JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

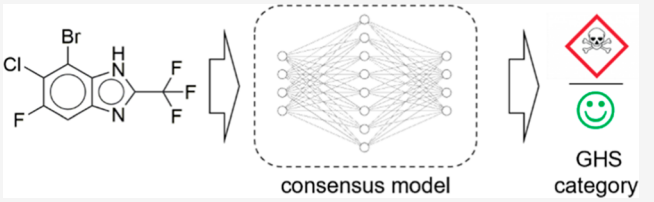
In Silico Prediction of Oral Acute Rodent Toxicity Using Consensus Machine Learning

Sebastian Schieferdecker,*,† Florian Rottach,† and Esther Vock

Cite This: https://doi.org/10.1021/acs.jcim.4c00056 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Acute oral toxicity (AOT) is required for the classification and labeling of chemicals according to the global harmonized system (GHS). Acute oral toxicity studies are optimized to minimize the use of animals. However, with the advent of the three R_s principles and machine learning in toxicology, alternative in silico methods became a reasonable alternative approach for addressing the AOT of new chemical matter. Here, we describe the compilation of AOT data from a commercial database and the development of a consensus classification model after evaluating different combinations of molecular representations and machine learning algorithms. The model shows significantly better performance compared to publicly available AOT models. Its performance was evaluated on an external validation data set, which was compiled from the literature, and an applicability domain was deduced.



INTRODUCTION

The determination of acute oral toxicity (AOT) is an initial test to evaluate the toxicological characteristics of a chemical. The Organisation for Economic Co-operation and Development (OECD) test guideline 423 (December 2001) for the determination of oral acute toxicity mandates a stepwise approach with a minimum number of animals per step and the rat as a preferred rodent species. One important goal of acute oral toxicity studies is the classification and labeling of the tested compounds.[1] The Globally Harmonized System for Classification and Labeling of Chemicals (GHS) allows for the categorization of the acute oral toxicity of a substance into five categories (Table S1).[2,3] Today, the determination of the acute toxicity of pharmaceuticals is no longer needed;[4] however, it is still required for the assessment of chemicals and agrochemicals and can provide guidance for worker safety for intermediate handling during the pharmaceutical manufacturing processes. Historically, in vivo acute oral toxicity determination could utilize ten rats dosed with different concentrations of test compounds.[5] Today, the OECD guidelines implemented the stepwise procedure with 3 animals of a single sex per step. However, it is still the goal to follow the rule of the Three R_s[6] (replacement, reduction, and refinement) with in vitro[7] or in silico methodologies.

In recent times, machine learning has become increasingly prominent in predicting toxicological end points[8,9] as toxicological data became more available and several classification and regression models for predicting acute oral toxicity were published as a result.[10−17]

In 2009, Zhu et al.[18] published a consensus regression model trained on 7385 compounds and employing more than 800 descriptors. The model is made available in the toxicity estimation software tool (TEST) by the US Environmental Protection Agency (EPA).[19] Later, Lei et al.[20] described an AOT consensus regression model based on the relevance vector machine algorithm to predict logarithmic median lethal dose (LD_{50}) values. The model is trained on 7413 compounds and employs molecular descriptors and substructure fingerprints. A data set of 12,200 compounds for AOT prediction was compiled by Lai et al.[16] The authors employed three graph convolution neural networks to predict LD_{50} values and GHS categories. One year later, Li et al.[21] published a classification model for EPA acute oral toxicity classes[22] based on MACCS[23] and FP4 fingerprints and a data set of 12,204 compounds. Wu and Wei[24] applied element-specific topological and physical descriptors to predict AOT with single and multitask deep neural networks. Extremely randomized trees were combined with daylight fingerprints[25,26] to prepare an AOT model named eToxPred by Pu et al.[14] One year later, Minerali et al.[5] published an AOT model which was developed with their proprietary software, Assay Central. The model utilizes Bayesian models in combination with extended connectivity fingerprints (ECFP).[27] The model was trained on 8994 compounds, and the authors reported Matthews correlation coefficients[28] (MCC) values between 0.358 and 0.489 for GHS classes 1 to 4. A consensus QSAR model for organophosphates AOT was

Received: January 10, 2024
Revised: February 20, 2024
Accepted: March 5, 2024

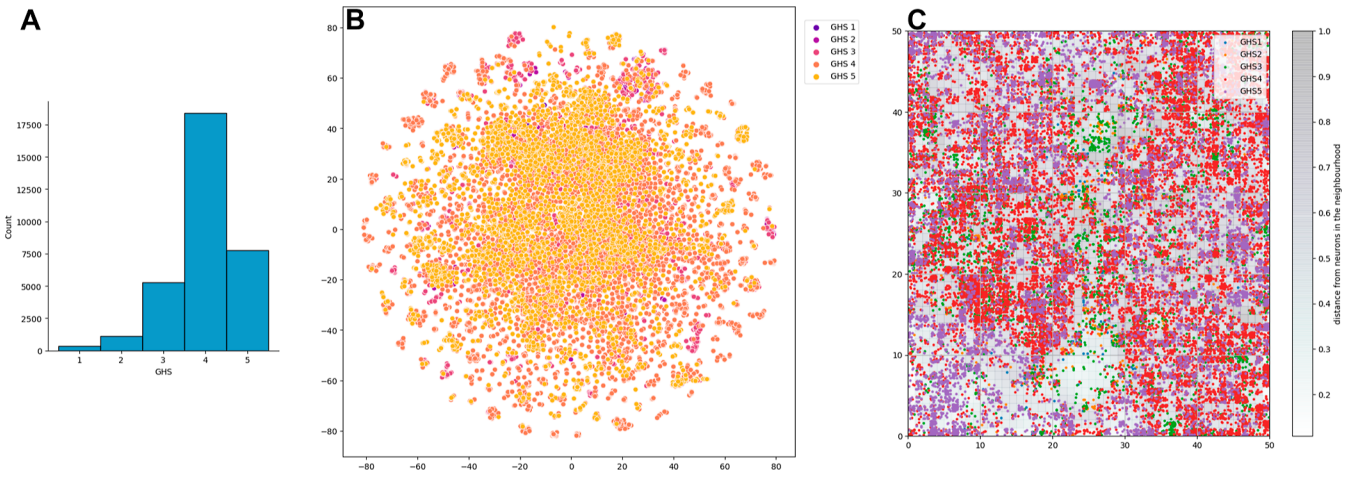


Figure 1. Distribution of GHS classes in the data set (A), visualization of the chemical space distribution of the compound data set by t-SNE (calculated from 1024 bit ECFP4 fingerprints) labeled by GHS class (B), and self-organizing map of the compound data set visualizing compound clustering labeled by GHS class (C).

