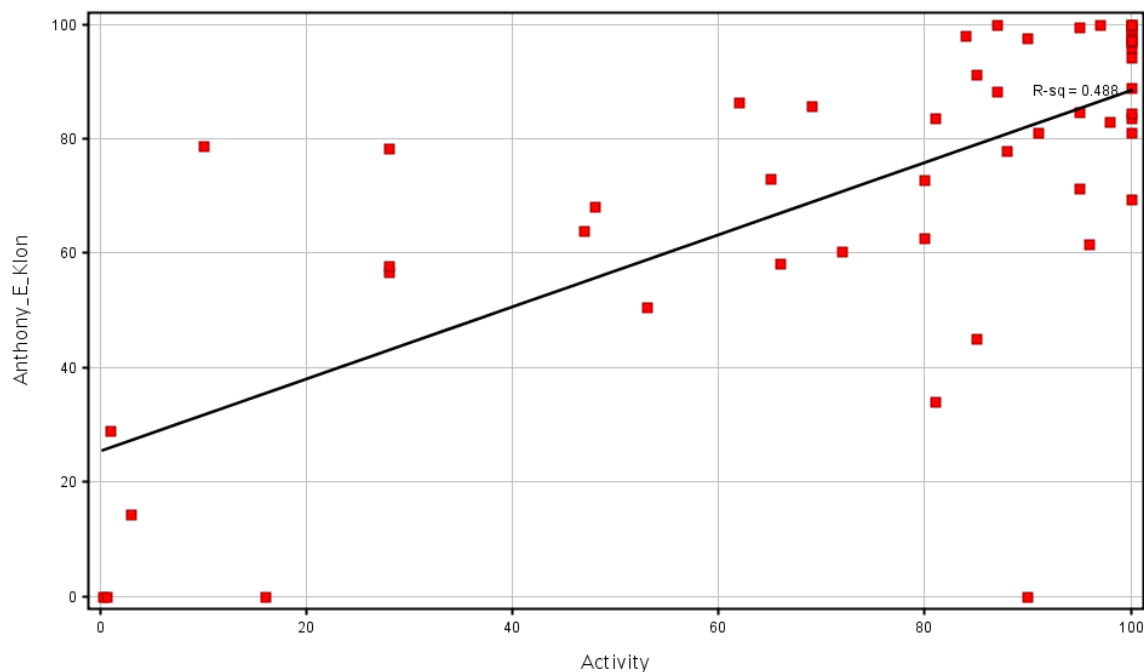


(Submissions follow in alphabetical order below. Figures were created using Sarchitect Designer 2.3.0 (<http://www.strands.com/sarchitect/>) while RMSE values were calculated using a Jython script.)

Anthony_E_Klon *RMSE: 23.84* <aklon@pcop.com>



Dataset Preparation:

All compounds were imported into a Molecular Operating Environment (MOE) database using v.2006.08. The compounds were minimized in MOE using the MMFF94x force field after the assignment of partial atomic charges.

Descriptor Calculation:

A total of 387 2D and 3D descriptors were calculated using MOE, including 166 MACCS public keys used as feature counts. 3D descriptors that required the compounds to be aligned or which were calculated using external coordinates were not included in those calculated. The descriptor values and activities for the set of training compounds were exported in csv format.

Descriptor Selection:

The descriptor values and activities for the training set were imported into Weka 3.5.6. A correlation-based subset evaluator (CfsSubsetEval) was used with the GeneticSearch algorithm to identify the most significant descriptors. CfsSubsetEval evaluates the predictive ability of each feature (descriptor) individually as well as the degree of correlation between descriptors. Descriptors with a high degree of correlation with the response value (activity) that have a low correlation with other descriptors are preferred. The evaluator iteratively adds descriptors with higher correlation to the model on the condition that the model does not already contain one or more descriptors with a higher correlation to the descriptor being considered. The parameters for the genetic search algorithm were as follows:

Crossover probability: 0.6
Maximum number of generations to evaluate: 20
Mutation probability: 0.033
Population Size: 20

This process resulted in 64 descriptors for further consideration.

Manual triaging was carried out on the reduced set of 64 descriptors to remove redundancies. For example, two logP calculators (SlogP and logP) were represented in the set of 64 descriptors. Using a dataset of 14,000 compounds from the AquaSol dataset, SlogP was found to have a higher correlation and logP was removed. Individual descriptors that were part of a larger descriptor set where the majority of the other descriptors were not found to be relevant were also evicted from further consideration. This included two BCUT descriptors, four GCUT descriptors, one SlogP_VSA descriptor, and PEOE_VSA descriptors. We also found some descriptors that existed twice in the descriptors list such as constitutional descriptors (# of Cl, F, I, or P atoms) that were already represented by MACCS public keys and so were removed. This process resulted in a reduced set of 29 descriptors for model generation:

Descriptor List:

b_triple
PEOE_VSA_FHYD
PEOE_VSA_FPPOS
logs
MACCS(-17)
MACCS(-27)
MACCS(-29)
MACCS(-30)
MACCS(-40)
MACCS(-43)
MACCS(-49)
MACCS(-50)
MACCS(-62)
MACCS(-69)
MACCS(-72)
MACCS(-76)
MACCS(-78)
MACCS(-90)
MACCS(103)
MACCS(109)
MACCS(116)
MACCS(123)
MACCS(125)
MACCS(135)
MACCS(145)
MACCS(155)
SlogP
TPSA
density

Model Creation:

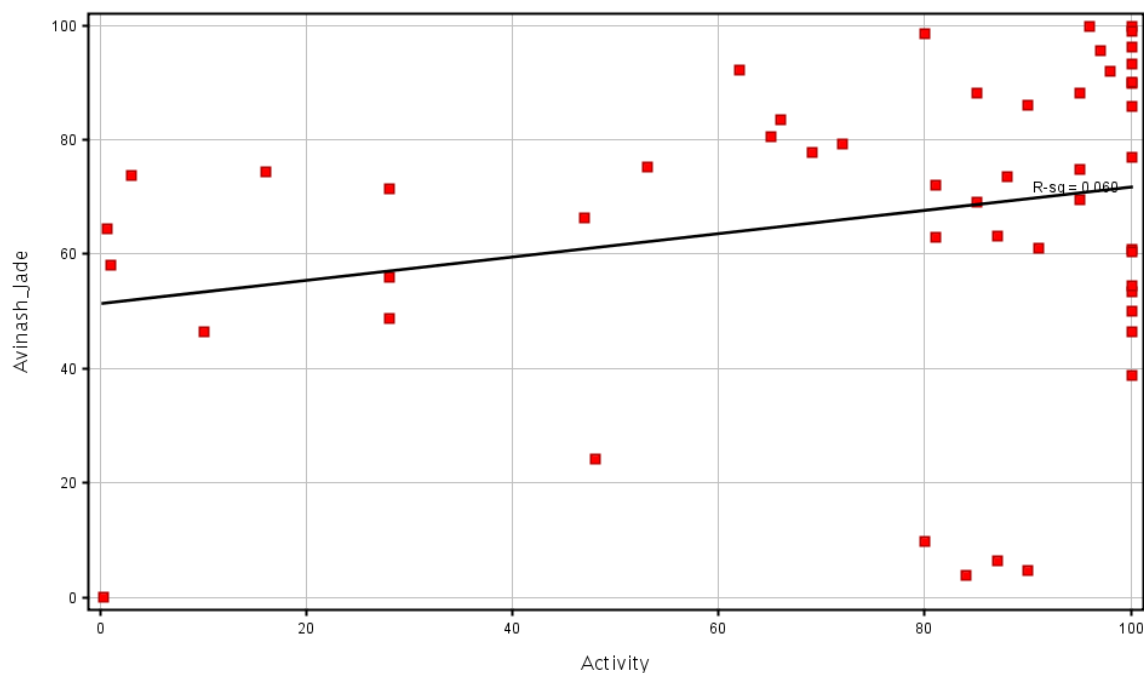
A support vector machine for regression (SVMreg) was used to create a model on the training data with a radial basis function (RBF) kernel. The complexity parameter was set to 6.0, and the descriptor values for the training data were standardized. The model was validated using a ten-fold cross validation procedure.

Training Model Statistics:

R²: 0.8777
RMSE: 14.9765
Cross-validated R²: 0.7367
Cross-validated RMSE: 21.1443

Anthony E. Klon
Research Scientist
Molecular Modeling
Pharmacopeia
P.O. Box 5350
Princeton, NJ 08543-5350
609.452.3676

Avinash_Jade RMSE: 36.93 <avinash_jade@persistent.co.in>

**Description of the method used for Activity Prediction**

All of the descriptors (1081) for the molecules in training and test data were taken from the Sarchitect. We have added six more descriptors viz. Lipinski's rule of five, Rotatable bonds count, Hydrogen bond acceptors, Hydrogen bond donors, Atomic polarizabilities and Bond polarizabilities. Molecules having activity more than and equal to 99 from the training data were removed. Thus training comprises 149 molecules with 1087 features.

Out of these features we got 22 best features that were used by regression tree model to predict the activity. These descriptors are mentioned in the file "Selected Features" table below. These 22 features were then used to build Support Vector Regression (SVR) model to predict the activity. A set of SVR model parameters that has given the least RMSE (21.60) on 10 fold cross validation on the training data were selected. SVR Model built on the training data with the optimal parameters were then used to predict the activities for the molecules in the test data.

Based on our COSMO-RS method we previously set up and published a QSPR model for the prediction of intestinal absorption [1]. This model was trained by a set of 38 chemical diverse compounds with reliable experimental data [2].

We applied the same computational procedure, to generate a QSPR model for the compounds of the training set, see details below for details.

The QSPR equation in terms of σ -moments turned out to be basically the same in both models. However the RMSD value of the new model was significantly worse. We therefore decided to apply our existing model to predict the intestinal absorption of the compounds of the challenge.

