Machine Learning for Acute Oral System Toxicity Regression and Classification

Conor Parks[1], Zied Gaieb[1], Rommie E. Amaro[1*]

[1]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093

Abstract

In vivo toxicity testing remains a costly and time-consuming component of any pre-clinical drug development campaign. In particular, LD50 measurements require the loss of animal life but remain a critical component in preventing lethal compounds from entering the clinic. With advances in machine learning, *in silico* LD50 prediction now has the potential to greatly reduce this burden. We study various types of machine learning models to predict acute oral LD50 measurements in rats as regression and classification problems. We demonstrate that transfer learning a ResNet34 model pretrained on ImageNet with test time augmentation generates the best performing regression model and that random forest augmented with conformal prediction provides a robust methodology to perform classification.

Introduction

Chemical compounds must pass a battery of *in vitro* and *in vivo* tests to assess potentially lethal or adverse effects prior to human administration. While essential, these tests are not only time consuming and costly, but also require the loss of animal life. As such, LD50 measurements remain a particularly difficult end point in the preclinical drug development pipeline. As part of the "Toxicity Testing in the Twenty First Century" initiative[1,2], the development of accurate and fast in silico models are increasingly being looked towards to relieve this burden[3]. The machine learning field in particular was emboldened when deep neural networks recently demonstrated state of the art predictive capability on 12 *in vitro* toxicology assay endpoints in the blinded Tox21 challenge[4]. However, the ability to predict LD50 measurements has remained comparatively understudied to date due to the lack of quality curated datasets for machine learning training. Prior works suggest that machine learning is suitable for LD50 prediction[5–10], but they have largely used data sets curated by academics or have used random splits which are known to inflate model performance metrics[11].

Deep convolutional neural networks (ConvNets) have revolutionized the field of image analysis, with many applications to cell based imaging arising in drug discovery[12,13]. One of the main strengths of ConvNets is the ability to perform feature extraction in a data driven fashion directly from image data[14,15]. However, this method is sparingly used in the field of cheminformatics for quantitative structure activity relationship (QSAR) prediction, in part due to data quantity limitations. Rather, model types such as gradient boost, random forest, support vector machine, k-nearest neighbor, and fully connected neural networks remain the prominent models of choice for QSAR applications[16,17]. To date, ConvNet studies have employed either semi supervised pretraining strategies or trained directly from scratch on the data set at hand. For example, Chemception[18] was trained from scratch and showed comparable performance to feed forward networks (FFN) on various QSAR benchmark data sets, such as free energy of solvation, inhibition of HIV replication, and Tox21. The same authors then developed Chemnet[19], a semi-supervised pretraining approach to reproduce a suite of directly computable chemical properties where the pre-trained model was subsequently trained via transfer learning on various QSAR benchmark datasets. With no pre training at all, Toxic Colors demonstrated the utility of ConvNets on the Tox 21 dataset[20]. KekuleScope[21] was the first application of ConvNets pretrained on ImageNet for QSAR applications via transfer learning[21]. KekuleScope[21] transfer learned AlexNet[22], DenseNet-201[23], ResNet152[24] and VGG-19[25] models pretrained

on ImageNet for the purpose of predicting compound-cell line pIC50 measurements. It was demonstrated that transfer learned ConvNets obtained competitive performance to state of the art random forest models for predicting cell line sensitivity directly from compound images. However, the impact of test time augmentation (TTA), where data augmentation is applied to items in the validation set, and the final prediction being the average over the augmented items, was not explored. In addition, Meyer et al. have recently compared RF models and ConvNets to predict drug classification and again show the competitive performance of ConvNet models for drug discovery applications[26].

Herein, we describe machine learning results for LD50 regression and classification using the high quality, recently released, EPA LD50 dataset. In particular, we compare the performance of random forest (RF), gradient boost (GB), fully connected neural networks (NN) models, and a ResNet34 ConvNet model transfer learned on the training data. In total, we find that the ResNet34 model with TTA achieves superior performance to all other model types for regression, but an ensemble model averaging the RF and ResNet34 predictions yields the highest predictive accuracy. This demonstrates the utility of transfer learning pretrained ConvNet models to low data set size scenarios typical in drug discovery. Finally, RF models with conformal prediction proves to be a promising methodology for classification, potentially allowing the prioritization of compound based off a toxin/non-toxic classification.

