EPA ToxCast LELPredictor Marathon Match Results Summary

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

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 RMSE 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

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.

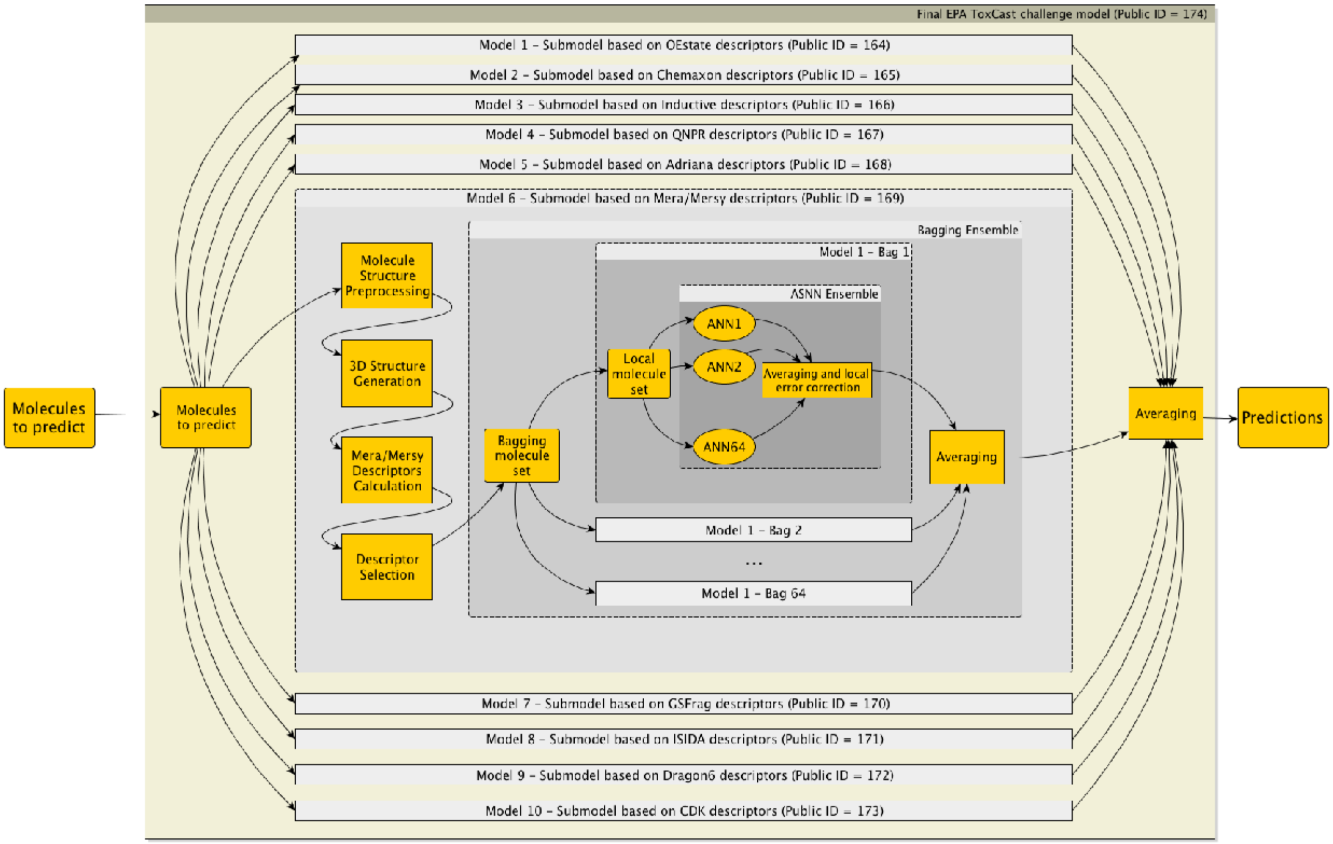

Final EPA ToxCast challenge model (Public ID = 174)
Model 1 – Submodel based on OEstate descriptors (Public ID = 164)
Model 2 – Submodel based on Chemaxon descriptors (Public ID = 165)
Model 3 – Submodel based on Inductive descriptors (Public ID = 166)
Model 4 – Submodel based on QNPR descriptors (Public ID = 167)
Model 5 – Submodel based on Adriana descriptors (Public ID = 168)
Model 6 – Submodel based on Mera/Mersy descriptors (Public ID = 169)
Bagging Ensemble
Model 1 – Bag 1
ASNN Ensemble
Molecules to predict
Molecules to predict
Molecule Structure Preprocessing
3D Structure Generation
Mera/Mersy Descriptors Calculation
Descriptor Selection
Bagging molecule set
Local molecule set
ANN1
ANN2
ANN64
Averaging and local error correction
Averaging
Model 1 – Bag 2
...
Model 1 – Bag 64
Averaging
Predictions
Model 7 – Submodel based on GSFrag descriptors (Public ID = 170)
Model 8 – Submodel based on ISIDA descriptors (Public ID = 171)
Model 9 – Submodel based on Dragon6 descriptors (Public ID = 172)
Model 10 – Submodel based on CDK descriptors (Public ID = 173)

