JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2017, 57, 2672-2685

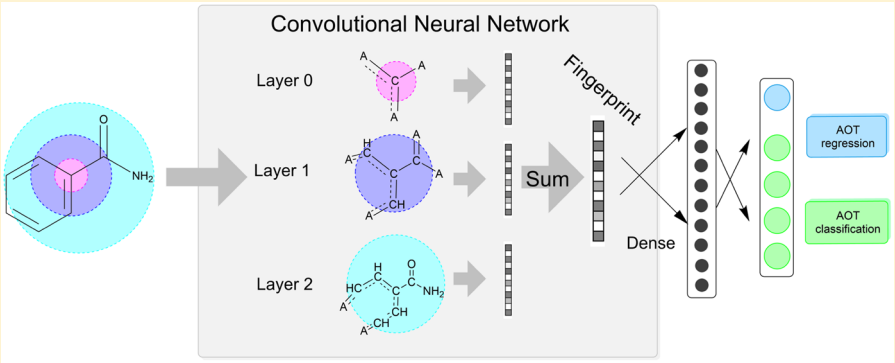
pubs.acs.org/jcim

Deep Learning Based Regression and Multiclass Models for Acute Oral Toxicity Prediction with Automatic Chemical Feature Extraction

Youjun Xu,[†] Jianfeng Pei,*[,†] and Luhua Lai*[,†,‡,¶]

[†]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, [‡]Beijing National Laboratory for Molecular Sciences, State Key Laboratory for Structural Chemistry of Unstable and Stable Species, College of Chemistry and Molecular Engineering, and [¶]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China

Supporting Information



ABSTRACT: Median lethal death, LD_{50}, is a general indicator of compound acute oral toxicity (AOT). Various in silico methods were developed for AOT prediction to reduce costs and time. In this study, we developed an improved molecular graph encoding convolutional neural networks (MGE-CNN) architecture to construct three types of high-quality AOT models: regression model (deepAOT-R), multiclassification model (deepAOT-C), and multitask model (deepAOT-CR). These predictive models highly outperformed previously reported models. For the two external data sets containing 1673 (test set I) and 375 (test set II) compounds, the R^2 and mean absolute errors (MAEs) of deepAOT-R on the test set I were 0.864 and 0.195, and the prediction accuracies of deepAOT-C were 95.5% and 96.3% on test sets I and II, respectively. The two external prediction accuracies of deepAOT-CR are 95.0% and 94.1%, while the R^2 and MAE are 0.861 and 0.204 for test set I, respectively. We then performed forward and backward exploration of deepAOT models for deep fingerprints, which could support shallow machine learning methods more efficiently than traditional fingerprints or descriptors. We further performed automatic feature learning, a key essence of deep learning, to map the corresponding activation values into fragment space and derive AOT-related chemical substructures by reverse mining of the features. Our deep learning architecture for AOT is generally applicable in predicting and exploring other toxicity or property end points of chemical compounds. The two deepAOT models are freely available at http://repharma.pku.edu.cn/DLAOT/DLAOThome.php or http://www.pkumdl.cn/DLAOT/DLAOThome.php.

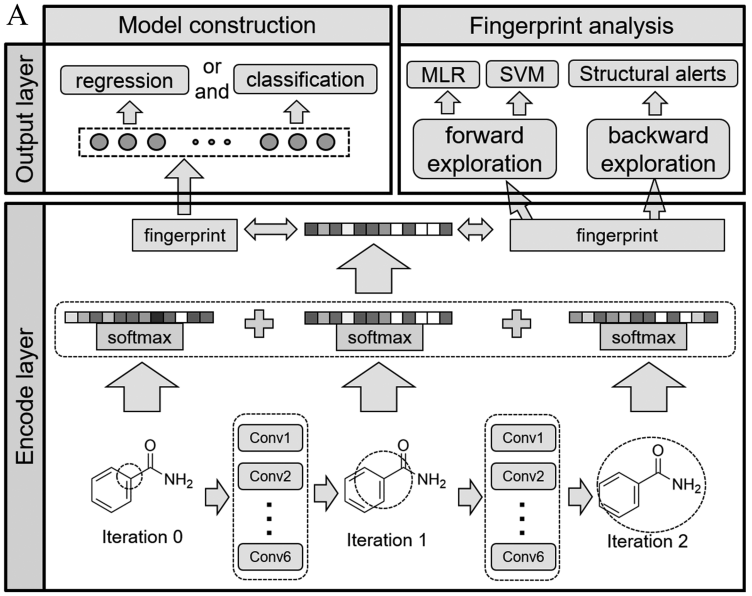
INTRODUCTION

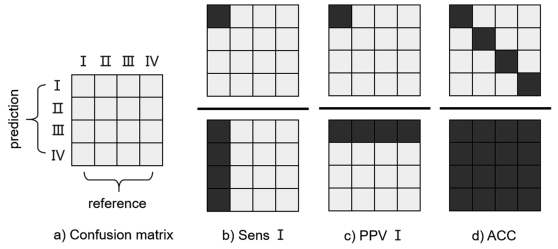
Evaluating chemical acute toxicity is important in avoiding potential harmful effects of compounds on human health. LD_{50}, the dose of a chemical that causes a 50% death rate in test animals after administration of a single dose,[1] is a general indicator used to measure the acute toxicity of a compound. *In vivo* experiments of animal tests are required to accurately determine acute chemical toxicity, although these procedures are complicated, costly, and time-consuming. In addition, due to animal rights, LD_{50} testing on animals is highly controversial.[2] Therefore, new reliable *in silico* methods need to be developed in comparison to standard *in vivo* experiments in predicting chemical acute toxicity.

Currently, many quantitative structure−property relationship (QSPR) models have been developed to predict acute rodent toxicity of organic chemicals. In these studies, there are various mathematical methods applied to construct regression models (RMs) and classification models (CMs), such as multiple linear regression (MLR),[3−6] linear regression,[7,8] neural network (NN),[9−12] *k* nearest neighbors,[13,14] random forest (RF),[13,14] hierarchical clustering,[13] support vector machine (SVM),[14,15] relevance vector machine (RVM),[14] and local lazy learning (LLL).[16] In terms of RMs, Lu et al.[16] constructed prediction models using the LLL method, which yielded a maximized linear correlation coefficient (R^2) for large test sets. The R^2 of consensus RM based on LLL was 0.608 for "Set_3874". Lei et al.[14] argued that this method relies on prior knowledge of the

Received: May 3, 2017
Published: October 11, 2017

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685





B

Algorithm 1 MGE-CNN feedforward process

Input: molecule x which contains m atoms, FPD, FPL, HLS, CKW
Given: trained network parameters
$W_{L_1}^{encode} \in \mathbb{R}^{d_{L_1}^{atom} \times |FPL|}$, $b_{L_1}^{encode} \in \mathbb{R}^{1 \times |FPL|}$,
$W_{L_i}^{encode} \in \mathbb{R}^{d_{L_i}^{CKW} \times |FPL|}$, $b_{L_i}^{encode} \in \mathbb{R}^{1 \times |FPL|}$ $(i = 2, ..., FPD + 1)$;
$W_{L_1 \to L_2}^{self} \in \mathbb{R}^{d_{L_1}^{atom} \times |CKW|}$,
$W_{L_i \to L_{i+1}}^{self} \in \mathbb{R}^{|CKW| \times |CKW|}$, $(i = 2, ..., FPD)$,
$W_{L_i}^{convk} \in \mathbb{R}^{d_{L_{i-1}}^{bond+atom} \times |CKW|}$, $b_{L_i}^{conv} \in \mathbb{R}^{1 \times |CKW|}$,
$(i = 2, ..., FPD + 1; k = 1, 2, ..., 6)$;
$W_H^{output} \in \mathbb{R}^{|FPL| \times |HLS|}$, $W_O^{output} \in \mathbb{R}^{|HLS| \times d_{all}^{out}}$,
$b_H^{output} \in \mathbb{R}^{1 \times |HLS|}$, $b_O^{output} \in \mathbb{R}^{1 \times d_{all}^{out}}$
// generate the representations of x
$z_x = [\,]$
for each atom a in x **do**
  Z_{atoms}.add($g_{atoms\_L_1}(a)$) // $g_*(*)$ is lookup operation
end for
$z_{L_1} = sum(softmax(Z_{atoms} W_{L_1}^{encode} + b_{L_1}^{encode}), axis = 0) \in \mathbb{R}^{1 \times |FPL|}$
Z_{layer}.add(z_{L_1})
// extract features from each layer
for $l = 2$ to $FPD + 1$ **do**
  for each atom a in x **do**
    $n = getdegree(a)$
    $r_1, r_2, ..., r_n = neighbors(a)$
    for $i = 1$ to n **do**
      $Z_{neighbor-bonds}[i, :] = g_{bonds}(bond(a \to r_i))$
      $Z_{neighbor-atoms}[i, :] = g_{atoms\_L_{l-1}}(r_i)$
    end for
    $Z_{both} = concatenate(Z_{neighbor-bonds}, Z_{neighbor-atoms})$
    $Z_{neighbors}$.add($sum(Z_{both}, axis = 0) W_{L_l}^{convn}$)
  end for
  $Z_{atoms} = f(normalize(Z_{atoms} W_{L_{l-1} \to L_l}^{self} + Z_{neighbors} + b_{L_l}^{conv}))$
  // $f(*)$ is activation function;
  $z_{L_l} = sum(softmax(Z_{atoms} W_{L_l}^{encode} + b_{L_l}^{encode}), axis = 0)$
  Z_{layer}.add(z_{L_l})
end for
$z_x = sum(Z_{layer}, axis = 0)$
//hidden and output layers
$score = f(z_x W_H^{output} + b_H^{output}) W_O^{output} + b_O^{output}$
if regression model (1-output) **then**
  $score = score$
else if multiclass model (4-output) **then**
  $prob = softmax(score)$
end if
return *score* or *prob*