Computational details

We generated the QSPR model as follows:

1. 2D \rightarrow 3D conversion by CORINA [3]
2. Geometry optimization on the AM1/COSMO level using MOPAC7 [4]
3. Calculation of the σ -surface on the BP-SVP level applying TURBOMOLE 5]
4. Calculation of the σ -moments by the current COSMO $therm$ version C2.1-01.07 [6]
5. Generation of the QSPR model in terms of five σ -moments M_0 , M_2 , M_3 , M_{acc} , M_{don}

In order to set up a QSPR model, partition coefficients are needed. However, the experimental data of percent intestinal absorption (%Abs) is not themselves partition coefficients. They can be converted into partition coefficients K_{ia} by the relationship

$$K_{ia} = (\%Abs)/(100 - \%Abs)$$

However, this leads to a problem for the experimental values reported as 0% and 100%, respectively, for which K_{ia} cannot be estimated accurately. To use the full data set we define two thresholds %Abs(min) = 3% and %Abs(max)=97%. Experimental values beyond the thresholds are set equal to %Abs(min) and %abs(max), respectively.

For the subsequent regression analysis of $\log K_{ia}$ with respect to the σ -moments, we consider initially the interval between the thresholds. Outside this range the error of the prediction is set to zero if the predicted value is beyond the interval. As a consequence of the treatment of errors outside the interval, standard regression analysis can not be applied. Thus a self written quasi-Newton nonlinear optimization procedure to minimize the sum of the squared residual errors is applied.

This procedure results in the QSPR equation

$$\log K_{ia} = 0.004M_0 - 0.053M_2 - 0.0024M_3 - 0.113M_{acc} - 0.11656M_{don} + 1.37$$

Prediction of intestinal absorption of the set of test compounds.

In order to predict the intestinal absorption of the test compounds the five σ -moments from a COSMO-RS calculation are needed. COSMO $logic$ offers two methods for the calculation of the σ -moments.

1) DFT/COSMO

The σ -moments are calculated from the screening charge densities (σ -profiles) generated from a DFT/COSMO quantum-mechanic calculation which is the time consuming step.

The 3D coordinates, as given in the multi SDF file of the test set, were taken as start geometry. Conformational searches with respect to internal H-bonds were performed with a subsequent AM1/COMSO geometry optimisation. A modified MOPAC7 [4] version was utilized. Subsequently the σ -surface on the BP-SVP level applying TURBOMOLE [5] was calculated.

Finally, the σ -moments are calculated by COSMO $therm$ and the intestinal absorption was predicted by our QSPR model.

This procedure failed in case of compound 28, therefore the value of a COSMOfrag calculation was taken.

2) COSMOfrag

In order to overcome the time consuming step we introduced COSMOfrag, a fast shortcut for high throughput COSMO-RS calculations [7]. COSMOfrag generates a composition of the σ -profile for each individual molecule out of partial σ -profiles taken from a database of precalculated compounds. This algorithm is very effective thus it takes less than five minute to calculate the complete test set.

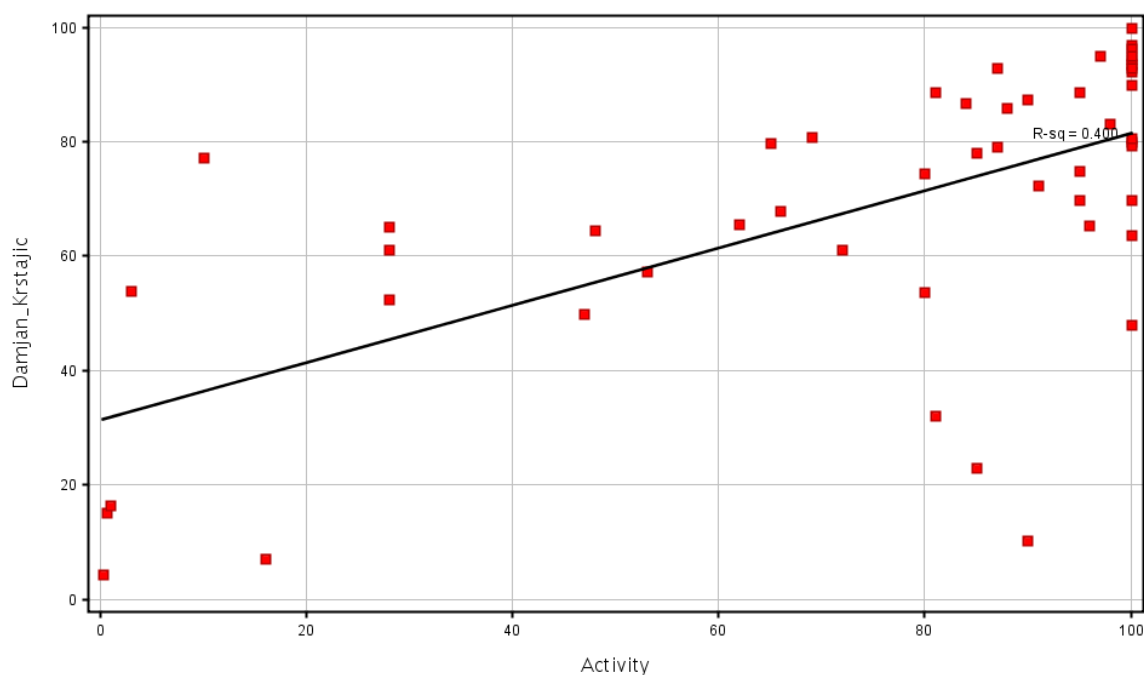
References:

- [1] Jones, R., Connolly, P. C., Klamt, A. and Diederhufen, M. Use of Surface Charges from DFT Calculation To Predict Intestinal Absorption, *J. Chem. Inf. Model* **2005**, *45*, 1337 -1342
- [2] Zhao Y. H., Le J., Abraham M. H. , Hersey A., Eddershaw P. J. , Luscombe C. N., Butina D., Beck G., Sherborne B., Cooper I., Platts J. A. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.* **2001**, *90*, 749
- [3] COSMOtherm, version C2.1 release 01.06; COSMOlogic GmbH &Co. KG: Leverkusen, Germany, **2006**
- [4] CORINA version 2.6, Molecular Networks GmbH, Erlangen, Germany **2001**
- [5] MOPAC7 own modified public domain version
- [6] R. Ahlrichs, M. Bär, M. Häser, H. Horn, C. Kölmel *Chem. Phys. Lett.* **1989**, *162*, 165-169
- [7] Hornig, M., Klamt A., COSMOfrag> A Novel Tool for High Throughput ADME Property Prediction and Similarity Screening Based on Quantum Chemistry. *J. Chem. Inf. Model.* **2005**, *45*, 1169 - 1177

Dr. Carsten Wittekindt
COSMOlogic GmbH & Co. KG
Burscheider Str. 515
D-51381 Leverkusen
Germany

Phone +49-2171-363667
Fax +49-2171-73168-9
Email wittekindt@cosmologic.de
Web <http://www.cosmologic.de>

Damjan_Krstajic RMSE:25.87 <dkrstajic@discoverybus.com>



The models and predictions were automatically generated by Discovery Bus.

For calculating chemical descriptors we used following descriptor calculators:

- a) CDL
- b) PETRA
- c) SArchitect provided by open <http://www.qsarworld.com/> for the competition

We applied our algorithm for feature selection which is based a lot on Mark Hall's CFS (correlation feature based selection) algorithm.

During the processing of the competition data the Bus created several hundred different models and we have chosen 6 different models with different input descriptors. The predictions we are sending are median values of the predictions of the six models. The four models are neural net and there is one PLS and one linear models. They differ in the number of descriptors from 4 - 38. For the attached predictions the number of used descriptors in total is 40.

Our approach is to leave 10% of the training set for later testing and to perform 10 fold cross-validation on the 90% of the training set. We choose the best models on their performance on the 10% of the training set which was not used during any time of model building.

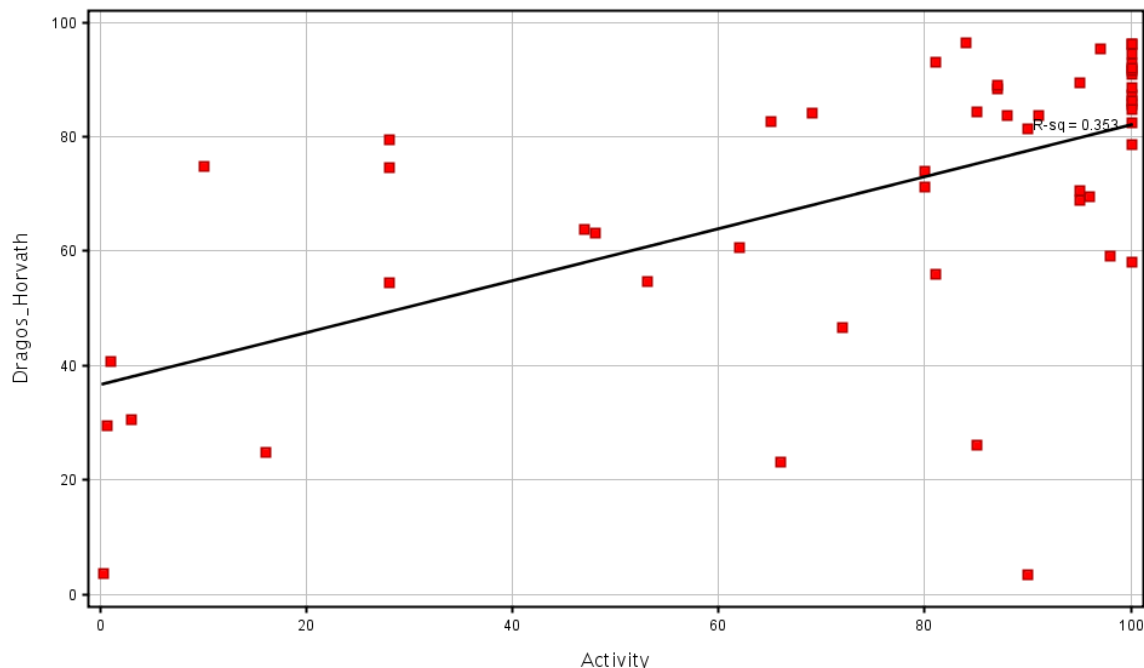
Our best models have still high RMSE on the test set but we thought to give it a go. All the models and predictions were done automatically and only for TEST45 the median was above 100 and I had to manually change it to 100.