Materials and Methods.

Data Set Source

As part of its push to replace *in vivo* animal studies with predictive models for acute oral systemic toxicity determination, the EPA recently released a dataset containing measured LD50 values obtained from literature data provided by the Dow Chemical Company, REACH data from eChemportal, HSDB (Hazardous Substances Data Bank), RTECS data from Leadscope, and the training set used by TEST (Toxicity Estimation Software Tool) (https://ntp.niehs.nih.gov/go/tox-models). The dataset provides 11,854 data points classified into either toxic or non-toxic categories, along with 8891 annotated LD50 values.

Data Set Preparation

Molecule SMILES strings were first standardized and canonicalized using the charge parent function in MolVS[27]. Only molecules with molecular weights in the range of 75 to 800 Da were retained. Molecules without an explicit LD50 measurement were dropped from further analysis. For tabular model learning (RF, GB, and FFN), SMILES strings were featurized using 4096 bit length ECFP6 fingerprints, molecular weight, topological surface area, number of hydrogen donors, number of hydrogen acceptors, LogP, heavy atom count, number of rotatable bonds, and ring count with RDKit[28]. Non ECFP6 bit values were scaled using the standard scaler function in Scikit-learn[29]. For the ResNet34 model, SMILES strings were converted to png files using RDKit using the procedure outlined by Cortes-Ciriano et al[21]. Molecules were split into training and validation sets following a Murcko scaffold split. Finally, the LD50 target values were log transformed, and scaled using the robust scaler in RDKit[28].

Machine Learning Model Training

Random Forest models were trained using the Scikit-learn[29] library via a grid search hyperparameter optimization strategy. The following hyperparameter values were explored: 250,500,750, and 1000 for the number of estimators; sqrt, log2, 0.3, 0.5, and 1.0 for max features; 1,3,5,10,25 for min samples leaf; and 'balanced' and 'None' for class_weight. Gradient boost models were trained using the XGBoost library via Bayesian optimization with the following possible ranges for hyperparameters: 1e-6 to 1 For the learning

rate; 0 to 5 for gamma; 0 to 1 for colsample by tree; 1 to 16 for max depth; number of estimators either 50, 100, 500 or 1000; and 1 to 10 for the min child weight. The ResNet34 model was trained using methods in Fast.ai. Specifically, we transfer learned a ResNet34 model pretrained on ImageNet directly on molecular images. The Fast.ai default image augmentation methods were used throughout, with the addition of vertical flipping and rotation (20.0 degrees). During test time on the validation set, TTA was also explored to see the impact on performance metrics. Here, test time data (validation set) are augmented and the predictions averaged to yield the final prediction. To train the ResNet34, the head of the network, i.e the fully connected layers appended to the ResNet34 architecture, was trained for 200 epochs with the upper layers frozen. The learning rate was determined via the lr_find function in Fast.ai which yielded 6e-3 to be a suitable learning rate for the fully connected layers during this stage. Specifically, this was determined by choosing a learning rate approximately 10 times smaller than the learning rate at which the loss obtains a minimum from the lr_find function[30]. After this initial head training, all layers of the network were unfrozen, and the optimal learning rate for the higher layers was determined again using the lr_find function again resulting in a lr of 5e-5. The full model was then trained for 1000 epochs using a discriminative learning rate across the layers, varying from 1e-5 at the initial layers to 0.006/5 at the lower layers. During all phases of training the ConvNet, a cyclical learning rate was employed using the fit_one_cycle function in Fast.ai[30]. The mean squared error (MSE) loss was used as the optimization function. FFN models were trained using Fast.ai as well. Here, ECFP6 bit vectors were treated as categorical variables whose embeddings were optimized via backpropagation during model training. A 2-layer model was employed with 1000 nodes in the first layer and 500 nodes in the second layer. The weight decay was set to 0.2, and dropout of 0.001 and 0.01 was used in the first and second layer respectively. All other Fast.ai tabular model defaults were used. Again, the FFN model was trained with the fit_one_cycle[30] method with a max_lr value of 0.0005. All models were then analyzed using MSE, Pearson correlation coefficient, and Kendall's Tau ranking. The ResNet34 and FFN tabular models were trained using fp16 precision.

