QSAR Modeling of Human Liver Microsomal Stability

Alexey Zakharov

CADD Group

Chemical Biology Laboratory

Frederick National Laboratory for Cancer Research

National Cancer Institute, National Institutes of Health, DHHS

Causes of major drug failures

- Global Business Intelligence Research released a report related to the causes of major drug failures during 2005-2010, in which more than 20 drug failures were analyzed.

  The main reasons for failures were: 68% related to efficacy, 21% to safety

- Efficacy and safety is strongly influenced by metabolic degradation and excretion

Metabolic stability assessment

- The most important value that can be measured to quantify metabolic excretion and thus stability of compounds is their half-life time ($t_{1/2}$) determined in human liver microsomes

- High-throughput in vitro metabolic stability assays are widely used for investigation of the stability of compounds

- An alternative are computational approaches (QSAR methods), which can be applied to prioritize large numbers of compounds for in vivo measurements

Microsomal stability data sets

1) Evolvus database (commercial), Elvolvus Group, India

2) ChEMBL (public), Assay ID 1614674

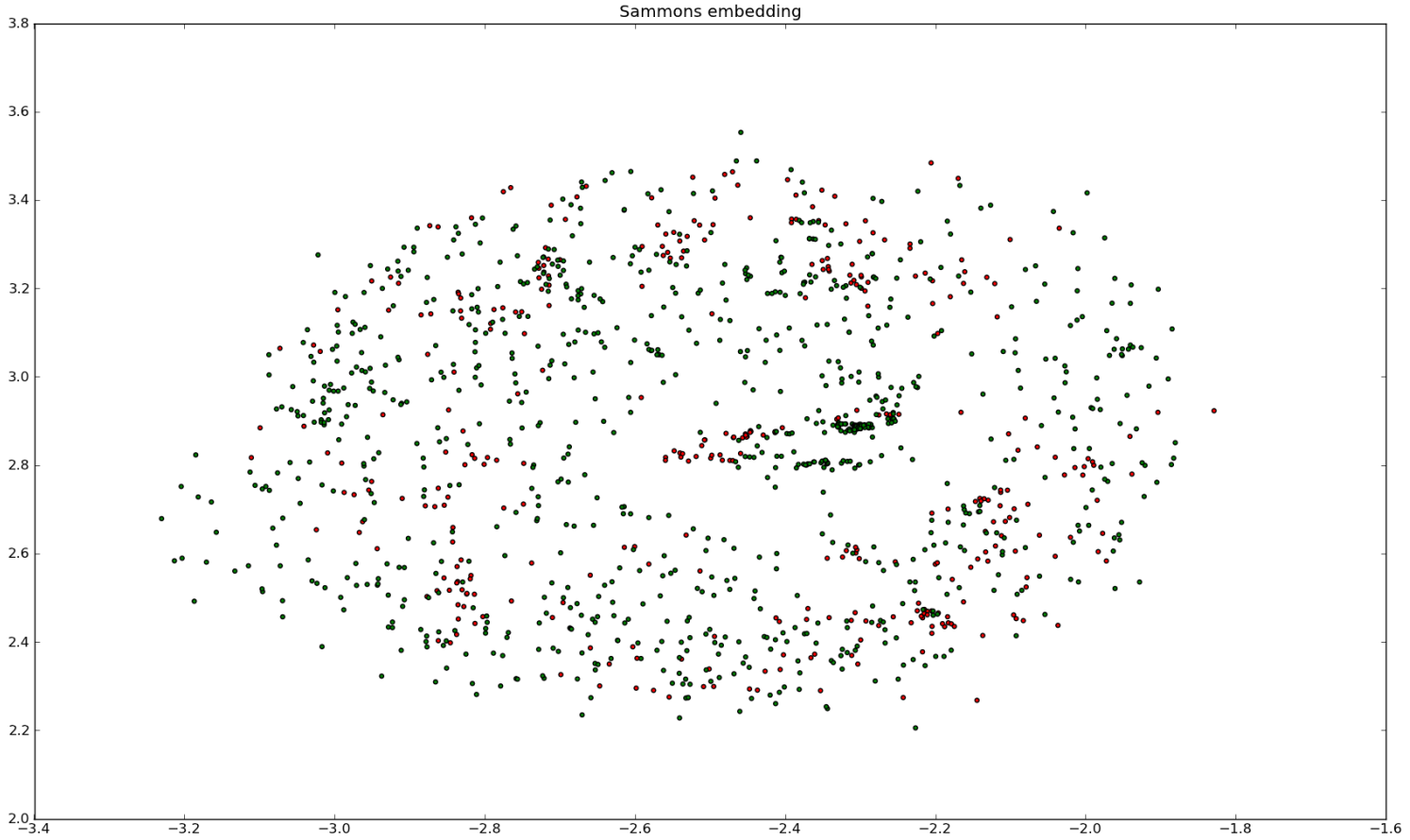
3) Goodman & Gilman's book (public), Gilman AG (Ed.), McGraw-Hill, pp. 1917-2023 (2001).

