Supporting Information

ToxCast EPA in vitro to in vivo challenge: Insight into Rank-I model

- Sergii Novotarskyi,^{a,1} Ahmed Abdelaziz,^{b,c} Yurii Sushko,^{a,2} Robert Körner,^{a,3} Joachim Vogt^a and Igor V. Tetko^{d,e,*}
- ^aeADMET GmbH, Lichtenbergstraße 8, D-85748 Garching, Munich, Germany
- ^bRosettastein Consulting (UG), D-85354 Freising, Germany
- ^cWissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt, TUM-Technische Universität München, Freising, Germany
- ^dHelmholtz Zentrum München Research Center for Environmental Health (HMGU), Ingolstädter Landstraße 1 b. 60w, D-85764 Neuherberg, Germany
- ^eBigChem GmbH, Ingolstädter Landstraße 1 b. 60w, D-85764 Neuherberg, Germany ¹ All research performed at eADMET GmbH.

¹Currently employed by Facebook U.K. Ltd., London, UK ²Currently at Google Inc., Munich, Germany ³Currently at OSB AG, Regensburg, Germany

* Address for correspondence
Dr. Igor V. Tetko,
Institute of Structural Biology,
Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1,
D-85764 Neuherberg, Germany

itetko@vcclab.org Tel.: +49-89-3187-3575 Fax: +49-89-3187-3585

Table of Contents:

 1) EPA ToxCast LELPredictor Marathon Match Results Summary
 1

 (Reprinted with Permission from the U.S. Environmental Protection
 1

 Agency)
 19

2) Description of the consensus model for prediction of Lowest Effect Level

(LEL) concentrations (Technical description)	25
3) Table S1	
4) References	27



EPA ToxCast LELPredictor Marathon Match Results Summary

ToxCast MM Results Summary © TopCoder, Inc. 2014

Page 1 of 18



1.0 Introduction

This document aims at providing a thorough analysis of the top placed (winner) submissions that were received as a part of the recently completed EPA ToxCast Marathon Match challenge. The document starts by outlining the testing and scoring set ups that were used in the challenge. This exercise is to provide a complete background of the results and outcomes that are discussed further. In the next section, a concise summary of the provisional and system scores of the winner submissions along with the same scores for algorithm provided by EPA are described. The document then discusses various methods and approaches that were implemented by the winners to solve the challenge and glosses over the features selection methods and feature combinations that were used. Further, it provides a detailed statistical analysis done by our internal team on the results from the challenge and draws insights on various important aspects of the outcome. Next, it shows comparisons between various dimensions ranging from differences in modeling approach between submissions to scoring. Finally, it draws conclusions from the results discussed earlier and provides a concise view on the outcome of the challenge.

Please note:

1. The provided submission details and results are based on only top-4 submissions (based on final score) which are the prize winners in this contest.

2. All the analysis done here is based purely on the outcome of the Marathon Match. No insights presented here have been drawn from the follow-up round which is under-way currently. A separate analysis will be provided at the end of that round.

3. There have been few issues related to 1st place submission, specifically concerning the use of external data in the model which was not allowed in the match. But at this point we have not reached any conclusion and we are still in discussions with the winner to better understand the model and see if it actually uses any external data. It may be disqualified in such a scenario but at present, we are considering it as qualified submission and below analysis provides full details about that submission.

4. All the details provided in Section 4 related to the methods implemented by each submission has been extracted and summarized using the full documentation provided by the winners as a part of this match.

2.0 Testing and Scoring Mechanism

ToxCast MM Results Summary

© TopCoder, Inc. 2014

Page 2 of 18



The data in this match consisted of 1854 chemicals. Internally, this data had been split into 3 groups:

- Group A contains 483 chemicals,
- Group B contains 143 chemicals, and
- Group C contains remaining 1228 chemicals.

.This problem had 1 example, 1 provisional and 1 system test case. The score for each test case was evaluated using <u>RMSE</u> metric. However, this metric was applied only to a certain subset of data:

- For example test case this subset is training data (the data for which we had shared LEL values with members). It contains all chemicals from group A.
- For provisional test case this subset contains 63 chemicals randomly selected from group B.
- For system test case this subset contains remaining 80 chemicals from group B.

For n chemicals, if x_i is the ground truth LEL and y_i is the predicted LEL value for a chemical i, then the score will be calculated as:

Score = 1,000,000.0 * (2 - SquareRoot(($(x_0-y_0)^2 + (x_1-y_1)^2 + ... + (x_{n-1}-y_{n-1})^2) / n$))

Note: if any of y_i values is outside [-50, 50] range (and thus is obviously a very bad prediction), it was changed to -50 (if it is negative) or 50 (if it is positive) before the score was calculated.

3.0 Provisional and System Scores Summary (Top-4)

Table 3-1 provides system scores, final provisional scores and best provisional scores for top-4 submissions. It also provides system and provisional scores for algorithms provided by EPA. The number in the parenthesis signifies the rank obtained by submissions corresponding to the test case. For EPA provided submissions, it means the rank they would have achieved in current setup.

Submission	System Score	Final Prov. Score	Best Prov. Score
noveserj	880,600.27 (#1)	973,447.55 (#8)	973,447.55
NobuMiu	869,008.54 (#2)	972,777.97 (#9)	972777.97
a9108tc	865,655.93 (#3)	948,126.87 (#16)	1080134.36
klo86min	860,859.77 (#4)	906,363.11 (#27)	1054988.54

ToxCast MM Results Summary

© TopCoder, Inc. 2014



EPA_Assay	750421.99 (#14)	747443.85 (#36)	747443.85
EPA_BP	795412.15 (#11)	861629.73 (#31)	861629.73

Table 3-1. System Score, Final Provisional Score and Best Provisional Score

4.0 Algorithms, Tools and Feature Selection Methods (Top-4)

The first success of this match is that we have been able to receive different machine learning/ scientific approaches and variety of feature combinations use which was one of the major expectations. Analyzing the approaches of top-4 submissions, we observed a striking dichotomy in their roadmaps for solving this challenge. While the 1st place submission has used much more domain specific approach with the use of sophisticated models available in field of cheminformatics, the next two submissions in rank have used a more traditional machine learning (numerical-driven) approach. The 4th place submission used a mixture of both approaches. And it was an extremely interesting outcome to see that both these approaches competed very closely without much difference in their final performance in the provided set-ups. Such an outcome helps to instate the confidence that data science approach to such problems does perform as strong as any domain specific approach.

Below, we will provide a concise summary of each submission - algorithms and feature selection approaches - and refer to more details wherever required. The full version of all the materials that have been discussed below will be available to you along with code modules. Also, please ask for any references that have been used in the description and that will also be provided if not available online.

Rank-I Submission

As mentioned earlier, this submission was based more on domain-specific approach in which molecular descriptors were used as the only set of features and all these features were calculated on-the-fly from the SMILES code provided by EPA. SMILES codes and LEL values were the only data (from EPA provided data set) that this submission used for obtaining challenge predictions.

The model used by this submission was built using OCHEM. OCHEM (online chemical database and modeling environment, https://ochem.eu) is a publicly available web-based platform for QSAR research, developed and maintained by Helmholtz Zentrum Muenchen (http:// www.helmholtz-muenchen.de) and eADMET GmbH (http://eadmet.com).

Ten individual LEL models were built using different in-silico descriptor packages available in OCHEM. **No EPA in-vitro features were used in this model.** Although several models with EPA



in-vitro features were built, their inclusion into the final model did not result into a statistically significant increase in prediction accuracy. The final model was built as a simple average consensus over these ten individual models.

Figure 4-1 illustrates the full architecture of the resulting LEL model.



Figure 4-1. Final Consensus Model Diagram

Following is the brief outline of the process that this submission followed in achieving the predictions (excerpt created from full submission by the winner):

- All the molecular structures were preprocessed following a standardized OCHEM protocol in which the molecules were cleaned, neutralized, standardized and desalted.

- Some of the descriptors require a 3D representation of the molecule. These descriptors were calculated based on an optimized 3D structure representation using Corina tool by Molecular Networks GmbH.

- Ten different descriptor packages implemented in the public platform OCHEM were used individually to create ten models for the resulting consensus – here is the list of packages:

ToxCast MM Results Summary © TopCoder, Inc. 2014

Page 5 of 18



- Electrotopological state indices (E-State indices)
- Chemaxon descriptors
- Inductive descriptors
- QNPR Descriptors
- ADRIANA.Code
- Mera, Mersy
- GSfrag
- ISIDA SMF descriptors
- Dragon6 descriptors
- CDK descriptors

- Within each individual model, basic unsupervised descriptor selection procedure was performed. First, descriptors with constant values for the dataset were removed. Next, descriptors with pair-wise correlation of more than 0.95 were eliminated.

- Bootstrap aggregation (bagging) meta-learning approach with the bag size of 64 was used in each individual model. That is, each of the ten individual models in the challenge are an ensemble of 64 models, built on different training sets, which were obtained from the original one through resampling with replacement.

- For all ten of the individual models, the Associative Neural Networks (ASSN) machine learning method was used. ASNN is the author's implementation of the algorithm that uses local error correction in neural network ensemble prediction space.

- The Final EPA ToxCast challenge model was applied to the full set of 1854 molecules. In the resulting predictions file there are 37 errors (descriptors for these molecules could not be calculated for various reasons). Predictions for these molecules were taken as a mean value for the training set = 3.2602 -log(M).

- The final model is freely available from the OCHEM web site and can be accessed and used by the EPA without any limitations.

Rank-II Submission

This submission use a more numerical-driven approach and completely base its solution on Random Forest algorithm. It does not use any other libraries or helper tools and strictly uses the data provided by EPA. Following is the outline of this solution:

Page 6 of 18



- It uses Random Forest Algorithm of R package.

- Number of trees is 16000.

- Features from only following four files are used:

- ToxRefDB_Challenge_Training.csv
- TOX21S_v4a_8599_11Dec2013.csv
- toxprint_v2_vs_TOX21S_v4a_8599_03Dec2013.csv
- ToxCast_Summary_AC50_2013_12_10_NO_BSK.csv

- All the data in these files **except** the following is used as features: "CASRN", "chemical_name", "ShortName", "TS_ChemName", "TS_ChemName_Synon yms", "ChemNote", "STRUCTURE_Formula", "STRUCTURE_IUPAC", "STRUCTURE_SMILES", "STR UCTURE_SMILES_Desalt".

- All NA data is replaced to -987654321 because it is a bit tricky to handle NA on random forest of R.

- And then features from STRUCTURE_Formula and STRUCTURE_MW are added.

- The complete source code that was used to locally compute the prediction is available.

Rank-III Submission

This submission also use Random forest algorithm as the base but with different selections of features from Rank-II submission. This submission also uses ChemminR – a cheminformatics package available in R – to calculate some of the features.

Following is the outline of the submission:

- Random Forest is used for solving this problem, which generate a bunch of random decision trees according to the training data, and then merge the result from those trees into the overall prediction.

- Maximum depth of tree = 30 and minimum number of instances for each node = 5.

- Split selection for features and merging of predicts is done using statistical formulas which are available in the full version.

- Following guideline to select features was used: select the features that is mostly available for all data instances, and is rather clean with less noise.

- The following features were used in this solution.

ToxCast MM Results Summary © TopCoder, Inc. 2014

Page 7 of 18



- o MW
- 0 IntendedTarget
- o UseCategory
- O ChemType
- O ToxPrint
- o ToxCastSum
- o MW Desalted (generated by ChemmineR)
- O Chemical Property (generate by ChemineR)
- o Group Information (generated by ChemmineR)
- 0 Ring Information (generated by ChemmineR)

- All the code used in this solution is available for use.

Rank-IV Submission

*several scores are shown in this description. These scores were obtained on training set by the member locally. These scores should not be compared in any manner with the provisional or system scores obtained from final submission. All the details provided here are extracted from final documentation provided by winner – klo86min,

This submission combines the earlier approaches, specifically from feature selection point of view and provides a hybrid flavor of the solution. The solution uses Random Forest approach at the base but also uses a couple of chemical descriptors libraries to calculate on-the-fly features for chemicals. Although, this solution has achieved a 4th place score, it has a very rich experiment base as it provides insights into different feature combinations and shows empirical evaluation of various approaches.

Following is the outline of the submission:

- A Random Forest solution has been chosen for its ability to handle "large p (features) small n (samples)" cases such as this one and also different variable importance it could compute.

- Table 4-1 shows short list of final features (30 highest importance – full set is available in Table A-1 of appendix), sorted by decreasing "removal" importance.

