



**S-IN Soluzioni  
Informatiche**

*for Chemistry and Pharmaceutical Chemistry*

# MVDA - Multivariate Data Analysis - applied to analysis and prediction of physico-chemical properties

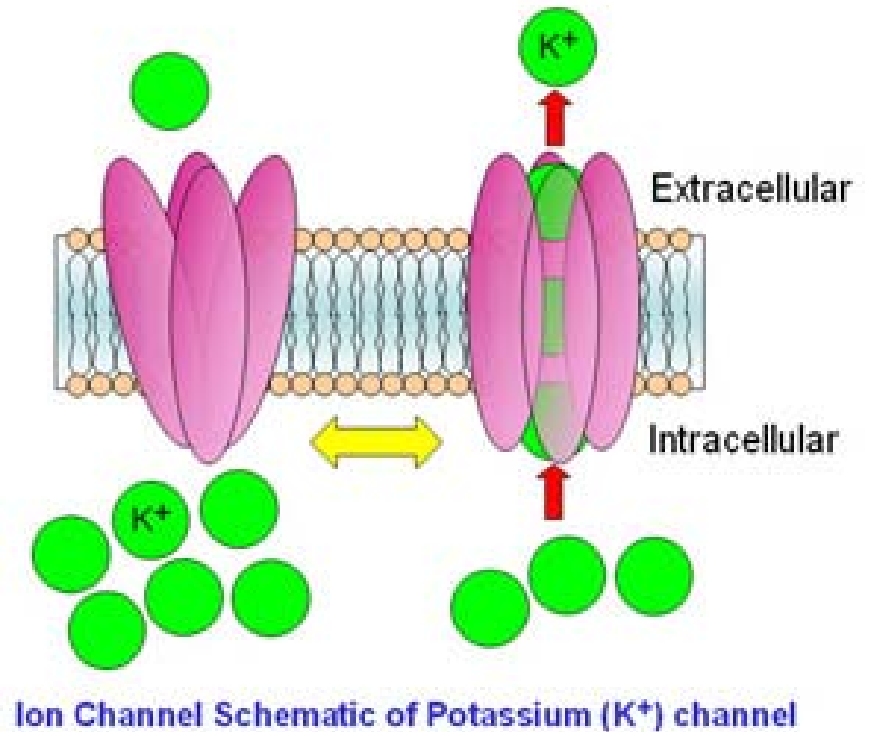
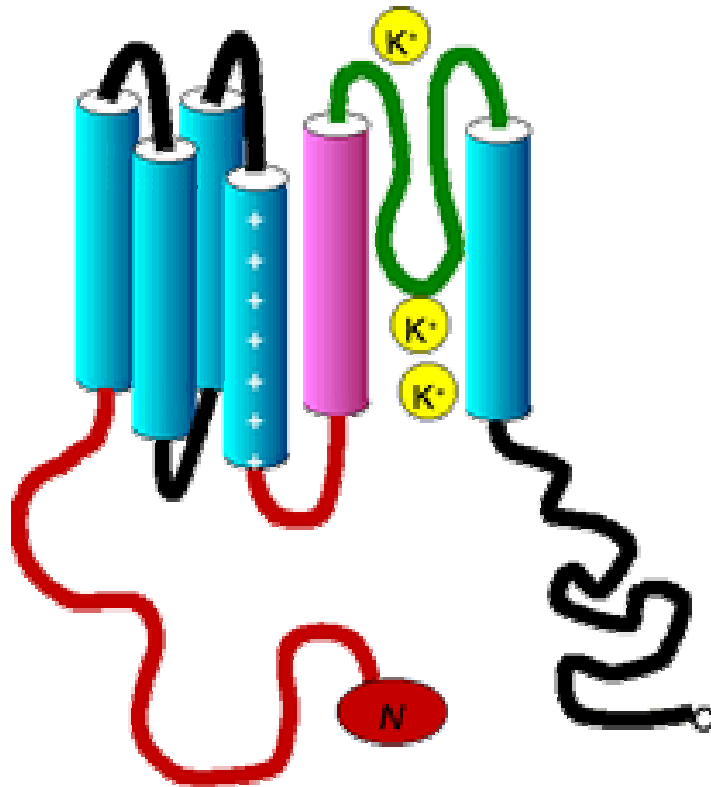
Elena Fioravanzo



