JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING
Cite This: J. Chem. Inf. Model. 2019, 59, 1062−1072
Article
pubs.acs.org/jcim

Comparative Study of Multitask Toxicity Modeling on a Broad Chemical Space

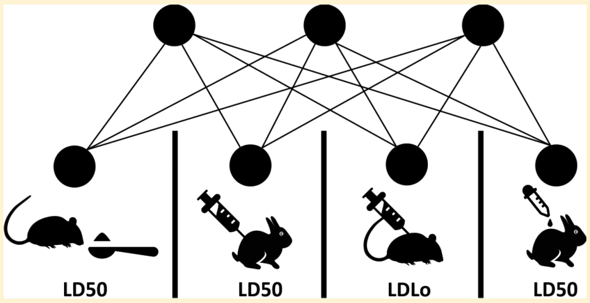
Sergey Sosnin,*[†] Dmitry Karlov,[†] Igor V. Tetko,*[‡] and Maxim V. Fedorov*[†,¶]

[†]Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Moscow 143026, Russia

[‡]Helmholtz Zentrum München—Research Center for Environmental Health (GmbH), Institute of Structural Biology and BIGCHEM GmbH, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany

[¶]University of Strathclyde, Department of Physics, John Anderson Building, 107 Rottenrow East, Glasgow, U.K. G40NG

Supporting Information



ABSTRACT: Acute toxicity is one of the most challenging properties to predict purely with computational methods due to its direct relationship to biological interactions. Moreover, toxicity can be represented by different end points: it can be measured for different species using different types of administration, etc., and it is questionable if the knowledge transfer between end points is possible. We performed a comparative study of prediction multitask toxicity for a broad chemical space using different descriptors and modeling algorithms and applied multitask learning for a large toxicity data set extracted from the Registry of Toxic Effects of Chemical Substances (RTECS). We demonstrated that multitask modeling provides significant improvement over single-output models and other machine learning methods. Our research reveals that multitask learning can be very useful to improve the quality of acute toxicity modeling and raises a discussion about the usage of multitask approaches for regulation purposes. Our MultiTox models are freely available in OCHEM platform (ochem.eu/multitox) under CC-BY-NC license.

INTRODUCTION

Toxicity is defined as the potential for a chemical compound to cause injury.[1] Accurate prediction of toxicity of organic compounds is one of the most challenging tasks in medicinal chemistry and pharmacology. According to a study,[2] nearly 30% of drug candidates fail in the first stage of clinical trials due to a presence of nondesired side effects, which results in a cost increase for the pharmaceutical industry. This fact emphasizes that current methods for "*in silico*" toxicity estimation, as well as experimental techniques, have serious shortcomings and that development of the new methods is of the utmost interest. Because it is involved in many organism's systems and metabolic pathways, toxicity can not be easily modeled solely by calculation. Moreover, the methodology of the experiments that measure toxicity, and the statistical analysis of the data obtained is under criticism.[3,4]

Toxicity estimation can be performed in two main ways: *in vivo* using rodent models or clinical trials data and *in vitro* using cell-based bioassays. The former approach allows for the estimation of the toxic effect, at the organism level, producing comprehensive results, and is widely used in preclinical tests. It should be noted that rodent models are not fully representative of humans and their use can thus result in unexpected side effects, which can be observed during clinical trials or even after drug approval.[5] The fact that *in vitro* tests are relatively inexpensive facilitates automation and makes their use possible in high-throughput screening (HTS).[6] The different types of toxicity mechanisms can be detected by using different assay types. Currently, there is a great demand for development of new reliable relevant assays for, e.g., nephrotoxicity.[7] However, due to their biological complexity, the *in vitro* tests do not also always provide a reliable estimation of the *in vivo* toxicity because human cell-based data used in *in vitro* testing may not take into account the general systemic toxicity for the whole organism. At the same time, the *in vivo* based rodent models do not always correctly represent human toxicity. Thus, there is a strong interest in and hope that the development of computational techniques could account for the drawbacks of these methods and help to reliably predict toxicity.[8]

Currently, a large amount of information has been accumulated and kept in commercial and open source databases. Some examples of the open source databases are the TOXNET database[9] and DSSTox,[10] which includes Tox21 high throughput data and the ChEMBL[11] database containing approximately 15 million of bioactivities. Among the proprietary databases, the Registry of Toxic Effects of Chemical Substances (RTECS)[12] database is the most valuable, and it contains information about 187 000 chemical substances. It has *in vivo* data for acute toxicity, skin irritation,

Special Issue: Machine Learning in Drug Discovery

Received: October 3, 2018
Published: December 27, 2018

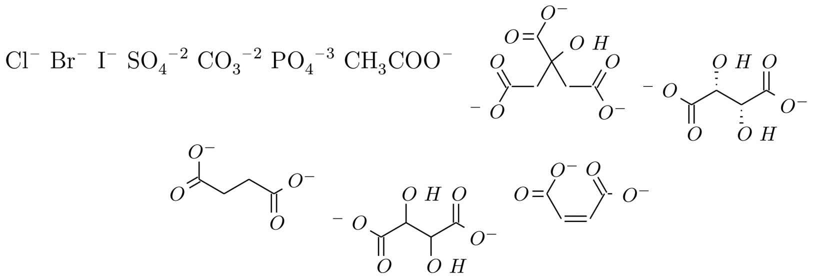

Cl^- Br^- I^- SO_4^{-2} CO_3^{-2} PO_4^{-3} CH_3COO^-

Figure 1. Ions considered to be nontoxic.

tumorogenic properties, and other effects measured for different organisms such as rodents, rabbits, and many others.

The recent advances in accessibility of bioactivity data in these and other databases prompted the development of high quality prognostic models created using various machine learning methods. For example, the PASS software (and web-service)[13] based on the Naive Bayes approach and trained using ChEMBL demonstrates good reliability when performing the classification task on a set of more than 2500 protein targets. The EMolTox web-service[14] predicts different types of toxicity using random forests and conformational prediction as measure of confidence and simultaneously visualizes the ToxAlert substructures on the molecular graph. The ProTox Web server is another tool for prediction of acute toxicity and other types of toxicity,[15] which utilizes a nearest neighbor approach combined with fingerprint similarity assessment. There are a number of models constructed for a narrow class of chemical compounds[16−18] or the certain model organism;[19,20] however, the applicability domain of such models is limited. The toxicity of chemical compounds is estimated using different types of biological assays which describe various toxic effects (neurotoxicity, cardiotoxicity, etc.), model organisms (rodents, dogs, monkeys), or the toxicity outcome (LD50, LD100). Only a few compounds are investigated in several assays and unavailability of experimental data in all assays may prevent detection of their toxicity. However, since the toxicity data sets are correlated, we can expect that such correlations could help to develop models with higher predictivity for each data point by modeling such data sets simultaneously (multitask learning). The previously mentioned RTECS data set, which contains data for different species and end points, is especially interesting for such a study. However, this data set is not widely used for the development of predictive models. We are only aware that part of it was used for mapping and chemical space visualization of the IDDB data set.[21] In this study we have addressed this question by using multitask learning[22−24] with state of the art machine learning methods.

■ MATERIALS AND METHODS

Data Set. We used RTECS[12] database version 2018.1 to extract organic compounds with acute toxicity records available. Since the structures of organic compounds are not presented in the database, we extracted them from PubChem[25] using Chemical Abstract Structure (CAS) registration numbers. The nonorganic compounds, plant extracts, parts of biological compounds, and compounds containing elements other than (C, H, O, N, P, S, F, Cl, Br, I) were ignored.