Conformal Prediction

Conformal prediction defines a machine learning model's applicability domain, i.e the region of chemical space where predictions are reliably accurate[31–37]. Conformal prediction produces a confidence region where the true value lies with a probability determined by a user specified confidence threshold. This stands in contrast to traditional machine learning models that produce point predictions, giving the user no sense of what degree of confidence should be placed on the prediction made. A conformal predictor is considered valid if for a chosen confidence level 1-ε, the number of prediction errors made does not exceed ε. For classification problems, the conformal predictor will assign a class label to a new molecule if the new molecule is similar enough to prediction outcomes made on the calibration set. Here, similarity is determined by a non-conformity measure which measures similarity to previously seen data. If the non-conformity measure is greater than 1-ε , a class label is assigned. This is done for all classes in the calibration set. Non-conformity can be quantified using auxiliary models, or directly from the model itself[38]. For binary classification, a new molecule could be labeled either, neither, or both of the two classes. We use conformal prediction to gather model statistics for molecules with only one class label, hence defining the applicability domain for a given ε. Mondrian conformal prediction (MCP) was subsequently developed to deal with data sets with class imbalance, a frequent issue in cheminformatics[33]. A MCP model sets a significance level for each individual class, hence guaranteeing their respective validity[33]. All conformal prediction code was generated using the nonconformist python library[39]. For a full thorough discussion of the theory and application of conformal prediction in QSAR, we direct the reader to a recent review of the material[37].

Results

Regression

For each of the models studied herein, we present the resulting performance metrics on the validation set in Table 1:

Table 1: Regression performance metrics on the validation set for all model types

Model type	MSE (e-2)	Pearson correlation	Kendall's Tau
ResNet34	5.27	0.462	0.308
ResNet 34 (TTA)	5.04	0.481	0.313
Random Forest	5.16	0.460	0.283
FFN	5.55	0.420	0.246
Gradient Boost	5.6	0.393	0.218
Ensemble	4.86	0.512	0.333

In total, LD50 regression on this data set proved to be a challenging problem, with the best models achieving only 0.512 Pearson correlation between predictions and the scaled LD50. This suggests that further improvement in model quality is still needed which may come from training on larger in house pharma ADME data sets with more diverse scaffolds or the inclusion of biological information (transcriptomic etc.). In total, the ResNet34 model with TTA is the best performing individual model, outperforming all others in all metrics. TTA of the ResNet34 model led to an approximately 5% performance increase on the validation set. Both the RF and ResNet34 models outperform the gradient boost and FFN models by 10-20% in the Pearson metric. In all, the performance of the ResNet34 demonstrates their ability to learn mappings from color coded pixels and target LD50 values in a data driven fashion. This work further demonstrates the utility of transfer learning and data augmentation to generate state of the art models with the limited data scenarios present in drug discovery.

The correlation between the TTA ConvNet and RF model residuals is high (r2=0.84). Despite the observed correlation, an improvement in performance is still observed with model ensemble averaging. Here, the predictions from the TTA ConvNet and RF are averaged. As reported in Table 1, this resulted in roughly a 10% boost in performance across all metrics, similar to the performance boost seen in the work of Cortés-Ciriano et al[21]. The RF and TTA ConvNet model distribution of residuals displays long tails, conveying that only a few compounds in the validation set contribute extensively to the reported MSE in Table. In fact, the median squared error across the full validation set is only 0.014 for the TTA ConvNet model, which is a 72% reduction relative to the value reported in Table 1. The Kendall's Tau ranking of the target values in the validation set by the residuals is 0.74 for the RF model and 0.68 for the TTA ResNet34, respectively, further demonstrating the non-random distribution of residuals. The long tailed residual distribution is a result of the skewed distribution of LD50 values in the original training set, with a few compounds possessing exceedingly large LD50 values relative to the rest of the distribution. These compounds with large LD50 values contribute most heavily to the overall MSE. The correlation plot of TTA ConvNet and RF residuals, distribution of residuals for the RF and TTA ConvNet individual models, and the distribution of scaled LD50 values in the training set can be found in Figure 1.