# Outline

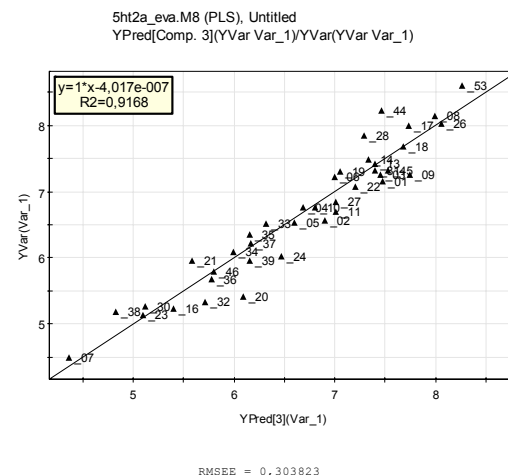
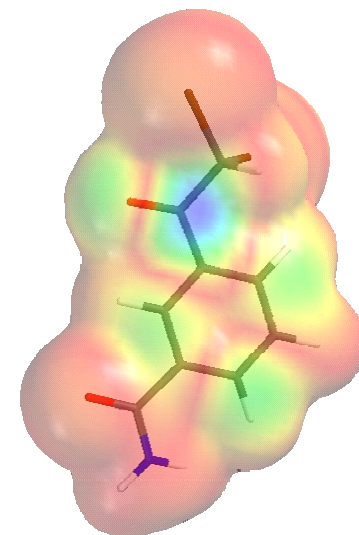
- Case Study
  - hERG K<sup>+</sup> channels
- Methods
  - QSAR
  - Other approaches
    - PASS
    - QikProp
  - Classification