outlined by Wang et al.[29] The model employed 50 quantum chemical descriptors in combination with 206 2D descriptors and used the extreme gradient boosting (XGBoost) algorithm for prediction. Mansouri et al.[30] reported a consensus AOT model named CATMoS from a collaborative modeling initiative, which was trained on 11,992 compounds and made freely available in the Open QSAR App (OPERA).[31,32] Wijeyesakere et al.[33] reported a mechanistic AOT QSAR model in 2023 to predict LD_{50} values. The model is based on molecular initiating events describing mechanistic fingerprints in combination with MACCS fingerprints and uses a random forest trained on 6234 compounds for prediction of LD_{50} values. A regression and classification model for AOT was described recently by Bo et al.[34] The model is based on extended connectivity fingerprints, which were calculated for 8391 compounds with AOT determined in rats and 5571 compounds with AOT measured in mice. The authors report MCC values for the classifier between 0.45 and 0.47. Finally, Ryu et al.[35] published a model named predAOT, which is based on multiple random forest models and classifies compounds into toxic and nontoxic. Extended connectivity fingerprints were employed as features, and the model was trained on 6226 compounds with AOT determined in mice and 6238 compounds with rat AOT.

While many models are described in the literature, the small training data size used for model creation limits their applicability domain to a certain chemical space. However, due to the complexity of interactions between chemicals and biological systems and several root causes of AOT, bigger data sets are needed for modeling AOT than for traditional QSAR modeling. Furthermore, many of the cited models are not made publicly available, limiting their practical application in the pharmaceutical industry. Here, we describe the development of an AOT consensus model for the prediction of GHS categories, which is trained on a data set of 31,215 compounds.

RESULTS

For the compilation of data used for modeling, the Biovia toxicity database[36] was queried for oral acute toxicity measured in rats. All containing data originated from the Registry of Toxic Effects of Chemical Substances (RTECS).[37] Reported LD_{50} values were converted to the GHS classes. Molecular structures were preprocessed by salt stripping and removal of metal−organic and inorganic structures, as well as compounds with no defined stereochemistry. Duplicated structures were removed from the data set, and an initial 3D conformation was generated for each molecule using CORINA.[38,39] These structures were used as starting points for conformational search using MacroModel[40] with the OPLS4 force field.[41] The lowest energy conformer was further geometry-optimized with the semi-empirical GFN2-xTB method.[42] This workflow resulted in a data set of 31,215 unique molecules suitable for training machine learning models.

Class imbalance is a commonly observed problem in toxicological data sets, which is also present in the data described. The compiled data set is imbalanced toward the less toxic GHS classes, with 16.4% of compounds classified as GHS class 3, 56.4% of compounds associated with GHS class 4, and 22.8% belonging to GHS class 5, while only 1.1 and 3.3%, respectively, belong to GHS classes 1 and 2 (Figure 1A).

Since the generalizability of a molecular machine learning model highly depends on the diversity of the training data, special attention was given to the chemical space covered by the training data. For the description of the chemical diversity of the data set, the simple molecular descriptors heavy atom count (HAC), molecular weight (MW), *A* log *P*,[43] number of hydrogen bond acceptors and donors, number of rings, quantitative estimation of drug likeliness (QED),[44] topological polar surface area, and the fraction of sp^3-hybridized carbons were calculated (Table 1, Figure S1). These descriptors are easily interpretable and can facilitate the decision of whether novel chemicals are covered in the training data of the AOT models. The broad distribution of the calculated features in the data set suggests a broad coverage of chemical space. Among compounds belonging to each GHS class, the distributions are similar (Table S2).

Additionally, chemical diversity was encoded by the calculation of 1024 bit extended connectivity fingerprints with diameter 4 (ECFP4) for all compounds, and pairwise Tanimoto distances were calculated (Figure S2A, the distance matrix of the data set, and Figure S2B−F, the Tanimoto distance distribution of compounds in the individual GHS classes). The mean Tanimoto distance of 0.447 over the whole data set confirms the overall high chemical diversity, which is necessary for a broad applicability domain of its derived models. Additionally, ECFP4

https://doi.org/10.1021/acs.jcim.4c00056
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Description of the Chemical Space of the Dataset Used for AOT Modeling[a]

descriptor	min	max	mean
HAC	4	44	20
MW	53	600	293
A log *P*	−4.8	7	2.8
#HBA	0	16	4
#HBD	0	10	1
CN	0	10	2
QED	0.05	0.95	0.62
TPSA	0	285.15	54.25
Fsp^3	0	1	0.41

[a]HAC heavy atom count, MW molecular weight, #HBA number of hydrogen bond acceptors, #HBD number of hydrogen bond donors, CN cyclomatic number, QED quantitative estimation of drug likeliness, TPSA topological polar surface area, and Fsp^3 fraction of sp^3-hybridized carbons.

fingerprints were transformed with the t-SNE algorithm[45] for the visual depiction of the distribution of chemical diversity in the data set (Figure 1B). While clustering of individual compounds into structural classes can be observed in the figure, the overall covered chemical space of the data set seems large. To better visualize the cluster distribution of the data set, a self-organizing map[46] was calculated based on the same fingerprints (Figure 1C). While compounds belonging to GHS classes 4 and 5 are widely distributed over the map, the more toxic compounds of GHS classes 1 to 3 are aggregated into clusters, which can be associated with specific toxicophoric patterns. To further investigate these structural classes, the compounds belonging to GHS classes 1 and 2 were clustered using the Taylor−Butina algorithm,[47] and the largest clusters were analyzed for the maximum common substructure (Figure 2A) using the FMCS algorithm.[48] The largest class of acute oral toxic compounds are organophosphorus compounds, with 372 structural representatives in the data set. These compounds are known inhibitors of acetylcholinesterase (AChE)[49,50] and are widely used in agriculture as insecticides, with compounds like fosmethilan (**1**) as a well-known representative. The second cluster of acute oral toxic compounds shares the 2-trifluoromethylbenzimidazole partial structure, which is found in 128 compounds in the data set. These compounds are known uncoupling agents of oxidative phosphorylation[51] and commonly used as fungicides and herbicides [for example, chlorflurazole (**2**)]. The third cluster consists of 114 compounds that share an (aminooxy)(oxo)methanamine partial structure. These compounds are known acetylcholinesterase (AChE) inhibitors and are commonly used as insecticides, acaricides, and nematicides for plant protection, with aldicarb (**3**) as a representative compound. The fourth largest cluster of acute oral toxic compounds consists of 41 structural analogues of adenosine. These compounds were investigated as antitumor agents[52] and a representative example structure of this group is compound **4**. Cluster 5 is populated with 33 *N*-nitrosamines like tauromustine (**5**) followed by 28 chlorinated cyclodienes in cluster 6. These compounds were developed as insecticides, piscicides, and rodenticides and possess high neurotoxicity by inhibiting calcium ATPases[53,54] and GABA receptor.[55,56] A representative compound of this group is endrin (**6**). The next most abundant compound class belongs to hexahydroazepino indole derivatives (e.g., compound **7**). The last analyzed cluster consists of 19 compounds of the oxicam family. These compounds are cyclooxygenase inhibitors and are used as nonsteroidal anti-inflammatory agents, with lornoxicam (**8**) as an example structure. In summary, this analysis confirms that the compiled data set covers a broad chemical diversity, and models trained on it should be applicable to a wide range of substance classes.