Page 8 of 18



Label	Removal Importance	Provider	Comment
High toxicity target	15281	EPA	ACHE/ESR1/ACACA/lonchannel
BSK_Sag_CD38_down	13676	EPA	Level 8 Hitcall
Low Toxicity target	8729	EPA	ALS/PPARA/PYGM
Missing target	7125	EPA	
nNO	5191	Rdkit	
Morgan Fingerprint Acid	4918	Rdkit	Morgan Fingerprint, SMARTS: [\$([C,S](=[O,S,P])-[O;H1,-1])]
Halogen	4681	Rdkit	
Aromatic N	4585	Rdkit	
Morgan Fingerprint Aromatic	4078	Rdkit	Morgan Fingerprint, SMARTS: [a]
GRAV-6	3887	CDK	PaDel-Descriptors
nC=O	3607	Rdkit	
nNH	3432	Rdkit	
BCUTp-1h	3364	CDK	CDK Desc GUI
BCUTw-1h	3281	CDK	CDK Desc GUI
Insecticide	3193	EPA	
Mild toxicity target	3170	EPA	Mitochondria/Sterolsynthesis/Microtuble/IonchannelNa/PPO/CHRNA/cyp19a1
ATSc1	3147	CDK	CDK Desc GUI
BCUTc-1h	3086	CDK	CDK Desc GUI
ATSc2	3058	CDK	CDK Desc GUI
BCUTp-1I	3017	CDK	CDK Desc GUI
ALogP	2987	CDK	CDK Desc GUI
GRAV-4	2737	CDK	PaDel-Descriptors
ATSp5	2656	CDK	CDK Desc GUI
MOMI-YZ	2613	CDK	PaDel-Descriptors
NVS_ADME_hCYP2C19	2425	EPA	Level 8 Hitcall
ATSm5	2405	CDK	CDK Desc GUI
MOMI-XY	2398	CDK	PaDel-Descriptors
BCUTc-1I	2297	CDK	CDK Desc GUI
Morgan Fingerprint Basic	2247	Rdkit	Morgan Fingerprint, SMARTS: [#7;+,\$([N;H2&+0][\$([C,a]);
Unknown target	2238	EPA	

Table 4-1. Short list of 30 most important final features

- "Removal Importance" is the difference of 2 scores, computed with and without the feature of interest.

- CDK and RDKit descriptor features have been used in addition to features obtained from EPA data. Some descriptors have also been build based on MACCS keys. All the input and output of the process of this feature extraction is available for use. All features are also available in various csv files.

- Following method was used for extracting EPA features:

- o For numerical or binary categorical variable, original data was used.
- For missing values, it is set to zero. This is arbitrary and unfortunate but it scores better than using the mean.
- This flaw is partly compensated by the fact that in the final set of feature, there are not so many missing values and the RF method could still propagate those value and find a good split on the next depth.
- o The only non-binary categorical variable that were parsed are the column of the file "ToxCast_Generic_Chemicals_2013_12_10.csv", namely "IntentedTarget" and "UseCategory".
- o "UseCategory" was grouped into more general classes: "Herbicide", "Pesticide",

ToxCast MM Results Summary © TopCoder, Inc. 2014

Page 9 of 18



- o "Flavor/Flagrance", "Insecticide", "Pharma/Drugs".
- o "IntentedTarget" was grouped into different toxicity classes: "High", "Mild", "Low", "Unknown", and "Missing".
- Each class have been linked to targets according to what have been observed on training data (with a potential over-fit).
- o Figure 4-2 shows the breakdown of this classification with their relative importance:

Label	Removal Importance	Comment
Herbicide	-855	
Pesticide	-1211	
Flavor/Flagrance	958	
Insecticide	3193	
Pharmaceutical	1982	
High toxicity target	15281	ACHE / ESR1 / ACACA / Ionchannel
Mild toxicity target	3170	Mitochondria / Sterolsynthesis / Microtuble / IonchannelNa / PPO / CHRNA / cyp19a1
Low Toxicity target	8729	ALS / PPARA / PYGM
Unknown target	2238	Labeled target with a label that do not fall in the 3 previous categories
Missing target	7125	Unlabeled target

Table 4-2. Breakdown of Classification sorted by removal importance

- An ensemble of randomly drawn decision trees (a Random Forest variation) is built using the scikit-learn libraries (Machine Learning in Python).

- The splitting criteria of the random forest regression is based on variance reduction, which is consistent with the root-mean-square error (RMSE) used for this contest.

- The best results were obtained with maximal randomness: random subset of sample drawn with replacement (bootstrap, bagging), random subset of feature is drawn for each split, randomly capped interval for the pool of tested "best split" values of selected features. The scikit function used is "ExtraTreesRegressor".

- To illustrate the performances of the ExtraTreesRegressor, the member benchmarked few others models on the final set of features and Table 4-3 shows the results:

Scikit-learn model	Score	Comment
ExtraTreesRegressor	1,194,467	Random forest with extra randomness on feature's domain
RandomForestRegressor	1,166,786	Classical Random forest
KNeighborsRegressor	1,124,801	K-Nearest Neighbors regressor
Ridge	1,095,464	Ridge regression

Table 4-3. Performance of different models on final set of features

ToxCast MM Results Summary © TopCoder, Inc. 2014

Page 10 of 18



- For error management, the member used both out-of-bag error (oob) of the RF, and average error over 3 random 8-folds cross validation sets. The oob errors has many advantages - it is a direct output of the RF training, it is unbiased as the cross-validation sets helps manage errors: poor convergence/high variance of results could be detected, as cross-validation score and oob score shall be equivalent. 24 cross-validation testing sets were used - they provide diverse configurations and allows to see if score changes/improvement were spread over all different testing sets or concentrated on only few ones.

- There is a long discussion provided in this submission about the process of selection of features and various experiments that were performed by the member. It will be available for use whenever required. Table A-2 (Appendix) summarizes the training performance of various feature combinations that were used by the member.

- The solution is to be considered as work-in-progress and some of suggested measures for improvement include adding more descriptors, SMARTS pattern, more level8 hitcalls and recursive feature augmentation process.

5.0 Statistical Analysis

- Once the match results were available, we used several statistical methods to analyze the outcome of the results and to quantify the substance in the provided solutions and achieved scores. In this section, we describe various metrics that we calculated based on the system score for all the available solutions and also for top-4 placed submission. This analysis is expected to give you a quick idea on the outcome of the match.

- As an initial step, a quick investigation was completed by building a graph of Final-Y vs Provisional-X scores. Figure 5-1 shows that graph which essentially shows the presence of predictive power in the dataset. Each point in the graph corresponds to one competitor. The X-Axis shows the provisional score of this competitor and the Y-Axis shows his/her final score.



- Following points are to be noted from the above graph:

• The placement is non-random, so algorithms definitely see some predictive power in the data, what is already a great outcome for such a challenge.

Q

- O In competition where provisional and system datasets are similar enough, submissions that are not overfitted to provisional dataset are expected to have similar provisional and final scores. Therefore points would concentrate around the diagonal (dotted line in the figure above). However, we clearly see that majority of the dots fall below the diagonal. This suggests significant overfitting effects and/or statistical bias between provisional and scoring datasets.
- O We think that in fact both overfitting and statistical bias took place. The provisional and system datasets were obtained by randomly splitting one single dataset which suggests that there should be no statistical bias. However, this is true only for sufficiently large datasets and in our case we had two pretty small datasets (with 63 and 80 points, respectively). Additional study shows that both datasets have almost the same LEL mean (abound 2.791), but final dataset happens to have much larger LEL variance (1.314 against 1.083 in provisional).



- Large LEL variance in final dataset means that higher RMSE values (and thus lower scores) should be expected. This explains why almost all dots on the figure above fall below the diagonal. Furthermore, LEL variance gives the maximum possible score for a solution that returns a constant value (and thus sees no predictive value in datasets). RMSE of 1.314 corresponds to a score of 686K. Therefore, all dots above Y = 686K (line Q) correspond to submissions that were able to find some predictive power in the data. (Solutions below line Q could also find some predictive power, but had some problems using their findings, or fighting overfitting, etc.)
- We can see that points lie on different distance from the diagonal. This shows that despite statistical bias between provisional and final datasets, there were definitely overfitting effects in the contest as well. However, it's important to note that almost all (more exactly, all except one) top (largest Y) points lie pretty close to the diagonal. This means that among the highest scoring group, all competitors did a very good job fighting overfitting. (One outlier dot among the top ones corresponds to 8-th place competitor.)

- As a next step, we calculated various metrics for a thorough analysis to gain more insights about the solutions and to understand the significance of the results. The metrics that were used for this analysis include:

- O RMSE: Root-Mean Squared Error (lower the value, better the result)
- O RMSE (80%): RMSE on 80% best predictions (lowest prediction error)
- o Pearson: just Pearson correlation coefficient (higher the value, better the result)
- O Pearson (80%): the same on 80% best predictions
- O AUC: percentage of pairs where predicted1 < predicted2 among those where ground_truth1 < ground_truth2 (higher the value, better the result)</p>
- P-value: defines significance of the prediction, evaluated on 1 million random permutations.

Table 5-1 shows the above metrics for the top 4 solutions and also provides a comparison with the two solutions provided by EPA.

Solution	RMSE	RMSE (80% best)	Pearson	Pearson (80% best)	AUC	P-value
noveserj	1.119	0.657	0.561	0.661	0.701	0
NobuMiu	1.13	0.649	0.551	0.672	0.683	0
a9108tc	1.134	0.642	0.542	0.621	0.677	4.00E-006
klo86min	1.139	0.667	0.538	0.673	0.682	0
EPA_Assay	1.249	0.733	0.336	0.518	0.616	1.51E-003
EPA_BP	1.204	0.748	0.436	0.5	0.636	1.21E-004

Table 5-1. Various Metric Results on outcome of match

Comments based on the above results:

ToxCast MM Results Summary

© TopCoder, Inc. 2014



- The colored values in the table shows the statistically top performers in that particular metric. Those columns for which first 4 are not green signifies that some other submissions performed statistically better than the top-4 winners for that particular metric. But overall, top-4 have performed significantly better.

- As it is seen, the values of AUC and Pearson correlation coefficient are not very high overall showing that prediction in general is not an easy task.

- With a random AUC baseline of 50%, there is a significant difference between winner's 70% and EPA's solutions results of 62-64%.

- Generally, if one has 50% AUC it means that all the guesses are absolutely random. If one assumes that her subset consists of x fraction of chemicals that you predict 100% accurately, and 1-x fraction of chemicals you predict absolutely randomly, then your expected AUC value is:

 $AUC\% = 100\% * [x^2 + 1/2*(1-x^2)] = 50\% (1+x^2)$

x = SQRT ((AUC - 50%)/50%)

Based on the above formula:

Solution	AUC (%)	x
EPA_Assay	61.6	48%
EPA_BP	63.6	52%
#1 Solution	70.1	63%

- Considering the above statistic, 63% over the other two solutions is a pretty good result.

- This is one and simple way to look at AUC. There are various ways to look at AUC and an extreme case model with a double Gaussian assumption would work as follows:

- O Assume that the LELs are distributed as Gauss with mean square deviation 1.
- Assume that the LELs, predicted by algorithm, have equal precision, meaning they are distributed around true values with deviation x.
- O Derive AUC as a function of x. It is easy, as all the integrals can be taken analytically.
- Turn it around as x vs AUC, so 1/x would get you the level of granulation of the LEL distribution by algorithm (simply speaking, in how many categories the algorithm's precision allows to split the chemicals)

- The answer is simple (all integrals can be taken analytically): N=1/x = -ctg(Pi*AUC/100%). Using this, the above three cases now turns out to be:

ToxCast MM Results Summary © TopCoder, Inc. 2014 Page 14 of 18



Solution	AUC (%)	Ν
EPA_Assay	61.6	0.38
EPA_BP	63.6	0.46
#1 Solution	70.1	0.73

- N<1 means that the ability of solution to intentionally localize the value LEL of chemical is near but still within the natural deviation of LEL. But as one can see, comparatively the winner solution adds much more than just the random guess to the predictions (as N is comparable to 1).

6.0 Conclusions

The size of the data set, the natural variability within each set, and the subject domain of the problem combined to present a very ambitious challenge for a machine learning solution. The variety of techniques observed in solutions submitted to this [topcoder] contest, and the statistically significant performance of the solutions, lead us to conclude that the match successfully produced solutions that were in fact predictive. That these solutions also out-performed the algorithms supplied by the EPA, reinforced our confidence in both the applicability of data science techniques to this problem, and the immediate results of the solutions.

Following are the key highlights of what makes these results a positive outcome:

- The provided ToxCast data have been successfully used to predict statistically relevant LEL values, which verified the predictive power of the data, and reinforced confidence that data science can provide promising solutions to these problems.

- Two very different approaches to the same problem were discovered which provides a promising outlook on future directions in this undertaking. Specifically, a great comparative analysis is now available to gauge:

- o Numerical driven data science approach (Random forest) vs. highly sophisticated domain-specific approach (OCHEM tool)
- o Use of in-vitro features vs. use of molecular descriptors and also their combination.
- o EPA's current algorithmic approach to the problem vs. solution available from community.

- Statistical analyses have shown that prediction was indeed tough with the given data set and overfitting issues but the winner solutions have combated such issues and shown significantly successful results. It has also shown that the overall outcome of the contest is strong with results that can prove to be baseline for further research.



- Having said that, we would also like to add a word of caution that these results must be considered as the first step towards building a comprehensive and sophisticated machine learning approach for predicting LEL values.