# hERG K<sup>+</sup>channels

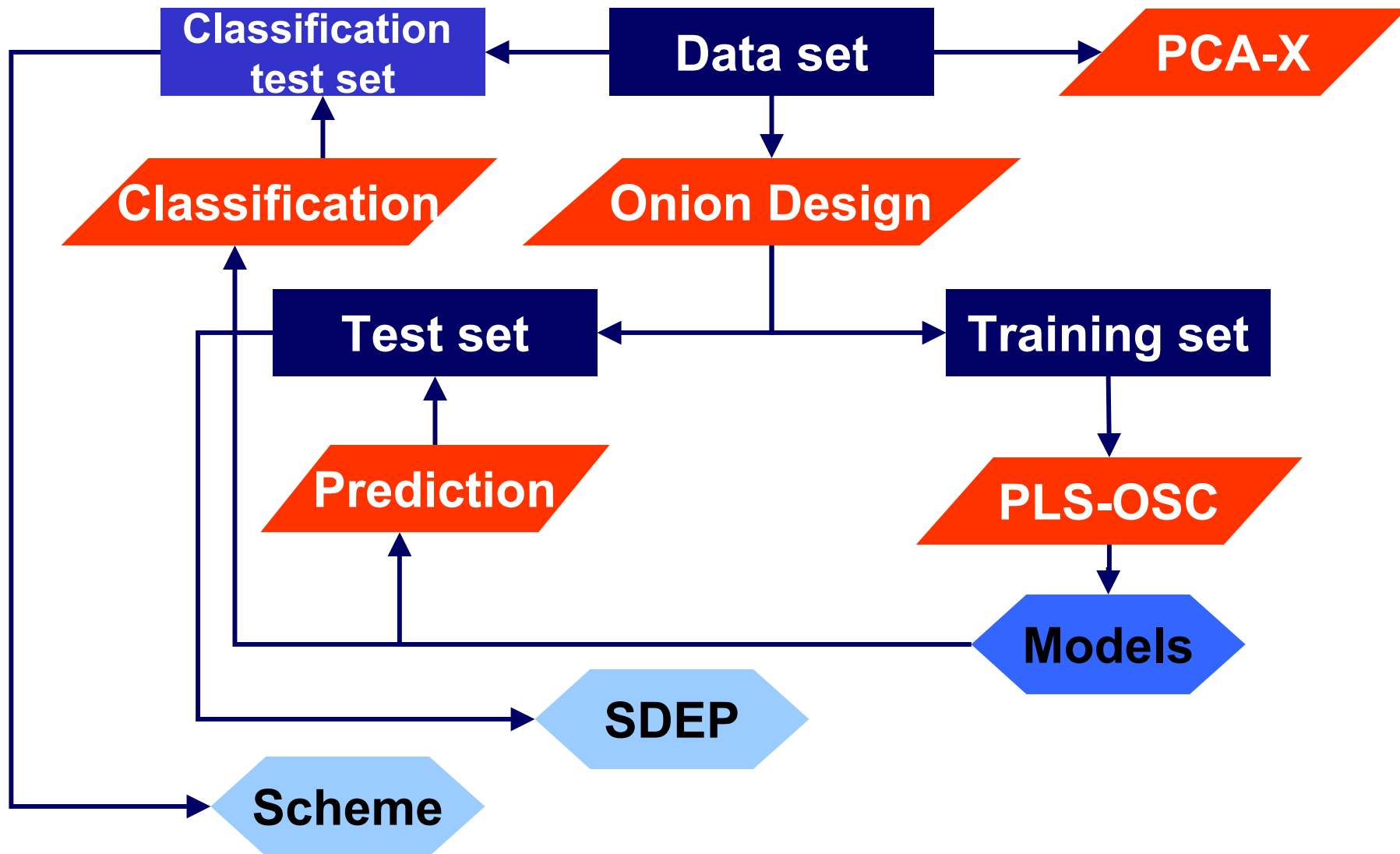


# Methods

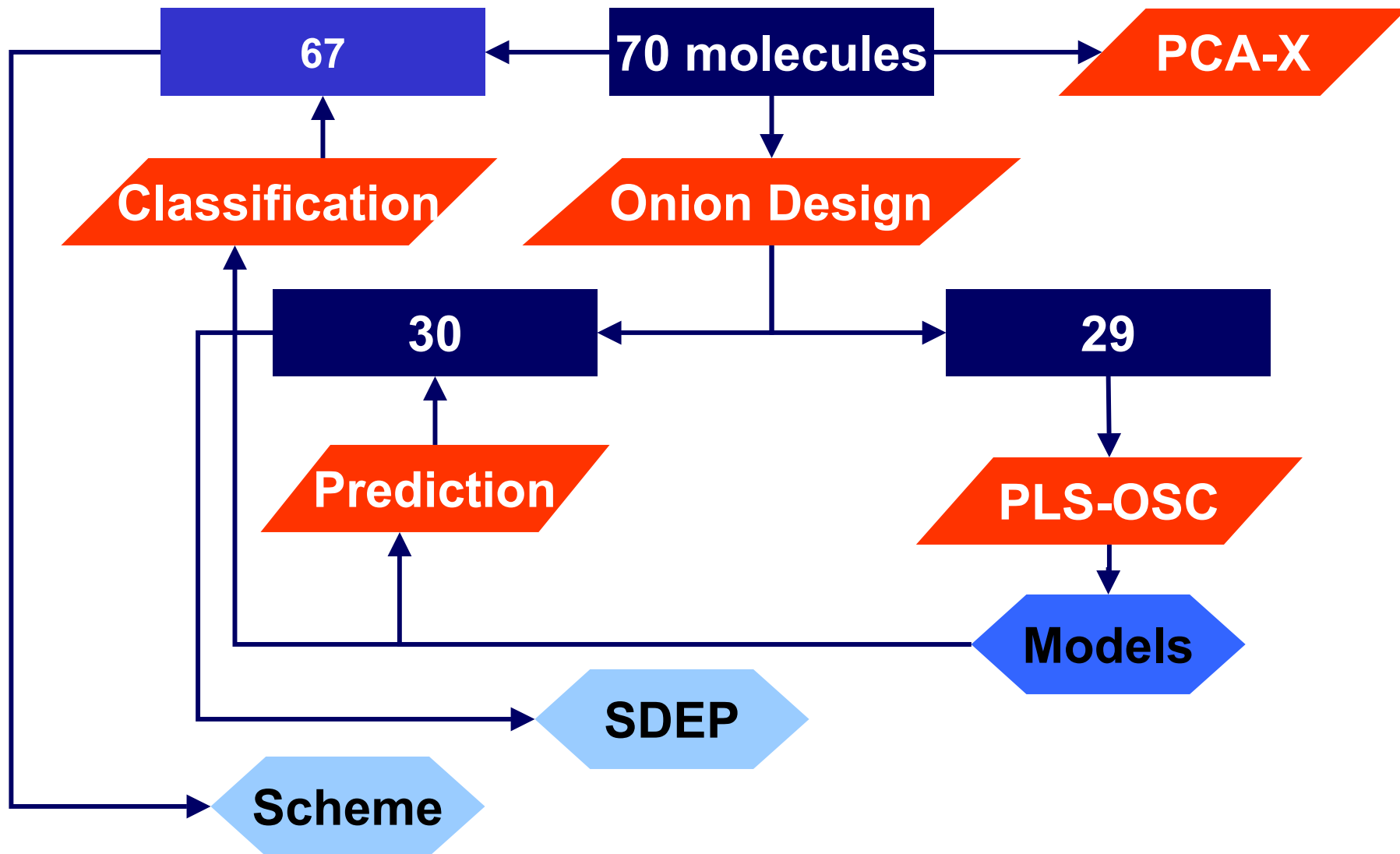
- QSAR
  - Descriptors
    - EVA Spartan (Wavefunction, Inc.) EVA (S-IN)
    - Dragon (Talete)
  - Algorithms
    - Onion Design MODDE (Umetrics AB)
    - PLS-OSC SIMCA-P+ (Umetrics AB)
  