Figure 1. (A) Schematic diagram of MGE-CNN architecture. "Conv" represents the convolution kernel, and the six kernels rely on the degree of each atom. (B) Overview of pseudocode in Algorithm 1. (C) Assessment method of Sens and PPV for each of the classes and ACC of all the classes. Sens I is equal to the number of the higher black region divided by the sum of the bottom black region, which was identical with PPV I. The roman letters "I, II, III, IV" represent toxicity categories.

query neighbor experimental data, such that the actual prediction capability was associated with the chemical diversity and structural coverage of the training set. However, machine learning methods have demonstrated potential in establishing complex QSPRs for data sets that contain diverse ranges of molecular structures and mechanisms. Thus, Lei et al. employed RVM combined with other methods (k nearest neighbor, RF, SVM, etc.) to construct a consensus RM for predicting AOT in rat. The predictive R^2 for the external test set (of 2736 compounds) was 0.69. Li et al.[15] suggested that a multiclassification model (MCM) might be more intuitive in toxicity estimation than a regression model (RM), as a toxic classification is easier to interpret. According to the classification criterion of the U.S. Environmental Protection Agency (EPA) (category I (0, 50]; category II (50, 500]; category III (500, 5000]; category IV (5000, + ∞); mg/kg), MCM with one-vs-one (SVM_{OAO}) and binary tree SVM methods were developed based on different molecular fingerprints or descriptors, yielding an accuracy of 83.2% for validation set (2049 compounds), 83.0% for test set I (1678 compounds), and 89.9% for test set II (375 compounds). In cheminformatics research, high-quality QSPR models with interpretable relationships between chemical properties and chemical features are especially welcome. However, predictive power and interpretability of QSPR models are two different objectives that are difficult to achieve simultaneously.[17] Some important features can be identified through the analysis of linear-based models (MLR, linear-SVM, etc.) with low predictive power. These features may be mapped into the corresponding fragments in chemical structural space, contributing to extracting structural fragments related to target objects.[18−22] The "black box" models (NN, kernel-SVM, RF) with high predictive power may learn more or less human understandable knowledge from the view of statistics and sensitivity analysis[23] of input features (rather than intuitive analysis of the well-constructed models). And the frequency-based fragment analysis could promote the identification of more meaningful structural fragments.[24,25] Generally, these methods depend on complicated molecular representation (MR) using chemical knowledge and intuition.

Appropriate MRs that are related to biological activities or other experimental end points[26,27] are crucial in developing accurate prediction models. Automatic representation would greatly simplify and accelerate the development of QSPR models. The emergence of deep learning techniques[28−30] may provide possible solutions to this problem. Recent deep learning architectures, like Neural Turing Machines,[31] Memory Networks,[32] Neural Stacks/Queues,[33,34] Neural Programmer (Interpreters),[35,36] Neural Theorem Provers,[37] replace discrete algorithms and data structures with end-to-end differentiable counterparts that operate on real-valued vectors. These real-valued vectors combined with subsequent NN modules can be trained with gradient descent to implement data-driven automatic representations, instead of general discrete molecular representations (e.g., extended-connectivity fingerprints (ECFP),[38] MACCS,[39] etc.). In addition, this kind of MR has

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685

a certain transfer ability on related tasks. Using automatic representations, the AOT issue can be resolved without manual intervention or selection of complicated features. As we know, the two-dimensional (2D) structure of a small molecule is equivalent to an undirect graph, with atoms as nodes and bonds as edges. Encoding different molecules with NN modules can be thought as fitting different graphs into fix-sized vectors representing different molecules. Currently, two types of end-to-end differentiable architectures, sink-based and source-based architecture, were designed for implementing automatic representations of different molecules.[40] For the sink-based approach, by defining a root node, all the other nodes in the graph proceed toward the root. The internal process is embedded with multiple NNs in representing the information transmission between nodes, after which the final information is extracted from the root node and translated into real-valued vectors. The sink-based method was demonstrated to be feasible and practical.[41,42] However, there are no reasonable explanation for hidden-layer features such that the model seems "black". For the source-based approach, similar to ECFP, when starting from an initial node and diffusing outward layer-by-layer with multiple NNs, the distributed information can be extracted stepwise from each layer, then summed as the real-valued vectors. Recently, Duvenaud et al.,[43] Kearnes et al.,[44] and Coley et al.[45] used CNNs to successfully implement similar architectures. It is very convenient and practical for directly mapping SMILES strings into target properties in such end-to-end differentiable architectures. The state-of-the-art performance on some public data sets[46−51] suggests that these approaches for molecular graph encoding (MGE) with NN modules have potential in the field of cheminformatics. In principal, this kind of MGE is an ideal representation of chemical structures without information loss.

Actually, intermediate features within deep learning models are far from random, uninterpretable patterns. By visualizing the activity of hidden layers based on well-performed models from ImageNet 2012, Zeiler et al. presented a nested hierarchy of concepts, with each concept defined in relation to simpler concepts (pixels → edges → corners and contours → object parts → object identity),[52] which is an efficient illustration of a deep learning-based CNN model. Different compounds may play different functions in the living organisms. Simple concepts of atoms and bonds are combined into more complex concepts of structural fragments and, then, integrated into high concepts of different functions (atoms and bonds→ fragments → functions). By designing ECFP-based CNN architecture, the internal features were visualized by Duvenaud et al. as the corresponding activation fragments,[43] providing a better understanding of a deep learning-based QSPR model. Despite of a number of successful application examples using MGE, the following points on AOT need to be improved for better prediction and easy interpretation: (1) hyperparameters, (2) training and prediction strategy, (3) multioutput problem, and (4) model interpretation. The architecture based on CNN with these above improvements was referenced hereafter as "MGE-CNN".

Here we used a MGE-CNN architecture to construct AOT models. Forward and backward exploration were carried out to interpret these models and visualize the learned fragments. In view of end-to-end learning, the MGE-CNN architecture can also be applied to predict and explore other toxicity end points induced by small molecules in complex systems.

MATERIALS AND METHODS

MGE-CNN. The MGE-CNN architecture takes the canonical SMILES string of a small molecule as input and produces a score capable of describing a value or label about toxicity. Figure 1A and B show this architecture and its high-level pseudocode with the steps of MGE-CNN feedforward process. First, given an input SMILES string (x), a molecular structural graph is converted by the RDKit toolbox.[53] The subgraph from each layer (or iteration) is encoded into a fixed-sized vector $z_{L_l} \in \mathbb{R}^{|\mathrm{FPL}|} l \in \{1, 2, ..., |\mathrm{FPD}|\}$ (|FPL| is the length of fingerprint, |FPD| is the depth of fingerprint), then these vectors are summed as $z_x \in \mathbb{R}^{|\mathrm{FPL}|}$ representing this molecule. More detailed description for the encoding of molecules is depicted in Figure S1. Then z_x is used as input of the subsequent neural network in the output layer for executing the following operation:

$$\mathrm{score} = f(z_x W_{\mathrm{H}}^{\mathrm{output}} + b_{\mathrm{H}}^{\mathrm{output}}) W_{\mathrm{O}}^{\mathrm{output}} + b_{\mathrm{O}}^{\mathrm{output}} \quad (1)$$

where $W_{\mathrm{H}}^{\mathrm{output}} \in \mathbb{R}^{|\mathrm{FPL}|\times|\mathrm{HLS}|}$ (|HLS| is the size of the hidden layer) is the weight matrix of hidden layer in the output layer, $W_{\mathrm{O}}^{\mathrm{output}} \in \mathbb{R}^{|\mathrm{HLS}|\times d_{\mathrm{all}}^{\mathrm{out}}}$ is the weight matrix of output layer in the output layer, and $b_{\mathrm{H}}^{\mathrm{output}} \in \mathbb{R}^{1\times|\mathrm{HLS}|}$ and $b_{\mathrm{O}}^{\mathrm{output}} \in \mathbb{R}^{1\times d_{\mathrm{all}}^{\mathrm{out}}}$ are bias terms. $d_{\mathrm{all}}^{\mathrm{out}} = 1$ for RMs, $d_{\mathrm{all}}^{\mathrm{out}} = 4$ for MCMs. The 4-dimensional vector is transformed with the *softmax* function representing the probability of four classes. $p(i|x) = \frac{e^{\mathrm{score}(x)_i}}{\sum_{j=1}^{4} e^{\mathrm{score}(x)_j}}$ is the probability of category i, where $\mathrm{score}(x)_i$ is the score for category i.

The MGE-CNN has three main advantages: (1) The input information on initial atoms and bonds is very similar to that of ECFP. The atom information contains atomic type, its degree, its implicit valence, the number of attached H atoms, and aromatic atoms. The bond information relies on bond type (single, double, triple, aromatic, conjugated, or in-a-ring). These atom and bond-level information is used to characterize the surrounding chemical environment of each atom as completely as possible. All of this information can be calculated using RDKit. (2) Molecular graphs are encoded with CNN, which makes information transmission become continuous and constructs an end-to-end differential system. In such a case, we can perform gradient descent with a large number of labeled data to optimize this system. During the training process, automatic feature learning is implemented, avoiding manual feature selection. (3) Both processes of feature learning and model construction are integrated together. Once the model is well-trained with supervised learning, these fingerprints are also learned.

The MGE-CNN architecture was used to construct AOT models, shown in Figure 1A. In order to develop high-quality deep learning models, namely deepAOT, RMs were constructed using the reported largest AOT data set from Li et al.,[15] including experimental oral LD_{50} values for chemicals in rat. Based on the U.S. EPA criterion for the AOT category, MCMs were also developed to predict chemical toxicity categories. Two external test data sets were used to estimate the predictive power of RMs and MCMs. The consensus RM and the best MCM were called "deepAOT-R" and "deepAOT-C", respectively. We demonstrated that the deepAOT-R and deepAOT-C models outperformed the previous reported models whether it was a regression or classification problem.

