QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach

Dmitry V. Zankov, Maria Matveieva, Aleksandra V. Nikonenko, Ramil I. Nugmanov, Igor I. Baskin, Alexandre Varnek, Pavel Polishchuk, and Timur I. Madzhidov*

Cite This: J. Chem. Inf. Model. 2021, 61, 4913−4923

ABSTRACT: Modern QSAR approaches have wide practical applications in drug discovery for designing potentially bioactive molecules. If such models are based on the use of 2D descriptors, important information contained in the spatial structures of molecules is lost. The major problem in constructing models using 3D descriptors is the choice of a putative bioactive conformation, which affects the predictive performance. The multi-instance (MI) learning approach considering multiple conformations in model training could be a reasonable solution to the above problem. In this study, we implemented several multi-instance algorithms, both conventional and based on deep learning, and investigated their performance. We compared the performance of MI-QSAR models with those based on the classical single-instance QSAR (SI-QSAR) approach in which each molecule is encoded by either 2D descriptors computed for the corresponding molecular graph or 3D descriptors issued for a single lowest energy conformation. The calculations were carried out on 175 data sets extracted from the ChEMBL23 database. It is demonstrated that (i) MI-QSAR outperforms SI-QSAR in numerous cases and (ii) MI algorithms can automatically identify plausible bioactive conformations.

INTRODUCTION

A typical QSAR model establishes a relationship between bioactivity and molecular structure represented by a vector of molecular descriptors. Meanwhile, one can consider descriptors of different dimensionality: 0D (derived from the empirical formula), 1D (derived from a vector of values, e.g., fingerprints), 2D (derived from a molecular graph), 3D (derived from a single conformation), and 4D (usually derived from a molecular-dynamic trajectory). Although 2D descriptors are a gold standard in QSAR modeling because of the simplicity of their calculation, 2D representation does not directly encode the spatial structure of molecules which is important for protein−ligand recognition. Ignoring this information may reduce the performance of QSAR models. This motivates the development of 3D-QSAR methods which consider explicitly the spatial structure of the molecules.

The first proposed 3D-QSAR method was Comparative Molecular Field Analysis (CoMFA), which correlates the biological activity of organic molecules with their electrostatic and “steric” fields represented as interaction energies with special probes placed at grid nodes around an aligned set of molecules. To build a CoMFA model, a single conformation (3D structure) should be chosen for each molecule, followed by their alignment in space and calculation of interaction energies considered as descriptors. Choosing irrelevant conformations and/or alignment may result in a substantial decrease in model performance. This issue becomes critical for flexible molecules possessing several rotatable bonds and, as a consequence, many possible conformations. Following CoMFA, most 3D-QSAR methods rely on the choice of a single “bioactive” conformation for a molecule, which can be determined from the structures of protein−ligand complexes. Although such conformations can be determined in X-ray, NMR, and EM (electron microscopy) studies or computed using molecular modeling techniques (docking, molecular dynamics, etc.), there is a clear indication that using “receptor-bound” conformations might be a bad choice for building QSAR models. Although multiple 3D-QSAR approaches have

Received: June 17, 2021
Published: September 23, 2021

https://doi.org/10.1021/acs.jcim.1c00692

See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.
been developed so far, all of them suffer from the issue of selection of relevant conformations.4−12

The concept of 4D-QSAR,13,14 in which a molecule is represented by an ensemble of conformations, was introduced to overcome the limitations of the 3D-QSAR approach associated with the choice of a single conformation for each molecule. Such ensembles may be extracted from molecular-dynamics trajectories, sampled from Monte Carlo simulations, or obtained using conformer generators. Most 4D-QSAR approaches compute 3D alignment-independent descriptors for individual conformations and combine them by some schemes to obtain a single vector of descriptors for each molecule to be used in conventional machine learning methods. The most widely used schemes are summation or averaging of 3D descriptors of individual conformations15 and summation of 3D descriptors weighted by the Boltzmann factor estimated for conformations in vacuum or water solution.16−18 Considering that the (a) energy assessment of the conformations is subjected to high errors in the parametrization of force fields and (b) energy of the receptor-bound conformations of ligands may be rather high and hence their Boltzmann factor may be very low, the Boltzmann averaging schemes may introduce significant noise to the data.

