**CERAPP: Collaborative Estrogen Receptor Activity Prediction Project**

**Combining Estrogen Receptor Structure-based Models in a Large Collection of Environmental Chemicals**

**Contact**: Kamel Mansouri, Richard Judson, U.S. EPA

[mansouri.kamel@epa.gov](mailto:mansouri.kamel@epa.gov)

[judson.richard@epa.gov](mailto:judson.richard@epa.gov)

919-541-3085

**Experimental data for evaluation set**

**Subgroup**:

* U.S.EPA/NCCT: Kamel Mansouri, Jayaram Kancherla, Ann Richard, Richard Judson
* UMEA/Chemistry: Aleksandra Rybacka, Patrik Andersson
* FDA/NCTR/DBB: Huixiao Hong
* NIH/NCATS: Ruili Huang
* Helmholtz/ISB: Igor Tetko

**Collected data:**

In vitro experimental data for this step of the CERAPP project was collected from different overlapping sources including EPA’s high-throughput screening assays, online databases and other data sets used by participants to train models.

1. Tox21, ~8000 chemicals in 4 assays;
2. The FDA DEAC / EDKB database of ~8000 chemicals from the literature;
3. METI database, ~2000 chemicals;
4. ChEMBL database, ~2000 chemicals.

**Data cleaning procedure:**

All molecular structures from these sources were curated and standardized using the previously developed KNIME workflow. Generated standard InChI codes were used to uniquely identify the chemicals.

The resulting data were combined into a single file after unifying all information fields from the different sources using data-mining tools implemented in KNIME. The full dataset consisted of more than 60,000 entries for ~15,000 unique chemicals. (ALL\_exp\_data.xls)

Only 7,600 chemicals overlapped with the CERAPP prediction set, for a total of 45,000 entries. The non-overlapping chemicals were excluded from the following process.

After that, *in-vivo* data, cytotoxicity data and all ambiguous entries (missing values, non-defined/standard endpoints, not clear units…) were removed from the dataset. Then all data was categorized according to the following assay classes: binding, reporter gene or cell proliferation.

In order to have an activity concentration value (as an equivalent to IC50) for each entry, endpoint values of entries with “ENDPOINT\_NAME” of:

* IC50, AC50, REC10, GI50, EC50, Kd and Ki were converted uM.
* -log(IC50), -log(Ki), logPC50, logRA, logRBA, logRA10, logRE, logRPE and logRPP were antilog transformed and converted to uM.

Chemicals with cell proliferation assays explained in percentage were considered as actives if they exceed an arbitrary threshold of 125% proliferation. The reported testing concentration mentioned in the “ASSAY\_NAME” field was considered as the activity concentration and normalized to unit in uM. All inactive compounds were assigned to a value of 1M.

The obtained data for 7547 CERAPP compounds from 44641 entries was saved into a second file. (CEARAPP\_exp\_data\_in.xls).

**Reference Chemicals:**

A list of 36 known positive and negative reference chemicals was used to categorize the collected chemicals from the literature. These same chemicals have been used to validate the ER model for the training set of CERAPP project (shared manuscript by Judson et al).