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685

Given the relevance of both tasks, the multitask deepAOT-CR model was developed for improving the consistency of regression and classification models. Further analysis was performed by forward and backward exploration (Figure 1A) of internal features (referred to as deep fingerprints) directly extracted from our models to interpret the RMs and MCMs. The forward exploration was used to test the predictability of fingerprints, while the backward exploration was used to understand and explore structural fragments concerning toxicity.

The following improvements for better prediction and easy interpretation in our system were adopted: (1) For hyperparameter optimization in the AOT system, we empirically found that the default settings (β_1 = 0.9, β_2 = 0.999) for adaptive moment estimation (Adam) would be more helpful than those provided by Duvenaud et al. (2) To avoid providing the training examples in a meaningful order (which may bias the optimization algorithm and lead to overfitting), the trick of "shuffling"[54] was added into the whole training process. (3) The popular methods of *softmax* and *cross-entropy* loss functions were introduced to meet the requirements of multiclassification tasks. (4) Regression and classification tasks were taken into consideration simultaneously for developing the multitask model. (5) To further explain the rationality of our models, deep fingerprints directly extracted from well-built models were used to construct shallow machine learning models. The structural fragments with the largest contribution (argmin(linear regression coefficient × activation values)) to chemical toxicity were drawn out for comparison with the reported toxicity alerts, while the original MGE only considered those coefficients. (6) The mean and standard deviation of the training set for each layer are calculated for normalizing validation or external test set, reducing the bias caused by different distributions. Based on these, the MGE-CNN was employed to construct RMs and CMs for estimating AOT in rat, as shown in Figure 1A. During "Model construction", these models were trained, validated, and externally challenged. During "Fingerprint analysis", the well-trained deep fingerprints of small molecules were used to develop shallow models, such as MLR and SVM, to predict AOT values or labels. Simultaneously, the most relevant feature among deep fingerprint for each compound was calculated based on linear regression with least-squares fitting, then traced back to the atomic level, and mapped onto AOT activation fragments. These activated fragments were then used to compare with structural alerts from ToxAlerts[55] to validate the inference capability for TAs.

Training deepAOT Models. The approach for training deepAOT models includes hyperparameter optimization methods and gradient descent optimization algorithms.

Hyperparameter Optimization. Deep learning is a dramatic improvement in many fields,[28] in particular for CNNs,[56−58] which are often able to automatically learn useful features with little manual intervention of data through multiple layers of abstraction. However, these successes do not detract from the advantages of hyperparameter optimization. An appropriate set of hyperparameters must be selected before applying deep learning architecture for a new data set, which is a time-consuming and tedious task.[59] The hyperparameters of MGE-CNN include the length of fingerprint (FPL), the depth of fingerprint (FPD), the width of convolution kernel (CKW), the size of hidden units in the output layer (HLS), the L2 penalty of cost function (L2P), the scale of initial weights (IWS), and the step size of learning rate (LRS). The ranges of these parameters are shown in Table S1, as recommended by Duvenaud et al. (github.com/HIPS/neural-fingerprint/issues/2) In order to reduce computational costs, a simplified parameter range was used as follows: FPL ∈ {16, 32, 48, 64, 80, 96, 112, 128}; FPD ∈ {1, 2, 3, 4}; CKW ∈ {5, 10, 15, 20}; HLS ∈ {50, 60, 70, 80, 90, 100}; log(L2P) ∈ {−6, −5, −4, −3, −2, −1, 0, 1, 2}; log(IWS) ∈ {−6, −5, −4, −3, −2}; log(LRS) ∈ {−8, −7, −6, −5, −4}.

Usually, the three most popularly used methods for hyperparameter optimization are manual search, grid search, and random search. Of these methods, random search was demonstrated to outperform a combination of manual and grid search when applied to a set of problems.[60] Therefore, random search was used to generate 500 sets of hyperparameters for RMs and CMs and all hyperparameter sets were evaluated with the validation set (2045 compounds, Table 1). The top 10 models were then applied to the next step in selecting the model with the lowest root-mean-square error (RMSE) for RMs, eventually selecting models with the highest accuracy (ACC) for MCMs.

Table 1. Statistical Description of the Training, Validation, and External Test Sets

category	I	II	III	IV	total
training set	794	1933	4303	1050	8080
validation set	224	463	1155	203	2045
test set I	92	341	1099	141	1673
test set II	57	93	183	42	375
total	1167	2830	6740	1436	12173

Gradient Descent Optimization. Gradient descent is one of the most popular algorithms to optimize deep learning-based networks. Every state-of-the-art deep learning library contains implementations of various algorithms to optimize gradient descent.[61] Adaptive Moment Estimation (Adam)[62] is a popular method that computes adaptive learning rates for each weight. It takes an exponentially decaying average of past gradients and past squared gradients into consideration and demonstrates empirically that Adam works well for adaptive learning-method algorithms. The shuffle of training set after each epoch was also applied to the training process for avoiding bias of the optimization algorithm. Therefore, the training strategy was implemented by a pseudocode of Algorithm 2 in the Supporting Information.

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \alpha\|\theta\|_2 \tag{2}$$

$$J(\theta) = -\frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} 1\{y^{(i)} = j\}\log\frac{e^{\theta_j^{T}x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^{T}x^{(i)}}}\right] + \alpha\|\theta\|_2 \tag{3}$$

where $J(\theta)$ is the loss function added L2 penalty described in eqs 2 and 3, which were used to evaluate RMs and MCMs, respectively. A flexible automatic differentiation package called Autograd (https://github.com/HIPS/autograd) was easily adopted for computing gradients of weights.

Experimental Setup. *Data Collection and Preparation.* The AOT database provided by Li et al.,[15] the largest data set for oral LD_{50} in rat, was used in this study. All data was from

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685

three sources: (1) the admetSAR database;[63] (2) the MDL Toxicity Database (version 2004.1),[64] and (3) the Toxicity Estimation Software Tool (TEST version 4.1)[65] program from the U.S. EPA. The preparation of the data set had been executed by Li et al.[15] The "Structure Checker" and "Standardizer" modules from ChemAxon Inc. (evaluation version)[66] were used to fix some error valence and standardize all the SMILEs strings in the data sets. The workflow is shown in Figure S2. Finally, the training and validation sets included 8080 and 2045 compounds, respectively, with measured LD_{50} values adopted from the admetSAR database. Two external data sets contained 1673 (from MDL Toxicity Database) and 375 (from TEST) compounds. Based on the U.S. EPA definition of toxicity,[67] all compounds were divided into four categories based on their levels of toxicity. The statistical description of the entire data set is shown in Table 1. The entire data set was consistent with observations made by Li et al.'s (training set 8102; validation set 2049; test set I 1678; test set II 375). Test set II only had category labels without exact experimental values of acute oral LD_{50}.

Construction Strategy of RMs and MCMs. RMs and MCMs were constructed by MGE-CNN. For RMs, the training target was a $\log(LD_{50})$ (unit: log(mg/kg)) value for each compound. The loss function of eq 2 was adopted in the MGE-CNN. In order to select appropriate sets of hyperparameters, each set of 500 random combinations was run for 750 iterations with a mini-batch gradient descent and Adam optimization algorithm. We selected the top 10 sets of hyperparameters with lowest RMSE values of the validation set. Generally, the purpose of 10 well-trained models is to quantitatively predict $\log(LD_{50})$ of unknown compounds. Therefore, the 10 models needed to be challenged by an external data set (test set I) (note: test set II lacks the LD_{50} values). The consensus RM (deepAOT-R) was constructed with averaging the previous 10 models and the classification capacity of the deepAOT-R model was estimated and analyzed.

For MCMs, the training target was a defined label of compound toxicity. According to the category criterion, four categories also meant four outputs in the MGE-CNN architecture. The *softmax* loss function (eq 3) was used as the object function for MCMs. Initially, each of the 500 random sets of hyperparameters was run for 1000 iterations to select the top 10 sets with highest ACC of the validation set. Next, the top 10 models were run for an additional 1000 iterations. Finally, the best-trained weights were selected out with the highest ACC of the validation set. Consequently, the best 10 MCMs were challenged by the two external test sets. Meanwhile, the consistency between MCMs and RMs was analyzed according to their prediction outcomes.

Construction Strategy of Multitask Model. The multitask model (deepAOT-CR) was constructed for simultaneous prediction of regression and classification problems, and help to improve the consistency of RMs and MCMs. It also could be applied by other multiproperty prediction task. Considering the inconsistency for the scale of loss function between regression and classification problems, the modified cost function is as follows.

$$J(\theta) = J_C(\theta) + \beta J_R(\theta) + \alpha\|\theta\|_2$$

Here, $J_C(\theta)$ and $J_R(\theta)$ are the loss of classification task and regression task, respectively. $\beta \in (0, 1]$ is another weight parameter to be trained, setting the initial value to 1. The Autograd toolkit was used to compute the gradients of β. The weight updating was implemented by Adam algorithm with the smaller learning rate. The hyperparameters from the best MCM were straightly adopted in the deepAOT-CR to simplify multitask training. The consistency of regression and classification outcomes was analyzed.

Forward and Backward Exploration of Fingerprints. In order to determine what these models actually predict, the forward and backward exploration approach was applied for the "Fingerprint" layer. The forward exploration was implemented by extracting the values of the Fingerprint layer (deep fingerprints) to construct MLR and SVM models. This could demonstrate the support degree that these features provided in the shallow machine learning decision-making system. The performance of shallow models with deep fingerprints was accessed and, then, compared with previous reported models using application-specific fingerprints or descriptors.