QSAR models rely on the simple principle that compounds with similar chemical structures possess similar biological activity. This is often the case in structure activity relationships (SAR), where the activity relies on a specific molecular recognition event. However, the cause of acute oral toxicity can be due to multiple specific and unspecific mechanisms (see Figure S3 for a selection of compounds with assigned modes of action), which can result in a flat SAR profile. In addition, small structural changes sometimes lead to big activity changes.[57,58] These activity cliffs are regularly observed in molecular data sets and can have a huge impact on QSAR model performance.[59,60] To investigate the presence of activity cliffs in a molecular data set, the structure−activity landscape index (SALI) was introduced.[61] When depicted in a heatmap (Figure S4), SALI

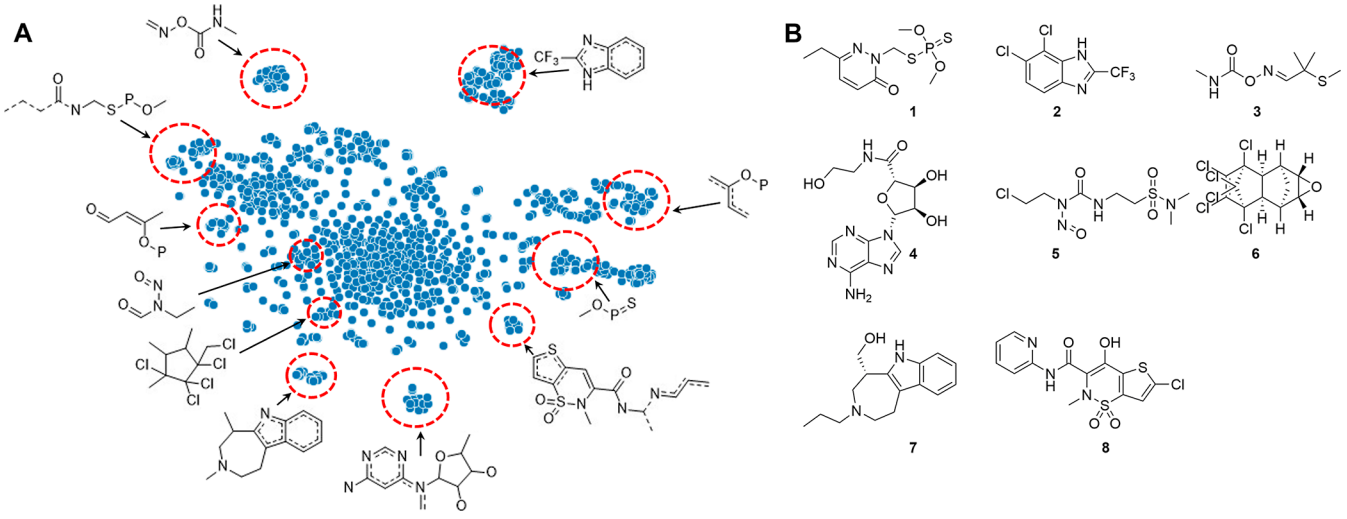


Figure 2. Visualization of the clustering of highly toxic (GHS class 1 and 2) compound structures and the corresponding maximum common substructure of some clusters (A) and exemplary representatives for substance classes from the largest clusters of highly toxic compounds (B).

https://doi.org/10.1021/acs.jcim.4c00056
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

allows for quick identification of challenging compound pairs in the AOT data set. This plot reveals a rather challenging structure-activity landscape to model. Quantitatively, the quality of a QSAR data set can be described with the modelability index (MODI).[62] MODI depends on the fraction of activity cliffs in the data set, and a cutoff of 0.65 was described for the developability of high-quality QSAR models. Again, the MODI of 0.55 calculated for the AOT data set reflects the challenging task of predicting in vivo toxicity.

When developing molecular machine learning models, the type of chemical representation has a strong influence on the predictive performance of the model.[63] To investigate a suitable molecular description for an AOT model, we investigated molecular descriptors and fingerprint vectors in combination with gradient-boosted ensemble decision trees and multilayer perceptrons (MLPs) as well as graph neural network (GNN) architectures and a large language model trained on the simplified molecular input line entry system (SMILES). For the molecular descriptors, three individual sets of descriptors were employed for modeling. The first set consisted of 2D descriptors, including simple count-based descriptors, topological indices, and Shannon entropy descriptors (Table S4).[64] The second descriptor set consisted of 3D descriptors such as molecular surfaces and energies (Table S5). Finally, CATS3D topological atom-pair descriptors were employed as the third set of chemical descriptors.[65] Additionally, several feature vector representations were chosen for model generation. These included extended connectivity fingerprints,[27] Mol2Vec,[66] continuous and data-driven molecular descriptors (CDDD),[67] molecular field points,[68] and pharmacophore feature-map vectors.[69] ECFP4 are circular fingerprints, which map the chemical substructures around a heavy atom by a defined radius using the Morgan algorithm.[70] The Mol2Vec algorithm is inspired by natural language processing and maps molecules into a high-dimensional embedding space, where vectors of similar molecules are grouped closely in the vector space. CDDD vectors are embeddings of an autoencoder neural network which was trained to translate SMILES to InChI line notations. Molecular field points are extrema of electrostatic, steric, and hydrophobic fields derived from the eXtended Electron Distribution (XED) force field.[71,72] Finally, pharmacophore feature-map vectors are generated by the alignment of compound conformers to a reference feature map. While tree-based learners and MLPs require manual feature engineering, GNNs are a class of representation learning algorithms that intrinsically generate feature vectors as embeddings from convolution and aggregation steps during the processing of the molecular graph. Based on the kind of aggregation function employed and the possibility to use both node and edge (corresponding to atom and bond) features (see Table S6), seven GNN architectures were tested in this work. While molecules are usually viewed as an undirected graph, the directed acyclic graph (DAG) model converts this graph to a series of directed graphs.[73] For each atom in the molecule, a DAG is created, which uses this atom as a vertex of the DAG and all edges pointing "inwards" to this vertex. The second evaluated GNN architecture is based on neural graph fingerprints, as described by Duvenaud et al.[74] The graph convolution starts with a set of descriptors for each atom in the molecule that are then combined over two convolutional layers. The third GNN model relies on spectral graph convolutions.[75] First, a weighted sum of the transformed node representations in the graph is calculated. Afterward, max pooling is applied to the node representations, the outputs are concatenated, and the final prediction is done using a MLP. The Weave GNN architecture employs fuzzy histograms for each dimension of the feature vector, which are concatenated on the molecule-level representation.[76] Graph attention networks (GAT) employ the attention mechanism, which allows them to focus on the most relevant parts of a molecular graph.[77] Message passing neural networks (MPNN) combine node and edge features with several rounds of message passing.[78] Finally, the GINE architecture is a GNN model which satisfies the Weisfeiler–Lehmann test.[79] Next to GNNs, large language models (LLM), and in particular the transformer architecture, have made significant progress in natural language processing.[80] The ChemBERTa model is a BERT-style architecture which was pretrained on 77 million canonicalized SMILES taken from the PubChem database.[81,82] Subsequent fine-tuning was performed to predict GHS classes.