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:

- 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:

- 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_Synonyms", "ChemNote", "STRUCTURE_Formula", "STRUCTURE_IUPAC", "STRUCTURE_SMILES", "STRUCTURE_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.

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

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

Rank-IV Submission

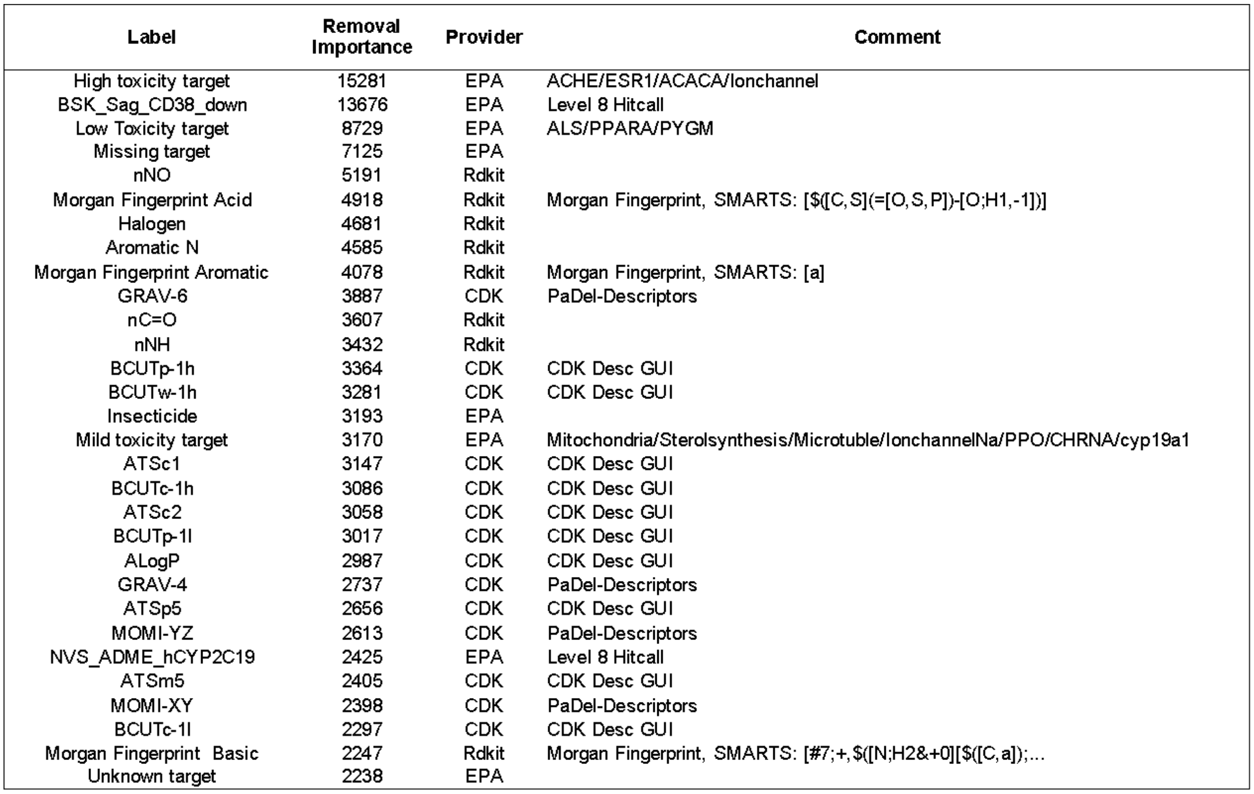
**several scores are shown in this description. These scores were obtained on training set by the member locally. These scores should not be compared in any manner with the provisional or system scores obtained from final submission. All the details provided here are extracted from final documentation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