- We are also doing some more statistical analysis and gauging the scientific relevance of these submissions through our follow-up round which will be presented separately. Such an analysis will help to build further confidence in this already promising outcome.

Appendix – Feature Sets used by Rank-IV submission



5 High touring traget 16.28 EPA (String Lange) ACHELESNI LACAC/Anonhamel Level B Histal (Level B Histal) 60 BKK, Sup (LDS) (String Lange) PA (Level B Histal) ALSPPRAAP/VGM 9 Maining Langet PD (Level B Histal) PA (Level B Histal) 94 Maining Langet PD (Level B Histal) PD (Level B Histal) 94 Maining Langet PD (Level B Histal) PD (Level B Histal) 94 Maining Langet PD (Level B Histal) PD (Level B Histal) 94 Halogen 488 Rdatit 96 Anomatic M Aromatic 470 Rdatit 96 Anomatic M Aromatic 470 Rdatit 96 BCUTp-th 384 CDK CDK Chee CU 91 BCUTp-th 384 CDK CDK Chee CU 91 Insteticide 3133 EPA Maintonial/Sterolsynthesis/MicrotublefonchamelNa/PPOCHRNA/cyp19a1 92 ATSc1 31 ATSc5 CDK CDK Chee CU 93 ATSc6 CDK CDK Cheee CU CDK </th <th>Index</th> <th>Label</th> <th>Removal Importance</th> <th>Provider</th> <th>Comment</th>	Index	Label	Removal Importance	Provider	Comment
BD BFA Law I B Haculi 7 Low Toucity target 872 EPA 9 Missing target 712 EPA 9 Missing target 712 EPA 9 Missing target 712 EPA 9 Missing target 400 Rdat 9 Missing target 401 Rdat 9 Missing target 401 Rdat 9 Missing target 402 Rdat 10 GRAv6 802 Rdat 11 B DUTwith 303 EFA Missing target 12 A75:01 307 EFA Missing target 516 13 Interacticide 303 EFA Missing target 516 14 Missing target 317 CFK CFK Dec GUI 14 BUTwith 303 EFA Missing target 516 14 BUTwith 302 CFK CFK Dec GUI 14 BUT	5	High toxicity target	15281	EPA	ACHE/ESR1/ACACA/Ionchannel
7 Low Toxicity targit 623 EPA ALS/PPARA/PYGM 9 Masen Targetymin Acid 619 Rdat Margen Fingetymin Acid 819 61 Margen Fingetymin Acid 419 Rdat Margen Fingetymin Acid 819 62 Margen Fingetymin Acid 419 Rdat Margen Fingetymin Acid 818 60 Margen Fingetymin Acid 318 CDK Pub-Ubecriptors 61 MAR 323 Rdat Margen Fingetymin, SMARTS: [s] 62 nAF 323 Rdat CDK Pub-Ubecriptors 63 NMH 323 EPA Margentymin, SMARTS: [s] Pub-Ubecriptors 64 Margen Fingetymin, Margentymin, SMARTS: [s] Pub-Ubecriptors Pub-Ubecriptors Pub-Ubecriptors 7 BOLTP-11 317 CDK CDK Cocc GU CDK Cocc GU 7 ATSci Target 226 CDK CDK Cocc GU 7 ADSC ZDSC CDK CDK Cocc GU ZDSC 7	63	BSK_Sag_CD38_down	13676	EPA	Level 8 Hitcall
9 Missing inget 7125 EPA 64 nNO 6191 Rdut Margan Fingerprint, Acid 4211 62 Mingan Fingerprint Acid 4211 Rdut Margan Fingerprint, SMARTS: [\$[(C,S)[+[O,S,P]+(O,H,+1])] 61 Accratic N 4235 Rdut Margan Fingerprint, SMARTS: [a] 0 Morgan Fingerprint Aconstic 4237 Rdut Margan Fingerprint, SMARTS: [a] 0 more NH 4232 Rdut Margan Fingerprint, SMARTS: [a] 19 BCUTp-1h 334 CDK CDK Cesc GU 3 Insecticite 339 EPA Machandia/Serol synthesia/Microtuble/onchamelNa/PPO/CHRNA/cyp19a1 10 ATSci 317 CDK CDK Cesc GU CDK 11 AcapP 237 CDK CDK Cesc GU CDK 11 AcapP 237 CDK CDK Cesc GU CDK 11 AcapP 237 CDK CDK Cesc GU CDK CDK CDK CDK CDK CDK CDK CD	7	Low Toxicity target	8729	EPA	ALS/PPARA/PYGM
Set nNO State State 20 Morgan Fingerprint 4881 Ridat Margan Fingerprint, SMARTS: [8](C.S)[-[0,S,P]-[0,H],1]]) 80 Morgan Fingerprint Aromatic 4881 Ridat Margan Fingerprint, SMARTS: [8](C.S)[-[0,S,P]-[0,H],1]]) 80 Morgan Fingerprint Aromatic 4881 Ridat Margan Fingerprint, SMARTS: [a] 90 BOUTp-1h 384 CDK CDK Cesc GUI 19 BOUTp-1h 384 CDK CDK Cesc GUI 11 Reschride 319 EPA Matchondria/Stanlaynthesia/Microtuble/IonchanelNa/PPO/CHRN4/cyp19a1 10 ATS2 3078 EDA CDK Cesc GUI 11 ALogP 2077 CDK CDK Cesc GUI 11 ALogP 2077 CDK CDK Cesc GUI 12 ATS4 237 CDK CDK Cesc GUI 13 ATS6 2665 CDK CDK Cesc GUI 14 MO/V12 2131 CDK CDK Cesc GUI 14 Morgan Fingerprint, SMARTS: [#/,*,\$([N+128-0][\$	9	Missing target	7125	EPA	
EC Morgan Fingerprint Acid 4918 Rdda Morgan Fingerprint, SMARTS: [4][C.S.P[-](.C.H1]] 6 Margan Fingerprint Aconatic 4038 Rdda Margan Fingerprint, SMARTS: [a] 60 Margan Fingerprint Aconatic 387 CDK Publ-Descriptors 61 Margan Fingerprint Aconatic 387 CDK Publ-Descriptors 63 Margan Fingerprint Aconatic 387 CDK Publ-Descriptors 64 GRAVE 387 CDK CDK Desc CU 75 BCUTp-1h 3364 CDK CDK Desc CU 76 Mart Instructure 3170 EPA Marchordina/Sterolsynthesis/Micratuble/onchannelNa/PPO/CHRN4/cyp19a1 77 BCUTp-1h 3170 EPA Marchordina/Sterolsynthesis/Micratuble/onchannelNa/PPO/CHRN4/cyp19a1 78 GATS 3164 CDK CDK Desc CU Marchordina/Sterolsynthesis/Micratuble/onchannelNa/PPO/CHRN4/cyp19a1 78 GATS 2372 CDK CDK Desc CU Marchordina/Sterolsynthesis/Micratuble/onchannelNa/PPO/CHRN4/cyp19a1 78 GATS 2372 CDK <td>54</td> <td>nNO</td> <td>5191</td> <td>Rdkit</td> <td></td>	54	nNO	5191	Rdkit	
Halogen 4681 Ridit Construction Absence 64 Morgan Fingeprint, SMARTS: [a] Morgan Fingeprint, SMARTS: [a] 64 Arrowalic N 4078 Ridit 65 Arrowalic N 3007 Ridit 66 Arrowalic N 3007 Ridit 61 Morgan Fingeprint, SMARTS: [a] Construction Construction 62 nC=0 3007 Ridit Construction 63 nHit 3284 CDK CDK Desc GU 75 Mild toxicity target 3170 EPA Macchandrin/Sterolymthesin/Microtuble/onchannelNa/PPO/CHRNA/cyp19a1 77 BCUTo-th 306 CDK CDK Desc GU 78 Mild toxicity target 3170 CDK CDK Desc GU 78 BCUTo-th 306 CDK CDK Desc GU 78 ArSpid 277 CDK Pabe-Descriptors 78 ArSpid 277 CDK Pabe-Descriptors 78 ArSpid 2022 CDK <t< td=""><td>62</td><td>Morgan Fingerprint Acid</td><td>4918</td><td>Rdkit</td><td>Morgan Fingerprint, SMARTS: [\$([C,S](=[O,S,P])-[O;H1,-1])]</td></t<>	62	Morgan Fingerprint Acid	4918	Rdkit	Morgan Fingerprint, SMARTS: [\$([C,S](=[O,S,P])-[O;H1,-1])]
66 Acomains 400 Karpan Fingerprint, SMARTS: [a] 60 Margan Fingerprint, SMARTS: [a] Adde-Descriptors 61 Margan Fingerprint, SMARTS: [a] Adde-Descriptors 62 n.C=0 387 CDK Pade-Descriptors 63 nH4 343 Relat CDK CDK Cost Cost 7 BCUTp-1h 384 CDK CDK Cost Cost CDK 8 Instactively traget 3170 EDK Marchand-Display-Indersin-Micratuble/Inchanne/Na/PPO/CHRN4/cyp19a1 7 BCUTp-1h 3170 EDK CDK Cost Cost CDK 7 BCUTp-1h 3170 EDK CDK Cost Cost CDK 7 BCUTp-1h 3170 CDK CDK Cost Cost CDK Descriptors 7 ATSx2 366 CDK CDK Cost Cost CDK Descriptors 8 ACopf 286 CDK CDK Descriptors Descriptors 14 ALogF 2256 CDK CDK Descriptors Descriptors	59	Halogen	4681	Rdkit	
BD Morgan Fingerprint Avomatic 407 Ridet Morgan Fingerprint, SMARTS: [a] 40 GRAV-5 307 CNC P20-Descriptors 52 nC=0 307 Ridet 53 nNH 332 Ridet 53 nSUTp-1h 3364 CDK CDK Desc GUI 54 misraticide traget 339 EPA Microtadin/Steoleynthesia/Microtable/onchannelNa/PPO/CHRN4/cynl9a1 64 MiS1 in/y traget 317 EPA Microtadin/Steoleynthesia/Microtable/onchannelNa/PPO/CHRN4/cynl9a1 71 BCUTp-1h 307 CDK CDK Desc GUI 71 BCUTp-1h 307 CDK CDK Desc GUI 72 ATSs2 308 CDK CDK Desc GUI 73 ATSs6 266 CDK CDK CDK Desc GUI 74 MOM-72 213 CDK Pade-Descriptors 74 MOM-72 213 CDK CDK Desc GUI 74 MOM-72 213 CDK CDK Desc GUI 74 <td>56</td> <td>Aromatic N</td> <td>4585</td> <td>Rdkit</td> <td></td>	56	Aromatic N	4585	Rdkit	
AD GRAVE 387 CDK PaDeDescriptors 62 nC=O 397 Ridkt 63 nNH 3122 Ridkt 79 BCUTp-1h 3384 CDK CDK Desc GUI 16 BCUTp-1h 3384 CDK CDK Desc GUI 21 ATSc1 3170 EPA Micchordina/Sterolsynthesis/Microtublefonchanne/Na/PPO/CHRN4/cyp19a1 21 ATSc1 3170 EPA Micchordina/Sterolsynthesis/Microtublefonchanne/Na/PPO/CHRN4/cyp19a1 21 ATSc1 3170 EPA Micchordina/Sterolsynthesis/Microtublefonchanne/Na/PPO/CHRN4/cyp19a1 21 ATSc1 3170 CDK CDK Desc GUI 21 ATSc5 2560 CDK CDK Desc GUI 31 ATSp6 2560 CDK CDK Desc GUI 34 MOM-YZ 2310 CDK PaDe/Descriptors 34 MOM-YY 2380 CDK CDK Desc GUI 34 MOM-YY 2380 CDK CDK Desc GUI 34	80	Morgan Fingerprint Aromatic	4078	Rdkit	Morgan Fingerprint, SMARTS: [a]
Construction State Relation 53 nNH 3322 Relation 19 BCUTp-1h 3342 Relation 19 BCUTp-1h 3341 CDK CDK Desc GUI 31 Insecticide 3139 EPA Micohandria/Sterolsynthesis/Micratuble/IonchannelNa/PPO/CHRN4/cyp19a1 20 ATSc1 3147 CDK CDK Desc GUI 21 ATSc2 3068 CDK CDK Desc GUI 31 RAVA 2277 CDK CDK Desc GUI 41 MOMAY 2869 CDK Pabe-Descriptors 42 MOSapa Fingerprint, Basic 2247 RMit Moran Fingerprint, GMARTS [M7+s.([N+D2+0][s[(C,a]) 43 Moran Fingerprint, Basic 2133 Rdkit Moran Fingerprint, GMARTS [M7+s.([N+D2+0][s[(C,a]) 44 Pham	40	GRAV-6	3887	CDK	PaDel-Descriptors