For model training, the data set was split into a training and test portion while stratifying the GHS classes. Splitting was done by two means: random and scaffold based. While random splitting does not account for the distribution of chemical scaffolds in the individual data splits, scaffold splitting employs Bemis–Murcko clustering to ensure that the individual splits do not share similar scaffolds.[83] This enables evaluation of the extrapolation capability of the models in different regions of chemical space unknown to the model. Each splitting strategy was used in triplicate, using different random seeds. To account for class differences, an oversampling strategy of the minority classes was evaluated. For model training, initial hyperparameter optimization was done using a grid search strategy together with 5-fold cross-validation of the training data. Models were evaluated based on their Matthews correlation coefficient (MCC) and the area under the receiver operating characteristics curve (AUROC) for each GHS class. Individual model performance is depicted in Tables S7–S10.

The best performance for gradient-boosted decision trees could be achieved with 2D-molecule vector representations, followed by 2D descriptors. On the contrary, models based on 3D molecular representation showed lower predictive power, with molecular field points having the lowest MCC. Additionally, the combination of a 2D vector representation with 3D descriptors did not improve the predictive power of the 2D vector-based models, and oversampling to balance the class distribution of the data did not result in any improvement of the models. When trained and evaluated on scaffold splits, the performance of most models dropped significantly except for 2D pharmacophore vectors, which showed the same MCC as when trained on random split data.

The MLP models trained on descriptors were significantly worse compared to their ensemble tree pendants, but when trained on vector representations, comparable results could be obtained. Again, the compensation for the class imbalance did not significantly improve the model's performance. When evaluated on scaffold splits, a similar trend could be observed as for the gradient-boosted decision trees.

Most graph neural networks either outperformed the ensemble tree and MLP models or showed a similar performance. Only the DAG and GAT architectures were not able to predict AOT with an MCC similar to or higher than that of the best decision tree or MLP models. Again, oversampling did not influence the model performance on most architectures and, in fact, decreased model performance in some cases. When evaluated on scaffold splits, most architectures showed similar

https://doi.org/10.1021/acs.jcim.4c00056
J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

MCC values compared to random split data, indicating the good generalizability of these models. Only the two GINE models' performance decreased significantly.

Lastly, the ChemBERTa transformer model showed predictive power similar to that of the best-performing ensemble tree and MLP models and slightly lower performance compared to most GNN architectures when evaluated on random split data. On scaffold splits, the model's performance decreased. Finally, oversampling of the minority classes did not show any influence on the predictive power of the model.

Ensemble models are known to improve the overall prediction quality over single models.[84] Here, we evaluated both stacking and majority voting as ensemble classifiers. While majority voting simply returns the most common class prediction, stacking employs a metamodel to output a prediction based on the different single model predictions. Both a logistic regression and an MLP were evaluated as meta-models. The following best-performing models were evaluated for ensemble modeling: gradient boosting decision trees with ECFP4 fingerprints, MLP with ECFP4 or CDDD fingerprints, AttentiveFP GNN, GCN GNN, MPNN GNN, Weave GNN, and ChemBERTa transformer. All possible combinations with three to ten models were assessed. While model-stacking could not significantly increase predictive performance, employing a majority voting strategy with five individual model predictions yielded the best predictive performance (Tables 2 and S11).

Table 2. Metrics of the Best-Performing Ensemble Models

model	Acc	MCC	AUC
majority voting classifier[a]	0.741 ± 0.016	0.553 ± 0.033	0.717 ± 0.029
MLP stacking classifier[b]	0.695 ± 0.008	0.487 ± 0.009	0.708 ± 0.021
logistic regression stacking classifier[b]	0.693 ± 0.004	0.479 ± 0.003	0.701 ± 0.015

[a]Ensemble model using ChemBERTa transformer, MLP with ECFP4, AttentiveFP GNN, GCN GNN, and MPNN GNN predictions. [b]Ensemble model using ChemBERTa transformer, gradient boosting decision trees with ECFP4, MLP with CDDD, AttentiveFP GNN, GCN GNN, MPNN GNN, and Weave GNN predictions.

To validate the majority voting classifier and investigate the possible chance correlation of the single employed classification models, *y*-randomization was chosen as a strategy.[85] For this purpose, the connection between molecular features and the GHS class was interrupted by randomly permutating the compound label while keeping the feature vectors untouched. Using these data, model building was repeated with the same parameters as before. Three random permutations were employed for each model. As shown in Table S12, each individual model lost its predictiveness for all permutations, indicating that no chance correlations between features and class labels are present, and the models learnt meaningful coherences.

To get an impression of the majority voting model's performance in comparison to public domain models, we decided to evaluate the FDA T.E.S.T. tool and CATMos on the same test splits of the AOT data set used for model building. In this evaluation, both tools had significantly poorer performance, with big differences for the individual splits as expressed by the MCC values of 0.143 ± 0.154 and 0.193 ± 0.173, respectively.