On behalf of the Discovery Bus team

Damjan Krstajic

The DiscoveryBus team (David E Leahy, Damjan Krstajic and Vladimir J Sykora).

Dragos_Horvath RMSE:26.54 <horvath@chimie.u-strasbg.fr>



Brief description of the modeling approach:

1. Descriptor preselection procedure: Whole-molecule lipophilicity descriptors (LogP, LogD, polar surface area, BCUT descriptors – all calculated by the ChemAxon tool *generatemd*¹), topological pharmacophores (2-point ChemAxon fingerprints and fuzzy triplets²) and ISIDA³ fragment descriptors were all deemed to be potentially relevant for oral absorption modeling. Using versions FPT-1 for the fuzzy triplets and II(Hy) for ISIDA fragments, the number of candidate descriptors would have exceeded 6000. Direct descriptor selection out of such a large pool being unfeasible, three independent preselection campaigns were conducted for ChemAxon (global lipophilicity + 2-point topological pharmacophore fingerprints), fuzzy triplets and ISIDA fragments. To this purpose, the training set (187 compounds) was further split into a Learning (148) and a Validation (39) subset⁴, and subjected to separate linear and non-linear model building attempts using the Stochastic QSAR Sampler⁵ (SQS) on three independent workstations. Out of the thousands of successfully cross-validating models, the ones having both the Learning and Validation set correlation coefficients above 0.6 were retained, and 262 descriptors entering either of these equations were kept for the main model building tournament. All initial categories were represented within the pool of 262 chosen terms.

¹ www.chemaxon.com

² Bonachera, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 1 - Topological Fuzzy Pharmacophore Triplets and adapted Molecular Similarity Scoring Schemes. J. Chem. Inf. Mod. 2006, 46, 2457-2477

³ Solov'ev, V. P.; Varnek, A.; Wipff, G. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. J. Chem. Inf. Comput. Sci. 2000, 40, 847-858

⁴ Splitting was random, but ensured a similar distributions of high and low oral absorption values in both Learning and Validation sets

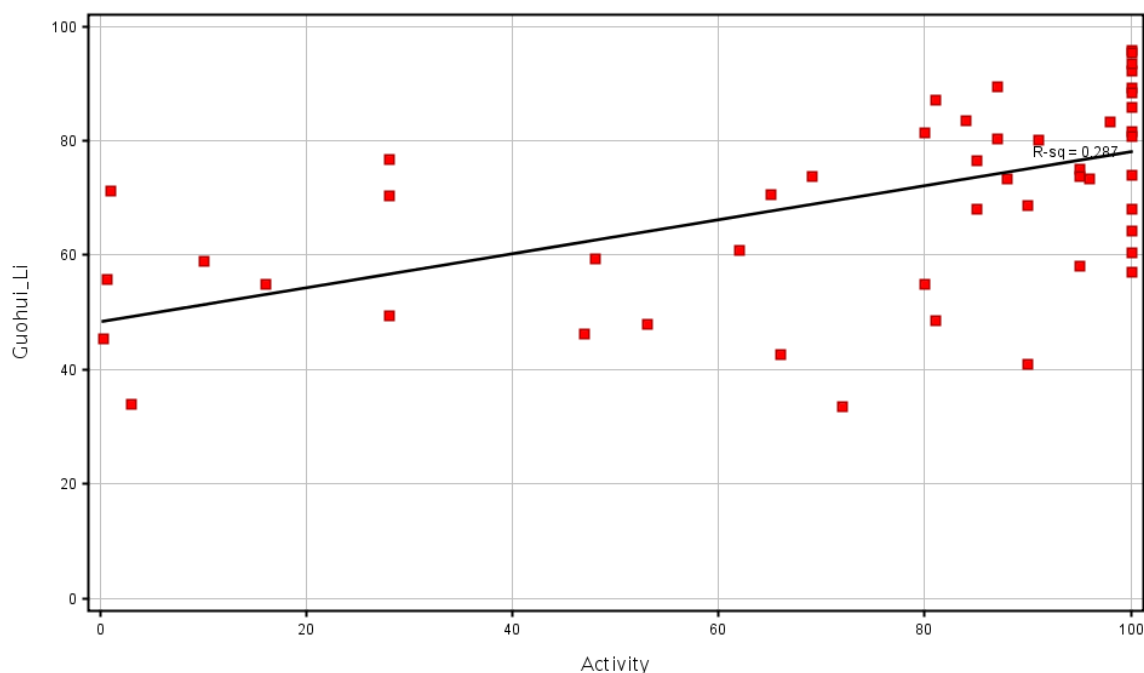
⁵ Horvath, D.; Bonachera, F.; Solov'ev, V.; Gaudin, C.; Varnek, A. Stochastic versus Stepwise Strategies for Quantitative Structure-Activity Relationship Generation - How much effort may the mining for successful QSAR models take?. J. Chem. Inf. Model. 2007, 47, 927-939.

2. **Main model building:** The training set was subjected to a systematic five-fold splitting into Learning and Validation sets, such as to ensure that each compound is left once in the Validation set. Linear, polynomial and fully non-linear models were then generated using the SQS engine for each of the five splitting schemes. Globally, the evolutionary sampling of the activity-structure landscape (all splitting schemes and non-linearity policies confounded, and including the SQS runs at the preselection stage) led to the discovery of 32009 individual QSAR equations selected due to their high fitness (cross-validated correlation coefficients on Learning sets) – out of many millions explored. Out of these, a final pool of 3733 models assembled all equations with a proper validation behavior ($R^2_{\text{Learn}} > 0.6$ and $R^2_{\text{Validation}} > 0.5$). Each model was assigned a weight, proportional to $(R^2_{\text{Learn}} - 0.6) * (R^2_{\text{Validation}} - 0.5)$, and rescaled such that the best scored model weighs 10 times as much as the worst.

3. **Prediction:** Oral absorptions for the test compounds were calculated according to each model from the final pool. Individual predictions were weighed according to the above outlined model relevance score. Reported values for test compounds represent weighted averages, accompanied by weighted prediction variances. Lower variances obviously represent higher prediction confidence. Encouragingly, for 28 compounds out of 51, the 3733 independent models basically agree on the predicted values, with an average dispersion (variance) below 10%. For 20 more, the variance still remains below 10%, whereas the remaining 3 molecules witness some significant disagreement among the various models.

Laboratoire d'Infochimie, UMR 7177 CNRS-ULP
Strasbourg, France

Guohui_Li RMSE:27.23 <lgh19721031@hotmail.com>



The following is my summary for my modeling:

Parameters:

Descriptors: Entended Connectivity Finger Print 2

No feature selection used.

I used SVM to do modeling.

Number of Support Vectors: 168

Cross-validation results:

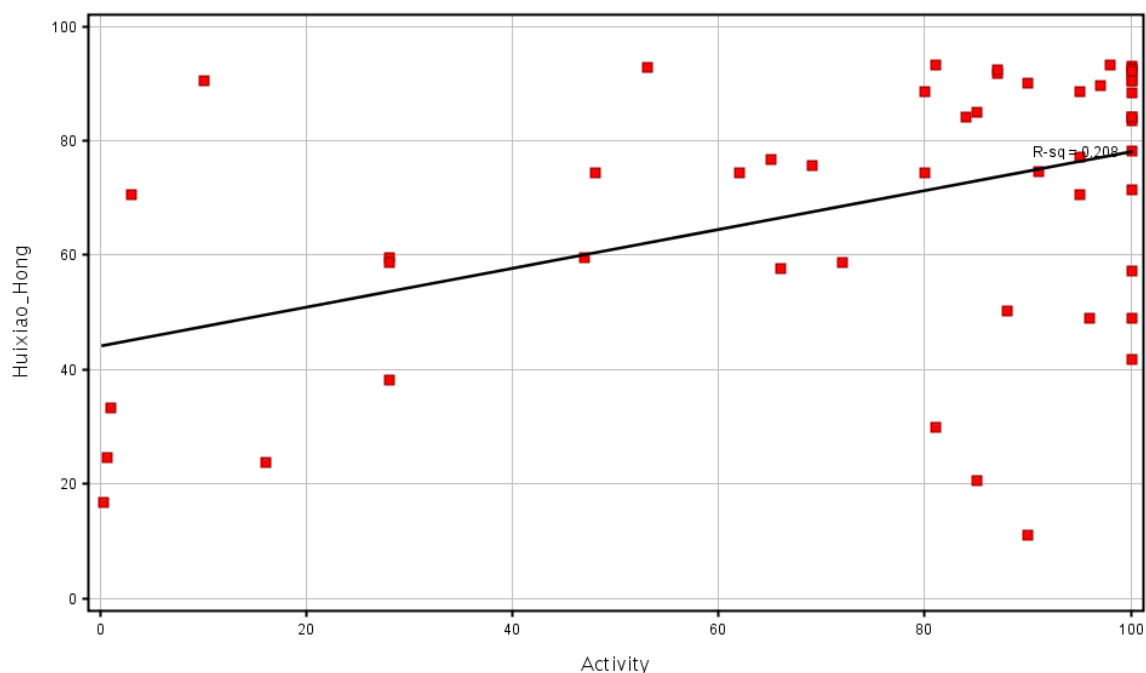
RMS Error: 27.29

Q-Squared: 0.2534

All-data model results (non-cross-validated):