Multi-instance (MI) machine learning approaches can be used to solve the issues of representation of each molecule by multiple conformations (instances) and automatic selection of the most relevant ones (Figure 1). In the multi-instance approach, an example (i.e., a molecule) is presented by a bag of instances (i.e., a set of conformations), and a label (a bioactivity value) is available only for a bag (a molecule) but not for individual instances (conformations). MI learning was first introduced for recognizing handwritten numbers19 but became better known after the paper by Dietterich et al.,20 where the authors developed a model to predict the odors of compounds, which were represented by multiple conformations. The Compass algorithm21 is another example where MI learning significantly improved the performance of models in comparison with single-instance (SI) learning on the task of predicting the bioactivity of the compounds. The Compass algorithm implemented the idea of representing a molecule by multiple conformations, which were used to train a neural network. Though MI learning was initially developed for modeling the properties/activities of chemical compounds, this methodology has not found wide application in the QSAR area, although it has become widely adopted in other fields.22 Only a few studies with the application of MI learning to predict the bioactivity of the compounds have been published so far in mathematics and bioinformatics journals.20,23−25 Moreover, recently proposed deep learning-based multi-instance approaches have not been used in the chemistry domain except in our recent work.26 Recently, we demonstrated the applicability of unsupervised27 and supervised clustering-based MI approaches to bioactivity predictions on several data sets.28 However, a proper comparison of MI learning approaches to conventional ones has not been made so far.

The main goal of MI learning algorithms is to predict a label for an object represented by a bag of instances. However, it is often desirable not only to predict the bag label (in our case, to assess the bioactivity of a given molecule) but also to identify the key instances in the bag (i.e., to assess bioactive conformations). This problem, called Key Instance Detection (KID), was first formalized in a prior publication.29 The
identification of the conformation responsible for the observed bioactivity of the molecule provides deeper insight into the interaction mechanism between the ligand and the target protein.

In this study, we show that the application of MI learning can be used to solve the long-standing problem of 3D-QSAR—the selection of relevant (or biologically active) conformations for modeling. Instead of a single conformation, MI learning considers the whole conformational ensemble, which significantly improves the predictive performance of the models based on 3D descriptors. Here, several 3D MI QSAR approaches were implemented and their performances compared on numerous data sets. It has been demonstrated that in most cases, the 3D MI QSAR models outperformed conventional 2D models. We also identified the physicochemical characteristics of compounds impacting the performance of 3D MI or 2D models. In addition, we studied the ability of MI models based on attention neural networks to identify relevant bioactive conformations.

## METHODS

### Data Sets

One hundred seventy-five data sets of compounds with measured pKᵢ or pIC₅₀ values were extracted from the ChEMBL23 database. The size of the data sets varied from several hundred to several thousand compounds (Figure 2a). Molecules with a molecular weight greater than 700 (3% of the total number of molecules) were discarded. Because the performance of the 3D models may depend on the flexibility of the studied compounds, the average number of rotatable bonds for molecules in each data set was calculated using RDKit (Figure 2b). Most molecules in the data sets can be considered as low to moderately flexible with the average number of rotatable bonds within 3–6.

In addition, in the collected data sets, we identified compounds deposited in the Protein Data Bank (PDB) and retrieved their conformations. These PDB conformations were used as references to compare with the conformations predicted by MI models to provide the largest contribution to biological activity.

### Conformation Generation

Conformations representing each molecule were generated using the algorithm implemented in RDKit, which is claimed by its authors to be able to reproduce bioactive conformations observed for ligands in PDB complexes with reasonable accuracy. This ability is important because it may improve the performance of the obtained models, may make them more reasonable, and in the case of MI modeling approaches would increase the probability of identifying the most relevant/contributed conformations to the studied end point. It also increases the chance to find conformations similar to those observed in the X-ray structures of protein–ligand complexes if the latter are available. In our study, we generated up to 100 conformations and removed conformations with RMSD values below 0.5 Å to the remaining ones to reduce redundancy.