26 of these chemicals overlapped with the list of chemicals of CERAPP prediction set and the collected data from the literature that will be used as validation set.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CERAPP ID | CHEMICAL NAME | AUC | Agonist (AC50) | Antagonist (AC50) | Binding (AC50) | Known Potency Class |
| 10979 | meso-Hexestrol | 0.879 | 0.020 | 100000 | 0.020 | Strong |
| 10017 | 17beta-Estradiol | 0.792 | 0.059 | 100000 | 0.059 | Strong |
| 10121 | Estrone | 0.74 | 0.099 | 100000 | 0.099 | Moderate |
| 15909 | Bisphenol AF | 0.684 | 0.182 | 4406.137492 | 0.182 | Weak |
| 13660 | Genistein | 0.574 | 0.545 | 100000 | 0.545 | Weak |
| 10719 | Bisphenol B | 0.548 | 0.730 | 4428.118177 | 0.730 | Weak |
| 10862 | Bisphenol A | 0.458 | 1.741 | 100000 | 1.741 | Weak |
| 13993 | Apigenin | 0.433 | 2.222 | 1493.911442 | 2.222 | Weak |
| 13804 | Daidzein | 0.407 | 3.069 | 100000 | 3.069 | Weak |
| 14488 | 4-Cumylphenol | 0.394 | 3.492 | 100000 | 3.492 | Weak |
| 13989 | Kaempferol | 0.383 | 3.851 | 100000 | 3.851 | Weak |
| 11369 | Butylparaben | 0.364 | 4.470 | 5497.966751 | 4.470 | Weak |
| 12892 | 2,2',4,4'-Tetrahydroxybenzophenone | 0.347 | 5.054 | 100000 | 5.054 | Moderate |
| 15023 | o,p'-DDT | 0.335 | 5.938 | 100000 | 5.938 | Weak |
| 12548 | Ethylparaben | 0.214 | 20.359 | 100000 | 20.359 | Very Weak |
| 10556 | Methoxychlor | 0.203 | 24.645 | 100000 | 24.645 | Very Weak |
| 13180 | Kepone | 0.0724 | 150.158 | 100000 | 150.158 | Weak |
| 10255 | Progesterone | 0.0167 | 861.680 | 100000 | 861.680 | Inactive |
| 10013 | Corticosterone | 0.0121 | 1201.287 | 100000 | 1201.287 | Inactive |
| 19287 | Tamoxifen | 0.417 | 3710.082 | 2.742938463 | 2.743 | Weak |
| 16215 | Atrazine | 0 | 100000 | 100000 | 100000 | Inactive |
| 13503 | Linuron | 0 | 100000 | 100000 | 100000 | Inactive |
| 10097 | Spironolactone | 0 | 100000 | 100000 | 100000 | Inactive |
| 10115 | Haloperidol | 0 | 100000 | 100000 | 100000 | Inactive |
| 23243 | Ketoconazole | 0 | 100000 | 100000 | 100000 | Inactive |
| 10460 | Cycloheximide | 0 | 100000 | 100000 | 100000 | Inactive |

This categorization procedure was applied on entries with endpoint names: AC50, PC50, IC50, GI50 and EC50.

According to the classes of the reference chemicals, the following thresholds were applied on the endpoint values:

* Strong: 0-0.09
* Moderate: 0.09-0.18
* Weak: 0.18-20
* Very Weak: 20-800
* Inactive: 800>

The classification of the chemicals using these 26 chemicals was not applied on the following endpoints for inconsistency: REC10, EC50, Kd, Ki, -log(Ki), logRA, logRBA, logRA10, logRE, logRPE and logRPP.

The 5 classes were assigned to scores from 0 (inactive) to 1 (strong). Then, for each chemical, the average of the scores of the merged entries from different literature sources was calculated.

A new class was generated for the merged entries according to the following thresholds:

* Average score=0 => Inactive;
* 0 < Average score < = 0.25 => Very Weak
* 0.25 < Average score < = 0.5 => Weak
* 0.5< Average score < = 0.75 => Moderate
* Average score > 0.75 => Strong

**Summary file**

The summary file was created by grouping the entries into Agonist, Antagonist or Binding class so that each chemical is represented by one, two or three rows. If a chemical is an active agonist or antagonist, it was considered also as an active binder. This final summary file contains 17,118 rows. (CEARAPP\_exp\_data\_in\_Sum.xls)

For each row the average and the median concentration values for activity were calculated in order to characterize the potency of the compounds. This was achieved by converting the concentrations to Molar units, calculating the average and median of the log10 values then applying the antilog to convert back to Molar units.

**Binary classification Active/Inactive**

To analyze the consistency between the different sources of data, a new field was added to flag the chemicals where all assays are consistent. In other words, for these chemicals, if there are multiple data points, all of them agree as to whether the chemical is active or inactive. 7,511 chemicals in 14,948 rows were flagged. These are summarized in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Active | Inactive | Total |
| Binding | 1999 | 259 | 2258 |
| Agonist | 338 | 5856 | 6194 |
| Antagonist | 275 | 6221 | 6496 |
| Total | 2612 | 12336 | 14948 |

To be able to use more of the data where most of the reports are consistent, another field was added to flag chemicals where the number of assays in one class (active, inactive) is at least 5 times larger than the assays from the opposite class.

