JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

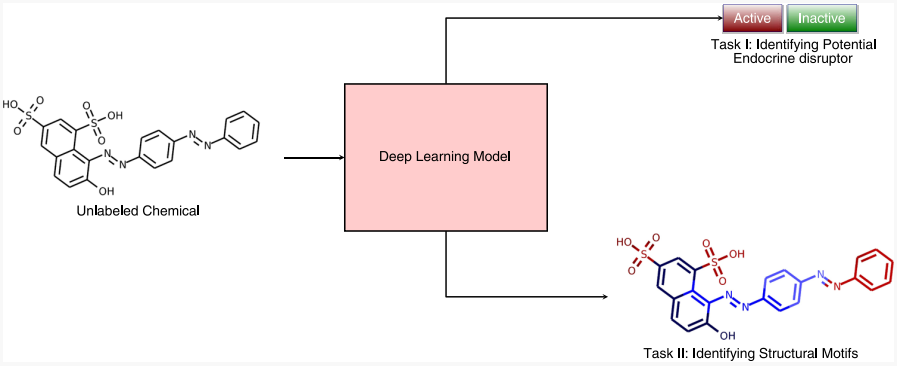
pubs.acs.org/jcim Article

Deep Learning Model for Identifying Critical Structural Motifs in Potential Endocrine Disruptors

Arpan Mukherjee, An Su, and Krishna Rajan*

Cite This: J. Chem. Inf. Model. 2021, 61, 2187−2197 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: This paper aims to identify structural motifs within a molecule that contribute the most toward a chemical being an endocrine disruptor. We have developed a deep neural network-based toolkit toward this aim. The trained model can virtually assess a synthetic chemical's potential to be an endocrine disruptor using machine-readable molecular representation, simplified molecular input line entry system (SMILES). Our proposed toolkit is a multilabel or multioutput classification model that combines both convolution and long short-term memory (LSTM) architectures. The toolkit leverages the advantages of an active learning-based framework that combines multiple sources of data. Class activation maps (CAMs) generated from the feature-extraction layers can identify the structural alerts and the chemical environment that determines the specificity of the structural alerts.

INTRODUCTION

Synthetic chemicals are leading to existing impacts and potential risks to human health and the environment. Among them, there is a notorious group of chemicals called the endocrine-disrupting chemicals (EDC) that are known to disrupt the hormonal regulation and endocrine system of humans and animals.[1] To identify the potential toxicity of hazardous chemicals like EDC, structural alerts (SAs) have been a widely accepted way in the fields of chemical toxicology and regulatory decision support.[2] SAs are functional groups or molecular substructures based on human expertize that reflect the chemical basis of activity or properties. SAs are easy to generate and explain, but an increasing number of studies have shown that the accuracy in the toxicity assessment using structural alerts is limited.[2] Many SAs can be found in both toxic and nontoxic chemicals. In other words, SAs do not always lead to toxicity—they have specificities that usually depend on other groups in the molecule.[3] Hence, to achieve toxicity estimation with higher accuracy, methods that can determine the specificities of SAs are urgently needed.

Since 2008, collaboration programs for large-scale in vitro toxicity screening of chemicals such as Tox21[4−6] and ToxCast[7−10] have been initiated, which provides a vast amount of toxicity testing data (e.g., Tox21 10K chemicals[11−14]). Meanwhile, the rising of big data and artificial intelligence have brought attention to the application of machine learning technologies to toxicity prediction in a data-driven fashion.[15−17] In 2016, DeepTox, a fully connected deep learning model developed by Mayr et al., demonstrated that toxicity-related structures could be encoded in the hidden units of deep learning.[18] A few studies have used machine learning or deep learning algorithms with different formats of molecular representations (e.g., molecular descriptors, fingerprints, two-dimensional (2D) graphical representations) to extract chemical structure features that lead to toxicity.[19−21] A recent study from Webel et al. suggested that using SMILES representation as molecular encoding can help generalize the applicability domain to the entire chemical space and avoid bit collision, which is a limitation of fingerprints.[21]

Received: December 7, 2020
Published: April 19, 2021

https://doi.org/10.1021/acs.jcim.0c01409
J. Chem. Inf. Model. 2021, 61, 2187−2197

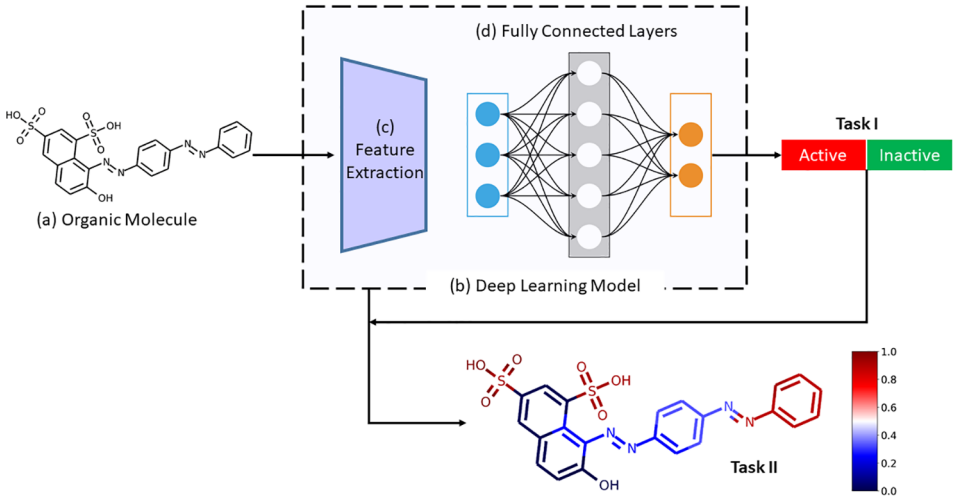


Figure 1. Overall workflow of (Task I) predicting a potential endocrine disruptor and (Task II) identifying structural motifs in the chemicals. This involves (b) a Deep Learning Model that comprises (c) feature extractions and (d) fully connected layers.

Simplified molecular input line entry system (SMILES) is a single line notation of the molecular graph of the chemicals using a finite set of ASCII characters.[22] Compared to the graphical representation or molecular descriptors, using SMILES as the molecular representation for the input of machine learning models can accurately represent the 2D molecular structure with machine-readable strings and retain all of the necessary structural information from the molecule.[23−25] A few recent studies have used different neural network models to read SMILES representation of compounds, predict the structure−activity relationships or structure−property relationships (SAR/SPR), and identify the chemical motifs or atoms that can interpret the predicted results.[26−28] Meanwhile, gradient-weighted class activation mapping (Grad-CAM),[29−31] an explainable AI method, has been applied to explain neural network classification tasks in the fields of image recognition,[32,33] radiology,[34−37] and microstructure recognition.[38] Grad-CAM is a visual tool supported by mathematical formulations that quantifies the statistical significance of all of the high-level features in an image-type data and thereby provides both qualitative and quantitative assessment of the working of a fitted model. A deep learning model containing convolution filters is suited to the Grad-CAM technique because of its ability to extract layer-by-layer geometric features from such a machine-readable molecular descriptor such as SMILES. A convolution neural network (CNN) can be used to extract high-level features that correspond to individual subgroups from the SMILES string. Hence, Grad-CAM can be applied on SMILES to visually represent critical chemical substructures that contribute toward a specific classification task, and to the best of our knowledge, Grad-CAM has not been implemented in explaining the toxicity of chemicals as endocrine disruptors.