The goal of backward exploration is to explore the most prominent part of the |FPL|-dimensional Fingerprint contributing to AOT. Once the |FPL|-dimensional vectors were extracted, linear regression operation with the known log LD_{50} value was adopted to determine the most negatively correlated bit on the fingerprint. The biggest activation value for this key bit could be mapped into the fragment space. The activation value was actually corresponding to the activation degree of an fragment centroid. And the depth of the encoding layer where this value is located was used to determine the size of the fragment. The activation fragments were highlighted in a drawing of each compound presented in category I (794 compounds). These highlighted fragments were considered by prediction models to be substructures most related to AOT, which was an inference to toxicity fragments. Meanwhile, these fragments were used to make comparisons with the reported structural features from ToxAlerts[55] for validating the inference capability of MGE-CNN-based deepAOT models.

Evaluation Metrics. All of the models were evaluated using the validation set, then challenged by two external test sets. The three indexes of root-mean-square error (RMSE, eq 4), mean absolute error (MAE, eq 5), and square of Pearson correlation coefficient (PCC^2, eq 6) were used as evaluation indexes for the RMs. The MCMs were assessed in accordance with the multiclass confusion matrix, in which the sensitivity (Sens), positive predictive value (PPV), and ACC were calculated as shown in Figure 1C. In addition, the consensus deepAOT-R model was used to assess classification performance. The PCC is a description of linear correlation and a regression line estimates the average value of target y for each value of input X, but actual y values differ from the predicted values $\hat{y}$ unless the correlation is perfect. These differences are called prediction errors or residuals, which means that it is reasonable and valuable for a predicted value accomplished by a wiggle room to judge this prediction. Thus, 1-fold RMSE for the validation set was added into the outcomes of RMs. For the two external test sets, deepAOT-R predicted the output values, which were then mapped into the category space and transformed into the output labels. The ranges of output labels were calculated with the output values within 1 RMSE. Assuming that the range of a predicted label contains the actual target label, this prediction was considered to be correct.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (4)$$

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685

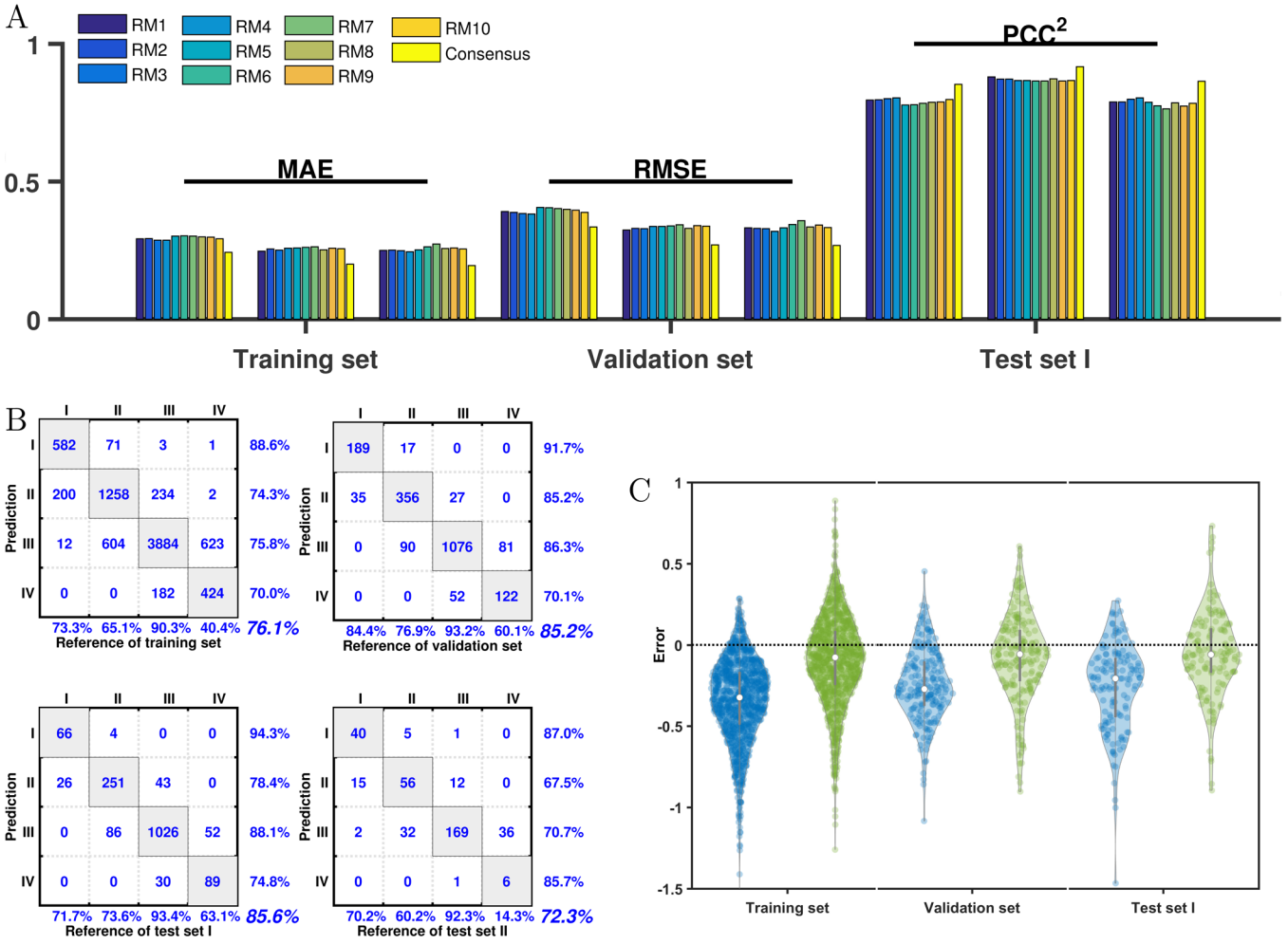


Figure 2. Performance overview of the top 10 RMs and the consensus deepAOT-R model. (A) Overview of MAE, RMSE, and PCC^2 index for all the RMs. (B) Confusion matrix for assessing deepAOT-R's classification capacity. (C) Distribution comparison of regression prediction errors from category IV. Blue: deepAOT-R. Green: deepAOT-CR.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{5}$$

$$\mathrm{PCC}^2 = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}\right]^2 \tag{6}$$

RESULTS AND DISCUSSION

Performance Evaluation of RMs. The RMs help to quantitatively predict the $\log(LD_{50})$ values in rat for compounds, reflecting their toxicity: the smaller the value, the more toxic the compound. The 500 random sets of hyperparameters were fed into the MGE-CNN architecture and those 500 models were trained with different hyperparameters for 750 iterations to construct the RMs.

The RMSE and PCC indexes of the training and validation sets from 500 models after gradient-based optimization training were shown in Figure S3. For the training and validation sets, decreased RMSE was accompanied by a progressive increase of PCC, which completely conformed to the logical law of gradient descent. The three indexes of RMSE, MAE and PCC^2 over 500 models with different hyperparameters had a wide range of changes and the whole performance of the top 10 RMs is shown in Table S2 and Figure 2A, in which MAE, RMSE, and PCC^2 on the three sets are described. Preference analysis for hyperparameters is shown in Figure S4. Among the 10 RMs, RM4 had the best MAE (0.287), RMSE (0.382), PCC^2 (0.804) for the training set, but a suboptimal performance for the validation set (MAE of 0.258, RMSE of 0.337, PCC^2 of 0.867). For test set I, RM4 also has the optimal performance of 0.245 for MAE, 0.319 for RMSE, 0.804 for PCC^2. The consensus outcomes display a further improvement of the three indexes for the three data sets. For example, PCC^2 was 0.853 for the training set (with a 0.049 increase), 0.917 for the validation set (with a 0.037 increase), and 0.864 for test set I (with a 0.060 increase). These deepAOT-R outcomes outperformed the consensus model from Lei et al.[14] (0.487 for MAE, 0.646 for RMSE, 0.690 for PCC^2). The distribution of prediction errors (prediction targets) for the three sets is shown in Figure S5, which was a reasonable distribution for training and prediction results. Therefore, it was necessary for the MGE-CNN architecture to optimize hyperparameters, which would help to boost the performance. Moreover, the ensemble strategy demonstrated that the deepAOT-R had the optimal performance.

In order to investigate classification abilities of the RMs, the consensus model, deepAOT-R, was used to predict the toxicity labels for all of the data sets (test set II had toxicity labels, but

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685

Table 2. Hyperparameters and Performances of the Top 10 Multiclassification Models[a]

model	CM1	CM2	CM3	CM4	CM5	CM6	CM7	CM8	CM9	CM10
Hyperparameter										
FPL	80	128	128	48	112	48	16	128	16	112
FPD	4	2	3	2	3	3	3	4	3	2
CKW	20	20	20	20	20	20	15	15	20	15
HLS	90	60	90	70	50	60	50	50	70	90
log(L2P)	−4	−5	−3	−3	−4	−2	−4	−6	−3	−5
log(IWS)	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
log(LRS)	−4	−4	−4	−4	−4	−5	−4	−5	−5	−4
Evaluation Index[a]										
preTrainACC	0.902	0.866	0.902	0.802	0.891	0.788	0.764	0.810	0.768	0.810
preValACC	0.940	0.914	0.942	0.869	0.941	0.881	0.845	0.883	0.839	0.880
TrainACC	0.921	0.920	0.908	0.841	0.934	0.802	0.790	0.855	0.812	0.881
ValACC	0.958	0.959	0.942	0.905	0.963	0.891	0.866	0.922	0.887	0.931
TestIACC	0.955	0.958	0.953	0.914	0.965	0.886	0.881	0.911	0.897	0.950
TestIIACC	0.963	0.928	0.965	0.851	0.947	0.811	0.816	0.901	0.835	0.883

[a]The abbreviations preTrainACC and preValACC represent the pretraining ACC of the training and validation sets. TrainACC, ValACC, TestIACC, and TestIIACC stand for the ACC predicted by the models on the training, validation, test I, and test II sets.

lacked LD_{50} values). The predicted values $\log(LD_{50})$ were transformed into LD_{50} values and mapped into category space, then the multiclass confusion matrix is summarized in Figure 2B, where the Sens, PPV, and ACC index for each class are shown at the bottom of the box, the right of the box, and as a number in the bottom right corner, respectively. The overall performance was at an acceptable level, although there were poor levels among the four sets when examining the Sens IV index, dividing the compounds with category IV into category III, which suggested that deepAOT-R could not distinguish well between categories III and IV. The prediction error distribution of category IV is presented in Figure 2C (in blue), suggesting that most prediction errors of category IV were lower than zero and might cause such phenomena. However, when the 1-fold RMSE of validation set (ValRMSE, 0.270) wiggle room was taken into consideration, the classification performance significantly improved (Figure S6), which revealed that the deepAOT-R outcomes were still relatively close to the actual target values. Hence, deepAOT-R had a certain distinguishing power of classification, indicating that a wiggle room of 1-fold ValRMSE could be useful for prediction results.