This increased the number of considered chemicals to 7,522 chemicals in 15,141 rows summarized in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Active | Inactive | Total |
| Binding | 2024 | 259 | 2283 |
| Agonist | 350 | 5969 | 6319 |
| Antagonist | 284 | 6255 | 6539 |
| Total | 2658 | 12483 | 15141 |

**Classification to 5 classes**

Entries with equivalent endpoints (AC50, PC50, IC50, GI50 and EC50) were merged. The mean, median and standard deviation were calculated for the activity concentrations (AC50eq in uM) and log values (AC50eq in M).

The number of merged entries is also shown in the column AC50 (eq) (Count\*).

This resulted in 16372 rows (Agonist, Antagonist, and Binding) 7272 unique chemicals with score values.

Then the classification procedure based on the reference chemicals and the 5 class thresholds was applied. The generated classes were checked to be in accordance with the binary classification (active/inactive) in the previous step.

The total number of categorized rows is now 14399 for 7253 unique chemicals summarized in the following table.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Inactive | V. Weak | Weak | Moderate | Strong | Total |
| Binding | 0 | 688 | 894 | 72 | 77 | 1731 |
| Agonist | 5892 | 19 | 179 | 31 | 42 | 6163 |
| Antagonist | 6221 | 76 | 188 | 10 | 10 | 6505 |
| Total | 12113 | 783 | 1261 | 113 | 129 | 14399 |

This procedure helped understanding and categorizing many cases shown to be inconsistent in the previous step. (For example when 4 entries for inactive and 2 for very weak)

**Final evaluation set**

5221 chemicals were noticed to be inactive in agonist and antagonist assays with both classification methods (binary and 5 potency classes with reference chemicals). 5000 of these chemicals are with no available binding assays. So these chemicals can be considered as non-binders and will be added to the binding assays as inactives.

Between the 221 overlapping chemicals, there are 42 that are reported as actives in binding assays but inactives in both agonists and antagonists. These 42 chemicals will be considered as false positives because for most of them only one source is available while the number of agonist and antagonist assays is higher.

With the additional 5000 entries to the binding assays we obtain the following tables.

**Evaluation set for binary classification models**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Active | Inactive | Total |
| Binding | 1982 | 5301 | 7283 |
| Agonist | 350 | 5969 | 6319 |
| Antagonist | 284 | 6255 | 6539 |
| Total | 2616 | 17525 | 20141 |

**Evaluation set for quantitative models**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Inactive | V. Weak | Weak | Moderate | Strong | Total |
| Binding | 5042 | 685 | 894 | 72 | 77 | 6770 |
| Agonist | 5892 | 19 | 179 | 31 | 42 | 6163 |
| Antagonist | 6221 | 76 | 188 | 10 | 10 | 6505 |
| Total | 17155 | 780 | 1261 | 113 | 129 | 19438 |

**Procedure for the evaluation of the predictions**

The proposed procedure for the evaluation of the predictions received from the different groups will be mainly based on the qualitative experimental data since most of the models are classification. For that we can use the binary classification dataset of 20141 rows.

**Classification/ qualitative models**

The performance of classification models will be evaluated using statistical indices proposed in the literature. These indices are calculated from the confusion matrix which collects the number of samples of the observed and predicted classes in the rows and columns, respectively.

For a two-class dataset, the classification parameters are defined using the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).

The most important parameter that will be considered during the evaluation step is the Non-Error Rate (𝑁𝐸𝑅) or balanced accuracy. It is usually expressed in percentage and given by:

where the Sensitivity (𝑆𝑛), or True Positive Rate (𝑇𝑃𝑅):

and the Specificity (𝑆𝑝), or True Negative Rate (𝑇𝑁𝑅):

The different models will be ranked according to their respective NER.

**Regression/ quantitative models**

The quantitative models can be evaluated as classification models using a threshold (active/inactive) or categorizing the predictions into 5 classes and evaluate them using the second subset.

Models with AC50 predictions can be evaluated using the (median) experimental values for activity concentrations (AC50eq) and the correlation with the potency level scores. The commonly used parameter Root Mean Square Error, with a non-unit slope and ROC curves can be calsulated as an indication of the accuracy of predictions.