Choosing a suitable classification model is motivated by the type of data available and the research question. In a lot of data-driven toxicology studies, the type of data is often imbalanced, i.e., there is an unequal distribution of the chemicals across both the active (toxic) and inactive (nontoxic) class,[39−41] or from multiple sources.[42] Traditional and advanced deep learning models also fail to generate a discriminative function that can map the imbalanced data into a separable space leading to inaccurate toxicity prediction.[43] While learning from imbalanced data is still an open challenge, numerous researchers have attempted to solve this problem using different preprocessing techniques. Additionally, training a classifier on multiple data sources is a challenge due to heterogeneity in the data. Quantifying the statistical divergence between data is an intractable problem due to high dimensionality,[44] and a classic approach is often to train models on individual data sets.[42] Thus, a predictive and interpretable data-driven model needs suitable data-handling methodologies that address the problems associated with the class imbalance and multiple sources and work in sync with the chosen molecular representation and the machine learning model.

In our current study, an important concept, "critical structural motif" (CSM), that is different from SAs or "toxicophores/pharmacophores",[45−47] is introduced. Unlike SAs or toxicophores, the CSMs from our chemical activation maps consist of both the existing or potential SAs and the chemical environment that determines the specificity of SAs. We have used the technique of Grad-CAM to extract and visualize these CSMs in our study that requires training an accurate CNN-based deep learning model to predict the activity (toxicity) of chemicals. Our proposed model is trained on two different sources of chemical data represented by their SMILES strings using the Active Learning paradigm.[48] We have adopted techniques of undersampling[49−52] and rule-based oversampling[53] to tackle the imbalance in our training and testing data. Our proposed model consists of convolution layers to extract layer-by-layer features from low-level to high-level chemical subgroups. The convolution layers are followed by long short-term memory (LSTM) layers that learn the interdependencies between the high-level features extracted by the convolution layers.[54−56] The Grad-CAM quantifies the relative importance of the high-level subgroups identified by the last convolution layers along with the feedback from the predicted class label to show the CSMs.

METHODS

Overview. Our proposed deep learning model, as shown in Figure 1, is capable of performing two tasks: (I) classifying the binding and agonist estrogen receptor (ER) activity of a chemical to active or inactive (Figure 1a) and (II) highlighting

https://doi.org/10.1021/acs.jcim.0c01409
J. Chem. Inf. Model. 2021, 61, 2187−2197

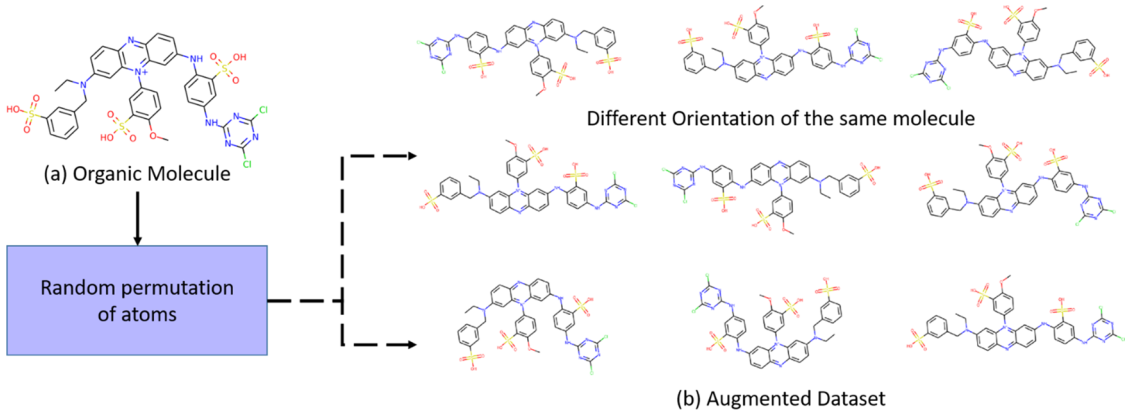


Figure 2. Data augmentation: The database of SMILES strings is augmented to give different orientations of the same chemical (b). This is achieved by permuting the atoms in each chemical (a) to give us an augmented data set (b).

Table 1. Final Training and Testing Data Set after Random Undersampling, Train-Test Split, and Rule-Based Data Augmentation

			Tox21 (primary)			literature (secondary)		
training	agonist	active	unique chemicals: 331	1488	size: 3236	unique chemicals: 456	2304	size: 4492
		inactive		1748			2188	
	binding	active		1598			2559	
		inactive		1638			1933	
testing	agonist	active	unique chemicals: 143	675	size: 1419	unique chemicals: 196	924	size: 1930
		inactive		744			1006	
	binding	active		739			1016	
		inactive		680			914	
total				4655			6432	

the subgroups that contribute to the chemical's ER activity classification. The proposed model (Figure 1b) consists of (Figure 1c) feature-extraction layers and (Figure 1-d) fully connected layers. The geometric features in a chemical identified while performing the specific classification task are used to generate the (Figure 1 Task II) activation maps that illustrate the contributions of subgroups. The feedback from the predicted label is also used along with the feature maps to generate the activation maps. A step-by-step methodology has been detailed in Figure S3 in the Supporting Information.

Data Sets. The original data sets for this study are the training set and evaluation set of the US EPA's Collaborative Estrogen Receptor Activity Prediction Project (CERAPP)[a42] There are two data sets that we have used in our current study. In both data sets, each entry of chemical structures is grouped into three binary (active/inactive) classes: binders, agonists, and antagonists. The primary source of data, derived from ToxCast and Tox21 programs,[7,8,12] is a collection of 18 in vitro HTS assays on the different sites of mammalian ER pathways.[42,57] The secondary source of data is the CERAPP's evaluation set collected from a variety of overlapping sources, including additional US EPA's HTS assays,[5,11,12,58] and estrogenic activity data from other online databases.[59,60] Rules are applied to solve the inconsistency of the results from different sources, and more details are introduced in the description of CERAPP.[42]

Our data analysis shows a high imbalance in the CERAPP's training set. While there are moderate imbalances in the class of binder (14.1% active) and agonist (13.1% active), only 2.4% are active antagonists (Figure S1). Hence, the class of antagonists is not included in our study. Also, only the chemicals having valid data in both binder and agonist classifications are selected, leaving the final data sets in our study with 1,677 chemical structures from the training set (Tox21) and 6206 from the evaluation set (Literature). The overall preprocessing of the data set involves three steps: undersampling, train-test splitting, and data augmentation. We have balanced the data sets using undersampling[49−52] that otherwise may wrongly interfere with the performance of the classifier.[61] We have used Random undersampling without replacement[62] to randomly sample the active chemicals from the original data set to match the number of inactive chemicals. Thus, we will have more active data instances and fewer inactive instances, which solve data imbalances. The method is simple and easy to apply.

We have performed a random train-test split of ratio 7:3 on the balanced data set. Each chemical is represented by its Canonical SMILES **s** and is assigned two binary labels $y = \{y1, y2\}$, $y1, y2 \in [0, 1]$, where 0 represents activity and 1 represents inactivity. For each canonical SMILES, a rule-based data augmentation technique[53] is adopted that randomly changes the order of the atoms in the SMILES leading to different molecular orientations of the same chemical (see Figure 2). A rule-based SMILES enumeration[53] is adopted for two purposes. First, through SMILES enumeration, a molecule is represented by different SMILES instead of a single canonical SMILES, while using canonical SMILES may lead to the latent chemical space representing the grammar rules of canonical SMILES rather than the underlying chemical structure.[63] Second, SMILES enumeration is a useful data augmentation technique that generates more data to train our machine learning model and help overcome data insufficiency. After random undersampling, train-test split, and rule-based data augmentation, the final data set contains 4655 data points

https://doi.org/10.1021/acs.jcim.0c01409
J. Chem. Inf. Model. 2021, 61, 2187−2197