S niNe 3422 Ridai 19 BCUTp-1h 354 CDK CDK Desc GUI 15 BCUTp-1h 351 CDK CDK Desc GUI 16 Mild toxicly target 370 EPA 17 BCUTp-1h 366 CDK CDK Desc GUI 18 BCUTp-1h 366 CDK CDK Desc GUI 18 BCUTp-1h 367 CDK CDK Desc GUI 18 BCUTp-1h 367 CDK CDK Desc GUI 31 ATSp5 266 CDK CDK Desc GUI 34 MOM-YZ 287 CDK CDK Desc GUI 34 ATSp5 2455 EPA Level B Hreal 24 MOM-XY 286 CDK CDK Desc GUI 34 MOM-XY 286 CDK Pabe-Descriptors 35 BCUTp-1h 287 CDK CDK Desc GUI 36 MOM-XY 288 CDK Pabe-Descriptors 36 Mom	52	nC=O	3607	Rdkit	
9 CC/U + 1h 338.4 CD/C CCK Desc GUI 15 BC/U + 1h 3281 CD/C CCK Desc GUI 3 Inserticide 3139 EPA Michobandia/Sterolsynthesis/Microtuble/IonchannelNa/PPO/CHRN4/cyp19a1 20 ATSc1 3147 CD/C CDK Desc GUI 17 BC/U - 1h 3166 CDK CDK Desc GUI 18 BC/U - 1h 3170 CDK CDK Desc GUI 18 BC/U - 1h 3170 CDK CDK Desc GUI 19 GRAV.4 2737 CDK PDeb-Descriptors 31 ATSp5 286 CDK PDeb-Descriptors 24 ATSn5 286 CDK PDeb-Descriptors 34 MOM-XY 238 CDK PDeb-Descriptors 24 ATSn5 2237 CDK CDK Desc GUI 35 Top-PSA 2138 Rditt Moran Engerprint Basic 2247 36 Hdroons target 2238 Rditt Moran Engerprint Basic	53	nNH	3432	Rdkit	
15 DCUT-Vin 2281 CDX CCK Desc GUI 3 Inserticide 319 EPA Machandia/Stendsynthesiz/Micratuble/InchannelNa/PPO/CHRN4/cyp19a1 4 ATSc1 317 CDK CDK Desc GUI 17 BCUT-In 306 CDK CDK Desc GUI 18 BCUT-P1 3017 CDK CDK Desc GUI 18 BCUT-P1 3017 CDK CDK Desc GUI 39 GRAV-4 2737 CDK CDK Desc GUI 31 ATSp5 2866 CDK CDK Desc GUI 34 MOM-YZ 2313 CDK PADe-Descriptors 34 MOM-YZ 2313 CDK PADe-Descriptors 34 MOM-YX 2386 CDK PADe-Descriptors 35 GrapPSA 2105 CDK CDK Desc GUI 36 Moran Fingerprint Basic 2237 Rdkit Moran Fingerprint, SMARTS [M7+*,\$[(NH28-0][\$(C,a]) 36 Horans 1732 CDK CDK Desc GUI	19	BCI III n-1h	3364	CDK	CDK Desc GUI
3 Inserticide 313 EPA Microscol 6 Militoxicity target 3170 EPA Microscol 17 BCUTo-In 3170 EPA Microscol 17 BCUTo-In 306 CDK CDK Desc GUI 18 BCUTo-In 307 CDK CDK Desc GUI 11 ALog ¹ 297 CDK CDK Desc GUI 31 ATSc1 297 CDK PDebescriptors 31 ATSp5 256 CDK CDK Desc GUI 44 MOM-Y2 2737 CDK CDK Desc GUI 43 MOM-Y2 203 CDK CDK Desc GUI 44 MOM-Y2 298 CDK CDK Desc GUI 43 MOM-Y2 238 CDK CDK Desc GUI 44 Momary Fingerprint Basic 247 Rikit Margan Fingerprint, SMARTS: [#7,+\$(N/H24+0]]\$([C,a]): 61 Morgan Fingerprint Basic 247 Rikit Margan Fingerprint, SMARTS: [#7,+\$(N/H24+0]]\$([C,a]):	15	BCLITw-1b	3281	CDK	CDK Desc GUI
6 Mild toxicity target 317 CEA Micchondria/Surphysics/MicrotubleAnonhanne/Na/PPOYCHRNA/cyp19a1 20 ATSc1 3147 CDK CDK Dees GUI 21 ATSc2 3988 CDK CDK Dees GUI 21 ATSc2 3988 CDK CDK Dees GUI 21 ATSc2 3988 CDK CDK Dees GUI 21 ATSc5 3988 CDK CDK Dees GUI 21 ATSc5 2977 CDK CDK Dees GUI 31 ATSp5 2977 CDK CDK Dees GUI 32 ATSp5 2966 CDK CDK Dees GUI 34 MOMAYZ 2936 CDK PADe/Descriptors 34 MOMAY 2986 CDK PADe/Descriptors 35 TopoPSA 2977 CDK CDK Dees GUI 34 Morgan Fingerprint Basic 2247 RMich Morgan Fingerprint, SMARTS: [#7,+\$(NHQ&+0][\$(C,a]) 36 TopoPSA 2195 CDK CDK Dees GUI DE	3	Insecticide	3193	FPA	
On ATS:1 OK	6	Mild toxicity target	3170	EPA	Mitochondria/Sterolsynthesis/Microtuble/IonchannelNa/PPO/CHRNA/cyn19a1
17 BCUTI-Ih 308 CDK CDK DEX 12 ATSx2 398 CDK CDK DEX DEX 18 BCUTp-11 3017 CDK CDK DEX DEX DEX 11 AL.op ² 2377 CDK CDK DEX DEX <t< td=""><td>วัก</td><td>ATSc1</td><td>3147</td><td>CDK</td><td>CDK Dest GUI</td></t<>	วัก	ATSc1	3147	CDK	CDK Dest GUI
21 ATSv2 398 CDX CDX Desc GUI 11 ALog ¹ 2987 CDX CDX Desc GUI 13 ALog ¹ 2987 CDX CDX Desc GUI 13 ATSv6 2987 CDX CDX Desc GUI 13 ATSv6 2987 CDX Pubel Descriptore 14 MOM-X2 293 CDX Pubel Descriptore 15 MOM-X1 293 CDX Pubel Descriptore 16 Morgan Fingerprint 2937 CDX CDX Desc GUI 16 Morgan Fingerprint 2937 CDX CDX Desc GUI 16 Morgan Fingerprint 2937 CDX CDX Desc GUI 17 Haceptors 2138 Rdit 104 18 Unknown target 2236 EPA Level Besc GUI 19 ATSp4 2002 CDX Desc GUI 10 ATSp4 2002 CDX Desc GUI 11 Atoms 1732 Rdit 1173 28 ATSm4 1501 CDX DX Desc GUI </td <td>17</td> <td>BCLITe-1h</td> <td>3086</td> <td>CDK</td> <td>CDK Desc GUI</td>	17	BCLITe-1h	3086	CDK	CDK Desc GUI
BCUT_P11 DOT CDK CDK Dase GUI 11 ALogP 297 CDK CDK Dase GUI 33 GRAV.4 2737 CDK Pabe/Descriptors 31 ATSp6 2856 CDK Pabe/Descriptors 44 MOM-YZ 2813 CDK Pabe/Descriptors 29 ATSr6 2405 CDK CDK Desc GUI 31 MOM-YZ 2813 CDK Pabe/Descriptors 32 ATSr6 2405 CDK CDK Desc GUI 34 MOM-YY 298 CDK CDK Desc GUI 43 MOM-YY 298 CDK CDK Desc GUI 44 Pharmoreutical 293 CDK CDK Desc GUI 57 H acceptors 2133 Ridt Mogan Fingerprint, SMARTS: [#7,+\$(NH28+0)[\$(C,a)] 40 Pharmaceutical 1982 CDK CDK Desc GUI 56 H donors 1732 Ridt Mogan Fingerprint, SMARTS: [#7,+\$(NH28+0)[\$(C,a)] 51 nOH	21	ATSc2	3058	CDK	
11 ALog ¹ 2897 CDK CDK Desc GUI 33 GRAV.4 2737 CDK Pable/Descriptors 31 ATSp6 2866 CDK DCK Desc GUI 44 MOM+YZ 2813 CDK Pable/Descriptors 64 MVS, ADME_DCYP2C19 2425 EVA Leviel Bittcall 29 ATSm6 2050 CDK Pable/Descriptors 16 BCUT-11 297 CDK CDK Desc GUI 31 ATSm6 2045 CDK Pable/Descriptors 16 BCUT-11 297 CDK CDK Desc GUI 33 TopoPSA 2195 CDK CDK Desc GUI 4 Phamaceutical 1982 EPA Leviel 30 ATSp4 2002 CDK CDK Desc GUI 4 Phamaceutical 1982 EPA Leviel 31 AMGP 1027 CDK Pable/Descriptors 31 AMF 1030 CDK <td< td=""><td>18</td><td>BCIII::</td><td>3017</td><td>CDK</td><td>CDK Desc GUI</td></td<>	18	BCIII::	3017	CDK	CDK Desc GUI
30 GFA 203 CDK Pable/Descriptors 31 ATSp5 2256 CDK CDK Desc GUI 44 MOM-YZ 2615 CDK CDK Desc GUI 44 MOM-YZ 2615 CDK CDK Desc GUI 24 ATSn5 2405 CDK PABe/Descriptors 25 ATSn5 2405 CDK CDK Desc GUI 43 MOM-YY 2386 CDK PABe/Descriptors 16 BCUTc+11 2297 CDK CDK Desc GUI 31 TopPSA 2195 CDK CDK Desc GUI 32 TopPSA 2195 CDK CDK Desc GUI 33 TopPSA 2195 CDK CDK Desc GUI 4 Pharmaceutical 1982 EPA 33 TopPSA 2195 CDK CDK Desc GUI 4 Patamaceutical 1982 CPA PaBel-Descriptors 51 nOH 1650 Rdkit PaBel-Descriptors	10	ALogP	2987	CDK	CDK Desc GUI
39 ONA-A 2.0 CDK Packeroscilluiss 31 ATSp5 2865 CDK CDK Desc GUI 44 MOM4-YZ 2813 CDK PADe-Descriptors 64 MVS_ADME_hCYP2C19 2425 CDK CDK Desc GUI 33 ATSm5 2405 CDK CDK Desc GUI 44 MOM4/Y 2396 CDK CDK Desc GUI 43 Moman Fingerprint Basic 2247 CDK CDK Desc GUI 44 Moran Fingerprint Basic 2238 EPA CDK Desc GUI 45 Hacceptors 2133 Rikit CDK Desc GUI 44 Pharmaceutical 1982 EPA CDK Desc GUI 44 Pharmaceutical 1982 EPA CDK Desc GUI 45 H donors 1732 Rikit PaDel-Descriptors 46 H donors 1732 Rikit PaDel-Descriptors 47 GRAVH-1 1717 CDK CDK Desc GUI 48 H donors	20		2307	CDK	PoPol Descriptoro
J. A. (30) 200 CDK Deck USB (0) 44 MOM-YZ 2613 CDK Pabe-Descriptors 64 NVS_ADME_hCYP2C19 2425 EPA Level 8 Hrcall 29 ATSn6 2405 CDK Pabe-Descriptors 16 BCUTc-11 2395 CDK CDK Desc GUI 61 Morgan Fingerprint Basic 2247 Rukit Morgan Fingerprint, SMARTS: [#7:+,\$[[N]+D2&+0][\$[(C,a]) 7 H acceptors 2133 Rukit Morgan Fingerprint, SMARTS: [#7:+,\$[[N]+D2&+0][\$[(C,a]) 8 Unknown target 2135 CDK CDK Desc GUI 9 H acceptors 1732 Rukit EPA 30 ATSp4 2002 CDK CDK Desc GUI 41 Hadnors 1732 Rukit EPA 37 GRA/H-11 1717 CDK Descriptors 51 nOH 1660 Rukit 24 MLogP 1274 CDK CDK Desc GUI 31 AMR	20	ATChE	2/3/	CDK	CDK Deep CUI
Holm 12 2013 CD/k Padde Descriptions 64 NVS_ADME_hCVP2C19 2425 CPA Level B Hitcail 29 ATSm6 2405 CDK CDK Desc GUI 43 MOMLXY 238 CDK Padde Descriptors 16 BCUTc11 2297 CDK CDK Desc GUI 61 Morgan Fingerprint Basic 2247 Rdkit 70 TapoPSA 2195 CDK CDK Desc GUI 71 Hacceptos 2133 Rdkit 73 TapoPSA 2195 CDK CDK Desc GUI 74 Hacseptos 1732 Rdkit 75 nOH 1660 Rdkit 76 RAVH-1 1717 CDK Pade-Descriptors 71 MCM-1 1650 Rdkit 73 RTSM4 1651 CDK Desc GUI 74 MoMeratina 1037 CDK CDK Desc GUI 74 MOM-Z 1173 CDK CDK Desc GUI <	44		2000	CDK	ReDel Descriptore
bit INVS_ADME_INCTP2CIS 24.25 EPA Level of Initial 29 ATSmE 2405 CDK Pabel-Discriptors 16 BCUTc-11 2297 CDK CDK Desc GUI 16 Morgan Fingerprint Basic 2247 Rdkit Morgan Fingerprint, SMARTS: [#7;+\$([NH28+0][\$(C,a)); 8 Unknown target 2238 EPA CDK Desc GUI 30 ATSp4 2012 CDK CDK Desc GUI 31 TopoPSA 2195 CDK CDK Desc GUI 32 ATSp4 2012 CDK CDK Desc GUI 33 TopoPSA 2193 Rdkit CDK 34 Phamaceutical 1982 EPA 37 GRAVH-1 177 CDK Pabel-Descriptors 51 nOH 1660 Rdkit 23 ATSm6 1051 CDK CDK Desc GUI 34 BCUTw-11 149 CDK Desc GUI CDK Desc GUI 35 ML0gP 1274 CDK <td></td> <td>NUC ADME LOVDOCID</td> <td>2013</td> <td>CDK</td> <td>Faberbeschptors</td>		NUC ADME LOVDOCID	2013	CDK	Faberbeschptors
25 ATSrb 2405 CDK CDK Descriptors 43 MOM-XY 238 CDK Pabe-Descriptors 16 BCUTc-11 2297 CDK CDK Desc GUI 61 Morgan Fingerprint Basic 2247 Rdkit 30 TopoPSA 2195 CDK CDK Desc GUI 57 H acceptors 2133 Rdkit 30 ATSp4 2002 CDK CDK Desc GUI 4 Pharmaceutical 1982 EPA 30 H donors 1732 Rdkit 37 GRAVH-1 1717 CDK Pabe-Descriptors 38 H donors 1732 Rdkit 39 ATSm4 1551 CDK CDK Desc GUI 14 BCUTv-11 1459 CDK CDK Desc GUI 14 BCUTv-11 1459 CDK CDK Desc GUI 14 BCUTv-11 1459 CDK CDK Desc GUI 15 nOH 1060 Rdkit 2004 26 MLogP 1207 EPA 27 ATSm4 1551 CDK CDK Desc GUI 38 KLogP 1081 CDK CDK Desc GUI <	04 30	ATCHE	2425	CDK	CDK Deep CUI
