Using fractional factorial designs, evaluate individual contribution of each variable to predictive ability of model**
  - **Remove X-variables that do not contribute**
  - **Compute new PLS model**

D-optimal Design

- Aim:
  - Preselection to remove variables containing little or redundant information
- Criterion: position in Loading space
  - Remove variables from center of PLS loadings (weights)
- User-defined number of variables
  - Stepwise
  - E.g. retain half and rebuild PLS model
  - Ensure no deterioration of model
- D-optimal preselection may not be necessary

Fractional Factorial Design (FFD)

- **Aim:**
  - Remove variables that do not assist prediction
- **Criterion: Predictive ability of PLS model**
  - Build large number of "reduced" PLS models each with different sets of X-variables removed
  - Evaluate predictive ability by CV
  - Remove X-variables and rebuild PLS model
  - Repeat as necessary
- **Design-matrix:**
  - X-variable combinations
  - Dummy variables – significance of real variables
    - E.g. Real: dummy variables: 4:1
  - X-variables assigned as:
    - Fixed – decrease SDEP, increase predictive ability
    - Excluded – increase SDEP, decrease predictive ability
    - Uncertain

Cross-validation (CV)

- Choose the optimal number of LVs for the PLS model
- Evaluate the quality of the model

- In “internal” CV,
  - A reduced model is built
    - I.e. a model derived with some of the objects removed
  - The reduced model is used to predict the Y values of the removed objects
  - The predicted and experimental Y values are compared The process is repeated several times with different models with different LV dimensionality and with different objects removed.

PLS Model quality measures

- **SDEP: Standard Deviation of Errors of Prediction**

$$SDEP = \sqrt{\sum \frac{(Y - Y')^2}{N}} = \sqrt{\frac{PRESS}{N}}$$

- **Q^2: Predictive correlation coefficient**

$$Q^2 = 1 - \left[\frac{\sum (Y - Y')^2}{\sum (Y - \overline{Y})^2} \right]$$

Y: Experimental value
Y': Predicted value
$\overline{Y}$: Average value
N: Number of objects

- **Similarly for fitting:**
  - **R^2 : Fitting correlation coefficient**
  - **SDEC**

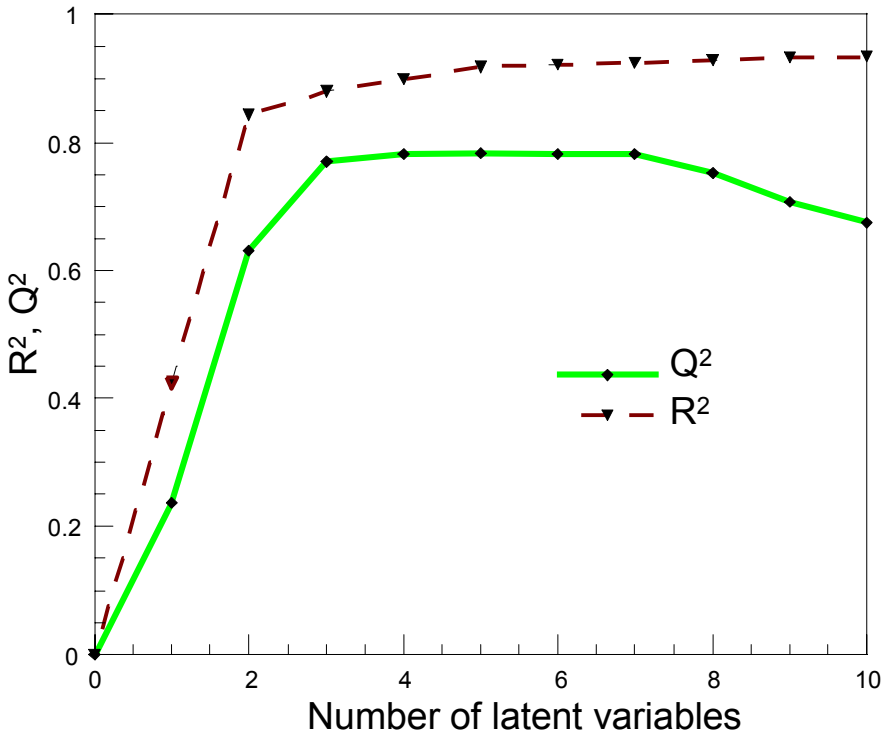
Cross-validation (CV)