RMS Error: 14.84

R-Squared: 0.8387

Huixiao_Hong *RMSE: 30.26* <huixiao.hong@fda.hhs.gov>

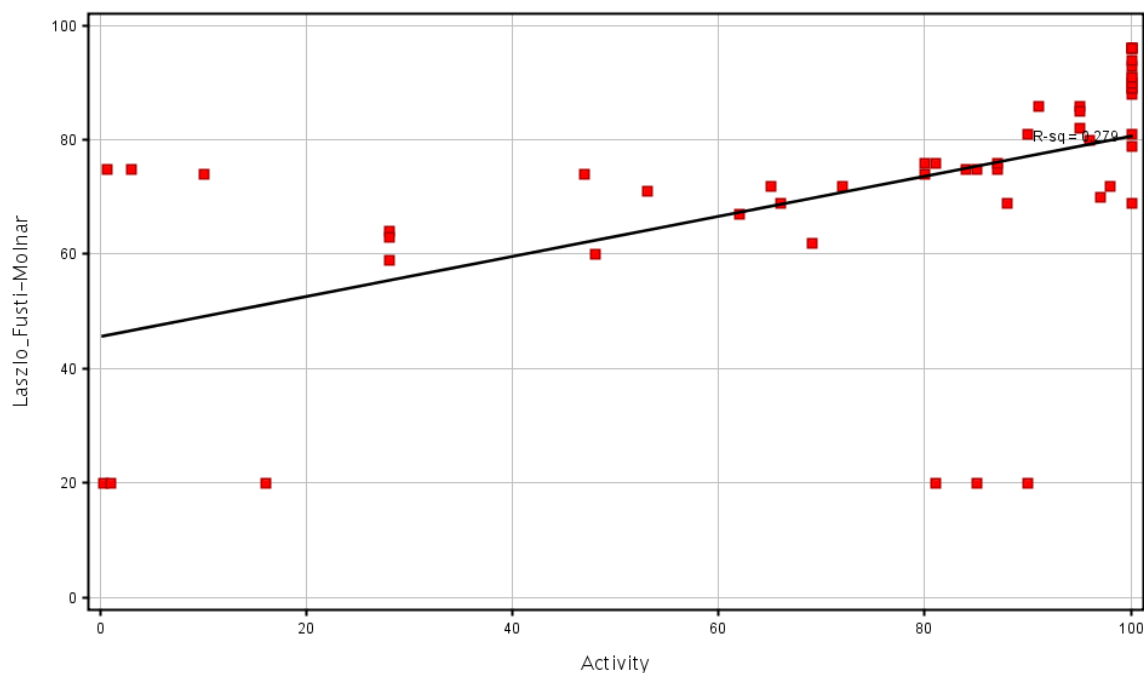
The modeling procedures are described below:

- 1) Sarchitect descriptors (1081) were used, which were obtained from QSAR-World;
- 2) Weighted kNN model with Euclidean distance;
- 3) Descriptors filtering:
 - 3.A) Remove 128 descriptors with more than 50% of values of zero;
 - 3.B) Remove 166 descriptors with Shannon Entropy less than 1.5;
 - 3.C) Remove 302 descriptors with correlation coefficient to the activity less than 0.01;

- 4) Data transformation: As the descriptors have different units, transforming to same unit by auto-scaling was conducted on the training data set, and the scaling parameters from training data set were applied to the tests data set.
- 5) Descriptor (variable) selection: This process was integrated into model parameter tuning. Sequential forward selection method was used;
- 6) Model parameter (k) was tuned simultaneously with descriptor selection, using leave-one-out cross validation;
- 7) Optimized model: k=13, one descriptor (tHAccVSA) was used.
- 8) Results: q2 of leave-one-out is 0.319.

Huixiao Hong, Ph.D.
Division of Systems Toxicology
National Center for Toxicological Research
U.S. Food and Drug Administration
3900 NCTR Road, Building 5, Room 5B-112C
Jefferson, AR 72079
Phone: 870-543-7296
Email: Huixiao.Hong@FDA.HHS.GOV

Laszlo_Fusti-Molnar RMSE: 27.58 <fusti@qtp.ufl.edu>



1. Model description:

I have chosen one dimensional distribution functions of the atomic electrostatic potentials as my descriptor. The descriptors are obtained by using my newly developed qsar code that is interfaced with a first principle (ab initio) package. The molecular orbitals are obtained from the ab initio results by using B3LYP/3-21G* geometry optimizations followed by B3LYP/6-31G* single point calculations. The statistical calculations and the activity estimations are based on the pls fit.

2. Programs used:

- a. Qchem ab initio package for geometry optimization and single point DFT energy calculations.
- b. Own program package to obtain the electrostatic potentials from the molecular orbitals and to create the qsar descriptors
- c. Own pls implementation based on the original paper of Wolg.

3. Results:

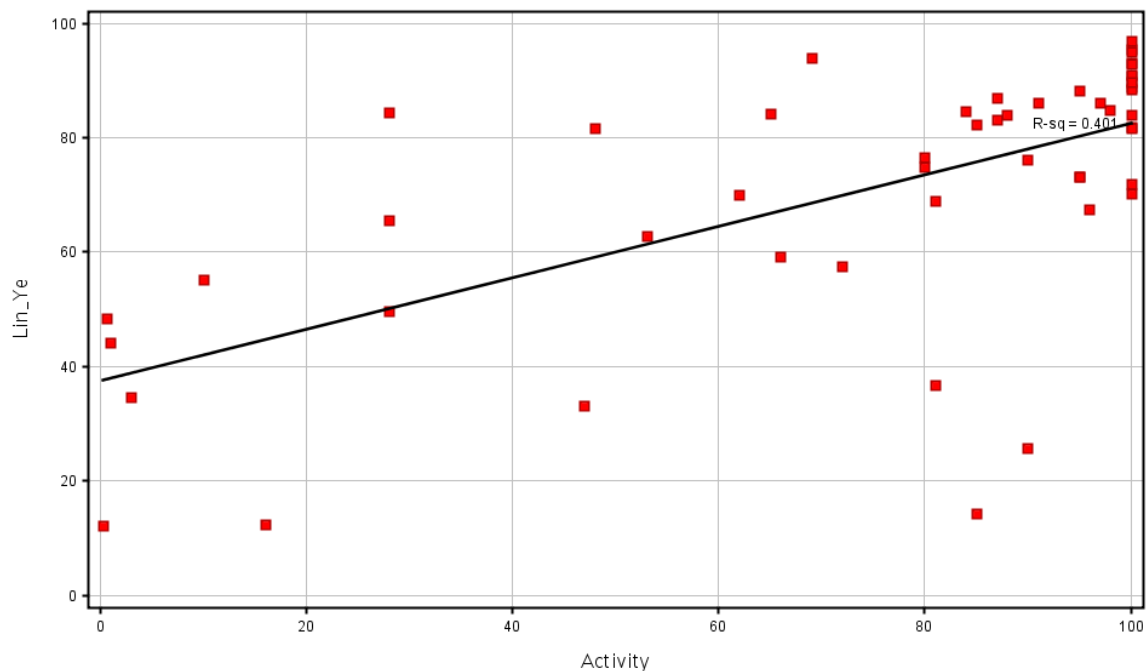
I had to throw away 3 systems from the training set because of convergence problem. Thus I have used 184 molecules for the training set. The geometry optimizations were truncated to 50 cycles from the initial geometries, about 30% did not converged yet but those resulting geometries are considered to be good enough. Due to the lack of time and the necessary program package there was no ab initio based search performed for alternative conformations, PH states etc.

The pls fit for the training set resulted $R^2=0.55$ and $Q^2(\text{leave one out})=0.50$ by using 5 PCs and no outliers. I have realized that my estimation for the very low activity compounds have huge inaccuracies perhaps due to the lack of bad conformations, lack of internal H bonds etc. I have worked out a new filter program based on the Oxygen, Nitrogen and Sulfur contributions total electrostatic potentials of the individual systems to try to filter out those low activity compounds which I cannot predict at the moment. From the training set I filtered out 14 compounds out of 184, 8 of them has less than 30 activity values and there is only one compounds with more than 50 activity value. The averaged activity of those compounds is about 20. The same filter program was used for the test set and six compounds out of 51 test molecules were filtered out.

To maximize the predictability of my model I have decided to throw away some outliers from the training set by maximizing the Q^2 values. Leaving out 7 compounds resulted of $R^2=0.76$ and $Q^2=0.70$ and leaving out 11 compounds resulted of $R^2=0.85$ and $Q^2=0.78$. Most of the outliers are either very low activity compounds or activities with the truncated 100 values. I have made five different predictions by using the training set with 7,8,9,10,11 outliers and my final predictions for the test set is the equally weighted average of those. They are listed in Table 1.

*Laszlo Fusti-Molnar PhD
University of Florida,
Quantum Theory Project
PO BOX 118435
Gainesville, FL, 32611-8435
Phone: (352)-392-7554
Fax: (352)-392-8722
e-mail: fusti@qtp.ufl.edu*

Lin_Ye RMSE: 25.06 <lye@email.unc.edu>



Report Summary:

Outliers are removed from the original modeling set based on the similarity search (descriptor-based). Final 153 molecules are used to build the model. The applicability domain (AD) is calculated according the 153 molecules. This AD is used to screen the 51 external test set. 39 molecules fall within this AD.