- Other approaches
  - PASS (Prof. Poroikov) PAD-Viewer (S-IN)
  - QikProp (Schrödinger L.C.C.)



# Protocol



# Protocol



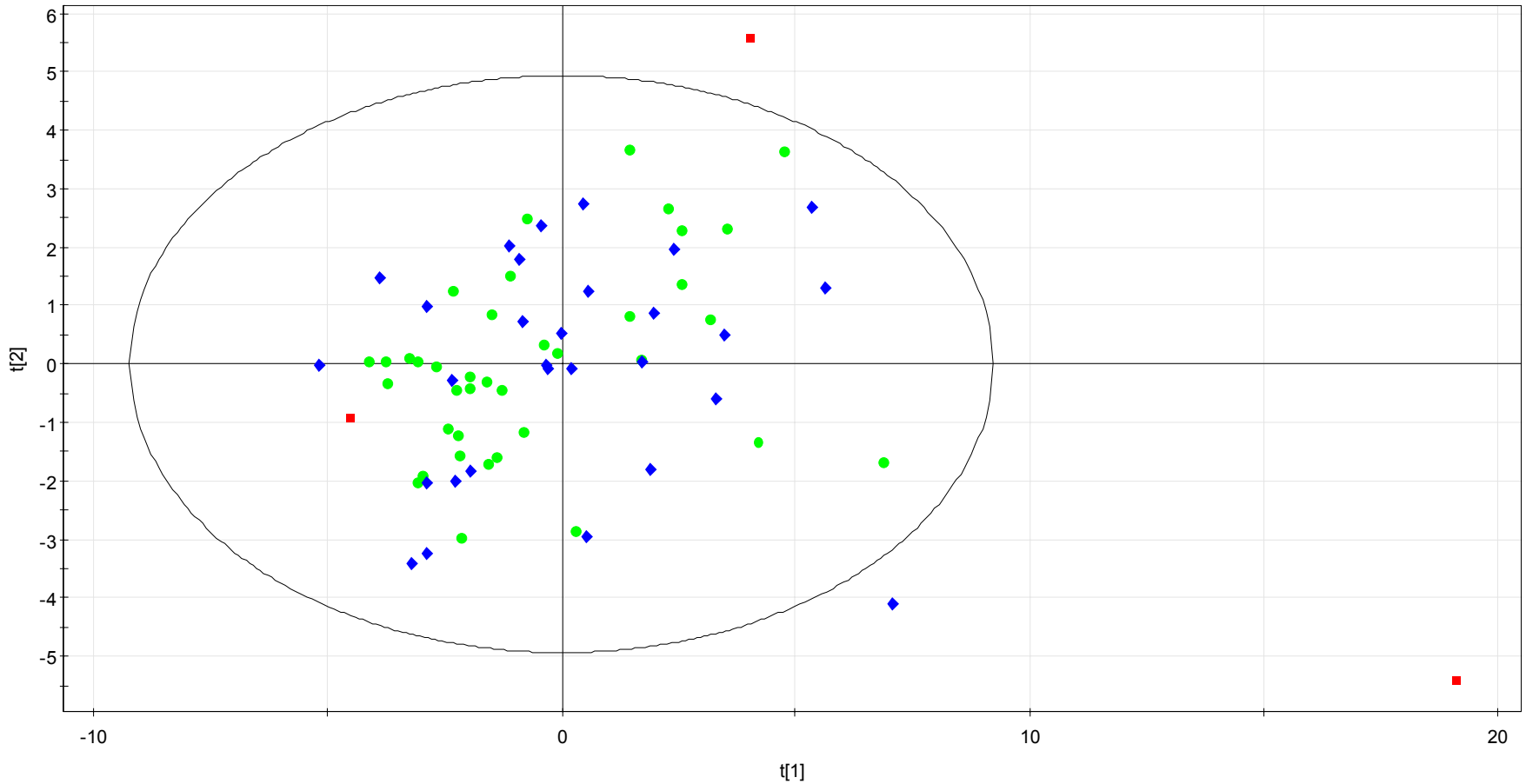
# SIMCA-P+

- Advanced data analysis and user-friendliness combined together
- Multivariate Data Analysis
  - To extract valuable information from complicated data matrices
- PCA & PLS
- OSC – Orthogonal Signal Correction
- OPLS – Orthogonal PLS
  - is a modification of the usual PLS method that filters out variation that is not directly related to the response. The result is more transparent models which are easier to interpret. OPLS is of particular value in spectroscopic calibration, QSAR modelling and "omics" analysis but is set to become the standard in all areas of data analysis. OPLS is a patented technology by Umetrics.

# PCA-X

EVA.M9 (PCA-X)  
t[Comp. 1]/t[Comp. 2]  
Colored according to Obs ID (Primary)