4) Sanford Burnham Medical Research Institute (SBMRI), (see PubChem AID 1555, AID1940)

Data Set	Number of Compounds	Unstable	Stable
Evolvus data set	1242	345	897
ChEMBL human external set	669	5	664
Goodman & Gilman human external set	246	5	241
SBMRI human external set	80	21	59

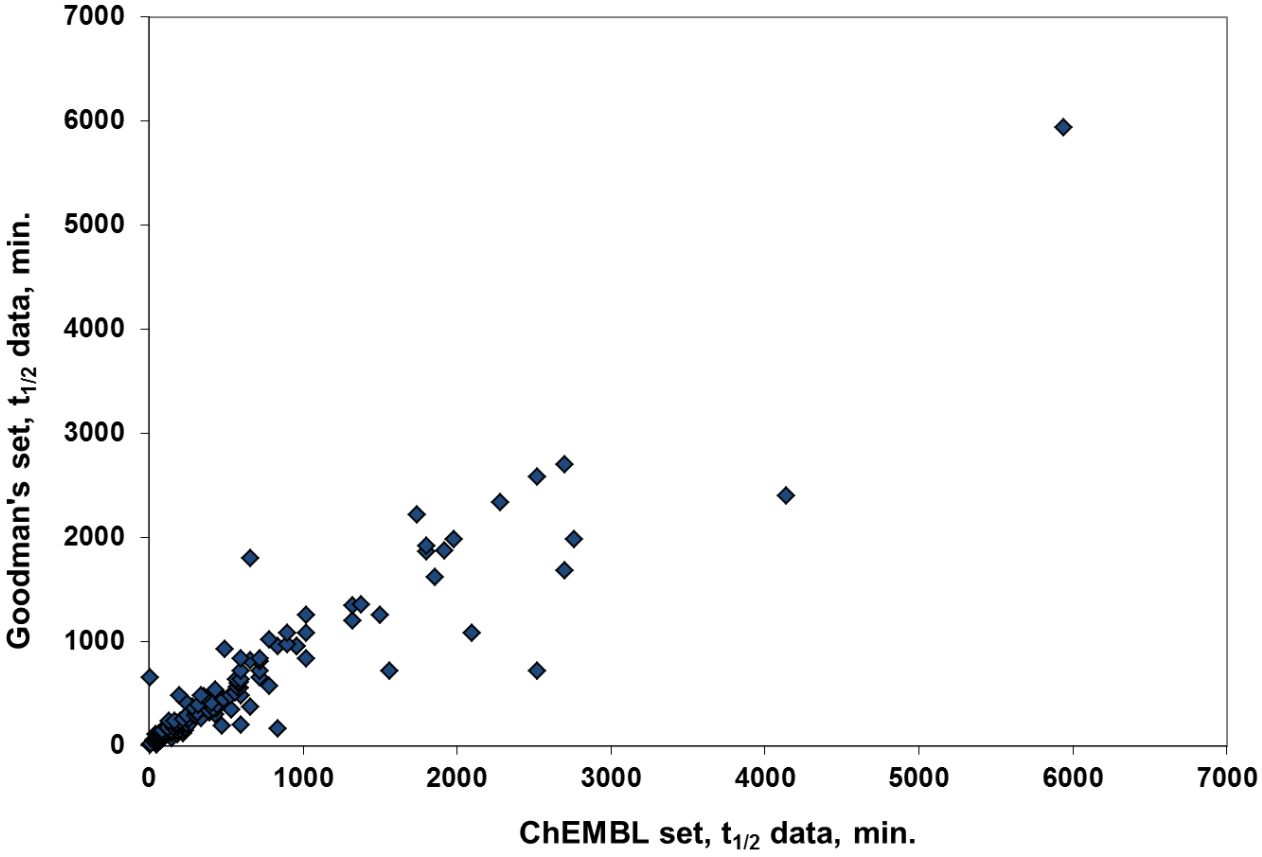
Unstable: $t_{1/2} \leq 15$ min; stable: $t_{1/2} > 15$ min.