Performance Evaluation of MCMs. The MCM, as a semiquantitative description for AOT, is more intuitive in toxicity estimation than the more simplistic numbers predicted by RMs, which creates difficulty in understanding chemical toxicity.

In order to develop high-level MCMs, the 500 random sets of hyperparameters were set in the MGE-CNN, as were the 500 models with different topological networks that were pretrained with 1000 iterations. Preference analysis for hyperparameters is described in Figure S7. After pretraining, the top 10 models were selected with the highest ACC of the validation set (Table 2). Of these, different sets of hyperparameters resulted in large differences on ACC of the validation set (83.9−94.2%). After the next 1000 iterations were finished, all of the 10 sets of well-trained weights were selected and stored for external predictions. The satisfactory results are displayed in the rows of "TrainACC", "ValACC", "TestIACC" and "TestIIACC" of Table 2. Of these values, ACC in the validation set was between 86.6 and 96.3%, while the ACC range for the two test sets were from 81.1 to 96.5%. The CM1 (deepAOT-C) had the best external prediction ability (with fewer feature dimension) for test set I (ACC of 95.5%) and test set II (ACC of 96.3%) among the 10 models. The confusion matrix of deepAOT-C is portrayed in Figure 3A. The high Sens and PPV index for each class and the high ACC demonstrated that deepAOT-C performed better than the previously reported MCM of Li et al.[15] for the validation set (ACC of 83.2%) and the two external test sets (ACC of 83.0%, ACC 0f 89.9%, respectively). These data indicate that deepAOT-C has an excellent generalization ability. In addition, it is suggestive that the MGE-CNN architecture could be successfully extended to multiclassification problems.

Performance Evaluation of Multitask Models. As for multitask model, the comparable performance of deepAOT-CR with that of deepAOT-C and deepAOT-R is shown in Figure 3D and Figure S8. Although it is slightly lower than the single-task deepAOT-C and deepAOT-R, deepAOT-CR was demonstrated to outperform each of all the single models (shown in Table S2) for regression task. More importantly, based on weight share, it could be used for simultaneous predictions of the classification and regression tasks, which suggested that it was appropriate for the MGE-CNN architecture to achieve multitask problems.

Consistency Analysis of RMs and MCMs. In order to examine the consistency between RMs and MCMs, the deepAOT-R and deepAOT-C were analyzed together. The outcomes of deepAOT-R were assigned to the category space. The consistent prediction outcomes of both models was counted for each data set (Figure 3B). For the consistent prediction, the percentages on the four data sets were 76.8%, 85.2%, 86.1%, and 71.7%, respectively. The accurate classification prediction of deepAOT-R was 76.1%, 85.2%, 85.6% and 72.3%, respectively. Meanwhile, the consistent and accurate predictions respectively occupied 72.8%, 83.2%, 83.7%, and 70.1% for each data set. Such comparisons suggested that most of the consistent predictions were corresponded to correct labels, which meant there was a high consistency between the deepAOT-R and deepAOT-C. For the deepAOT-CR, the consistent outcomes of regression and classification were 82.6%, 83.1%, 84.1%, and 84.5%, respectively, which improve the overall consistency for the four data sets. The consistent and accurate predictions respectively occupied 77.9%, 79.9%, 80.6%, and 80.8% for each data set, shown in Figure 3B. From

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685

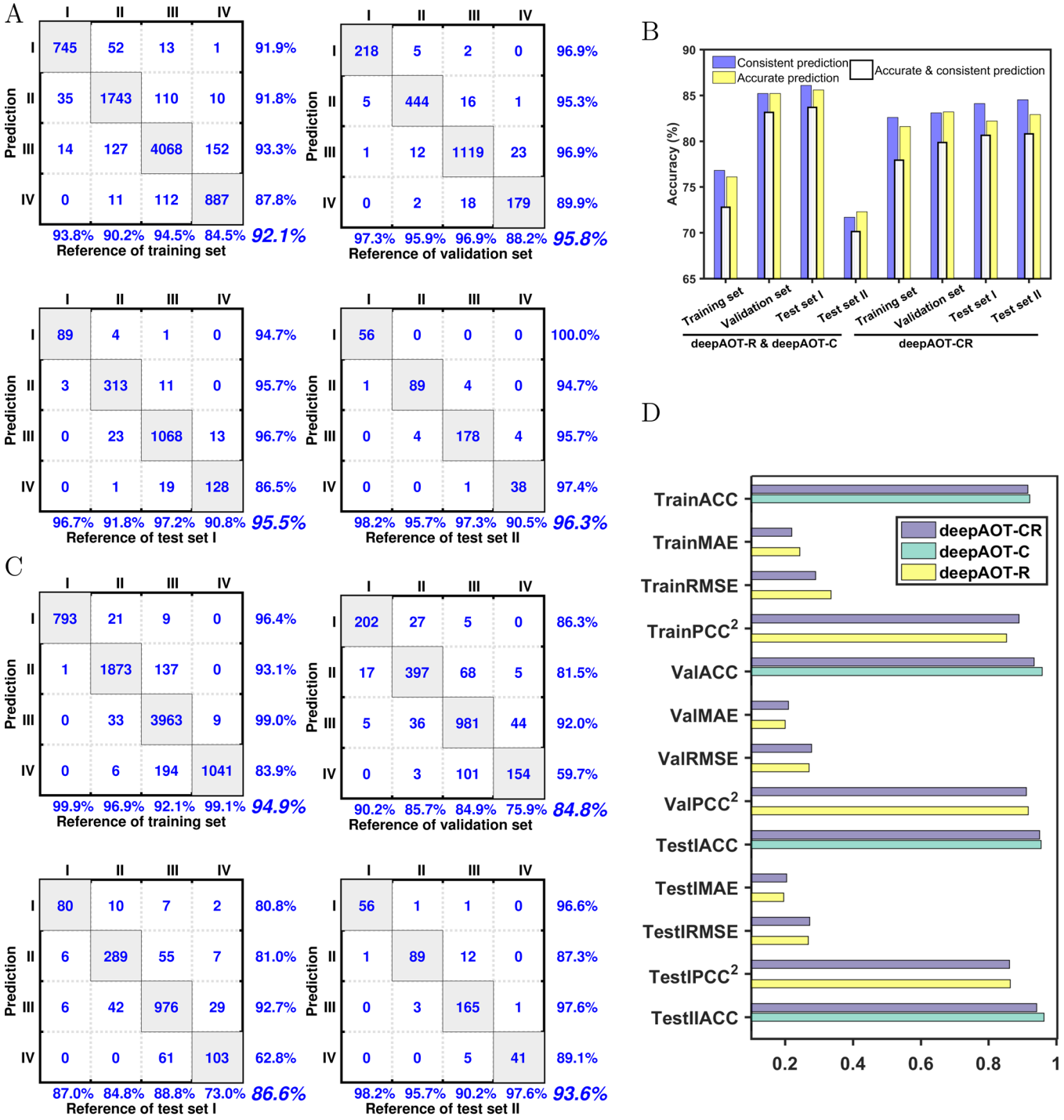


Figure 3. (A) Confusion matrix of deepAOT-C. (B) Consistency comparison between deepAOT-R, deepAOT-C, and deepAOT-CR. (C) Confusion matrix for SVM_CM1, which is an SVM model with deep fingerprints from CM1. (D) Performance comparison of deepAOT-CR, deepAOT-R, and deepAOT-C. TrainACC, ValACC, TestIACC, and TestIIACC mean the ACC index of the training, validation, test I, and test II sets, respectively. Different suffixes represents different indicators.

the view of Figure 2C (green) and Figure S9, deepAOT-CR could significantly (p-value of paired *t* test <0.001) improve the distinguishing capability for category IV.