Red outliers  
Blue training set  
Green test set

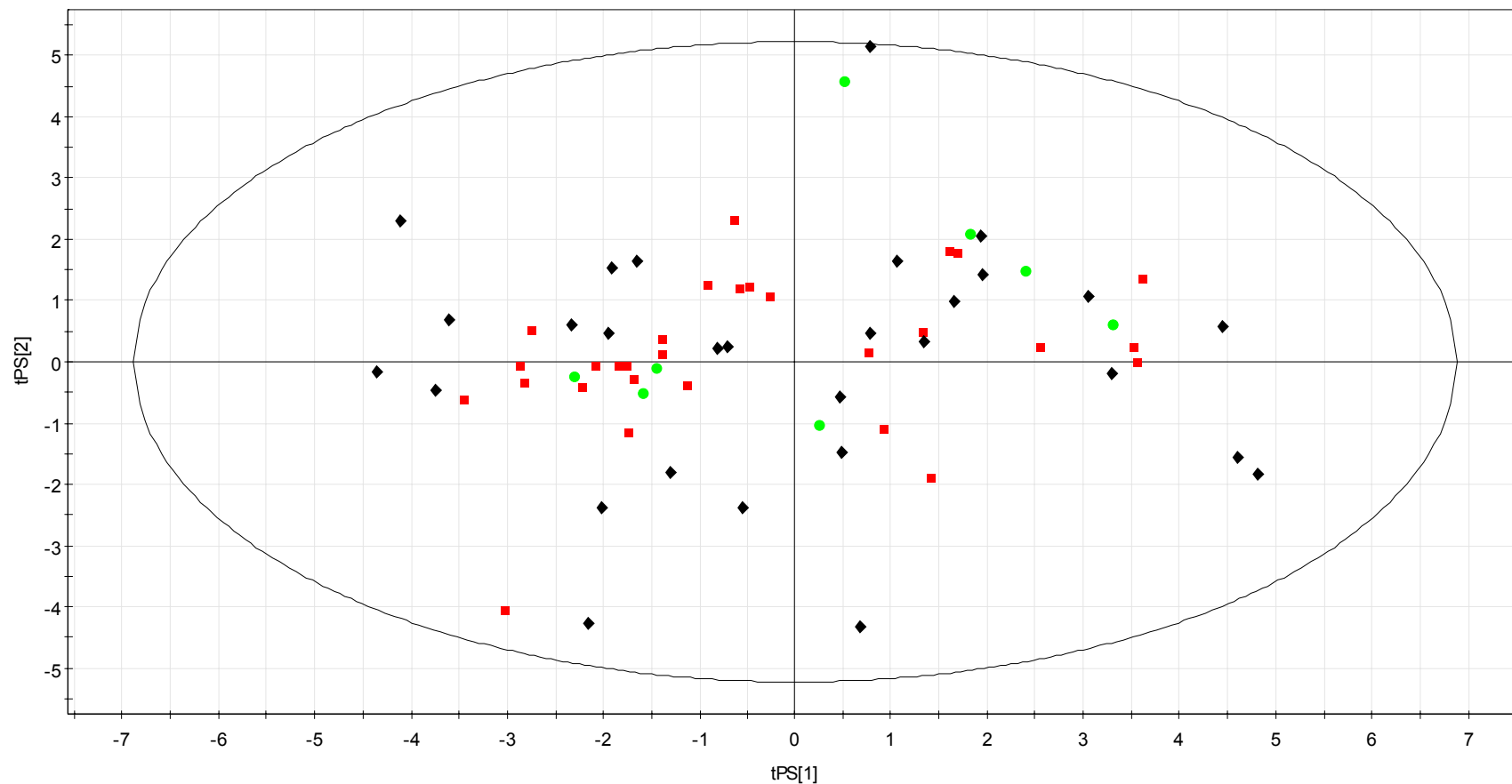


R2X[1] = 0.435241      R2X[2] = 0.124516  
Ellipse: Hotelling T2 (0.95)

# PLS-OSC

mol70\_eva\_OSC.M1 (PLS), PS-SETA  
 tPS[Comp. 1]/tPS[Comp. 2]  
 Colored according to Obs ID (Primary)

■ test1  
 ● test2  
 ◆ train



R2X[1] = 0.156674                      R2X[2] = 0.0930566  
 Ellipse: Hotelling T2PS (0.95)

# Results

70 compounds

3 outliers + 29 training set + 30 test set

OSC models	X	PCs	R <sup>2</sup>	Q <sup>2</sup>	SDEC	SDEP
DRAGON 2D	741	1	0.990	0.972	0.138	0.842
DRAGON 3D	1441	1	0.994	0.994	0.103	1.001
EVA	615	2	0.998	0.974	0.060	0.991

# PASS & QikProp

- PASS
  - Prediction of **Activity Spectra** for Substances
  - List of probabilities for a given compound to be active or inactive towards specific **macromolecular targets** and **pharmacological effects**.
- QikProp
  - Based on physically meaningful **3D descriptors** best suited to describe ADME properties
  - Calculates **novel properties**, such as binding to human serum albumin, skin permeability, reactive functional groups, likely metabolic processes, HERG

# Classification

- **PASS** was trained using molecules with  $pIC_{50}$  values  $\geq 5.0$ .
- **QikProp and QSAR models**
  - threshold value  $pIC_{50} = 5$
- **New set: 67 compounds**
  - 59 compounds from PLS data set
  - 8 compounds from literature
  - threshold value  $pIC_{50} = 5$ 
    - 47 active compounds
    - 20 inactive compounds