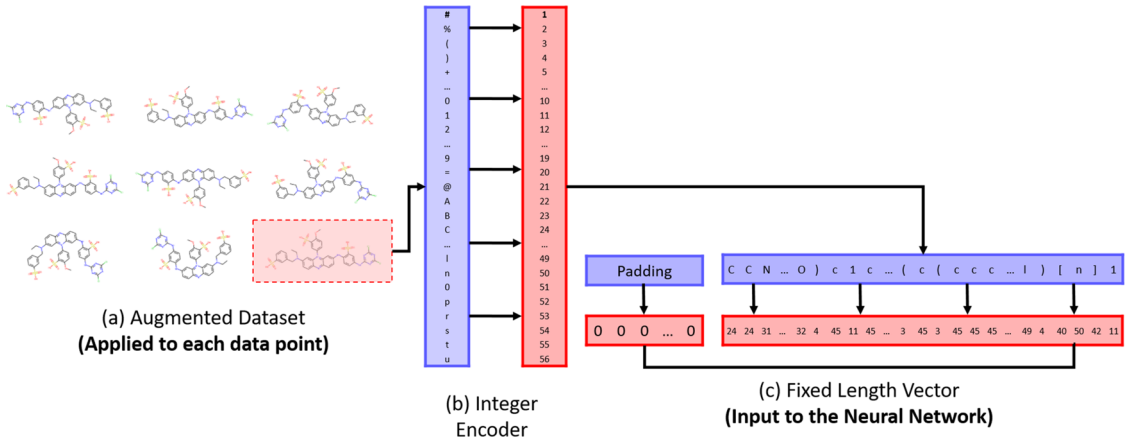


Figure 3. Encoding and padding: Each enumerated SMILES is encoded using the bag-of-words model. The bag-of-words represents a pool of characters that are used to represent the SMILES code for all of the chemicals in the data set. A non-negative integer is assigned to each character such as {C: 23}, {#:1}. The SMILES is mapped to a string of integers using the bag-of-words. Furthermore, the encoded SMILES is padded with zeros in the beginning, to convert it into a fixed-length vector of length 130 that acts as a feature vector for the Deep Learning model. The length is chosen arbitrarily based on the longest SMILES code in the database. Thus, a SMILES of length 80 is padded with 50 zeros to convert it into a 130 length vector. Usually, a higher value of the maximum length is chosen, so as to incorporate chemicals with SMILES length longer than that present in the database.

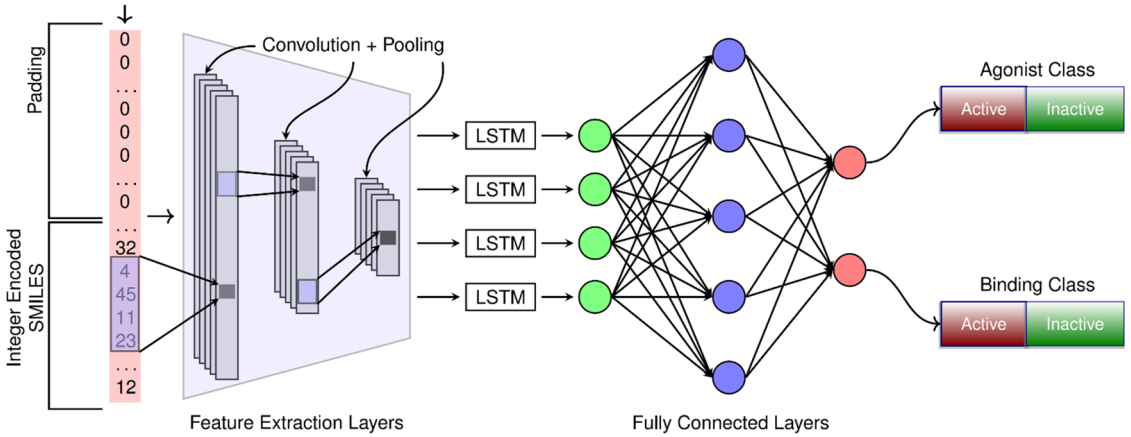


Figure 4. Feed-forward architecture of our proposed classification model (Task I). The input to the neural network (see Figure 3) is used to predict the toxicity of an unlabeled chemical as a potential endocrine disruptor. Detailed architecture is given in Table S1.

for Tox21 and 6432 data points for the literature data set. The detailed counts including the number of unique chemicals in each split set are given in Table 1. We have not used stratification to generate random splits to avoid overlapping since a chemical from the augmented data set can fall into multiple subgroups. Our data processing scheme allows us to work with a moderately large but balanced data set.

Each SMILES sequence $\mathbf{s}$ is a finite ordered list of symbols generated from a library of a finite set of characters D, $0 < n(D) < \infty$. Unlike traditional pattern classification, each sequence $\mathbf{s}$ is not a real-valued sequence, and all of the sequences are not of equal length. Each pair $(\mathbf{s}, l)$ is assumed to be independently and identically distributed (iid). The SMILES sequences are converted into fixed-length real-valued vectors following the framework shown in Figure 3. A well-known technique known as the bag-of-words[64] is adopted to encode the SMILES $\mathbf{s}$ into an integer-valued representation of the same length.

The characters of the set D are numbered from 1 to 57. Even though it is a random order, preserving this numbering order is essential to interpret the following classification framework results. The character vector SMILES strings $\mathbf{s}$ are converted into integer sequences $\mathbf{x}_e$ using the index of the characters from the set D using a one-to-one encoder mapping E, E: $\mathbf{s} \rightarrow \mathbf{x}_e$ whereby

$$\mathbf{s} = \{s_1, s_2, ..., s_{n_i}\},\ \mathbf{x}_e = \{x_1, x_2, ..., x_{n_i}\} D_{x_j} = s_j\ \forall\ j = 1 \text{ to } n_i \quad (1)$$

Additionally, each vector $\mathbf{x}_e$ is padded with $n - n_i$ 0 s to convert into a fixed-length vector as

$$\mathbf{x} = \{0, 0, ..., 0, x_1, x_2, ..., x_{n_i}\} \quad (2)$$

The index of D starts from 1. Thus, padding the encoded SMILES with 0 does not make any difference in the analysis but is done to apply the pattern recognition algorithm that requires a fixed-length vector. The encoded and padded SMILES and its toxicity label form the input−output pair $(\mathbf{x}, \mathbf{y})$ for the classifier.

Model. We have employed an Active Learner[65−69] that enables a classifier to perform inference by combining information drawn from both the available data sets.

https://doi.org/10.1021/acs.jcim.0c01409
J. Chem. Inf. Model. 2021, 61, 2187−2197

Simultaneously, an active learner controls the amount and quality of the data needed to extract to reach the desired accuracy. Iteratively, the learner combines a base classifier with a query strategy that extracts qualitative data from the evaluation (literature) data set into the primary training data set (Tox21) and thereby improves the accuracy of the classifier.

We have developed a deep neural network, "VisualTox", as our base classifier for the Active Learner. Neural networks are nonlinear stochastic approximation that uses an empirical but differentiable function from an input to an output and is often termed as a universal approximator.[70] Sequence-based neural networks commonly follow two types of architecture. While recurrent neural network or RNN[54−56] is used for sequence prediction, further feature extractions are performed using a combination of convolution neural network or CNN followed by an RNN.[71−74] Our base model follows a selective symbolic sequence classification model $C(\theta)$: $\mathbf{x} \rightarrow \mathbf{y}$ combining both the capabilities of a CNN and an RNN. It comprises a stack of convolution layers followed by a long short-term memory (LSTM) unit. Here, θ represents the parameter of C. The classification architecture of VisualTox is displayed in Figure 4.