### Descriptors

For the descriptor representation of the conformations, we used previously developed 3D pharmacophore signatures. Each conformation is represented by a set of pharmacophore features (H-bond donor/acceptor, center of positive/negative charge, hydrophobic, and aromatic) determined by applying the corresponding SMARTS patterns. All possible quadruplets of features of a particular conformation were enumerated. Distances between features were binned to allow fuzzy matching of quadruplets with small differences in the position of features. Here, we used the 1 Å bin step as it demonstrated reasonable performance in our previous studies.

Three-dimensional pharmacophore signatures were generated for each quadruplet according to the algorithm described in our previous publication. These signatures consider distances between features and their spatial arrangement to recognize the stereoconfiguration of the quadruplets. We counted the number of identical 3D pharmacophore quadruplet signatures for each conformation and used the obtained vectors as descriptors for model building. The 3D pharmacophore descriptors used in this study were implemented in the pmapper Python package (https://github.com/DrrDom/pmapper). Since the pharmacophore descriptors were very sparse, we kept only those quadruplets that occurred in at least 5% of all conformations of the data set molecules.

To build 2D models, we chose binary Morgan fingerprints (MorganFP) of radius 2 and size 2048 calculated with RDKit because they are widely used 2D descriptors and demonstrated high performance in previous benchmarking studies. For comparative purposes, we also used 2D physicochemical descriptors (PhysChem) and binary 2D pharmacophore fingerprints (PharmFP) calculated with RDKit. The former included EState indexes, the number of different pharmacophore features, rings systems, functional groups, and fragments (the full list is provided in the Supporting Information). To calculate the 2D pharmacophore descriptors, we used the same definitions of the pharmacophore features as in pmapper to make the comparison more robust. Afterward, pharmacophore triplets were enumerated using default binning of the topological distances (0–2, 2–5, 5–8, 8+).

### Algorithms

In conventional SI-QSAR, each molecule is represented by a single vector of 2D descriptors computed for
the corresponding molecular graph or 3D descriptors for its lowest energy conformation. In MI-QSAR, a molecule is represented by a set of conformations and a set of associated vectors of descriptors which forms a bag of instances. To build a model in this case special algorithms should be applied. All of the considered MI algorithms can be divided into two groups:

Figure 3. MI wrapper algorithms: (a) Instance-Wrapper and (b) Bag-Wrapper. Learning algorithm (SI Algorithm in the figure) was a three-layer fully connected neural network having 256, 128, and 64 neurons in hidden layers.

Figure 4. Multi-instance neural networks: (a) Instance-Net, (b) Bag-Net, and (c) Bag-AttentionNet.
instance based and bag based.\textsuperscript{35} Instance-based algorithms consider each conformation as a separate training instance. Bag-based algorithms, on the contrary, represent a molecule by a single vector of descriptors, which is produced from the vectors of the conformation descriptors.

Single-Instance Algorithms. We considered traditional SI learning as a baseline approach, where a molecule is described by a vector of 2D descriptors or a vector of 3D descriptors associated with the lowest energy conformation. We used a three-layered fully connected neural network with ReLU activation to construct SI-QSAR models. Our tests show that such architecture gives quite high and stable results across different data sets and does not require additional hyperparameter adjustment for a particular data set.

Multi-Instance Wrappers. The learning process in instance-based algorithms occurs at the instance level. Instance-level learning is applicable if it is possible to assign a label to individual instances in a bag. Also, it is assumed that there is a rule that aggregates the predictions for each instance to get the prediction for the entire bag. The simplest instance-based MI algorithm is Instance-Wraper, where each training instance of a bag is assigned the same label as for the whole bag. This means, for example, that if a molecule is bioactive, it is assumed that all of its conformations are bioactive. As a result, one gets a data set where each conformation is an individual training object and any conventional ML algorithms can be applied to build the model. Given a new molecule, the bioactivity is predicted for each conformation and predictions are averaged to get the final predicted bioactivity of the molecule (Figure 3a). This approach has an obvious drawback because assigning the same bioactivity to all conformations of a molecule in a training set can bring some noise into the learning process because the fact that a molecule is bioactive does not mean that all of its conformations are biologically relevant and responsible for protein—ligand recognition.