Diversity analysis of Evolvus data set



- Fingerprints from KNIME
- Tanimoto distance
- Sammons embedding approach

Significant differences in $t_{1/2}$ measurements



Experimental $t_{1/2}$ data in the ChEMBL set vs. the G&G set

These two sets have 156 structures in common

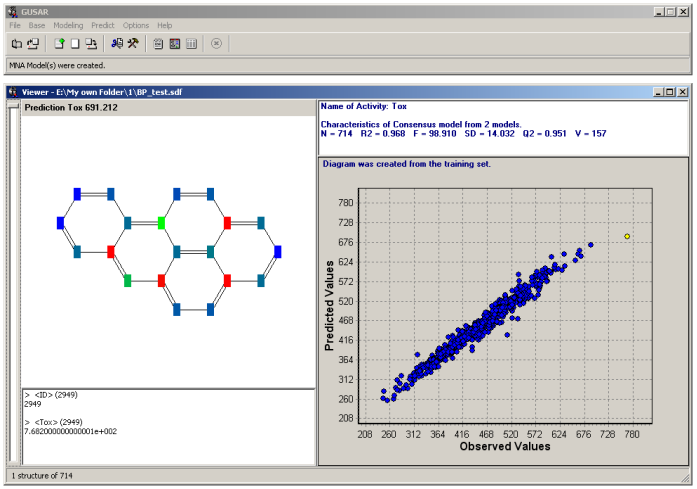
Significant differences in $t_{1/2}$ measurements

The ChEMBL and G&G sets have the same end-point data for 156 structures, but obtained from different sources

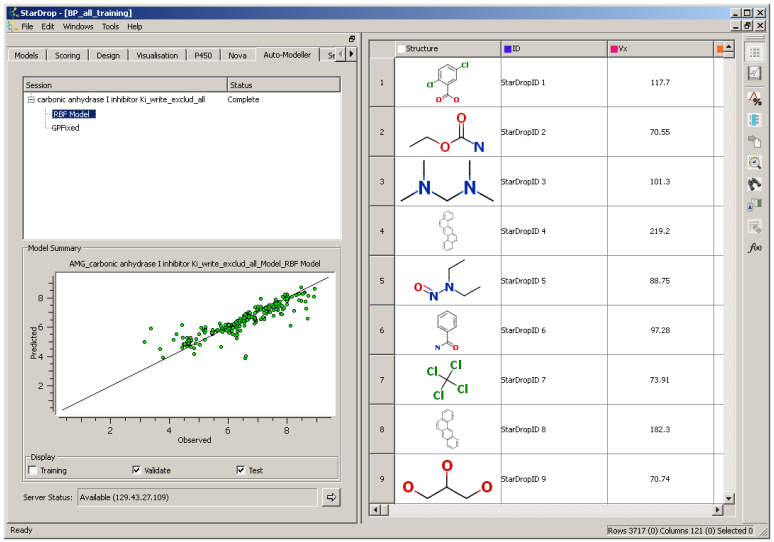
$t_{1/2}$, min.	**0 -15**	**15 – 60**	**60 - 360**	**360-720**	**720-1440**	**1440-6000**
$t_{1/2}$, hours	0.25	0.25 – 1	1 - 6	6 - 12	12 - 24	24 - 100
Number of compounds	3	10	82	28	19	14
Average difference, min.	12	9	50	197	188	411
Average difference (%)	43	18	23	31	18	20

QSAR Methods

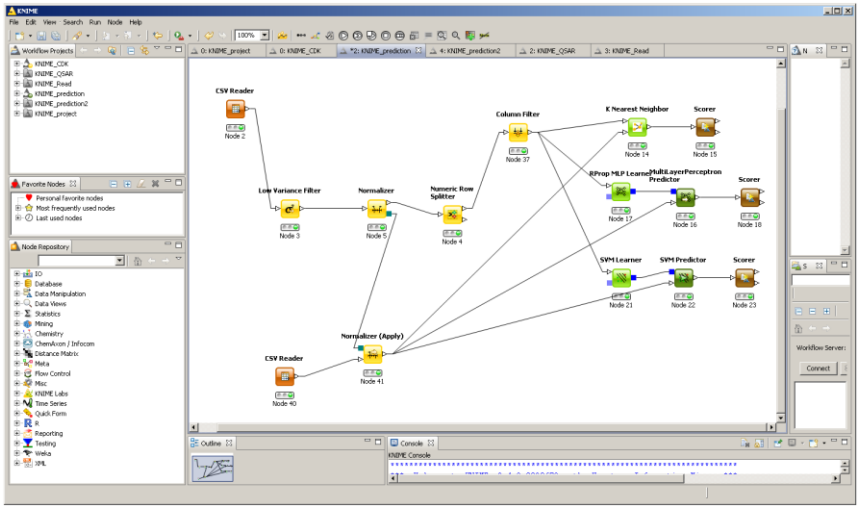
GUSAR (commercial), Ver. 2011



StarDrop (commercial), Ver. 5.0



KNIME (public), Ver. 2.4.2



QSAR Methods

GUSAR software

QNA (Quantitative Neighborhoods of Atoms) descriptors

Filimonov D.A. et al. Proceedings of the 15th European Symposium on Structure-Activity Relationships (QSAR) and Molecular Modeling, Ed. by Esin Aki (SENER), Ismail Yalcin, Istanbul, 2004, p.98-99