The feature extractor of VisualTox comprises two major components: convolution layers and pooling. In a sequence type data such as a SMILES string, the property of the chemical is not only related to a particular substructure but also its neighboring substructures. A convolution layer produces an output $z(\mathbf{x}) \in \mathbb{R}^{n-2K}$ map from the input $\mathbf{x} \in \mathbb{R}^{n}$, where the ith element of z is related to the x_i atom and its neighboring atoms as

$$z_i = \sum_{j=-K}^{K} x_{i+j} w_j + b_j \tag{3}$$

where $\mathbf{w} = \{w_{-K}, w_{K+1}, ..., w_{K-1}, w_K\}$ is the convolution kernel and $\mathbf{b} = \{b_{-K}, b_{K+1}, ..., b_{K-1}, b_K\}$

both of sizes $2K + 1$ and K is a positive integer. Each one-dimensional (1D) convolution layer is followed by a rectified linear unit (ReLU) activation unit to overcome the vanishing gradient problem. The pooling layer that follows each convolution layer extracts rotation and position invariant features and further reduces the size of the output feature map $z(\mathbf{x})$. The output of the last convolution and pooling layers fed into the LSTM layer, and the output 1D vector serves as an input for successive fully connected layers.

VisualTox is a multioutput classification model that predicts the toxicity of a chemical with respect to both the agonist and binding class. While training the network, it assumes a weighted loss function of the individual loss functions for the two types of toxicities given as

$$L = 0.5L_1 + 0.5L_2 \tag{4}$$

where the individual losses for $i \in [1,2]$ are given as

$$L_i = -\frac{1}{N_{\text{train}}} \sum_j y_j \log(p(\hat{y}_j)) + (1 - y_j)\log(1 - p(\hat{y}_j))\hat{y}_j$$
$$= \frac{e^{\beta_j z}}{1 + e^{\beta_j z}} \tag{5}$$

The weight θ of the classifier C is estimated by solving the following optimization problem

$$\theta = \min_{\theta} L(\mathbf{x}, \mathbf{y}, \theta) \tag{6}$$

The Active Learner (Figure 5), as mentioned earlier, improves the accuracy of the VisualTox by combining the information

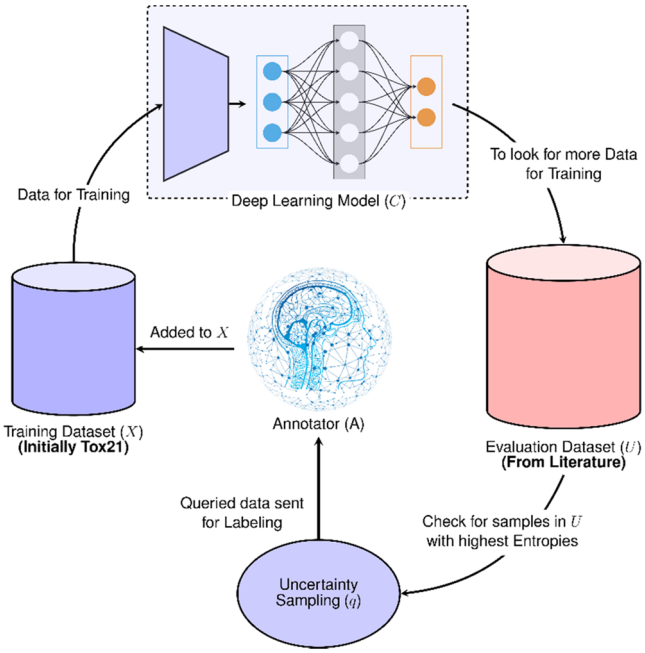


Figure 5. Active learning framework begins with the initial X data set for training that is Tox21. The whole data set is used for training (C) the Deep Learning model for a certain number of iterations iter_k = 150. After iter_k iterations, the evaluation data set U is looked into for more training data and the uncertainty sampling technique q is used to extract only those samples with k = 100 highest entropies. We use the literature as our annotator (A) to extract the labels for new k data points and add them to the training set X for further training of C.

from both the Tox21 and the literature data. It is a pool-based active learner that comprises three major components (X, C, q). C is a classifier C: $\mathbf{x} \rightarrow [0, 1]$ trained on a set of chemicals $\mathbf{x} \in X$ labeled for toxicity. The learning framework has a virtual annotator (A) that labels a set of unlabeled chemicals. Here, the annotator (A) is simply the literature that labels the literature data set. The query strategy (q) looks for chemicals in the unlabeled data set (U) or the literature data set that has the highest entropy with respect to the fitted model C. The samples with higher entropy reduce the variance of the prediction of the learner and hence reduce overfitting as shown in Section S4 of the Supporting Information. The labels of the queried data are extracted from the literature. The set of newly queried chemicals, along with the existing labeled database (X), is the new training data set for the classifier (C). This process continues until the query strategy reaches a specific decision-making step (in terms of accuracy or maximum iteration).

Active learning overcomes data insufficiency by controlling the amount of data that the VisualTox has so far classified unreliably. With each iteration interval, the learner draws samples from the secondary data, and the classifier is shown to improve accuracy for the same hyperparameters, such as the architecture of the VisualTox, optimizer learning rate, and optimality conditions. The theoretical foundation of the query strategy and the detailed algorithm of the Active Learner are given in the Supporting Information section.

https://doi.org/10.1021/acs.jcim.0c01409
J. Chem. Inf. Model. 2021, 61, 2187−2197

RESULTS

Performance on the Classification. We have listed down the different classification metrics in the Supporting Information Section S5. We have taken the classification accuracy or simply accuracy as the ratio of correct predictions to the total number of testing samples, and balanced accuracy (BA) that is the average of specificity and sensitivity, as our choices of metrics to compare model performances. A higher value of nearly 1 for both the metrics implies a good classification for a balanced data set. However, for an imbalanced data set, higher classification accuracy might not imply higher BA. We begin our analysis by taking the VisualTox architecture from Figure 4 and train the model using the unbalanced Tox21 data set consisting of 1677 chemicals and test on the 6206 chemicals from the literature data set. The training data set has 14.1% active binders and 13.1% active agonists, while the testing data set has 16.28% binders and 5.25% agonists. Training on highly imbalanced data by mini-batches enforces the learner to learn the over-represented class. As a result, either the sensitivity or specificity approaches the value of 1, but the other approaches the value 0. Our trained model produces accuracies and BAs of 0.8694 and 0.6369 for the Agonist class and 0.8587 and 0.6267 for the Binding class against the Tox21 data set. Meanwhile, the model produces accuracies and BAs of 0.9475 and 0.6779 for the Agonist class and 0.8372 and 0.5735 for the Binding class against the literature data set. While neural networks can be modified to learn data imbalance by controlling the representation of classes in the mini-batch and a weighted error measure, we have attempted to improve the accuracy of VisualTox by balancing and oversampling the data set as described in our Data Sets section. Furthermore, since training purely on Tox21 will show poor generalization error on the literature data set, we wish to further improve the network by performing Active Learning.

To obtain a more accurate and robust model, we have trained the VisutalTox architecture described in Figure 4 on different splits of the data set mentioned in Table 1. The names and descriptions of the three variations of VisualTox model based on different training and testing dataset are given in Table 2. We have also listed the balanced accuracies for these different variations of our proposed model in Table 2. All of the models are compared against a combined testing data set of 1419 samples from the Tox21 testing set and 1930 samples from the literature testing set. The BAs for different machine learning models are listed in ref 42 and are listed in the Supporting Information. All of the VisualTox models listed in Table 2 have been tested against the 1419 chemicals from the Tox21 testing set and the 1930 chemicals from the literature testing set, as mentioned in Table 1. It is shown that except for VisualTox_AL and VisualTox_Literature, the rest of the models perform well only on the Tox21 data set, while VisualTox_Literature performs poorly on the Tox21 data set. Also, the methods from ref 42 are trained on the complete Tox21 data set; hence, the generalization error for Tox21 is missing in these analyses. On the contrary, VisualTox_AL is trained on a combined training data set of both the Tox21 and the literature. Hence, it performs equally well on both the testing data sets. Active Learning achieves the comparable performance of VisualTox on the different data sets with significantly low generalization error.