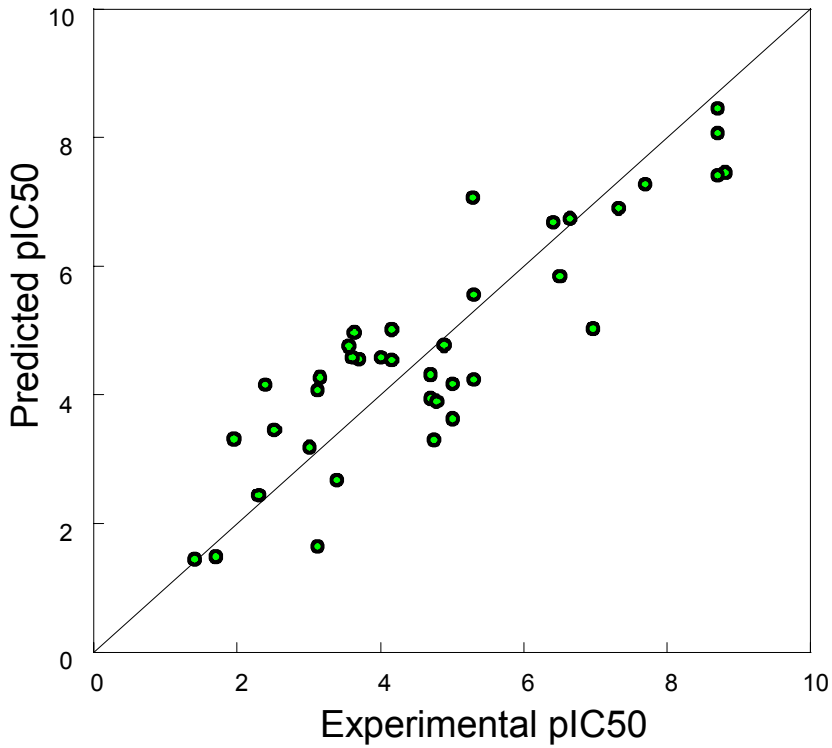
- Choice of groups for building reduced models
  - No standard
  - Should delete each object once over model ensemble
    - LOO: Leave-one-out
      - Do for all objects
    - LTO: Leave-two-out
    - Fixed groups:
      - Approx equal size (e.g. 5 objects), do for each group
    - Random groups:
      - Approx equal size, do for each group, reassign groups, repeat
  - Predictive measures are generally better for LOO than LTO than fixed/random groups

Validation of a QSAR model

- **Requirements for model validity:**
  - **Internal cross-validation: $Q^2 > 0.3$, SDEP**
    - LOO, LTO, LSO
    - Fixed or random groups, e.g. of 5
  - **Robustness to randomization of Y-values**
    - Random values of Y
    - Scrambled values of Y
    - Neither should produce significant models
  - **Predictions for external test set:**
    - Q^2, SDEP
    - Plot predicted vs experimental to spot outliers