The goal of this study was to examine the toxic effects of the organic compounds. However, many compounds were reported in the database as salts or as mixtures, and some of the counterions are toxic themselves, e.g. methylsulfate ion ($CH_3OSO_4^-$). Their toxicity could interfere with the interpretation of the toxicity of the organic part. Therefore, only compounds with nontoxic counterions listed in Figure 1 were kept in the database. The compounds with other counterions and compounds with mixtures were eliminated. We also eliminated all polymeric substances. For the salts which were kept in the database, only the organic part was used to generate descriptors. If after salts removal there were more than one compound with the same structure, we used the median values of measurements.

After the preprocessing stage, all compounds were grouped for the same toxicity type by two parameters: the route of administration and the animal species used for the experiment. We removed all records that had less than 300 reported measurements for each group to reduce the dimensionality of the output. As the result, a database with 129 142 toxicity measurements was created. It consists of 87 064 unique molecular structures and 29 unique end points. The sparsity (the percentage of the filled values) of the data matrix is 5.12%. The information on the end points is summarized in Table 1

Molecular Descriptors. Different descriptor sets may have different performance in the modeling of toxicity.[26,27] The investigation of different sets of descriptors for the performance of single and multitask models could help better to understand whether the performance of models depends on the used descriptor sets. Therefore, we calculated a number of molecular descriptor sets which are provided by the OCHEM platform. A short description of the descriptors used is given in Table 2.

Machine Learning Methods. In this work we used deep neural networks (DNN) as well as several other machine methods that are gaining a lot of popularity in the machine learning community. Below, we provide a brief overview of these methods:

Deep Neural Networks. DNNs are now the state of the art methods for the development of models across different areas of science and technology. Their efficiency was confirmed for bioinformatics, medicinal applications (e.g., tissue image analysis and recognition of pathologies from voice analysis), prediction of chemical compounds properties, and other areas. In some cases these approaches provided higher prediction accuracy compared to all previously published models. Moreover, it was shown that further improvement of these methods could come from their application to multitask data sets. An artificial neural network (ANN) is a function that maps points from the input space to the output space. Deep ANNs commonly consist of several sequential layers where each layer represents a linear vector transformation $\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$ where $\boldsymbol{W}$ is a matrix of tunable weights and $\boldsymbol{b}$ is a bias vector, followed by a nonlinear transformation function (i.e., sigmoid). The training procedures use several techniques, such as batch

DOI: 10.1021/acs.jcim.8b00685
J. Chem. Inf. Model. 2019, 59, 1062−1072

Table 1. End Points Extracted from RTECS Dataset

species	administration	type of toxicity	no. of records
guinea pig	oral	lethal dose fifty	799
mammal, species unid.	unreported	lethal dose fifty	1121
man	oral	toxic dose low	512
mouse	intraperitoneal	lethal dose fifty	37202
mouse	intraperitoneal	lethal dose low	2965
mouse	intraperitoneal	toxic dose low	1057
mouse	intravenous	lethal dose fifty	17742
mouse	oral	lethal dose fifty	24355
mouse	oral	lethal dose low	1565
mouse	oral	toxic dose low	646
mouse	subcutaneous	lethal dose fifty	7221
mouse	subcutaneous	lethal dose low	921
mouse	unreported	lethal dose fifty	1804
rat	intraperitoneal	lethal dose fifty	5041
rat	intraperitoneal	lethal dose low	1029
rat	intraperitoneal	toxic dose low	1117
rat	intravenous	lethal dose fifty	2538
rat	intravenous	toxic dose low	608
rat	oral	lethal dose fifty	10743
rat	oral	lethal dose low	966
rat	oral	toxic dose low	955
rat	subcutaneous	lethal dose fifty	2014
rat	subcutaneous	toxic dose low	555
rat	skin	lethal dose fifty	930
rat	unreported	lethal dose fifty	838
rabbit	intravenous	lethal dose fifty	764
rabbit	oral	lethal dose fifty	910
rabbit	skin	lethal dose fifty	1734
woman	oral	toxic dose low	490

Table 2. Descriptors Used in Our Experiment[a]

descriptor	short description
PyDescriptor (3D)[28]	a PyMOL-based plugin for calculations different groups of descriptors
Dragon6 (3D)[29]	descriptors provided by Dragon 6 program
SIRMS[30]	calculates simplexes, which are *n*-atoms fragments of a fixed composition, structure, chirality, and symmetry
StructuralAlerts[31]	presence of certain subfragments in molecular graphs which are believed to be related to toxicity of organic compounds
QNPR[32]	uses substrings of SMILES as a representation of molecules
Spectrophores (3D)[33]	spectrophores are one-dimensional descriptors that describe the three-dimensional molecular fields surrounding a molecule
Adriana (3D)	descriptors provided by the Adriana.CODE program
Inductive (3D)[34]	descriptors based on inductive and steric effects of atoms
Chemaxon (3D)	subset of descriptors calculated by Chemaxon (www.chemaxon.com) module in OCHEM
Mera and Mesry (3D)[35]	3D descriptors of molecules
GSFrag[35]	descriptors calculated by GSFRAG program (the occurrence numbers of certain special fragments on *k* = 2, ..., 10 vertices in a molecular graph)
Fragmentor[36]	molecular fragments which contains from 2−4 atoms generated by the ISIDA module in OCHEM
ALogPS,[37,41] OEstate[42]	prediction of logP by ALogPS2.1 program in combination with OEstate descriptors which are based on electrostatic properties of atoms and bonds
CDK2 (3D)[43]	chemistry development kit descriptors, version 2.0
Morgan fingerprints[38,39]	Morgan (circular) fingerprints of radius two (which corresponds to ECFP4[38]) calculated by RDKit[40]

[a] Several descriptor blocks that are indicated by “(3D)” required 3D representation of molecules, which was calculated by using 2D to 3D structure conversion using the *Corina* program.

normalization[44] and dropout,[45] which help to achieve faster convergence and prevent overfitting. Deep neural networks are also a good choice for multitasked approaches due to their simplicity of implementation and ability to handle a loss function explicitly. In our model each end point was represented as separate output of a DNN as it is shown in Figure 2.

The architecture and training parameters are given in the Supporting Information. We implemented our DNN in the Chainer[46] framework and included one into the OCHEM[47] platform.

XGBoost. Gradient boosted trees is one of the most prominent approaches in data mining. This is frequently among the leading algorithms in the Kaggle data science competition. It was shown that XGBoost can be very efficient for processing large chemical data set in terms of accuracy and speed of computation.[48] On each iteration of XGBoost a new decision tree is constructed to fit the residuals of the model obtained at the previous stage.

K-Nearest Neighbors. This is a popular method for QSAR/QSPR modeling in which the prediction is calculated as a mean (or weighted sum) of N compounds that are the nearest ones to the compound under investigation in some descriptor space.[49] The idea is close to chemical paradigm that similar compounds have alike properties. This method is frequently used in chemical modeling especially for small data sets.[50,51]

Random Forest. This method uses the set (forest) of the simple classifiers or regressors, namely decision trees.[52] This method has been heavily used in chemoinformatics for the past decade before the rise of deep learning due to a long list of advantages, particularly the performance of modeling, the speed of computation, and the ability to use default parameters or parameters with minimal tuning. It should be mentioned that this method has a long history of usage for toxicity prediction.[53,54]

Consensus. These models frequently improve the quality of predictions of toxicity of single models by combining top-ranked models.[55] We constructed consensus models by averaging of the predictions of top five individual models.