The learning process of bag-based algorithms occurs at the bag level. In bag-based algorithms, there is no need to identify a label for each instance in a bag. Instead, there is an operation that aggregates the instances to get a single vector representing the entire bag. Our implementation of the Bag-Wrapper algorithm averaged descriptor values across all conformations and supplied this single vector of descriptors to a conventional SI machine learning method—a three-layer fully connected neural network (Figure 3b). The Bag-Wrapper algorithm has a drawback similar to Instance-Wraper because aggregation of the descriptor vectors of all conformations to the resultant vector may introduce additional noise due to the contribution of irrelevant conformations.

Multi-Instance Neural Networks. Multi-instance neural networks learn in an end-to-end way and take a bag of instances as input and directly output bag prediction. All parameters in MI networks are optimized via back-propagation. Wang et al.\textsuperscript{36} revisited MI neural networks and proposed a series of novel neural network frameworks for MI learning. They considered two types of MI neural networks: mi-Net (hereafter Instance-Net) and MI-Net (hereafter Bag-Net). We implemented both of these neural network architectures. In Instance-Net (Figure 4a), instances are running through fully connected layers and an output neuron. Then, instance predictions are averaged in the pooling layer to obtain a bag prediction, and its error is calculated and backpropagated to adjust model weights. Bag-Net (Figure 4b) consists of three fully connected layers followed by one pooling layer. The pooling layer averages instance representations learned by previous layers into a single embedding vector as a bag representation. The last fully connected layer takes the embedding vector as input and outputs the bag prediction. Wang et al.\textsuperscript{36} examined three typical pooling operators—max pooling, mean pooling, and log-sum-exp pooling—and concluded that all of them provided a similar performance on benchmark data sets. Our tests also supported this conclusion for bioactivity prediction; thus, only mean pooling was applied.

The Bag-Net uses an unlearnable mean pooling function, and as mentioned above, the irrelevant conformations can contribute noise to the prediction and reduce model performance. This drawback can be eliminated using more flexible types of pooling, such as weighted averaging pooling, known as attention. This type of pooling was proposed in another publication,\textsuperscript{37} where an additional two-layered neural network was used to obtain weights of instances. In the Bag-AttentionNet (Figure 4c), all instances are first fed to three fully connected layers. Then, the learned instance representations are used by the attention network with a single hidden layer. In the attention network, the number of output neurons is equal to the number of instances. The output layer of attention has the Softmax activation function and predicts instance weights. Finally, the instance weights given by the attention network are used for weighted averaging of instance representations to get the embedding vector that is used to produce the bag prediction. Implementation of weighted pooling enables the Bag-AttentionNet to automatically identify probable bioactive conformations.

Experimental Setup. A large-scale comparative analysis of the above MI approaches was carried out using 175 data sets extracted from the ChEMBL database. Each data set was randomly divided into a training, validation, and test set. The test set comprised 20% of the molecules of the initial data set; the rest was used as a modeling set. In turn, the latter was divided into a training set (80% of modeling set) and a validation set (20%) used for hyperparameter adjustment.

The Bag-AttentionNet provides attention weights that determine the contribution of each conformation to the predicted bioactivity. We applied regularization of attention weights to force the Bag-AttentionNet network to more strongly highlight key conformations during training. In each training epoch, instances (conformations) were ranked by the attention unit of Bag-AttentionNet. Then X percent of instances ($X \in \{10\%, 20\%, 40\%, 60\%, 80\%, 90\%, 95\%\}$) was discarded and followed by recalculation of the attention weights for the remaining key instances. The number of discarded instances was a hyperparameter adjusted during the training.

To compare several algorithms on multiple data sets, we follow the recommendations from refs 38 and 39. First, we performed the Friedman test to reject the null hypothesis, which is that there were no significant differences in the performance of the models. Then, we performed a pairwise comparison of models using the Wilcoxon–Holm test with a significance level of 5%. The results of a pairwise comparison of the models were visualized with a critical difference diagram.\textsuperscript{38,40} The horizontal lines on the critical difference diagrams connect models that are not significantly different in performance. The pipeline of construction of critical difference diagrams is reported in more detail in the Supporting Information.
For clarity, a special name was assigned to each type of model. It consists of several parts: Representation Level/ Learning Type/Algorithm. The Representation Level denotes the descriptors type (2D or 3D). The Learning Type distinguishes between single- and multi-instance learning schemes (SI or MI) or in the case of 2D models the type of descriptors used. The Algorithm is the machine learning method that was used to build the model. For example, 2D/MorganFP/Net denotes a neural network model based on 2D Morgan fingerprints. The 3D/SI/Net model was trained on the single lowest energy conformations of molecules represented by 3D pharmacophore descriptors using an ordinary multilayer neural network as the learning algorithm.