Table 2. Descriptions of Model Names Based on the Choice of Data Set Used for Training and Their Testing Accuracies

		testing accuracy	
model name	description	agonist	binding
VisualTox_AL	trained using active learning (see Figure 5) using 3236 Tox21 training chemicals as a primary source of data and 4492 chemicals from literature training set as a secondary source of data.	0.9073	0.8958
VisualTox_Tox21	training on the 3236 chemicals from Tox21 training set only	0.7990	0.7932
VisualTox_Literature	training on the 4492 chemicals from literature training set only	0.8171	0.8014

The detailed accuracy values of VisualTox on the testing data set from Tox21 is presented in Table 3. VisualTox shows high values of classification accuracy, ROC-AUC, and PR-AUC for both the classes. A detailed description of the metrics used for quantifying the performance is given in the Supporting Information.

We have performed 10 random train-test splits on our data set before performing the data augmentation. Higher random trials help remove any selection bias that may arise due to choosing a fixed split set. We have calculated the mean (μ) and the standard deviation (σ) for the classification accuracies for VisalTox_AL on the testing data sets mentioned in Table 1 of the main text. The $\mu \pm 3\sigma$ limits are reported in Table 4.

Performance on the Visual Explanation. The stochastic approximation model C is a replica of an actual physical process. Thus, we need to explore the physical meaning of the estimated parameters or weight θ of the stochastic model. The fitted CNN breaks down the chemical into a collection of substructures that are statistically important toward determining the toxicity of the chemical. A clear understanding of how the intermediate layers are functioning leads us to more conclusive evidence than just performing the classification. Recent designs of interpretation techniques have allowed us to explore the deep nonlinear methods. One such approach is the use of feature maps[75] that involve visualizing all of the convolution filters and how the heatmap of a chemical behaves when it passes through the convolution and dense layers during the feed-forward propagation. Feature maps are excellent tools in identifying layer-by-layer unfolding of latent chemical space and chemical patterns. It shows the subregions of the SMILES string that get activated during the feed-forward propagation of the CNN model (see Figure 6b). It brings out the structural subgroups that contribute toward the classification of the chemical as active or inactive. However, feature maps alone cannot capture the variation in the information captured by CNN. Additionally, the activation of each filter may differ, and each activated filter's contribution is not also well captured.

We have used the gradient-weighted class activation mapping (Grad-CAM)[29] to extract the discriminating image regions and generate the molecular motifs that contribute toward the toxicity label predicted by the VisualTox (see Figure 6c). The chemical activation map is a weighted aggregate of all of the subgroups that have been identified by the feature-extraction layers of a CNN (see Figures S4, S5, and S6 for details). Grad-CAM is a modification of the class activation maps (CAM),[76] whereby it produces a weighted global average of the K feature maps $A^k \in \mathbb{R}^u$, $u < n$ from the last convolution layer. Grad-CAMs are computed for the

https://doi.org/10.1021/acs.jcim.0c01409
J. Chem. Inf. Model. 2021, 61, 2187−2197

Table 3. Performance of VisualTox_AL on Tox21 and Literature Testing Data Set

			precision	recall	F1-score	support[a]	accuracy[b]	ROC-AUC	PR-AUC
Tox21	agonist	active	0.91	0.92	0.92	675	0.9203	0.9751	0.9785
		inactive	0.92	0.92	0.92	744			
	binding	active	0.91	0.94	0.93	739	0.9225	0.9744	0.9748
		inactive	0.93	0.90	0.92	680			
literature	agonist	active	0.90	0.93	0.92	926	0.9187	0.9710	0.9691
		inactive	0.93	0.91	0.92	1004			
	binding	active	0.93	0.89	0.91	1016	0.9062	0.9651	0.9560
		inactive	0.88	0.93	0.90	914			

[a]Support refers to the number of occurrences of the class in a data set. [b]Accuracy is defined as the ratio of correct predictions to the total number of testing samples. It is different from balanced accuracy (BA).

Table 4. Mean (μ) and Standard Deviation σ Calculated for 10 Trials on the Testing Data Set[a]

	agonist class		binding class	
	Tox21	literature	Tox21	literature
$\mu + 3\sigma$	0.8995 ± 0.0372	0.9450 ± 0.0946	0.9040 ± 0.0327	0.9418 ± 0.1099

[a]The $\mu \pm 3\sigma$ limits are reported.

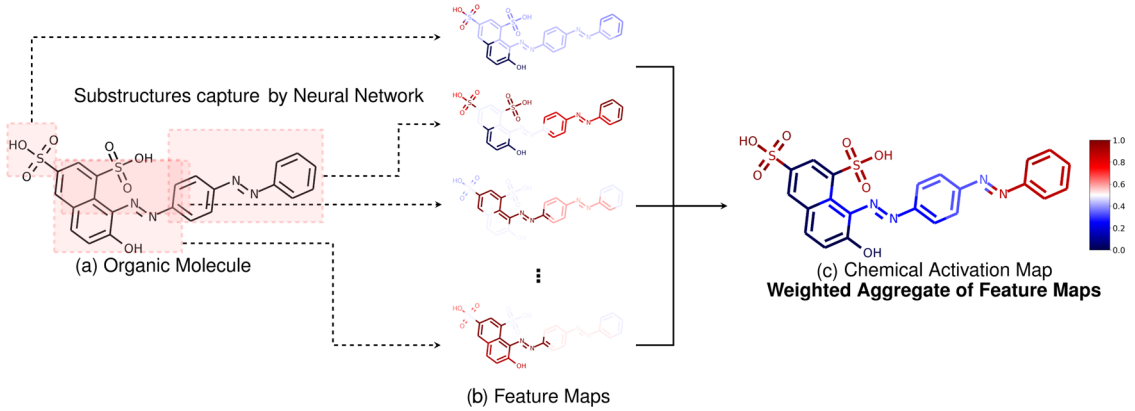


Figure 6. Chemical activation map is a weighted aggregate of all of the important geometric substructures of an organic molecule that are captured by a CNN while classifying it as an active compound. These substructures are captured as feature maps, which the CNN produces while performing its feed-forward propagation.

toxicity class c using the output score y^c that is predicted by the VisualTox. We consider the gradient of the output score y^c with respect to each feature map A^k. Thus, mathematically, the expression for Grad-CAM S^c for class c is given as

$$S^c = \mathrm{ReLU}\left(\sum_k^K w_k A^k\right) \quad w_k = \sum_i^w \frac{\partial y^c}{\partial A_i^k}$$

The scores $S^c \in \mathbb{R}^u$ obtained for a particular input image $\mathbf{x} \in \mathbb{R}^n$ are linearly interpolated to $\mathbb{R}^n$ using an open-source tool OpenCV.[77]

Structural alert, also known as "toxicophore", is a substructure of a chemical that accounts for the chemical's toxicity. The traditional way to identify structural motifs is to extract common functional groups from a large set of compounds by human expertize. One of the shortages of traditional structural alerts is the high possibility of false-positive, which means many compounds containing structural alerts are inactive (not having the specific toxicity). The false-positive rate of the structural alerts for endocrine disrupters can range from 0 to 100%.[78] On the other hand, the activation maps from VisualTox highlight the subgroups that contribute the most to the active or inactive classification. Simultaneously, they deemphasize the subgroups that are insignificant toward the classification. Hence, by comparing the activation maps of both the active and inactive compounds that contain the same or similar structural alerts, we will be able to understand how the "chemical environment" subgroups contribute to the compound's classification in addition to the structural alerts and how they influence the effectiveness of structural alerts.