Model Validation and Statistical Performance Measurement. A number of common metrics to evaluate a statistical performance have been used: *root mean square error* (*RMSE*), *mean absolute error* (*MAE*), and R^2 in accordance with the formulas below:

$$RMSE = \sqrt{\frac{\sum_i^T (\hat{y}_i - y_i)^2}{T}}$$

$$MAE = \frac{\sum_i^T |\hat{y}_i - y_i|}{T}$$

$$R^2 = 1 - \frac{\sum_i^T (\hat{y}_i - y_i)^2}{\sum_i^T (y_i - \bar{y})^2}$$

where $\hat{y}_i$ is a predicted value, y_i is a real value, $\bar{y}$ is a mean value over all samples, and T is the number of samples. Overfitting of machine learning algorithms is a well-known problem resulting in inadequate performance estimations.[56] To combat this

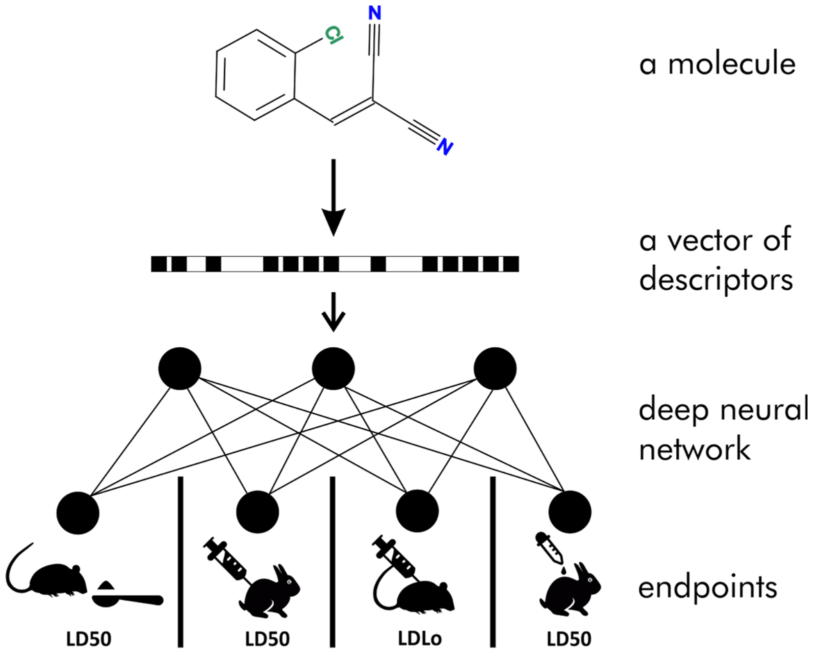


Figure 2. Representation of end points as outputs of a deep neural network.

problem and estimate the statistical performance in a robust way, a 5-*fold* cross-validation routine has been carried out for all models in this study. A graphical explanation of a cross-validation procedure is given in Figure 3.

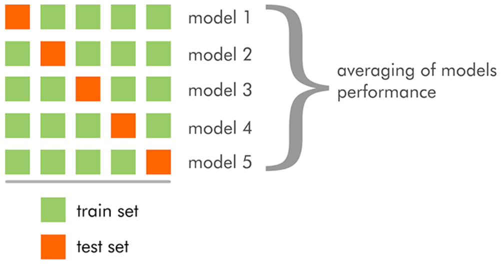


Figure 3. Scheme of 5-fold cross-validation procedure. On each fold, $\frac{4}{5}$ of a data set becomes a training set and $\frac{1}{5}$ becomes a test set, sliding over folds. The cross-validation is done based on molecules, and thus, all toxicity values for the same molecules are within the same set always.

It should be noted that OCHEM develops a new model on each validation step without using any information about the test compounds, which are only predicted following model developments. This provides correct validation (identical to the use of so-called "external sets") since no information about the test molecules is used to guide model development.

RESULTS AND DISCUSSION

Description of the Data Set Chemical Space. For the description of the whole data set, we took the highest value across all end points for each molecule. For the generation of the 2D chemical space representation the calculated RDKit[40] circular fingerprints (4096 bit vectors) based on the standardized SMILES molecular representation (molvs python package) were embedded into the 2D space using the t-SNE method.[57] A pairwise distance matrix was calculated using the Dice metric, and the default values were chosen for parameters of the algorithm. Figure 4 shows the results of the chemical space embedded in the 2D space. Each point corresponds to a chemical structure and the color denotes the toxicity values according to the palette. Some of the toxic clusters are highlighted by the rectangular shapes and their representative members are visualized in Figure 4. We provide the description of several clusters composed from the relatively toxic molecules. The enlarged image of cluster **K** is given for clarity and demonstrates its composition from the hydroxytriptamine derivatives. Arylcarbamate (neostigmine derivative is shown as a representative cluster member) derivatives are embedded into cluster **A** and their toxic effects may be explained by the cholinesterase inhibition. Cluster **B** is composed of possible nicotinic acetylcholine receptor ligands. The derivatives of the 3-quinuclidinyl benzilate which is a potent muscarinic anticholinergic agent are the major members of cluster **C**. Cluster **D**, similarly to cluster **B**, is composed of compounds based on the two quaternary amine groups connected by a linker. Phenothiazine derivatives acting on a number of different targets and widely used as antipsychotic agents earlier are the major components of cluster **E**. Phencyclidine derivatives (NMDA-receptor channel blocker) are included in cluster **F**. Possible alkylating agents and organophosphorus compounds were grouped in clusters **G** and **H**, respectively. Cluster **I** is composed of the adrenoreceptor ligands and the propranolol structure is shown for example in Figure 4. And isoquinoline derivatives belong to cluster **J**. This result shows that toxic compounds are grouped by similar structural features and neighbor compounds tend to have similar toxicity.

Correlation Analysis of End Points. Previous studies[58] pointed out that the efficiency of multitask modeling depends on correlations between targets. To examine it, a correlation analysis of end points was performed. Pearson correlation coefficients between each pair of end points were calculated. Mutual correlations as heatmaps are presented on Figure 5. For the objective evaluation of correlations, we set a number of thresholds. If the corresponding end points have the number of simultaneous measurements less than a threshold, the color on