Influenza neuraminidase inhibitors COMBINE model





N=39 (N2 + N9), 3LV,
R^2=0.88, Q^2=0.77,
SDEP=0.96, SDEPext =1.2

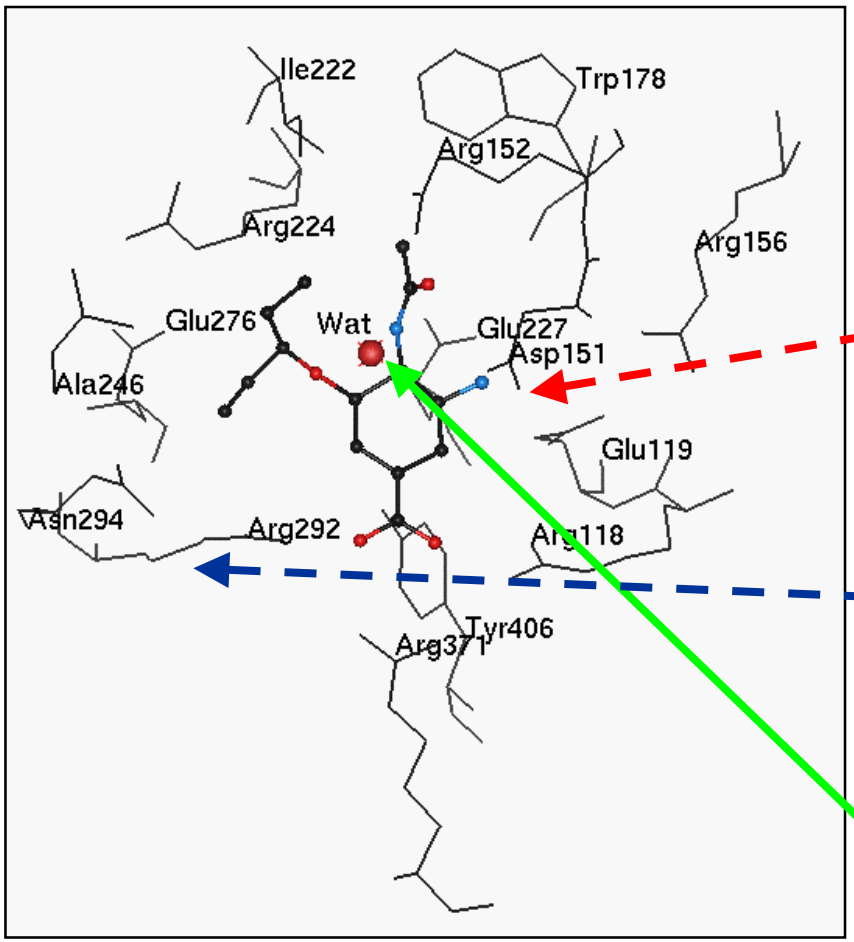
Validation of docked ligand alignments

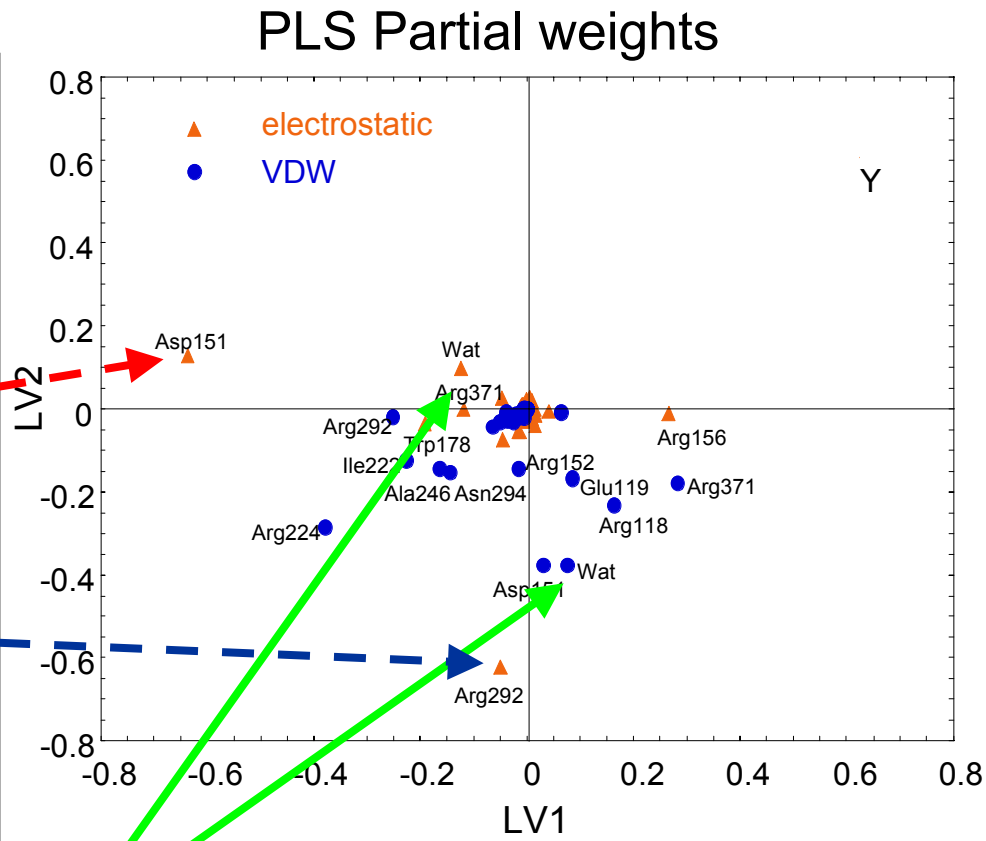
Influenza neuraminidase inhibitors COMBINE model