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

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

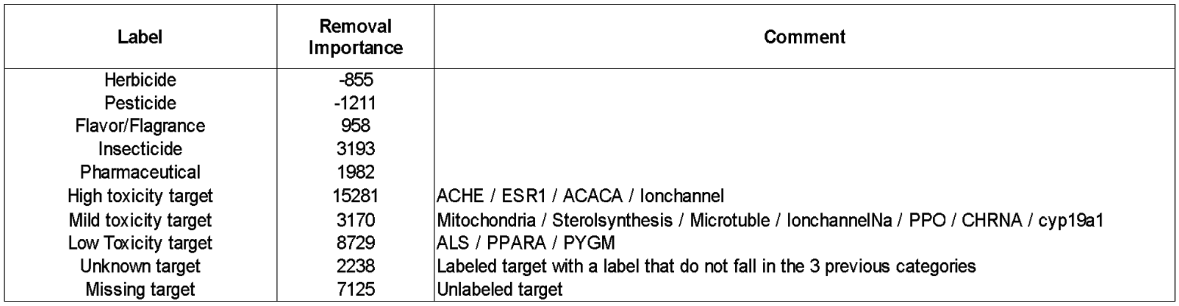
- "Removal Importance" is the difference of 2 scores, computed with and without the feature of interest.

- 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:

  - 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.
  - 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".
  - "UseCategory" was grouped into more general classes: "Herbicide", "Pesticide",

- “Flavor/Flagrance”, “Insecticide”, “Pharma/Drugs”.
- “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).
- 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

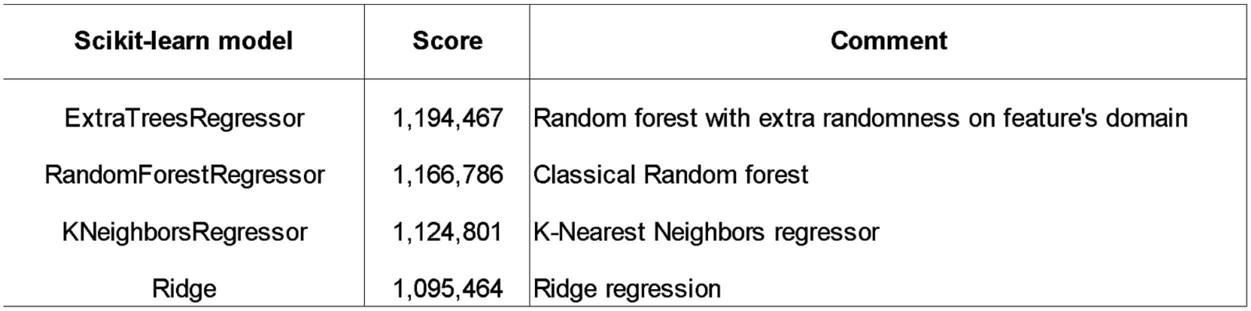
Table 4-2. Breakdown of Classification sorted by removal importance

- 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

- 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.