Molecular machine learning models often perform well in their trained chemical space but struggle when extrapolation into an unknown chemical space is necessary. A possibility to address this issue is to define an applicability domain for the model that quantifies whether a prediction for a certain compound can be made or not. To investigate a possible applicability domain for the majority voting classifier AOT model, we decided to employ two literature data sets.[33,86] Both data sets were combined, and compounds which also appeared in the training data were removed. The resulting external validation data set consists of 3793 compounds with a similar GHS class distribution as the training data (Figure 3A). The highest Tanimoto similarity of each compound in the external validation data to all training data compounds was computed employing either ECFP4 or RDKit fingerprints, and compounds were binned based on their structural similarity (Figure 3B,C). Noteworthy similarities were much higher if calculated with the RDKit fingerprint compared to ECFP4. Predictions on the validation data were done for the whole data and for subsets, where compounds were removed based on similarity thresholds. When the AOT model was evaluated on the complete validation set, a MCC of 0.453 was achieved. When highly dissimilar structures were removed, prediction quality improved (Figure 3D) and reached a MCC of 0.5 when compounds were removed with a Tanimoto similarity smaller than 0.5 (ECFP4 fingerprint) or smaller than 0.6 (RDKit fingerprint). Using these similarity boundaries would allow predictions for 55.45 and 65.6%, respectively, of the external validation data set (Figure 3E) and present a good measure for the applicability domain of the model.

Evaluating the reliability of machine learning model predictions is a challenging and active research field. The final Softmax layer in the classification neural networks produces an estimate of a probability distribution over the output classes. These models are, however, well known for being overconfident in the reported class probabilities. Intuitively, for a perfectly calibrated model, we expect for 100 predictions with a confidence level of 80% 80 correctly classified samples. This perfect relationship can oftentimes not be observed, which can mainly be explained by the characteristics of modern architectures, like the number of model parameters, regularization, normalization layers, and loss functions.[87] Generally speaking, the class estimates can only be as good as the model itself, and if the model is wrong, class probabilities can be quite misleading.

Many approaches have been developed to calibrate these output probabilities to get more trustworthy predictions that better align with human intuition.[87] These techniques are mainly based on postprocessing to refine the output distribution of the final Softmax layer in neural networks.

Even though it is an imperfect measure, Softmax output has been discussed as performing moderately well in many real-world scenarios, and therefore, we perform an analysis on the calibration of the class estimates of the AOT model to assess if any postprocessing is required.[88]

In addition to confidence, the measurement of model uncertainty plays a crucial role in the deployment of machine learning models into real-world applications. In contrast to confidence, uncertainty quantification additionally allows the detection of out-of-distribution (OOD) data, by estimating the uncertainty in the model parameters (epistemic uncertainty). Class probabilities are mainly reflecting aleatoric uncertainty, which is the uncertainty in the data, and therefore they are not suited for detecting data that originates from a different data distribution.[89] Despite this, it has been shown empirically that Softmax distributions are suitable as the effective baseline to

https://doi.org/10.1021/acs.jcim.4c00056
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

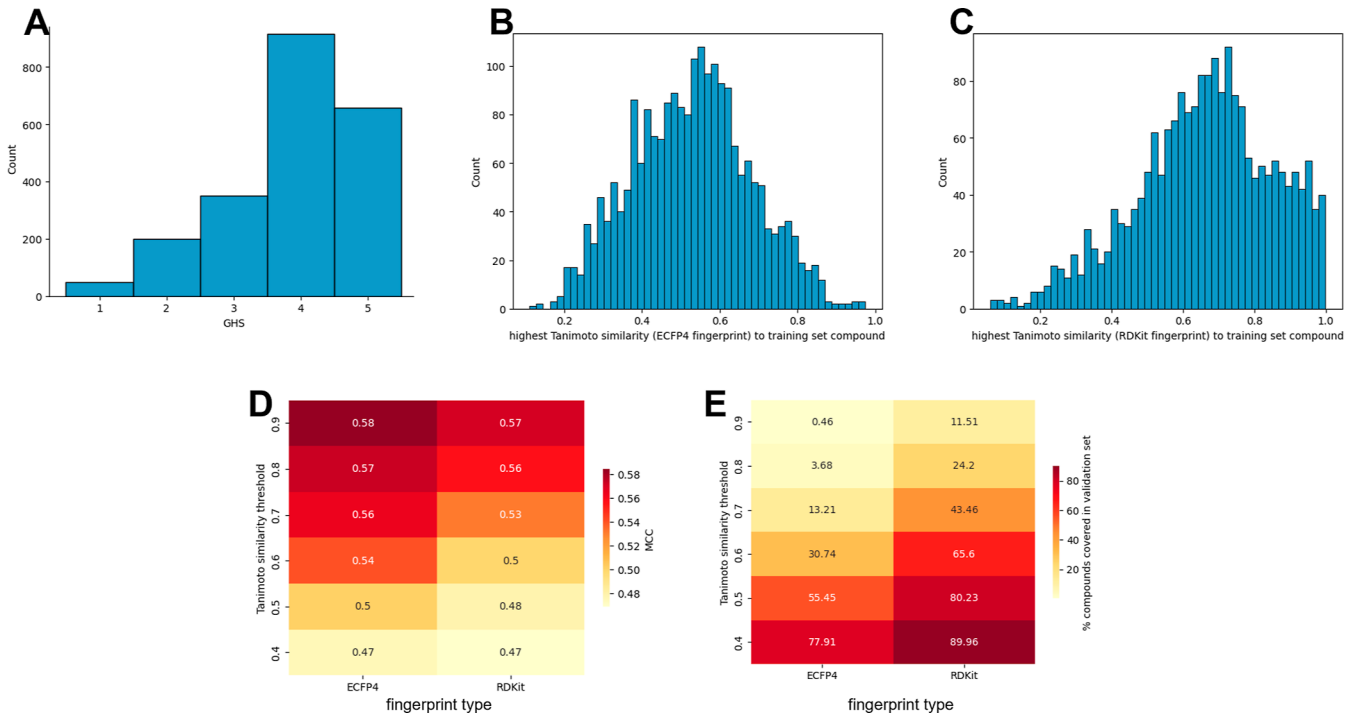


Figure 3. GHS class distribution of the external validation data set (A), distribution of the highest Tanimoto similarity of the external validation data set to the compounds from the data set used for model training calculated with ECFP4 (B) and RDKit fingerprints (C), MCC score (D), and percent of compounds in the applicability domain (E) of the majority voting consensus model evaluated on the external validation data set using Tanimoto similarity thresholds to the training data set.

determine out-of-distribution data.[90] In many cases, the prediction probability of incorrect and OOD samples tends to be lower than that for correct examples.

Based on these findings, we aggregate the class probabilities of the consensus model and empirically analyze the calibration and uncertainty prediction abilities of a holdout set with 3793 samples. Figure 4 shows a clear linear trend that suggests that the model outputs are correctly calibrated. We only display the confidence levels between 0.3 and 0.9, as no other values were observed in the test set. The results also indicate that the model is consistently overconfident, as highlighted by the blue bars, which is expected based on the previous paragraph.