DOI: 10.1021/acs.jcim.8b00685

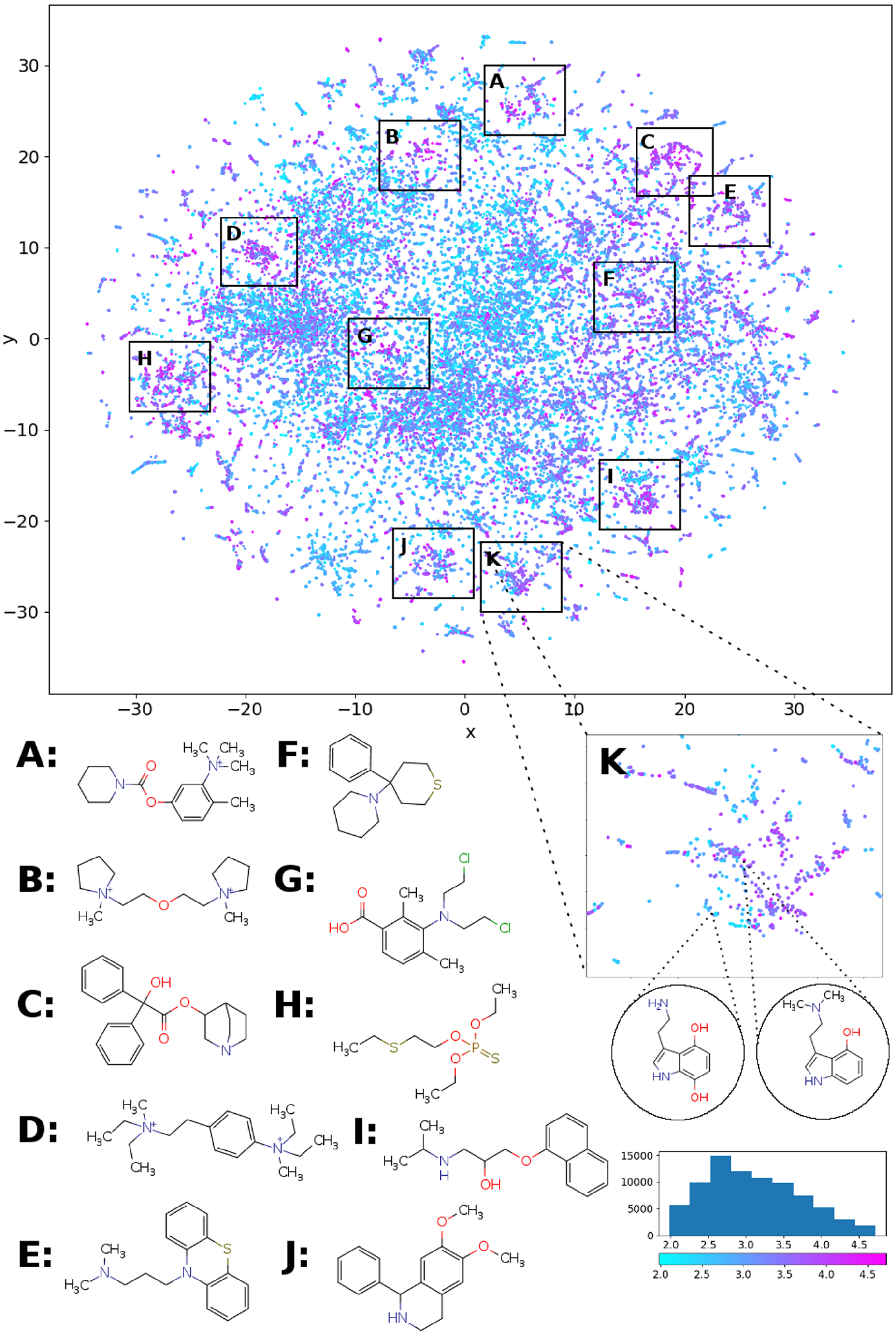


Figure 4. RTECS chemical space visualization. Each point stands for one molecular structure, and its color indicates the acute toxicity values (−log(mol/kg)).

the heatmap is absent. It is possible to observe that the correlations between end points are significantly high and it can explain the success of multitask modeling. The high correlations between end points also can reflect the good quality of the data presented in RTECS on the assumption that the provided measurements were independent.

Comparison of Models. Our main goal was to compare models of toxicity built for different end points. In this paper we defined each end point according to the conditions of the experiments. For example the LD50 toxicities measured when using intraperitoneal administration to mouse belong to the same end point. As a counterexample LD50 records with oral and intraperitoneal admission belong to different end points. However, due to hidden relations between end points we can expect that the multitask (multiend point) models should achieve better quality than single-task models. To prove the hypothesis we built multitask DNN models (*MT_DNN*), single target DNN models (*ST_DNN*), and several models with other aforementioned machine learning algorithms, namely: *XGBoost*, *random forest*, *K nearest neighbors*. In order to show that the observed relationships are not specific to a single set of descriptors, we used all sets of descriptors reported in Table 2. The performance of different models is given in Figure 6.

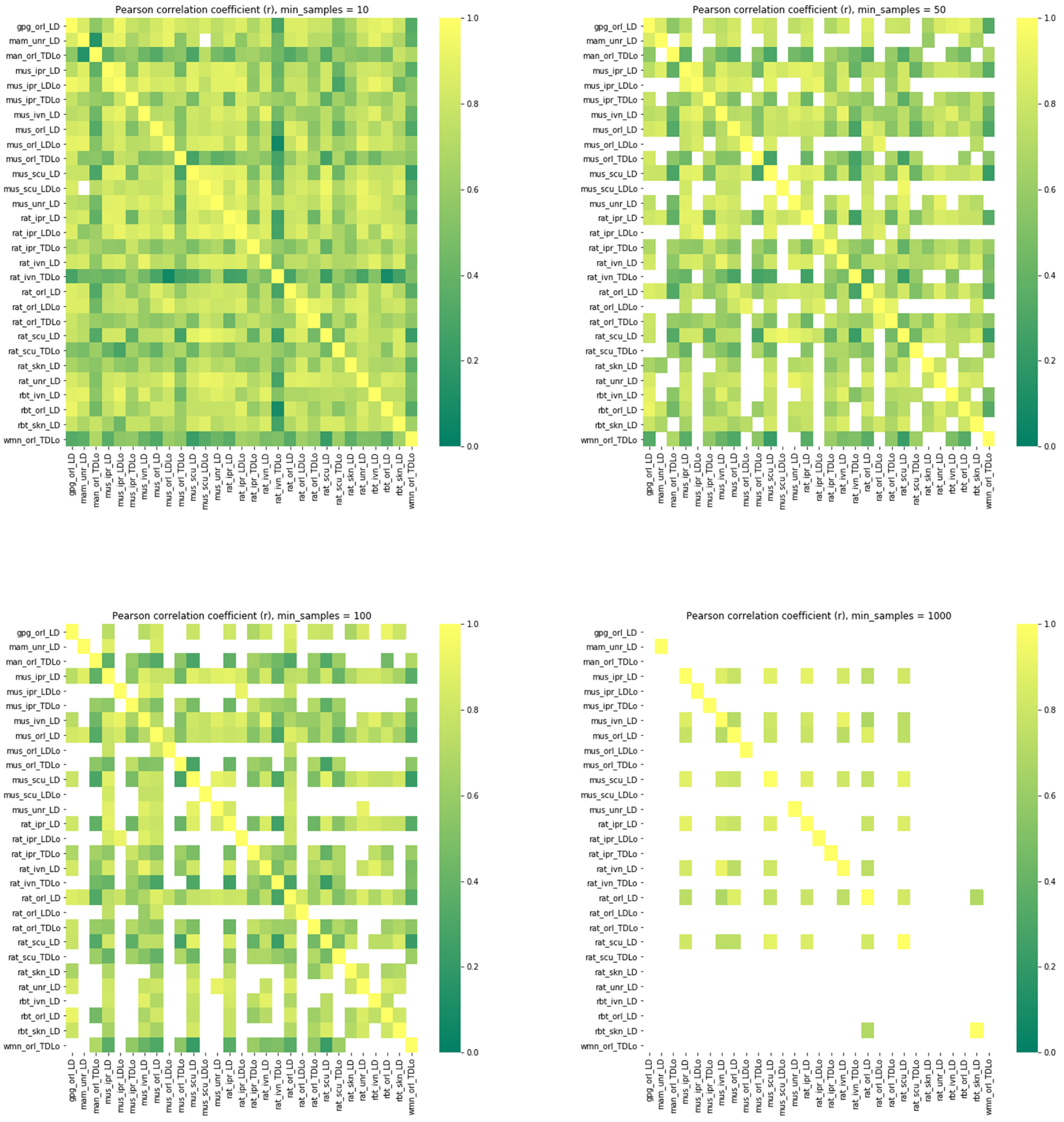


Figure 5. Correlation matrices for end points with various thresholds (*min_samples*) values. The toxicity end points demonstrate their correlation notwithstanding the number of compared samples.