Data set	#Complexes	# LV	R^2	SDEC	Q^2	SDEP	$SDEP^{ext}$ ±SD
N2+ N9	39	**3**	**0.88**	**0.69**	**0.77**	**0.96**	**1.22 ± 0.20**
		4	0.89	0.64	0.78	0.94	1.20 ± 0.22
		5	0.92	0.57	0.78	0.94	1.19 ± 0.21
N9	28	3	0.91	0.56	0.76	0.93	
		4	**0.94**	**0.45**	**0.84**	**0.76**	
		5	0.95	0.42	0.85	0.72	

Abbreviations: LV, latent variable; R^2 , correlation coefficient; SDEC, standard deviation of errors of correlation; Q^2 , predictive correlation coefficient; SDEP, standard deviation of errors of prediction; $SDEP^{ext}$, average SDEP of 15 external validation tests (10 complexes removed); SD: standard deviation.

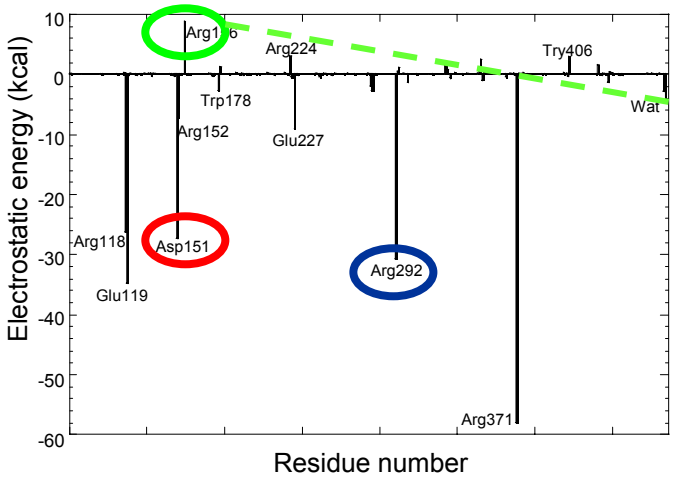
Important contributions to neuraminidase inhibition