43 MOM:A T 2390 CUK Pade/Descriptors 16 ECUTo-11 2390 CUK CUK Cec GUI 61 Morgan Fingerprint Basic 2247 Rdkt Morgan Fingerprint, SMARTS: [#7:+\${[N}+2&+0][\${(C,a);} 8 Unknown target 238 EPA CUK Desc GUI 30 ATSp4 2002 CDK CDK Desc GUI 4 Pharmaceutical 1962 EPA 51 nOH 1660 Rdkt 32 ATSm4 1561 CDK CDK Desc GUI 34 Pharmaceutical 1962 EPA CDK Desc GUI 35 ATSm4 1561 CDK CDK Desc GUI 36 ATSm4 1561 CDK CDK Desc GUI 37 GRAVH-1 1473 CDK PaDel-Descriptors 38 ATSm4 1501 CDK PaDel-Descriptors 39 XLogP 1081 CDK CDK Desc GUI 40 MOMar weight 1007 CDK	29	ATSm5	2405	CDK	CDK Desc GUI
b BCUTC-11 2297 CUK CUK Desc GUI 61 Morgan Fingerprint, SMARTS: [#7,+,\$[[N]+2&+0][\$([C,a]), 8 Unknown target 238 EPA 30 TopaPSA 2195 CDK CDK Desc GUI 57 H acceptors 2133 Rdkit 30 ATSp4 2002 CDK CDK Desc GUI 4 Pharmaceutical 1962 EPA 60 H donors 1732 Rdkit 37 GRAVH-1 1717 CDK PaDel-Descriptors 51 nOH 1660 Rdkit 28 ATSm4 1551 CDK CDK Desc GUI 14 BCUTw-11 1469 CDK CDK Desc GUI 10 Molar weight 1007 EPA 42 MOM-Z 1173 CDK PaDel-Descriptors 13 AMR 1003 CDK CDK Desc GUI 24 MOM-Z 1019 DMK CDK 35	43	MONEX 1	2398	CDK	PaDe-Descriptors
bit Morgan Fingerprint Basic 2247 Rdkit Morgan Fingerprint, SWARTS: [#7,+3([]\[R]/R-24-U]]\$([C,4]) 38 TopoPSA 2195 CDK CDK Desc GUI 30 ATSp4 2002 CDK CDK Desc GUI 40 Pharmaceutical 1962 EPA CDK Desc GUI 41 Pharmaceutical 1962 EPA CDK Desc GUI 42 Pharmaceutical 1962 EPA CDK Desc GUI 33 GRAVH-1 1717 CDK CDK Desc GUI 34 BCUTw-11 1469 CDK CDK Desc GUI 35 MLogP 1274 CDK CDK Desc GUI 36 MLogP 108 CDK CDK Desc GUI 37 GAMR 1103 CDK CDK Desc GUI 38 XLogP 1081 CDK CDK Desc GUI 39 XLogP 1081 CDK CDK Desc GUI 40 MOM-R 1093 CDK CDK Desc GUI 51 ATSc5 <td>16</td> <td>BCUICII</td> <td>2297</td> <td>CDK</td> <td>CUK Desc GUI</td>	16	BCUICII	2297	CDK	CUK Desc GUI
8 Unknown target 223 EPA 33 TopdPSA 2193 Rdkit 57 H acceptors 2133 Rdkit 30 ATSp4 2002 CDK CDK Desc GUI 4 Phamaceutical 1962 EPA EPA 58 H donors 1732 Rdkit EVA 37 GRAVH-1 1717 CDK PaDel-Descriptors 61 nOH 1660 Rdkit 28 ATSm4 1561 CDK CDK Desc GUI 14 BCUTw-11 1469 CDK CDK Desc GUI 32 MLogP 1207 EPA EPA 42 MOM-Z 1173 CDK CDK Desc GUI 10 Molar weight 1207 EPA EPA 42 MOM-Z 1173 CDK CDK Desc GUI 36 XLogP 1081 CDK CDK Desc GUI 24 ATSc5 834 CDK CDK Desc GUI	61	Morgan Fingerprint Basic	2247	Rakit	Morgan Fingerprint, SMARTS: [#/;+,\$([N;H2&+U][\$([U;a]);
33 lopor SA 2195 CUK CUK Desc GUI 57 H saceptors 213 Rdkt 30 ATSp4 2002 CDK CDK Desc GUI 4 Phamaceutical 1962 EPA 58 H donors 1732 Rdkt 37 GRAVH-1 1717 CDK Pabel-Descriptors 51 nOH 1650 Rdkt 28 ATSm4 1551 CDK CDK Desc GUI 32 MLogP 1274 CDK CDK Desc GUI 34 MOM-Z 1173 CDK Pabel-Descriptors 35 MLogP 1274 CDK DK Desc GUI 42 MOM-R 1103 CDK CDK Desc GUI 36 XLogP 1081 CDK CDK Desc GUI 45 MOM-R 1019 CDK CDK Desc GUI 46 MS-5 820 CDK CDK Desc GUI 47 ATSc5 820 CDK CDK D	8	Unknown target	2238	EPA	
by Flaceptors 2133 Rdkit 30 ATSp4 202 CDK CDK Desc GUI 4 Pharmaceutical 1982 EPA 58 H donors 1732 Rdkit 37 GRAVH-1 1717 CDK PaDel-Descriptors 51 nOH 1660 Rdkit 28 ATSm4 1561 CDK CDK Desc GUI 14 BCUTw-11 1469 CDK CDK Desc GUI 10 Molar weight 1207 EPA 42 MOMI-Z 1173 CDK CDK Desc GUI 36 XLogP 1081 CDK CDK Desc GUI 45 MOMI-R 1091 CDK PaDel-Descriptors 24 ATSc5 834 CDK CDK Desc GUI 25 ATSc4 801 Rdkit 143 46 MOMI-R 109 CDK CDK Desc GUI 26 ATSc5 820 CDK CDK Desc GUI	33	TopoPSA	2195	CDK	CDK Desc GUI
3J A1 Sp4 2002 CDK CDK Desc GUI 44 Pharmaceutical 192 EPA 58 H donors 1732 Rdkit 37 GRAVH-1 1717 CDK PaDel-Descriptors 51 nOH 1660 Rdkit 28 ATSm4 1551 CDK CDK Desc GUI 30 MLogP 1274 CDK CDK Desc GUI 31 Molar weight 1207 EPA EPA 42 MOMI-Z 1173 CDK PaDel-Descriptors 33 AMR 1103 CDK CDK Desc GUI 44 MOMI-Z 1173 CDK PaDel-Descriptors 33 AMR 1103 CDK CDK Desc GUI 44 MOMI-Z 1103 CDK PaDel-Descriptors 2 Flavor/Flagrance 958 EPA EPA 44 ATSc5 834 CDK CDK Desc GUI 20 Estatemin 801 Rdkit EPA 34 VABC 756 <td< td=""><td>5/</td><td>H acceptors</td><td>2133</td><td>Rakit</td><td>2817 B 011</td></td<>	5/	H acceptors	2133	Rakit	2817 B 011
4 Phamaceutical 1962 EPA 58 H donors 1732 R&kit 37 GRAVH-1 1717 CDK PaDel-Descriptors 51 nOH 1660 Rdkit 28 ATSm4 1551 CDK CDK Desc GUI 32 MLogP 124 CDK CDK Desc GUI 34 BCUTw-11 1499 CDK CDK Desc GUI 35 MLogP 127 EPA 42 MOMI-Z 1173 CDK CDK Desc GUI 36 XLogP 1081 CDK CDK Desc GUI 36 XLogP 1081 CDK CDK Desc GUI 45 MOMI-R 1019 CDK CDK Desc GUI 46 MOMI-R 1019 CDK CDK Desc GUI 50 Estatemin 801 Rdkit 51 VABC 789 CDK CDK Desc GUI 52 ATSm1 764 CDK Desc GUI CDK Desc GUI	30	AISp4	2002	CDK	CDK Desc GUI
BB H data 37 GRAVH-1 1717 CDK PaDel-Descriptors 51 nOH 1660 Rdkit 28 ATSm4 1551 CDK CDK Desc GUI 32 MLogP 1274 CDK CDK Desc GUI 32 MLogP 1274 CDK CDK Desc GUI 42 MOM-Z 1173 CDK PaDel-Descriptors 33 AMR 1103 CDK CDK Desc GUI 36 XLogP 1081 CDK CDK Desc GUI 45 MOMI-R 1019 CDK PaDel-Descriptors 2 Flavor/Flagrance 958 EPA 24 ATSc3 834 CDK CDK Desc GUI 22 ATSc3 820 CDK CDK Desc GUI 24 ATSc3 820 CDK CDK Desc GUI 25 ATSm1 764 CDK CDK Desc GUI 26 VABC 747 CDK CDK Desc GUI <td>4</td> <td>Pharmaceutical</td> <td>1982</td> <td>EPA</td> <td></td>	4	Pharmaceutical	1982	EPA	
3/ GRAVH-1 17.1 CUK Paule-Descriptors 51 nOH 1660 Rdkit 28 ATSm4 1551 CDK CDK Desc GUI 14 BCUTw-11 1469 CDK CDK Desc GUI 32 MLogP 1274 CDK CDK Desc GUI 10 Molar weight 1207 EPA 42 MOM-Z 1173 CDK CDK Desc GUI 36 X.LogP 1081 CDK CDK Desc GUI 36 X.LogP 1081 CDK CDK Desc GUI 2 Flavor/Flagrance 958 EPA 24 ATSc5 834 CDK CDK Desc GUI 23 ATSc4 820 CDK CDK Desc GUI 35 VABC 789 CDK CDK Desc GUI 36 X.Logp2 747 CDK CDK Desc GUI 37 ATSm1 764 CDK CDK Desc GUI 36 VABC 789 C	58	H donors	1732	Rdkit	
b1 NDH Ibbol Rdwit 28 ATSm4 1551 CDK CDK Desc GUI 14 BCUTw-11 1469 CDK CDK Desc GUI 32 MLogP 1274 CDK CDK Desc GUI 10 Molar weight 1207 EPA 42 MOM-Z 1173 CDK Pabe-Descriptors 13 AMR 103 CDK CDK Desc GUI 36 X LogP 1081 CDK Pabe-Descriptors 2 Flav or/Flagrance 958 EPA EPA 24 ATS-G 834 CDK CDK Desc GUI 36 X LogP 831 Rdwit EDA 21 ATS-G 834 CDK CDK Desc GUI 35 V ABC 789 CDK CDK Desc GUI 36 V ABC 789 CDK CDK Desc GUI 37 ATSm3 736 CDK Debe-Descriptors 48 geomShape 1	3/	GRAVH-1	17.17	CDK	Papel-Descriptors
AB Alsma Test CUK CUK Desc GU 14 BCUTw-11 1469 CDK CDK Desc GU 32 MLogP 1274 CDK CDK Desc GU 10 Molar weight 1207 EPA 42 MOM-Z 1173 CDK CDK Desc GU 33 AMR 103 CDK CDK Desc GU 36 XLogP 1081 CDK CDK Desc GU 36 MOM-R 1019 CDK PaDel-Descriptors 2 Flav or/Flagrance 958 EPA 24 ATSc5 834 CDK CDK Desc GU 25 ATSc3 820 CDK CDK Desc GU 26 Estatemin 801 Rdkit 35 VABC 789 CDK CDK Desc GU 25 ATSm1 736 CDK CDK Desc GU 26 VABC 737 CDK PaDel-Descriptors 38 GRAVH-3 -1 CDK	51	nUH	1660	Rakit	201/ D 01/
14 BCUIW-11 1469 CDK CDK Desc GUI 32 MLogP 1274 CDK CDK Desc GUI 10 Molar weight 1207 EPA 42 MOMI-Z 1173 CDK Pabel-Descriptors 13 AMR 1103 CDK CDK Desc GUI 36 XLogP 1081 CDK CDK Desc GUI 45 MOMI-R 1019 CDK Pabel-Descriptors 2 Flav or/Flagrance 958 EPA 24 ATSc5 834 CDK CDK Desc GUI 25 ATSc3 820 CDK CDK Desc GUI 35 VABC 789 CDK CDK Desc GUI 26 ATSm1 764 CDK CDK Desc GUI 27 ATSm3 736 CDK CDK Desc GUI 27 ATSm3 736 CDK Pabel-Descriptors 41 MOM-Y 13 CDK Pabel-Descriptors 28 GRAVH-3	28	AISm4	1551	CDK	CDK Desc GUI
32 MLogP 12/4 CUK CUK bes GU 10 Molar weight 1207 EPA 42 MOMI-Z 1173 CDK PaDel-Descriptors 13 AMR 1103 CDK CDK Desc GU 36 XLogP 1081 CDK CDK Desc GU 45 MOMI-R 1019 CDK PaDel-Descriptors 2 Flav or/Flagrance 958 EPA 24 ATSc5 834 CDK CDK Desc GU 20 Estatemin 801 Rdkit	14	BCUIW-11	1469	CDK	CDK Desc GUI
IU Molar weight 12/2 EPA 42 MOM-Z 1173 CDK PaDel-Descriptors 13 AMR 1103 CDK CDK Desc GUI 36 XLogP 1081 CDK CDK Desc GUI 45 MOM-R 1019 CDK PaDel-Descriptors 2 Flav or/Flagrance 958 EPA 24 ATSc5 834 CDK CDK Desc GUI 25 ATSc3 820 CDK CDK Desc GUI 26 Estatemin 801 Rdkit 36 VABC 789 CDK CDK Desc GUI 25 ATSm1 766 CDK CDK Desc GUI 26 ALsgp2 747 CDK CDK Desc GUI 27 ATSm3 736 CDK PaDel-Descriptors 41 MOM-Y 13 CDK PaDel-Descriptors 38 GRAVH-3 -1 CDK PaDel-Descriptors 26 ATSm2 -180 <td>32</td> <td>MLogP</td> <td>1274</td> <td>CDK</td> <td>CDK Desc GUI</td>	32	MLogP	1274	CDK	CDK Desc GUI