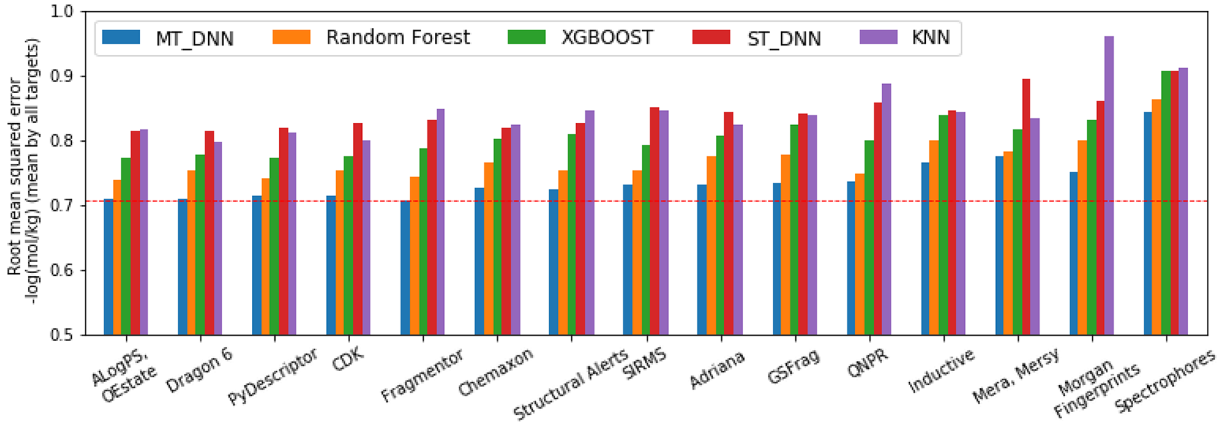


Figure 6. Average *RMSE* of predictions of toxicity for all end points (−log(mol/kg)) units by different methods and descriptor sets. Descriptors were arranged in accordance with mean values of predictions by all methods (the best are on the left). Methods are ordered by the mean *RMSE* over all descriptors (*MT_DNN* and *KNN* demonstrated highest and lowest overall performances, respectively).

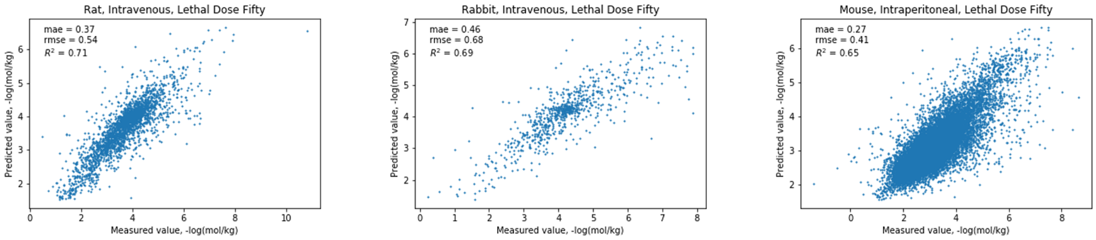


Figure 7. Prediction charts for a number of selected end points.

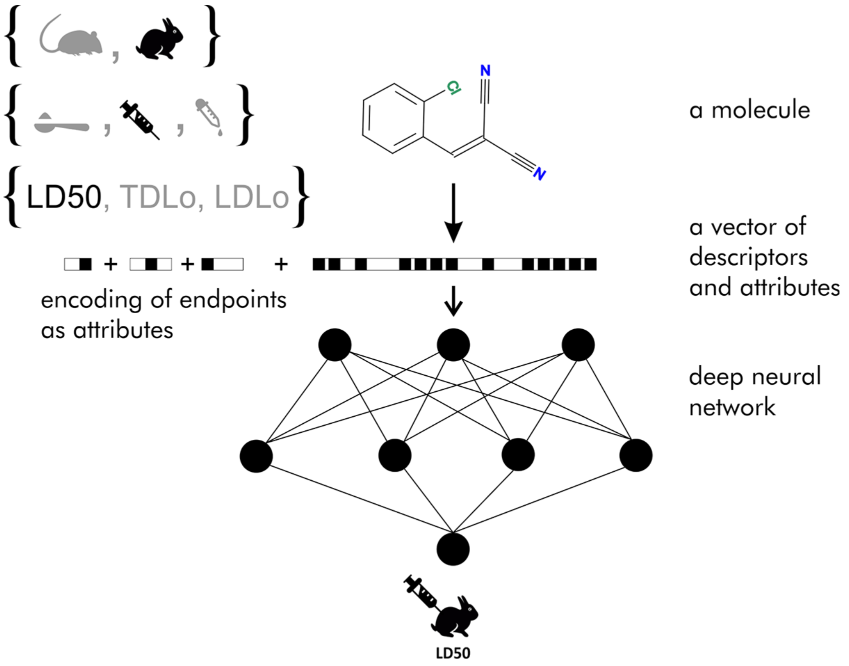


Figure 8. Representation of end points as attributes in STL modeling. The encoding of end points as input descriptors allows their simultaneous prediction using neural network with one output neuron.

The *MT_DNN* models outperformed both *ST_DNN* models and all other methods used for all analyzed sets of descriptors. Models, which are based on *ALogPS* combined with *OEstate* descriptors achieved the best average performances across all studied methods. The red dashed line on Figure 6 corresponds to average RMSE = 0.71 ± 0.01, which was calculated using *MT_DNN* for several sets of descriptors, namely *Fragmenter*, *CDK*, *Dragon*, and *ALogPS, OEstate*. The performances of *ST_DNN* models were comparable with *XGBoost* and *random forest* models. This result is not surprising, and it confirmed observations of previous studies[48,59] that the efficiency of these methods is similar. The *random forest* method achieved a better average performance compared to the *XGBoost* method. This can be related to the robustness of this method in comparison to that of *XGBoost*. One should carefully select the *XGBoost* parameters to achieve close to optimal solution, while *random forest* usually provides high quality results for the models "out-of-the-box". We also experimented with other ANN types. Associative neural networks (ASNN)[60] required long computational time, because they used CPU and not GPU computing. This algorithm, which was based on a so-called "shallow neural network" which used just one hidden layer, provided a lower accuracy presumably due to the absence of latent representation of the molecules (in deep neural networks latent representation is commonly regarded as the outputs of second-to-last layer).

End Point Modeling. We also compared the quality of models for each individual end point. To do that, a consensus model which averages of the outcomes of the top five descriptor models were created. There were 29 end points which represent four animal species (mouse, rat, rabbit, guinea pig), one unspecified class, and two classes of humans (man and woman), several types of administration, and three outcomes (lethal dose fifty (**LD50**), toxic dose low (**TDLo**), lethal dose low (**LDLo**)). The numbers of records for each end point are given in Table 1. Our automatic data extraction procedure keeps the extracted end points unchanged, that is the reason why the human toxicity is reported separately for man and woman and an "unspecified" animal class is also present. There is the significant gap in the quality of prediction of toxicity for different end points. **LD50** values were predicted with relatively good quality for several species and several types of admission: for **mouse intravenous**, **oral**, **subcutaneous**, and **LD50** type of toxicity had the value of $R^2 \geq 0.65$ for corresponding models. The same model quality is observed for **rat** and **rabbit intravenous LD50** toxicity. It should be noted that **LDLo** was predicted with lower accuracy than **LD50** toxicity for all species and admission types. For **TDLo** the prediction accuracy is inferior: R^2 values for those targets are in the range 0.26−0.43 which is fairly low. The low accuracy of

DOI: 10.1021/acs.jcim.8b00685

the prediction of these end points can be explained by the limited amount of data for these types of toxicity. Moreover, **TDLo** and **LDLo** measurements are less reliable due to disproportionately inaccurate experimental conditions (e.g., could be contributed by other sources of toxicity not directly related to the analyzed compounds): the instrumental errors during measurements were higher for these end points, since both of these toxicities have lower values compared to **LD50**. The target with the lowest error is **rat**, **intravenous**, **LD50** with R^2 = 0.71 and *RMSE* = 0.54. Toxicity for humans is represented only by **TDLo** values, and the quality of prediction of models for this target is unsatisfactory. This is related to the factors mentioned above, and it should inspire new developments because of the extreme importance of such modeling to drug development. On Figure 7 we demonstrate some representative prediction charts, and a full set of prediction charts (for each end point) can be found in the Supporting Information.

Attributed Modeling. Multitask modeling can be approximated as single-task where the end point tags are provided to the input of the model as attributes. For example, species of animal, a type of administration and type of toxicity can be encoded by one-hot encoding and concatenated with a vector of chemical descriptors. The scheme of the attributed modeling is given in Figure 8. The advantage of the attributed modeling is the possibility to use any machine-learning algorithm without additional modifications of a loss function. We compared the performance of consensus XGBoost attributed model with consensus multitask DNN model and consensus attributed DNN model. XGBoost method has been chosen due to both quickness and its ability to achieve the good quality among single-target models. Our experiments revealed that there is no significant discrepancy between the performance of the multitask DNN and the attributed XGBoost model. The statistical performance of different modeling schemes is given in Table 3.

Table 3. Comparison of Quality of Two Consensus Attributed Models with Consensus Multitask Model (Averaged over All End Points)

model	MAE	R^2	RMSE
DNN (attributed)	0.49	0.54	0.69
XGBoost (attributed)	0.49	0.55	0.68
DNN (multitask)	0.49	0.55	0.68

Feature Net Approach. The Feature Net approach has been proposed as a variant of multitask learning by some of us. The main idea of this approach is a usage of predictions of one (or a group) model as additional descriptors for the resulting ST models. It was shown that the Feature Net approach can achieve better accuracy than single-task learning[61] and can provide models with similar accuracy to MT models. We used results of *ST_DNN* as the feature nets to train the models and after that we used these predictions as additional descriptors to develop final models. The statistical performance of these models are given in Table 4.

We observed that for all descriptors the general trend remains the same. The accuracy of Feature Net models is between that of single-task models and multitask models. We believe that Feature Net models partially regard latent correlations in the data; however, the multitask models have significantly better performance. Taking into account that fact that the Feature Net approach requires significantly more time compared to MT models, the feasibility of usage of this approach is questionable.

Table 4. Comparison of RMSE for Models Based on Feature Net Approach with Multi- and Single-Task Models (Averaged over All End Points)

descriptors	feature net	ST DNN	MT DNN
Dragon 6	0.77	0.85	**0.74**
ALogPS, OEstate	0.75	0.86	**0.74**
Fragmentor	0.77	0.88	**0.74**
PyDescriptor	0.76	0.85	**0.74**

Processing of Intervals. Toxicity data sets frequently include a significant number of records reported as intervals e.g., ">" (greater than), in cases where the exact value of toxicity has not been measured. This frequently happens for non- or low toxic compounds or for compounds for which larger concentrations can not be achieved due to solubility or availability. The existence of this large number of the records without exact toxicity values is a special problem in automatic data analysis. The most common approach in this case is to set the maximal toxicity dose observed in the whole data set for these types of records, considering them to be nontoxic. But in the case of particularly heterogeneous data, this discussed approach is not optimal due to the large variations in the toxicity values for different end points. We propose a modification of a loss function which allows the correct processing of such records; the formula for a RMSE loss function over a batch which regards intervals is given below:

$$L(y, \hat{y}) = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - \max(\hat{y}_i, y_i))^2 & \text{if } > \\ \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - \min(\hat{y}_i, y_i))^2 & \text{if } < \end{cases}$$

where $\hat{y}_i$ is a predicted value, y_i is a real value, and n is the total number of samples in a batch.

To estimate the efficiency of the training with our modified loss function we trained two models: one with modified loss function and the other with the standard RMSE loss. Then a comparison of those models applied only to compounds with exact values of toxicity was performed. The motivation for this kind of experiment was to find out if the training on ranged data can improve the quality of models or not. However, no significant difference between models trained with modified loss functions and with RMSE loss were shown. This showed that, despite the simplicity of the idea to modify the loss function, this method is not efficient for the data set under study and the use of standard loss function i.e. RMSE or MAE during training is preferable. Nonetheless, the problem of correct and efficient processing of ranged data, especially for large diverse data sets, is still open and we hope that our research will stimulate interest to this problem.

Latent Representation of Compounds. Neural networks generate a hidden representation of data on their hidden layers by processing the data. We visualized this process directly by performing projection of the latent representations of the compounds onto the 2D plane by the t-SNE method, using the same approach for mapping of toxicity data to compounds from "the description of the dataset chemical space". The neuron's activation on the last-to-last ANN layer for the molecule was used as their hidden representations. The visualization of this latent space on Figure 9 shows that ANN on the last hidden layer achieves good separation of toxic and nontoxic compounds but generally does not group structurally similar compounds together. One can notice three areas containing the most toxic compounds and each of these groups are composed of different compounds: organophosphorus compounds, sterane derivatives, etc. It should be noted that the least toxic compounds are grouped in one cluster: iodine-containing contrast agents, perfluorinated alkanes, and compounds with a β-lactam ring.

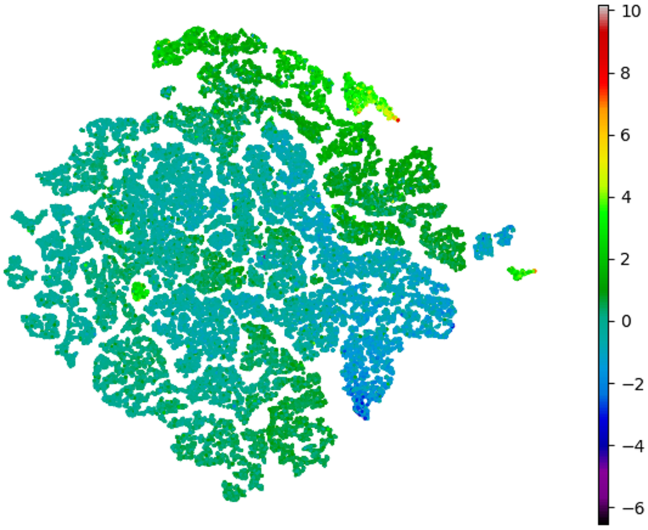


Figure 9. Results of the application of the t-SNE method to deep features generated by the multitarget DNN, values are minus logarithms of maximal end point (greater values correspond to larger toxicity). Several clusters with high toxicity, which presumably reflect different mechanisms of actions (MOAs), are observed.

Regulations in the Light of Multitask Learning. Recent progress in QSAR/QSPR modeling raises questions about the correspondence of newer methods to guidelines established and approved by authorities. In this section we would like to put forward the discussion about the Organization for Economic Co-operation and Development (OECD) principles for the validation, for regulatory purposes of QSAR models. The "Guidance Document on the Validation of (Quantitative) Structure−Activity Relationship [(Q)SAR] Model" summarized the collective opinion of OECD specialists to QSAR modeling.[63] In this document peculiar attention is given to principle N 1−a defined end point. Despite the uncertainty of formalizing a defined end point, the authors of the guideline warned researchers from usage of end points which are not clearly defined. We agree with the authors that for a QSAR model, the end point should be clearly defined, but we believe that the current description of the defined end point is insufficient. For example, item 68 states that "4. The chemical end point of the (Q)SAR should be contained within the chemical end point of the test protocol. 5. The end point being predicted by a (Q)SAR should be the same as the end point measured by a defined test protocol that is relevant for the purposes of the chemical assessment." The interpretation of this formulation may prohibit the usage of multitask learning. At the same time, we are at the beginning of a "big data" time[62] in chemistry and biology. The appearance of these data promotes development of powerful multitask models that could significantly increase quality of models for individual end points. But these methods can break the paradigm "one accurate dataset" → "one model for a narrow end point". It should be mentioned that the *Feature Net* approach, in principle, can still allow us to use the OECD principles by treating predictions of STL models as additional descriptors. However, as we have shown in our studies this approach may not allow us to use the full advantages of multitask modeling.

■ CONCLUSIONS

In this work the efficiency of several methods of machine learning and several types of descriptors was estimated on a large multitask data set. The statistical analysis of the data extracted from the largest toxicity data set the Registry of Toxic Effects of Chemical Substances (RTECS) was performed. We demonstrated that multitask deep neural networks can significantly improve prediction of toxicity by comparing them to single-output types of models investigated including: single-task deep neural network, *XGBoost*, *random forest*, *K nearest neighbors*. The models with highest prediction abilities were those obtained for *rabbit* and *rat* species.

Interestingly, the attributed models (target end points are encoded with additional descriptors) and multitask models (each end point corresponded to one output) demonstrated similar accuracy. While the *Feature Net* approach contributed better models than single-task models, it performed worse than the multitask models. Our results demonstrated that a multitask approach can be beneficial for toxicity prediction due to its ability to process a heterogeneous data set containing different end points. In conclusion, we would like to raise a discussion about applications of multitask learning methods for regulatory purposes and, possibly, to provide a correct interpretation of the Organization for Economic Co-operation and Development (OECD) guidelines which will allow the use of models developed with such methodologies for legal purposes.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00685.

Information about the neural network architecture used in this study (Table S1). Information about the optimizer parameters for ANN training. Prediction charts for all end points (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: m.fedorov@skoltech.ru (M.V.F.).
*E-mail: sergey.sosnin@skoltech.ru (S.S.).
*E-mail: itetko@vcclab.org (I.V.T.).

ORCID

Sergey Sosnin: 0000-0002-3042-7369
Dmitry Karlov: 0000-0002-7194-1081
Igor V. Tetko: 0000-0002-6855-0012

Notes

The authors declare the following competing financial interest(s): IVT is CEO of BIGCHEM GmbH, which licenses the OCHEM. The other authors declared that they have no actual or potential conflicts of interests.

ACKNOWLEDGMENTS

This work has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC "Kurchatov Institute", http://ckp.nrcki.ru. The Quadro P6000 and Titan Xp used for this research were donated by the NVIDIA Corporation. Parts of the figures were prepared using Freepik from www.flaticon.com. The authors are thankful to Allison Keys from Stanford University for fruitful discussions. This study was partially supported by Russian Science Foundation, grant No 14-43-00024.

REFERENCES

(1) Katzung, B. G.; Trevor, A. J. *Basic and Clinical Pharmacology 13 E*, 13th ed.; McGraw-Hill Education/Medical: New York, 2014.

(2) Wong, C. H.; Siah, K. W.; Lo, A. W. Estimation of Clinical Trial Success Rates and Related Parameters. *Biostatistics* **2018**, DOI: 10.1093/biostatistics/kxx069.

(3) Festing, M. F. W. The Extended Statistical Analysis of Toxicity Tests Using Standardised Effect Sizes (SESs): A Comparison of Nine Published Papers. *PLoS One* **2014**, *9*, e112955.

(4) Martel, B. *Chemical Risk Analysis: A Practical Handbook*; Butterworth-Heinemann: Oxford, UK, 2004.

(5) Alden, C. L.; Lynn, A.; Bourdeau, A.; Morton, D.; Sistare, F. D.; Kadambi, V. J.; Silverman, L. A Critical Review of the Effectiveness of Rodent Pharmaceutical Carcinogenesis Testing in Predicting for Human Risk. *Vet. Pathol.* **2011**, *48*, 772−784.

(6) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative High-Throughput Screening: A Titration-Based Approach That Efficiently Identifies Biological Activities in Large Chemical Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 11473−11478.

(7) Huang, J. X.; Blaskovich, M. A.; Cooper, M. A. Cell- and Biomarker-Based Assays for Predicting Nephrotoxicity. *Expert Opin. Drug Metab. Toxicol.* **2014**, *10*, 1621−1635.

(8) 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 - Alternatives to animal experimentation* **2018**, *35*, 163−168.

(9) Institute of Medicine (US) Committee on Internet Access to the National Library of Medicine's Toxicology and Environmental Health Databases. In *Internet Access to the National Library of Medicine's Toxicology and Environmental Health Databases*; Liverman, C. T., Fulco, C. E., Kipen, H. M., Eds.; The National Academies Collection: Reports funded by National Institutes of Health; National Academies Press (US): Washington (DC), 1998.

(10) Richard, A. M.; Williams, C. R. Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network: A Proposal. *Mutat. Res., Fundam. Mol. Mech. Mutagen.* **2002**, *499*, 27−52.

(11) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.

(12) *Registry of Toxic Effects of Chemical Substances*. http://www.3dsbiovia.com/products/collaborative-science/databases/bioactivity-databases/rtecs.html (accessed November 23, 2018).

(13) Pogodin, P. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. PASS Targets: Ligand-Based Multi-Target Computational System Based on a Public Data and Naïve Bayes Approach. *SAR QSAR Environ. Res.* **2015**, *26*, 783−793.

(14) Ji, C.; Svensson, F.; Zoufir, A.; Bender, A. eMolTox: Prediction of Molecular Toxicity with Confidence. *Bioinformatics* **2018**, *34*, 2508−2509.

(15) Drwal, M. N.; Banerjee, P.; Dunkel, M.; Wettig, M. R.; Preissner, R. ProTox: A Web Server for the in Silico Prediction of Rodent Oral Toxicity. *Nucleic Acids Res.* **2014**, *42*, W53−W58.