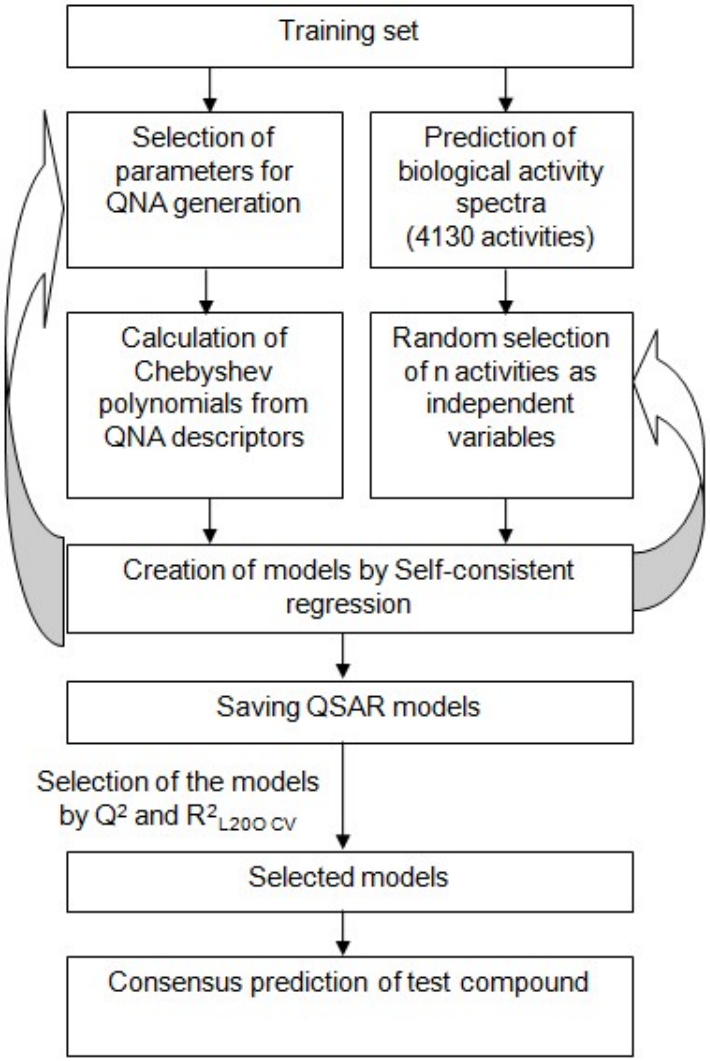
PASS* Predictions as independent variables

Filimonov D. et al. J. Chem. Inf. Comput. Sci. (1999), Vol. 39, P. 666-670.

Self-Consistent Regression (SCR)

Filimonov D. et al. Pharm. Chem. J., 2004, 38, 21-24

*PASS: Prediction of Activity Spectra for Substances



QNA: Quantitative Neighborhoods of Atoms descriptors

$$P_i = B_i \sum_k (\exp(-\tfrac{1}{2}C))_{ik} B_k$$

$$Q_i = B_i \sum_k (\exp(-\tfrac{1}{2}C))_{ik} B_k A_k$$

$$A = \tfrac{1}{2}(IP + EA),$$

$$B = (IP - EA)^{-\frac{1}{2}},$$

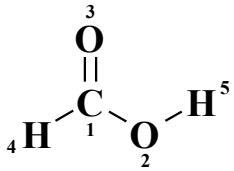
IP is the first ionization potential,

EA is the electron affinity.

Feynman R. Ph. *Phys. Rev.*, 1939, 56, 340-343.
Robert G. Parr et al. *J. Chem. Phys.*, 1978, 68(8), 3801-3807.
Gasteiger J, Marsili M. *Tetrahedron*, 1980, 36, 3219-3228.
Rappe A K and W A Goddard III. *J. Ph. Ch.*, 1991, 95, 3358-3363.

D. Filimonov et al. in Proceedings of the QSAR 2004, Ankara, 2005, pp. 98-99.
D. Filimonov et al. Abstr. 3rd Internat. Symp. CMTPI 2005, Shanghai, 2005.
A. Lagunin et al. SAR and QSAR in Environmental Research 18 (2007), pp. 285-298.