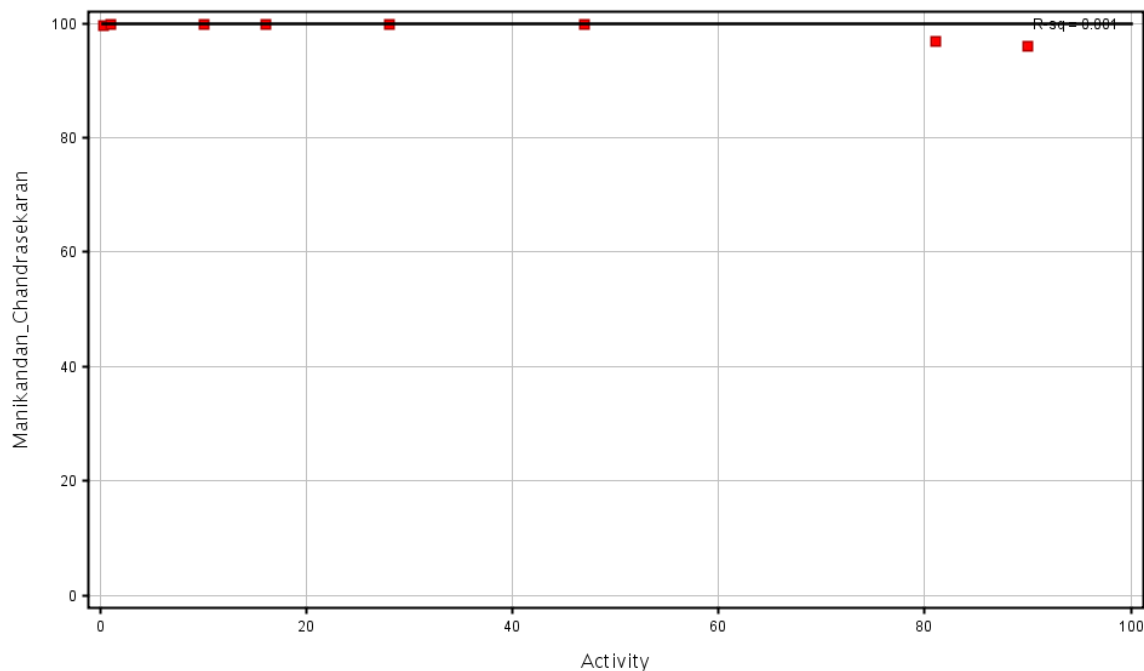
Several different approaches are tried (KNN, SVM, PLS with different descriptor package Dragon, MolConnZ and MOE). Two results (see excel file) are reported here

Approach	Descriptors	# of Descriptor	Computational time	Model development method	Models used for prediction
1	Dragon	736	Two hours	KNN	Consensus model
2	MOE	184	Within minutes	PLS	Single model

If you can consider both results, please pick up the better one. If you only can take one result, please take the first one. **(We took only the first one – QSAR-World)**

Lin Ye
 Molecular Modeling Lab at UNC
lye@email.unc.edu

Manikandan_Chandrasekaran *RMSE: 41.04* <chandrasedkaran.manikandan@gmail.com>



ALGORITHM FOLLOWED

Two methods were developed as possible solutions. Both of these below mentioned methods have been worked out. However the complete solution to the problem became possible with the second method.

The first method involved the tedious process of obtaining the molecular descriptors from e-Dragon (www.vcclab.org) and screening for the desired molecular descriptors. And finally obtaining a regression equation or developing a neural network that will help in predicting the unknown. The process is termed as tedious, as there are 1600 molecular descriptors and also high throughput process became quite impossible in the provided sdf file format. Hence there is a need for conversion of the file to SMILES notation, for which E-BABEL, again from the www.vcclab.org was used. Each of the 187 training set molecules were converted to SMILES format and finally they were sent to the server for the desired properties. However the process of screening for the desired properties was planned as by means of regression analysis, for which the 7-day evaluation version of the NCSS-2007 was downloaded. When regression analysis was performed with the obtained set of results, erroneous results (noise) occluded the process of arriving at a proper conclusion. Thus due to the time constraint and others, the search for a solution utilizing free online resources ended up. The second alternative means was worked out.

The second method involved the process of using the older version V6 (evaluation copy) of Viewer Pro software. This software provides the critical molecular descriptors of compounds, when provided with the compound's sdf files. Here the screening process was done manually i.e. the critical features required for describing a compound were analyzed and picked up. The manual curation of data also involved the process of picking up two molecular descriptors which essentially contain the same information. These features are molecular weight and exact molecular weight. The process was done so, in order to analyze the importance of these independent variables. Thus a neural network was designed and optimized using the NeuralTools evaluation version. After the process of optimization, the test data whose activities are to be determined are provided to the network and the results were obtained.

THE 13 MOLECULAR DESCRIPTORS USED TO DEVELOP NEURAL NETWORK

- Exact Molecular Weight
- Molecular Weight
- Molecular Volume
- Jurs-RNCS – Jurs Relative Negative Charge Surface Area
- Jurs-RPCG- Jurs Relative Positive Charge
- Jurs FPSA- PPSA (partial positive surface area = sum of surface area on positive parts of molecule) / total molecular surface area
- Jurs WPSA –PPSA * total molecular surface area / 1000
- Jurs PPSA
- Number of atoms
- Lipinski's violations
- Number of Hydrogen acceptors
- Number of Hydrogen donors
- Number of Chiral Centers
- Number of rotatable bonds

NEURALTOOLS QUICK SUMMARY (TRAIN-TEST)**Net Information**

Name: Net Trained on train

Configuration: MLFN Numeric Predictor (nodes: 2, 2)

Location: NET1.ntf

Independent Category Variables: 0

Independent Numeric Variables: 14

Dependent Variable: Numeric Var. (ACTIVITY)

Training

Number of Cases: 150

Training Time (h:min:sec): 08:30:58

Number of Trials: 71763154

% Bad Predictions (30% Tolerance): 12.6667%

Root Mean Square Error: 4.899

Mean Absolute Error: 3.129

Std. Deviation of Abs. Error: 3.770

Testing

Number of Cases: 37

% Bad Predictions (30% Tolerance): 16.2162%

Root Mean Square Error: 10.71

Mean Absolute Error: 6.709

Std. Deviation of Abs. Error: 8.351

Data Set

Name: train

Number of Rows: 187

Manual Case Tags: NO

Variable Impact Analysis

Molecular Weight: 19.1410%

Exact Mol. Weight: 15.9264%

Jurs-WPSA-1: 11.4340%

Number of Atoms: 11.3608%

Molecular Volume: 9.7799%

Number of Hydrogen Bond Acceptors: 3.2505%

Jurs-PPSA-2: 2.2766%

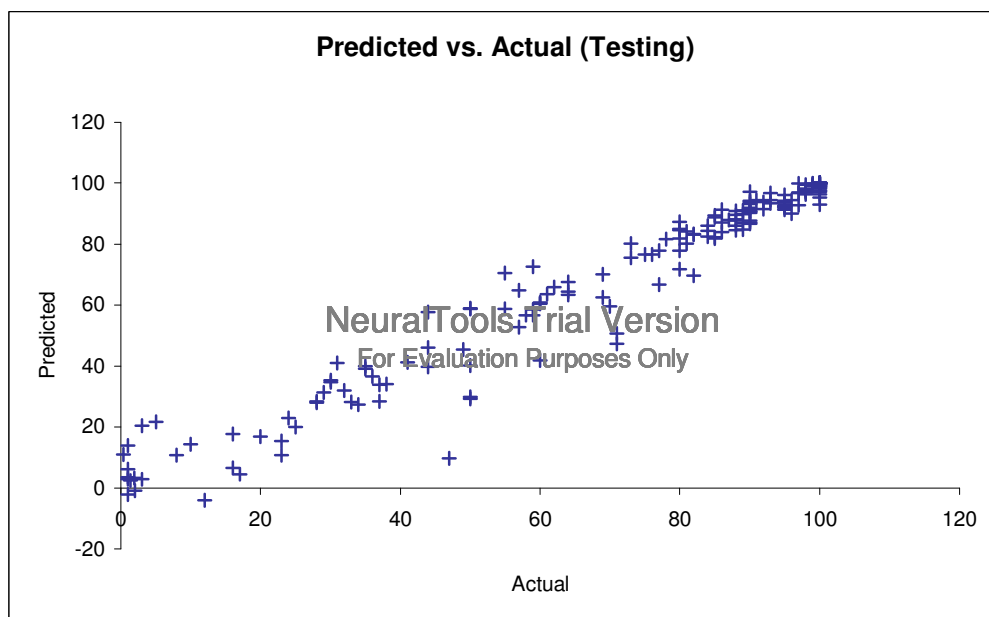
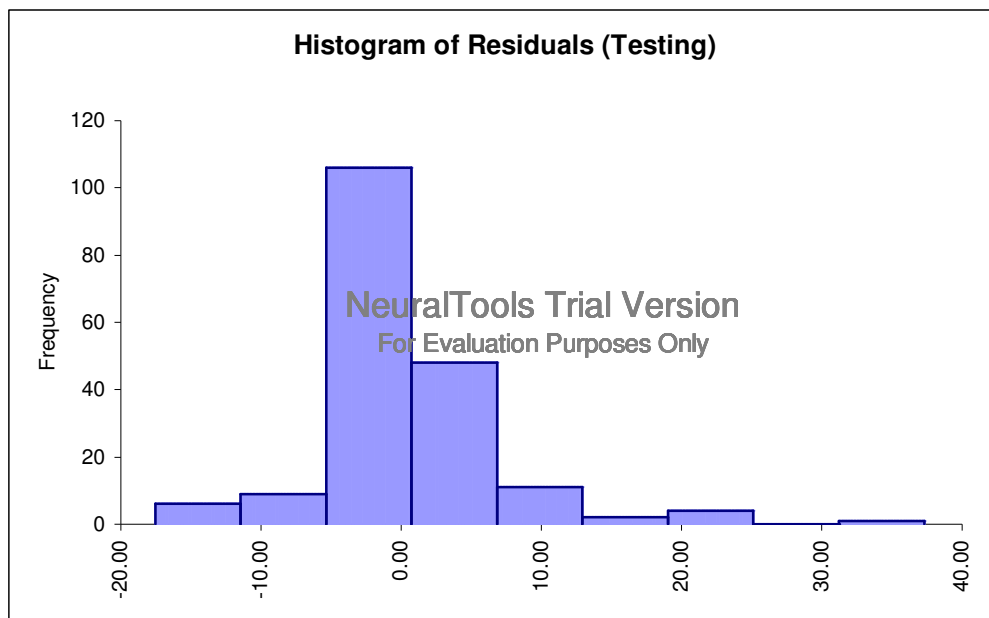
Jurs-FPSA-3: 2.2417%