42 MUM-Z 11/3 CUK Pabel-Descriptors 13 AMR 1103 CDK CDK Desc GU 36 XLogP 1081 CDK Pabel-Descriptors 2 Flav or/Flagrance 958 EPA 24 ATSc5 834 CDK CDK Desc GU 25 ATSc3 820 CDK CDK Desc GU 35 VABC 789 CDK CDK Desc GU 25 ATSm1 764 CDK CDK Desc GU 26 ATSm3 736 CDK CDK Desc GU 27 ATSm3 736 CDK CDK Desc GU 27 ATSm3 736 CDK CDK Desc GU 27 ATSm3 736 CDK CDK Desc GU 28 genShape 173 CDK Pabel-Descriptors 38 GRAVH-3 -1 CDK Pabel-Descriptors 26 ATSm2 -180 CDK CDK Desc GU 34 VAdjMat -217 CDK CDK Desc GU 34 VAdjMat -217 CDK CDK Desc GU 34 VAdjMat -217 CDK DK Desc GU 46 gennDiameter -9	10	Molar weight	1207	EPA	
13 AMR 1103 CDK CDK Desc GUI 36 XLogP 1081 CDK PaDel-Descriptors 2 Flav or/Flagrance 958 EPA 24 ATSc5 834 CDK CDK Desc GUI 22 ATSc3 820 CDK CDK Desc GUI 23 ATSc5 834 CDK CDK Desc GUI 24 ATSc5 820 CDK CDK Desc GUI 25 ATSm1 801 Rdkit 36 VABC 788 CDK CDK Desc GUI 27 ALogp2 747 CDK CDK Desc GUI 28 ATSm3 736 CDK CDK Desc GUI 27 ATSm3 736 CDK PaDel-Descriptors 48 geomShape 173 CDK PaDel-Descriptors 48 geomShape -1 CDK PaDel-Descriptors 38 GRAVH-3 -1 CDK DK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 34 VAdjMat -21	42	MOMI-Z	1173	CDK	PaDel-Descriptors
36 XLogP 1081 CDK CDK Desc GUI 45 MOM-R 1019 CDK PaDeI-Descriptors 2 Flav or/Flagrance 958 EPA 24 ATSc5 834 CDK CDK Desc GUI 20 Estatemin 801 Rdkit 36 VABC 789 CDK CDK Desc GUI 25 ATSm1 764 CDK CDK Desc GUI 27 ALsgp2 747 CDK CDK Desc GUI 28 geomShape 173 CDK CDK Desc GUI 27 ATSm3 736 CDK CDK Desc GUI 27 ATSm3 736 CDK CDK Desc GUI 48 geomShape 173 CDK PaDeI-Descriptors 41 MOM-Y 13 CDK PaDeI-Descriptors 38 GRAVH-3 -1 CDK DK Desc GUI 34 VAdjMat -217 CDK DK Desc GUI 49 Estatemax -280 Rdkit E 56 Phenol -487 <td>13</td> <td>AMR</td> <td>1103</td> <td>CDK</td> <td>CDK Desc GUI</td>	13	AMR	1103	CDK	CDK Desc GUI
45 MOM-R 1019 CDK Pabe/Descriptors 2 Flav or/Flagrance 958 EPA 24 ATSc5 834 CDK CDK Desc GUI 22 ATSc3 820 CDK CDK Desc GUI 50 Estatemin 801 Rdkit 35 VABC 789 CDK CDK Desc GUI 25 ATSm1 764 CDK CDK Desc GUI 27 ATSm3 736 CDK CDK Desc GUI 28 geomShape 173 CDK Pabel-Descriptors 41 MOM-Y 13 CDK Pabel-Descriptors 38 GRAVH-3 -1 CDK DK Desc GUI 26 ATSm2 -180 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 34 VAdjMat -217 CDK DK Desc GUI 49 Estatemax -280 Rdkit 56 Phenol -487 Rdkit 57 Phenol -487 Rdkit 58 Pabel-Descriptors -280 Rdkit 54 Phenol -487 Rdkit 56 Phenol -487	36	XLogP	1081	CDK	CDK Desc GUI
2 Flav or/F lagrance 958 EPA 24 ATSc5 834 CDK CDK Desc GUI 22 ATSc3 820 CDK CDK Desc GUI 50 Estatemin 801 Rdkit 36 VABC 789 CDK CDK Desc GUI 25 ATSm1 764 CDK CDK Desc GUI 27 ATSm3 736 CDK CDK Desc GUI 28 geomShape 173 CDK PaDel-Descriptors 48 geomShape 173 CDK PaDel-Descriptors 41 MOM-Y 13 CDK PaDel-Descriptors 38 GRAVH-3 -1 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 35 Phenol -487 Rdkit 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDe	45	MOMI-R	1019	CDK	PaDel-Descriptors
24 A1Sc5 834 CDK CDK Desc GU 22 ATSc3 820 CDK CDK Desc GU 50 Estatemin 801 Rdkit 35 VABC 789 CDK CDK Desc GU 25 ATSm1 764 CDK DEsc GU 12 ALogp2 747 CDK CDK Desc GU 27 ATSm3 736 CDK CDK Desc GU 48 geomShape 173 CDK PaDel-Descriptors 41 MOM-Y 13 CDK PaDel-Descriptors 28 ATSm2 -180 CDK CDK Desc GU 29 Estatemax -280 Rdkit 20 ATSm2 -180 CDK CDK Desc GU 34 VAdjMat -217 CDK CDK Desc GU 49 Estatemax -280 Rdkit 55 Phenol -487 Rdkit 0 Herbicide -855 EPA 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GU 1 Pesticide -1211 EPA 46 geomRadius -13	2	Flavor/Flagrance	968	EPA	
22 A1Sc3 82U CDK CDK Desc GU 50 Estatemin 801 Rdkit 35 VABC 789 CDK CDK Desc GU 25 ATSm1 764 CDK CDK Desc GU 12 ALogp2 747 CDK CDK Desc GU 27 ATSm3 736 CDK CDK Desc GU 48 geomShape 173 CDK Pabel-Descriptors 41 MOM-Y 13 CDK Pabel-Descriptors 38 GRAVH-3 -1 CDK CDK Desc GU 26 ATSm2 -180 CDK CDK Desc GU 34 VAdjMat -217 CDK CDK Desc GU 49 Estatemax -280 Rdkit 56 Phenol -487 Rdkit 56 Phenol -487 Rdkit 61 geomDiameter -922 CDK Pabel-Descriptors 23 ATSc4 -1178 CDK Desc GU 1 Pesticide -127 CDK Desc GU 1 Pesticide -1327 CDK	24	ATSc5	834	CDK	CDK Desc GUI
50 Estatemin 801 Rdkt 36 VABC 789 CDK CDK Desc GUI 25 ATSm1 764 CDK CDK Desc GUI 12 ALopp2 747 CDK CDK Desc GUI 27 ATSm3 736 CDK CDK Desc GUI 48 geomShape 173 CDK PaDel-Descriptors 41 MOMI-Y 13 CDK PaDel-Descriptors 38 GRAVH-3 -1 CDK PaDel-Descriptors 26 ATSm2 -180 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 49 Estatemax -280 Rdkit 56 Phenol -487 Rdkit 0 Herbicide -855 EPA 47 geomDiameter -922 CDK	22	ATS03	820	CDK	CDK Desc GUI
36 VABC /89 CDK CDK Desc GUI 25 ATSm1 764 CDK CDK Desc GUI 12 ALogp2 747 CDK CDK Desc GUI 27 ATSm3 736 CDK CDK Desc GUI 48 geomShape 173 CDK PaDel-Descriptors 41 MOM-Y 13 CDK PaDel-Descriptors 38 GRAVH-3 -1 CDK PaDel-Descriptors 26 ATSm2 -180 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 49 Estatemax -280 Rdkit 55 Phenol -487 Rdkit 0 Herbicide -855 EPA 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1327 CDK PaDel-Descriptors	50	Estatemin	801	Rdkit	
25 A ISm1 /64 CDK CDK CDK Desc GU 12 ALogp2 747 CDK CDK Desc GU 27 ATSm3 736 CDK CDK Desc GU 48 geomShape 173 CDK PaDeI-Descriptors 41 MOM-Y 13 CDK PaDeI-Descriptors 38 GRAVH-3 -1 CDK PaDeI-Descriptors 26 ATSm2 -180 CDK CDK Desc GUI 49 Estatemax -280 Rdkit 56 Phenol -487 Rdkit 56 Phenol -487 Rdkit 47 geomDiameter -922 CDK PaDeI-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -121 EPA 46 geomRadius -1327 CDK PaDeI-Descriptors	36	VABC	789	CDK	CDK Desc GUI
12 ALogp2 747 CDK CDK Desc GUI 27 ATSm3 736 CDK CDK Desc GUI 48 geomShape 173 CDK PaDel-Descriptors 41 MOM-Y 13 CDK PaDel-Descriptors 38 GRAVH-3 -1 CDK PaDel-Descriptors 26 ATSm2 -180 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 49 Estatemax -280 Rdkit 55 Phenol -487 Rdkit 0 Herbicide -855 EPA 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDel-Descriptors	25	AI Sm1	/64	CDK	CDK Desc GUI
27 ATSm3 736 CDK CDK Desc GUI 48 geomShape 173 CDK PaDeI-Descriptors 41 MOM-Y 13 CDK PaDeI-Descriptors 38 GRAVH-3 -1 CDK PaDeI-Descriptors 26 ATSm2 -180 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 49 Estatemax -280 Rdkit 56 Phenol -487 Rdkit 0 Herbicide -855 EPA 47 geomDiameter -922 CDK PaDeI-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDeI-Descriptors	12	ALogp2	747	CDK	CDK Desc GUI
48 geomShape 173 CDK PaDel-Descriptors 41 MOM-Y 13 CDK PaDel-Descriptors 38 GRAVH-3 -1 CDK PaDel-Descriptors 26 ATSm2 -180 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 49 Estatemax -280 Rdkit 56 Phenol -487 Rdkit 0 Herbicide 855 EPA 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1217 CDK PaDel-Descriptors 46 geomRadius -1327 CDK PaDel-Descriptors	27	ATSm3	736	CDK	CDK Desc GUI
41 MOM-Y 13 CDK PaDel-Descriptors 38 GRAVH-3 -1 CDK PaDel-Descriptors 26 ATSm2 -180 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 49 Estatemax -280 Rdkit 55 Phenol -487 Rdkit 0 Herbicide -855 EPA 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDel-Descriptors	48	geomShape	173	CDK	PaDel-Descriptors
38 GRAVH-3 -1 CDK PaDel-Descriptors 26 ATSm2 -180 CDK CDEX Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 49 Estatemax -280 Rdkit 56 Phenol -487 Rdkit 0 Herbicide -855 EPA 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDel-Descriptors	41	MOMI-Y	13	CDK	PaDel-Descriptors
26 ATSm2 -180 CDK CDK Desc GUI 34 VAdjMat -217 CDK CDK Desc GUI 49 Estatemax -280 Rdkit 55 Phenol -487 Rdkit 0 Herbicide -855 EPA 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDel-Descriptors	38	GRAVH-3	-1	CDK	PaDel-Descriptors
34 VAdjMat -217 CDK CDK Desc GUI 49 Estatemax -280 Rdkit 56 Phenol -487 Rdkit 0 Herbicide -855 EPA 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSC4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDel-Descriptors	26	ATSm2	-180	CDK	CDK Desc GUI
49 Estatemax -280 Rdkit 56 Phenol -487 Rdkit 0 Herbicide -855 EPA 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDel-Descriptors	34	VAdjMat	-217	CDK	CDK Desc GUI
56 Phenol -487 Rdkit 0 Herbicide -855 EPA 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDel-Descriptors	49	Estatemax	-280	Rdkit	
0 Herbicide -855 EPA 47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDel-Descriptors	55	Phenol	-487	Rdkit	
47 geomDiameter -922 CDK PaDel-Descriptors 23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDel-Descriptors	0	Herbicide	-855	EPA	
23 ATSc4 -1178 CDK CDK Desc GUI 1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDel-Descriptors	47	geomDiameter	-922	CDK	PaDel-Descriptors
1 Pesticide -1211 EPA 46 geomRadius -1327 CDK PaDel-Descriptors	23	ATSc4	-1178	CDK	CDK Desc GUI
46 geomRadius -1327 CDK PaDel-Descriptors	1	Pesticide	-1211	EPA	
	46	geomRadius	-1327	CDK	PaDel-Descriptors