QNA: Quantitative Neighborhoods of Atoms descriptors



a)

$$\mathbf{C} = \begin{vmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{vmatrix}$$

b)

$$\mathrm{Exp}\left(-\tfrac{1}{2}\mathbf{C}\right) = \begin{vmatrix} 1.40 & -0.59 & -0.57 & -0.57 & 0.14 \\ -0.59 & 1.27 & 0.14 & 0.14 & -0.54 \\ -0.57 & 0.14 & 1.13 & 0.13 & -0.02 \\ -0.57 & 0.14 & 0.13 & 1.13 & -0.02 \\ 0.14 & -0.54 & -0.02 & -0.02 & 1.13 \end{vmatrix}$$

c)

	EA	**IP**	*A*	*B*	**P**	**Q**
C	1.263	11.26	6.262	0.316	-0.00218	-0.1820
O	1.461	13.62	7.541	0.287	0.02944	0.3019
O	1.461	13.62	7.541	0.287	0.06199	0.5297
H	0.754	13.60	7.177	0.279	0.05812	0.4706
H	0.754	13.60	7.177	0.279	0.05304	0.3533

d)

(a) structural formula;
(b) connectivity matrix;
(c) exponent of the connectivity matrix;
(d) electron affinities (EA), ionization potentials (IP), parameters A and B, P and Q values for each of the atoms of *formic acid* molecule.

QSAR Methods

StarDrop software

Descriptors used:

• 2D SMARTS-based descriptors
• Whole molecule properties: LogP, TPSA, Molecular weight, McGowan volume, Flexibility index, Number of positive, negative and overall charges, Number of aromatic rings
• Total number of descriptors - 330

StarDrop uses several different **QSAR techniques**:

• Partial Least Squares
• Radial Basis Function fitting (RBF SD)
• Gaussian Processes (GP SD)
• Decision Trees (DT SD)

QSAR Methods

KNIME software

Descriptors used:

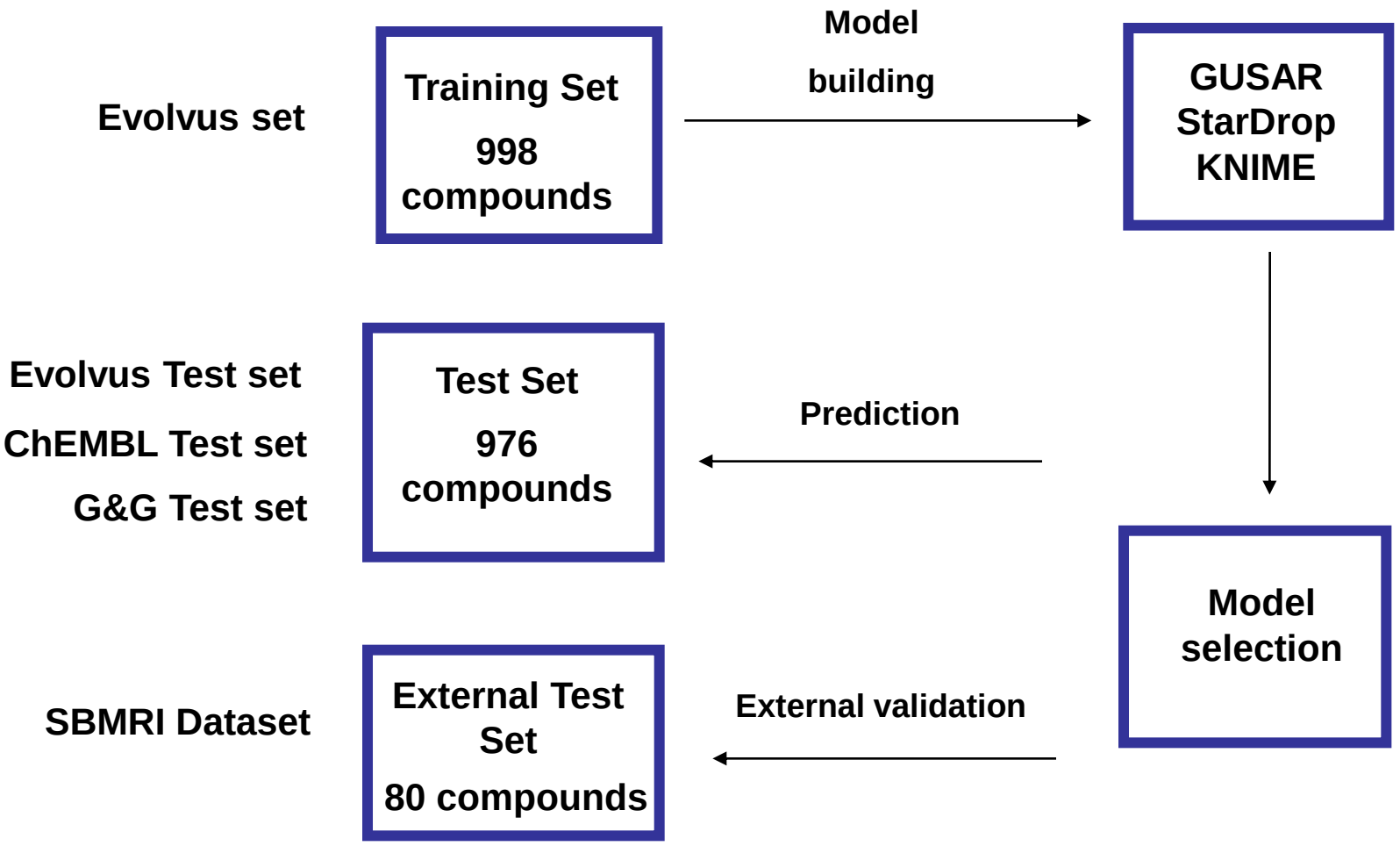
MOLD2 descriptors*:
• Physicochemical properties, fragmental descriptors, structural features and functional groups
• Total number of descriptors: 777