# Parameters

$$\textit{Sensitivity} = 100 \frac{TP}{TP + FN}$$

the percent of positives  
correctly predicted positives

$$\textit{Specificity} = 100 \frac{TN}{TN + FP}$$

the percent of negatives  
correctly predicted negatives

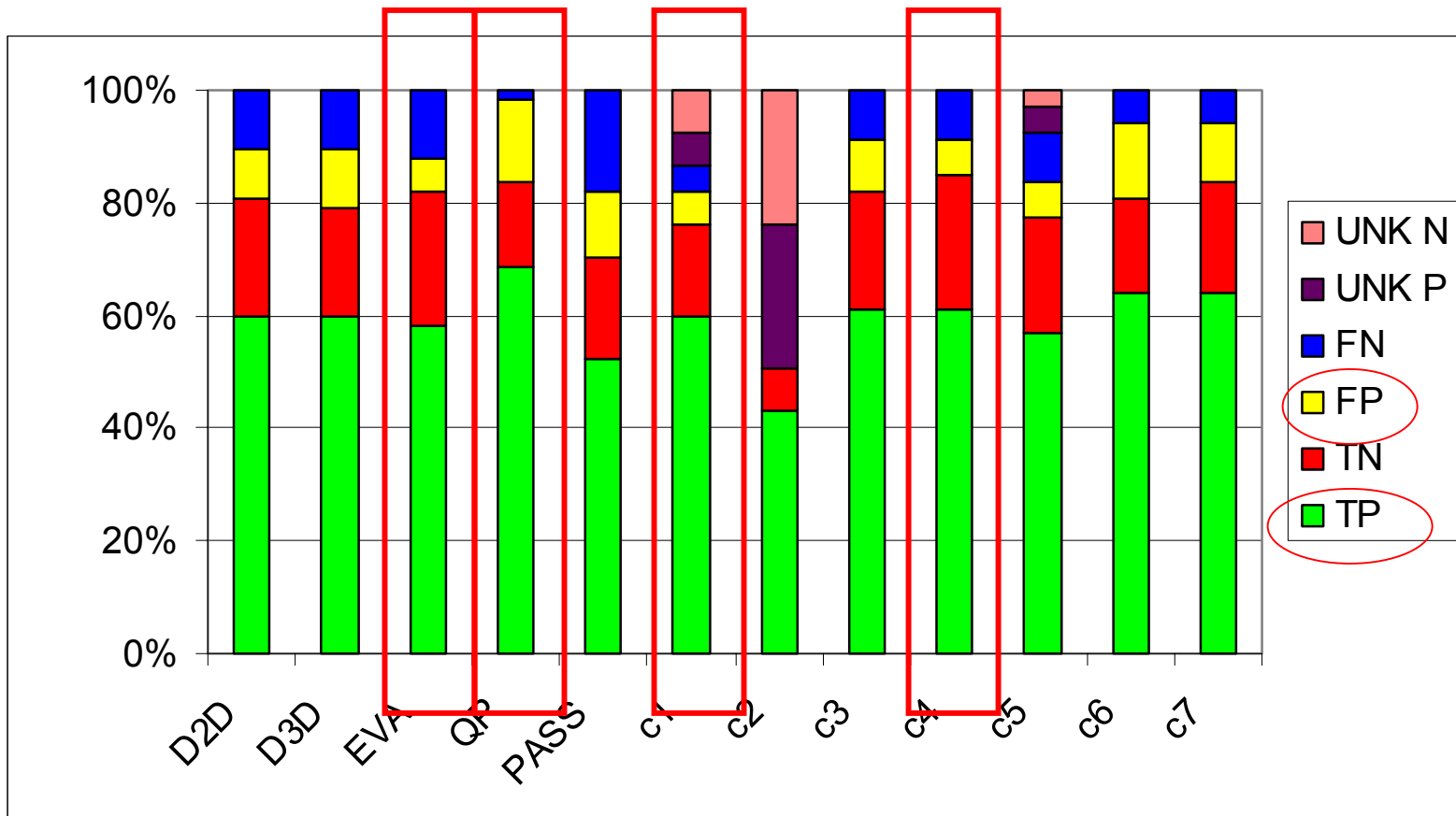
$$\textit{MCC} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (FP + TN) * (TN + FN) * (FN + TP)}}$$

Mathews Correlation Coefficient, a weighted statistic. When MCC equals 0, the model is equivalent to that expected by chance. When it equals 1, there is perfect agreement between actual and predicted values. The stronger this agreement, the higher the value of MCC

# Results

	<b>TP</b>	<b>TN</b>	<b>FP</b>	<b>FN</b>	<b>Sen.</b>	<b>Spe.</b>	<b>MCC</b>
D2D	40	14	6	7	85	70	0.54
D3D	40	13	7	7	85	65	0.50
EVA	39	16	4	8	83	80	0.60
QP	46	10	10	1	98	50	0.59
PASS	35	12	8	12	74	60	0.33