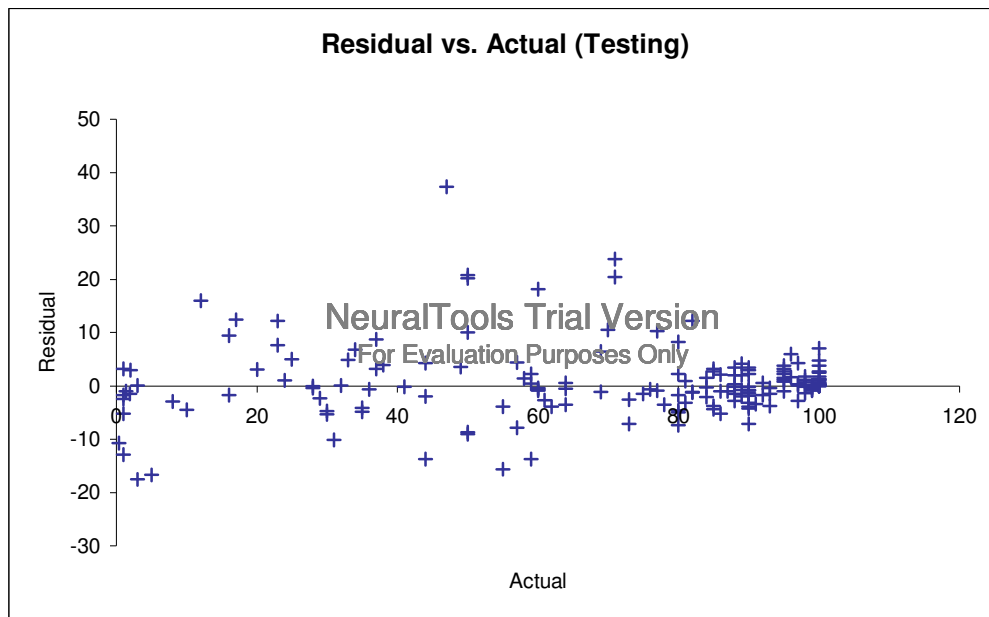
Number of Hydrogen Bond Donors: 1.9853%

Jurs-RPCG: 1.6140%

Number of Rotatable Bonds: 1.5209%
 Lipinski Violations: 1.5151%
 Jurs-RNCS: 1.4420%
 Number of Chiral Centers: 1.4248%

GRAPHICAL REPRESENTATIONS (3):

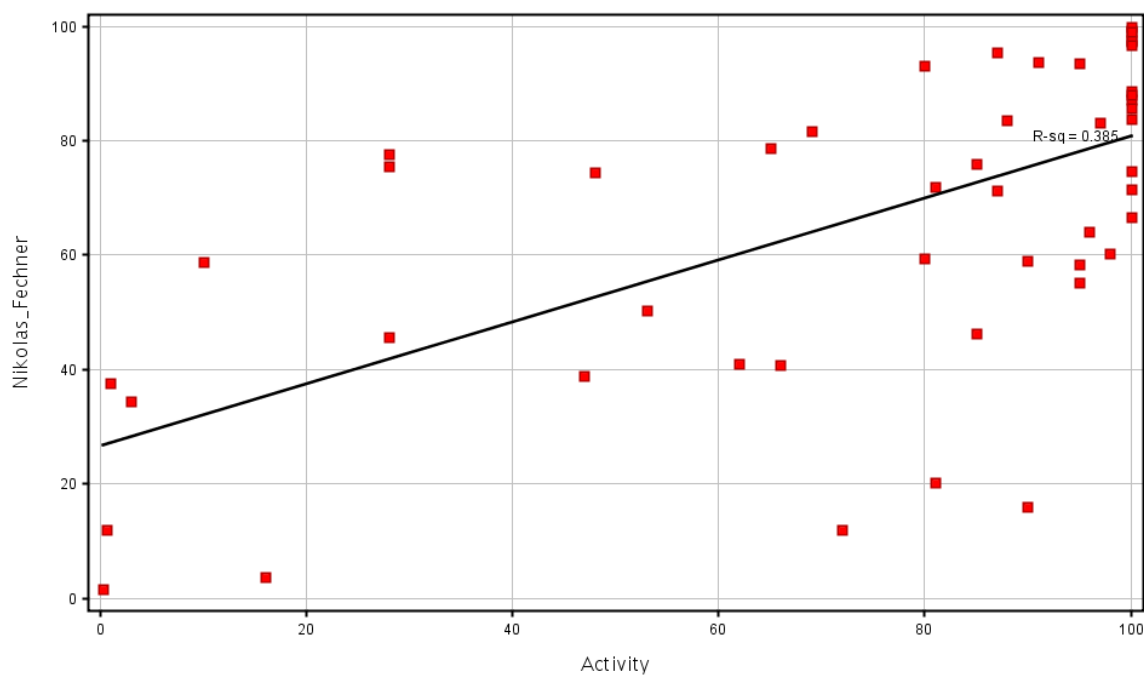




Manikandan Chandrasekaran
Government College of Technology
Coimbatore

Nikolas_Fechner RMSE:27.32

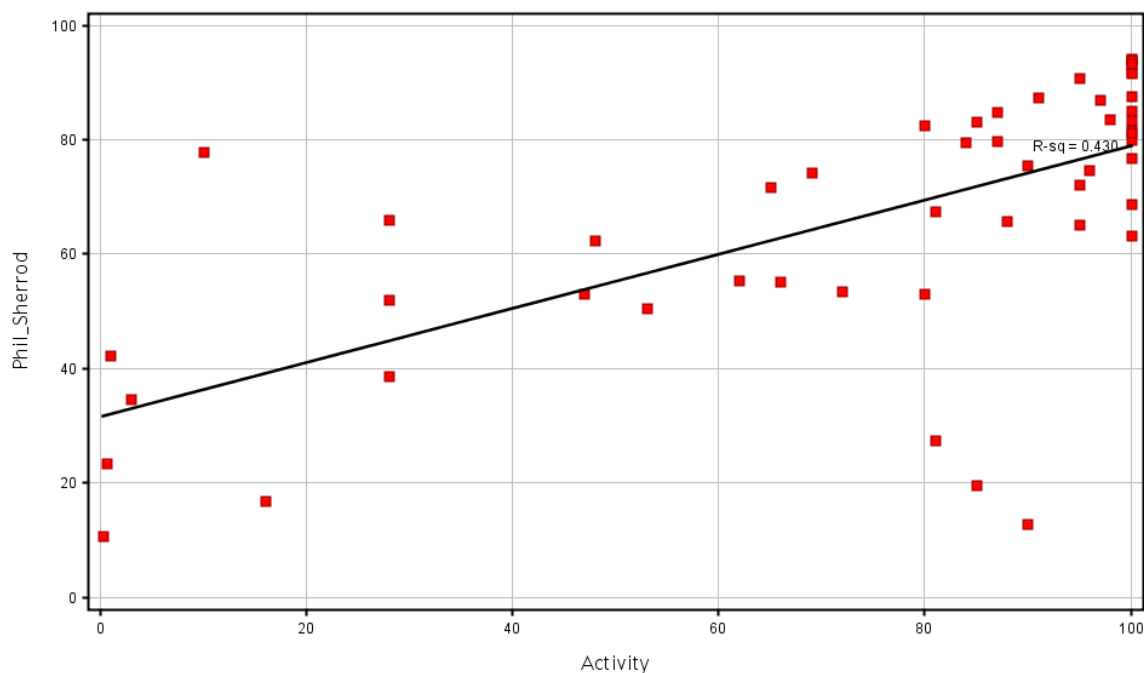
<nikolas.fechner@uni-tuebingen.de>



We used a nu-Support Vector Regression (LibSVM) with a precomputed kernel matrix. The matrix was calculated using an improved version of the optimal assignment kernel (Fröhlich et al. "Kernel function for attributed molecular graphs - a similarity based approach to ADME prediction in classification and regression", QSAR Comb. Sci, 25:317-326, 2006), that incorporates the molecular flexibility into the measure. The method has been submitted (Fechner et al. "Incorporating molecular flexibility into a structured kernel for ligand-based virtual screening", Bioinformatics, submitted) but not published yet. The soft margin parameter has been optimized via a nested cross-validation on the training set (Squared correlation coefficient on the training set: 0.91).

Nikolas Fechner and Georg Hinselmann
Dipl.-Inform. (Bioinformatik) Nikolas Fechner
Center of Bioinformatics - Dept. Prof. Zell
University of Tuebingen
Sand 1
72076 Tuebingen
Tel. +49 (0) 7071 29 77174
Fax. +49 (0) 7071 29 5091
Mail: nikolas.fechner@uni-tuebingen.de
<http://www-ra.informatik.uni-tuebingen.de>

Phil_Sherrod RMSE: 25.35 <philsherrod@comcast.net>



A TreeBoost model was created consisting of 380 boosted decision trees. TreeBoost is an implementation of Stochastic Gradient Boosting. My DTREG program (<http://www.dtreg.com>) was used to perform the analysis.

Feature selection was performed automatically by the model building program. Here is a list of the relative importance of the most important 20 predictor variables:

Variable	Importance
tPSA	100.000
tHAccSA	68.772
MLOGP	67.722
fHplVSA	56.820
nHAcc	33.613
tHDonorVSA	29.766
Hy	27.209
H4m	26.923
GATS2p	25.392
tHDonorSA	23.739
Hy	27.209
H4m	26.923
GATS2p	25.392
tHDonorSA	23.739
MATS1e	22.551
HATS7p	16.162
R5p+	14.530
Mor10m	13.496
tHplSA	13.134
GATS1e	12.160
R5v+	11.476
SEigZ	11.319
ATS8e	10.261
RDF060e	9.673

For a TreeBoost model, the relative importance of the variables is calculated based on the reduction in variance (error) for the splits performed using the variable.