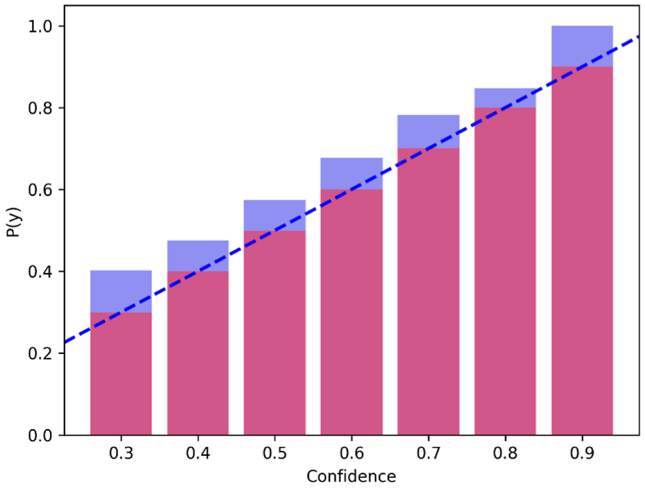


Figure 4. Confidence calibration of the consensus model on a holdout set with 3793 samples indicated a slight overconfidence in the probability estimates. Blue bars represent the estimated probability $P(y)$ by the model, and red bars depict the ground truth confidence levels.

In summary, the naïve interpretation of Softmax confidence as a confidence measure provides a useful proxy for faithful predictions. We also calculate the Tanimoto distance to the training data and provide this as Supporting Information to detect the OOD samples.

CONCLUSIONS

In summary, we conducted a comprehensive in silico analysis of rodent acute oral toxicity data from the Biovia toxicity database and prepared classification models for the GHS categories. We could show that modeling AOT is challenging because of the class imbalance toward highly toxic GHS classes, the presence of different specific and unspecific modes of action causing AOT, and the presence of activity cliffs in the data. While the choice of molecular representation had a huge impact on model performance, addressing class imbalance by oversampling did not offer any performance improvements. As expected, models trained on randomly split data showed better performance than models trained on scaffold splits, indicating the limited extrapolability of molecular machine learning models. Only models trained on 2D pharmacophore vectors showed the same predictive performance on both splitting types, indicating that this kind of descriptor is an abstract enough description of chemical space, allowing more reliable predictions on highly dissimilar compounds. Models trained on features based on 3D representations of molecules in general performed worse than 2D-based descriptors. This could have its cause in the fact that a bioactive conformation for the compounds is unknown, and several conformations could contribute to their toxicity. Using an ensemble modeling strategy, the predictive performance of the single models could be improved by applying a majority

https://doi.org/10.1021/acs.jcim.4c00056
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

voting strategy. This model reached a MCC of 0.55 on randomly split data, and its performance was compared with two public domain models on the same data splits, which both performed significantly worse. Finally, the model was evaluated on an external validation data set compiled from the literature. Using this data set, an applicability domain was established based on fingerprint similarity to the model training data. When the boundaries of the applicability domain were applied to the external validation data set, a similar predictive performance to the test splits of the data used for modeling could be reached.

Overall, this study presents a machine learning-based modeling strategy for rodent AOT that can help ensure that 3R principles are adhered to and GHS categories for novel chemical matter can be provided faster and more cost effectively.

■ ASSOCIATED CONTENT

Data Availability Statement

The developed model code is available at https://github.com/Boehringer-Ingelheim/TOXPR. The AOT training data set was extracted from the commercial Biovia Toxicity Database, which can be acquired from Dassault Systemes SE. The validation data set is available in the Supporting Information.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00056.

Compound structures (SMILES) and GHS categories of the validation data set (XLSX)

Details about molecular structure curation, calculation of molecular descriptors and fingerprints, data set splitting, t-SNE calculation, SOM calculation, calculation of SALI and MODI, machine learning, figures describing the training data set descriptor distribution, Tanimoto-distance matrix, example compounds from the training data set with different modes of action for AOT and pairwise SALI map for each compound of the training data set, table of GHS categories and corresponding LD_{50} bins and compound counts for each GHS category in the training data set, List of used 2D descriptors, list of used 3D descriptors, featurization of GNNs, individual model performance, and y-scrambling results (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Sebastian Schieferdecker − *Boehringer Ingelheim Pharma GmbH & Co. KG, 88397 Biberach, Germany;* orcid.org/0000-0002-4016-7409; Email: sebastian.schieferdecker@boehringer-ingelheim.com

Authors

Florian Rottach − *Boehringer Ingelheim Pharma GmbH & Co. KG, 88397 Biberach, Germany*

Esther Vock − *Boehringer Ingelheim Pharma GmbH & Co. KG, 88397 Biberach, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00056

Author Contributions

†S.S. and F.R. contributed equally. S.S. and F.R. carried out method design, ML modeling, and wrote the manuscript. E.V. supervised the project and revised the manuscript. All authors have given their approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Seidle, T.; Robinson, S.; Holmes, T.; Creton, S.; Prieto, P.; Scheel, J.; Chlebus, M. Cross-Sector Review of Drivers and Available 3Rs Approaches for Acute Systemic Toxicity Testing. *Toxicol. Sci.* **2010**, *116*, 382−396.

(2) United Nations. *Globally Harmonized System of Classification and Labelling of Chemicals: GHS*, 6th ed.; United Nations: New York, NY, 2015.

(3) OECD. *OECD Guideline for the Testing of Chemicals*. No. 423: Acute Oral Toxicity — Acute Toxic Class Method, 14pp; Organisation for Economic Co-Operation and Development: Paris, France, 2001. Https://Www.Oecd.Org/Chemicalsafety/Risk-Assessment/1948378.Pdf.

(4) ICH Expert Working Group. *ICH Topic M3 (R2): Non-clinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorization for Pharmaceuticals, Step 4, June 2009, CPMP/ICH/286/95, 25pp*; European Medicines Agency: London, UK, 2009. Https://Database.Ich.Org/Sites/Default/Files/M3_R2__Guideline.Pdf.

(5) Minerali, E.; Foil, D. H.; Zorn, K. M.; Ekins, S. Evaluation of Assay Central Machine Learning Models for Rat Acute Oral Toxicity Prediction. *ACS Sustain. Chem. Eng.* **2020**, *8*, 16020−16027.

(6) Russell, W. M. S.; Burch, R. L. *The Principles of Humane Experimental Technique*; Methuen & Co. Ltd.: London, 1959.

(7) Schrage, A.; Hempel, K.; Schulz, M.; Kolle, S. N.; van Ravenzwaay, B.; Landsiedel, R. Refinement and Reduction of Acute Oral Toxicity Testing: A Critical Review of the Use of Cytotoxicity Data. *Altern. Lab. Anim.* **2011**, *39*, 273−295.