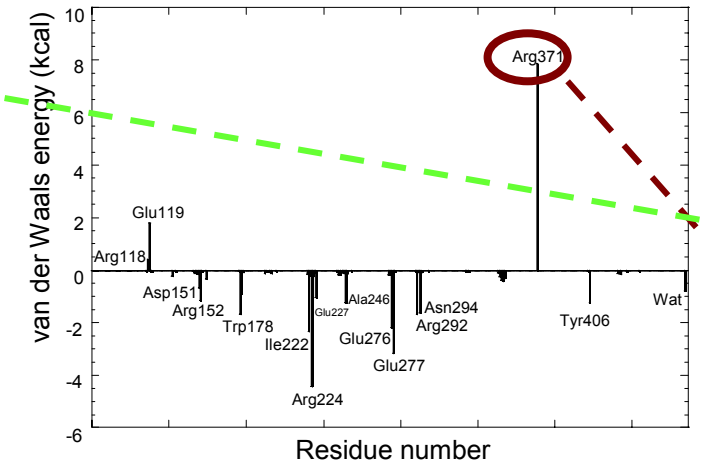


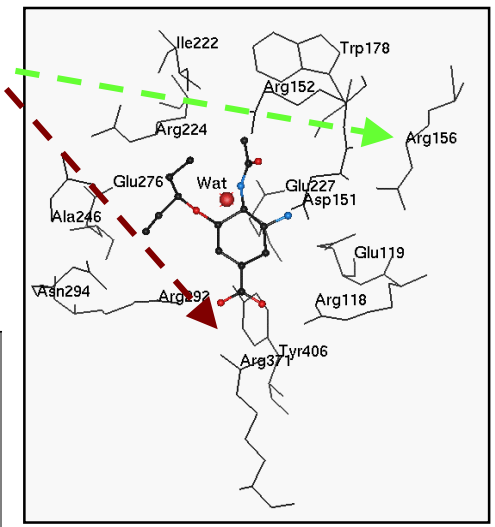


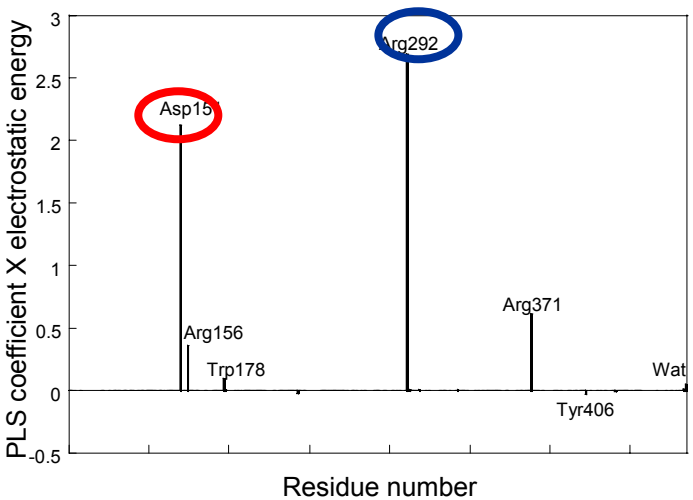
1 bridging water molecule

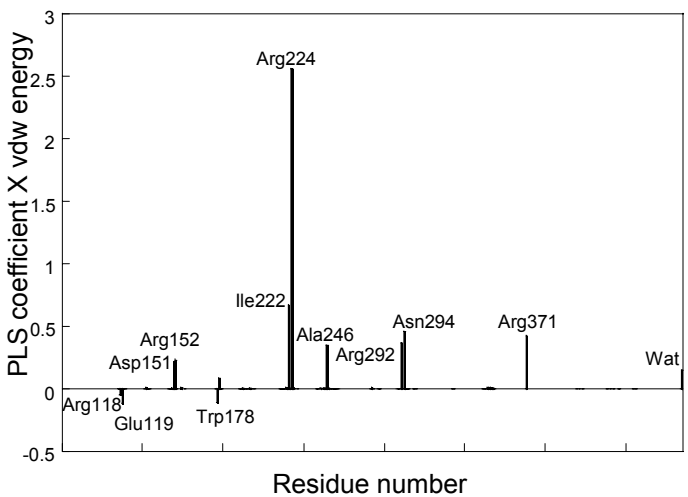
Influenza neuraminidase inhibitors COMBINE model