For analysis of groups of models and pairwise model comparison, we excluded the data sets for which all compared approaches resulted in models with low performance on the test set ($R_{	ext{test}}^2 < 0.4$). Thus, the total number of compared data sets differs as a function of the list of compared models.

We consider that the threshold 0.4 is reasonable for several reasons. (i) We did not tweak every model too much; therefore, we believe there is a room to improve them with tight tuning. (ii) We performed only a comparison between models and do not suggest using them for predictive modeling. (iii) The results and conclusions do not change if we will choose threshold 0.5, but this will decrease the number of considered data sets. Nevertheless, all data are disclosed in the Supporting Information (Tables S1 and S4), and these conclusions can be verified.

## RESULTS AND DISCUSSION

In this section, we present the results of a comparative analysis of single- and multi-instance learning approaches. For clarity, we first present the results of benchmarking MI learning algorithms to choose the best MI models. Then, we compare the best 3D MI models with 3D SI models as well as with conventional 2D SI models and evaluate the ability of MI models to identify relevant conformations in comparison to docking.

**Benchmarking of Multi-Instance Algorithms.** For 45 data sets out of 175, no MI models achieved the required performance of $R_{	ext{test}}^2 > 0.4$. These “non-modellable” data sets were excluded from further consideration, and benchmarking analysis was performed on the remaining 130 data sets. Among the two simplest approaches represented by wrapper algorithms, Instance-Wrapper performs significantly better than Bag-Wrapper (Figure 5). Thus, considering each conformation as an individual training example represents a better strategy than averaging descriptor vectors of individual conformations.

MI neural networks represent a group of methods specially modified to solve MI problems. In Bag-Net, the mean pooling operation is performed not on descriptors of particular conformations (as in Bag-Wrapper) but on their embeddings, resulting from descriptors transformation by three fully connected layers of the neural network. Comparative analysis shows that there is no significant difference in performance between the Bag-Net and the Bag-Wrapper models. To increase the contribution of the relevant conformations during training of the model, the Bag-Net architecture was enhanced by the attention mechanism (Bag-AttentionNet). This, however, does not lead to a significant increase in the predictive performance of the model (Figure 5).

Overall, the analysis shows that the Instance-Wrapper algorithm largely outperforms all other studied MI algorithms. Other algorithms demonstrated comparable performance, despite the substantial differences in their architecture.

**Comparison of 2D and 3D Models.** There is an ongoing discussion about the preference of 2D and 3D descriptors in QSAR. An important step in building QSAR models with 3D descriptors concerns the selection of the bioactive conformation, which is hard to do reasonably without some additional information. An MI model is free from the problem of arbitrary selection of conformations. It considers all conformations and automatically selects the most relevant ones. We compared MI models with 2D models to estimate the importance of accounting for 3D information in bioactivity prediction and to assess contributions of particular conformations.

Six approaches were compared: three classical approaches based on 2D molecular descriptors, a 3D single-instance approach based on 3D pharmacophore descriptors calculated for the lowest energy conformations, and two 3D multi-instance approaches based on all generated conformations of each molecule represented by 3D pharmacophore descriptors. Among the MI approaches we chose the best performing Instance-Wrapper algorithm and the most advanced Bag-AttentionNet algorithm. For the sake of clarity, 33 “non-modellable” data sets for which none of the considered 2D and 3D models had $R_{	ext{test}}^2 > 0.4$ were excluded, and the analysis was performed based on the remaining 142 data sets.

