PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2020, **22**, 8373

Are 2D fingerprints still valuable for drug discovery?†

Kaifu Gao,[a] Duc Duy Nguyen,[a] Vishnu Sresht,[b] Alan M. Mathiowetz,[b] Meihua Tu[b] and Guo-Wei Wei *[acd]

Received 17th January 2020,
Accepted 18th March 2020

DOI: 10.1039/d0cp00305k

rsc.li/pccp

Recently, molecular fingerprints extracted from three-dimensional (3D) structures using advanced mathematics, such as algebraic topology, differential geometry, and graph theory have been paired with efficient machine learning, especially deep learning algorithms to outperform other methods in drug discovery applications and competitions. This raises the question of whether classical 2D fingerprints are still valuable in computer-aided drug discovery. This work considers 23 datasets associated with four typical problems, namely protein–ligand binding, toxicity, solubility and partition coefficient to assess the performance of eight 2D fingerprints. Advanced machine learning algorithms including random forest, gradient boosted decision tree, single-task deep neural network and multitask deep neural network are employed to construct efficient 2D-fingerprint based models. Additionally, appropriate consensus models are built to further enhance the performance of 2D-fingerprint-based methods. It is demonstrated that 2D-fingerprint-based models perform as well as the state-of-the-art 3D structure-based models for the predictions of toxicity, solubility, partition coefficient and protein–ligand binding affinity based on only ligand information. However, 3D structure-based models outperform 2D fingerprint-based methods in complex-based protein–ligand binding affinity predictions.

I. Introduction

Drug discovery is a multi-parameter optimization process, which involves a long list of chemical, biological, and physiological properties.[1] For a drug candidate, numerous drug-related properties must be assessed, including binding affinity, toxicity, octanol–water partition coefficient ($\log P$), aqueous solubility ($\log S$), *etc.* Binding affinity assesses the strength of a drug's binding to its target,[2,3] while, toxicity is a measure of the degree to which a chemical compound can damage an organism adversely.[4] In addition, a partition coefficient is defined as the ratio of concentrations of a solute in a mixture of two immiscible solvents at equilibrium and, in the case of $\log P$, represents the drug-relatedness of a compound as well as its hydrophobic effect on human bodies.[5] Another relevant drug attribute is aqueous solubility which plays a vital role in distribution, absorption, and biological activity, among other processes because 65–90% of body mass is water.[6,7] Their importance to drug design and discovery has been emphasized by many recent surveys.[8,9] Indeed, unsatisfactory toxicity or pharmacokinetic properties are responsible for approximately half of drug candidate failures to reach the market.[10]

Traditional experiments for measuring drug properties are conducted either *in vivo* or *in vitro*. Such experiments are quite time consuming and expensive. Additionally, testing with animals can raise important ethical concerns. Therefore, various computer-aided or *in silico* methods become more attractive since they can produce quick results without sacrificing much accuracy in many situations. Among them, one of the most popular approaches is the quantitative structure–activity/property relationship (QSAR/QSPR) analysis. It assumes that similar molecules have similar bioactivities or physicochemical properties.[11] Based on this assumption, activities and properties of new molecules can be predicted by studying the correlation between chemical or structural features of molecules and their activities or properties, reducing the need for time-consuming experiments.

Molecular fingerprints are one way of encoding the structural features of a molecule. They play a fundamental role in QSAR/QSPR analysis, virtual screening, similarity-based compound search, target molecule ranking, drug ADMET prediction, and other drug discovery processes. Molecular fingerprints are property profiles of a molecule, usually in the form of vectors with each vector element indicating the existence, the degree or the frequency of one particular structure feature.[12–14] Various fingerprints have been

[a] *Department of Mathematics, Michigan State University, MI 48824, USA. E-mail: wei@math.msu.edu*
[b] *Pfizer Medicine Design, 610 Main St, Cambridge, MA 02139, USA*
[c] *Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA*
[d] *Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d0cp00305k

developed for molecular feature encoding in the past few decades.[15–17] Most fingerprints are 2D fingerprints which can be extracted from molecular connection tables without 3D structure information. However, high dimensional fingerprints have also been developed to utilize 3D molecular structure and other information.[18]

There are four main categories of 2D fingerprints, namely substructure key-based fingerprints, topological or path-based fingerprints, circular fingerprints, and pharmacophore fingerprints. Substructure key-based fingerprints are bit strings representing the presence of certain substructures or fragments from a given list of structural keys in the compound. Molecular access system (MACCS)[19] is one of the most popular substructure key-based fingerprint methods. Topological or path-based fingerprints are based on analyzing all the fragments of a molecule following a (usually linear) path up to a certain number of bonds, and then hashing every one of these paths to create one fingerprint. The most prominent ones in this category are FP2,[20] Daylight[21] and electro-topological state (Estate)[22] fingerprints. Circular fingerprints are also hashed topological fingerprints but rather than looking for paths in a molecule, they record the environment of each atom up to a pre-determined radius. A well-known example for this class is extended-connectivity fingerprint (ECFP).[15] Pharmacophore fingerprints include the relevant features and interactions needed for a molecule to be active against a given target, including 2D-pharmacophore,[23] 3D-pharmacophore[24] and extended reduced graph (ERG)[25] fingerprints as examples. Since 2D fingerprints only rely on the 2D structures, their generation is easy, fast and convenient.

In addition to the four categories mentioned above, recent improvements in deep learning have enabled the creation of neural fingerprints[26,27] – where the mapping between fingerprints and 2D structures is learned simultaneously with the parameters of the regression/classification model that maps fingerprints to targets. These 'learned' fingerprints can potentially improve predictive performance on QSAR/QSPR tasks, but they must be relearned when trying to predict new properties across significantly different regions of chemical space. Since the focus of this work is on comparing 2D and 3D descriptors across a number of disparate tasks and chemically diverse datasets, we have chosen not to consider neural fingerprints.

Most commonly used 2D molecular fingerprints were derived over a decade ago and their validation was carried out using classical regression or classification algorithms, such as linear regression, logistic regression, logistic classification, naive Bayes, k-nearest neighbors, support vector machine, *etc.* On the other hand, new 3D structure-based fingerprints built from algebraic topology,[28,29] differential geometry,[30] geometric graph theory,[31,32] and algebraic graph theory[33] have been developed in recent years. In particular, these new fingerprints were mostly paired with advanced machine learning algorithms, such as random forest (RF),[34] gradient boosting decision tree (GBDT),[35] single-task deep neural networks (ST-DNNs),[36] multi-task deep neural networks (MT-DNNs),[37] convolutional neural network (CNN), recurrent neural network (RNN), *etc.*, which are now easily accessible to the scientific community *via* user-friendly deep learning frameworks in popular programming languages.[38,39] Often, these new methods have demonstrated higher accuracy or better performance than earlier methods in the literature, which are typically based on 2D fingerprints and/or simple machine learning algorithms for drug discovery related applications, such as protein–ligand binding,[28] virtual screening,[29] toxicity,[4] solubility,[5] partition coefficient,[5] as well as protein folding stability change upon mutation.[40] Additionally, recent results from D3R Grand Challenges, a community-wide annual competition series in computer-aided drug design, indicate that structure-based methods using sophisticated 3D structure-based fingerprints have an advantage over ligand-based methods using 2D fingerprints in scoring and free energy predictions.[33,41] These developments raise an interesting question of whether 2D fingerprints are still valuable for drug design and discovery. Therefore, there is pressing need to reassess 2D fingerprints with advanced machine learning algorithms and compare their performance with the state-of-the-art 3D structure-based fingerprints for drug discovery related applications.

The objective of the present work is to reassess the predictive power of eight popular 2D fingerprints for four important drug-related problems, namely, toxicity, binding affinity, log *P*, and log *S*, involving a total of 23 datasets. These problems are selected for the availability of reference results generated by the state-of-the-art 3D structure-based fingerprints in the literature. To optimize 2D fingerprints' performance, advanced machine learning algorithms, including RF, GBDT, ST-DNN, and MT-DNN, are employed in the present study. Additionally, consensus models are constructed from appropriate combinations of 2D fingerprint-based predictions to further enhance their performance. The predictive power of each 2D fingerprint for certain functional groups is analyzed. Extensive numerical studies over 23 datasets using eight 2D fingerprints and four different machine learning algorithms indicate that the combination of appropriate machine learning algorithms and 2D fingerprint-based models, particularly consensus models, can bring significant improvements over previous 2D QSPR approaches especially on toxicity predictions.[42] Moreover, 2D fingerprint-based models perform as well as the state-of-the-art 3D structure-based fingerprints in the predictions of toxicity, log *S*, log *P* and ligand-based protein–ligand binding affinity. Finally, topology-based fingerprints extracted from 3D protein–ligand complexes have a significant advantage over 2D fingerprints in complex-based protein–ligand binding affinity predictions. This is because 2D models can only take care of relatively simple geometry, so do not work well for macromolecules that have complex 3D structures.[43] We believe that the present performance analysis and assessment will provide a useful guideline on how to choose appropriate fingerprints and machine learning methods for drug discovery related applications.

II Methods

II.A 2D fingerprints

In the present work, we investigate eight popular 2D fingerprints, including FP2 fingerprint, MACCS fingerprint, Daylight

Table 1 A introduction of eight fingerprints used in the present study

Fingerprint	Description	Number of features	Package
FP2	A path-based fingerprint which indexes small molecule fragments based on linear segments of up to 7 atoms[20]	256	Openbabel[20]
Daylight	A path-based fingerprint consisting 2048 bits and encoding all connectivity pathways in a given length through a molecule[21]	2048	RDKit[44]
MACCS	A substructure keys-based fingerprint with 166 structural keys based on SMARTS patterns[19]	166	
Estate1	A topological fingerprint based on electro-topological State Indices, which encodes the intrinsic electronic state of the atom as perturbed by the electronic influence of all other atoms in the molecule within the context of the topological character of the molecule. Estate1 represents the number of times each atom type is hit[22]	79	
Estate2	Similar to Estate1, however it contains the sum of the Estate indices for atoms of each type[22]	79	
ECFP4	The *de facto* standard circular fingerprint based on the Morgan algorithm,[45] which uses an iterative process to assign numeric identifiers to each atom[15]	2048	
Pharm2D	Each bit corresponds to a particular combination of features and interactions needed for a molecule to be active against a given target[23]	990	
ERG	A Pharmacophore fingerprint, which is an extended reduced graph approach using pharmacophore-type node descriptions to encode the relevant molecular properties[25]	315	

fingerprint, Estate1 fingerprint, Estate2 fingerprint, ECFP4 Fingerprint, 2D-pharmacophore (Pharm2D), and extended reduced graph fingerprint (ERG). They are chosen to represent four main 2D molecular fingerprint categories, namely key-based fingerprints, topological or path-based fingerprints, circular fingerprints, pharmacophore fingerprints. These features are some of the most popular and commonly used ones. Table 1 summarizes the information related to these fingerprints. All 2D fingerprints were generated by Openbabel (version 2.4.1)[20] and RDKit (version 2018.09.3).[44]

II.B Ensemble methods

Two popular ensemble methods were used in our work. The first method is random forest (RF), which constructs a multitude of decision trees during a training process. RF can be used to predict a classification label (classification model) or a mean prediction (regression model) of the individual trees. The second method is gradient boosting decision tree (GBDT). In this approach, individual decision trees are combined in a stage-wise fashion to achieve the capability of learning complex features. It uses both gradient and boosting strategies to reduce model errors. Compared to deep neural network (DNN) approaches, these two ensemble methods are robust against overfitting, relatively insensitive to hyper parameters, and easy to implement, moreover, they are much faster to train than DNN. In fact, for small datasets, RF and GBDT can perform even better than DNN or other deep learning algorithms.[4,5] Therefore, these methods have been applied to a variety of QSAR prediction problems, such as toxicity, solvation, and binding affinity predictions.[4,28,42,46,47]

II.C Single-task deep neural network (ST-DNN)

A DNN mimics the learning process of a biological brain by constructing a wide and deep architecture of numerous connected neuron units. A typical deep neural network often includes multiple hidden layers. In each layer, there are hundreds or even thousands of neurons. During learning stages, weights on each layer are updated by backpropagation. With a complex and deep network, DNN is capable of constructing hierarchical features and model complex nonlinear relationships.