Table A-1. Full feature set used by Rank-IV submission

ToxCast MM Results Summary

© TopCoder, Inc. 2014



Descriptors	Score
CDK + Rdkit	1,104,812
72 Level8 hitcalls (shortlist after recursive selection)	1,080,170
EPA features: Categories, MW, 72 level8 hitcalls	1,148,333
EPA with 72 level8 hicalls + CDK + Rdkit	1,187,827
MACCS Keys	1,080,908
CDK + Rdkit + Morgan Fingerprints	1,121,801
EPA without any level8 hitcalls + CDK + Rdkit + Morgan Fingerprint features	1,180,151
EPA with 72 level8 hicalls + CDK + Rdkit + Morgan Fingerprint features	1,191,101
EPA with only 2 level 8 hitcalls + CDK + Rdkit + Morgan Fingerprint features (Final)	1,194,467

Table A-2. Scores for various feature combinations

© TopCoder, Inc. 2014

Description of the consensus model for prediction of Lowest Effect Level (LEL) concentrations (Technical description)

Brief summary

The model used in the EPA ToxCast challenge was built using OCHEM (helper tool described in the *EPA ToxCast - Chemical Structural Descriptors (pdf version), pdf* document under section "Databases" - "Online Chemical database", page 51).

OCHEM(online chemical database and modeling environment, <u>http://ochem.eu</u>)¹ is a publicly available web-based platform for QSAR research, developed and maintained by Helmholtz Zentrum München (<u>http://www.helmholtz-muenchen.de</u>) and BIGCHEM GmbH (http://bigchem.de).

Within the scope of the challenge, the 483 molecules with experimental LEL values were uploaded to OCHEM. They were used as the training set for the model (http://ochem.eu/basket/101620). Additionally, the whole challenge set of 1854 molecules values **OCHEM** and without were uploaded to used for predictions (http://ochem.eu/basket/102518). According to the EPA ToxCast challenge rules, no additional data from OCHEM-Database were used in this challenge, except for the data provided in the challenge.

The profile of the model used in the challenge is publicly available from the following link (<u>http://ochem.eu/model/174</u>), the article profile with links to the model and data are available from the following link (<u>http://ochem.eu/article/68104</u>).

Data preparation

The dataset from the *ToxRefDB_Challenge_Training.csv* file was used as a reference for creating training and test sets for the challenge. The file contains 1854 rows, each row representing one molecule. For 483 rows the experimental LEL values were provided, the rest were considered as predicted (test) set.

The file *ToxRefDB_Challenge_Training.csv* was appended by the SMILES column. The SMILES information was taken from the files containing structural information - *ToxCast_Generic_Chemicals/csv/ToxCast_Generic_Chemicals_2013_12_10.csv*.

Pairwise correspondence between the rows of the files was established using the CASRN column, which was present in both mentioned files.

The resulting file was uploaded to the OCHEM using the batch data upload tool (http://docs.ochem.eu/display/MAN/Batch+data+upload). Besides the LEL values and SMILES structural information, molecule names and CASRN were uploaded for convenience purposes.

OCHEM Modeling Workflows

Within the scope of the challenge, default OCHEM workflows were used for individual model creation process, as well as for the process of applying model to the final test set. Figure S1 illustrates the simplified OCHEM workflows for the modeling process.



Model creation workflow



Figure S1. Simplified model creation and prediction workflows used in the challenge.

The model training involves the following steps (described in greater detail in the next section):

- preprocessing of the initial molecular structures, cleaning, neutralization and salts removal
- generation of 3D coordinated (for the workflows that use 3D descriptors)
- calculation of the full descriptor matrix
- unsupervised descriptor selection based on descriptor correlation
- training of the model using the machine learning model, this step often includes meta-learning techniques like bootstrap aggregation
- output of the resulting model (machine learning method specific data)

Application of this model shares a significant amount of steps with the training process and involves:

- preprocessing of the initial molecular structures, cleaning, neutralization and salts removal
- generation of 3D coordinated (for the workflows that use 3D descriptors)
- calculation of the full descriptor matrix
- descriptor selection based on the list of required descriptors for the model being applied
- application of the machine learning model using the model data and descriptor matrix
- output of resulting predictions

Model description

Within the scope of the challenge, ten individual LEL models were built according to the previously mentioned workflows using different in-silico descriptor packages available in OCHEM. No EPA in-vitro features were used in this model. Although within the scope of the challenges several models with EPA in-vitro features were built, their inclusion into the final model did not result into a statistically significant increase in prediction accuracy. The final model was built as a simple average consensus over these ten individual models. Figure S2 illustrates the full architecture of the resulting LEL model.



Figure S2. Final consensus model structure diagram

The rationale to use the consensus model is based on the earlier observations,²⁻⁵ that such models usually provides higher accuracy compared to the individual approaches. Indeed, the models developed using different descriptors have different biases and their combination decreases their variances thus providing the optimal performance.^{6, 7}

Data preprocessing. All the molecular structures were preprocessed following a standardized OCHEM protocol. The molecules were:

- *Cleaned* the molecule structure was exported to SMILES format and imported back, the 2D coordinates were then recalculated from scratch.
- *Neutralized* charged atoms in the molecules were neutralized by attaching additional hydrogen atoms to them.
- Standardized inconsistent representation of Nitro groups was unified.
- *Remove salts* all salts or counter-ions were removed, only the largest molecular fragment was kept.

More information on OCHEM data preprocessing protocol is available at the OCHEM manual page - <u>http://docs.ochem.eu/display/MAN/Molecule+preprocessing</u>

3D Structure generation. Some of the descriptors require a 3D representation of the molecule. These descriptors were calculated based on an optimized 3D structure representation. The optimization was performed using OCHEM recommended optimization protocol that utilizes Corina⁸ tool by Molecular Networks GmbH.

Descriptor packages. Ten different descriptor packages implemented in the public

platform OCHEM were used individually to create ten models for the resulting consensus.

- *Electrotopological state indices (E-State indices)* molecular descriptors introduced by Kier and Hall.^{9,10} The descriptors combine electronic and topological properties of the described molecules. More information can also be found on the OCHEM manual page for E-State indices http://docs.ochem.eu/display/MAN/ESTATE
- Chemaxon descriptors also known as Chemaxon calculator plugins, is a set of insilico prediction tools provided by Chemaxon. More information can be found on the manual page (<u>http://docs.ochem.eu/display/MAN/Chemaxon+descriptors</u>) or on Chemaxon calculator plugin page (<u>http://www.chemaxon.com/products/calculatorplugins/</u>).
- *Inductive descriptors*¹¹ the descriptors calculated based on models of inductive and steric effects, inductive electronegativity and molecular capacitance. These molecular parameters are accessible from electronegativities and covalent radii of the constituent atoms and interatomic distances and can reflect a variety of aspects of intra- and intermolecular interactions (http://docs.ochem.eu/display/MAN/Inductive+Descriptors).
- *QNPR Descriptors* quantitative name-property relationship descriptors are derived directly from the compounds name or SMILES strings. For each molecule either canonical SMILES or IUPAC name are split into fragments of a specified length, and the fragments become the descriptor names. The values are calculated as the number of occurrences of a particular fragment in a string for a molecule. In the challenge the OCHEM implementation of QNPR was used, calculated over SMILES of length 1 3, only descriptors that occur in more than 5 molecules in the set were kept (http://docs.ochem.eu/display/MAN/QNPR).
- *ADRIANA.Code* is a descriptor package provided by Molecular Networks GmbH. It contains a series of methods for the generation of 3D structures and the calculation of physicochemical descriptors and molecular properties based on rapid empirical models (http://docs.ochem.eu/display/MAN/Adriana.CODE).
- *Mera, Mersy*¹² descriptor package developed by Prof. V.A. Potemkin, ModelChem group (http://docs.ochem.eu/display/MAN/MERA+descriptors).
- $GSfrag^{13, 14}$ a descriptor package that allows to calculate the occurrence numbers of certain special fragments on k=2, ..., 10 vertices in a molecular graph (http://docs.ochem.eu/display/MAN/GSFrag).
- *ISIDA SMF descriptors*¹⁵ the ISIDA implementation of the substructural molecular fragments (SMF) method, which is based on the splitting of a molecule into fragments. The fragment type is then a descriptor, and the number of occurrences of this fragment in a molecule is the value for this descriptor (http://docs.ochem.eu/display/MAN/ISIDA+Fragments).

Fragment lengths 2 - 4 were used in the challenge.

- *Dragon6 descriptors*¹⁶ a descriptor set from a popular descriptor tool Dragon developed by Kode s.r.l. (<u>http://docs.ochem.eu/display/MAN/Dragon</u>).
- *CDK descriptors* a set of descriptors implemented in the open-source Chemistry Development Kit. The constitutional, topological, geometrical, electronic and hybrid descriptor classes were used in the challenge

(http://docs.ochem.eu/display/MAN/CDK).

Descriptor selection. Within each individual model, basic unsupervised descriptor selection procedure was performed. First, descriptors with constant values for the dataset were removed. Next, descriptors with pair-wise correlation of more than 0.95 were eliminated.

Bootstrap aggregation (bagging) meta-learning approach with the bag size of 64 was used in each individual model. That is, each of the ten individual models in the challenge are an ensemble of 64 models, built on different training sets, which were obtained from the original one through resampling with replacement (http://docs.ochem.eu/display/MAN/Bagging).

Machine learning method. For all ten of the individual models, the Associative Neural Networks^{17, 18} (ASSN) machine learning method was used. ASNN is the author's implementation of the algorithm that uses local error correction in neural network ensemble prediction space.

The approach has a deep neurophysiological background and was inspired by the work of Dr. Tetko on the analysis of neuronal coding.^{19,20} It exploits an idea of ensemble learning. The ASNN approach has been used to introduce concepts of local libraries and self-learning, property-based similarity, applicability domain of models.^{4,18,21} The models developed with the ASNN were top-ranked in several benchmarking studies^{2,4,5,22} and that is why this method was selected for the EPA challenge.

In the challenge, neural network ensembles of size 64 were used, with the individual neural networks having 3 neurons in a hidden layer, trained by the SuperSAB algorithm. These are default parameters, which were optimized in the previous studies. Additional info can be also found on the OCHEM manual page -

http://docs.ochem.eu/display/MAN/ASNN+-+Associative+Neural+Networks.

It should be noticed that all sub-models were built using default settings as specified on Comprehensive Modeling template web page, as illustrated on Figure S3. (<u>http://ochem.eu/multiplemodels/create.do</u>). No optimization of parameters was used for the challenge.