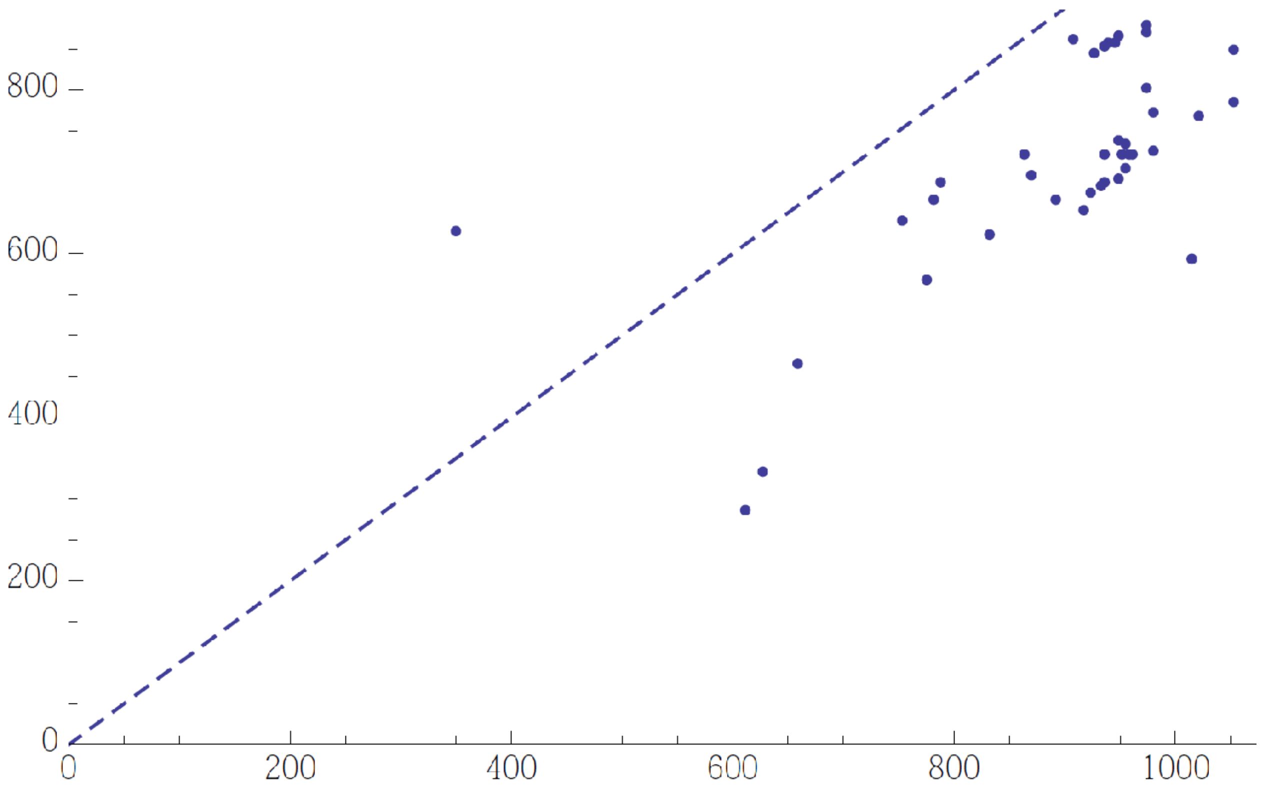


Figure 5-1. Final (Y-Axis) vs. Provisional (X-Axis) Scores

- 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.
- 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.
- 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:

- RMSE: Root-Mean Squared Error (lower the value, better the result)
- RMSE (80%): RMSE on 80% best predictions (lowest prediction error)
- Pearson: just Pearson correlation coefficient (higher the value, better the result)
- Pearson (80%): the same on 80% best predictions
- AUC: percentage of pairs where predicted1 < predicted2 among those where ground_truth1 < ground_truth2 (higher the value, better the result)
- 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:

- 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:

  - 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.
  - 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:

Solution	AUC (%)	N
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:

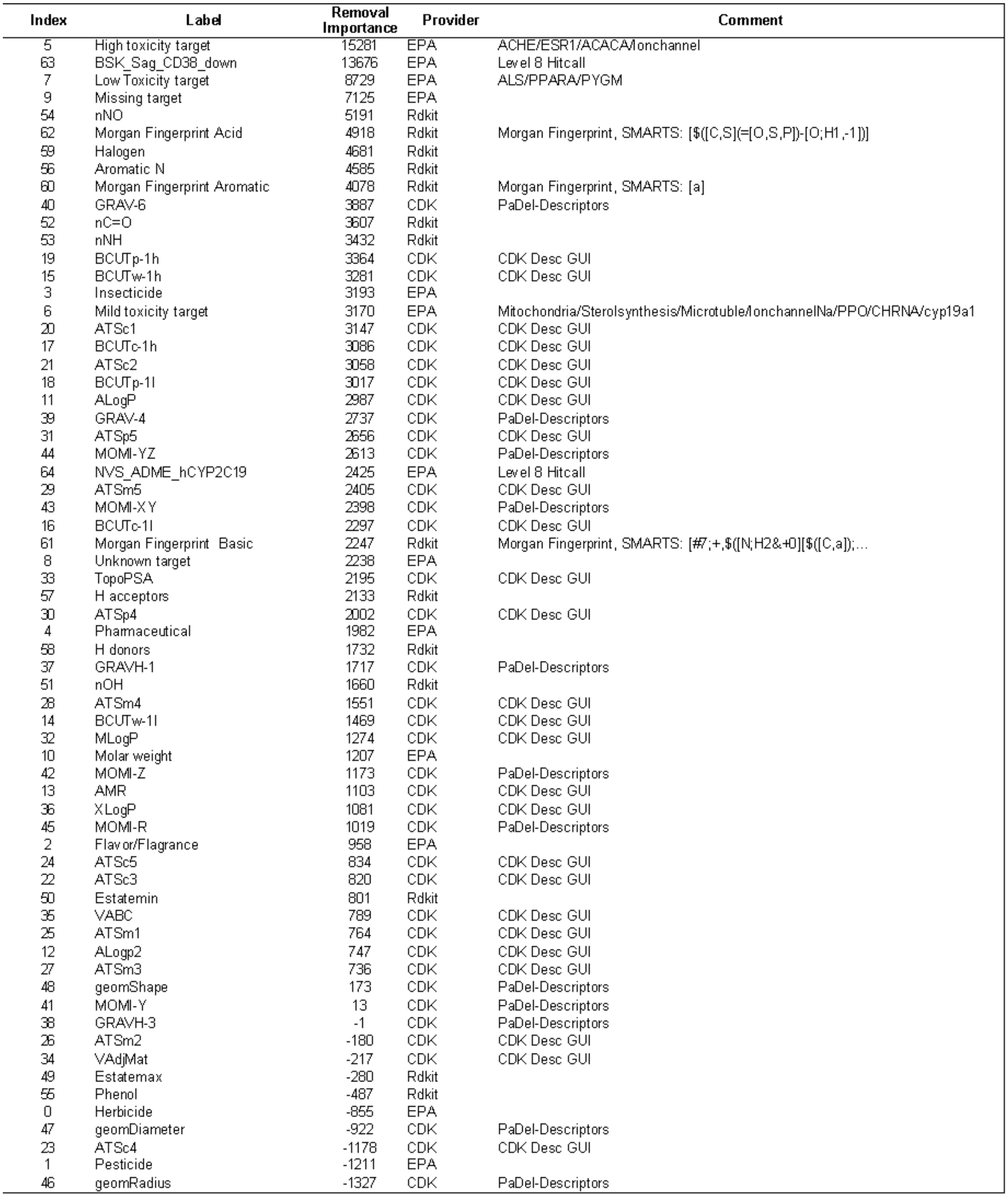
  - Numerical driven data science approach (Random forest) vs. highly sophisticated domain-specific approach (OCHEM tool)
  - Use of in-vitro features vs. use of molecular descriptors and also their combination.
  - EPA's current algorithmic approach to the problem vs. solution available from community.