KNIME uses several different **QSAR techniques**:

- K Nearest Neighbor (kNN)
- Multilayer Perceptron (MLP)
- Support Vector Machine (SVM)
- Bayes Network (BayesNet)
- Radial basis function network (RBF network)
- Logistic regression (Logistic)

* Hong H, Xie Q, Ge W, Qian F, Fang F, Shi L, Su Z, Perkins R, Tong W. Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J. Chem. Inf. Comput. Sci.* 48, 1337–1344 (2008).

QSAR Workflow

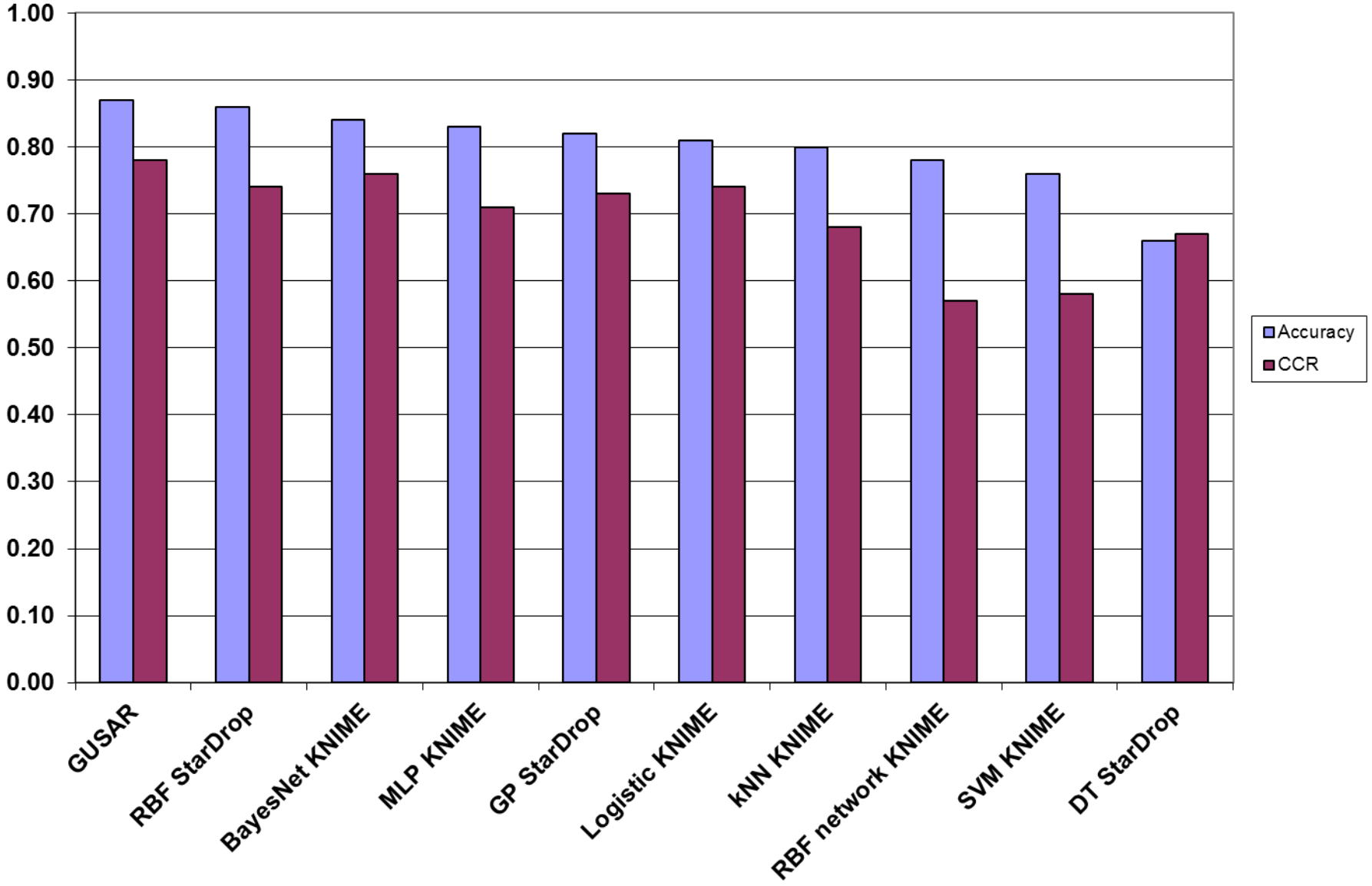


Calculation of prediction accuracy

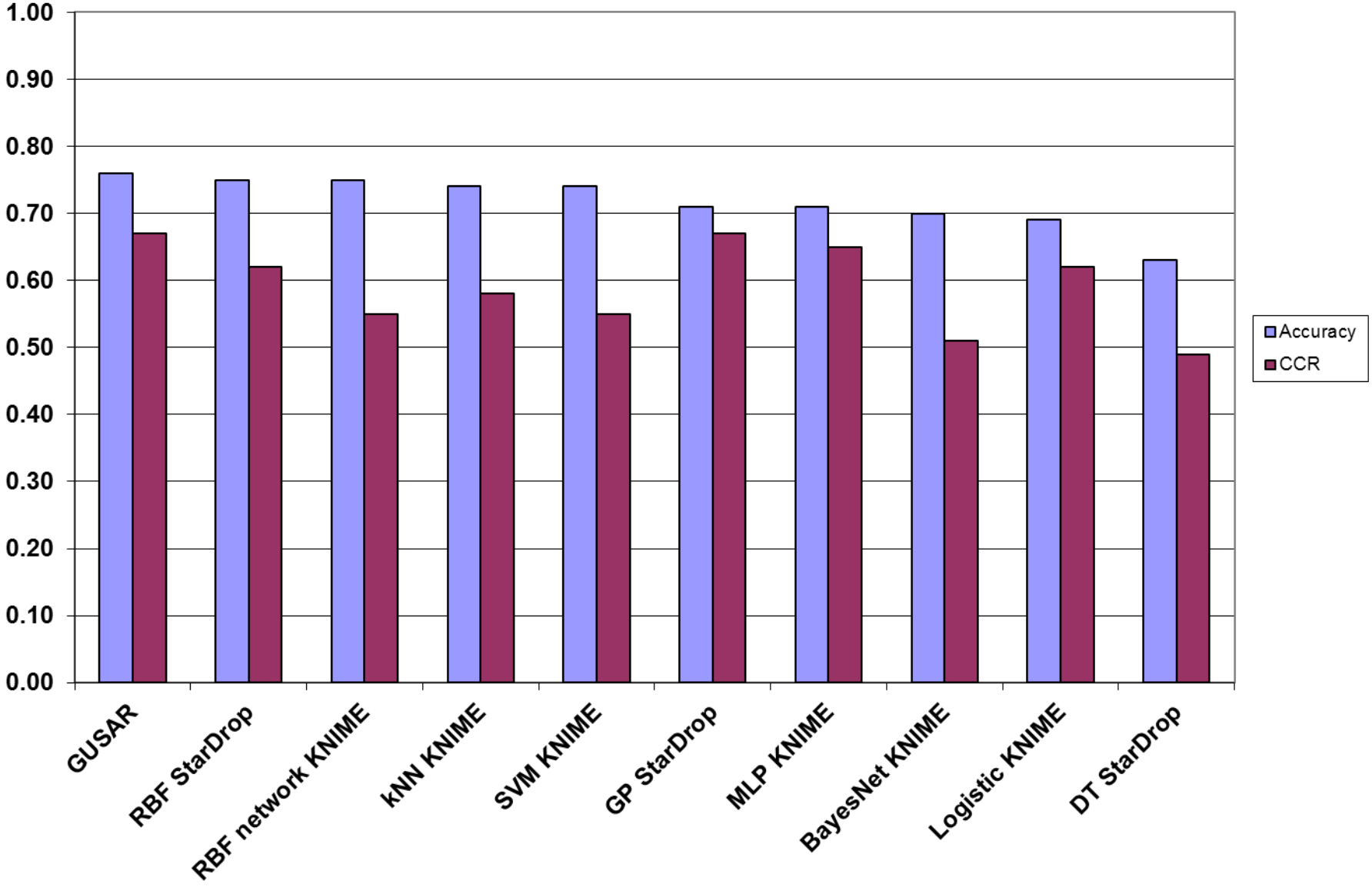
$Accuracy = \frac{TN + TP}{TN + TP + FP + FN}$	Accuracy: probability of correctly classifying compounds.
$Sensitivity = \frac{TP}{FN + TP}$	Sensitivity: probability of predicting positive (unstable) when true outcome is positive.
$Specificity = \frac{TN}{TN + FP}$	Specificity: probability of predicting negative (stable) when true outcome is negative.
$CCR = (Sensitivity + Specificity)/2$	Correct Classification Rate: shows balance between Sensitivity and Specificity.

where TN – true negatives, TP – true positives, FN – false negatives, FP – false positives.