All models will be evaluated on the training set and the evaluation set from the literature taking into consideration the applicability domain of each model.

**Consistency of the evaluation set**

To investigate the concordance between the two receptor subtypes (alpha and beta) and the different species, a concordance analysis has been conducted on the different sources.

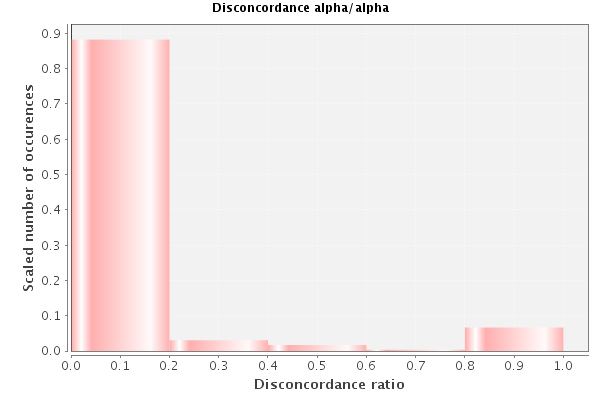
For each chemical, the disconcordance ratio was calculated. A ratio of zero means that all sources are concordant while a ratio of 1 means that the number of sources reporting that a specific chemical is active is equal to the number of chemicals reporting that it’s inactive.

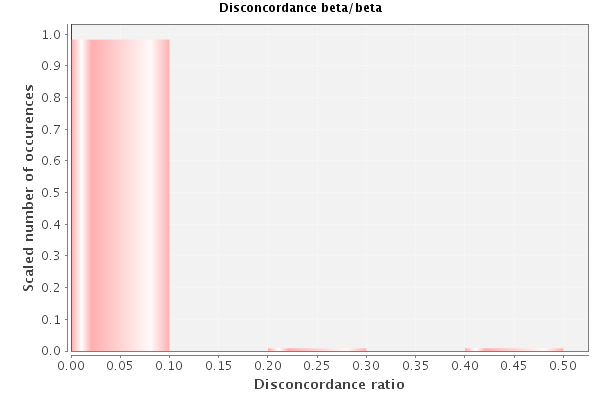
The disconcordance ratios are shown in histogram plots for interpretation.

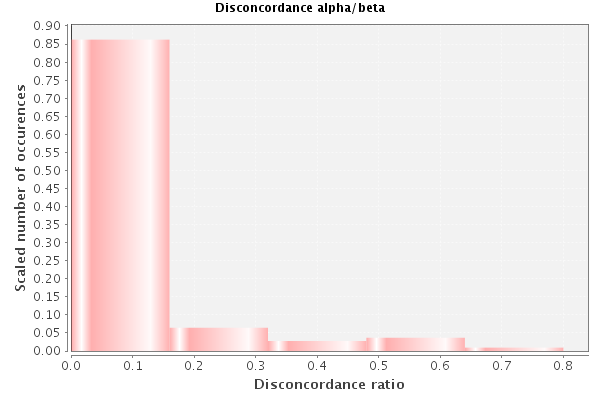
**Consistency alpha/beta**

As shown by the following 3 figures, the concordance alpha/beta is almost as high as the concordance alpha/alpha and beta/beta. The concordance beta/beta is higher than alpha/alpha due to the low number of available assays for beta compared to alpha.

These histograms confirm the possibility of mixing alpha and beta assays.

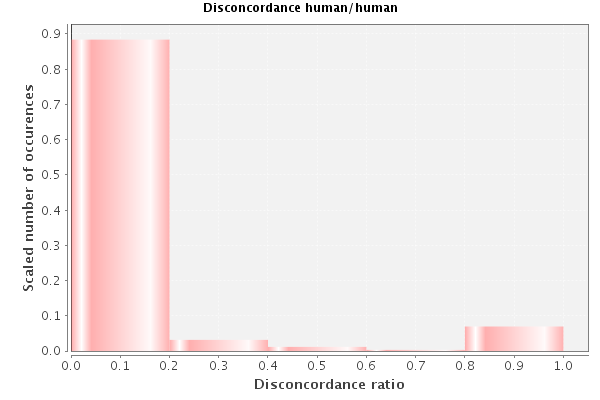






**Consistency between species**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Species** | Human | Rat | N.A | Mouse | Sheep | Trout | Cattle | Lizard | Chicken | Rabbit |
| **Number** | 47046 | 1260 | 764 | 134 | 44 | 36 | 25 | 19 | 16 | 10 |