Analysis of the RF feature importance demonstrates that a small set of features are of immense importance, as shown in figure 2. Here, 5/10 physical property features rank in the top 10 features by importance including MolLogP, MolWt, HeavyAtomCount, TPSA, NumRotatableBonds. A csv file containing the feature importance of each feature used herein is provided in Supplementary Table 1. These physical properties correlate with promiscuity, a known risk factor for toxicity.

Classification

Encouraged by the performance and ease of training of the RF model for regression, we elected to only optimize RF models for classification. We optimized the RF parameters to obtain the maximum ROC-AUC score on the validation set. This was achieved with 250 estimators**,** 0.3 for max_features, and a balanced class weight**.** the RF model displays quality performance on the full validation set with the following performance metrics: 0.704 ROC-AUC; 0.292 MCC; and 0.476 F1-score. We next trained a MCP model with an RF model as the base estimator with the optimal set of hyperparameters. Table 2 and Figure 3 show the MCP model results as a function of the confidence threshold and the MCP validities respectively. We define the local/total recall to be the recall of compounds in the applicability domain for a given confidence threshold, or the full data set, respectively. All statistics for the MCP model can be found in Supplementary Table 1.

Table 2: Classification performance classification metrics on validation set for varying MCP thresholds

Model (threshold)	**Efficiency**	**Total recall**	**Local Recall**	**ROC-AUC**	**MCC**	**F1-score**
MCP (0.05)	0.13	0.20	0.94	0.83	0.54	0.81
MCP (0.1)	0.35	0.39	0.90	0.77	0.41	0.68
MCP(0.15)	0.59	0.52	0.81	0.73	0.33	0.62

In total, the MCP performance metrics increase monotonically with decreasing confidence threshold, obtaining a final ROC-AUC of 0.83 at a 0.05 confidence level. At a 0.1 confidence level, we are able to obtain a 0.77 ROC-AUC while retaining 35% of the validation set for predictions. Most importantly, we recall 39% of all actives from the full validation set. Within the applicability domain, the recall is 90% of all positive molecules. These classification statistics show that the MCP model could assist in selecting predicted non-toxic molecules at high accuracy for advancement in the clinical pipeline in a data driven manner. This has the potential to reduce the financial and animal loss of life burdens during preclinical LD50 measurement testing.

Conclusion

Acute oral toxicity testing remains a time consuming, cost intensive portion of preclinical drug discovery. Herein, we analyze the performance of machine learning models for both rat oral LD50 regression and classification. In particular, we demonstrate that ResNet34 models pretrained on ImageNet and RF models are the best models for this data set and split. TTA and ensemble averaging of the ConvNet and RF models led to a 10% increase in the Pearson correlation coefficient for LD50 regression. For regression, the work herein demonstrates that ConvNets are able to determine relevant chemical features in a data driven fashion that are complementary to those in the traditional ECFP6 fingerprint. However, regression remains a challenging problem with the highest Pearson correlation coefficient obtained on this data set with a Murcko split only being 0.51. For classification, we demonstrated the utility of MCP with a RF base estimator for LD50 classification into toxic/non-toxic classes. Specifically, we demonstrated that model performance can be calibrated through suitable tuning of the MCP confidence threshold.

Disclosures

REA has equity interest in and is a co-founder and scientific advisor of Actavalon, Inc.

References

(1) Andersen, M. E.; Krewski, D. Toxicity Testing in the 21st Century: Bringing the Vision to Life. *Toxicol Sci* 2009, *107* (2), 324–330.

(2) Toxicity Testing in the 21st Century: A Vision and a Strategy: Journal of Toxicology and Environmental Health, Part B: Vol 13, No 2-4 https://www.tandfonline.com/doi/full/10.1080/10937404.2010.483176 (accessed Jun 14, 2019).