ST-DNN is a regular deep learning algorithm. It only takes care of one single prediction task, therefore, it only learns from one specific training dataset. A typical four-layer ST-DNN is showed in Fig. 1, where N_i (i = 1, ..., 4), represents the number of neurons in the ith hidden layer.

II.D Multitask deep neural network (MT-DNN)

The multitask (MT) learning technique has achieved much success in qualitative Merck and Tox21 prediction challenges.[48–51] In the MT framework, multiple tasks share the same hidden layers. However, the output layer is attached to different tasks. This framework enables the neural network to learn all the data simultaneously for different tasks. Thus, the commonalities and differences among various datasets can be exploited. It has been showed that MT learning typically can improve the prediction accuracy of relatively small datasets if it combines with relatively larger datasets in its training.

Fig. 2 is an illustration of a typical four-layer MT-DNN for training four different tasks simultaneously. Suppose there are totally T tasks and the training data for the tth task are $\left(X_i^t, y_i^t\right)_{i=1}^{N_t}$, where t = 1,...,T, i = 1,...,N_t, N_t is the number of samples in the tth task, and X_i^t is the feature vector for the ith sample in the tth task, y_i^t is the label value of the ith sample in the tth task, respectively. The purpose of MT learning is to simultaneously minimize the loss function:

$$\operatorname{argmin} \sum_{t=1}^{T} \sum_{i=1}^{N_t} L(y_i^t, f^t(X_i^t, \theta^t))$$

where f^t is the prediction for the ith sample in the tth task by our MT-DNN, which is a function of the feature vector X_i^t, L is the loss function, and θ^t is the collection of machine learning hyperparameters. A popular cost function for regression is the mean squared error, which can be defined as:

$$L\left(y_i^t, f^t(X_i^t, \theta^t)\right) = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(y_i^t - f^t(X_i^t, \theta^t)\right)^2.$$

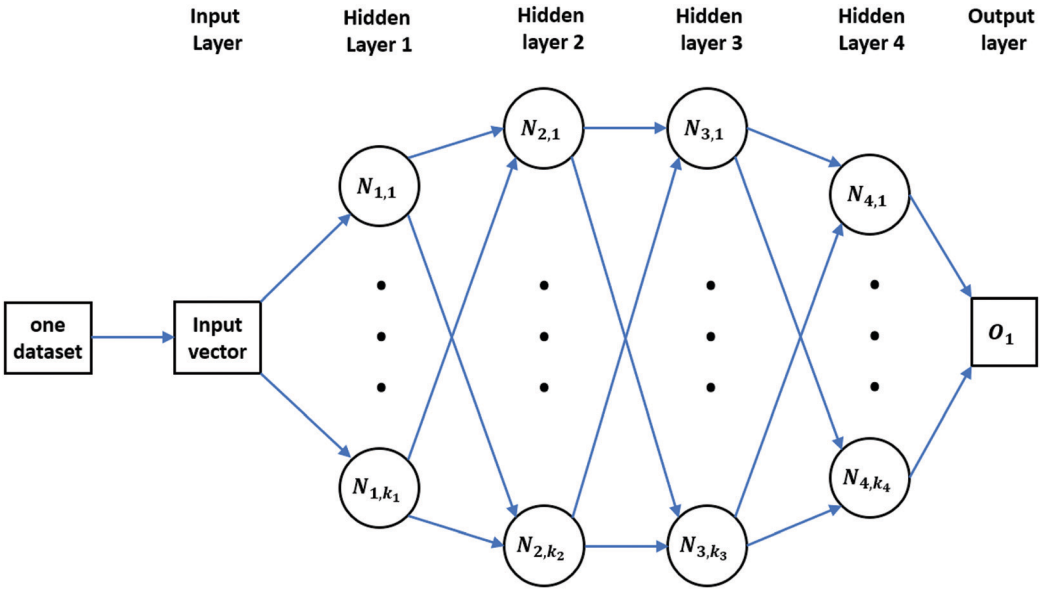


Fig. 1 An illustration of a typical ST-DNN. Only one task (data set) is trained in this network. Four hidden layers are included, k_i (i = 1, 2, 3, 4) represents the number of neurons in the ith hidden layer and $N_{i,j}$ is the jth neuron in the ith hidden layer. Here, O_1 is the single output for the task.

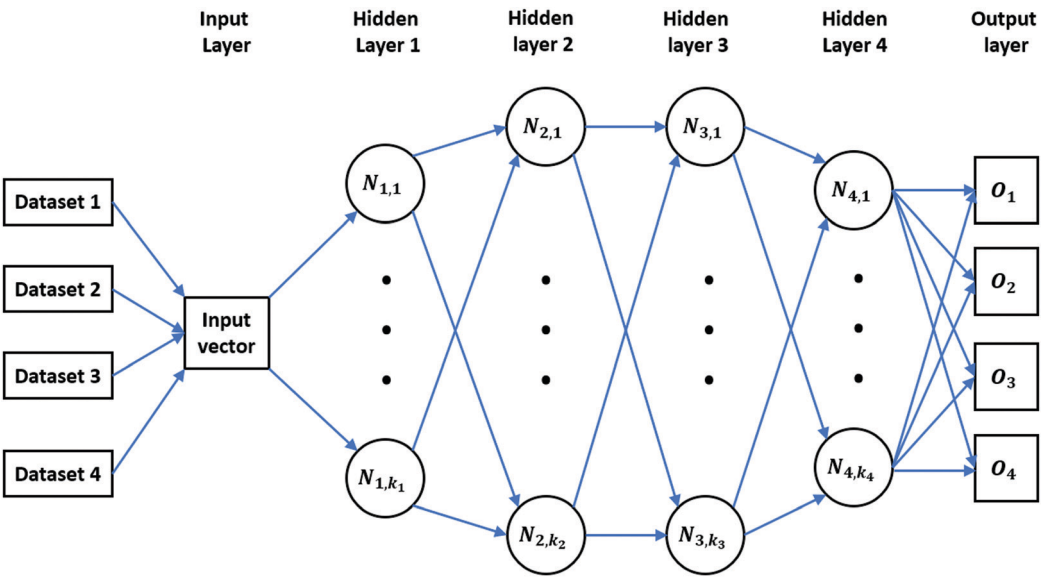


Fig. 2 An illustration of a typical MT-DNN training four tasks (datasets) simultaneously. Four hidden layers are included in this network, k_i (i = 1, 2, 3, 4) represents the number of neurons in the ith hidden layer and $N_{i,j}$ is the jth neuron in the ith hidden layer. Here O_1 to O_4 represent four predictor outputs for four tasks.

In this study, MT learning technology is applied to the toxicity prediction. The ultimate goal of this MT learning is to potentially improve the overall performance of multiple toxicity prediction models, especially for the smallest dataset that performs relatively poorly in the ST-DNN. More concretely, it is reasonable to assume that different toxicity indexes share a common pattern so that these different tasks can be trained simultaneously when their feature vectors are constructed in the same manner. For our toxicity prediction, four different tasks (LD_{50}, IGC_{50}, LC_{50}, LC_{50}-DM data sets) are trained together. This leads to four output neurons in the output layer (see O_1 to O_4 in Fig. 2), with each neuron being specific to one of four tasks.

II.E Consensus of multiple model predictions

Consensus means the average value from multiple model predictions, which typically enhances the results from individual models.

II.F Hyperparameters

Ensemble hyperparameters. Both RF and GBDT were implemented with the scikit-learn package (version 0.20.1).[52] In this

Table 2 RF and GBDT parameters for different training-set sizes

Training-set size	RF parameters	GBDT parameters
<800	n_estimators = 1000, criterion = 'mse', max_depth = none, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0	n_estimators = 2000, max_depth = 9, min_samples_split = 3, learning_rate = 0.01, subsample = 0.1, max_features = 'sqrt'
800 to 5000		n_estimators = 10 000, max_depth = 7, min_samples_split = 3, learning_rate = 0.01, subsample = 0.3, max_features = 'sqrt'
5000 to 10 000		n_estimators = 20 000, max_depth = 7, min_samples_split = 3, learning_rate = 0.01, subsample = 0.3, max_features = 'sqrt'

work, there are a total of 23 datasets with their training data size varying from 94 to 8199. RF has been showed to be consistent and robust with various datasets. However, if its parameters are carefully tuned based on the size of a given training set, GBDT can attain better performance than RF does in most cases. For all experiments in this work, the most essential parameters of GBDT are chosen as learning rate = 0.01, min_samples_split = 3, max_features = sqrt. Detail values of other parameters are given in Table 2.

Network hyperparameters. Since the numbers of features differ much in different 2D fingerprints, different network architectures have to be adopted. For example, Estate1 fingerprint has only 79 bits. Therefore a 4-layer network with the number of neurons in various hidden layers are chosen as 500, 1000, 1500, and 500. However, the Daylight fingerprint has as many as 2048 features, and thus a much larger network is needed. The network for this fingerprint still has 4 layers but there are 3000, 2000, 1000, and 500 neurons in the first, second, third and fourth hidden layers, respectively. Other network parameters are as followed: the optimizer is stochastic gradient descent (SGD) with momentum of 0.5. 2000 epochs were run for all the networks. Mini-batch size is set to 4. The learning rate is set to 0.01 in the first 1000 epochs and 0.001 for the rest epochs. Our tests indicate that adding a dropout or using L_2 decay does not necessarily improve the accuracy, and thus, we omit these two techniques. All the network hyperparameters are summarized in Table 3. These hyperparameters are applied to both ST-DNN and MT-DNN. All the DNN training is performed with Pytorch (version 1.0).[53]

III Results

III.A Toxicity prediction

Four toxicity datasets were studied in our work, namely oral rat LD_{50} (LD_{50}), 40 h *Tetrahymena pyriformis* IGC_{50} (IGC_{50}), 96 h fathead minnow LC_{50} (LC_{50}), and 48 h *Daphnia magna* LC_{50} (LC_{50}-DM). Among them, LD_{50} measures the amount of chemicals that can kill half of rats when orally ingested. IGC_{50} records the 50% growth inhibitory concentration of *Tetrahymena pyriformis* organism after 40 h. LC_{50} reports at the concentration of test chemicals in water in milligrams per liter that cause 50% of fathead minnows to die after 96 h. The last one is LC_{50}-DM, which represents the concentration of test chemicals in water in milligrams per liter that cause 50% *Daphnia magna* to die after 48 h. The unit of toxicity reported in these four datasets is $-\log_{10}$ mol L^{-1}. All of them are accessible from the recent publications[42,54,55] and the public database (https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test). The sizes of these four datasets vary from 353 to 7413 (see Table 4), which raises a challenge for a predictive model to achieve a consistent accuracy and robustness.

III.A.1 The performance of ensemble methods. Because it is easy to implement and fast to train, two ensemble methods, RF and GBDT, were first tested. Since four datasets have very different sizes, different numbers of estimators in RF and GBDT models should be used. Specifically, for two relatively small sets, LC_{50} and LC_{50}-DM, the numbers of estimators are set to 2000. For IGC50, 10 000 estimators are used. For the largest set LD_{50}, we have used 20 000 estimators.

The accuracy is measured in term of the square of Pearson correlation coefficient (R^2). Overall, GBDT's performance is always better than that of RF, which agrees with the early publication.[4] Among all the eight fingerprints we tested, Estate2, Estate1, Daylight, FP2, ECFP and MACCS usually work well on these four sets. Thus the consensus of these six fingerprints or say the average prediction of the six fingerprints, was also considered (''Top 6-cons'' in Fig. 4). The consensus model typically gives rise to a further improvement over all single fingerprints in most cases.

(a) LD_{50} test set. LD_{50} dataset is the largest set having as many as 7413 compounds. However, the set has a higher experimental uncertainty of the values (see ''Max value'' and ''Min value'' in Table 4) and more importantly, as revealed in Fig. 3(a), the ranges of the training set and test set are almost the same. The boundary values of the training set overlap with those of the test set, which brings difficulty to machine learning models. In our GBDT model, the best single fingerprint

Table 3 The network hyperparameters for both ST-DNN and MT-DNN

Fingerprint	Number of features	Number of hidden layers	Number of neurons in each hidden layer	Optimizer	Mini-batch	Learning rate
Estate1	79	4	500, 1000, 1500, 500	SGD with a momentum of 0.5	4	First 1000: 0.01; then: 0.001
Estate2	79					
Daylight	2048		3000, 2000, 1000, 500			

Table 4 The quantitative summary of four toxicity datasets. The original datasets and prediction results are available at https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

Data set	Total size	Train set size	Test set size	Max value	Min value
LD_{50}	7413	5931	1482	7.201	0.291
IGC_{50}	1792	1434	358	6.36	0.334
LC_{50}	823	659	164	9.261	0.037
LC_{50}-DM	353	283	70	10.064	0.117

(MACCS) yields an R^2 of 0.643, while the consensus of the top 6 fingerprints increases R^2 to 0.679.

(b) IGC_{50} test set. IGC_{50} set is the second largest set (1792 compounds) among the four sets we investigated. As indicated in Table 4, this set has the smallest range of label. Moreover, Fig. 3(b) shows that the test set has a smaller range than that of the training set, indicating a relatively easy case for machine learning models. Our results show that Estate2 is the best single fingerprint with an R^2 of 0.742, and the consensus of the top 6 fingerprints leads to an R^2 of 0.785.

(c) LC_{50} test set. LC_{50} set is a relative smaller set (823 compounds). Fig. 3(c) indicates that the ranges of the training set and test set are almost the same. In our GBDT model, Estate2 fingerprint achieves the top performance, which yields an R^2 of 0.662. The consensus of the top 6 fingerprint improves the R^2 to 0.715.

(d) LC_{50}-DM test set. Among the four sets, LC_{50}-DM test set is the smallest one with only 283 training molecules and 70 test molecules, which is troublesome to build a robust model. Moreover, as revealed in Fig. 3(d), not only the boundary values of the training set overlap with those of the test set, but also the test set has a higher distribution at the left boundary, rendering a difficult case for machine learning. Specifically, the best single fingerprint Estate1 only has an R^2 of 0.520. The consensus model even lowers the R^2 a little bit to 0.486. Similar difficulty is also faced by other recent work, such as the R^2 of the 3D-topology based GBDT model only reaches 0.505.[4] Thus, there is a need for multitask deep learning when dealing with such a small dataset.

III.A.2 The performance of single-task and multitask deep learning. On average, Estate2, Estate1, and Daylight are the top three fingerprints when using GBDT models in all the four sets. Thus, these three fingerprints are picked up to perform higher-level ST-DNN and MT-DNN.

Since the lengths of the three fingerprints differ much, different DNN architectures are needed. Four hidden layers with 500, 1000, 1500, and 500 neurons are used for Estate1 and Estate2, whose fingerprints have 79 features. Four hidden

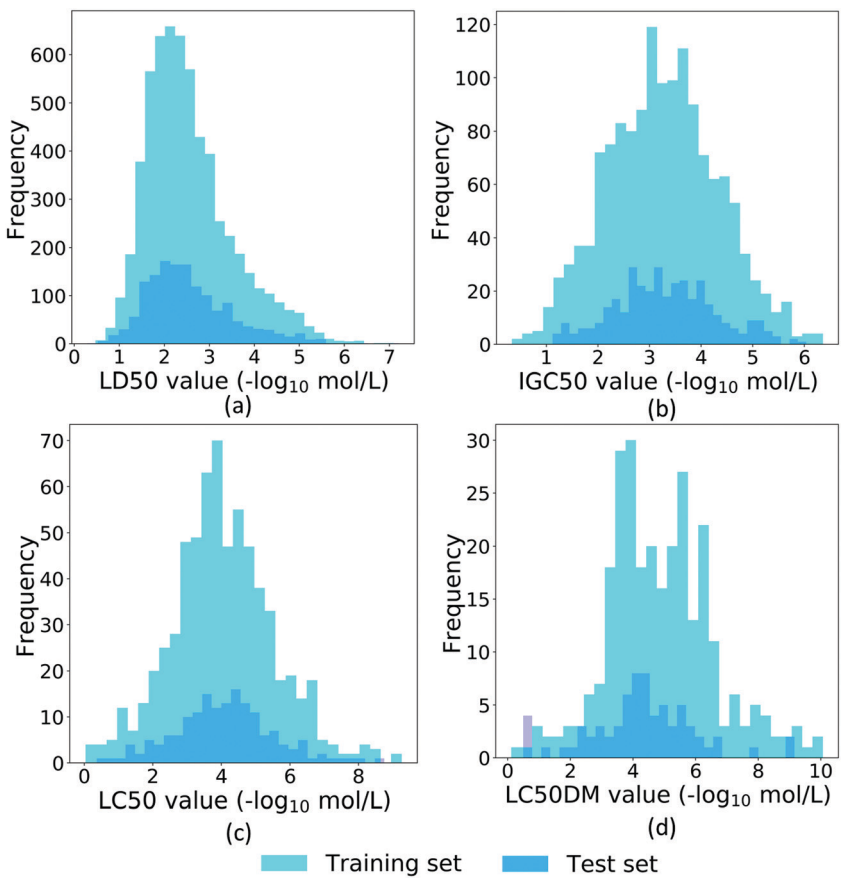


Fig. 3 The sample distributions of LD_{50}, IGC_{50}, LC_{50}, LC_{50}-DM training and test sets.

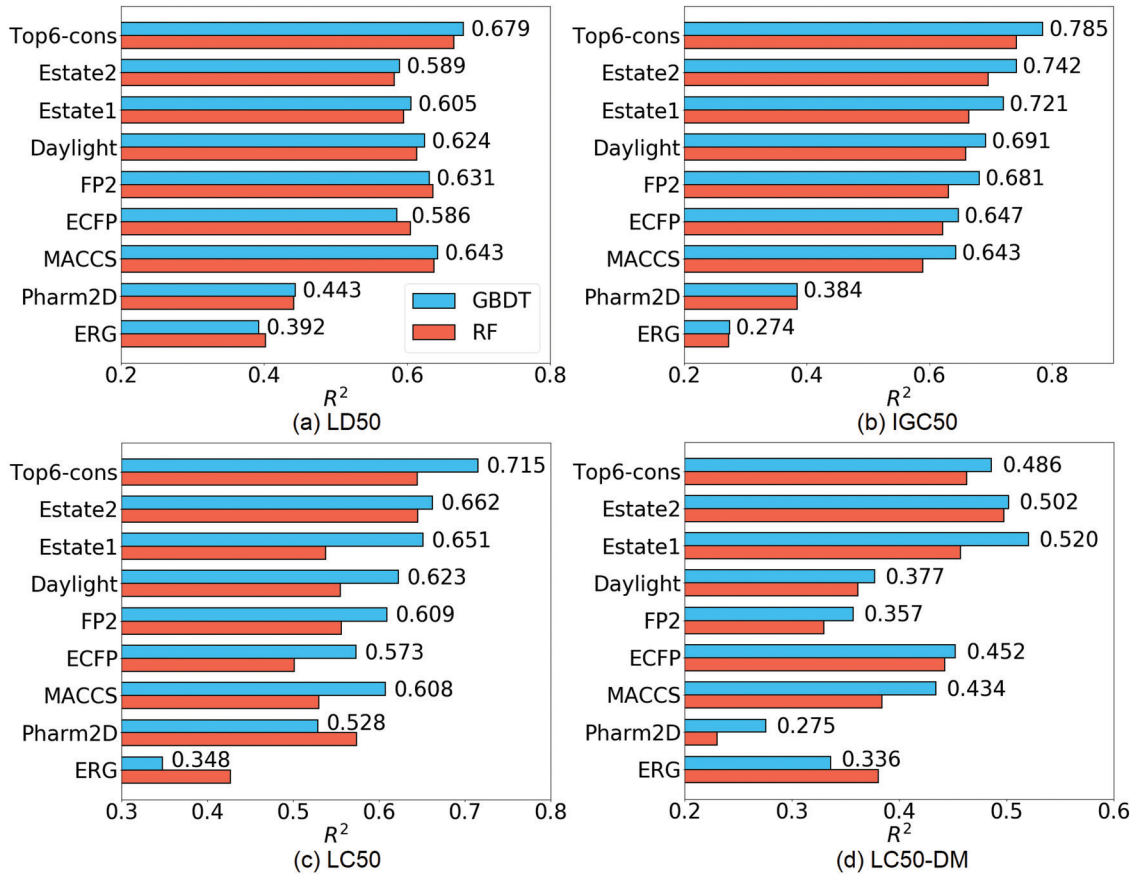


Fig. 4 The R^2 on LD_{50}, IGC_{50}, LC_{50}, LC_{50}-DM test sets yielded by eight fingerprints and the consensuses of the top 6 features. Two ensemble methods were adopted (GBDT: blue, RF: red). The values shown in the figure are the R^2 of GBDT.

layers with 3000, 2000, 1000, and 500 neurons are used for Daylight, whose fingerprint has 2048 bits.

The pattern of ST-DNN results is similar to that of GBDT results. On four data sets, a ST-DNN consensus model yields an average R^2 of 0.658 (0.632, 0.791, 0.687, and 0.523 respectively). As a comparison, the average R^2 by a GBDT consensus model is 0.666 (0.679, 0.785, 0.715, and 0.486 respectively). However, the performance can be largely enhanced by the multitask strategy because the two relatively smaller sets LC_{50} and LC_{50}-DM can benefit much from two larger sets LD_{50} and IGC_{50}. As shown in Table 5, while the MT-DNN model performance seldom changes on LD_{50} and IGC_{50}, it gives rise to a dramatic improvement on LC_{50} and LC_{50}-DM, especially on LC_{50}-DM. The consensus R^2 are lifted from 0.523 to 0.725.

Table 5 The R^2 of ST-DNN and MT-DNN based on the top 3 fingerprints in GBDT (Estate2, Estate1, Daylight) and their consensuses

Method	R^2 of LD_{50}	R^2 of IGC_{50}	R^2 of LC_{50}	R^2 of LC_{50}-DM
Estate2 ST-DNN	0.484	0.715	0.569	0.433
Estate2 MT-DNN	0.489	0.696	0.660	0.623
Estate1 ST-DNN	0.569	0.733	0.650	0.601
Estate1 MT-DNN	0.566	0.735	0.694	0.684
Daylight ST-DNN	0.619	0.701	0.570	0.346
Daylight MT-DNN	0.617	0.717	0.724	0.694
Consensus ST-DNN	0.632	0.791	0.687	0.523
Consensus MT-DNN	0.639	0.794	0.765	0.725

III.A.3 Systematic comparison with other toxicity predictions. A systematic comparison with other methods was provided in Table 6. The same datasets are also used to develop the Toxicity Estimation Software Tool (T.E.S.T), so many related results can be found in its user's guide,[42] including hierarchical, single model, FDA, group contribution, nearest neighbor, and T.E.S.T consensus.

Since T.E.S.T is also based on 2D descriptors, the comparison between the results from the present models and T.E.S.T can largely reflect the predictive power of the present models. As shown in Table 6, on the LD_{50}, IGC_{50} and LC_{50} sets, the present MT-DNN consensus always leads to a higher R^2 than T.E.S.T consensus. Especially, on the IGC_{50} and LC_{50} sets, the present MT-DNN consensus models largely beat T.E.S.T (0.794 *vs.* 0.764 and 0.765 *vs.* 0.728), and the present GBDT results also quite outperform T.E.S.T (0.679 *vs.* 0.626) on the LD_{50} set. Even on the LC50-DM set, because the training set is so small (283), ensemble methods (RF and GBDT) and DNN methods are not suitable for it: R^2 of ST-DNN and GBDT are, respectively, 0.523 and 0.486. However, the R^2 of MT-DNN is as high as 0.725 for LC_{50}-DM dataset, which is quite comparable to the T.E.S.T result with an R^2 of 0.739.

2D MT-DNN consensus has an average R^2 of 0.731 for these four datasets, while the average of T.E.S.T model is 0.714, and the recent 3D structure-based topological MT-DNN consensus result is also 0.731.[4] These results confirm that 2D fingerprints

Table 6 Comparison to other toxicity prediction methods. The prediction results for Hierarchical, Single model, FDA, Group contribution, Nearest neighbor, and T.E.S.T consensus are available in ref. 44 and at https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

LD_{50}

Method	R^2	RMSE	Coverage
The present 2D MT-DNN consensus	0.639	0.549	1.000
The present 2D GBDT consensus	0.679	0.580	1.000
Hierarchical[42]	0.578	0.650	0.876
FDA[42]	0.557	0.657	0.984
Nearest neighbor[42]	0.557	0.656	0.993
T.E.S.T consensus[42]	0.626	0.594	0.984
3D MT-DNN consensus[4]	0.653	0.568	0.997

IGC_{50}

Method	R^2	RMSE	Coverage
The present 2D MT-DNN consensus	0.794	0.457	1.000
The present 2D GBDT consensus	0.785	0.457	1.000
Hierarchical[42]	0.719	0.539	0.933
FDA[42]	0.747	0.489	0.978
Group contribution[42]	0.682	0.575	0.955
Nearest neighbor[42]	0.600	0.638	0.986
T.E.S.T consensus[42]	0.764	0.475	0.983
3D MT-DNN consensus[4]	0.802	0.438	1.000

LC_{50}

Method	R^2	RMSE	Coverage
The present 2D MT-DNN consensus	0.765	0.718	1.000
The present 2D GBDT consensus	0.715	0.783	1.000
Hierarchical[42]	0.710	0.801	0.951
Single model[42]	0.704	0.803	0.945
FDA[42]	0.626	0.915	0.945
Group contribution[42]	0.686	0.810	0.872
Nearest neighbor[42]	0.667	0.876	0.939
T.E.S.T consensus[42]	0.728	0.768	0.951
3D MT-DNN consensus[4]	0.789	0.677	1.000

LC_{50}-DM

Method	R^2	RMSE	Coverage
The present 2D MT-DNN consensus	0.725	0.935	1.000
The present 2D GBDT consensus	0.486	1.239	1.000
Hierarchical[42]	0.695	0.979	0.886
Single model[42]	0.697	0.993	0.871
FDA[42]	0.565	1.190	0.900
Group contribution[42]	0.671	0.803	0.657
Nearest neighbor[42]	0.733	0.975	0.871
T.E.S.T consensus[42]	0.739	0.911	0.900
3D MT-DNN consensus[4]	0.678	0.978	1.000

integrated with MT-DNN model surpass the previous 2D models and are as good as the recent 3D structure-based topological model.[4]

III.B Aqueous solubility (log *S*)

For log *S*, following the previous literature,[5,56] we test Klopman's test set[57] with the original train set. The unit of log *S* in these sets is $\log_{10}$ mol L^{-1}. Since the size of the training set is 1290, 10 000 estimators were used in the GBDT model (Table 7).

In the log *S* test, the top 6 fingerprints are MACCS, FP2, Daylight, Estate1, Estate2, and ECFP, which perform much better than the other two fingerprints, Pharm2D and ERG. The consensuses of the top 6 fingerprints results in R and RMSE of 0.944 and 0.684, respectively. The consensus of top 3 is even better, which improves R and RMSE to 0.955 and 0.648 (see Table 8). A systematic comparisons to other methods are included in Table 9. It indicates the present method outperforms all other state-of-the-art 3D and 2D methods.

Table 7 The sizes of log *S* training set and Klopman's test set

Training set	Klopman's test set
1290	21

Table 8 The R and RMSE of predicting log *S* by eight fingerprints and the consensuses of the top 3 and top 6 on Klopman's test set

Fingerprint	R	RMSE
Cons-top 3	0.955	0.648
Cons-top 6	0.944	0.684
MACCS	0.958	0.664
Estate1	0.932	0.791
Daylight	0.923	0.780
FP2	0.908	0.853
ECFP	0.904	0.875
Estate2	0.897	0.907
Pharm2D	0.832	1.114
ERG	0.811	1.202

III.C Partition coefficient (log *P*)

Three log *P* data sets were tested using the GBDT model. The training set has 8199 molecules, which was originally compiled by Cheng *et al.*[58] There are three test sets, namely FDA,[58] Star,[59] and Non-star[59] respectively, which are given in Table 10. The log *P* in these sets is by the unit of $\log_{10}$. Due to the size of the training set, 20 000 estimators are used in the GBDT model.

In order to easily compare to the earlier literatures, accuracies on these three test sets are reported by R^2 or acceptable rate. The acceptable rate here is defined as the percentage of molecules within error range <0.5.[60] Of all the three sets, the 2D fingerprints of Estate2, Estate1, MACCS, and ECFP are always the top 4. The consensuses of the top 4 fingerprints produce R^2 up to 0.901 on the FDA set and attain an acceptable rate on Star set at 71.3%. On the Non-star set, the top 4 consensus is somehow worse than the best single fingerprint

Table 9 Comparison of prediction results on the log *S* data set

Method	R	RMSE
Cons-top 3	0.955	0.648
Cons-top 6	0.944	0.684
MT-ESTD$^+$-1 (3D)[5]	0.94	0.69
Drug-LOGS (2D)[56]	0.94	0.64
Klopman MLR (2D)[57]	0.92	0.86

Table 10 The sizes of log *P* training set and test sets

	Test set		
Training set	FDA	Star	Non-star
8199	406	223	43

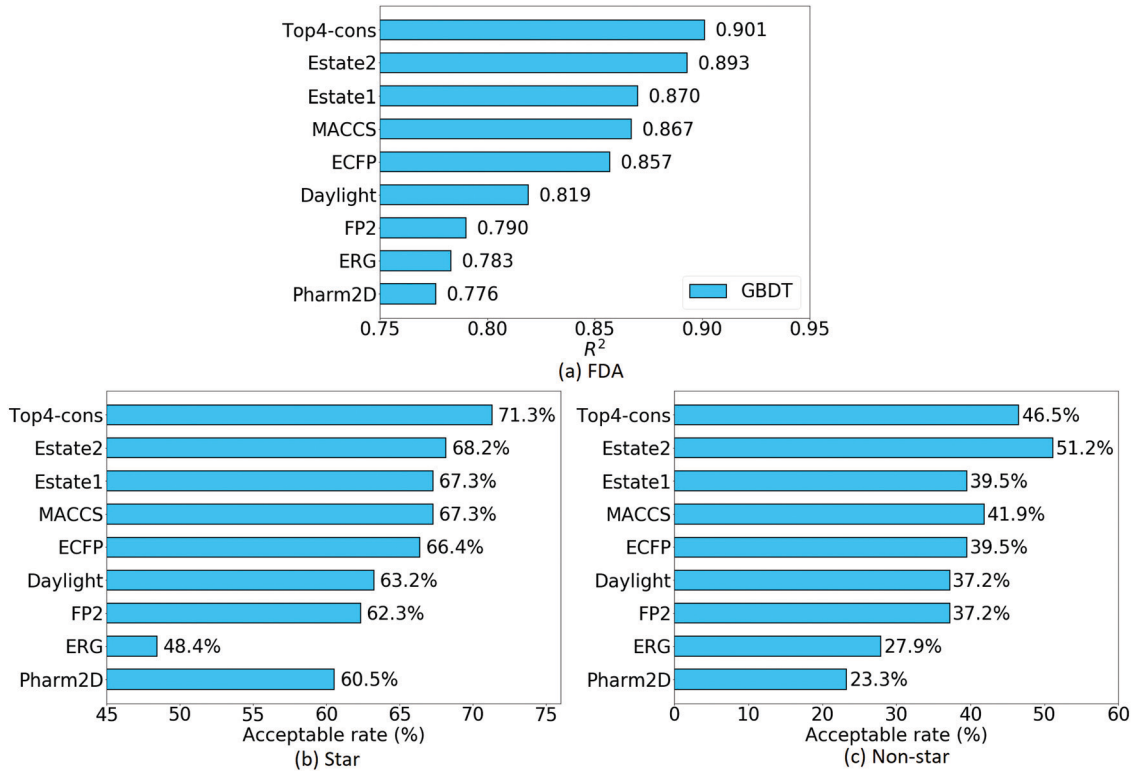


Fig. 5 The performance of eight fingerprints and the consensuses of the top 4 on the FDA, Star and Non-star data sets of log *P*. To be consistent with previous results, on the FDA set, R^2 is given, while on star and non-star datasets, acceptable rate is given.

Estate1 but it is still in the second place with an acceptable rate of 46.5% (see Fig. 5).

A detailed comparison with other log *P* prediction methods was shown in Table 11. On the FDA data set, GBDT-ESTD$^+$-2-AD[5] and MT-ESTD-1[5] are based on 3D descriptors. GBDT-ESTD$^+$-2-AD model includes some molecules from the NIH-dataset in its training set. Therefore, its performance is slightly better than the present one. The 2D method ALOGPS[58] also performs slightly better (0.908 *vs.* 0.901) than the present one. However, a previous study[56] has pointed out that for the PHYSPROP database,[61] the training set of ALOGPS actually contains all of the compounds in the FDA set. It is unclear how well it will perform if the overlapping compounds are removed from the training set. Unlike ALOGPS, XLOGP3's training data is completely independent of the test set.[58]

Table 11 Comparison of log *P* predictions on the FDA set

Method	R^2	RMSE
GBDT-ESTD$^+$-2-AD (2D + 3D)[5]	0.935	0.51
MT-ESTD-1 (3D)[5]	0.920	0.57
ALOGPS (2D but the training set contains test set)[58]	0.908	0.60
Our Cons-top 4 (2D)	0.901	0.63
XLOGP3 (2D)[58]	0.872	0.72
XLOGP3-AA (2D)[58]	0.847	0.80
CLOGP (2D)[58]	0.838	0.88
TOPKAT (2D)[58]	0.815	0.88
ALOGP98 (2D)[58]	0.80	0.90
KowWIN (2D)[58]	0.771	1.10
HINT (2D)[58]	0.491	1.93

In this case, the present prediction is more accurate than that of XLOGP3 (0.901 *vs.* 0.872).

The present results on the Star and Non-star sets are also systematically compared with other stat-of-the-art models as shown in Table 12. For the Star set, 71% of total number of molecules have the predicted error less than 0.5 (acceptable rate 71%). This result is quite satisfactory and is comparable to the 3D structure-based model developed by Wu *et al.*[5] with an acceptable rate of 72% on the same training set ("MT-ESTD-1" in Table 12). There are many commercial software packages developed to predict log *P* such as AB/log *P*,[60] *S*/log *P*,[60] ACD/log *P*,[60] *etc.* However, we cannot validate whether the training sets used in these software packages overlap with the Star set. It is more meaningful when comparing the present model to XLogP3 software[60] since its training dataset does not contain any molecules in the test set. Again, the present model outperforms XLogP3 package on the Star set with the acceptable rates being 71% and 60%, respectively. In the Non-star set, all of the published methods perform as accurate as those in the FDA and Star data set, since the structures in the Non-star set are relatively new and complex. Thus, our model also only achieves an acceptable rate of 47%. However, it is still tied for the third place among all predictors. This result is even better than some 3D structure-based models, though RMSE is relatively high due to a few large outliers.

III.D Protein–ligand binding affinity prediction

III.D.1 The S1322 dataset. To assess the predictive power of 2D-fingerprint based models, two protein–ligand binding affinity

Table 12 Comparison of log *P* predictions of the Star and Nonstar sets

	Star set (N = 223)			Non-star set (N = 43)		
	% of molecules within error range			% of molecules within error range		
Method	<0.5	<1	RMSE	<0.5	<1	RMSE
AB/log P[60]	84	12	0.41	42	23	1.00
MT-ESTD$^+$-1-AD[5]	77	16	0.49	49	19	0.98
S + log P[60]	76	22	0.45	40	35	0.87
ACD/log P[60]	75	17	0.50	44	32	1.00
CLOGP[60]	74	20	0.52	47	28	0.91
MT-ESTD-1[5]	72	18	0.55	33	28	1.01
ALOGPS[60]	71	23	0.53	42	30	0.82
Our cons-top 4	**71**	**18**	**0.625**	**47**	**16**	**1.233**
MiLogP[60]	69	22	0.57	49	30	0.86
KowWIN[60]	68	21	0.64	40	30	1.05
TLOGP[60]	67	16	0.74	30	37	1.12
CSLogP[60]	66	22	0.65	58	19	0.93
SLIPPER-2002[60]	62	22	0.80	35	23	1.23
XLOGP3[60]	60	30	0.62	47	23	0.89
XLOGP2[60]	57	22	0.87	35	23	1.16
QLOGP[60]	48	26	0.96	21	26	1.42
VEGA[60]	47	27	1.04	28	30	1.24
SPARC[60]	45	22	1.36	28	21	1.70
LSER[60]	44	26	1.07	35	16	1.26
CLIP[60]	41	25	1.05	33	9	1.54
MLOGP (Sim+)[60]	38	30	1.26	26	28	1.56
HINTLOGP[60]	34	22	1.80	30	5	2.72
NC + NHET[60]	29	26	1.35	19	16	1.71

datasets were investigated. The first one is denoted as the S1322 set. It is a high quality data set with 1322 protein–ligand complexes involving 7 protein clusters (labeled as CL1, CL2, …, CL7).[29,47] It is a subset of the refined set of PDBbind v2015.[62] The other dataset is PDBbind v2016,[63] in which the refined set excluding the core set in PDBbind v2016 is used as a training data. The core set is a test set. These two sets are summarized in Table 13.

The ligand-based model is used in the present work. For the S1322 set, a 5-fold cross validation was conducted with the GBDT method. To be consistent with the results in the previous literature, accuracy is measured in term of Pearson correlation coefficient (R). Because the results from Daylight and Pharm2D fingerprints are relatively poor, their results are omitted here. The performance of the other six fingerprints (ECFP, FP2, Estate2, MACCS, Estate1, ERG) and their consensus are shown in Fig. 6.

Fig. 6 indicates that for all the seven clusters, the consensuses of the six fingerprints largely achieve better performance than that of any single fingerprint. Specifically, the R values of consensus models are 0.717, 0.847, 0.708, 0.718, 0.831, 0.777, and 0.760 on each of 7 clusters, respectively and 0.765 on average. These results are comparable to ones achieved by a ligand-based 3D topology and GBDT model (Fig. 7).[28]

Table 13 The quantitative summary of the S1322 and PDBbind v2016 data sets

S1322 set							PDBBind v2016 refined set		
CL1	CL2	CL3	CL4	CL5	CL6	CL7	Refined set	Training set	Core set (test set)
333	264	219	156	134	122	94	4057	3767	290

III.D.2 PDBbind v2016 refined set and core set. The present ligand-based model was also tested on PDBbind v2016. Rather than cross validation, this time the core set is regarded as a test set. Quite consistent with core validation on the S1322 set, the consensus of the six fingerprints leads to a large improvement than any single one, with an R of 0.747. These results indicate that the present model has a stable and reliable performance on different protein–ligand binding affinity data sets.

For protein–ligand binding affinity prediction, the present 2D fingerprint-based model is not competitive, because protein–ligand binding not only depends on the ligand, but also on the protein. Therefore, for a more accurate prediction, the information of the protein, at least the information of the binding site should be included. State differently, a complex based model is recommended. Recently, Wójcikowski *et al.*[64] reports 2D fingerprint-based complex models. In their work, a recently developed 2D fingerprint model is used to encode protein–ligand complex information. When combined with DNN, their method gives rise to an R of 0.817 on the PDBBind v2016 core set. Table 14 lists these results.

IV Discussion

IV.A General analysis

In the present work, the predictive power of eight popular 2D fingerprints as well as their consensuses on four important drug-related properties (*i.e.*, toxicity, log S, log P, binding affinity) was investigated. The present study reveals that with a proper machine learning algorithm, the 2D fingerprint-based models including their consensuses outperform other 2D QSPR approaches in most cases, especially on the toxicity predictions. Additionally, 2D fingerprint-based models are comparable to state-of-the-art 3D structure-based models in most drug-related property predictions, except for protein–ligand binding affinity prediction. Considering 2D fingerprints are very ''cheap'' molecular descriptors that are easy and fast to generate, our results are very impressive. It means that 2D fingerprints with appropriate machine learning algorithms are still very valuable for practical problems, such as the prediction of toxicity, the aqueous solubility (log S), and the partition coefficient (log P). However, for protein–ligand binding affinity prediction, complex-based models using 3D topological fingerprints have a major advantage over the present 2D fingerprints, *i.e.*, a GBDT model based on 3D topological fingerprints can achieve about 15% more accurate.[28]

IV.B The performance analysis of 2D fingerprints

IV.B.1 Analysis of 2D fingerprints for PDBbind v2016 core set predictions. The performance of each 2D fingerprint can be systematically analyzed by comparing the difference between prediction errors of every pair of fingerprints as follows.

(1) The relative absolute error for the fth fingerprint on the ith sample (molecule) in the test set is defined by

$$\mathrm{Error}_{f,i} = \frac{|\text{prediction value}_{f,i} - \text{experimental value}_i|}{|\text{experimental value}_i|}$$

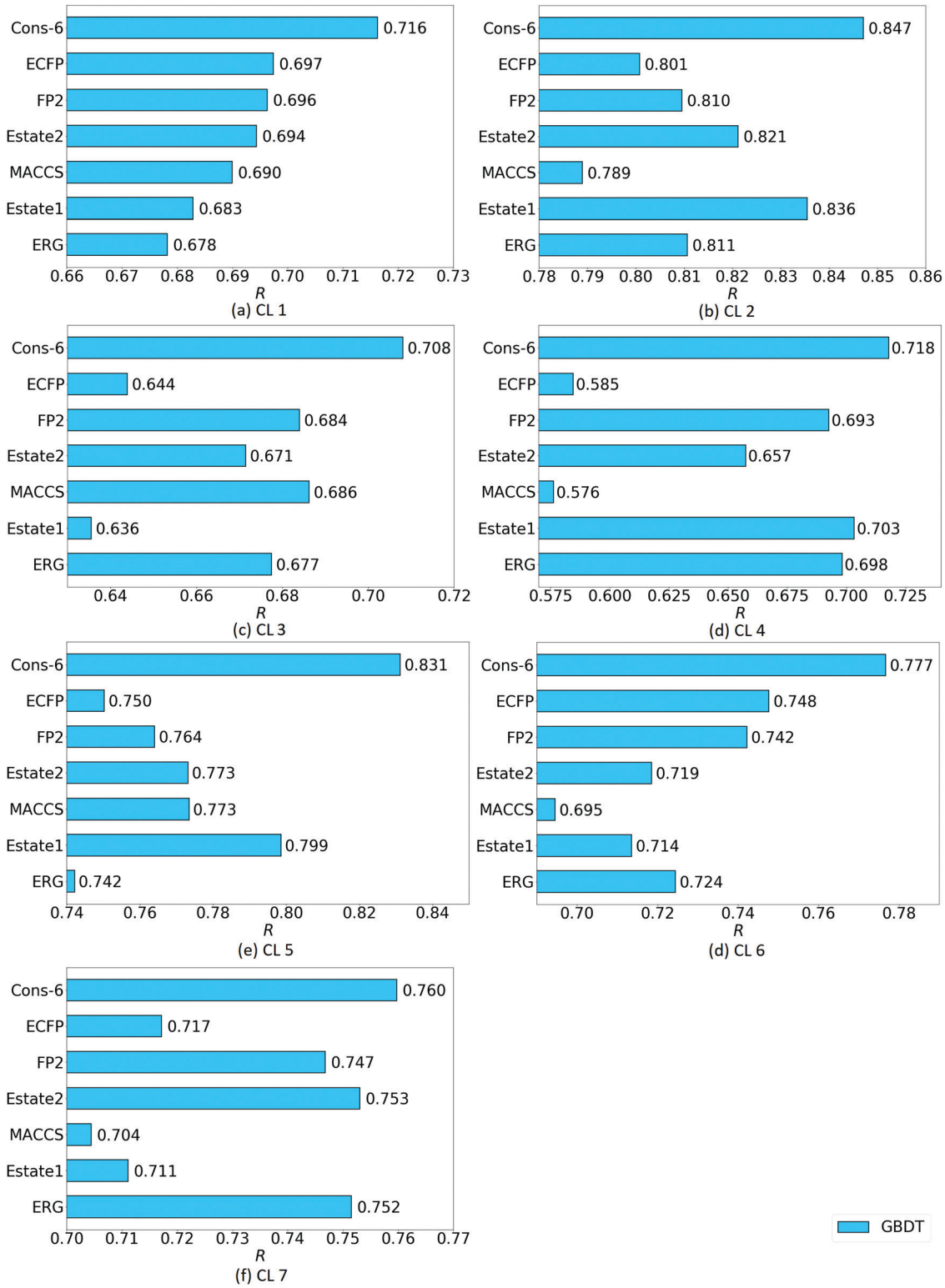


Fig. 6 Pearson correlation coefficient (R) on the seven clusters of the S1322 data set yielded by the six fingerprints (ECFP, FP2, Estate2, MACCS, Estate1, ERG) and their consensuses.

(2) For each molecule, the error difference between each pair of fingerprints is calculated.

(3) Then, the differences for all molecules are ranked from the largest to smallest. The result for PDBbind v2016 core set of 290 complexes is plotted in Fig. 8. We have shown all of 6 pairs for the top four 2D fingerprints.

(4) To further analyze the strength of each fingerprint on certain molecules, we collect those molecules on which a

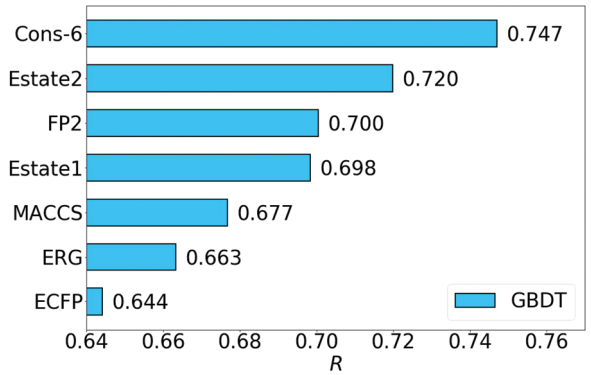


Fig. 7 The R on the PDBbind v2016 binding affinity set yielded by the six fingerprints (ECFP, FP2, Estate2, MACCS, Estate1, ERG) and their consensus.

Table 14 Comparison of protein–ligand binding affinity predictions PDBbind v2016 core set

Method	R	RMSE (kcal mol^{-1})
TopBP (complex)[29]	0.861	1.65
PLEC FP (complex)[64]	0.817	1.71
Our cons-top 6 (ligand)	0.747	2.02

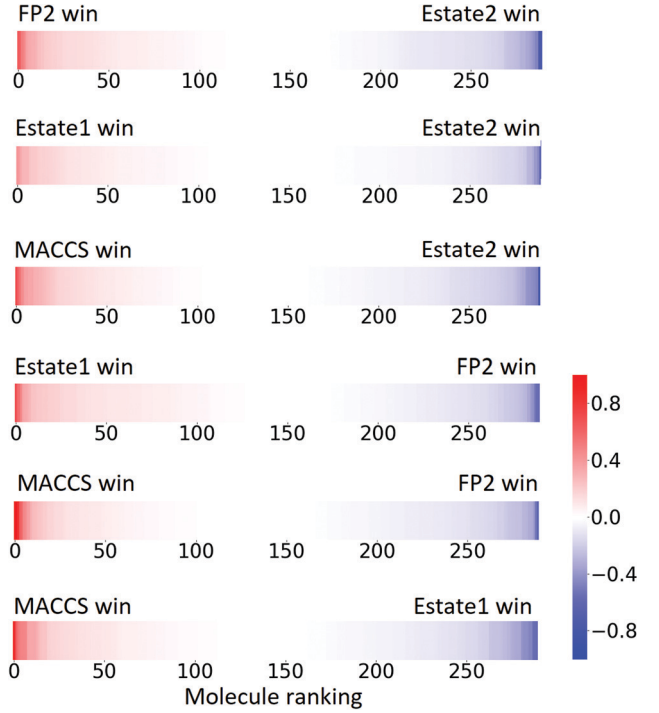


Fig. 8 The ranked error differences between pairs of fingerprints for PDBbind v2016 core set of 290 molecules. Only the top 4 fingerprints (*i.e.*, Estate2, FP2, Estate1, MACCS) are considered.

fingerprint is able to outperform another fingerprint by 0.4 in the error difference.

(5) Among these molecules for each fingerprint, we identify the top 10 most frequently occurring functional groups. The frequency of the occurrence of each functional group, along with the total of number of molecules, are given in Table 15.

This analysis is quite significant as shown in Table 15. It indicates that different fingerprints have different performance on certain functional groups: some fingerprints perform better on some functional groups, while other fingerprints perform better on other functional groups. Our explanation to this is, the different fingerprints are based on different chemical features, since different functional groups have different chemical properties, different fingerprints are sensitive to different functional groups. One can find, in the columns for different fingerprints, the number of functional groups are different, this is because, in the table those molecules on which a fingerprint is able to outperform another fingerprint by 0.4 in the error difference are collected, for different fingerprints, the number of such molecules are different.

One can select an appropriate fingerprint to represent a certain class of functional groups based on Table 15. For the FP2, Estate1, and Estate2 fingerprints, the top two functional groups are carbonyl groups and unfused benzene rings. However, the MACCS fingerprint is different. Its top two functional groups are bicyclic compounds and pyridine. The third top functional groups differ much for four fingerprints: bicyclic compounds for FP2, aniline for Estate1, carboxylate ion for Estate2, and ether for MACCS, which gives us more information to choose fingerprints. Such as, if one has a molecule including aniline, then Estate1 should be selected. Noticeably, some functional groups occur exclusively for one or two types of fingerprints. For example, F, Cl, Br, I is only on the lists of FP2 and Estate1. While azole appears only on the list of FP2 and MACCS and multiple non-fused benzene rings are only for FP2 and Estate2. Moreover, phenol occurs only for Estate1 and furan occurs only for MACCS.

IV.B.2 Analysis of 2D fingerprints for the IGC50 toxicity data set prediction and also other data sets. Using the same 5-step procedure outlined above, we carry out a performance analysis for toxicity dataset IGC_{50}, which is shown in Fig. 9 and Table 16. The molecules in the toxicity data set are typically small and simple, leading to the functional groups in Table 16 also small. Moreover, since there are not too many functional groups in these relatively simple molecules, only top 8 functional groups are presented in the table. Similar to the performance on the binding affinity, for the top 4 fingerprints on the toxicity set, the carbonyl group is in the first place. Unfused benzene rings also have a high occurrence frequency, resulting in the second or third ranking. The difference between the performance of various fingerprints is mainly located on sulfide and aliphatic chains with 8 or more members. FP2 fingerprint works well on sulfide, whereas, Daylight, Estate1 and Estate2 work well on aliphatic chains with 8 or more members.

The same performance analyses were also conducted for other toxicity and log$P$ data sets, the results are shown in Tables S1 to S4 (ESI†). These tables indicate, for the toxicity data sets of LD_{50}, LC_{50}, LC_{50}-DM, the performance of the Estate1 and Estate2 fingerprints are similar, they both work well on bicycle compounds; comparing to it, the FP2 fingerprint

Table 15 The top 10 frequently occurred functional groups in PDBbind v2016 core set for each fingerprint. For each fingerprint, the occurrence frequency and the total number of molecules are also given

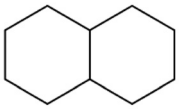
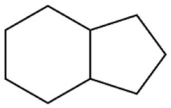
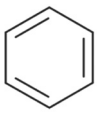
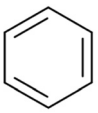
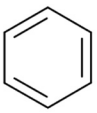
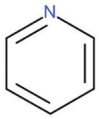
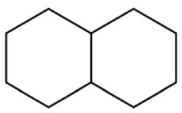
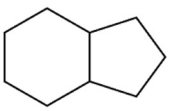
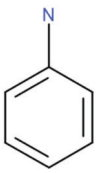
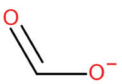
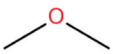
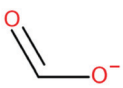
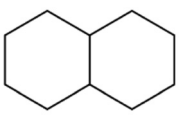
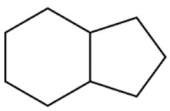
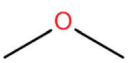
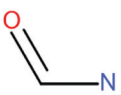
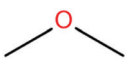
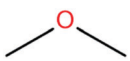
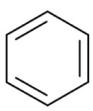
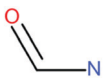
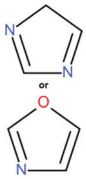
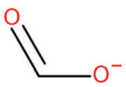
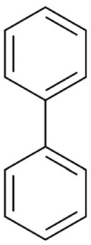
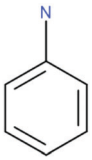
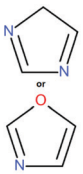
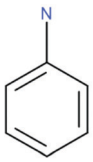
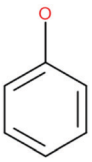
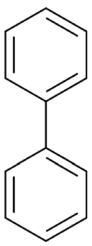
Ranking	FP2	Estate1	Estate2	MACCS
1	 carbonyl group: 24/41	 carbonyl group: 25/42	 carbonyl group: 23/41	 or  ..... bicyclic compounds: 17/36
2	 unfused benzene ring: 21/41	 unfused benzene ring: 18/42	 unfused benzene ring: 22/41	 pyridine: 17/36
3	 or  ..... bicyclic compounds: 19/41	 aniline:14/42	 carboxylate ion: 16/41	 ether: 16/36
4	 hydroxyl: 16/41	 carboxylate ion: 14/42	 or  ..... bicyclic compounds: 15/41	 carbonyl group 15/36
5	 ether: 14/41	 hydroxyl: 14/42	 carbonyl group with N: 13/41	 hydroxyl: 15/36
6	 12/41	 ether: 14/42	 ether: 13/41	 unfused benzene ring: 12/36
7	 amide: 10/41	 carbonyl with Nitrogen: 13/42	 hydroxyl: 11/41	 amide: 10/36

Table 15 (*continued*)

Ranking	FP2	Estate1	Estate2	MACCS
8	 azole: 8/41	 amide: 11/42	 amide: 11/41	 carboxylate ion: 9/36
9	 ..... multiple non-fused benzene rings: 7/41	 11/42	 aniline: 10/41	 azole: 7/36
10	 ..... aniline: 7/41	 phenol: 8/42	 ..... multiple non-fused benzene rings 8/41	 furan: 5/36

works better on aliphatic chains with 8 or more members, the daylight fingerprint has a better performance on amide. For log *P* data set, the ECFP and Estate2 fingerprints lead to a good performance on aniline, the Estate1 fingerprints works better on bicycle compounds; MACCS fingerprint works better on unfused benzene ring.

IV.C The predictive power of the consensus of 2D fingerprints

The consensus of several different fingerprints typically further enhances the performance of a single fingerprint. This enhancement can be quite significant. However, on the datasets of different drug-related properties, the best fingerprint combinations for the consensus are not consistent. One possible explanation is that different fingerprints are good at encoding certain functional groups, and datasets for different drug-related properties have different functional group distributions. This is also the reason why a consensus can enhance performance. The consensus can capture more functional groups and counterbalance the systematical bias from different fingerprints.

On toxicity prediction, the best combination for consensus is obtained with Estate2, Estate1, Daylight, FP2, ECFP, and MACCS. On the log *S* prediction, the best combination is achieved with MACCS, Estate1, and Daylight. While on the log *P* prediction, the best consensus involves Estate2, Estate1, ECFP, and MACCS. Finally, on the binding affinity prediction, the best consensus uses Estate2, Estate1, FP2, ECFP, MACCS, and ERG. It is worth noting that, Estate related (Estate1, Estate2 or both) models are always included in the best combinations. In fact, their single performances are also relatively good. This finding is not surprising since Estate fingerprints encode the intrinsic electronic state of the atom as perturbed by the electronic influence of all other atoms. It is well-known that electronic state is important to drug-related properties.

IV.D Multitask deep learning

Multitask deep learning was utilized on our toxicity prediction. It turns out that the smallest set LC_{50}-DM with only 283 training samples benefits dramatically from the multitask deep learning strategy. Its R^2 value rises from 0.523 to 0.725. This is because, in the frame of multitask deep learning, different data sets (tasks) share similar structure–function relationships. When a small dataset is trained with a large dataset through shared neural networks, the statistics learned from the large datasets in the shared neurons can help predict the small dataset property. As a result, the other three large toxicity sets can share their patterns learned from training with the small toxicity set, enhancing its prediction. Therefore, multitask deep learning could be a useful strategy to train relatively small datasets.

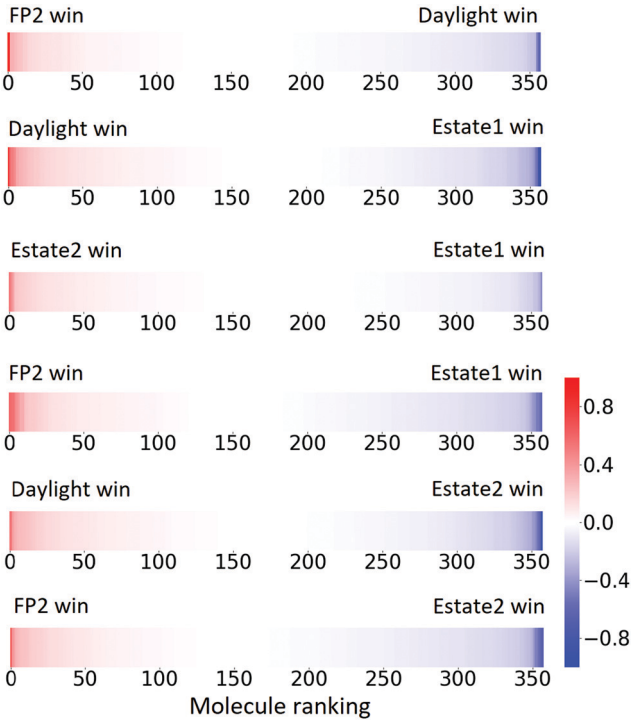


Fig. 9 The ranked error differences between pairs of fingerprints for IGC_{50} toxicity set of 358 molecules. Only the top 4 fingerprints (*i.e.*, Estate2, FP2, Estate1, Daylight) are considered.

IV.E The limitation and advantage of 2D fingerprints

Typically, 2D fingerprints only encode small molecules, such as ligands, although high level 2D fingerprint models including both proteins and ligands have also been developed.[64,65] Theoretically, 2D fingerprints are more suitable for target-independent or target-unspecific problems involving small molecules, such as toxicity, solvation free energy, aqueous solubility, partition coefficient, permeability, *etc.* The current investigation confirms this point. For toxicity, aqueous solubility and partition coefficient, the present 2D-fingerprint based methods perform quite similar to or even somewhat better than 3D structure-based methods in some cases.

For protein–ligand binding affinity predictions, both ligand-based approaches and complex-based are examined. For ligand-based approaches, 2D-fingerprint based methods can perform as well as 3D structure-based models. However, 3D structure-based topological models[29] outperform 2D-fingerprint based methods (*i.e.*, *R*: 0.861 *vs.* 0.747 for PDBbind v2016 core test). In fact, more sophisticated 2D fingerprint models that utilize the protein–ligand complex information and DNN[64,65] are still not as accurate as 3D topology-based models[29] (*i.e.*, *R*: 0.817 *vs.* 0.861 for PDBbind v2016 core test and 0.774 *vs.* 0.808 for PDBbind v2013 core test). Essentially, algebraic topology is designed to simplify the geometric complexity of biological macromolecules. Therefore, it is able to extract vital information from protein–ligand complexes to predict their binding affinities.

When there is no available 3D experimental structure, 3D models can still largely outperform 2D models on the binding affinity prediction. An example occurs in D3R Grand Challenges (D3R GC),[66] in which binding affinities are to be predicted without given 3D experimental structures. Therefore, 3D models can only be built from docking. Even in this circumstance, from GC1 to recent GC4,[41,66–68] 3D models has been always proven to be more reliable than 2D models. For example, in recent GC4,[66] our 3D model (receipt ID ar5p6) achieved the smallest $RMSE_c$ at 0.47 kcal mol^{-1}, while the best 2D model in that competition attained $RMSE_c$ as high as 0.53 kcal mol^{-1}. These results confirm the 3D structured-based model is superior to 2D counterpart in binding affinity prediction even there is no crystal structure.

Moreover, binding affinities typically depend on target (protein). The same ligand can have quite different binding affinities on different targets. The 3D models can take care of binding affinities on different targets but for most 2D models, because of lacking protein information, they work only for a single target.

In general, 2D models can only take care of simple geometry and do not work as well as 3D models do for macromolecules that have complex 3D structures.[43] The complexity of biomolecular structure, function, and dynamics often makes 2D models inconclusive, inadequate, inefficient and sometimes intractable. In contrast, 3D models can easily handle the complexity of biomolecular structures.

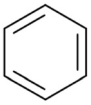
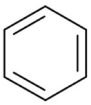
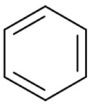
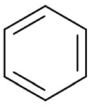
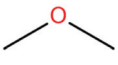
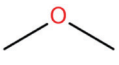
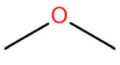
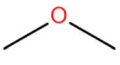
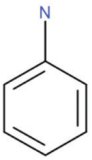
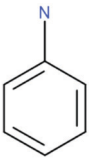
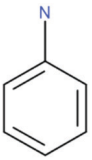
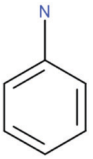
However, the advantage of 2D fingerprints is, they are much easier to generate than 3D structure-based fingerprints built from algebraic topology, differential geometry or various graph theory. Therefore, 2D-fingerprint based models can be useful tools for preliminary drug screening studies.

V Conclusion

Two-dimensional molecular fingerprints, or 2D fingerprints, refer to molecular structural patterns, such as elemental composition, atomic connectivity, functional groups, 2D-pharmacophores *etc.* extracted from a molecule without taking into account the 3D-structural representation of these properties. 2D fingerprints have been a main workhorse for cheminformatics and bioformatics for decades. However, their validations in various datasets were typically carried out long time ago with earlier machine learning algorithms. Recently, new 3D structure-based molecular fingerprints built from algebraic topology,[28,29] differential geometry,[30] geometric graph theory,[31,32] and algebraic graph theory[33] have found much success in drug discovery related applications,[4,5,28,29] including D3R Grand Challenges.[33,41] It raises an interesting issue whether 2D fingerprints are still competitive in drug discovery related applications.

This work reassesses 2D fingerprints for their performance in drug discovery related applications. We consider a total of eight commonly used 2D fingerprints, namely FP2, Daylight, MACCS, Estate1, Estate2, ECFP, Pharm2D, and ERG. Four types of drug discovery related applications with 23 datasets, including solubility ($\log S$) and partition coefficient ($\log P$) that are independent of a target protein, toxicity that may depend on certain

Table 16 The top 10 frequently occurred functional groups in IGC_{50} toxicity set for each fingerprint. For each fingerprint, the occurrence frequency and the total number of molecules are also given

Ranking	FP2	Daylight	Estate1	Estate2
1	 carbonyl group: 14/33	 carbonyl group: 17/34	 carbonyl group: 16/39	 carbonyl group: 14/37
2	 unfused benzene ring: 21/41	 hydroxyl: 9/34	 hydroxyl: 15/39	 hydroxyl: 14/37
3	 amide: 9/33	 unfused benzene ring: 9/34	 unfused benzene ring: 9/39	 unfused benzene ring: 10/37
4	 hydroxyl: 9/33	 amide: 7/34	 ether: 7/39	 ether: 8/37
5	 ether: 8/33	 ether: 7/34	 6/39	 8/37
6	 5/33	 6/34	 amine: 6/39	 amine: 6/37
7	 sulfide: 3/33	aliphatic chains with 8 or more members: 5/34	aliphatic chains with 8 or more members: 6/39	aliphatic chains with 8 or more members: 5/37
8	 aniline: 3/33	 aniline: 4/34	 aniline: 3/39	 aniline: 5/37

unknown target proteins, and protein–ligand binding affinity that depend on known target proteins, are designed to validate 2D fingerprints. Advanced machine learning algorithms, including random forest (RF), gradient boosting decision trees (GBDT), single-task deep neural network (ST-DNN), and multitask deep neural network (MT-DNN) are used to optimize the performance of the above 2D fingerprints in the aforementioned four types of datasets. In particular, MT-DNN is designed to enhance the performance of 2D fingerprints on relatively small datasets by a simultaneous training with relatively large datasets that share a similar pattern. Since each fingerprint may have an explicit bias on certain functional groups or 2D patterns, we carry out various consensus to further boost the performance of 2D fingerprints in all the datasets. Finally, the strengths of top four 2D fingerprints for predicting protein–ligand binding affinity and quantitative toxicity are analyzed in detail.

Our general findings are as follows. (1) 2D fingerprint-based models are as good as 3D structure-based models for various toxicity, $\log S$ and $\log P$ datasets under the same training-test condition. (2) For ligand-based protein–ligand binding affinity predictions, 2D fingerprint-based models perform equally well as 3D structure-based models that are based only on ligand 3D structures. (3) 3D structure-based models that utilize 3D protein–ligand complex information outperform 2D fingerprints that based on either ligand information or protein–ligand complex information. (4) Advanced machine learning algorithms, such as DNN and MT-DNN, are crucial for 2D fingerprints to achieve optimal performance. (5) There is no 2D fingerprint that outperforms all other 2D fingerprints in all applications. However, Estate related (Estate1, Estate2 or both) models always perform well. (6) Appropriate consensus of a few 2D models typically achieves better performance. Therefore, if combined with advanced machine learning algorithms, the 2D fingerprints are still competitive in most drug discovery related applications except for those that involve macromolecular structures.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported in part by NSF Grants DMS-1721024, DMS-1761320, and IIS1900473 and NIH grant GM126189.

References

1 L. Di and E. H. Kerns, *Drug-like properties: concepts, structure design and methods from ADME to toxicity optimization*, Academic Press, 2015.
2 N. M. Henriksen, A. T. Fenley and M. K. Gilson, *J. Chem. Theory Comput.*, 2015, **11**, 4377–4394.
3 K. Gao, J. Yin, N. M. Henriksen, A. T. Fenley and M. K. Gilson, *J. Chem. Theory Comput.*, 2015, **11**, 4555–4564.
4 K. Wu and G.-W. Wei, *J. Chem. Inf. Model.*, 2018, **58**, 520–531.
5 K. Wu, Z. Zhao, R. Wang and G.-W. Wei, *J. Comput. Chem.*, 2018, **39**, 1444–1454.
6 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 1997, **23**, 3–25.
7 L. Di and E. H. Kerns, *Drug Discovery Today*, 2006, **11**, 446–451.
8 A. L. Hopkins, G. M. Keserü, P. D. Leeson, D. C. Rees and C. H. Reynolds, *Nat. Rev. Drug Discovery*, 2014, **13**, 105.
9 P. Atallah, K. B. Wagener and M. D. Schulz, *Macromolecules*, 2013, **46**, 4735–4741.
10 H. Van De Waterbeemd and E. Gifford, *Nat. Rev. Drug Discovery*, 2003, **2**, 192.
11 C. Hansch, P. P. Maloney, T. Fujita and R. M. Muir, *Nature*, 1962, **194**, 178.
12 H. Geppert, M. Vogt and J. Bajorath, *J. Chem. Inf. Model.*, 2010, **50**, 205–216.
13 K. Roy and I. Mitra, *Curr. Comput.-Aided Drug Des.*, 2012, **8**, 135–158.
14 M. Tareq Hassan Khan, *Curr. Drug Metab.*, 2010, **11**, 285–295.
15 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
16 Y.-C. Lo, S. E. Rensi, W. Torng and R. B. Altman, *Drug Discovery Today*, 2018, **23**, 1538–1546.
17 A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé and G. Pujadas, *Methods*, 2015, **71**, 58–63.
18 J. Verma, V. M. Khedkar and E. C. Coutinho, *Curr. Top. Med. Chem.*, 2010, **10**, 95–115.
19 J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1273–1280.
20 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 33.
21 I. Daylight Chemical Information Systems, Daylight, https://hadoop.apache.org.
22 L. H. Hall and L. B. Kier, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 1039–1045.
23 M. J. McGregor and S. M. Muskal, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 569–574.
24 J. S. Mason and D. L. Cheney, *Biocomputing 2000*, World Scientific, 1999, pp. 576–587.
25 N. Stiefl, I. A. Watson, K. Baumann and A. Zaliani, *J. Chem. Inf. Model.*, 2006, **46**, 208–220.
26 D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, *Advances in neural information processing systems*, 2015, 2224–2232.
27 K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley and M. Mathea, *et al.*, 2019, arXiv preprint arXiv:1904.01561.
28 Z. Cang and G.-W. Wei, *Int. J. Numerical Methods Biomed. Eng.*, 2018, **34**, e2914.
29 Z. Cang, L. Mu and G.-W. Wei, *PLoS Comput. Biol.*, 2018, **14**, e1005929.
30 D. D. Nguyen and G.-W. Wei, *Int. J. Numerical Methods Biomed. Eng.*, 2019, **35**, e3179.
31 D. D. Nguyen, T. Xiao, M. Wang and G.-W. Wei, *J. Chem. Inf. Model.*, 2017, **57**, 1715–1721.
32 D. Bramer and G.-W. Wei, *J. Chem. Phys.*, 2018, **148**, 054103.
33 D. D. Nguyen, Z. Cang, K. Wu, M. Wang, Y. Cao and G.-W. Wei, *J. Comput.-Aided Mol. Des.*, 2019, **33**, 71–82.
34 V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1947–1958.
35 R. E. Schapire, *Nonlinear estimation and classification*, Springer, 2003, pp. 149–171.
36 I. A. Basheer and M. Hajmeer, *J. Microbiol. Methods*, 2000, **43**, 3–31.
37 R. Caruana, *Mach. Learn.*, 1997, **28**, 41–75.
38 M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015, https://www.tensorflow.org/, Software available from tensorflow.org.
39 A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga and A. Lerer, NIPS Autodiff Workshop, 2017.
40 Z. Cang and G.-W. Wei, *Bioinformatics*, 2017, **33**, 3549–3557.
41 Z. Gaieb, C. D. Parks, M. Chiu, H. Yang, C. Shao, W. P. Walters, M. H. Lambert, N. Nevins, S. D. Bembenek and M. K. Ameriks, *et al.*, *J. Comput.-Aided Mol. Des.*, 2019, **33**, 1–18.
42 T. Martin, User's guide for TEST (version 4.2) (Toxicity Estimation Software Tool): A program to estimate toxicity from molecular structure, 2016.
43 D. D. Nguyen, Z. Cang and G.-W. Wei, *Phys. Chem. Chem. Phys.*, 2020, **22**, 4343–4367.
44 G. Landrum, *et al.*, RDKit: Open-source cheminformatics, 2006.
45 H. Morgan, *J. Chem. Doc.*, 1965, **5**, 107–113.

46 B. Wang, C. Wang, K. Wu and G.-W. Wei, *J. Comput. Chem.*, 2018, **39**, 217–233.
47 B. Wang, Z. Zhao, D. D. Nguyen and G.-W. Wei, *Theor. Chem. Acc.*, 2017, **136**, 55.
48 S. J. Capuzzi, R. Politi, O. Isayev, S. Farag and A. Tropsha, *Front. Environ. Sci.*, 2016, **4**, 3.
49 B. Ramsundar, B. Liu, Z. Wu, A. Verras, M. Tudor, R. P. Sheridan and V. Pande, *J. Chem. Inf. Model.*, 2017, **57**, 2068–2076.
50 J. Wenzel, H. Matter and F. Schmidt, *J. Chem. Inf. Model.*, 2019, **59**, 1253–1268.
51 Z. Ye, Y. Yang, X. Li, D. Cao and D. Ouyang, *Mol. Pharmaceutics*, 2018, **16**, 533–541.
52 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *et al.*, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
53 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein and L. Antiga, *et al.*, *Advances in Neural Information Processing Systems*, 2019, 8024–8035.
54 K. S. Akers, G. D. Sinks and T. W. Schultz, *Environ. Toxicol. Pharmacol.*, 1999, **7**, 33–39.
55 H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica, T. Oberg, P. Dao, A. Cherkasov and I. V. Tetko, *J. Chem. Inf. Model.*, 2008, **48**, 766–784.
56 T. Hou, K. Xia, W. Zhang and X. Xu, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 266–275.
57 G. Klopman, S. Wang and D. M. Balthasar, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 474–482.
58 T. Cheng, Y. Zhao, X. Li, F. Lin, Y. Xu, X. Zhang, Y. Li, R. Wang and L. Lai, *J. Chem. Inf. Model.*, 2007, **47**, 2140–2148.
59 A. Avdeef, *Absorption and drug development: solubility, permeability, and charge state*, John Wiley & Sons, 2012.
60 R. Mannhold, G. I. Poda, C. Ostermann and I. V. Tetko, *J. Pharm. Sci.*, 2009, **98**, 861–893.
61 P. Howard and W. Meylan, Physical/chemical property database (PHYSPROP), 1999.
62 Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu and R. Wang, *Bioinformatics*, 2014, **31**, 405–412.
63 M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li and R. Wang, *J. Chem. Inf. Model.*, 2018, **59**, 895–913.
64 M. Wójcikowski, M. Kukiełka, M. M. Stepniewska-Dziubinska and P. Siedlecki, *Bioinformatics*, 2019, **35**, 1334–1341.
65 I. Kundu, G. Paul and R. Banerjee, *RSC Adv.*, 2018, **8**, 12127–12137.
66 C. D. Parks, Z. Gaieb, M. Chiu, H. Yang, C. Shao, W. P. Walters, J. M. Jansen, G. McGaughey, R. A. Lewis and S. D. Bembenek, *et al.*, *J. Comput.-Aided Mol. Des.*, 2020, **34**, 99–119.
67 S. Gathiaka, S. Liu, M. Chiu, H. Yang, J. A. Stuckey, Y. N. Kang, J. Delproposto, G. Kubish, J. B. Dunbar and H. A. Carlson, *et al.*, *J. Comput.-Aided Mol. Des.*, 2016, **30**, 651–668.
68 Z. Gaieb, S. Liu, S. Gathiaka, M. Chiu, H. Yang, C. Shao, V. A. Feher, W. P. Walters, B. Kuhn and M. G. Rudolph, *et al.*, *J. Comput.-Aided Mol. Des.*, 2018, **32**, 1–20.