Table 1 presents the mean $R_{	ext{test}}^2$ of models across the chosen 142 data sets. The 3D SI models built with one conformation per molecule demonstrated poor performance (mean $R_{	ext{test}}^2 = 0.024$) in comparison with the other models. The poor performance of 3D SI models can be explained by the ambiguous strategy when only one lowest energy conformation is considered. The lowest energy conformation might substantially differ from the actual bioactive conformation responsible for the observed bioactivity of the molecule. However, the performance of the 3D models drastically increases as soon as all available generated conformations are considered. Mean $R_{	ext{test}}^2$ values of 0.524 and 0.468 were

---

**Table 1. Performance Comparison of 2D and 3D Models**

<table>
<thead>
<tr>
<th>model</th>
<th>mean</th>
<th>median</th>
<th>top 1</th>
<th>top 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D/MI/Instance-Wrapper</td>
<td>0.524 ± 0.131</td>
<td>0.526</td>
<td>69</td>
<td>105</td>
</tr>
<tr>
<td>3D/MI/Bag-AttentionNet</td>
<td>0.468 ± 0.161</td>
<td>0.474</td>
<td>12</td>
<td>57</td>
</tr>
<tr>
<td>2D/MorganFP/Net</td>
<td>0.464 ± 0.199</td>
<td>0.502</td>
<td>39</td>
<td>66</td>
</tr>
<tr>
<td>2D/PhyChem/Net</td>
<td>0.450 ± 0.144</td>
<td>0.443</td>
<td>17</td>
<td>37</td>
</tr>
<tr>
<td>2D/PharmFP/Net</td>
<td>0.382 ± 0.216</td>
<td>0.404</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>3D/SI/Net</td>
<td>0.024 ± 0.372</td>
<td>0.089</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Table reports mean, standard deviations, and median of $R_{	ext{test}}^2$. Top 1 is the number of cases where the model was the best. Top 2 is the number of cases where the model was the first- or second-best one.
obtained, respectively, for the 3D/MI/Instance-Wraper and 3D/MI/Bag-AttentionNet models. The former even outperforms the 2D models built with Morgan fingerprints (mean $R^2_{\text{test}} = 0.464$). The 3D/MI/Instance-Wraper models displayed the highest $R^2_{\text{test}}$ in almost 49% of the cases (69 out of 142 data sets), and they were in the top 2 models for 105 data sets. The 2D/MorganFP/Net models were the best in 27% of the cases (39 out of 142 data sets). The other 2D models based on physicochemical or pharmacophore descriptors had poorer performance than those based on Morgan fingerprints.

The critical differences diagram of a pairwise comparison of the 2D and 3D models is shown in Figure 6. The 3D/MI/Instance-Wraper models have an average rank of 1.96 and outperform the 2D/MorganFP/Net model having an average rank of 2.77. The 3D/SI/Net model showed the worst performance across almost all data sets (average rank 5.72).

We calculated the number of wins and losses to perform a pairwise comparison of models (Table 2). Wins represent the number of tasks where the first model was higher than that of the second (model 1 vs model 2). Losses are counted as the number of data sets where the biological activity of the ligands is more accurately predicted by the second model. Ties are the number of data sets where the accuracy of both models is equal. Inconclusive is the number of data sets where $R^2_{\text{test}}$ of both models was less than 0.4.

Table 2. Pairwise Comparison of Models

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Wins</th>
<th>Losses</th>
<th>Ties</th>
<th>Inconclusive</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D/MI/Instance-Wraper vs 2D/MorganFP/Net</td>
<td>90</td>
<td>48</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>3D/MI/Bag-AttentionNet vs 2D/MorganFP/Net</td>
<td>50</td>
<td>72</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td>3D/SI/Net vs 2D/MorganFP/Net</td>
<td>3</td>
<td>99</td>
<td>0</td>
<td>73</td>
</tr>
<tr>
<td>3D/MI/Instance-Wraper vs 3D/SI/Net</td>
<td>122</td>
<td>2</td>
<td>0</td>
<td>51</td>
</tr>
<tr>
<td>3D/MI/Bag-AttentionNet vs 3D/SI/Net</td>
<td>97</td>
<td>4</td>
<td>0</td>
<td>74</td>
</tr>
</tbody>
</table>

“Wins are the number of data sets for which the accuracy of predicting the biological activity of ligands by the first model is higher than that of the second (model 1 vs model 2). Losses are counted as the number of data sets where the biological activity of the ligands is more accurately predicted by the second model. Ties are the number of data sets where the accuracy of both models is equal. Inconclusive is the number of data sets where $R^2_{\text{test}}$ of both models was less than 0.4.

number of tasks where the $R^2_{\text{test}}$ of the first model was higher than that of the second model and at least one model had $R^2_{\text{test}} > 0.4$. For example, 3D/MI/Instance-Wraper outperformed 3D/SI/Net in 122 out of 124 data sets (98%), and its $R^2_{\text{test}}$ was higher than that of 2D/MorganFP/Net in 90 out of 139 data sets (65%).

