Journal of Chemical Information and Modeling

Article

pubs.acs.org/jcim

Deep Learning for Drug-Induced Liver Injury

Youjun Xu,[†] Ziwei Dai,[†] Fangjin Chen,[†] Shuaishi Gao,[†] Jianfeng Pei,[*,†] and Luhua Lai[*,†,‡,§]

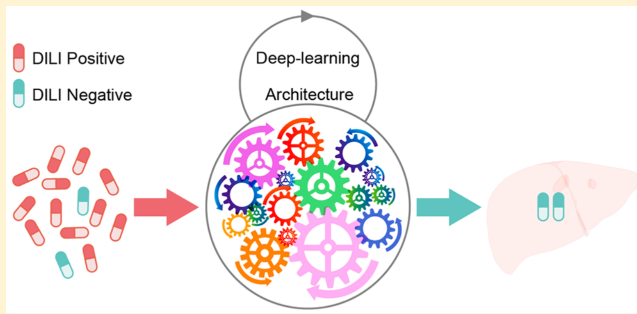
[†]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

[‡]Beijing National Laboratory for Molecular Sciences, State Key Laboratory for Structural Chemistry of Unstable and Stable Species, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

[§]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China

Supporting Information

ABSTRACT: Drug-induced liver injury (DILI) has been the single most frequent cause of safety-related drug marketing withdrawals for the past 50 years. Recently, deep learning (DL) has been successfully applied in many fields due to its exceptional and automatic learning ability. In this study, DILI prediction models were developed using DL architectures, and the best model trained on 475 drugs predicted an external validation set of 198 drugs with an accuracy of 86.9%, sensitivity of 82.5%, specificity of 92.9%, and area under the curve of 0.955, which is better than the performance of previously described DILI prediction models. Furthermore, with deep analysis, we also identified important molecular features that are related to DILI. Such DL models could improve the prediction of DILI risk in humans. The DL DILI prediction models are freely available at http://www.repharma.cn/DILIserver/DILI_home.php.



INTRODUCTION

More than 700 drugs have been found to be associated with liver injury.[1,2] Drug-induced liver injury (DILI) has been the single most frequent cause of safety-related drug marketing withdrawals for the past 50 years (e.g., iproniazid, etc.) continuing to the present (e.g., ticrynafen, benoxaprofen, bromfenac, troglitazone, nefazodone).[3] Drugs that cause severe DILI in humans typically do not show clear hepatotoxicity in animals, do not show dose-related toxicity, and cause low rates of severe injury (1 in 5000 to 10 000 or less).[1] The mechanisms underlying DILI are complicated and diverse and are far from being elucidated, making toxicological studies of DILI difficult.

Recently, *in silico* DILI prediction models based on the molecular structures of drug compounds have become a more convenient and practical approach to predicting DILI.[4] A DILI prediction model described by Cruz-Monteagudo et al., which involved radially distributed molecular descriptors and linear discriminant analysis for classification. It was trained on a 74-drug data set and achieved 82% prediction accuracy for an external test data set of 13 drugs.[5] A later model designed by Rodgers et al. used a K-nearest neighbor method and mixed molecular descriptors. It showed 74% sensitivity and 90% specificity for 37 drugs.[6] Although these two models seemed to perform well, their predictive abilities for large external data sets were never validated. Ekins et al. developed a Bayesian model with extended connectivity fingerprints and other interpretable descriptors based on 295 compounds. It had a 60% predictive accuracy on a test set of 237 compounds.[7] More recently, a model using mixed machine learning algorithms and PaDEL molecular descriptors developed by Liew et al., trained on a large data set of 1087 compounds, showed 63% predictive accuracy on an external test set of 120 compounds.[8] Chen et al. developed a model that used a Decision Forest algorithm[9] and Mold2 chemical descriptors[10] and that was trained on a reliable public data set from the U.S. Food and Drug Administration (FDA). When tested on three external data sets, its accuracies were between 62 and 69%.[4] In 2015, Muller at al. used a combination of theoretically calculated parameters and measured *in vitro* biological data for DILI predictions. Their best model achieved a training balanced accuracy of 88%, correctly identifying 9 out of 10 compounds of the external test set, but with a significantly decreased accuracy of predictions using solely theoretical molecular descriptors.[11] In this way, there remains considerable room for improvement in DILI prediction.

Deep learning (DL) techniques, which were developed based on deep artificial neural networks, have improved the state of the art in the fields of computer vision,[12−17] speech recognition,[18−23] and natural language processing.[24−29] These major breakthroughs made by DL push machine learning closer to one of its original goals, artificial intelligence.[30] A key advantage of deep learning is that features can be learned automatically using a general-purpose procedure. This procedure is usually implemented by a multilayer stack of simple neural networks with nonlinear input−output mappings,

Received: April 27, 2015
Published: October 6, 2015

DOI: 10.1021/acs.jcim.5b00238
J. Chem. Inf. Model. 2015, 55, 2085−2093

Table 1. Summary of Data Sets Used in This Study[a]

	data sets	DILI labels		total number
		DILI-positive	DILI-negative	
Training	NCTR training data set	81	109	190
	Combined training data set	236	239	475
	Liew training data set	648	417	1065
External validation	NCTR validation data set	95	90	185
	Greene data set	209	111	320
	Xu data set	128	108	236
	Combined validation data set	114	84	198
	Liew validation data set	70	49	119

[a]Note: Some of the data sets used in this study differed slightly from the original data sets due to data check.

involving deep neural networks (DNNs),[19,20,24,31] convolutional neural networks (CNNs),[15,32−35] recurrent or recursive neural networks (RNNs),[36] and other deep networks with more than one hidden layer and more neurons in each layer. It has been confirmed that deep-learning architectures have the power to handle big data with little manual intervention.[37,38] These practical and useful techniques also have been applied to chemoinformatics and bioinformatics[39−49] for dealing with different tasks, such as aqueous solubility prediction, quantitative structure−activity relationship analysis,[48] and predicting the sequence specificities of DNA/RNA binding proteins.[50]

Motivated by its great success in these fields mentioned above, we expect that DL is also practical and effective for DILI prediction. To predict molecular properties like DILI, if autoencoder-based[51] and convolutional[52] architectures can be adopted, the coding of given molecules can be represented by vectors of fixed length, referring to molecular fingerprints[53] and descriptors.[54] Although potentially useful for chemoinformatics,[55] these approaches still rely heavily on good encoding function,[44] which is a translation from structural information to vectors. Recently, Lusci et al. developed the novel undirected graph recursive neural networks (UGRNN) method used for molecular structure encoding and used this encoding approach to effectively predict the water solubility of compounds based on DL architectures.[44] One advantage of this UGRNN is that it relies only minimally on the identification of suitable molecular descriptors because suitable representations are learned automatically from the data.[44] In the present study, this UGRNN molecular encoding architectures combined with a line bisection method[56] were used to construct new DILI prediction models. These models (called DL DILI models) were trained and tested on large data sets and showed high performance, indicating their power and potential in terms of predicting DILI in the field of chemoinformatics.