(8) Wang, M. W. H.; Goodman, J. M.; Allen, T. E. H. Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models. *Chem. Res. Toxicol.* **2021**, *34*, 217−239.

(9) Rácz, A.; Bajusz, D.; Miranda-Quintana, R. A.; Héberger, K. Machine Learning Models for Classification Tasks Related to Drug Safety. *Mol. Divers.* **2021**, *25*, 1409−1424.

(10) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*, 80.

(11) 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. *J. Chem. Inf. Model.* **2018**, *58*, 1533−1543.

(12) Jiang, J.; Wang, R.; Wei, G.-W. GGL-Tox: Geometric Graph Learning for Toxicity Prediction. *J. Chem. Inf. Model.* **2021**, *61*, 1691−1700.

(13) Karim, A.; Mishra, A.; Newton, M. A. H.; Sattar, A. Efficient Toxicity Prediction via Simple Features Using Shallow Neural Networks and Decision Trees. *ACS Omega* **2019**, *4*, 1874−1888.

(14) Pu, L.; Naderi, M.; Liu, T.; Wu, H.-C.; Mukhopadhyay, S.; Brylinski, M. EToxPred: A Machine Learning-Based Approach to Estimate the Toxicity of Drug Candidates. *BMC Pharmacol. Toxicol.* **2019**, *20*, 2.

(15) Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, *59*, 1253−1268.

(16) 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*, 2672−2685.

(17) Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. Deep Learning for Drug-Induced Liver Injury. *J. Chem. Inf. Model.* **2015**, *55*, 2085−2093.

(18) 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*, 1913−1921.

(19) United States Environmental Protection Agency. Toxicity Estimation Software Tool (TEST). https://www.epa.gov/chemcial-research/toxicity-estimation-software-tool-test.
(20) 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. Cheminf.* **2016**, *8*, 6.
(21) 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*, 1061−1069.
(22) United States Environmental Protection Agency. *Label Review Manual, Chapter 7: Precautionary Statements*; U.S. EPA: Washington, DC, 2012.
(23) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(24) Wu, K.; Wei, G.-W. Quantitative Toxicity Prediction Using Topology Based Multitask Deep Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 520−531.
(25) Daylight. Daylight Fingerprints. https://www.daylight.com/meetings/summerschool101/course/basic/fp.thml.
(26) Daylight. Daylight Theory: Fingerprints. https://www.daylight.com/dayhtml/doc/theory/theory.finger.html.
(27) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(28) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta Protein Struct.* **1975**, *405*, 442−451.
(29) Wang, L.; Ding, J.; Shi, P.; Fu, L.; Pan, L.; Tian, J.; Cao, D.; Jiang, H.; Ding, X. Ensemble Machine Learning to Evaluate the in Vivo Acute Oral Toxicity and in Vitro Human Acetylcholinesterase Inhibitory Activity of Organophosphates. *Arch. Toxicol.* **2021**, *95*, 2443−2457.
(30) Mansouri, K.; Karmaus, A. L.; Fitzpatrick, J.; Patlewicz, G.; Pradeep, P.; Alberga, D.; Alepee, N.; Allen, T. E. H.; Allen, D.; Alves, V. M.; Andrade, C. H.; Auernhammer, T. R.; Ballabio, D.; Bell, S.; Benfenati, E.; Bhattacharya, S.; Bastos, J. V.; Boyd, S.; Brown, J. B.; Capuzzi, S. J.; Chushak, Y.; Ciallella, H.; Clark, A. M.; Consonni, V.; Daga, P. R.; Ekins, S.; Farag, S.; Fedorov, M.; Fourches, D.; Gadaleta, D.; Gao, F.; Gearhart, J. M.; Goh, G.; Goodman, J. M.; Grisoni, F.; Grulke, C. M.; Hartung, T.; Hirn, M.; Karpov, P.; Korotcov, A.; Lavado, G. J.; Lawless, M.; Li, X.; Luechtefeld, T.; Lunghini, F.; Mangiatordi, G. F.; Marcou, G.; Marsh, D.; Martin, T.; Mauri, A.; Muratov, E. N.; Myatt, G. J.; Nguyen, D.-T.; Nicolotti, O.; Note, R.; Pande, P.; Parks, A. K.; Peryea, T.; Polash, A. H.; Rallo, R.; Roncaglioni, A.; Rowlands, C.; Ruiz, P.; Russo, D. P.; Sayed, A.; Sayre, R.; Sheils, T.; Siegel, C.; Silva, A. C.; Simeonov, A.; Sosnin, S.; Southall, N.; Strickland, J.; Tang, Y.; Teppen, B.; Tetko, I. V.; Thomas, D.; Tkachenko, V.; Todeschini, R.; Toma, C.; Tripodi, I.; Trisciuzzi, D.; Tropsha, A.; Varnek, A.; Vukovic, K.; Wang, Z.; Wang, L.; Waters, K. M.; Wedlake, A. J.; Wijeyesakere, S. J.; Wilson, D.; Xiao, Z.; Yang, H.; Zahoranszky-Kohalmi, G.; Zakharov, A. V.; Zhang, F. F.; Zhang, Z.; Zhao, T.; Zhu, H.; Zorn, K. M.; Casey, W.; Kleinstreuer, N. C. CATMoS: Collaborative Acute Toxicity Modeling Suite. *Environ. Health Perspect.* **2021**, *129*, 047013.
(31) OPERA. https://npt.niehs.nih.gov/wgatwestudy/niceatm/comptox/ct-opera/opera.
(32) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *J. Cheminf.* **2018**, *10*, 10.
(33) Wijeyesakere, S. J.; Auernhammer, T.; Parks, A.; Wilson, D. Profiling Mechanisms That Drive Acute Oral Toxicity in Mammals and Its Prediction via Machine Learning. *Toxicol. Sci.* **2023**, *193*, 18−30.
(34) Bo, T.; Lin, Y.; Han, J.; Hao, Z.; Liu, J. Machine Learning-Assisted Data Filtering and QSAR Models for Prediction of Chemical Acute Toxicity on Rat and Mouse. *J. Hazard. Mater.* **2023**, *452*, 131344.
(35) Ryu, J. Y.; Jang, W. D.; Jang, J.; Oh, K.-S. PredAOT: A Computational Framework for Prediction of Acute Oral Toxicity Based on Multiple Random Forest Models. *BMC Bioinf.* **2023**, *24*, 66.
(36) BIOVIA Databases-BIOVIA-Dassault Systèmes. https://www.3ds.com/.
(37) Registry of Toxic Effects of Chemical Substances (RTECS). https://www.cdc.gov/niosh/docs/97-119/default.html.
(38) CORINA Classic. https://mn-am.com/products/corina/.
(39) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.
(40) *MacroModel*; Schrödinger, LLC: New York, NY, 2023.
(41) Lu, C.; Wu, C.; Ghoreishi, D.; Chen, W.; Wang, L.; Damm, W.; Ross, G. A.; Dahlgren, M. K.; Russell, E.; Von Bargen, C. D.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. *J. Chem. Theory Comput.* **2021**, *17*, 4291−4300.
(42) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-XTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.
(43) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(44) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90−98.
(45) van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(46) Kohonen, T. Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* **1982**, *43*, 59−69.
(47) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.
(48) Dalke, A.; Hastings, J. FMCS: A Novel Algorithm for the Multiple MCS Problem. *J. Cheminf.* **2013**, *5*, O6.
(49) Guo, J.-X.; Wu, J. J.-Q.; Wright, J. B.; Lushington, G. H. Mechanistic Insight into Acetylcholinesterase Inhibition and Acute Toxicity of Organophosphorus Compounds: A Molecular Modeling Study. *Chem. Res. Toxicol.* **2006**, *19*, 209−216.
(50) Morris, C. M.; Savy, C.; Judge, S. J.; Blain, P. G. *Basic and Clinical Toxicology of Organophosphorus Compounds*; Springer Science & Business Media, 2013; pp 45−78.
(51) Spencer, Y. E. *Guide to the Chemicals Used in Crop Protection*; Agriculture Canada: Ottawa, Canada, 1982.
(52) Fleysher, M. H.; Bernacki, R. J.; Bullard, G. A. Some Short-Chain N6-Substituted Adenosine Analogs with Antitumor Properties. *J. Med. Chem.* **1980**, *23*, 1448−1452.
(53) Mehrotra, B. D.; Bansal, S. K.; Desaiah, D. Comparative Effects of Structurally Related Cyclodiene Pesticides on ATPases. *J. Appl. Toxicol.* **1982**, *2*, 278−283.
(54) Mehrotra, B. D.; Reddy, S. R.; Desaiah, D. Effect of Subchronic Dieldrin Treatment on Calmodulin-regulated Ca2+ Pump Activity in Rat Brain. *J. Toxicol. Environ. Health* **1988**, *25*, 461−469.
(55) Glotfelty, D. E. The Atmosphere as a Sink for Applied Pesticides. *J. Air Pollut. Control Assoc.* **1978**, *28*, 917−921.
(56) Wafford, K. A.; Lummis, S. C. R.; Sattelle, D. B. Block of an Insect Central Nervous System GABA Receptor by Cyclodiene and Cyclohexane Insecticides. *Proc. R. Soc. London, Ser. B* **1989**, *237*, 53−61.
(57) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18−28.
(58) Cruz-Monteagudo, M.; Medina-Franco, J. L.; Pérez-Castillo, Y.; Nicolotti, O.; Cordeiro, M. N. D. S.; Borges, F. Activity Cliffs in Drug Discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today* **2014**, *19*, 1069−1080.
(59) Maggiora, G. M. On Outliers and Activity Cliffs—Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
(60) Sheridan, R. P.; Karnachi, P.; Tudor, M.; Xu, Y.; Liaw, A.; Shah, F.; Cheng, A. C.; Joshi, E.; Glick, M.; Alvarez, J. Experimental Error,