We have generated Chemical activation maps for both the ER active and the ER inactive compounds from the Tox21 training set. Example chemical activation maps for the compounds from the categories identified as EDC[1] (e.g., alkylphenols, bisphenols, phthalates, and organochlorines) are shown in Figure 7. VisualTox classifies the compounds on the left column as active and the ones on the right as inactive regarding their ER agonist activity. The chemical activation maps can provide important information on two sides apart from accurate active/inactive classification. First, the CSMs—the motifs with higher contribution to the classification, shown in colors from orange to red for active compounds or from blue to dark blue for the inactive compounds as the contribution go higher; second, the non-CSMs—the motifs that contribute little to the classification results, usually in yellow for the actives and green for the inactive compounds. By observing the CSMs in Figure 7, we can find the classic

https://doi.org/10.1021/acs.jcim.0c01409
J. Chem. Inf. Model. 2021, 61, 2187−2197

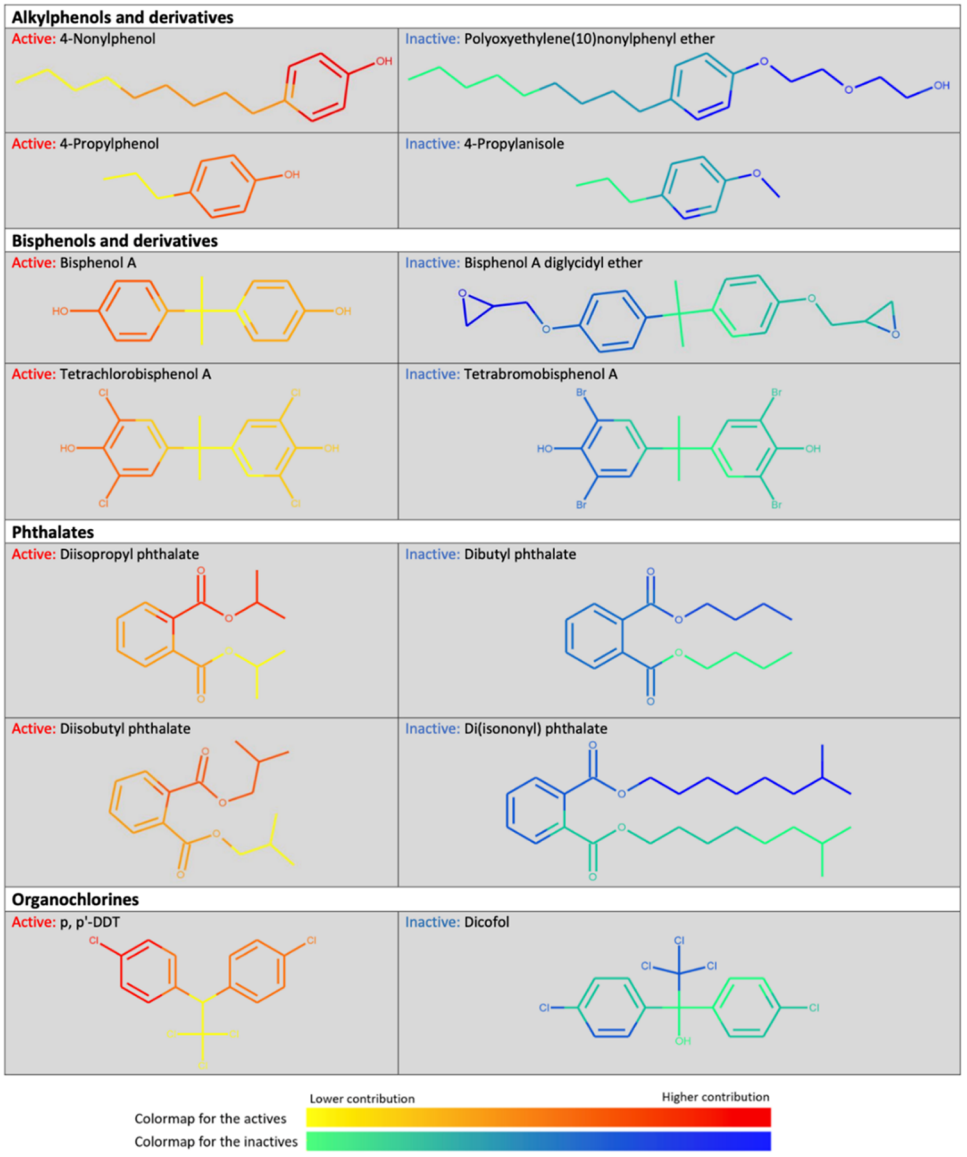


Figure 7. Examples of the activation maps of common endocrine disruptors from the CERAPP training set.

structural alerts (SAs), which are generalized by human expertize, can be part of the CSMs, while the CSMs can determine the specificity of SAs. For example, 4-nonylphenol and 4-propylphenol have their phenol groups and some alkyl groups shown in orange and red, which suggests an alkylated phenol group is a CSM when classified as an active chemical. On the contrary, the two inactive alkylphenol derivatives on the right have their ether groups shown in blue, suggesting that a phenol with hydroxyl hydrogen substituted is a CSM for inactive classification. The unsubstituted phenolic hydroxyl group is essential for estrogen receptors to distinguish between testosterone and estradiol/estrone, which has been confirmed by the X-ray structures of the complex of the estrogen receptor α ligand-binding domain with estradiol (ER-LBD-E_2 complex).[79] We compare the CSMs for the active and inactive compounds for alkylphenols and bisphenols and their derivatives. We can conclude that phenol, as a classic SA,[80] can be a hint for ER active compounds when the phenol hydroxyl hydrogen is not substituted, the phenyl ring is alkylated, and the hydroxyl group is not in-between two aryl bromides. Meanwhile, the CSMs can suggest new SAs with their specificity determined. From the CSMs identified from the phthalate esters (Figure 7), the phthalic acid ester can be a SA for estrogen disruptors. The specificity of the phthalic acid ester as a SA possibly depends on the length and linearity of its alkyl group: the active phthalates have one of their carboxylic and short branched alkyl groups as CSMs, while the CSMs for the inactive phthalates contain long-chain alkyls groups. We can conclude that the CSMs identified by VisualTox provide a new approach to understanding the molecular origin of chemical toxicity by highlighting the existing or suggested SAs as well as the chemical environments that determine the specificity of SAs.

CONCLUSIONS

Overall, we have developed a structural alert visualization toolkit using Grad-CAM that solves an underlying machine learning problem of multioutput classification. We have also shown how we can use a contradicting data set and strategically extract only a sufficient amount of data to reach the desired accuracy. The fitted statistical model combines feature-extraction and sequence prediction neural network

https://doi.org/10.1021/acs.jcim.0c01409
J. Chem. Inf. Model. 2021, 61, 2187−2197

architectures for a chemical descriptor that is made of elements from a finite set of characters. We have translated the patterns extracted from the subset of SMILES characters to meaningful chemical subgroups. The RDKit's inbuilt visualizer aids in drawing the chemical activation maps.

The chemical activation maps identify the critical structural motifs (CSMs) containing existing structural alerts (SAs) or suggest new SAs. Simultaneously they specify the conditions for the SAs to take effect, which can serve as a new approach to understanding the molecular origin of chemical toxicity.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c01409.

Tox21_train, Tox21_test, literature_train, literature_test (XLSX)

Complete list of unique and augmented chemicals used for training and testing as per Table 1 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Krishna Rajan − *Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260-1660, United States;* orcid.org/0000-0001-9303-2797; Email: krajan3@buffalo.edu

Authors

Arpan Mukherjee − *Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260-1660, United States;* orcid.org/0000-0001-5698-6268

An Su − *Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260-1660, United States;* orcid.org/0000-0002-6544-3959

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c01409

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the support from NSF Award #1640867—DIBBs: EI: Data Laboratory for Materials Engineering and the Collaboratory for a Regenerative Economy (CoRE center) in the Department of Materials Design and Innovation, University at Buffalo.

■ ADDITIONAL NOTE

[a]https://www3.epa.gov/research/COMPTOX/CERAPP_files.html.

■ REFERENCES

(1) Casals-Casas, C.; Desvergne, B. Endocrine disruptors: from endocrine to metabolic disruption. *Annu. Rev. Physiol.* **2011**, *73*, 135−162.