■ MATERIALS AND METHODS

Data Sets. Four publicly available data sets composed of annotated DILI-positive or DILI-negative properties of drugs or compounds were used in this work. The first three data sets were pharmaceutical compound data sets previously used by Chen et al.:[4] (1) a data set from the U.S. FDA's National Center for Toxicological Research (called the NCTR data set);[4,57] (2) a data set from Greene et al.[58] (called the Greene data set), which was used as a validation data set; and (3) and a data set from Xu et al.[59] (called the Xu data set), which was also used as a validation data set. For these three data sets, drugs with a high risk of DILI were labeled "DILI-positive," drugs with no risk of DILI were labeled "DILI-negative," and drugs with a low risk of DILI were not included due to their uncertainty. Though similar DILI labeling criteria were adopted, there were still inconsistencies in annotations between the three data sets (e.g., 17 inconsistent annotations between the NCTR and Greene data sets).[4] The fourth data set was a large data set from Liew et al.[8] (called the Liew data set), which included pharmaceutical and nonpharmaceutical compounds. For the Liew data set, a different labeling method was used: compounds with the potential to cause any adverse liver effects were labeled "DILI-positive," and compounds not associated with any adverse liver effects were labeled "DILI-negative." In this way, the Liew data set was very different from the first three data sets. A combined data set (called the combined data set) was also constructed. It was a combination (with duplication and annotation consistency check) of the NCTR, Greene, and Xu data sets. A summary of the five data sets is shown in Table 1. Details of data processing of the data sets are provided in the Supporting Information.

Molecular Encoding and DL Architecture. UGRNN architecture, which was described in detail by Lusci et al.,[44] was used in our models for molecular structural encoding. Typically, chemical structures of single molecules are represented by small undirected graphs[60] (UGs), in which heavy atoms are represented as nodes and bonds between atoms are represented as edges. Briefly, UGRNN encodes molecular structures from UGs into directed acyclic graphs (DAGs) for use in recursive neural networks[61] (RNN). In UGRNN architecture, to transform a UG into a DAG, each node is traversed sequentially and regarded as a root. Then, all edges are oriented toward the root with the shortest possible path. The architecture of the UGRNN approach is composed of two layers: an encoding layer and an output layer. A brief schematic diagram of UGRNN encoding of glycine is shown in Figure 1. In this example, each atom of glycine is encoded as a vector with information on atom types and bond types (e.g., atom types are encoded as C = (1,0,0), N = (0,1,0), and O = (0,0,1); bond types are similarly encoded).[44] The vector transfers its own and upper layer information to the next layer until the root layer (or root node) is reached. A schematic diagram of the DL architecture used in current study is shown in Figure S1. The final output node P in Figure S1 represents a binary annotation of DILI.

DL Architecture Settings and Models. To optimize the DL architecture settings, three parameters were set as variables and one parameter as a constant: the number of hidden layer cells in the UGRNN encoding layer (EH), which ranged from 3 to 12; the number of output layer cells in the UGRNN encoding layer (EO), which ranged from 3 to 12; the number of hidden layer cells in the output layer (OH), which ranged

DOI: 10.1021/acs.jcim.5b00238
J. Chem. Inf. Model. 2015, 55, 2085−2093

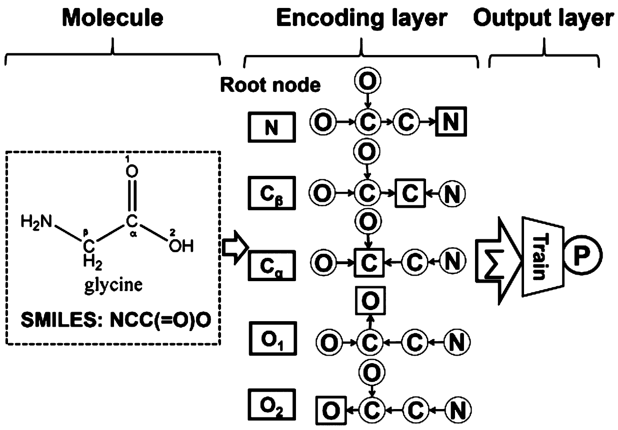


Figure 1. A brief schematic diagram of UGRNN encoding of glycine. First, the chemical structure of glycine is converted into a primary canonical SMILES structure. Second, each of the atoms in the SMILES structure is sequentially defined as a root node. Third, information for all other atoms is transferred along the shortest possible paths.

from 3 to 9; and the number of output layer cells in the output layer (OO), which was equal to 1. As a result, a total of 700 (10 × 10 × 7 × 1) DL architectures were obtained. We used a fast pretraining strategy with only 100 iterations of training to assess the 700 DL architectures. To evaluate the pretraining DL models, the scoring function S was defined as

$$S = -\min(V_{\mathrm{RMSE}}) - \min(V_{\mathrm{AAE}}) + \max(V_{\mathrm{R}}) \quad (1)$$

Here, S is the evaluation score of a DL model and V_{RMSE}, V_{AAE}, and V_{R} are the vectors of root-mean-square error (RMSE), average absolute error (AAE), and the Pearson correlation coefficient (R), respectively.

After architecture setting optimization, the 10 DL architectures with the highest scores and top potencies were selected for 5000 iterations of deep training. The best deeply trained models were selected as the final DL DILI models. In the process of deep training, a scoring function F was defined as

$$A = -\mathrm{RMSE} - \mathrm{AAE} + R \quad (2)$$

$$F = \sum_{i=1}^{10} A_{(i)}/10 \quad (3)$$

Here, A is the evaluation score of each iteration of training and F, used to evaluate the corresponding architecture, is the mean value of the top 10 values of A.