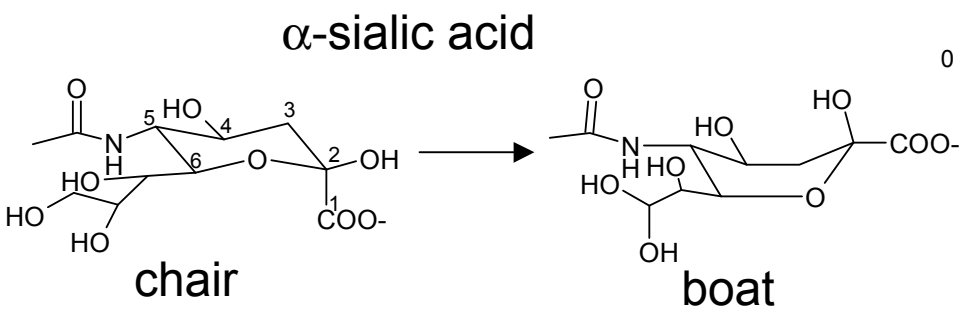


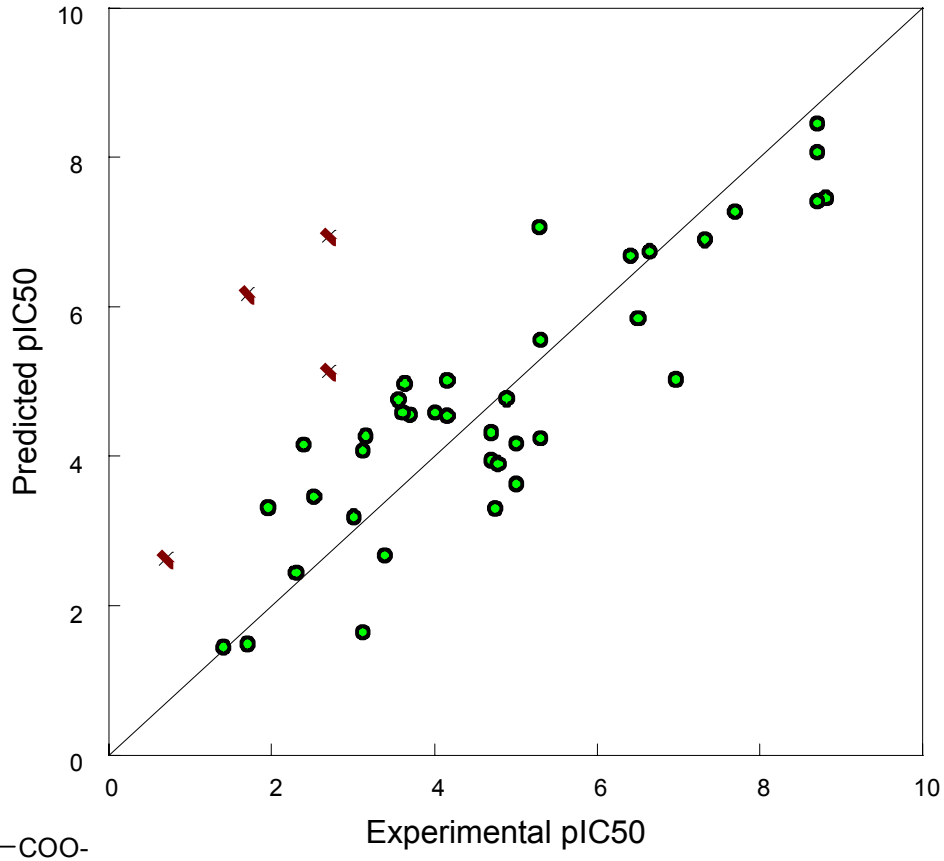
Influenza neuraminidase inhibitors COMBINE model outliers