(2) Alves, V. M.; Muratov, E. N.; Capuzzi, S. J.; Politi, R.; Low, Y.; Braga, R. C.; Zakharov, A. V.; Sedykh, A.; Mokshyna, E.; Farag, S. Alarms about structural alerts. *Green Chem.* **2016**, *18*, 4348−4360.

(3) Dagan, I.; Engelson, S. P. Committee-Based Sampling for Training Probabilistic Classifiers. In *Machine Learning Proceedings*; Elsevier, 1995, pp 150−157.

(4) Kavlock, R. J.; Austin, C. P.; Tice, R. R. Toxicity testing in the 21st century: implications for human health risk assessment. *Risk Anal.* **2009**, *29*, 485−487.

(5) Tice, R. R.; Austin, C. P.; Kavlock, R. J.; Bucher, J. R. Improving the human hazard characterization of chemicals: a Tox21 update. *Environ. Health Perspect.* **2013**, *121*, 756−765.

(6) Thomas, R. S.; Paules, R. S.; Simeonov, A.; Fitzpatrick, S. C.; Crofton, K. M.; Casey, W. M.; Mendrick, D. L. The US Federal Tox21 Program: A strategic and operational plan for continued leadership. *Altex* **2018**, *35*, 163−168.

(7) Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* **2007**, *95*, 5−12.

(8) Judson, R. S.; Houck, K. A.; Kavlock, R. J.; Knudsen, T. B.; Martin, M. T.; Mortensen, H. M.; Reif, D. M.; Rotroff, D. M.; Shah, I.; Richard, A. M.; et al. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ. Health Perspect.* **2010**, *118*, 485−492.

(9) Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; et al. ToxCast chemical landscape: paving the road to 21st century toxicology. *Chem. Res. Toxicol.* **2016**, *29*, 1225−1251.

(10) Kavlock, R.; Chandler, K.; Houck, K.; Hunter, S.; Judson, R.; Kleinstreuer, N.; Knudsen, T.; Martin, M.; Padilla, S.; Reif, D.; et al. Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. *Chem. Res. Toxicol.* **2012**, *25*, 1287−1302.

(11) Attene-Ramos, M. S.; Miller, N.; Huang, R.; Michael, S.; Itkin, M.; Kavlock, R. J.; Austin, C. P.; Shinn, P.; Simeonov, A.; Tice, R. R.; et al. The Tox21 robotic platform for the assessment of environmental chemicals−from vision to reality. *Drug Discovery Today* **2013**, *18*, 716−723.

(12) Huang, R.; Sakamuru, S.; Martin, M. T.; Reif, D. M.; Judson, R. S.; Houck, K. A.; Casey, W.; Hsieh, J.-H.; Shockley, K. R.; Ceger, P.; et al. Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Sci. Rep.* **2014**, *4*, No. 5664.

(13) Attene-Ramos, M. S.; Huang, R.; Michael, S.; Witt, K. L.; Richard, A.; Tice, R. R.; Simeonov, A.; Austin, C. P.; Xia, M. Profiling of the Tox21 chemical collection for mitochondrial function to identify compounds that acutely decrease mitochondrial membrane potential. *Environ. Health Perspect.* **2015**, *123*, 49−56.

(14) Hsu, C.-W.; Zhao, J.; Huang, R.; Hsieh, J.-H.; Hamm, J.; Chang, X.; Houck, K.; Xia, M. Quantitative high-throughput profiling of environmental chemicals and drugs that modulate farnesoid X receptor. *Sci. Rep.* **2014**, *4*, No. 6437.

(15) Tang, W.; Chen, J.; Wang, Z.; Xie, H.; Hong, H. Deep learning for predicting toxicity of chemicals: A mini review. *J. Environ. Sci. Health, Part C* **2018**, *36*, 252−271.

(16) Zhang, L.; Zhang, H.; Ai, H.; Hu, H.; Li, S.; Zhao, J.; Liu, H. Applications of machine learning methods in drug toxicity prediction. *Curr. Top. Med. Chem.* **2018**, *18*, 987−997.

(17) Wu, Y.; Wang, G. Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. *Int. J. Mol. Sci.* **2018**, *19*, No. 2358.

(18) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* **2016**, *3*, No. 80.

(19) Xu, Y.; Pei, J.; Lai, L. Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *J. Chem. Inf. Model.* **2017**, *57*, 2672−2685.

(20) Fernandez, M.; Ban, F.; Woo, G.; Hsing, M.; Yamazaki, T.; LeBlanc, E.; Rennie, P. S.; Welch, W. J.; Cherkasov, A. Toxic colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images. *J. Chem. Inf. Model.* **2018**, *58*, 1533−1543.

(21) Webel, H. E.; Kimber, T. B.; Radetzki, S.; Neuenschwander, M.; Nazaré, M.; Volkamer, A. Revealing cytotoxic substructures in