The 3D/MI/Instance-Wraper models outperformed the 3D SI models in almost all cases except for few data sets for which quite similar prediction accuracy between the two approaches was observed (see Figure 7a). The 3D/MI/Instance-Wraper models also outperformed 2D models in many cases. However, a large variability of model performances was observed for these two approaches. Most notably, in 38 out of 142 compared data sets the 3D/MI/Instance-Wraper models achieved $R^2_{\text{test}} > 0.4$, while the 2D models had $R^2_{\text{test}} < 0.4$. This means that using multiple conformations in the model building may significantly improve the model performance, and if the 2D models fail one may try to apply 3D MI-QSAR approaches.

We investigated which factors can distinguish cases where 3D SI models outperformed 2D conventional ones. We analyzed the distribution of the physicochemical characteristics of data sets where the 3D/MI/Instance-Wraper models outperform 2D models and vice versa. The data sets were divided into two groups. The first group consisted of 42 data sets, for which the 3D/MI/Instance-Wraper models were significantly ($\Delta R^2 > 0.1$) better than the 2D models. The second group included 18 data sets where the 2D models outperformed ($\Delta R^2 > 0.1$) the 3D/MI/Instance-Wraper models. We established that the smaller number of rotatable bonds is more favorable for the 3D/MI/Instance-Wraper models than for the 2D (Figure 8a). This may be caused by the poorer ability of the conformer generator to generate biologically relevant conformations for more flexible compounds. The 3D/MI/Instance-Wraper models were favorable in cases where the fraction of unique Murcko frameworks (the ratio of the number of unique scaffolds in the data set to the total number of molecules) was high (Figure 8b). This corresponds to data sets with higher scaffold diversity which are more difficult for the 2D models. Similar box plots were created for other characteristics of the data sets (see Supplementary Figure S3), but on average they cannot distinguish cases where the 3D/MI/Instance-Wraper models dominate.

## IDENTIFICATION OF BIOACTIVE CONFORMATIONS

The attention mechanism allows the 3D/MI/Bag-AttentionNet models to identify the most relevant conformations during the learning. The question arises of how accurately the attention mechanism recognizes the bioactive conformation? To answer this question, we chose the 3D/MI/Bag-AttentionNet models with $R^2_{\text{test}} > 0.4$. Then, the 3D structures of the ligands were extracted from the protein—ligand complexes retrieved from the PDB database. Since these data were sparse, four data sets having at least 10 test set molecules with available information about the bioactive conformation were chosen for the subsequent analysis. Experimental bioactive conformations were compared with the three top conformations that received the highest attention weights from the 3D/MI/Bag-AttentionNet. To measure the accuracy of identification of the bioactive conformations, we calculated the top 3 success rate as a proportion of compounds for which at least one of the three best conformations fits the experimental structure with RMSD < 2.0 Å.

To compare the accuracy of identification of relevant conformations with docking, we chose for each protein target a PDB complex with a binding site intersected with most of the binding sites of other complexes and used it for docking of the same test set compounds (CHEMBL2820-4Y8Y, CHEMBL3048-41MS, CHEMBL335-3EAX, and CHEMBL4802-4KCCQ). This was more fair than performing redocking to cognate receptor structures because in the case of machine learning we do not use information about the receptor conformation to select a relevant conformation. Docking was performed using AutoDock Vina. Three top-
scored poses were taken to calculate the top 3 statistics similarly as described above.

Since it was claimed that the RDKit conformer generator can reproduce bioactive conformations, we calculated two baseline statistics. The first one used three conformations with the lowest estimated energy. The second one corresponds to the top 3 metric value of three randomly chosen conformations for each molecule. We calculated the probability of choosing at least one conformation with the RMSD below 2 Å among the three randomly selected ones for each molecule and averaged these values across the test molecules.