Forward Exploration of Fingerprints from RMs. The forward exploration evaluated the extent by which the fingerprints from the MGE-CNN-based models favored of shallow decision making systems, such as MLR and SVM. For this purpose, fingerprints were extracted from the Fingerprint layer in the well-trained deep models, and then the whole data set was transferred into a matrix of N (number of compounds) × FPL, which was a featurization and vectorization process for compounds. This operation was executed for both RMs and MCMs. For the RM4, the matrix for the training set, 8080 (compounds) × 48 (features), was regarded as an input for MLR, fitting the target values of $\log(LD_{50})$ by minimizing the sum of the squares of the vertical deviations from each data point to the best-fitting line. The best-fitting line for the training set was calculated, and was used to predict the

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685

Table 3. Performance of MLR Models with Deep Fingerprints from MGE-CNN Architecture on the Training, Validation, and Test I Sets

model[a]	evaluation index[a]					
	TrainMAE	TrainRMSE	TrainPCC^2	Val&TestIMAE	Val&TestIRMSE	Val&TestIPCC^2
MLR_RM1	0.428	0.558	0.580	0.404	0.542	0.584
MLR_RM2	0.425	0.556	0.583	0.402	0.538	0.592
MLR_RM3	0.418	0.544	0.600	0.403	0.528	0.606
MLR_RM4	0.432	0.563	0.572	0.405	0.535	0.596
MLR_RM5	0.410	0.538	0.610	0.397	0.524	0.614
MLR_RM6	0.442	0.578	0.549	0.427	0.561	0.554
MLR_RM7	0.437	0.569	0.563	0.415	0.553	0.566
MLR_RM8	0.402	0.523	0.631	0.378	0.499	0.650
MLR_RM9	0.422	0.548	0.595	0.398	0.521	0.614
MLR_RM10	0.432	0.561	0.575	0.414	0.542	0.583
consensus	0.379	0.497	0.679	0.348	0.465	0.696

[a]MLR_RM*i* means the MLR model constructed by deep fingerprints from RM*i*, $i \in \{1, 2, ..., 10\}$. "Consensus" means the average outcomes of the above 10 models. Val&TestIMAE, Val&TestIRMSE, and Val&TestIPCC^2 are MAE, RMSE, and PCC^2 of the merged validation and test I set.

validation and test I sets (total of 3718 compounds). Performance of the MLR models with deep fingerprints are summarized in Table 3. In which, the MAE, RMSE and PCC^2 were calculated for the training set and the validation and test I sets. The MAE and RMSE range for the validation and test I sets was from 0.378 to 0.427 and from 0.499 to 0.561, respectively, while the PCC^2 was in the range of 0.554−0.650. The consensus model also demonstrated significant improvement for the training and external test sets, and the performances of MAE, RMSE, and PCC^2 were 0.348, 0.465, and 0.696, respectively. These prediction levels are completely acceptable for a MLR method. When the LLR (which was an improved MLR method) reported by Lu et al.[16] was challenged by "Set_3874", the PCC^2 and MAE of the consensus model (with different molecular fingerprints: ECFP4, FCFP4,[38] MACCS, and physicochemical descriptors from commercial software[68,69]) were 0.608 and 0.420, respectively (Figure S10). A pure MLR method based on deep fingerprints was used to ensure that PCC^2 and MAE would stay in a range of 0.554−0.650 and 0.378−0.427, respectively. Comparing the two, whether for a single model or the consensus model, the MLR models outperformed LLR models at a similar level test set size, which revealed that deep fingerprints were more useful than application-specific molecule descriptors or fingerprints for AOT prediction without an idea of "Clustering first, and then modeling".[70]

Forward Exploration of Fingerprints from MCMs. For the MCMs, fingerprints were also extracted, and the training part was used to construct multiclass SVM$_{OAO}$) models with the "scikitlearn" package[71] in Python 2.7. The Gaussian radial basis function kernel was used and the parameters C and γ were tuned with the validation set. The performance of SVM$_{OAO}$ models with deep fingerprints was assessed with ACC index (Table 4). The range for the training set was from 84.2 to 96.5% and the validation range was between 78.7 and 84.8%. For the two external sets, an acceptable ACC range is from 77.9 to 94.9%. Among the SVM models, SVM_CM1 had the best ACC of 94.9% for the training set, 84.8% for the validation set, 86.6% for test set I and 93.6% for test set II. Meanwhile, the confusion matrix for SVM_CM1 indicated that the three indexes of SVM_CM1 were better than those of SVM models developed by Li et al.,[15] shown in Figure 3C and S11. Therefore, deep fingerprints from MGE-CNN-based RMs and MCMs were better than standard fingerprints, which further

Table 4. Performance of SVM$_{OAO}$ Models with Deep Fingerprints from MGE-CNN Architecture

model[a]	evaluation index			
	TrainACC	ValACC	TestIACC	TestIIACC
SVM_CM1	0.949	0.848	0.866	0.936
SVM_CM2	0.934	0.816	0.799	0.880
SVM_CM3	0.950	0.840	0.849	0.912
SVM_CM4	0.959	0.800	0.793	0.885
SVM_CM5	0.958	0.839	0.830	0.928
SVM_CM6	0.942	0.812	0.806	0.891
SVM_CM7	0.857	0.787	0.779	0.840
SVM_CM8	0.988	0.809	0.805	0.949
SVM_CM9	0.847	0.801	0.780	0.837
SVM_CM10	0.965	0.791	0.780	0.909

[a]SVM_CM*i* means the SVM$_{OAO}$ model based on deep fingerprints from CM*i*, $i \in \{1, 2, ..., 10\}$.

demonstrated that the MGE-CNN implemented better MRs for AOT prediction with automatic feature extraction through supervised learning. With analysis of Tanimoto distance, Table 5 suggested that deep fingerprints had a high correlation to molecular topological structure-based ECFP4, FCFP4, and MACCS fingerprints and were different from randomly generated fingerprints. To a certain extent, it revealed the interpretability and rationality of these deep fingerprints.

Backward Exploration of Fingerprints. The backward exploration of the Fingerprint layer was expected to provide an understanding of fingerprint activation.

Herein, only the above RM4 and CM1 was further examined. After linear regression, the most negative correlation feature of the fingerprints was calculated, which represented the most toxic feature. Comparing activation values of this feature, nine values were determined to contribute most to feature activation. The nine values could be mapped into different substructures, thereby suggesting that these substructures were the most correlative to the explored toxicity feature (Figure 4A and B). There were mainly two classes of highlighted fragments, nitriles (TA1190, 42/794) and alyl (thio)phosphates (TA776, 185/794) for RM4, while TA776 and thicarbonyl (TA374, 21/794) for CM1. The three fragments have been reported to be TAs.[55]

Further analysis of RM4 and CM1 explored the highlighted fragments for each compound in category I, followed by the

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685

Table 5. Correlation Analysis of Tanimoto Distance between Different Fingerprints

fingerprint	random	ECFP4	FCFP4	MACCS	DeepAOT	PCC
correlation of tanimoto distance	+	+				0.193 ± 0.014
	+		+			0.264 ± 0.030
	+			+		0.237 ± 0.021
	+				+	**0.212 ± 0.030**
		+	+			0.984
		+		+		0.952
		+			+	**0.845**
			+	+		0.946
			+		+	**0.834**
				+	+	**0.867**

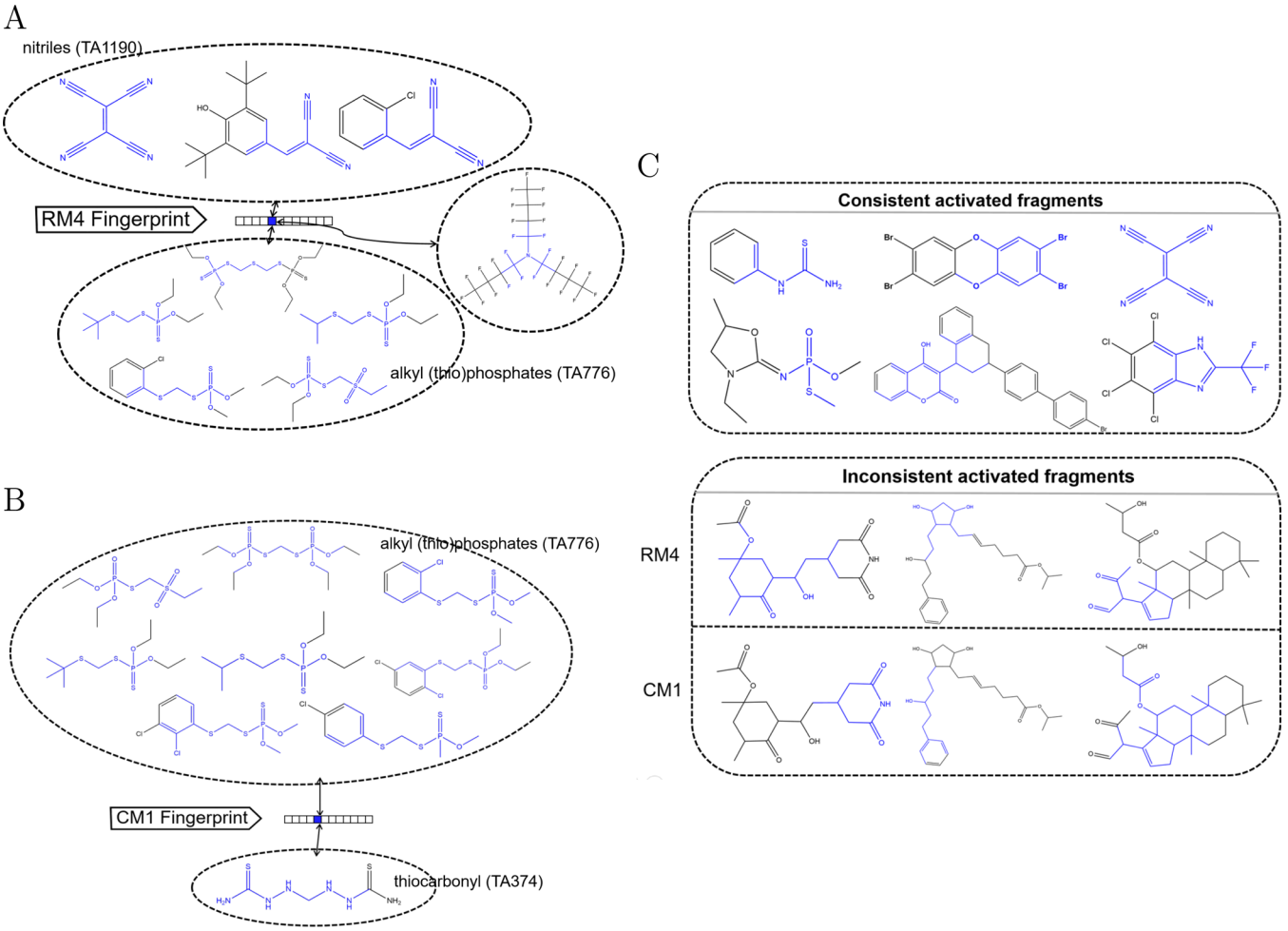
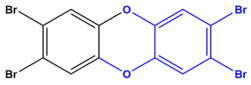
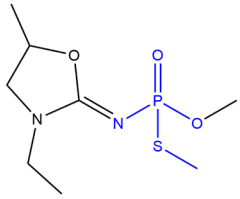
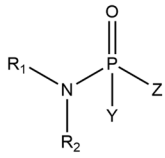
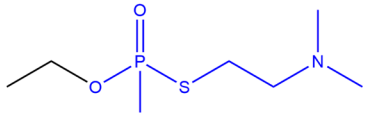
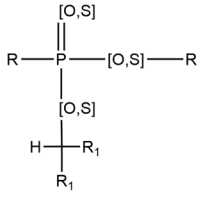
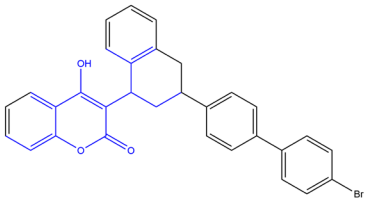
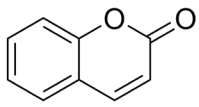
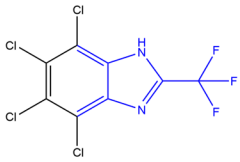
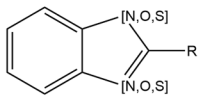
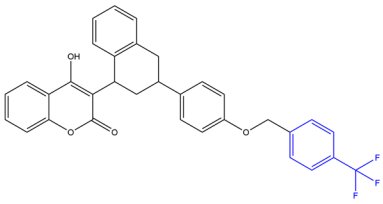
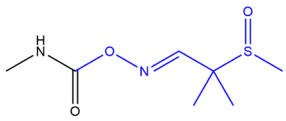
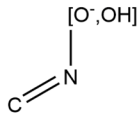
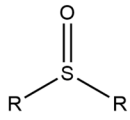