molecules using deep learning. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 731−746.
(22) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(23) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(24) Weininger, D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237−243.
(25) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(26) Hirohara, M.; Saito, Y.; Koda, Y.; Sato, K.; Sakakibara, Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinf.* **2018**, *19*, No. 526.
(27) Zheng, S.; Yan, X.; Yang, Y.; Xu, J. Identifying Structure−Property Relationships through SMILES Syntax Analysis with Self-Attention Mechanism. *J. Chem. Inf. Model.* **2019**, *59*, 914−923.
(28) Karpov, P.; Godin, G.; Tetko, I. V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J. Cheminf.* **2020**, 1−12.
(29) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. In *Grad-cam: Visual Explanations from Deep Networks via Gradient-Based Localization*, Proceedings of the IEEE International Conference on Computer Vision, 2017; pp 618−626.
(30) Gunning, D. *Explainable Artificial Intelligence (xai)*; Defense Advanced Research Projects Agency (DARPA), nd Web, 2017.
(31) Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V. N. In *Grad-cam++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks*, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018; pp 839−847.
(32) Lu, J.; Xiong, C.; Parikh, D.; Socher, R. In *Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
(33) Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I. S. In *CBAM: Convolutional Block Attention Module*, Proceedings of the European Conference on Computer Vision (ECCV), 2018.
(34) Li, L.; Qin, L.; Xu, Z.; Yin, Y.; Wang, X.; Kong, B.; Bai, J.; Lu, Y.; Fang, Z.; Song, Q.; Cao, K.; Liu, D.; Wang, G.; Xu, Q.; Fang, X.; Zhang, S.; Xia, J.; Xia, J. Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. *Radiology* **2020**, *296*, E65−E71.
(35) Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; Seekins, J.; Mong, D. A.; Halabi, S. S.; Sandberg, J. K.; Jones, R.; Larson, D. B.; Langlotz, C. P.; Patel, B. N.; Lungren, M. P.; Ng, A. Y. In *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*, Proceedings of the AAAI Conference on Artificial Intelligence, 2019; pp 590−597.
(36) Tiulpin, A.; Thevenot, J.; Rahtu, E.; Lehenkari, P.; Saarakkala, S. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. *Sci. Rep.* **2018**, *8*, No. 1727.
(37) Han, S. S.; Kim, M. S.; Lim, W.; Park, G. H.; Park, I.; Chang, S. E. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *J. Invest. Dermatol.* **2018**, *138*, 1529−1538.
(38) Kondo, R.; Yamakawa, S.; Masuoka, Y.; Tajima, S.; Asahi, R. Microstructure recognition using convolutional neural networks for prediction of ionic conductivity in ceramics. *Acta Mater.* **2017**, *141*, 29−38.
(39) Svensson, F.; Norinder, U.; Bender, A. Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. *Toxicol. Res.* **2017**, *6*, 73−80.
(40) Banerjee, P.; Dehnbostel, F. O.; Preissner, R. Prediction Is a balancing act: importance of sampling methods to balance sensitivity and specificity of predictive models based on imbalanced chemical data sets. *Front. Chem.* **2018**, *6*, No. 362.
(41) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.
(42) Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A. M.; Grulke, C. M.; Trisciuzzi, D.; Fourches, D.; Horvath, D.; Benfenati, E.; Muratov, E.; Wedebye, E. B.; Grisoni, F.; Mangiatordi, G. F.; Incisivo, G. M.; Hong, H.; Ng, H. W.; Tetko, I. V.; Balabin, I.; Kancherla, J.; Shen, J.; Burton, J.; Nicklaus, M.; Cassotti, M.; Nikolov, N. G.; Nicolotti, O.; Andersson, P. L.; Zang, Q.; Politi, R.; Beger, R. D.; Todeschini, R.; Huang, R.; Farag, S.; Rosenberg, S. A.; Slavov, S.; Hu, X.; Judson, R. S. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect.* **2016**, *124*, 1023−1033.
(43) Wang, J.; Xu, M.; Wang, H.; Zhang, J. In *Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding*, 2006 8th International Conference on Signal Processing, 2006.
(44) Crammer, K.; Kearns, M.; Wortman, J. Learning from multiple sources. *J. Mach. Learn. Res.* **2008**, *9*, 1757−1774.
(45) Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48*, 312−320.
(46) Williams, D. P.; Park, B. K. Idiosyncratic toxicity: the role of toxicophores and bioactivation. *Drug Discovery Today* **2003**, *8*, 1044−1050.
(47) Williams, D. P. Toxicophores: investigations in drug safety. *Toxicology* **2006**, *226*, 1−11.
(48) Yan, Y.; Rosales, R.; Fung, G.; Farooq, F.; Rao, B.; Dy, J. Active learning from multiple knowledge sources. *Artif. Intell. Stat.* **2012**, 1350−1357.
(49) Liu, X.-Y.; Wu, J.; Zhou, Z.-H. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst., Man, Cybern.* **2008**, *39*, 539−550.
(50) Yen, S.-J.; Lee, Y.-S. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Syst. Appl.* **2009**, *36*, 5718−5727.
(51) Bao, L.; Juan, C.; Li, J.; Zhang, Y. Boosted Near-miss Under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets. *Neurocomputing* **2016**, *172*, 198−206.
(52) Mani, I.; Zhang, I. In *kNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction*, Proceedings of Workshop on Learning from Imbalanced Datasets, 2003.
(53) Bjerrum, E. J. Smiles Enumeration as Data Augmentation for Neural Network Modeling of Molecules. 2017, arXiv:1703.07076. arXiv.org e-Print archive. http://arxiv.org/abs/1703.07076.
(54) Sutskever, I.; Vinyals, O.; Le, Q. V. Sequence to Sequence Learning with Neural Networks. 2014, arXiv:1409.3215. arXiv.org e-Print archive. http://arxiv.org/abs/1409.3215.
(55) Reimers, N.; Gurevych, I. Reporting Score Distributions Makes a Difference: Performance Study of lstm-Networks for Sequence Tagging 2017, arXiv:1707.09861. arXiv.org e-Print archive. http://arxiv.org/abs/1707.09861.
(56) Park, S. H.; Kim, B.; Kang, C. M.; Chung, C. C.; Choi, J. W. In *Sequence-to-Sequence Prediction of Vehicle Trajectory via LSTM Encoder-Decoder Architecture*, 2018 IEEE Intelligent Vehicles Symposium (IV), 2018; pp 1672−1678.
(57) Judson, R. S.; Magpantay, F. M.; Chickarmane, V.; Haskell, C.; Tania, N.; Taylor, J.; Xia, M.; Huang, R.; Rotroff, D. M.; Filer, D. L.; et al. Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor. *Toxicol. Sci.* **2015**, *148*, 137−154.
(58) Collins, F. S.; Gray, G. M.; Bucher, J. R. Transforming environmental health protection. *Science* **2008**, *319*, 906−907.

https://doi.org/10.1021/acs.jcim.0c01409
J. Chem. Inf. Model. 2021, 61, 2187−2197

(59) Shen, J.; Xu, L.; Fang, H.; Richard, A. M.; Bray, J. D.; Judson, R. S.; Zhou, G.; Colatsky, T. J.; Aungst, J. L.; Teng, C.; et al. EADB: an estrogenic activity database for assessing potential endocrine activity. *Toxicol. Sci.* **2013**, *135*, 277−291.
(60) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(61) Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A. *Classification and Regression Trees*; CRC Press, 1984.
(62) Lemaître, G.; Nogueira, F.; Aridas, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 559−563.
(63) Bjerrum, E. J.; Sattarov, B. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules* **2018**, *8*, No. 131.
(64) Zhang, Y.; Jin, R.; Zhou, Z.-H. Understanding bag-of-words model: a statistical framework. *Int. J. Mach. Learn. Cybern.* **2010**, *1*, 43−52.
(65) Olsson, F. *A Literature Survey of Active Machine Learning in the Context of Natural Language Processing*, 2009.
(66) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, No. 241733.
(67) Yuan, R.; Liu, Z.; Balachandran, P. V.; Xue, D.; Zhou, Y.; Ding, X.; Sun, J.; Xue, D.; Lookman, T. Accelerated discovery of large electrostrains in BaTiO3-based piezoelectrics using active learning. *Adv. Mater.* **2018**, *30*, No. 1702884.
(68) Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513−524.
(69) Reker, D.; Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today* **2015**, *20*, 458−465.
(70) Sonoda, S.; Murata, N. Neural network with unbounded activation functions is universal approximator. *Appl. Comput. Harmonic Anal.* **2017**, *43*, 233−268.
(71) Ma, X.; Hovy, E. End-to-End Sequence Labeling via Bi-Directional lstm-cnns-crf. 2016, arXiv:1603.01354. arXiv.org e-Print archive. http://arxiv.org/abs/1603.01354.
(72) Bae, S. H.; Choi, I.; Kim, N. S. In *Acoustic Scene Classification Using Parallel Combination of LSTM and CNN*, Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), 2016; pp 11−15.
(73) Ullah, A.; Ahmad, J.; Muhammad, K.; Sajjad, M.; Baik, S. W. Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE Access* **2017**, *6*, 1155−1166.
(74) Mishra, P.; Khurana, K.; Gupta, S.; Sharma, M. K. In *VMAnalyzer: Malware Semantic Analysis using Integrated CNN and Bi-Directional LSTM for Detecting VM-level Attacks in Cloud*, 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019; pp 1−6.
(75) Zeiler, M. D. ADADELTA: An Adaptive Learning Rate Method. 2012, arXiv:1212.5701. arXiv.org e-Print archive. http://arxiv.org/abs/1212.5701.
(76) Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. In *Learning Deep Features for Discriminative Localization*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016; pp 2921−2929.
(77) Bradski, G. *The OpenCV Library*, Dr Dobb's J. Software Tools, 2000.
(78) Müller, M.; Wenzel, A.; Nendza, M.; Lewin, G. *Assessment and Regulation of Environmental Hormones (Sub-Project 5): Development of Structure-and Risk-Based Methods for the Identification of Chemicals with Suspected Endocrine Disrupting Activities for the Prioritisation of Chemicals as Part of the Authorisation Procedure under REACH*; Forschungsprojekt im Auftrag des Umweltbundesamtes, FuE-Vorhaben FKZ, 2009; p 448.
(79) Brzozowski, A. M.; Pike, A. C. W.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engström, O.; Öhman, L.; Greene, G. L.; Gustafsson, J.-Å.; Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **1997**, *389*, 753−758.
(80) Stepan, A. F.; Walker, D. P.; Bauman, J.; Price, D. A.; Baillie, T. A.; Kalgutkar, A. S.; Aleo, M. D. Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States. *Chem. Res. Toxicol.* **2011**, *24*, 1345−1410.

https://doi.org/10.1021/acs.jcim.0c01409
J. Chem. Inf. Model. 2021, 61, 2187−2197