The calculated random baseline statistics was relatively high (Figure 9). This indicates that the RDKit conformer generator substantially enriches the set of conformations with those which are close to the experimental ones. This also makes it challenging to improve the statistics. Selection of three conformations with the lowest energy performed comparable to random choice, and in the case of CHEMBL2820, the performance was even worse. The 3D/MI/Bag-AttentionNet models could improve the baseline accuracy in the identification of the bioactive conformations and perform comparably well or better than the random choice. The most remarkable improvement was for coagulation factor XI (CHEMBL2820). For two targets, brain and endothelial nitric-oxide synthases (CHEMBL3048 and CHEMBL4802, correspondingly), 3D/MI/Bag-AttentionNet performed comparably to the baseline. Protein-tyrosine phosphatase 1B (CHEMBL335) was the most difficult target for the identification of relevant conformations, and all approaches demonstrated low performance. This was caused by the fact that only part of those compounds binds to the protein; the remaining part was pretty flexible and exposed to a water
medium. Therefore, even docking could not identify true poses. In general, docking performed relatively poorly and slightly worse than the random baseline in the case of CHEMBL2820 and CHEMBL335. Examples of the lowest energy conformations and the conformations predicted by the 3D/MI/Bag-Attention model in comparison with the experimental ones retrieved from the PDB are shown in Figure 10.

In addition, we considered subsets of “challenging” compounds with a mean RMSD to the bioactive conformation greater than 2 Å. These subsets were enriched by very flexible compounds for which diverse sets of conformations were generated. As expected, the performance of key conformation identification for these compounds was lower (Figure 9), but 3D/MI/Bag-Attention had a performance comparable with or higher than the random baseline, supporting an intelligent selection of relevant conformations.

**CONCLUSION**

This study reports a large-scale comparison of single- and multi-instance machine learning algorithms for predicting the biological activity of chemical compounds. The molecules were represented either by the lowest energy conformation (single-instance) or by a set of generated conformations (multi-instances). The multi-instance learning algorithms reduce the problem of ambiguous selection of a putative bioactive conformation and simultaneously consider all available conformations in the model building.

The present study is the first comprehensive comparison of MI approaches with traditional QSAR based on 2D and 3D descriptors. The results demonstrate that multi-instance models generally outperform both single-instance 3D models and traditional QSAR models built on 2D Morgan fingerprints (mean $R^2_{\text{test}} = 0.524, 0.024,$ and 0.464, respectively). Surprisingly, on average, the application of 3D descriptors of the lowest energy conformation for QSAR modeling was only slightly better than the null model. Thus, the highest accuracy in the bioactivity predictions is achieved by the multi-instance algorithm since it considers the whole conformational space of an individual training object. This result demonstrates the importance of accounting for the dynamic nature of chemical objects for QSAR modeling.

Contrary to a previous finding,43 our study shows that 3D descriptors of molecules in combination with the MI learning approach can compete with traditional 2D QSAR. Notably, there were 38 data sets where the MI learning approach showed reasonable performance while the traditional 2D QSAR model failed. This means that the MI learning approach can be applied in cases where 2D QSAR modeling fails.

Last but not least, a multi-instance neural network with an attention mechanism can correctly identify a “bioactive” conformation close to the experimental structure of a ligand retrieved from the PDB. However, it should be noted that the performance of the multi-instance models depends on the conformer generator used. The RDKit conformer generator demonstrated a good ability to generate biologically relevant conformations that was confirmed by a relatively high random choice baseline estimate.

To facilitate the community being able to apply the MI learning approach for QSAR modeling, a set of MI learning algorithms based on different MI neural network architectures as well as wrappers used in the work are available at https://github.com/cimm-kzn/3D-MIL-QSAR.
Ministry of Science and Higher Education of the Russian Federation (agreement no. 14.587.21.0049, unique identifier RFMEFI8718X0049). Development of the neural nets architectures for the prediction of chemical object properties was supported by the Russian Science Foundation (agreement no. 19-73-10137).

REFERENCES

(18) Kuz’mín, V. E.; Artemenko, A. G.; Polischuk, P. G.; Muratov, E. N.; Hromov, A. I.; Liahovskyj, A. V.; Andronati, S. A.; Makar, S. Y.