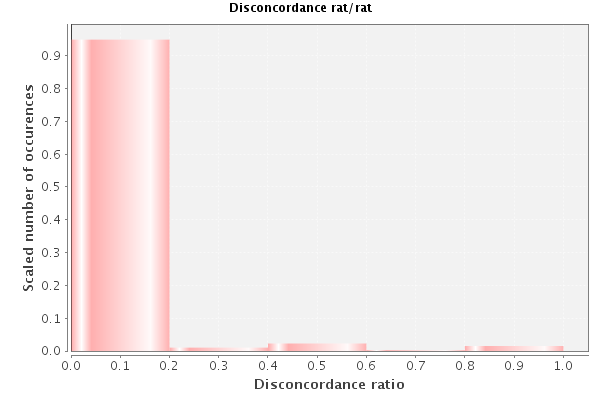
The only two species that have intraspecies disconcordance are human are rat. The other species presented no disconcordance due to the low number of available sources.

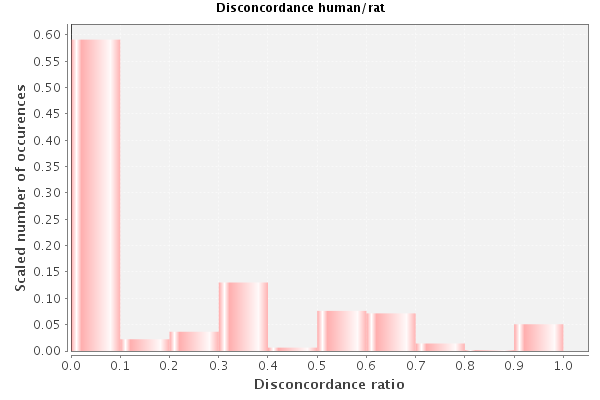
The human/human and rat/rat histograms show high intraspecies concordance. Rat is showing higher concordance due to the lower number of sources compared to human.

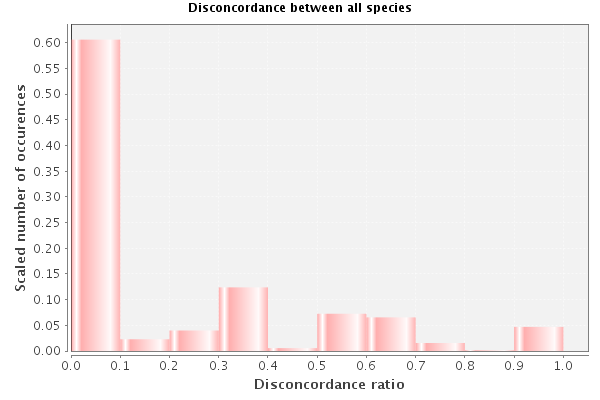
The human/rat and all-species combined concordances are also high (about 60%), which makes it possible to merge assays of all species together.

The interspecies concordance is lower than the intraspecies concordance because of the addition of noise from each one of them in addition to other sources of noise such as alpha/beta disconcordance.

Further intra and interspecies analysis will be conducted.







**Consistency training set ⬄ evaluation set**

*In classification/ qualitative data:*

In order to be able to use the collected dataset from the literature as an evaluation set for this project, the consistency with the training set used to build most of the models need to be checked.

The overlap between the training set and the evaluation set is 1659 chemicals. This overlapping dataset from the training set was evaluated against the literature data in both previously suggested AUC thresholds 0.01 and 0.1. Then, the very weak active chemicals that seemed to be ambiguous and origin of disconcordance between the two datasets were changed as inactives then removed to evaluate their effect on the overall consistency of the two datasets.

The non-error rate (balanced accuracy) as well as specificity and sensitivity were calculated and reported in the following table.