A 10-fold cross validation was used in all training processes. We constructed three final DL DILI prediction models based on three different training data sets, as shown by the flowchart in Figure 2. The three models were the DL model trained on the NCTR data set (DL-NCTR), the DL model trained on the combined data set (DL-Combined), and the DL model trained on the Liew data set (DL-Liew). All final constructed models were further evaluated using external validation tests.

DILI-Related Molecular Feature Motifs. One deficiency of DL models is that they are black-box models without apparent physical meaning. To further understand which molecular features are the most responsible for DILI risk, 1444 PaDEL molecular descriptors[62] were used to describe the molecules in the NCTR data set from Chen et al.[4,63] Assuming normal distribution of the two-class samples, a two-sample T-test was used to identify the descriptors that significantly differed between DILI-positive and -negative drugs. Next, a method of principal component analysis (PCA) was used to determine whether the samples could be distinguished. If the samples were well distinguished by certain combinations of descriptors, the important descriptors were excavated with an analysis of clustering and coefficient weights.

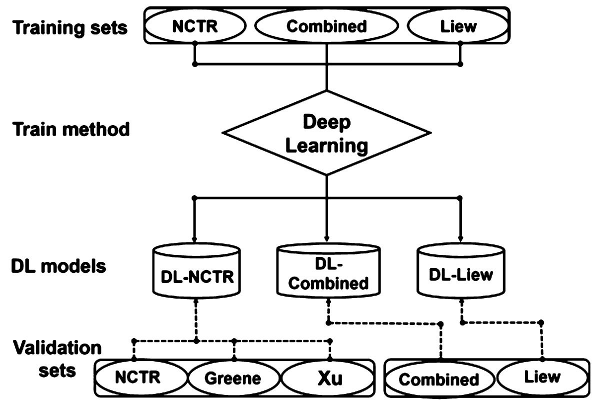


Figure 2. Flowchart for constructing DL DILI models. NCTR, NCTR data set; Greene, Greene data set; Xu, Xu data set; Combined, combined data set; Liew, Liew data set.

RESULTS AND DISCUSSION

Construction of DL Models. Pretraining performance of the 700 models is shown in Figure S2, and details of the top 10 DL architectures are listed in Table 2. The number of cells in each NN layer was a key factor influencing the models. The best 10 preselected architecture settings were then used to deeply train the models with 5000 iterations. The best-scoring model architectures were selected with best F values, as marked in bold in Table 2.

Because the outputs of the DL DILI prediction models are real numbers between 0 and 1, we employed cutoff values in the final models to distinguish DILI-positive and DILI-negative drugs. Cutoff values were calculated using the Daim tool in the R package[64] to estimate misclassification rate, sensitivity, and area under the curve (AUC) based on cross validation. The best cutoff values for the DL-NCTR, DL-Liew, and DL-Combined models were 0.4014, 0.6190, and 0.4811, respectively.

DL-NCTR DILI Model. The training and external validation results of the DL-NCTR model are shown in Table 3, in which the performance of recently described models is also shown for comparison. The training accuracy (ACC), sensitivity (SEN), and specificity (SPE) of the DL-NCTR model were 80.5%, 70.3%, and 88.2%, respectively. These results were much better than the original results reported by Chen et al. (ACC 69.7%, SEN 57.8%, and SPE 77.9%)[4] with nearly the same training set. The prediction results of the DL-NCTR model on the three external validation data sets were an ACC of 70.3%, SEN of 80.0%, and SPE of 60.0% for the NCTR validation data set; an ACC of 64.7%, SEN of 75.1%, and SPE of 46.0% for the Greene validation data set; and an ACC of 61.9%, SEN of 61.7%, and SPE of 62.1% for the Xu validation data set. The difference between training and prediction results suggested there might be overfitting in the training process due to the small data size. The AUC of the DL-NCTR model for the NCTR validation data set was 0.720. Although the prediction ACC and SEN of the current model on the NCTR validation data set was slightly better than that of Chen et al.'s model, the prediction performance on the Greene and Xu data sets was

DOI: 10.1021/acs.jcim.5b00238
J. Chem. Inf. Model. 2015, 55, 2085−2093

Table 2. Details of the Top 10 DL Architecture

ID	1	2	3	4	5	6	7	8	9	10
	DL architectures									
options	DL-NCTR model									
EH	11	12	12	**7**	9	11	12	4	9	7
EO	11	10	10	**9**	7	6	10	12	11	10
OH	3	6	7	**4**	3	4	3	8	4	4
OO	1	1	1	**1**	1	1	1	1	1	1
options	DL-Liew model									
EH	4	8	9	11	**10**	11	9	12	7	10
EO	12	9	12	6	**11**	11	7	4	10	4
OH	4	4	8	9	**9**	7	9	9	8	8
OO	1	1	1	1	**1**	1	1	1	1	1
options	DL-Combined model									
EH	9	10	9	9	10	7	10	6	**12**	4
EO	5	3	4	5	3	11	6	3	**3**	5
OH	3	8	5	4	4	9	7	8	**5**	8
OO	1	1	1	1	1	1	1	1	**1**	1

Table 3. Performance of the DL-NCTR Model

index[a]	internal cross-validation		external validation					
	NCTR training	original[b] NCTR model	NCTR test	original[b] NCTR model	Greene test	original[b] Greene model	Xu test	original[b] Xu model
ACC (%)	**80.5**	69.7	**70.3**	68.9	**64.7**	61.6	**61.9**	63.1
SEN (%)	**70.3**	57.8	**80.0**	66.3	**75.1**	58.4	**61.7**	60.6
SPE (%)	**88.2**	77.9	**60.0**	71.6	**46.0**	67.5	**62.1**	66.1
number of drugs	**190 (p/n = 81/109)**	197 (p/n = 81/116)	**185 (p/n = 95/90)**	190 (p/n = 95/95)	**320 (p/n = 209/111)**	328 (p/n = 214/114)	**236 (p/n = 128/108)**	241 (p/n = 132/109)

[a]ACC [(TP + TN)/(TP + TN + FP + FN)], SEN [TP/(TP + FN)], SPE [TN/(TN + FP)]. Here, TP = true positive; TN = true negative; FP = false positive; FN = false negative. [b]Note: The results of three original models are provided to compare with that of the DL-NCTR model. Internal and external validation results of the DL-NCTR model are marked in bold. p/n: positive/negative.

only comparable to that of the original models. One possible reason for this discrepancy may be that some drug DILI annotations were inconsistent between the NCTR, Greene, and Xu data sets,[4] which might be a serious problem for DILI prediction. Predicting external data sets with DILI annotations that are consistent with those in the training data set could solve this problem.