https://doi.org/10.1021/acs.jcim.4c00056
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure-Activity Relationship Models? *J. Chem. Inf. Model.* **2020**, *60*, 1969−1982.
(61) Guha, R.; Van Drie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646−658.
(62) Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data Set Modelability by QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 1−4.
(63) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29*, 476−488.
(64) Gregori-Puigjané, E.; Mestres, J. SHED: Shannon Entropy Descriptors from Topological Feature Distributions. *J. Chem. Inf. Model.* **2006**, *46*, 1615−1622.
(65) Reutlinger, M.; Koch, C. P.; Reker, D.; Todoroff, N.; Schneider, P.; Rodrigues, T.; Schneider, G. Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for 'Orphan' Molecules. *Mol. Inform.* **2013**, *32*, 133−138.
(66) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.
(67) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chem. Sci.* **2019**, *10*, 1692−1701.
(68) Cheeseright, T.; Mackey, M.; Rose, S.; Vinter, A. Molecular Field Extrema as Descriptors of Biological Activity: Definition and Validation. *J. Chem. Inf. Model.* **2006**, *46*, 665−676.
(69) Landrum, G. A.; Penzotti, J. E.; Putta, S. Feature-Map Vectors: A New Class of Informative Descriptors for Computational Drug Discovery. *J. Comput.-Aided Mol. Des.* **2007**, *20*, 751−762.
(70) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.
(71) Vinter, J. G. Extended Electron Distributions Applied to the Molecular Mechanics of Some Intermolecular Interactions. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 653−668.
(72) Vinter, J. G. Extended Electron Distributions Applied to the Molecular Mechanics of Some Intermolecular Interactions. II. Organic Complexes. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 417−426.
(73) Lusci, A.; Pollastri, G.; Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *J. Chem. Inf. Model.* **2013**, *53*, 1563−1575.
(74) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv* **2015**, arXiv:1509.09292.
(75) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
(76) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput. Aided Mol. Des.* **2016**, *30*, 595−608.
(77) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *arXiv* **2017**, arXiv:1710.10903.
(78) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *arXiv* **2017**, arXiv:1704.01212.
(79) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks? *arXiv* **2018**, arXiv:1810.00826.
(80) OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
(81) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
(82) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv* **2020**, arXiv:2010.09885.
(83) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(84) Hansen, L. K.; Salamon, P. Neural Network Ensembles. *IEEE Trans. Pattern Anal.* **1990**, *12*, 993−1001.
(85) Rücker, C.; Rücker, G.; Meringer, M. Y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345−2357.
(86) Banerjee, P.; Eckert, A. O.; Schrey, A. K.; Preissner, R. ProTox-II: A Webserver for the Prediction of Toxicity of Chemicals. *Nucleic Acids Res.* **2018**, *46*, W257−W263.
(87) Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q. On Calibration of Modern Neural Networks. *arXiv* **2017**, arXiv:1706.04599.
(88) Pearce, T.; Brintrup, A.; Zhu, J. Understanding Softmax Confidence and Uncertainty. *arXiv* **2021**, arXiv:2106.04972.
(89) Hüllermeier, E.; Waegeman, W. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Mach. Learn.* **2021**, *110*, 457−506.
(90) Hendrycks, D.; Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv* **2016**, arXiv:1610.02136.

https://doi.org/10.1021/acs.jcim.4c00056
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX