**SUPPLEMENTARY DATA**

**S1: Training sets and external test sets description**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Training | | | External test set | | |
| Models | Total  compound | active | inactive | Total  compound | active | inactive |
| Organ toxicity | 850 |  |  | 95 |  |  |
| Dili | 178 | 672 | 20 | 75 |
| Toxicity endpoints | 6156 |  |  | 685 |  |  |
| Mutagenicity | 3218 | 2938 | 346 | 339 |
| Carcinogenicity | 1391 | 711 | 680 | 155 | 82 | 73 |
| Cytotoxicity | 5487 | 1765 | 3722 | 610 | 205 | 405 |
| Immunotoxicity | 41883 | 11017 | 30166 | 93 | 14 | 79 |
| Toxicological pathways | 6901 |  |  | 610 |  |  |
| nr-ahr | 769 | 6132 | 73 | 537 |
| nr-ar | 7662 | 262 | 7400 | 586 | 12 | 574 |
| nr-ar-lbd | 7118 | 222 | 6896 | 582 | 8 | 574 |
| nr-aromatase | 6113 | 298 | 5815 | 528 | 39 | 489 |
| nr-er | 6414 | 680 | 5734 | 516 | 51 | 465 |
| nr-er-lbd | 6801 | 346 | 6455 | 600 | 20 | 580 |
| nr-ppar-gamma | 6820 | 190 | 6630 | 605 | 31 | 574 |
| sr-are | 6108 | 963 | 5145 | 555 | 93 | 462 |
| sr-hse | 7328 | 308 | 7019 | 610 | 23 | 588 |
| sr-mmp | 6106 | 937 | 5169 | 542 | 60 | 482 |
| sr-p53 | 7133 | 434 | 6699 | 616 | 41 | 575 |
| sr-atad5 | 7457 | 268 | 7189 | 621 | 37 | 584 |

**S2: Comparison with other models reported in literature**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **Models** | **(Pro) accuracy (%)** | **(Pro) AUC-ROC** | **Compared Models** | **Accuracy (%)** | **AUC-ROC** |
| **Organ toxicity** | 82.00 | 0.86 | Hong, et.al (2017) (1) | 71.30 | - |
| Dili |
| **Toxicity endpoints** | 84.00 | 0.90 | Xu , et.al (2012) (2) | 84.00 | 0.90 |
| Mutagenicity |
| Carcinogenicity | 81.24 | 0.85 | Zhang, et.al (2017)(3) | 71.00 | 0.77 |
| Cytotoxicity | 85.00 | 0.89 | Svensson, et.al  (2017) (4) | 80.00 | - |
| Immunotoxicity | 74.00 | 0.75 | Schrey, et.al (2017) (5) | 74.00 | 0.75 |
| **Toxicological pathways** | 91.00 | 0.90 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 85.00 | 0.92 |
| nr-ahr |
| nr-ar | 86.00 | 0.73 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 73.00 | 0.82 |
| nr-ar-lbd | 83.00 | 0.75 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 65.00 | 0.87 |
| nr-aromatase | 89.00 | 0.75 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 73.00 | 0.83 |
| nr-er | 91.00 | 0.79 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 74.00 | 0.80 |
| nr-er-lbd | 89.00 | 0.80 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 71.00 | 0.82 |
| nr-ppar-gamma | 85.00 | 0.84 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 78.00 | 0.86 |
| sr-are | 87.00 | 0.79 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 72.00 | 0.83 |
| sr-hse | 86.00 | 0.87 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 79.00 | 0.86 |
| sr-mmp | 91.00 | 0.92 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 90.00 | 0.95 |
| sr-p53 | 89.00 | 0.87 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 76.00 | 0.87 |
| sr-atad5 | 84.00 | 0.80 | [Tox21 Leaderboard](https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) | 74.00 | 82.00 |

\*(Pro) = ProTox-II methods

SFigure1.tif

S3: Application case: Etonogestrel is considered as input structure, which is predicted using different methodologies under five different classification schemes with confidence scores and overall toxicity radar plot.