Thus, we further developed the DL-Liew model and DL-Combined model, which were constructed from large data sets that may better reveal the learning advantages of DL. Nonetheless, from the view of the performance of the DL-NCTR model, the DL method showed a powerful learning ability relevant to DILI prediction.

DL-Liew DILI Model. The Liew data set is a large training data set of 1065 pharmaceutical and nonpharmaceutical compounds. The training and external validation results of the DL-Liew model are shown in Table 4, in which the performance of the original Liew model is also shown as a comparison. The training results of the DL-Liew model were an ACC of 70.1%, SEN of 70.0%, SPE of 70.0%, Matthew's correlation coefficient (MCC) of 0.394, and geometric mean (GMEAN) of 70.0%, which were better than results produced using Liew et al.'s model (ACC 63.8%, SEN 64.1%, SPE 63.3%, MCC 0.269, and GMEAN 63.7%).[8] The prediction performance of the DL-Liew model on the Liew validation set was an ACC of 74.8%, SEN of 81.4%, SPE of 65.3%, MCC of 0.474, and GMEAN of 72.9%, which was again better than the original results reported by Liew et al. (ACC of 62.2%, SEN of 62.4%, SPE of 61.8%, MCC of 0.240, and GMEAN of 61.8%).[8] The AUC of the DL-Liew model was 0.776 (Figure 3), which was better than that of the Liew et al. model (AUC of 0.668).[8] Therefore, when using the basically same data sets for training and prediction, the performance of the DL-Liew model was confirmed to be more powerful for DILI prediction than that of the model constructed by Liew et al.[8]

Table 4. Performance of the DL-Liew Model

index	internal 10-fold cross validation[a]	external validation
ACC (%)	70.1(63.8)	74.8 (62.2)
SEN (%)	70.0 (64.1)	81.4(62.4)
SPE (%)	70.0 (63.3)	65.3(61.8)
MCC[b]	0.394 (0.269)	0.474 (0.240)
GMEAN (%)[c]	70.0 (63.7)	72.9 (61.8)

[a]The items in parentheses are the results of Liew et al.'s original model. [b]MCC:[65] $((TP \times TN - FN \times FP)/(((TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP))^{1/2}))$. [c]GMEAN: $(SEN \times SPE)^{1/2}$.

DL-Combined DILI Model. The combined data set was constructed for two reasons: (1) The three data sets (NCTR, Greene, and Xu data sets) were all drug data sets and therefore could be combined, and (2) it was inferred that a large training data set would be beneficial to determining the advantages of deep learning. To reconcile the differences in drug DILI annotations between data sets, we followed the FDA-approved drug annotations[57] as much as possible. Thirteen drugs that showed different results in the NCTR and Greene data sets were reannotated. Fourteen other drugs that had different

DOI: 10.1021/acs.jcim.5b00238
J. Chem. Inf. Model. 2015, 55, 2085−2093

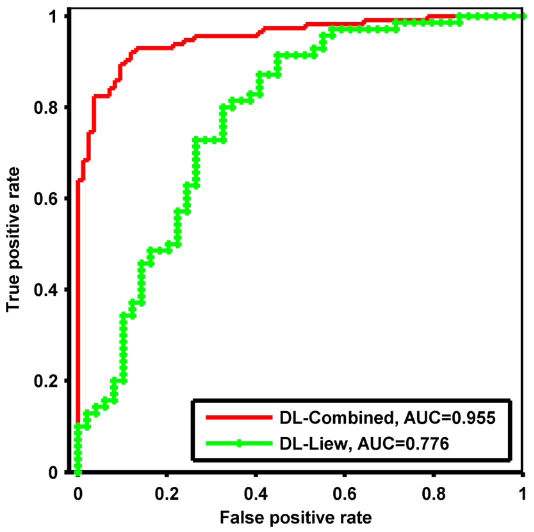


Figure 3. External prediction ROC curves for the DL-Combined and DL-Liew models.

results in the NCTR and Xu validation data sets were also reannotated. The other drugs were annotated using their original labeling. The training and external prediction results of the combined data set are shown in Table 5. The training ACC,

Table 5. Performance of the DL-Combined Model

index	internal 10-fold cross validation	external validation: combined validation set (198)
ACC (%)	88.4	86.9
SEN (%)	89.9	82.5
SPE (%)	87.0	92.9
MCC	0.771	0.746
GMEAN (%)	88.3	87.5

SEN, SPE, MCC, and GMEAN were 88.4%, 89.9%, 87.0%, 0.771, and 88.3%, respectively, showing that the model was quite well trained. The external predictions of ACC, SEN, SPE, MCC, and GMEAN were 86.9%, 82.5%, 92.9%, 0.746, and 87.5%, respectively. The AUC of the DL-Combined model was 0.955, as shown in Figure 3. These results demonstrated a state-of-the-art improvement of DILI computational prediction models by the DL methods.