(3) Rusyn, I.; Daston, G. P. Computational Toxicology: Realizing the Promise of the Toxicity Testing in the 21st Century. *Environmental Health Perspectives* 2010, *118* (8), 1047–1050.

(4) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* 2016, *3*.

(5) Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative Structure−Activity Relationship Modeling of Rat Acute Toxicity by Oral Exposure. *Chem. Res. Toxicol.* 2009, *22* (12), 1913–1921.

(6) Lu, J.; Peng, J.; Wang, J.; Shen, Q.; Bi, Y.; Gong, L.; Zheng, M.; Luo, X.; Zhu, W.; Jiang, H.; et al. Estimation of Acute Oral Toxicity in Rat Using Local Lazy Learning. *J Cheminform* 2014, *6* (1), 26.

(7) Lei, T.; Li, Y.; Song, Y.; Li, D.; Sun, H.; Hou, T. ADMET Evaluation in Drug Discovery: 15. Accurate Prediction of Rat Oral Acute Toxicity Using Relevance Vector Machine and Consensus Modeling. *J Cheminform* 2016, *8* (1), 6.

(8) Hamadache, M.; Benkortbi, O.; Hanini, S.; Amrane, A.; Khaouane, L.; Si Moussa, C. A Quantitative Structure Activity Relationship for Acute Oral Toxicity of Pesticides on Rats: Validation, Domain of Application and Prediction. *Journal of Hazardous Materials* 2016, *303*, 28–40.

(9) Li, X.; Chen, L.; Cheng, F.; Wu, Z.; Bian, H.; Xu, C.; Li, W.; Liu, G.; Shen, X.; Tang, Y. In Silico Prediction of Chemical Acute Oral Toxicity Using Multi-Classification Methods. *J. Chem. Inf. Model.* 2014, *54* (4), 1061–1069.

(10) Xu, Y.; Pei, J.; Lai, L. Deep Learning Based Regression and Multiclass Models for Acute Oral Toxicity Prediction with Automatic Chemical Feature Extraction. *J. Chem. Inf. Model.* 2017, *57* (11), 2672–2685.

(11) Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* 2018, *58* (5), 916–932.

(12) Simm, J.; Klambauer, G.; Arany, A.; Steijaert, M.; Wegner, J. K.; Gustin, E.; Chupakhin, V.; Chong, Y. T.; Vialard, J.; Buijnsters, P.; et al. Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. *Cell Chemical Biology* 2018, *25* (5), 611-618.e3.

(13) Hofmarcher, M.; Rumetshofer, E.; Clevert, D.-A.; Hochreiter, S.; Klambauer, G. Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. *J. Chem. Inf. Model.* 2019, *59* (3), 1163–1171.

(14) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J Comput Aided Mol Des* 2016, *30* (8), 595–608.

(15) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc., 2015; pp 2224–2232.

(16) Lavecchia, A. Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discovery Today* 2015, *20* (3), 318–331.

(17) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*; O'Reilly Media, Inc., 2019.