**S4. Performance analysis of small set with known activity**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Database** | **ID** | **endpoints** | **actual** | **predicted** | **confidence** |
| Drugbank | DB06817 | DILI | active | active | 0.65 |
| Drugbank | DB00573 | DILI | active | active | 0.72 |
| Drugbank | DB04743 | DILI | active | active | 0.77 |
| Drugbank | DB00549 | DILI | active | active | 0.83 |
| Drugbank | DB00675 | DILI | active | active | 0.73 |
| PubChem | 105111 | Immuno | active | active | 0.93 |
| PubChem | 40024 | Immuno | active | active | 0.82 |
| PubChem | 446541 | Immuno | active | active | 0.80 |
| PubChem | 9554 | Immuno | active | inactive | 0.99 |
| PubChem | 16682746 | Immuno | active | inactive | 0.93 |
| PubChem | 177 | Carcino | active | active | 0.87 |
| PubChem | 8266 | Carcino | active | active | 0.77 |
| PubChem | 2796 | Carcino | active | active | 0.79 |
| PubChem | 5281735 | Carcino | active | active | 0.92 |
| PubChem | 5770 | Carcino | active | active | 0.79 |
| PubChem | 163754 | Muta | active | active | 1.00 |
| PubChem | 51323 | Muta | active | active | 0.82 |
| PubChem | 151976 | Muta | active | active | 0.91 |
| PubChem | 2381519 | Muta | active | active | 0.91 |
| PubChem | 5281849 | Muta | active | active | 0.82 |
| PubChem | 71699310 | Cyto | active | active | 0.99 |
| PubChem | 532322742 | Cyto | active | active | 0.70 |
| PubChem | 53248909 | Cyto | active | active | 0.90 |
| PubChem | 71699310 | Cyto | active | active | 0.99 |
| PubChem | 51003445 | Cyto | active | active | 0.85 |
| PubChem | 170465670 | NR-AhR | active | active | 0.99 |
| PubChem | 170465712 | NR-AhR | active | active | 0.98 |
| PubChem | 5353853 | NR-AhR | active | active | 1.00 |
| PubChem | 5154 | NR-AhR | active | active | 1.00 |
| PubChem | 15789 | NR-AhR | active | active | 0.99 |
| PubChem | 28417 | NR-AR | active | active | 1.00 |
| PubChem | 42725 | NR-AR | active | active | 1.00 |
| PubChem | 54339 | NR-AR | active | active | 1.00 |
| PubChem | 32798 | NR-AR | active | active | 0.99 |
| PubChem | 5757 | NR-AR | active | active | 1.00 |
| PubChem | 7509 | NR-ER | active | active | 0.99 |
| PubChem | 1550489 | NR-ER | active | active | 1.00 |
| PubChem | 4807 | NR-ER | active | active | 1.00 |
| PubChem | 541 | NR-ER | active | active | 0.86 |
| PubChem | 11957723 | NR-ER | active | active | 1.00 |
| PubChem | 7188 | NR-ER-LBD | active | active | 1.00 |
| PubChem | 73864 | NR-ER-LBD | active | active | 1.00 |
| PubChem | 10229 | NR-ER-LBD | active | active | 0.96 |
| PubChem | 77328 | NR-ER-LBD | active | active | 1.00 |
| PubChem | 697993 | PPAR-Gamma | active | active | 1.00 |
| PubChem | 16759592 | PPAR-Gamma | active | active | 1.00 |
| PubChem | 95717 | PPAR-Gamma | active | active | 1.00 |
| PubChem | 162268 | PPAR-Gamma | active | active | 1.00 |
| PubChem | 9074 | PPAR-Gamma | active | active | 1.00 |
| PubChem | 28768 | Aromatase | active | active | 1.00 |
| PubChem | 44073 | Aromatase | active | active | 1.00 |
| PubChem | 70038 | Aromatase | active | active | 1.00 |
| PubChem | 12472902 | Aromatase | active | active | 0.87 |
| PubChem | 119182 | Aromatase | active | active | 1.00 |
| PubChem | 439501 | NR-AR-LBD | active | active | 1.00 |
| PubChem | 441207 | NR-AR-LBD | active | active | 1.00 |
| PubChem | 5281034 | NR-AR-LBD | active | active | 1.00 |
| PubChem | 6279 | NR-AR-LBD | active | active | 0.88 |
| PubChem | 3033968 | NR-AR-LBD | active | active | 0.96 |
| PubChem | 697993 | SR-ARE | active | active | 1.00 |
| PubChem | 22463 | SR-ARE | active | active | 1.00 |
| PubChem | 91771 | SR-ARE | active | active | 1.00 |
| PubChem | 16682736 | SR-ARE | active | active | 1.00 |
| PubChem | 17142 | SR-ARE | active | active | 1.00 |
| PubChem | 697993 | SR-HSE | active | active | 1.00 |
| PubChem | 16741 | SR-HSE | active | active | 1.00 |
| PubChem | 70414 | SR-HSE | active | active | 1.00 |
| PubChem | 4234241 | SR-HSE | active | active | 1.00 |
| PubChem | 102724 | SR-HSE | active | active | 1.00 |
| PubChem | 697993 | SR-MMP | active | active | 1.00 |
| PubChem | 6537489 | SR-MMP | active | active | 0.99 |
| PubChem | 727200 | SR-MMP | active | active | 1.00 |
| PubChem | 7737 | SR-MMP | active | active | 0.98 |
| PubChem | 521106 | SR-MMP | active | active | 1.00 |
| PubChem | 70414 | ATAD5 | active | active | 1.00 |
| PubChem | 10021362 | ATAD5 | active | active | 1.00 |
| PubChem | 5354198 | ATAD5 | active | active | 1.00 |
| PubChem | 16759592 | ATAD5 | active | active | 0.97 |
| PubChem | 93004 | ATAD5 | active | active | 1.00 |
| PubChem | 6537489 | p53 | active | active | 1.00 |
| PubChem | 3108 | p53 | active | inactive | 0.87 |
| PubChem | 76716 | p53 | active | active | 0.96 |
| PubChem | 73864 | p53 | active | active | 1.00 |
| PubChem | 676166 | p53 | active | active | 0.98 |