Prediction results for Test set



Prediction results for SBMRI test set



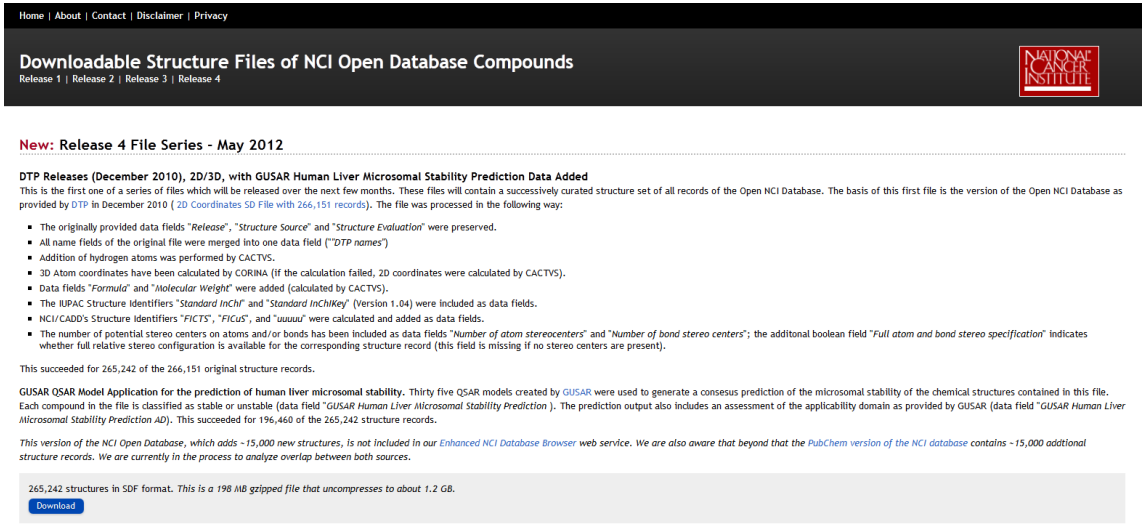
QSAR model application to public structures

```
-ISIS-  01180614322D

 7  7  0  0  0  0  0  0  0  0  0999 V2000
  -0.7107    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
  -0.2130   -0.8654    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
  -0.2130    0.8654    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
  -1.7174    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
   0.7824   -0.8654    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.7824    0.8654    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   1.2892    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
 1  2  2  0  0  0  0
 1  3  1  0  0  0  0
 1  4  1  0  0  0  0
 2  5  1  0  0  0  0
 3  6  2  0  0  0  0
 5  7  2  0  0  0  0
 6  7  1  0  0  0  0
M  END
>  <GUSAR: human liver microsomal stability prediction>
stable

>  <GUSAR_human liver microsomal stability prediction_AD>
in AD

$$$$
```



The NCI database includes more than 250,000 compounds, which are the publicly available part of the half-million structures assembled by the U.S. National Cancer Institute (NCI) in the course of its 55 year long efforts in screening compounds against cancer and AIDS.

Each compound from the NCI database was classified as stable or unstable using the best GUSAR models.

The prediction output also included an assessment of the applicability domain as provided by GUSAR (196460 comp.)

http://cactus.nci.nih.gov/download/nci/.

Summary

• Information about chemical structures and their half-life data was collected from several public and commercial sources and used for the construction of categorical QSAR models.

• Predictive QSAR models were developed using both commercial (StarDrop, GUSAR) and open-source software (KNIME).

• For estimation of the predictivity of the models, several external sets were used.

• The obtained QSAR models showed generally high accuracy of prediction.

• The best obtained model was used to predict metabolic stability of about 196,460 structures from the NCI database. These data have been made available for free download.

Acknowledgments

CADD Group, Chemical Biology Laboratory, Center for Cancer Research

Megan L. Peach Markus Sitzmann Igor V. Filippov Marc C. Nicklaus

Vanderbilt University

Heather J. McCartney

Sanford Burnham Medical Research Institute

Layton H. Smith

Computer-Aided Drug Design at Cancer Research UK, Beatson Laboratories

Angelo Pugliese