Influence of the Size of the Training Data set. To access the influence of the training data set size on model performance, we built DL models with different percentages (40%, 60%, 80%, and 100%) of the size of the combined training set by random data selection. Random selection was performed three times for each data set size to avoid selection bias. Training and external validation results are shown in Table S1. The standard deviations of the training represents the standard deviations from the 10-fold cross-validation processes. During 10-fold cross-validation, one tenth of the training data was reserved for prediction, and if the training set is small (for example, 40% of the Combined training set had 190 compounds, and then only 19 compounds were used for validation), the standard deviations tends to be large. It is also possible that the DL method has not still learned enough information to predict robustly when data size is small. We found that with an increasing size of the combined training set, the training accuracies are increasing (their corresponding standard deviations are decreasing), and the prediction performance is also increasing. In Figure 4, the ACC values

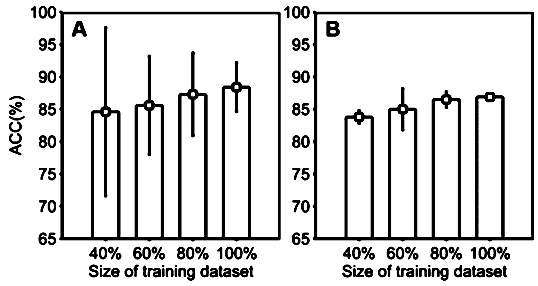


Figure 4. Training (A) and prediction (B) accuracies of the DL-Combined models trained on different sizes of training data sets (40%, 60%, 80%, and 100%).

became stable when 60−100% of the data were used, but the standard deviation values continued to decrease with increasing data set size. Therefore, the performance of the DL model became more accurate and robust with larger training data sets.

Influence of Different Splits on Training and Prediction Data. In order to investigate the stability of our DL models with different splits for training and prediction sets, we randomly divided the combined data set into training and validation sets with the proportion ~475/198 for 60 times. The computational results with the DL-Combined settings are listed in Table 6. The average training and prediction results were

Table 6. Performance of Random Splits of Training and Prediction Data on the Combined Data Set

index	internal cross-validation (60 runs)[a]	external validation: combined validation set (198)
ACC (%)	84.6 ± 1.1	84.3 ± 1.2
SEN (%)	84.4 ± 1.7	79.4 ± 2.1
SPE (%)	84.8 ± 1.5	90.9 ± 1.7
MCC	0.696 ± 0.021	0.695 ± 0.023
GMEAN (%)	84.5 ± 1.1	85.0 ± 1.2

[a]Cross-validated results are averaged values over 60 runs of 10-fold cross-validations.

slightly worse than the original values, but are still strong enough. These results provide more reasonable and interpretable assessment of our DL DILI models.

Comparison to Normal NN and DNN Models. DILI prediction models were built using PaDEL[62] and Mold2[10] descriptors with normal neural networks (NN) and deep neural networks (DNN)[48] with the combined data set. The details of the methods of the NN and DNN models are shown in the Supporting Information. The training and prediction results on the combined data set for these models are shown in the Table 7 and Figure 5, from which it is clear that models using deep neural networks perform slightly better than models using normal shallow neural networks for both training and prediction and both PaDEL and Mold2 descriptors, but they perform worse than the DL-Combined model using UGRNN. In UGRNN architecture, apparent descriptors are not needed, and the molecular structural information is automatically encoded by such a deep net. From these results, we can see that the good performance of our DL models comes from both the encoding method and the deep neural networks.

DOI: 10.1021/acs.jcim.5b00238
J. Chem. Inf. Model. 2015, 55, 2085−2093

Table 7. Performance of the Neural Network Model and Deep Neural Network Model

		neural network		deep neural network	
molecular descriptor	index	internal 10-fold cross validation	external validation	internal 10-fold cross validation	external validation
Mold2 descriptor	ACC (%)	82.5	82.3	83.2	83.3
	SEN (%)	78.4	71.1	83.1	79.0
	SPE (%)	86.6	97.6	83.3	89.3
	MCC	0.652	0.688	0.663	0.675
	GMEAN (%)	82.4	83.3	83.2	84.0
PaDEL descriptor	ACC (%)	81.6	79.1	82.3	81.1
	SEN (%)	75.8	72.3	85.2	82.1
	SPE (%)	87.5	88.1	79.4	79.8
	MCC	0.638	0.599	0.647	0.616
	GMEAN (%)	81.5	79.8	82.2	80.9

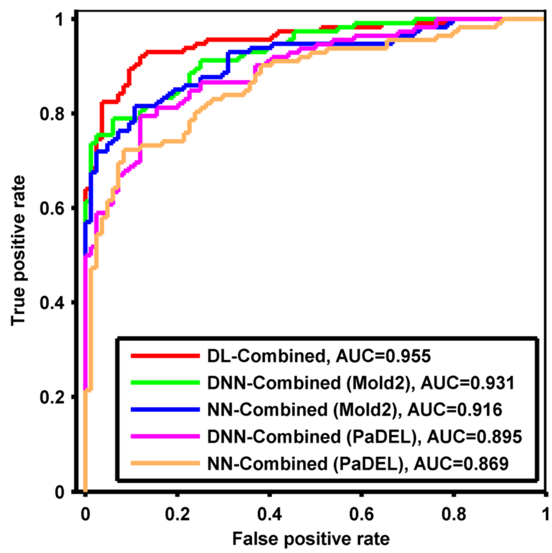


Figure 5. Prediction ROC curves for the DL-Combined, DNN-Combined, and NN-Combined models with different types of descriptors.

Further Discussion of the DL DILI Models. We showed that DL models performed better than previous models in DILI prediction. This was most obviously the case when a large data set was used for training. The DL DILI models performed well in predicting a large external validation set. As shown by the ROC curves in Figure 3, the AUC values for the corresponding external data sets were 0.955 for the DL-Combined model and 0.776 for the DL-Liew model, which were relatively high values for prediction. The DL-Combined and DL-Liew models were constructed for two reasons. First, there were differences in annotations between data sets. For the combined data set, the FDA-approved drug labeling method was used. Only two kinds of drugs (high DILI risk and no DILI risk) were included, and no uncertain data were included. For the Liew data set, a different strategy was used to label drugs with any DILI potential "DILI-positive," and drugs not associated with DILI "DILI-negative." This difference of labeling strategies resulted in 115 inconsistent annotations between the two data sets (Figure S3). Second, there were also differences in compound categories between data sets. Whereas the combined data set was a pharmaceutical data set, the Liew data set included pharmaceutical and nonpharmaceutical compounds.[8] As such, the DL-Combined model is more suitable for drug DILI risk prediction, and its prediction results were relatively extreme: DILI risk or no DILI risk. The DL-Liew model, on the other hand, is more suitable for compound (i.e., not only drug) DILI risk prediction, and its positive results are likely to indicate less DILI risk. Therefore, we established DL-Combined and DL-Liew models which have different scopes of application.