**Oversampling method**

A selective oversampling of minority class is introduced in the construction of the models. For each of the prediction end-points, the active (positive) and inactive (negative) data are fragmented using RECAP (6) and ROTBONDS fragmentation methods (7). The propensity score (PS) for each of the uniquely occurring fragments in both the sets is computed (8). Only those molecules having the highest propensity scores for fragments conserved for the active class are randomly chosen to be duplicated and added to the original data set. (This in turn reduces the variance). Both steps are repeated until the minority class consists of as many samples as the majority class for all the models.

**Identification of frequent features in active and inactive compounds**

To analyse the important and frequent features in active and inactive compounds. The percentage of occurrences of each feature from Morgan fingerprint (2,048 bits) in active and inactive compounds was calculated. The relative frequency of important features for a class (e.g., active) were calculated taking not only the feature position and occurrence within the active class into account but also the relative feature frequency of that feature in the inactive class and vice versa. The average relative frequency for each class were calculated, a feature was only considered important for a class, if it’s presence in one class is higher than the average relative frequency of that class as well as lower than the average relative frequency of the other class. The top features for each class were calculated using class-specific weighted bits/feature patterns in the fingerprints. The Bayesian based Feature detection applied in the study, calculates the probability of any compound containing a feature (F) from the Morgan -feature space belongs to a specific class (e.g., active or inactive), given that total number of the compounds containing the feature (F) and the number of compounds of feature (F) belong to that class. The dissimilarity (uncommon feature score between two classes) between the two features is calculated by the 1- Pearson correlation coefficient of their individual class specific scores. The distribution of top 10 most occurring features in respective classes and their relative frequency in each class are shown for each model, under the model description section of the webserver. This work has been reported in our recently accepted manuscript (Banerjee P and Preissner R (2018) BitterSweetForest: A Random Forest Based Binary Classifier to Predict Bitterness and Sweetness of Chemical Compounds. Front. Chem. 6:93. doi: 10.3389/fchem.2018.00093).

**Cross-validation**

The 10-fold cross-validation for the all the models were performed using fragment- based similarity of compounds. The fragment propensities were calculated for both active and inactive class, as continuous real-valued numbers in the range between 0= low and 1= high. The compounds were thus group in to 10 parts based on the fragment propensities. Thus, the group of compounds sharing the fragments propensities were distributed across the folds. The compounds with unassigned fragment propensity were then randomly assigned across the fold. The compounds assignment to the different folds was done ensuring fragment similarity of compounds and similar ratio of actives to inactives in all the folds, including training and test set.

References:

1. Hong,H., Thakkar,S., Chen,M. and Tong,W. (2017) Development of Decision Forest Models for Prediction of Drug-Induced Liver Injury in Humans Using A Large Set of FDA-approved Drugs. *Scientific Reports*, **7**, 17311.

2. Tang,Y. (2012) In silico Prediction of Chemical Ames Mutagenicity. 10.1021/ci300400a.

3. Zhang,L., Ai,H., Chen,W., Yin,Z., Hu,H., Zhu,J., Zhao,J., Zhao,Q. and Liu,H. (2017) CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. *Scientific Reports*, **7**, 1–14.

4. Svensson,F., Norinder,U. and Bender,A. (2017) Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. *Toxicol. Res.*, **6**, 73–80.

5. Schrey,A.K., Nickel-Seeber,J., Drwal,M.N., Zwicker,P., Schultze,N., Haertel,B. and Preissner,R. (2017) Computational prediction of immune cell cytotoxicity. *Food and Chemical Toxicology*, **107**, 150–166.

6. Lewell XQ, Judd DB, Watson SP,H.M. (1998) RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci.*, **38(3)**, 511–22.

7. Ahmed,J., Worth,C.L., Thaben,P., Matzig,C., Blasse,C., Dunkel,M. and Preissner,R. (2011) FragmentStore--a comprehensive database of fragments linking metabolites, toxic molecules and drugs. *Nucleic acids research*, **39**, D1049–54.

8. Drwal,M.N., Banerjee,P., Dunkel,M., Wettig,M.R. and Preissner,R. (2014) ProTox: a web server for the in silico prediction of rodent oral toxicity. *Nucleic acids research*, 10.1093/nar/gku401.