Figure 4. Overview of highlighted fragments. (A and B) Corresponding highlighted fragments that match the most toxic features (blue) of the RM4 and CM1 fingerprint. TA626, TA776, and TA374 are the registered numbers from the Online Chemical Database. (C) Consistency comparison of part of the highlighted (blue) fragments for RM4 and CM1.

approach demonstrated in Figure S12. The maximum contribution term to toxicity on the MLR model would be directly corresponding to a structural fragment, which was thought as a toxicity fragment inferred by deepAOT models. Table 6 and Table S3 describe the highlighted fragments which RM4 automatically generate. Moreover, we found most of the highlighted fragments could correspond to the reported TAs. For example, some of the corresponding reported TAs with structural diversity were TA636, TA1285, TA777, TA389, TA583, TA2815, TA1792, TA646, TA1199, TA321, TA279, TA1146, TA1190, TA584, TA374, and TA362. Due to the high consistency with the reported TA patterns, this approach had potential for inferring TAs for unknown compounds. For CM1, the activation fragments of each compound was almost similar to that from RM4, part of which shown in Figure 4C (in which some inconsistent highlighted fragments are also presented). With frequency analysis on the training set (shown in Table S4), these patterns with high frequency of category I and low frequency of category IV were suggested to be meaningful toxicity alerts for AOT. In addition to AOT prediction, MGE-CNN-based models was also able to automatically learn meaningful toxicity regions with analysis of internal activation

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685

Table 6. Comparison of TAs and Activity Fragments Inferred by RM4[a]

No.	Activation Fragment	Structural Alert	Alert ID (#)
1		 X = Br, F; Ar=any aromatic atom	TA636 (49)
2		 R1, R2 = any atom; Y, Z = any O, N, S, Hal residue;	TA1285 (20)
3		 R = any carbon atom; R1 = H or any carbon atom;	TA777 (54)
4			TA389 (12)
5		 R = C,N,O,S,Ar; Ar = any aromatic atom;	TA583 (68)
6		 R = any atom or group;	TA2815 (77)
7			TA1792 (71)
		 R = any atom;	TA646 (12)

[a]"#" indicates the number of compounds containing the TA among the class I training set.

layers. These physical meaningful substructures could not only be used to explore the mechanism of AOT, but also help to filter out toxicity-related molecules for drug discovery and development.

CONCLUSION

In this study, RMs and MCMs constructed by the MGE-CNN were used to estimate AOT in rat for chemical safety assessment. The consensus deepAOT-R model had an outstanding performance with higher PCC^2 (0.864), lower RMSE (0.268), and lower MAE (0.195) than the previous best models. When using the deepAOT-R to predict toxicity category, the performance was within an acceptable level and the recommended range (within 1-fold RMSE) may be important for prediction outcomes. In addition, the deepAOT-C also demonstrated an excellent performance (ACC of validation set 92.1%; ACC of test set I 95.8%; ACC of test set II 96.3%) when compared to the best reported MCMs. In addition, the multitask deepAOT-CR also presented a high

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685

predictive power for simultaneous assessment of regression and classification problems.

In consideration of high-level prediction models, more attention was focused on exploring and interpreting our models. In fact, deep fingerprints extracted from the RMs and the MCMs were able to better support the shallow decision making systems than application-specific molecular descriptors or fingerprints. The consensus MLR model based on these deep fingerprints had a high PCC^2 (0.696) and a low MAE (0.348) for the large external set (3718 compounds). Meanwhile, the best SVM model with deep fingerprints also performed very well, with ValACC of 84.8%, TestIACC of 86.6%, and TestIIACC of 93.6%. With correlation analysis of Tanimoto distance, we recognized that these deep fingerprints were highly correlated with topological structure-based fingerprints. The successes of deep fingerprints could potentially be applied to other tasks related to AOT. One toxicity-related feature of these fingerprints was tracked back to the atomic level and the highlighted toxicity fragments inferred by RM4 and CM1 were compared with the corresponding TAs. This surprising consistency suggests that the well-trained deep models are no longer "black" models and that these deep models advanced in AOT-related knowledge such that they can be used to roughly infer TAs. Without prior knowledge about fragments, only the information on atoms and bonds can be used to form the knowledge of fragments, all of which are due to the ability of automatic feature learning from deep learning.

The MGE-CNN is not limited to AOT and it could be applied for studying other end points induced by compounds in complex systems. Without understanding any mechanism, end-to-end (SMILES-to-end point) learning based on a known large data set with high quality can be useful in predicting this end point, extracting the end point-related fingerprints and inferring the end point-related fragments. This methodology is a promising strategy in developing and better understanding chemical information on compounds.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00244.

Tables S1−4, Algorithm 2, and Figures S1−12 (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*Fax: (+86)10-62759595. E-mail: jfpei@pku.edu.cn (J.P.).
*Fax: (+86)10-62751725. E-mail: lhlai@pku.edu.cn (L.L.).

ORCID

Jianfeng Pei: 0000-0002-8482-1185
Luhua Lai: 0000-0002-8343-7587

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Prof. Yun Tang, from the School of Pharmacy, East China University of Science and Technology, for providing the four mentioned datasets of rat oral LD_{50}.The authors also thank Prof. David Kristjanson Duvenaud for providing effective parameter ranges and Shuaishi Gao for web page design of the deepAOT prediction server. The work was partially carried out at Peking University High Performance Computing Platform, and the calculations were performed on CLS-HPC. This research was supported, in part, by the National Natural Science Foundation of China (grant numbers 21673010, 21633001) and the Ministry of Science and Technology of China (grant numbers 2016YFA0502303, 2015CB910302).

■ ABBREVIATIONS

LD_{50}, median lethal death; AOT, acute oral toxicity; CNN, convolution neural network; NN, neural network; RM, regression model; MCM, multiclassification model; QSPR, quantitative structure−toxicity relationship; MLR, multiple linear regression; LR, linear regression; kNN, k nearest neighbors; RF, random forest; SVM, support vector machine; RVM, relevance vector machine; LLL, local lazy learning; PCC^2 (R^2), square of linear correlation coefficient; EPA, Environmental Protection Agency; MR, molecular representations; FPL, the length of fingerprint; FPD, the depth of fingerprint; CKW, the width of convolution kernel; HLS, the size of hidden units in the output layer; L2P, the *L*2 penalty of cost function; IWS, the scale of initial weights; LRS, the step size of learning rate; RMSE, root-mean-square error; ACC, accuracy; MAE, mean absolute error; Sens, sensitivity; PPV, positive predictive value; MGE-CNN, molecular graph encoding with convolutional neural network