Here is the validation information for the training data using 10-fold cross validation. As you can see, the model explained about 47.9% of the variance:

--- Validation Data ---

Median target value for initial data sample = 85

Mean target value for initial data sample = 70.998396

Mean target value for predicted values = 70.539019

Average absolute error for initial data sample = 24.065775

Average absolute error after tree fitting = 17.734833

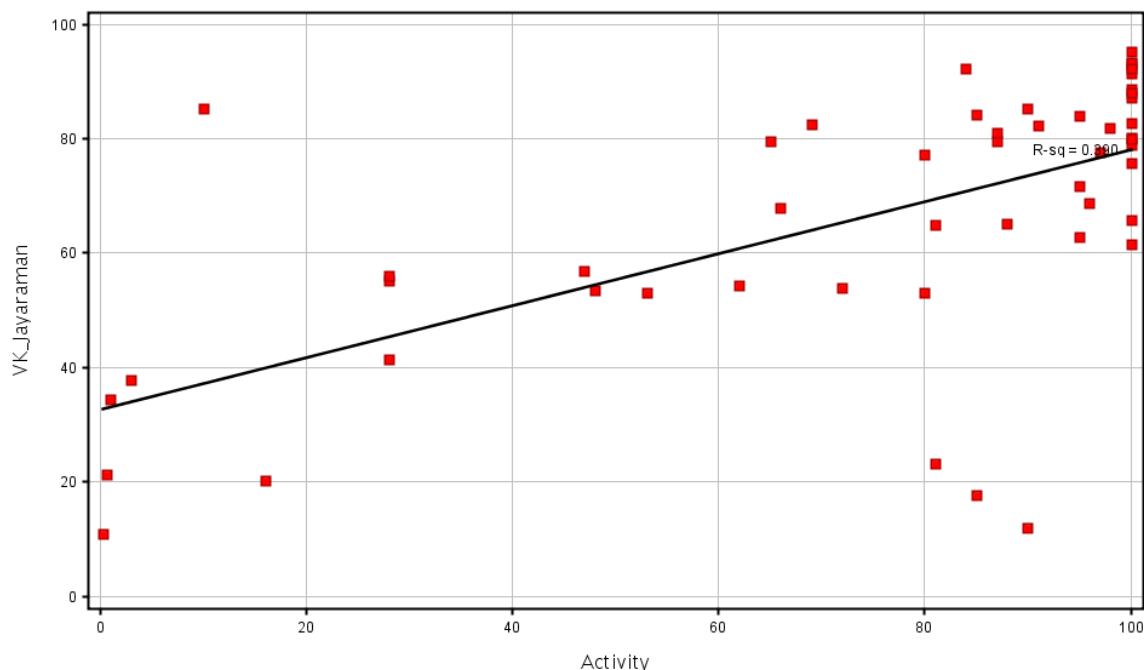
Reduction in absolute error = 26.307%

Variance in initial data sample = 1149.5226

Residual (unexplained) variance after applying TreeBoost model = 599.26279

Proportion of variance explained by TreeBoost model = 0.47869 (47.869%)

VK_Jayaraman RMSE: 26.36 <vk.jayaraman@ncl.res.in>



Team headed by: Dr. V.K. Jayaraman

Team members: Joydeep Mitra (email: mitra.joy@gmail.com), Prashant Shingade (email: prashant.shingade@gmail.com), Dr. V.K. Jayaraman (email: vk.jayaraman@ncl.res.in)

Descriptors

A set of 1801 descriptors was provided by the editor of QSAR World, on request. Smaller subsets of features were selected using randomForest raw importance score as described below.

Methodology used for regression model

Random Forest regression algorithm, originally developed by Breiman et al. [1,2] was employed. R scripts written using the "randomForest" [3] package were used for all the analyses.

Source:

Random Forest homepage: <http://stat-www.berkeley.edu/users/breiman/RandomForests>

R interface: <http://cran.r-project.org/src/contrib/Descriptions/randomForest.html>

Feature selection

Random Forest can also be used to get an estimate of the variables that are less important for classification. All the cases that are OOB for a particular tree are put down the tree to get a classification with some votes for each class. Now to get an estimate of variable importance, the value of each of the attributes is randomly permuted in the OOB cases of a particular tree and the decrease in the number of votes for the majority voted class is calculated. This decrease in the number of votes, when averaged over all the trees in the forest, gives the raw importance score for that variable. So higher the raw importance score, greater is the importance of that variable in classification. Thus the raw importance score can be employed for feature ranking. The best-ranked features are then selected and are used for classification.

Tuning for selection of best feature subset

Various subsets of features were extracted using the raw importance score from a model built using the training dataset. Each subset of this training data was then divided into five random splits of train and test, leaving out around 20% of the data in each set. A 10-fold cross-validation was performed on each of the five training subsets to tune the parameters for random forest. A random forest model was constructed using these parameters, which was then validated using the test subset. Finally, a 10-fold cross-validation was done using the best feature subset and all examples, with the tuned parameters. The value of the parameter **MTRY**, which gave the best cross validation Spearman Rank Correlation Coefficient (**SRCC**) was used to build the final random forest model, and the end point values for the blind dataset was predicted using this.

Results

On averaging out the results from all the five splits, the set of 25 descriptors was found to show the best performance, in terms of cross-validation mean squared error, cross-validation SRCC, mean squared error on test subset and SRCC on the test subset. These 25 descriptors were used to construct the final model for prediction of the Activity values on the blind dataset.

Selected Features:

Features Used

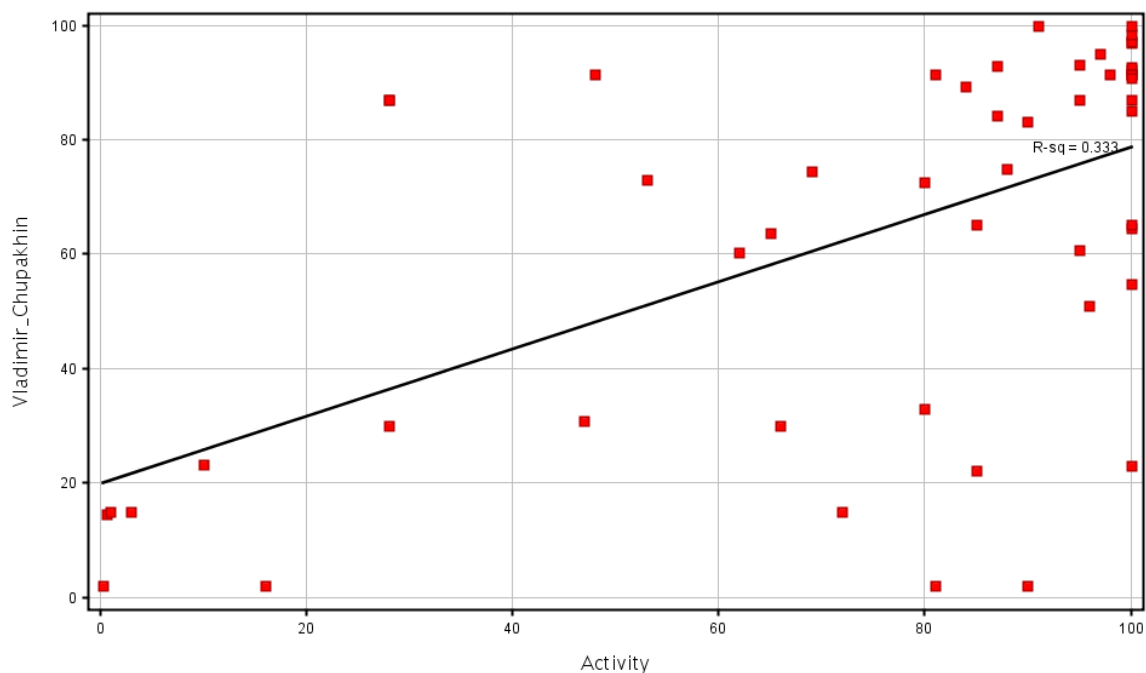
Hy
tHDonorSA
tHDonorVSA
MLOGP
tHAccVSA
tHAccSA
nHDon
tPSA
RDF060m
fHDonorVSA
fHDonorSA
H4m
T.O..O.
Mor01m
PW5
SEigv
fHplVSA
R7p
PW4
GATS2p
fHAccVSA
nHAcc
MATS1e
SEigp
Mor10m

References:

- [1] Breiman, L.; Cutler <http://www.stat.berkeley.edu/users/breiman/RandomForests/>. **2004**.
- [2] Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*, **1984**. Chapman & Hall, New York.

[3] Liaw, A.; Wiener, M.; *The randomForest package*, <http://cran.r-project.org/doc/packages/randomForest.pdf>

Vladimir_Chupakhin RMSE: 31.51 <chupvl@gmail.com>



I have used commercial solution ChemTree. Statistical Algorithm: Recursive Partitioning (Decision Trees)