In the current study, DILI properties were divided into only two categories: DILI-positive and DILI-negative. However, in many public and canonical data sets,[2,51,57] multiple levels of DILI are often used. Multiple-level prediction of DILI might be more accurate, but the lack of data makes it difficult to develop such prediction models at the current time. A unified criterion of drug DILI annotation is urgently needed to help develop better DL models to improve DILI risk prediction.

Analysis of DILI Feature Motifs. Three hundred and one descriptors (P value <0.05), including AlogP^2 ($(1 - \log P)^2$), Mannhold log P, and average molecular weight,[62] showed statistically significant differences after the two-sample t-test. Therefore, we performed PCA on these descriptors. Combinational distribution projected on the first and second principle components is displayed in Figure 6. Results showed that DILI-positive and -negative samples could be distinguished by a combination of structural features, suggesting the possibility of predicting DILI simply with the combination of these useful structural features. Results also showed that fragment complexity (FC; representing the complexity of a system), f(molecular framework) (FMF; representing promiscuity), and the Zagreb Index (representing the sum of squares of atom degree over all heavy atoms),[62] which are summarized and shown in Table S2, were the three most prominent features distinguishing DILI-positive from DILI-negative compounds. We combined, in

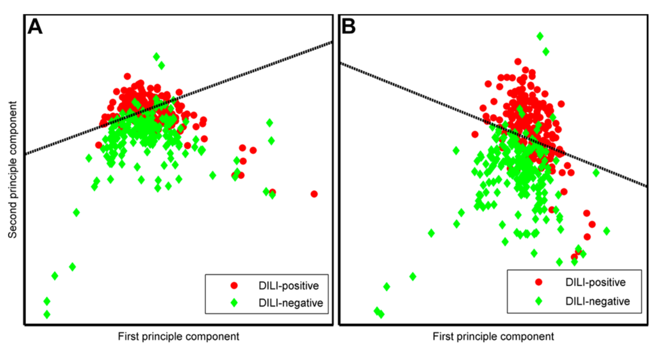


Figure 6. Plot of two top principle components. (A) The two principle components are based on the total descriptors (not treated with t test). The ACC based on the dotted line is 74.4%. (B) The two principle components are based on the selected descriptors (treated with t test). The ACC is improved to 83.1%.

DOI: 10.1021/acs.jcim.5b00238
J. Chem. Inf. Model. 2015, 55, 2085−2093

pairs, these three features and other confirmed features[63] (i.e., log *P* and daily dose) to analyze the properties of the NCTR data set. Figure 7 shows the results produced using six different

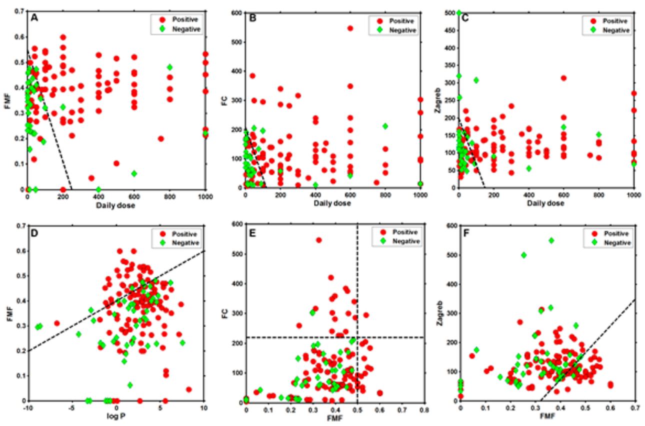


Figure 7. Combinations of FMF and daily dose, FC and daily dose, Zagreb index and daily dose, FMF and log *P*, FC and FMF, Zagreb index and FMF for DILI discrimination. The black dotted lines best distinguish DILI-positive and -negative samples. Details of these six combinations and their results are shown in Table 8.

feature combinations (log *P* and daily dose combination, which was analyzed by Chen et al.,[63] was not analyzed here). Results showed that the two categories were readily separated by daily dose, indicating that drug dose actually plays a key role in liver injury. Although the FC and log *P*, Zagreb and FC, Zagreb and log *P* combinations did not play rather well (Figure S4), the FMF and log *P*, FMF and FC, and Zagreb and FMF combinations identified the positive samples effectively, as shown in Table 8 and Figure 7. We inferred that these molecular structural features and their combinations were prominently associated with DILI and thus could be used for DILI prediction.

Table 8. Prediction DILI by Simple Pairwise Molecular Feature Combination

combination of features	precision (%)
daily dose/250 + FMF/0.55 ≥ 1	84.9
daily dose/120 + FC/210 ≥ 1	81.7
daily dose/150 + Zagreb/200 ≥ 1	81.7
log P/(−20) + FMF/0.4 ≥ 1	86.8
FC ≥ 220 or FMF ≥ 0.5	84.6
FMF/0.35 + Zagreb/(−295) ≥ 1	92.5

■ CONCLUSION

In conclusion, using large data sets and the UGRNN molecular encoding approach with the least information loss, we have developed DL models for predicting the DILI of drugs and small compounds. Mainly due to the powerful learning ability of DL, these models performed better than previous DILI prediction models. The DL-Combined model trained on 475 drugs predicted an external validation data set of 198 drugs with an accuracy of 86.9%, sensitivity of 82.5%, specificity of 92.9%, and AUC of 0.955, which were considerably high. With further analysis, we also found some important features of molecular structures that are closely related to DILI. Our DL models are expected to improve DILI risk prediction in humans and are freely available at http://www.repharma.cn/DILIserver/DILI_home.php. The deep learning methods may see widespread use in chemical and drug informatics studies covering subjects beyond DILI prediction.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00238.

Data processing; DL architecture; DL pretraining results; inconsistence of the DILI annotation between the combined and the Liew data sets; NN and DNN settings; Pairwise combinations of log P, FC, and Zagreb Index for DILI discrimination (PDF)

Detailed information about compound canonical SMILES strings and prediction results by the DL models (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

*Fax: (+86)10-62759595. E-mail: jfpei@pku.edu.cn.