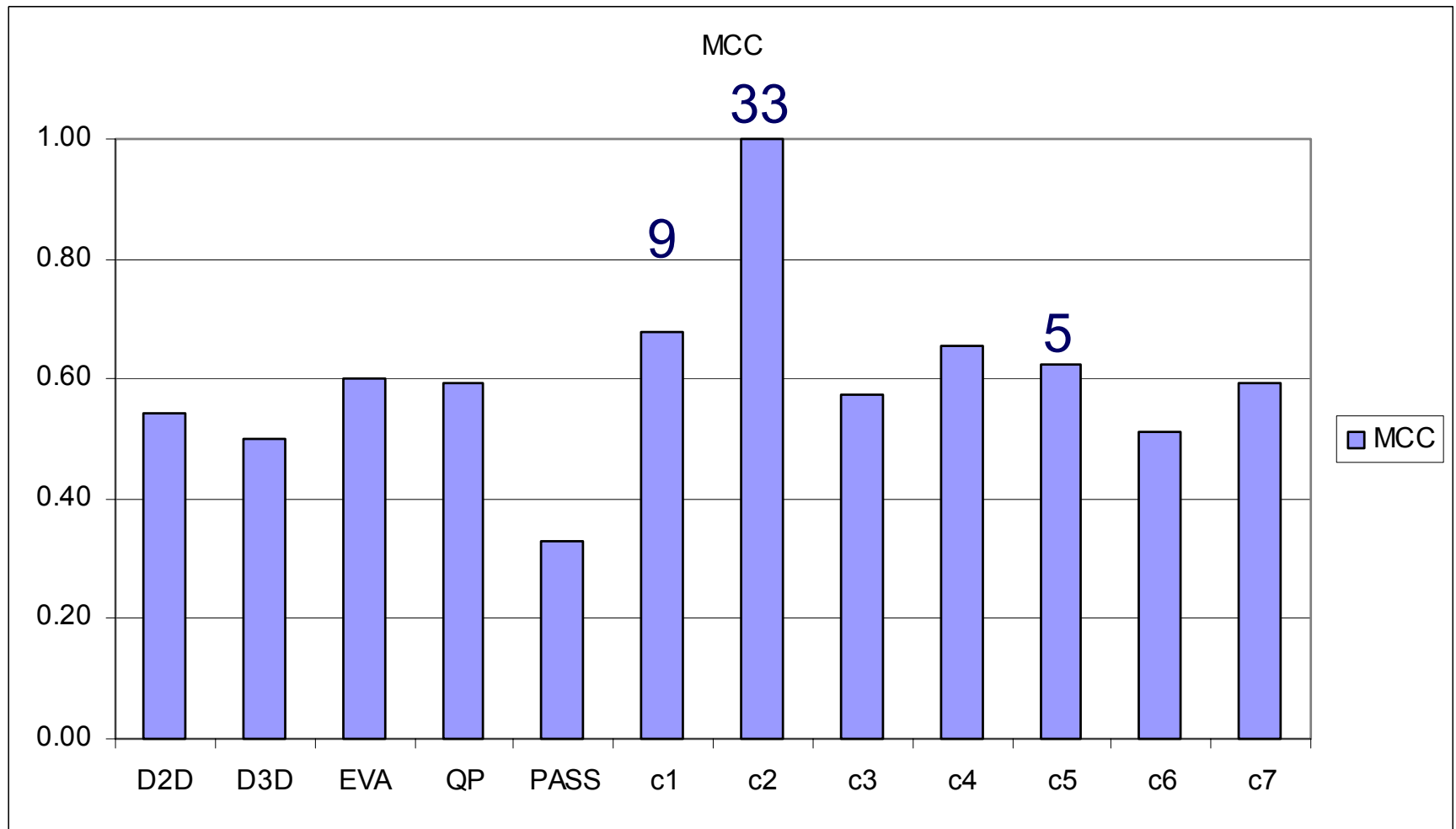
# Consensus



**C1:** 3 between D2D; EVA; QP; PASS  
**C2:** 4 between D2D; EVA; QP; PASS  
**C3:** D2D; EVA; QP **C4:** D2D; EVA; PASS

**C5:** D2D; EVA; **C6:** D2D; QP; PASS  
**C7:** EVA; QP; PASS

# Consensus



# Conclusion

- By employing different and independent approaches it is possible to obtain a consensus score, more reliable than any single method, to be used as a filter in the discovery process
- There are several ways in which ADMET models can be used to tailor predictions depending on one's needs, i.e. give the best overall models, or focus on recovering the positive or negative predictions
- These combined models allow the correct prediction between 80 and 90% of compounds for those classes of interest, either positive or negative

# Acknowledgments

- **ACRAF**
  - Lucia Durando
  - Nicola Cazzolla
  - Rosella Ombrato
- **S-IN**
  - Cristina Ferrari
  - Marco Parenti
  - Massimo Mabilia