■ REFERENCES

(1) Turner, R. Acute Toxicity: The Determination of LD50. *Screening Methods in Pharmacology* **1965**, *300*, 28−40.
(2) Test Guideline 401. *Trends Pharmacol. Sci.* **2001**, *22*.
(3) Enslein, K. A Toxicity Estimation Model. *J. Environ. Pathol. Toxicol.* **1977**, *2*, 115−121.
(4) Enslein, K.; Lander, T. R.; Tomb, M. E.; Craig, P. N. A Predictive Model for Estimating Rat Oral LD50 Values. *Toxicol. Ind. Health* **1989**, *5*, 265.
(5) 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.
(6) Toropov, A. A.; Rasulev, B. F.; Leszczynski, J. QSAR Modeling of Acute Toxicity for Nitrobenzene Derivatives Towards Rats: Comparative Analysis by MLRA and Optimal Descriptors. *QSAR Comb. Sci.* **2007**, *26*, 686−693.
(7) Jean, P. A.; Gallavan, R. H.; Kolesar, G. B.; Siddiqui, W. H.; Oxley, J. A.; Meeks, R. G. Chlorosilane Acute Inhalation Toxicity and Development of an LC50 Prediction Model. *Inhalation Toxicol.* **2006**, *18*, 515−522.
(8) Figueredo, A.; Wolf, P. S. Assortative Pairing and Life History Strategy - a Cross-Cultural Study. *Human Nature* **2009**, *20*, 317−330.
(9) Hamadache, M.; Benkortbi, O.; Hanini, S.; Amrane, A.; Khaouane, L.; Moussa, C. S. A Quantitative Structure Activity Relationship for Acute Oral Toxicity of Pesticides on Rats: Validation, Domain of Application and Prediction. *J. Hazard. Mater.* **2016**, *303*, 28−40.
(10) Hamadache, M.; Hanini, S.; Benkortbi, O.; Amrane, A.; Khaouane, L.; Moussa, C. S. Artificial Neural Network-Based Equation to Predict the Toxicity of Herbicides on Rats. *Chemom. Intell. Lab. Syst.* **2016**, *154*, 7−15.
(11) Zakarya, D.; Larfaoui, E. M.; Boulaamail, A.; Lakhlifi, T. Analysis of Structure-Toxicity Relationships for a Series of Amide Herbicides Using Statistical Methods and Neural Network. *SAR QSAR Environ. Res.* **1996**, *5*, 269−279.
(12) Eldred, D. V.; Jurs, P. C. Prediction of Acute Mammalian Toxicity of Organophosphorus Pesticide Compounds from Molecular Structure. *SAR QSAR Environ. Res.* **1999**, *10*, 75−99.

(13) 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*, 1913−21.
(14) 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*, 6−24.
(15) 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.
(16) Lu, J.; Peng, J.; Wang, J.; Shen, Q.; Bi, Y.; Gong, L.; Zheng, M.; Luo, X.; Zhu, W.; Jiang, H.; Chen, K. Estimation of Acute Oral Toxicity in Rat Using Local Lazy Learning. *J. Cheminf.* **2014**, *6*, 26−36.
(17) Duran-Frigola, M.; Rossell, D.; Aloy, P. A Chemo-Centric View of Human Health and Disease. *Nat. Commun.* **2014**, *5*, 5676−5686.
(18) Klopman, G. Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.
(19) Klopman, G. MULTICASE 1. A hierarchical computer automated structure evaluation program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176−184.
(20) Poroikov, V.; Filimonov, D.; Borodina, Y. V.; Lagunin, A.; Kos, A. Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1349−1355.
(21) Helma, C.; Cramer, T.; Kramer, S.; De Raedt, L. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1402−1411.
(22) Helguera, A. M.; González, M. P.; Cordeiro, M. N. D.; Pérez, M. Á. C. Quantitative Structure Carcinogenicity Relationship for Detecting Structural Alerts in Nitroso-Compounds. *Toxicol. Appl. Pharmacol.* **2007**, *221*, 189−202.
(23) Cortez, P.; Embrechts, M. J. Using Sensitivity Analysis and Visualization Techniques to Open Black Box Data Mining Models. *Inf. Sci.* **2013**, *225*, 1−17.
(24) Liu, R.; Yu, X.; Wallqvist, A. Data-Driven Identification of Structural Alerts for Mitigating the Risk of Drug-Induced Human Liver Injuries. *J. Cheminf.* **2015**, *7*, 4−11.
(25) Yang, H.; Li, J.; Wu, Z.; Li, W.; Liu, G.; Tang, Y. Evaluation of Different Methods for Identification of Structural Alerts Using Chemical Ames Mutagenicity Data Set As a Benchmark. *Chem. Res. Toxicol.* **2017**, *30*, 1355−1364.
(26) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; John Wiley & Sons, 2009; Vol. *41*.
(27) Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.* **2012**, *7*, 1399−1409.
(28) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436−444.
(29) Gawehn, E.; Hiss, J. A.; Schneider, G. Deep Learning in Drug Discovery. *Mol. Inf.* **2016**, *35*, 3−14.
(30) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*, 80−94.
(31) Graves, A.; Wayne, G.; Danihelka, I. Neural Turing Machines. *arXiv.org* **2014**, No. arXiv:1410.5401.
(32) Weston, J.; Chopra, S.; Bordes, A. Memory Networks. *arXiv.org* **2014**, No. arXiv:1410.3916.
(33) Grefenstette, E.; Hermann, K. M.; Suleyman, M.; Blunsom, P. Learning to transduce with unbounded memory. *NIPS* **2015**, 1828−1836.
(34) Joulin, A.; Mikolov, T. Inferring algorithmic patterns with stack-augmented recurrent nets. *NIPS* **2015**, 190−198.
(35) Neelakantan, A.; Le, Q. V.; Sutskever, I. Neural Programmer: Inducing Latent Programs with Gradient Descent. *arXiv.org* **2015**, No. arXiv:1511.04834.
(36) Reed, S. E.; de Freitas, N. Neural Programmer-Interpreters. *arXiv.org* **2015**, No. arXiv:1511.06279.
(37) Rocktäschel, T.; Riedel, S. End-to-end Differentiable Proving. *arXiv.org* **2017**, No. arXiv:1705.11040.
(38) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(39) 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.
(40) Baldi, P. The inner and outer approaches to the design of recursive neural architectures. *Data Min. Knowl. Disc.* **2017**, 1−13.
(41) 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.
(42) 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.
(43) 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. *NIPS* **2015**, 2224−2232.
(44) 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.
(45) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757−1772.
(46) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.
(47) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241−2251.
(48) Gamo, F.-J.; Sanz, L. M.; Vidal, J.; de Cozar, C.; Alvarez, E.; Lavandera, J.-L.; Vanderwall, D. E.; Green, D. V.; Kumar, V.; Hasan, S.; et al. Thousands of Chemical Starting Points for Antimalarial Lead Identification. *Nature* **2010**, *465*, 305−310.
(49) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49*, 169−184.
(50) Huang, R.; Xia, M. Editorial: Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways As Mediated by Exposure to Environmental Toxicants and Drugs. *Front. Environ. Sci.* **2017**, *5*, 3−5.
(51) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; et al. PubChem's BioAssay Database. *Nucleic Acids Res.* **2012**, *40*, D400−D412.
(52) Zeiler, M. D.; Fergus, R. Visualizing and Understanding Convolutional Networks. *arXiv.org* **2013**, No. arXiv:1311.2901.
(53) Landrum, G. *RDKit: Open-Source Cheminformatics Software*. http://www.rdkit.org (accessed October 21, 2017).
(54) Simard, P. Y.; LeCun, Y. A.; Denker, J. S.; Victorri, B. Transformation Invariance in Pattern Recognition−tangent Distance and Tangent Propagation. *Neural Networks: Tricks of the Trade* **1998**, *1524*, 239−274.
(55) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2310−2316.
(56) Lawrence, S.; Giles, C. L.; Tsoi, A. C.; Back, A. D. Face Recognition: A Convolutional Neural-Network Approach. *IEEE Trans. Neural Networks* **1997**, *8*, 98−113.

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685

(57) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet Classification with Deep Convolutional Neural Networks. *NIPS* **2012**, 1097−1105.
(58) Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-Scale Video Classification with Convolutional Neural Networks. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* **2014**, 1725−1732.
(59) Young, S. R.; Rose, D. C.; Karnowski, T. P.; Lim, S.-H.; Patton, R. M. Optimizing Deep Learning Hyper-Parameters Through an Evolutionary Algorithm. *Proc. Workshop Machine Learn. High-Perform. Comput. Environ.* **2015**, 4−8.
(60) Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn Res.* **2012**, *13*, 281−305.
(61) Ruder, S. An Overview of Gradient Descent Optimization Algorithms. *arXiv.org* **2016**, No. arXiv:1609.04747.
(62) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *An Overview of Gradient Descent Optimization Algorithms* **2014**, No. arXiv:1412.6980.
(63) Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P. W.; Tang, Y. AdmetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties. *J. Chem. Inf. Model.* **2012**, *52*, 3099−3105.
(64) Accelrys Inc., MDL Toxicity Database (presently Accelrys Toxicity Database). http://accelrys.com/products/databases/bioactivity/toxicity.html (accessed Febrary 14th, 2013).
(65) U. S. Environmental Protection Agency,. Quantitative Structure Activity Relationship. https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test (accessed October 21, 2017).
(66) ChemAxon Inc. Standardizer: Structure Canonicalization and Transformation. https://www.chemaxon.com/products/standardizer/ (accessed October 21, 2017).
(67) U. S. Environmental Protection Agency. Precautionary Statements. *Label Review Manual*; 2016; Chapter 7.
(68) Katritzky, A. R.; Lobanov, V. S.; Karelson, M.; Murugan, R.; Grendze, M. P.; Toomey, J. E. Comprehensive Descriptors for Structural and Statistical Analysis. 1: Correlations Between Structure and Physical Properties of Substituted Pyridines. *Rev. Roum. Chim.* **1996**, *41*, 851−867.
(69) Accelrys Software Inc. *Discovery Studio*, version 2.5; 2009.
(70) Yuan, H.; Wang, Y.; Cheng, Y. Local and Global Quantitative Structure-Activity Relationship Modeling and Prediction for the Baseline Toxicity. *J. Chem. Inf. Model.* **2007**, *47*, 159−169.
(71) 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. *J. Mach. Learn Res.* **2011**, *12*, 2825−2830.

DOI: 10.1021/acs.jcim.7b00244
J. Chem. Inf. Model. 2017, 57, 2672−2685