4 outliers overestimated binding: 3 sialic acid + axial-PANA complexes

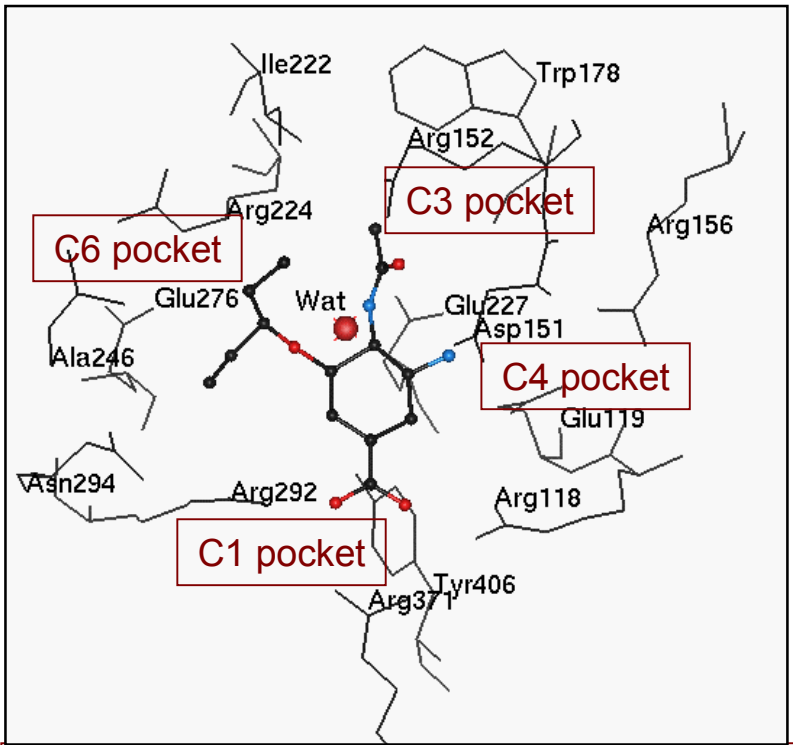
Sialic acid:

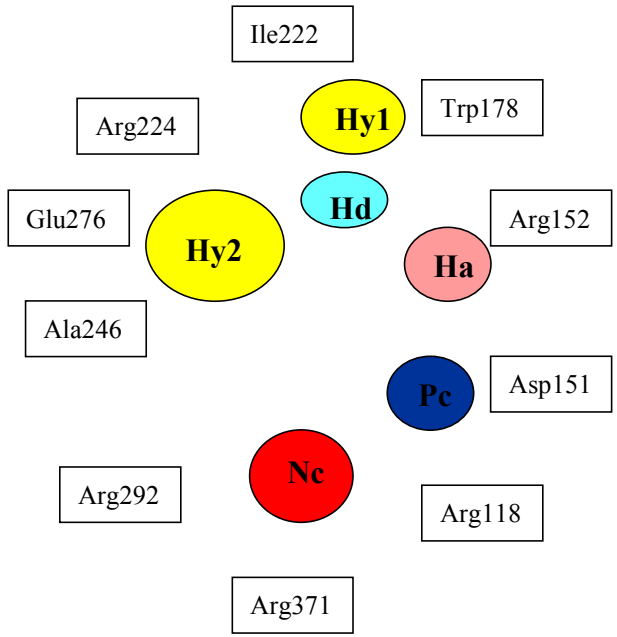
1. α: β anomers 1:10; α binds
2. Conformational change to boat form





Design of new Influenza neuraminidase inhibitors





Characteristics of high affinity ligand

1. Novel frameworks possible
   E.g. bcx-1812- new inhibitor with 5-membered ring framework
   Experimental pIC_{50} ~ 8.85-10; predicted pIC_{50} ~ 8.4 (N9), 7.1 (N2)
2. Ligand binding affinity optimization to maximum number of subtypes with minimum resistance due to mutation. Avoid R292; target residues whose mutation is highly deleterious, e.g.151, 276, 406

Lecture 4: Overview

- How can a COMBINE QSAR model be derived using the computed energy components?
  - Building an energy term/activity matrix
  - PCA: Principal Components Analysis
  - PLS: Projection to Latent Structures by means of Partial Least Squares
  - Variable selection
  - Model validation