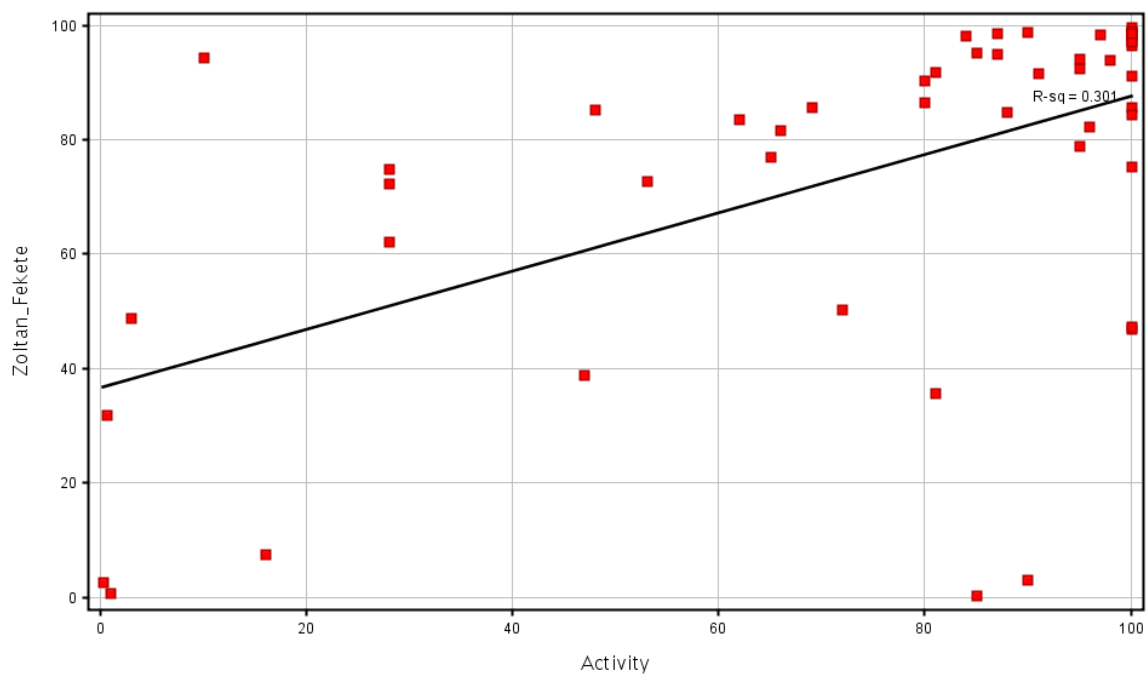
Molecular Descriptor: Atom Pair

Descriptor selection/reduction: minimal occurrence = 5

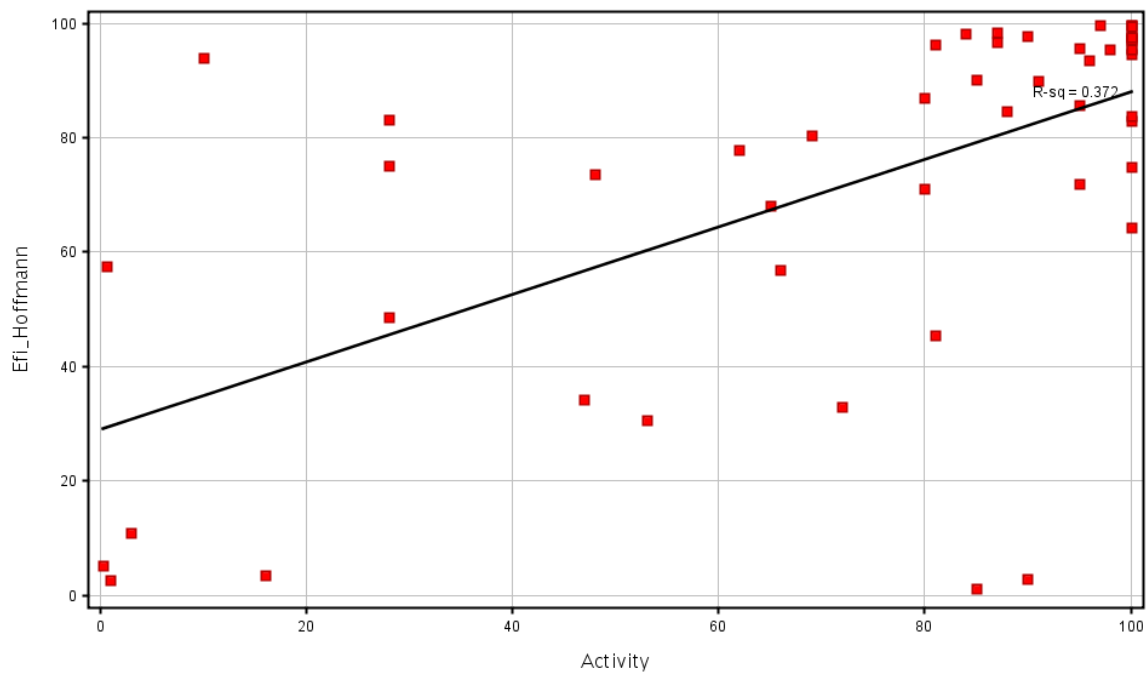
p value for model building is 0.99

I have made 100 trees and selected 5 best for prediction, so the result was the mean of the 5 best trees.

Zoltan_Fekete *RMSE: 29.36* <zoli.fekete@gmail.com>



Efi_Hoffmann *RMSE: 27.82*



Summary report on QSARWorld modeling challenge contribution

Zoltan A. Fekete[†] and Eufrozina A. Hoffmann (U. Szeged)

[†]email: QSARWORLD.ZAF@DFGH.NET

From the outset we decided to use ordinary Multivariate Linear Regression modeling, with descriptors accessible via freely available software. The basic tool applied for handling molecular data, as well as calculating many descriptors, is the OpenBabel package (version 2.1.1, <http://openbabel.sourceforge.net>), supplemented by our own Python scripts through its Python interface 'pybel'. This supplies built-in calculation of cLogP, MR and PSA values; extracting the parameters considered in Lipinski's 'Rules of 5' (Advanced Drug Delivery Reviews 1997, **23**, 3): HBA, HBD, RotBonds and molwt was done with a simple PyBel script (<http://baoilleach.blogspot.com/2007/07/pybel-hack-that-sd-file.html>).

We transform the %F values with the atanh (area tangent hyperbolic) function, to enforce the range to be 0-100% and provide a better distribution of points; *i.e.* instead of the conventional simple fit over raw %F, we do MLR over $\text{atanh}(\%F/100\% \cdot 2 - 1)$ as dependent variable, and the prediction is calculated from the regression back-transformed as $\%F_{\text{predicted}} = \tanh((Y+1)/2 \cdot 100\%)$. We also use a 0.2% boundary limit, substituting $\max(0.2\%, \min(99.8\%, \%F))$. After exploring the data we pruned the set: omitted train_32, as all other molecules have less than 150 atoms; deleted train_19 and train_147, which have identical structures but wildly different %F.

For variable selection we made use of a recently published human oral bioavailability database on 768 molecules by Tingjun Hou *et al.* (Journal of Chemical Information and Modeling, 2007, **47**, 460). Also utilized were the atomic contribution SMARTS patterns developed for the HlogS algorithm of Hou *et al.* (J. Chem. Inf. Comput. Sci. 2004, **44**, 266), as well as those suggested for pKa prediction by Roger Sayle (<http://www.daylight.com/meetings/emug00/Sayle/pkapedict.html>).

Furthermore, extensive use was made of the ALOGPS software (version 2.1, <http://vcclab.org/lab/alogps>) from Igor Tetko (J. Chem. Inf. Comput. Sci., 2002, **42**, 1136), as well as other tools from the VCCLAB suite (Mini Rev Med Chem, 2003, **3**, 809). Its standard operation yielded AlogP and AlogS values. We used its LIBRARY learning mode with the Hou database to generate AlogD values, too. We devised a special ALOGPS protocol for its neural network perception capability to handle bioavailability data in the following way: values from qsar-challenge_Train_Set.sdf were combined with those from Hou, and converted into an ALOGPS library file; this was used to replace the built-in AlogP and AlogS libraries (with blanked-out logs.bin and logp.bin) in learning mode - therefore, instead of logP or logS prediction, two new descriptors were generated as proxies for $\text{atanh}(\%F)$, which we designate athA_ANNP and athA_ANNS (for they are made with the associative neural nets contained in the ALOGPS software).

Based on our study of the Hou database, we had developed 4 composite descriptors as linear combinations of individual ones. These are all 2-D descriptors, which can be calculated from molecular topology alone (as encoded *e.g.* in SMILES format). COMP ALOGPS has 9 components: AlogP, AlogS, AlogD, athA_ANNP, athA_ANNS, AlogS², AlogP², AlogP·AlogS and AlogD². COMP_HLOGS incorporates counts of 15 atomic types according to the HLOGS specification (which lists a total of 64); we did not include calculated HLOGS values themselves, as they were found correlate too weakly with these %F data. COMP_pKaSayle incorporates counts of 14 atomic types, selected from among those (50) recommended by Roger Sayle for pKa prediction. Finally, COMP_ZAF is mixed by our own recipe from the following: calculated ClogP; its square ClogP²; topological polar surface area TPSA; molar refraction MR; rotatable bond count RotBonds; squared molar weight molwt²; count of hydrogen bond acceptors (as defined by Lipinski) HBA; logarithm of the total walk count (as defined by Rucker & Rucker: J. Chem. Inf. Comp. Sci. 1993, **33**, 683) $\ln(\text{tNON_Hwc})$; our quantitative measure of violating

Lipinski's 'Rules of 5' Q_LR05.Ind; and our quantitative measure of violating Veber's criteria (J. Med. Chem. 2002, **45**, 2615) Q_Veber.Ind.

Having fixed the composition of these descriptors by fitting over the Hou database, OMLR with the latter 3 independent variables then yields 27.8% RMSE (adjusted for DF=4) fitting error over the **QSARWorld** training set. Taking the residuals from this, we searched for other significant 2-D descriptors that are based on various walk-count measures and found $\ln(\text{mHFULLwc}10/(1+\text{HBA})/(1+\text{HBD}))$; this is the logarithm of molecular path count of order 10 (taken over fully hydrogen-populated rather than depleted graphs) divided by the product $(1+\text{HBA}) \cdot (1+\text{HBD})$. Including this gives 4-variable OMLR RMSE fitting error of 25.1%. The final parameter – and the only non-2D quantity used here - taken into our submitted parsimonious 5-variable model is the dielectric energy, calculated by the implicit solvation model COSMO implemented in MOPAC2007 by James J. P. Stewart ([HTTP://OpenMOPAC.net](http://OpenMOPAC.net)). With these, the fitting RMSE (adjusted for DF=6) is 22.5%. We have also submitted a larger model, which in addition to these 5 also includes refitting the 9 components of COMP ALOGPS (whose individual coefficients are varied separately here); for this 14-variable model the fitting RMSE (adjusted for DF=15) is 21.3%.

Due to lack of time we performed no cross-validation studies. It is hoped, however, that incorporating prior info from the large independent external dataset guards against over-fitting on the training set of this problem.