(18) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O.; Baker, N. Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-Developed QSAR/QSPR Models. *arXiv:1706.06689 [cs, stat]* 2017.
(19) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O. Using Rule-Based Labels for Weak Supervised Learning: A ChemNet for Transferable Chemical Property Prediction. *arXiv:1712.02734 [cs, stat]* 2017.
(20) Fernandez, M.; Ban, F.; Woo, G.; Hsing, M.; Yamazaki, T.; LeBlanc, E.; Rennie, P. S.; Welch, W. J.; Cherkasov, A. Toxic Colors: The Use of Deep Learning for Predicting Toxicity of Compounds Merely from Their Graphic Images. *Journal of Chemical Information and Modeling* **2018**, *58* (8), 1533–1543.
(21) Ciriano, I. C.; Bender, A. KekuleScope: Improved Prediction of Cancer Cell Line Sensitivity Using Convolutional Neural Networks Trained on Compound Images. *arXiv:1811.09036 [cs]* 2018.
(22) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*; Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q., Eds.; Curran Associates, Inc., 2012; pp 1097–1105.
(23) Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K. Q. Densely Connected Convolutional Networks; 2017; pp 4700–4708.
(24) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Las Vegas, NV, USA, 2016; pp 770–778.
(25) Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* 2014.
(26) Meyer, J. G.; Liu, S.; Miller, I. J.; Gitter, A.; Coon, J. J. Learning Molecule Drug Function from Structure Representations with Deep Neural Networks or Random Forests. *bioRxiv* 2018, 482877.
(27) MolVS: Molecule Validation and Standardization — MolVS 0.1.1 documentation https://molvs.readthedocs.io/en/latest/ (accessed Aug 13, 2019).
(28) RDKit https://www.rdkit.org/ (accessed Aug 13, 2019).
(29) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON* 6.
(30) Smith, L. N. A Disciplined Approach to Neural Network Hyper-Parameters: Part 1 -- Learning Rate, Batch Size, Momentum, and Weight Decay. *arXiv:1803.09820 [cs, stat]* 2018.
(31) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *Journal of Chemical Information and Modeling* 2014, *54* (6), 1596–1603.
(32) Giblin, K. A.; Hughes, S. J.; Boyd, H.; Hansson, P.; Bender, A. Prospectively Validated Proteochemometric Models for the Prediction of Small-Molecule Binding to Bromodomain Proteins. *J. Chem. Inf. Model.* 2018, *58* (9), 1870–1888.
(33) Sun, J.; Carlsson, L.; Ahlberg, E.; Norinder, U.; Engkvist, O.; Chen, H. Applying Mondrian Cross-Conformal Prediction To Estimate Prediction Confidence on Large Imbalanced Bioactivity Data Sets. *Journal of Chemical Information and Modeling* 2017, *57* (7), 1591–1598.
(34) Shafer, G.; Vovk, V. A Tutorial on Conformal Prediction. *J. Mach. Learn. Res.* 2008, *9*, 371–421.
(35) Svensson, F.; Norinder, U.; Bender, A. Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction. *Journal of Chemical Information and Modeling* 2017, *57* (3), 439–444.
(36) Cortés-Ciriano, I.; Bender, A.; Malliavin, T. Prediction of PARP Inhibition with Proteochemometric Modelling and Conformal Prediction. *Molecular Informatics* 2015, *34* (6–7), 357–366.

(37) Cortés-Ciriano, I.; Bender, A. Concepts and Applications of Conformal Prediction in Computational Drug Discovery. *arXiv:1908.03569 [cs, q-bio]* 2019.
(38) Svensson, F.; Aniceto, N.; Norinder, U.; Cortes-Ciriano, I.; Spjuth, O.; Carlsson, L.; Bender, A. Conformal Regression for Quantitative Structure–Activity Relationship Modeling—Quantifying Prediction Uncertainty. *J. Chem. Inf. Model.* 2018, *58* (5), 1132–1140.
(39) Linusson, H. *Python Implementation of the Conformal Prediction Framework.: Donlnz/Nonconformist*; 2019.

Figure 1: A) residual correlation B) TTA ConvNet residual distribution C) RF residual distribution D) Scaled LD50 distribution in training set