All sources combined

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Threshold | N. Chem | NER | Sn | Sp |
| All orig. | 0.01 | 1659 | 75.4 | 79.13 | 71.67 |
| All orig. | 0.1 | 1659 | 78.57 | 89.29 | 67.86 |
| VW to N | 0.01 | 1659 | 71.35 | 56.31 | 86.38 |
| VW to N | 0.1 | 1659 | 81.92 | 79.76 | 84.07 |
| No VW | 0.01 | 1424 | 78.5 | 72.96 | 84.04 |
| No VW | 0.1 | 1424 | 84.57 | 88.16 | 80.99 |

This table confirms that the threshold of 0.1 is the most reasonable to be applied on AUC values. It also shows that the very weak actives are source of disconcordance and removing them increases the consistency between the training set and the evaluation set.

The same calculations were applied on the subset of the evaluation set after removing the overlapping chemicals with only one literature source. The new overlapping subset consisted of 1410 chemicals.

Entries with only one source removed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Threshold | N. Chem | NER | Sn | Sp |
| All orig. | 0.01 | 1410 | 79.65 | 75.3 | 84.0 |
| All orig. | 0.1 | 1410 | 84.42 | 88.46 | 80.38 |
| VW to N | 0.01 | 1410 | 76.86 | 62.05 | 91.67 |
| VW to N | 0.1 | 1410 | 85.49 | 82.05 | 88.94 |
| No VW | 0.01 | 1306 | 81.26 | 71.53 | 90.98 |
| No VW | 0.1 | 1306 | 87.79 | 87.67 | 87.9 |

Removing the chemicals with only one source increased the consistency between the two datasets. This could be explained by the experimental uncertainty that is considerably lower in multi-sources data.

This suggests that the predictions from the participant models could be evaluated first considering the whole evaluation set then considering only entries with at least two sources.

*In regression/ quantitative data:*

The consistency of the collected quantitative data from the literature and the training set is evaluated using Pearson’s correlation (r2) between the AUC scores (training set) and the median value of log(AC50eq) (literature).

The overall correlation including both actives and inactives is 0.39.

The following tables summarized the correlation values for only the active chemicals in both thresholds:

* In training set
* Intersection of active chemicals in training and evaluation sets
* In training set, turning very weak chemicals in eval set to negatives
* Intersection of active chemicals in training set and eval set with very weak turned negatives.

All sources combined

|  |  |  |  |
| --- | --- | --- | --- |
|  | Threshold | N. Chem | r2 |
| Actives in tr | 0.01 | 236 | 0.51 |
| Actives in tr | 0.1 | 89 | 0.72 |
| Actives in tr & eval | 0.01 | 177 | 0.69 |
| Actives in tr & eval | 0.1 | 78 | 0.81 |
| Actives in tr (VW to N) | 0.01 | 236 | 0.58 |
| Actives in tr (VW to N) | 0.1 | 89 | 0.73 |
| Actives in tr & eval (VW to N) | 0.01 | 130 | 0.65 |
| Actives in tr & eval (VW to N) | 0.1 | 70 | 0.79 |

The highest correlations are obtained with the threshold 0.1 for the actives in both sets and after considering the very weak chemicals into negatives.

Entries with only one source removed

The same steps were applied after removing the literature data with only one source. The overall correlation including actives and inactives is 0.46.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Threshold | N. Chem | r2 |
| Actives in tr | 0.01 | 196 | 0.57 |
| Actives in tr | 0.1 | 83 | 0.75 |
| Actives in tr & eval | 0.01 | 140 | 0.74 |
| Actives in tr & eval | 0.1 | 72 | 0.83 |
| Actives in tr (VW to N) | 0.01 | 196 | 0.61 |
| Actives in tr (VW to N) | 0.1 | 83 | 0.76 |
| Actives in tr & eval (VW to N) | 0.01 | 118 | 0.71 |
| Actives in tr & eval (VW to N) | 0.1 | 67 | 0.82 |

The results are in concordance with those from the previous table with a considerable increase in correlation.

It can concluded that the qualitative predictions from the submitted models can be evaluated using the AC50 values of the evaluation set.

The correlation between AC50eq of the evaluation set from the literature and the converted AC50 values from AUC scores of the training set gives similar results.

The correlation of the 72 active compounds in both training and evaluation set is r2 =0.83.

R2=0.68

RMSE=0.74

corr_logAC50_actives.tif