- 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.

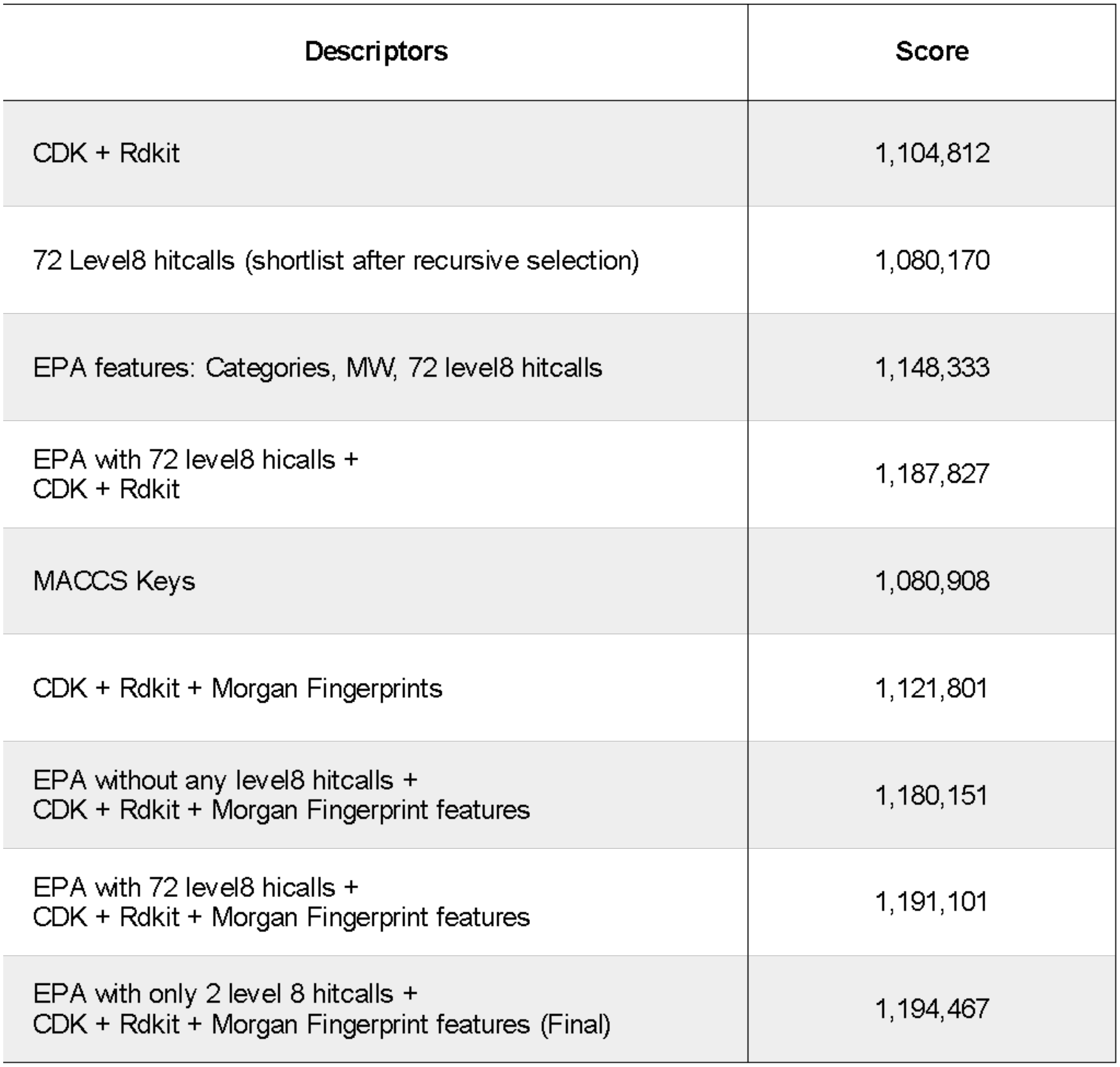
- We are also doing some more statistical analysis and gauging the scientific relevance of these submissions through our follow-up round which will be presented separately. Such an analysis will help to build further confidence in this already promising outcome.