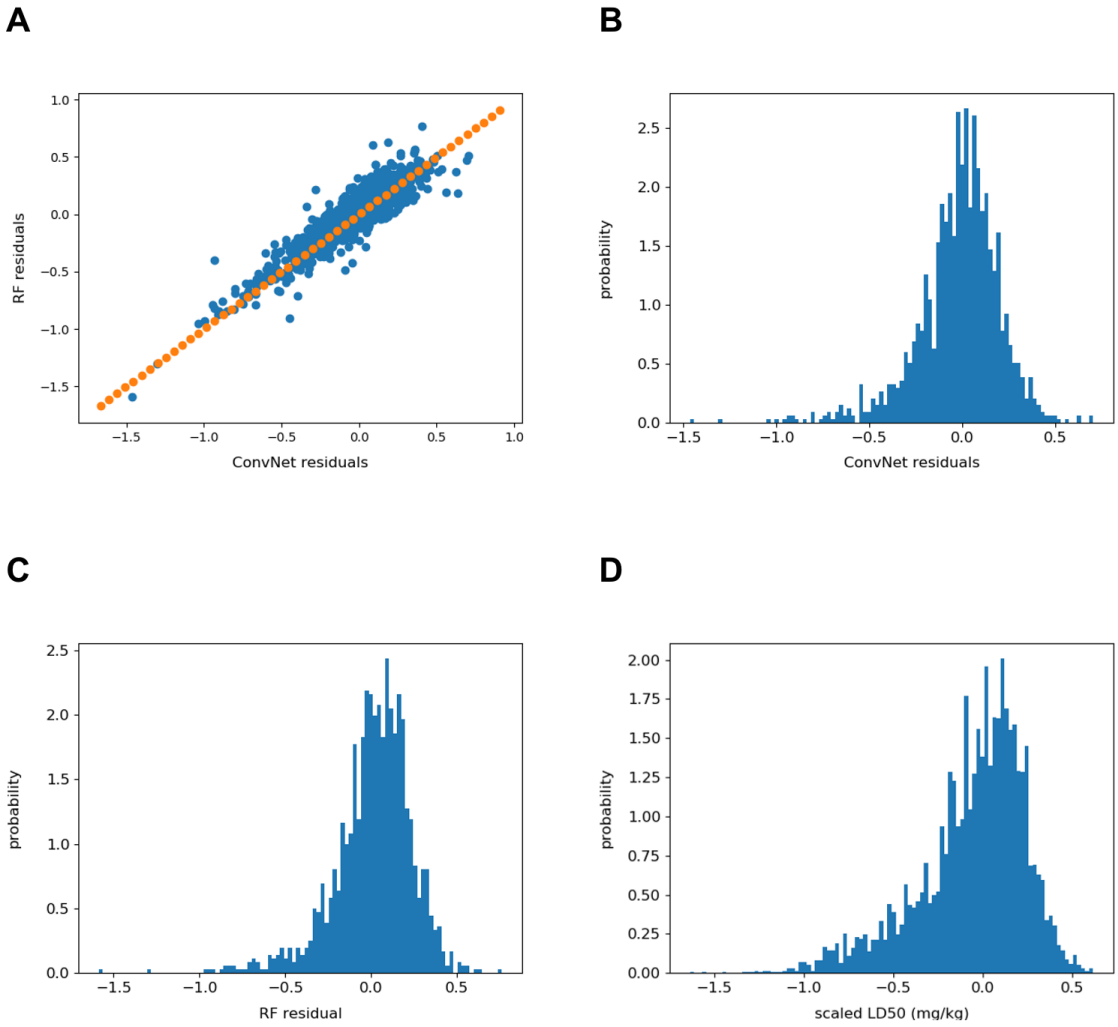

Figure 2: A) feature rank vs. feature importance as determined by the RF regression model

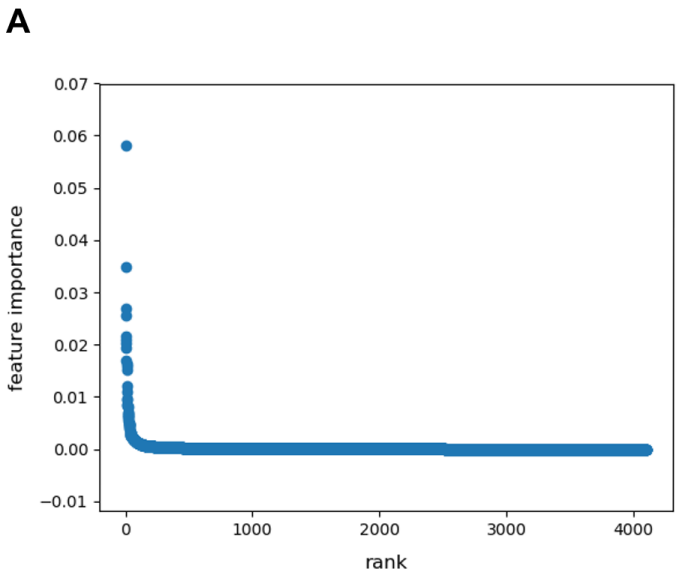

Figure 3: A) conformal prediction calibration plot of predicted vs observed error. Green dots are conformal predictor points, with red dots corresponding to y=x for comparison.