*Fax: (+86)10-62751725. E-mail: lhlai@pku.edu.cn.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was supported, in part, by the Ministry of Science and Technology of China (grant numbers: 2012AA020308, 2012AA020301) and the National Natural Science Foundation of China (grant numbers: 81273436, 91313302).

■ ABBREVIATIONS

DILI, drug-induced liver injury; DL, deep learning; UGRNN, undirected graph recursive neural networks

■ REFERENCES

(1) *Guidance for Industry Drug-Induced Liver Injury: Premarketing Clinical Evaluation*; Food and Drug Administration: Silver Spring, MD, 2009; pp 38035−38036.

(2) Hoofnagle, J. H.; Serrano, J.; Knoben, J. E.; Navarro, V. J. Livertox: A Website on Drug-Induced Liver Injury. *Hepatology* **2013**, *57*, 873−874.

(3) Assis, D. N.; Navarro, V. J. Human Drug Hepatotoxicity: A Contemporary Clinical Perspective. *Expert Opin. Drug Metab. Toxicol.* **2009**, *5*, 463−473.

(4) Chen, M.; Hong, H.; Fang, H.; Kelly, R.; Zhou, G.; Borlak, J.; Tong, W. Quantitative Structure-Activity Relationship Models for Predicting Drug-Induced Liver Injury Based on FDA-Approved Drug Labeling Annotation and Using a Large Collection of Drugs. *Toxicol. Sci.* **2013**, *136*, 242.

(5) Cruz-Monteagudo, M.; Cordeiro, M.; Borges, F. Computational Chemistry Approach for the Early Detection of Drug-Induced Idiosyncratic Liver Toxicity. *J. Comput. Chem.* **2008**, *29*, 533−549.

(6) Rodgers, A. D.; Zhu, H.; Fourches, D.; Rusyn, I.; Tropsha, A. Modeling Liver-Related Adverse Effects of Drugs Using K Nearest Neighbor Quantitative Structure- Activity Relationship Method. *Chem. Res. Toxicol.* **2010**, *23*, 724−732.

(7) Ekins, S.; Williams, A. J.; Xu, J. J. A Predictive Ligand-Based Bayesian Model for Human Drug-Induced Liver Injury. *Drug. Metab. Dispos.* **2010**, *38*, 2302−2308.

(8) Liew, C. Y.; Lim, Y. C.; Yap, C. W. Mixed Learning Algorithms and Features Ensemble in Hepatotoxicity Prediction. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 855−871.

DOI: 10.1021/acs.jcim.5b00238
J. Chem. Inf. Model. 2015, 55, 2085−2093

(9) Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. Decision Forest: Combining the Predictions of Multiple Independent Decision Tree Models. *J. Chem. Inf. Model.* **2003**, *43*, 525−531.
(10) Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold2, Molecular Descriptors from 2d Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48*, 1337−1344.
(11) Muller, C.; Pekthong, D.; Alexandre, E.; Marcou, G.; Horvath, D.; Richert, L.; Varnek, A. Prediction of Drug Induced Liver Injury Using Molecular and Biological Descriptors. *Comb. Chem. High Throughput Screening* **2015**, *18*, 315−322.
(12) Hinton, G. E.; Osindero, S.; Teh, Y.-W. A Fast Learning Algorithm for Deep Belief Nets. *Neural. Comput.* **2006**, *18*, 1527−1554.
(13) Bengio, Y. Learning Deep Architectures for AI. *J. Found. Trends. Mach. Learn.* **2009**, *2*, 1−127.
(14) Coates, A.; Ng, A. Y.; Lee, H. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *International Conference on Artificial Intelligence and Statistics*; 2011; pp 215−223.
(15) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, 2012; pp 1097−1105.
(16) Le, Q. V. Building High-Level Features Using Large Scale Unsupervised Learning. In Acoustics, Speech and Signal Processing (ICASSP). *IEEE International Conference on* **2013**, 8595−8598.
(17) Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. *In Proceedings of the ACM International Conference on Multimedia, ACM* **2014**, 675−678.
(18) Jaitly, N.; Nguyen, P.; Senior, A. W.; Vanhoucke, V. Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition. In *INTERSPEECH*; International Speech Communication Association: Baixas, France, 2012.
(19) Dahl, G. E.; Yu, D.; Deng, L.; Acero, A. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. Audio, Speech, and Language Processing. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 30−42.
(20) Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82−97.
(21) Graves, A.; Mohamed, A.-r.; Hinton, G. Speech Recognition with Deep Recurrent Neural Networks. In Acoustics, *Speech and Signal Processing (ICASSP). IEEE International Conference on, IEEE* **2013**, 6645−6649.
(22) Noda, K.; Yamaguchi, Y.; Nakadai, K.; Okuno, H. G.; Ogata, T. Audio-Visual Speech Recognition Using Deep Learning. *Applied Intelligence* **2015**, *42*, 722−737.
(23) Deng, L.; Yu, D.; Dahl, G. E. Deep belief network for large vocabulary continuous speech recognition. US20120065976 A1, 2015.
(24) Collobert, R.; Weston, J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. *In Proceedings of the 25th International Conference on Machine Learning, ACM* **2008**, 160−167.
(25) Yu, D.; Deng, L. Deep Learning and Its Applications to Signal and Information Processing [Exploratory Dsp]. *Signal Processing Magazine, IEEE* **2011**, *28*, 145−154.
(26) Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493−2537.
(27) Socher, R.; Lin, C. C.; Manning, C.; Ng, A. Y. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*; International Machine Learning Society: 2011; pp 129−136.
(28) Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, 2013; pp 3111−3119.
(29) Gao, J.; He, X.; Deng, L. Deep Learning for Web Search and Natural Language Processing. *Microsoft Research Technical Report*; Microsoft Corporation: Redmond, WA, 2015; MSR-TR-2015−7.
(30) Brooks, R. A. Intelligence without Representation. *Artificial intelligence* **1991**, *47*, 139−159.
(31) Ciresan, D.; Meier, U.; Schmidhuber, J. Multi-Column Deep Neural Networks for Image Classification. *In Computer Vision and Pattern Recognition (CVPR), IEEE Conference on* **2012**, 3642−3649.
(32) LeCun, Y.; Bengio, Y. Convolutional Networks for Images, Speech, and Time Series. *Handbook of Brain Theory and Neural Networks*; MIT Press: Cambridge, MA, 1995; p 3361.
(33) Lawrence, S.; Giles, C. L.; Tsoi, A. C.; Back, A. D. Face Recognition: A Convolutional Neural-Network Approach. *IEEE Trans. Neural Netw.* **1997**, *8*, 98−113.
(34) Sun, Y.; Wang, X.; Tang, X. Deep Convolutional Network Cascade for Facial Point Detection. *In Computer Vision and Pattern Recognition (CVPR), IEEE Conference on* **2013**, 3476−3483.
(35) Tompson, J. J.; Jain, A.; LeCun, Y.; Bregler, C. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, 2014; pp 1799−1807.
(36) Pineda, F. J. Generalization of Back-Propagation to Recurrent Neural Networks. *Phys. Rev. Lett.* **1987**, *59*, 2229−2232.
(37) LeCun, Y.; Bengio, Y.; Hinton, G. *Nature* **2015**, *521*, 436−444.
(38) Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks* **2015**, *61*, 85−117.
(39) Qi, Y.; Tastan, O.; Carbonell, J. G.; Klein-Seetharaman, J.; Weston, J. Semi-Supervised Multi-Task Learning for Predicting Interactions between Hiv-1 and Human Proteins. *Bioinformatics* **2010**, *26*, i645−i652.
(40) Plötz, T.; Hammerla, N. Y.; Olivier, P. Feature Learning for Activity Recognition in Ubiquitous Computing. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, 2011; vol. *22*, p 1729.
(41) Di Lena, P.; Nagata, K.; Baldi, P. Deep Architectures for Protein Contact Map Prediction. *Bioinformatics* **2012**, *28*, 2449−2457.
(42) Eickholt, J.; Cheng, J. Predicting Protein Residue−Residue Contacts Using Deep Networks and Boosting. *Bioinformatics* **2012**, *28*, 3066−3072.
(43) Eickholt, J.; Cheng, J. Dndisorder: Predicting Protein Disorder Using Boosting and Deep Networks. *BMC Bioinf.* **2013**, *14*, 88.
(44) 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.
(45) Leung, M. K.; Xiong, H. Y.; Lee, L. J.; Frey, B. J. Deep Learning of the Tissue-Regulated Splicing Code. *Bioinformatics* **2014**, *30*, i121−i129.
(46) Unterthiner, T.; Mayr, A.; ünter Klambauer, G.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Hochreiter, S. Deep Learning as an Opportunity in Virtual Screening. In *Deep Learning and Representation Learning Workshop, NIPS*; MIT Press: Cambridge, MA, 2014.
(47) Hughes, T. B.; Miller, G. P.; Swamidass, S. J. Modeling Epoxidation of Drug-Like Molecules with a Deep Machine Learning Network. *ACS Cent. Sci.* **2015**, *1*, 168−180.
(48) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure−Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263−274.
(49) Park, Y.; Kellis, M. Deep Learning for Regulatory Genomics. *Nat. Biotechnol.* **2015**, *33*, 825−826.
(50) Alipanahi, B.; Delong, A.; Weirauch, M. T.; Frey, B. J. Predicting the Sequence Specificities of DNA-and Rna-Binding Proteins by Deep Learning. *Nat. Biotechnol.* **2015**, *33*, 831−838.
(51) Lemme, A.; Reinhart, R. F.; Steil, J. J. Online Learning and Generalization of Parts-Based Image Representations by Non-Negative Sparse Autoencoders. *Neural Networks* **2012**, *33*, 194−203.