Training set (required): LEL_training [d Add a validation set	letails]		
The model will predict this property: LEL using unit: -log(M)			
Select the methods you want to use fe	or the modeling:		
Method	Descriptors	Descriptor selection	Model validation
[ali] [none] ANN [edit] [x] # ASNN (bias correction) [edit] [x] KNN [edit] [x] DibSVM [edit] [x] MLRA [edit] [x] PLS [edit] [x] +add a custom template	[all [none] © CDK [edit] [x] © Dragon v6 (all blocks) [edit] [x] © GEstate [edit] [x] © GEstate [edit] [x] © GEstrag [edit] [x] © Mera and Mersy [edit] [x] © Chemaxon descriptors [edit] [x] © Chemaxon descriptors [edit] [x] © Adriana [edit] [x] © Spectrophores [edit] [x] © QNPR (SMILES - length 1 - 3 threshold 5) [edit] [x] - Two simple descriptors (MW+Number of carbons) [edit] [x] +add a custom template	[all] [none] □ Unsupervised forward selection [adit] [x] Simple pairwise decorrelation (r < 0.95) [adit] [x] +add a custom template	[all] [none] □ 5-fold cross-validation (edit[k) □ 5-fold cross-validation (stratified) [edit[k] ♥ Bagging with 64 models [edit] [x] □ Bagging with 64 models (stratified) [edit] [x] +add a custom template
Show advanced options>>			

Considering the selection above, 10 models will be created

Figure S3. Comprehensive modeling settings, which were used to develop the models.

Obtaining challenge predictions

The Final EPA ToxCast challenge model (<u>http://ochem.eu/model/174</u>) was applied to the full set of 1854 molecules (<u>http://ochem.eu/basket/102518</u>). The calculation results can be obtained from the public task profile (<u>http://ochem.eu/pendingtasks/profile.do?id=587790</u>).

Predictions were exported in CSV format, the file *ModelPredictions_1854_compounds.csv* was obtained. The training set was also exported for convenience reasons, resulting in the file *LEL_training.csv*.

In the resulting predictions file there are 37 errors (descriptors for these molecules could not be calculated for various reasons). Predictions for these molecules were taken as a mean value for the training set = $3.2602 - \log(M)$. Finally, 483 values for the training set molecules were substituted by their experimental results. The result was converted to the required format, *LELPredictor.java*.

All manipulations over the exported CSV files were performed using a simple Perl script (*reader.pl*).

The final model <u>http://ochem.eu/model/174</u> is freely available from the OCHEM web site and can be accessed and used by the EPA as well as any other web users without any limitations. The development and publishing models on-line allows their wide dissemination to chemical community, allows a rigorous validation of all steps of model development and it will be increasingly used in the future.²³

Table S1. Statistical parameters of models developed using different descriptors, machine learning methods and validation protocols using On-line CHEmical database and Modeling environment (http://ochem.eu)

A) RMSE errors of cross-validation and bagging models.

Adriana

QNPR (SMILES - length 1 - 3 threshold 5)

OEstate

EPA_LEL

Predicted property: LEL Training set: LEL_training								
Metrics RMSE - Root Mean Square Error 📀 for Training set	Cross-Validation (55 models)							
	ASNN	KNN	LibSVM	PLS	PLS(2)			
CDK (constitutional, topological, geometrical, electronic, hybrid)	0.98	0.95	0.95	0.96	0.95			
Dragon6 (blocks: 1-29)	0.98	0.96	0.92	0.97	0.94			
Fragmentor (Length 2 - 4)	1.07	1	0.96	0.97	0.97			
GSFrag	1.07	1.02	1.06	1.03	1.02			
Mera, Mersy	1.04	0.94	0.93	0.96	0.96			
ChemaxonDescriptors (7.4)	0.98	0.99	0.98	1	0.99			
InductiveDescriptors	0.99	0.99	1	1.15	1.04			
Adriana	0.97	1	0.98	1.08	1.04			
QNPR (SMILES - length 1 - 3 threshold 5)	1.01	1.01	0.94	0.97	0.97			
OEstate	1	0.99	0.99	0.97	0.97			
EPA_LEL	1.09	1.01	1	8	1.16			
Predicted property: LEL Training set: LEL_training Metrics RMSE - Root Mean Square Error 🗘 for Training set 🗘 Validation: Bagging (55 models)								
	ASNN	KNN	LibSVM	PLS	PLS(2)			
CDK (constitutional, topological, geometrical, electronic, hybrid)	0.93	0.93	0.9	0.94	0.93			
Dragon6 (blocks: 1-29)	0.93	0.95	0.9	0.96	0.94			
Fragmentor (Length 2 - 4)	0.98	0.97	0.91	0.99	0.96			
GSFrag	0.97	0.98	0.94	0.98	0.97			
Mera, Mersy	0.93	0.93	0.91	1.2	0.95			
ChemaxonDescriptors (7.4)	0.93	0.98	0.94	0.97	0.94			
InductiveDescriptors	0.94	0.98	0.94	0.99	0.99			

0.92

0.92

0.91

0.97

0.93

0.95

0.96

0.99

0.94

0.99

0.95

1

0.98

0.93

0.93

1.05

1.01

0.93

0.94

7

Training set: LEL_training					
Metrics R2 of Training set	Validation: Bagging (55 models)				
	ASNN	KNN	LibSVM	PLS	PL S(2)
CDK (constitutional, topological, geometrical, electronic, hybrid)	0.22	0.19	0.24	0.19	0.2
Dragon6 (blocks: 1-29)	0.24	0.17	0.23	0.19	0.2
Fragmentor (Length 2 - 4)	0.18	0.17	0.22	0.15	0.17
GSFrag	0.16	0.1	0.16	0.12	0.12
Mera, Mersy	0.21	0.19	0.23	0.08	0.18
ChemaxonDescriptors (7.4)	0.22	0.1	0.2	0.14	0.17
InductiveDescriptors	0.18	0.11	0.17	0.09	0.08
Adriana	0.22	0.17	0.2	0.08	0.11
QNPR (SMILES - length 1 - 3 threshold 5)	0.2	0.16	0.2	0.21	0.2
OEstate	0.18	0.18	0.22	0.19	0.21
EPA_LEL	0.1	0.09	0.11	0	0.03

B) Square of Pearson correlation coefficient of cross-validation and bagging models

The published models which, contribute to the final consensus model, are shown in blue.

Predicted property: LEL Training set: LEL_training						
Metrics R2 O for Training set	Validation: Cross-Validation (55 models)					
	ASNN	KNN	LibSVM	PLS	PLS(2)	
CDK (constitutional, topological, geometrical, electronic, hybrid)	0.2	0.16	0.16	0.13	0.17	
Dragon6 (blocks: 1-29)	0.22	0.16	0.2	0.15	0.18	
Fragmentor (Length 2 - 4)	0.13	0.14	0.15	0.13	0.13	
GSFrag	0.08	0.06	0.07	0.04	0.041	
Mera, Mersy	0.12	0.2	0.19	0.14	0.14	
ChemaxonDescriptors (7.4)	0.18	0.08	0.11	0.07	0.08	
InductiveDescriptors	0.14	0.09	0.06	0	0.021	
Adriana	0.2	0.11	0.11	0	0.013	
QNPR (SMILES - length 1 - 3 threshold 5)	0.18	0.11	0.18	0.14	0.14	
OEstate	0.15	0.13	0.13	0.13	0.13	
EPA_LEL	0.05	0.08	0.07	0	0	

Predicted property: LEL

Predicted property: LEL

References

- Horvath, D., Lisurek, M., Rupp, B., Kuhne, R., Specker, E., von Kries, J., Rognan, D., Andersson, C. D., Almqvist, F., Elofsson, M., Enqvist, P. A., Gustavsson, A. L., Remez, N., Mestres, J., Marcou, G., Varnek, A., Hibert, M., Quintana, J., and Frank, R. (2014) Design of a general-purpose European compound screening library for EU-OPENSCREEN. *ChemMedChem* 9, 2309-2326.
- (2) Tetko, I. V., Sushko, I., Pandey, A. K., Zhu, H., Tropsha, A., Papa, E., Oberg, T., Todeschini, R., Fourches, D., and Varnek, A. (2008) Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* 48, 1733-1746.
- Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., Oberg, T., Dao, P., Cherkasov, A., and Tetko, I. V. (2008) Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena pyriformis. *J. Chem. Inf. Model.* 48, 766-784.
- (4) Sushko, I., Novotarskyi, S., Korner, R., Pandey, A. K., Cherkasov, A., Li, J., Gramatica, P., Hansen, K., Schroeter, T., Muller, K. R., Xi, L., Liu, H., Yao, X., Oberg, T., Hormozdiari, F., Dao, P., Sahinalp, C., Todeschini, R., Polishchuk, P., Artemenko, A., Kuz'min, V., Martin, T. M., Young, D. M., Fourches, D., Muratov, E., Tropsha, A., Baskin, I., Horvath, D., Marcou, G., Muller, C., Varnek, A., Prokopenko, V. V., and Tetko, I. V. (2010) Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. J. Chem. Inf. Model. 50, 2094-2111.
- (5) Tetko, I. V., Novotarskyi, S., Sushko, I., Ivanov, V., Petrenko, A. E., Dieden, R., Lebon, F., and Mathieu, B. (2013) Development of dimethyl sulfoxide solubility models using 163 000 molecules: using a domain applicability metric to select more reliable predictions. J. Chem. Inf. Model. 53, 1990-2000.
- (6) Geman, S., Bienenstock, E., and Doursat, R. (1992) Neural networks and the bias-variance dilemma. *Neural Comput.* 4, 1-58.
- Tetko, I. V., Livingstone, D. J., and Luik, A. I. (1995) Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.* 35, 826-833.
- (8) Online Demo Fast 3D Structure Generation with CORINA. http://www.molecular-networks.com/online_demos/corina_demo (March 6, 2016).
- (9) Hall, L. H., and Kier, L. B. (1995) Electrotopological State Indexes for Atom Types - a Novel Combination of Electronic, Topological, and Valence State Information. J. Chem. Inf. Comput. Sci. 35, 1039-1045.
- (10) Kier, L. B., and Hall, L. H. (1999) *Molecular Structure Description: The Electrotopological State*. Academic Press, London.
- (11) Cherkasov, A. (2005) 'Inductive' Descriptors: 10 Successful Years in QSAR. *Curr. Comp. Aid. Drug Design 1*, 21-42.
- (12) Potemkin, V. A., and Grishina, M. A. (2008) A new paradigm for pattern recognition of drugs. *J. Comput. Aided. Mol. Des.* 22, 489-505.
- (13) Skvortsova, M. I., Baskin, I. I., Skvortsov, L. A., Palyulin, V. A., Zefirov, N. S., and Stankevich, I. V. (1999) Chemical graphs and their basis invariants. *J. Mol. Struct.* 466, 211-217.
- (14) Skvortsova, M. I., Fedyaev, K. S., Baskin, I. I., Palyulin, V. A., and Zefirov, N. S. (2002) A New Technique for Coding Chemical Structures Based on Basis

Fragments. Dokl. Chem. 382, 33-36.

- (15) Varnek, A., Fourches, D., Horvath, D., Klimchuk, O., Gaudin, C., Vayer, P., Solov'ev, V., Hoonakker, F., Tetko, I. V., and Marcou, G. (2008) ISIDA -Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* 4, 191-198.
- (16) Todeschini, R., and Consonni, V. (2000) *Handbook of Molecular Descriptors*. WILEY-VCH, Weinheim.
- (17) Tetko, I. V. (2002) Neural network studies. 4. Introduction to associative neural networks. *J. Chem. Inf. Comput. Sci.* 42, 717-728.
- (18) Tetko, I. V. (2008) Associative neural network. *Methods Mol. Biol. 458*, 185-202.
- (19) Villa, A. E., Tetko, I. V., Hyland, B., and Najem, A. (1999) Spatiotemporal activity patterns of rat cortical neurons predict responses in a conditioned task. *Proc. Natl. Acad. Sci. U. S. A.* 96, 1106-1111.
- (20) Tetko, I. V., and Villa, A. E. (2001) A pattern grouping algorithm for analysis of spatiotemporal patterns in neuronal spike trains. 2. Application to simultaneous single unit recordings. *J. Neurosci. Methods* 105, 15-24.
- (21) Tetko, I. V., Poda, G. I., Ostermann, C., and Mannhold, R. (2009) Large-scale evaluation of log P predictors: local corrections may compensate insufficient accuracy and need of experimentally testing every other compound. *Chem. Biodivers.* 6, 1837-1844.
- (22) Tetko, I. V., Solov'ev, V. P., Antonov, A. V., Yao, X., Doucet, J. P., Fan, B., Hoonakker, F., Fourches, D., Jost, P., Lachiche, N., and Varnek, A. (2006) Benchmarking of linear and nonlinear approaches for quantitative structureproperty relationship studies of metal complexation with ionophores. *J. Chem. Inf. Model.* 46, 808-819.
- (23) Tetko, I. V. (2012) The perspectives of computational chemistry modeling. J. Comput. Aided. Mol. Des. 26, 135-136.