Appendix – Feature Sets used by Rank-IV submission



Index	Label	Removal Importance	Provider	Comment
5	High toxicity target	15281	EPA	ACHE/ESR1/ACACA/Ionchannel
63	BSK_Sag_CD38_down	13676	EPA	Level 8 Hitcall
7	Low Toxicity target	8729	EPA	ALS/PPARA/PYGM
9	Missing target	7125	EPA	
54	nNO	5191	Rdkit	
62	Morgan Fingerprint Acid	4918	Rdkit	Morgan Fingerprint, SMARTS: [$([C,S](=[O,S,P])-[O;H1,-1])]
59	Halogen	4681	Rdkit	
56	Aromatic N	4585	Rdkit	
60	Morgan Fingerprint Aromatic	4078	Rdkit	Morgan Fingerprint, SMARTS: [a]
40	GRAV-6	3887	CDK	PaDel-Descriptors
52	nC=O	3607	Rdkit	
53	nNH	3432	Rdkit	
19	BCUTp-1h	3364	CDK	CDK Desc GUI
15	BCUTw-1h	3281	CDK	CDK Desc GUI
3	Insecticide	3193	EPA	
6	Mild toxicity target	3170	EPA	Mitochondria/Sterolsynthesis/Microtuble/IonchannelNa/PPO/CHRNA/cyp19a1
20	ATSc1	3147	CDK	CDK Desc GUI
17	BCUTc-1h	3086	CDK	CDK Desc GUI
21	ATSc2	3058	CDK	CDK Desc GUI
18	BCUTp-1l	3017	CDK	CDK Desc GUI
11	ALogP	2987	CDK	CDK Desc GUI
39	GRAV-4	2737	CDK	PaDel-Descriptors
31	ATSp5	2656	CDK	CDK Desc GUI
44	MOMI-YZ	2613	CDK	PaDel-Descriptors
64	NVS_ADME_hCYP2C19	2425	EPA	Level 8 Hitcall
29	ATSm5	2405	CDK	CDK Desc GUI
43	MOMI-XY	2398	CDK	PaDel-Descriptors
16	BCUTc-1l	2297	CDK	CDK Desc GUI
61	Morgan Fingerprint Basic	2247	Rdkit	Morgan Fingerprint, SMARTS: [#7;+,$([N;H2&+0][$([C,a]);...
8	Unknown target	2238	EPA	
33	TopoPSA	2195	CDK	CDK Desc GUI
57	H acceptors	2133	Rdkit	
30	ATSp4	2002	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	1551	CDK	CDK Desc GUI
14	BCUTw-1l	1469	CDK	CDK Desc GUI
32	MLogP	1274	CDK	CDK Desc GUI
10	Molar weight	1207	EPA	
42	MOMI-Z	1173	CDK	PaDel-Descriptors
13	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	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
12	ALogp2	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
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

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



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