DOI: 10.1021/acs.jcim.5b00238
J. Chem. Inf. Model. 2015, 55, 2085−2093

(52) Lee, H.; Grosse, R.; Ranganath, R.; Ng, A. Y. In *Proceedings of the 26th Annual International Conference on Machine Learning*; 2009; pp 609−616.
(53) Marzorati, M.; Wittebolle, L.; Boon, N.; Daffonchio, D.; Verstraete, W. How to Get More out of Molecular Fingerprints: Practical Tools for Microbial Ecology. *Environ. Microbiol.* **2008**, *10*, 1571−1581.
(54) Randid, M. Novel Molecular Descriptor for Structure-Property Studies. *Chem. Phys. Lett.* **1993**, *211*, 478−483.
(55) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; John Wiley & Sons: Hoboken, NJ, 2009; p 41.
(56) Schenkenberg, T.; Bradford, D.; Ajax, E. Line Bisection and Unilateral Visual Neglect in Patients with Neurologic Impairment. *Neurology* **1980**, *30*, 509−509.
(57) Chen, M.; Vijay, V.; Shi, Q.; Liu, Z.; Fang, H.; Tong, W. FDA-Approved Drug Labeling for the Study of Drug-Induced Liver Injury. *Drug Discovery Today* **2011**, *16*, 697−703.
(58) Greene, N.; Fisk, L.; Naven, R. T.; Note, R. R.; Patel, M. L.; Pelletier, D. J. Developing Structure- Activity Relationships for the Prediction of Hepatotoxicity. *Chem. Res. Toxicol.* **2010**, *23*, 1215−1222.
(59) Xu, J. J.; Henstock, P. V.; Dunn, M. C.; Smith, A. R.; Chabot, J. R.; de Graaf, D. Cellular Imaging Predictions of Clinical Drug-Induced Liver Injury. *Toxicol. Sci.* **2008**, *105*, 97−105.
(60) Kamada, T.; Kawai, S. An Algorithm for Drawing General Undirected Graphs. Inform. *Process. Lett.* **1989**, *31*, 7−15.
(61) Jensen, F. V. *An Introduction to Bayesian Networks*; UCL press: London, 1996; p 210.
(62) Yap, C. W. Padel-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−1474.
(63) Chen, M.; Borlak, J.; Tong, W. High Lipophilicity and High Daily Dose of Oral Medications Are Associated with Significant Risk for Drug-Induced Liver Injury. *Hepatology* **2013**, *58*, 388−396.
(64) Potapov, S.; Adler, W.; Lausen, B. Daim: Diagnostic Accuracy of Classification Models. *R Package*, version 1.0.0; 2011.
(65) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405*, 442−451.

DOI: 10.1021/acs.jcim.5b00238
J. Chem. Inf. Model. 2015, 55, 2085−2093