(16) Asadollahi-Baboli, M. Exploring QSTR Analysis of the Toxicity of Phenols and Thiophenols Using Machine Learning Methods. *Environ. Toxicol. Pharmacol.* **2012**, *34*, 826−831.

(17) Auerbach, S. S.; Shah, R. R.; Mav, D.; Smith, C. S.; Walker, N. J.; Vallant, M. K.; Boorman, G. A.; Irwin, R. D. Predicting the Hepatocarcinogenic Potential of Alkenylbenzene Flavoring Agents Using Toxicogenomics and Machine Learning. *Toxicol. Appl. Pharmacol.* **2010**, *243*, 300−314.

(18) Liu, R.; Zhang, H. Y.; Ji, Z. X.; Rallo, R.; Xia, T.; Chang, C. H.; Nel, A.; Cohen, Y. Development of Structure−Activity Relationship for Metal Oxide Nanoparticles. *Nanoscale* **2013**, *5*, 5644−5653.

(19) Wang, Y.; Zheng, M.; Xiao, J.; Lu, Y.; Wang, F.; Lu, J.; Luo, X.; Zhu, W.; Jiang, H.; Chen, K. Using Support Vector Regression Coupled with the Genetic Algorithm for Predicting Acute Toxicity to the Fathead Minnow. *SAR QSAR Environ. Res.* **2010**, *21*, 559−570.

(20) Li, X.; Zhang, Y.; Chen, H.; Li, H.; Zhao, Y. Insights into the Molecular Basis of the Acute Contact Toxicity of Diverse Organic Chemicals in the Honey Bee. *J. Chem. Inf. Model.* **2017**, *57*, 2948−2957.

(21) von Korff, M.; Sander, T. Toxicity-Indicating Structural Patterns. *J. Chem. Inf. Model.* **2006**, *46*, 536−544.

(22) Unterthiner, T.; Mayr, A.; Klambauer, G.; Hochreiter, S. Toxicity Prediction Using Deep Learning. *arXiv.org* **2015**, 1503.01445.

(23) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-Task Neural Networks for QSAR Predictions. *arXiv.org* **2014**, 1406.1231.

(24) Sosnin, S.; Vashurina, M.; Withnall, M.; Karpov, P.; Fedorov, M.; Tetko, I. V. A Survey of Multi-Task Learning Methods in Chemoinformatics. *Mol. Inf.* **2018**, DOI: 10.1002/minf.201800108.

(25) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.

(26) Feng, J.; Lurati, L.; Ouyang, H.; Robinson, T.; Wang, Y.; Yuan, S.; Young, S. S. Predictive Toxicology: Benchmarking Molecular Descriptors and Statistical Methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1463−1470.

(27) Baskin, I. I. *Computational Toxicology; Methods in Molecular Biology*; Humana Press, New York, NY, 2018; pp 119−139.

(28) Masand, V. H.; Rastija, V. PyDescriptor: A New PyMOL Plugin for Calculating Thousands of Easily Understandable Molecular Descriptors. *Chemom. Intell. Lab. Syst.* **2017**, *169*, 12−18.

(29) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics; Methods and principles in medicinal chemistry*; Wiley-VCH: Weinheim, 2009.

(30) Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. Hierarchical QSAR Technology Based on the Simplex Representation of Molecular Structure. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 403−421.

(31) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2310−2316.

(32) Thormann, M.; Vidal, D.; Almstetter, M.; Pons, M. Nomen Est Omen: Quantitative Prediction of Molecular Properties Directly from IUPAC Names. *Open Appl. Inf. J.* **2007**, *1*, 28−32.

(33) Thijs, G.; Langenaeker, W.; De Winter, H. Application of Spectrophores to Map Vendor Chemical Space Using Self-Organising Maps. *J. Cheminf.* **2011**, *3*, P7.

(34) Cherkasov, A. Inductive QSAR Descriptors. Distinguishing Compounds with Antibacterial Activity by Artificial Neural Networks. *Int. J. Mol. Sci.* **2005**, *6*, 63−86.

(35) Potemkin, V.; Grishina, M. Principles for 3D/4D QSAR Classification of Drugs. *Drug Discovery Today* **2008**, *13*, 952−959.

(36) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and

Pharmacophoric Descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191−198.

(37) Tetko, I. V.; Tanchuk, V. Y. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136−1145.

(38) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(39) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(40) RDKit: Open-Source Cheminformatics. www.rdkit.org (accessed November 23, 2018).

(41) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.

(42) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Model.* **1995**, *35*, 1039−1045.

(43) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.

(44) Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, Lille, France, 2015; Vol. *37*, pp 448−456.

(45) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.

(46) Tokui, S.; Oono, K.; Hido, S.; Clayton, J. Chainer: A Next-Generation Open Source Framework for Deep Learning. *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-Ninth Annual Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada, Dec. 7−12, 2015.

(47) Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, II; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q. Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 533−554.

(48) Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme Gradient Boosting as a Method for Quantitative Structure−Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56*, 2353−2360.

(49) Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 468−481.

(50) Kauffman, G. W.; Jurs, P. C. QSAR and K-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553−1560.

(51) Gunturi, S. B.; Archana, K.; Khandelwal, A.; Narayanan, R. Prediction of hERG Potassium Channel Blockade Using kNN-QSAR and Local Lazy Regression Methods. *QSAR Comb. Sci.* **2008**, *27*, 1305−1317.

(52) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.

(53) Cao, D.-S.; Yang, Y.-N.; Zhao, J.-C.; Yan, J.; Liu, S.; Hu, Q.-N.; Xu, Q.-S.; Liang, Y.-Z. Computer-Aided Prediction of Toxicity with Substructure Pattern and Random Forest. *J. Chemom.* **2012**, *26*, 7−15.

(54) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(55) Novotarskyi, S.; Abdelaziz, A.; Sushko, Y.; Körner, R.; Vogt, J.; Tetko, I. V. ToxCast EPA in Vitro to in Vivo Challenge: Insight into the Rank-I Model. *Chem. Res. Toxicol.* **2016**, *29*, 768−775.

(56) Golbraikh, A.; Tropsha, A. Beware of Q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(57) van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(58) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57*, 2490−2504.

(59) Zhang, L.; Ai, H.; Chen, W.; Yin, Z.; Hu, H.; Zhu, J.; Zhao, J.; Zhao, Q.; Liu, H. CarcinoPred-EL: Novel Models for Predicting the Carcinogenicity of Chemicals Using Molecular Fingerprints and Ensemble Learning Methods. *Sci. Rep.* **2017**, *7*, 2118.

(60) Tetko, I. V. Neural Network Studies. 4. Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717−728.

(61) Varnek, A.; Gaudin, C.; Marcou, G.; Baskin, I.; Pandey, A. K.; Tetko, I. V. Inductive Transfer of Knowledge: Application of Multi-Task Learning and Feature Net Approaches to Model Tissue-Air Partition Coefficients. *J. Chem. Inf. Model.* **2009**, *49*, 133−144.

(62) Tetko, I. V.; Engkvist, O.; Koch, U.; Reymond, J.-L.; Chen, H. BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. *Mol. Inf.* **2016**, *35*, 615−621.

(63) IOMC. *Guidance Document on the Validation of (Quantitative) Structure−Activity Relationship [(Q)SAR] Model*; Environment Directorate, Organization for Economic Co-operation and Development: Paris, 2007.

DOI: 10.1021/acs.jcim.8b00685