Analysis of Drugs Molecular Graphs

Performed by Alexander Glushko and Alexander Krasnov

Abstract

Machine learning has become a crucial tool in drug discovery and chemistry at large. There are lots of different molecular compounds that have drug effects on humans and our reasters and lows couldn't check and process all the illegal substances. Contemporary ML approaches can help us to deal with it. There are several well-known papers using molecular graph analysis to identify certain chemical and physical properties of molecules. In this paper we would like to test the use of the package MolGraph to analyze substances for drug properties from the open compounds library ChEMBL and compare with traditional models.

Introduction

Analyses of compounds in hardly developed areas nowadays. There are many papers published every year with different approaches and benchmarks. In this work we are mainly interested in drug detecting models but this technique can be used in many different problems. One of the main challenges in finding new drugs is identifying compounds that can interact with specific biological targets in the body. Molecular graphs can be used to model the structure of these targets and the compounds that interact with them, providing valuable insights into the finding of new drugs.
Base idea behind popular graph base approaches is QSAR/QSPR models can be used to predict a wide range of biological activities, such as enzyme inhibition, receptor binding, and toxicity. QSAR /QSPR models are created by analyzing the relationship between the molecular descriptors and the biological activity/property of a set of molecules with known activity/property. The models can then be used to predict the biological activity of new molecules with similar chemical structures. Several ML models can be used to analyze molecular structures based on QSAR/QSPR. Not so long ago, deep neural networks were used as basic models, but recent work “Exposing the Limitations of Molecular Machine Learning with Activity Cliffs”[1] shows that SVM, GBM, and RF can perform well too. We tried to make a similar comparison but with classification problems of drug detecting. We tested popular packages MolGraph and MoleculeNet in drug detecting tasks and compared them with traditional RandomForest, GradientBoosting, XGBoost and SVC models. Also we tried different embedding strategies.

Literature Review

A good review focusing on the problem of identifying narcotics has been done by Pat Walters[2]. This review is very recent and looks at various interesting articles from the past year. Of particular interest are articles on large libraries of chemical compounds and current attempts to use Graph Neural Networks. Also, a detailed review for a longer period is presented in the article "Deep learning in drug discovery: an integrative review and future

challenges". Based on what we have read, we can conclude that to develop a successful model it is necessary to solve 3 problems: finding a sufficiently large data set with a suitable format, choosing a way to represent molecular graphs, and choosing an algorithm[3].
Currently there are two main types of models you can choose from: traditional models and neural networks. Traditional models hardly rely on descriptor based embeddings, that hardly improve performance of the model. For this work we chose ECFP[4], which was computed with a length of 1024 bits and a radius of 2 bonds, and MACCS[5] keys, with a length of 166, were computed with default settings. ECFP has many more dimensions that hardly affect weak algorithms, but this approach catches the structure of molecular graphs based on allowed length and radius and takes into account molecular graphs specific. MACCS is a descriptor based approach, counting many different descriptors from molecular graphs. This method has fewer dimensions that can be really helpful.

As for Graph Neural Networks (GNNs), they are a class of neural networks designed to operate on graph-structured data. They have gained significant attention in recent years due to their ability to capture and model complex relationships and dependencies present in graph data. Here is a general analysis of Graph Neural Networks. One of the key ideas in GNNs is the aggregation of information from neighboring nodes. At each layer of the GNN, the node representations are updated by aggregating and integrating information from its neighboring nodes. The update of node representations is typically performed using a node update function. This function takes into account the aggregated messages from neighboring nodes, the current node's features, and potentially other information. It transforms the input into an updated representation that captures the local and global information from the graph.Depending on the task at hand, GNNs can generate node-level or graph-level outputs. For node-level tasks, each node can predict some property or label based on its updated representation. For graph-level tasks, the final output is obtained by aggregating the representations of all nodes in the graph, often using pooling or readout operations.Graph neural models are used in various fields, for example, to analyze abnormal brain activity [6].

Anticipated Methods and Datasets

Representation

Currently one of the most well supported ways to represent molecular structures is SMILES[7] (Simplified molecular-input line-entry system). This is string representation that is used by many chemical databases and has a lot of tools for parsing and visualizing. One of the most popular tools for working with SMILES strings is RDKit[8]. This is a collection of cheminformatics and machine-learning software written in C++ and Python that work well for not only parsing data but also generating descriptors and fingerprints for machine learning and manipulating with 2D and 3D molecular operations. There are also several other representations of molecular structures, for example Mol, this type of structure not so well readable by humans but hardly used in computing different embeddings.

Dataset

ChEMBL is a handmade database of bioactive molecules with drug-like properties. The European Bioinformatics Institute created the dataset, which contains over 1.9 million compounds and provides users with information on both the chemical properties and activity of the compounds, as well as their genomics. Due to the extensive range of features available in ChEMBL, it is considered one of the most comprehensive public datasets in this domain. Users can query the dataset using the web interface or ElasticSearch, which offers more flexibility in selecting the desired features for their dataset. The dataset[9] is open source and can be easily used. We decided to use this data because it represents a format which is supported by existing algorithms. Official Python client for accessing the API was used for receiving data from ChEMBL[10].

One of the main properties of compounds in ChEMBL is “Max Phase”. This is the phase of discovering this compound and proving its properties. There aren’t a lot of fully approved compounds, so we took as much as we could from drugs and mixed them with the same amount of ordinary compounds. Finally we have a dataset with 9386 elements. We store molecular representation with smile and structural properties of this compound.

Traditional models

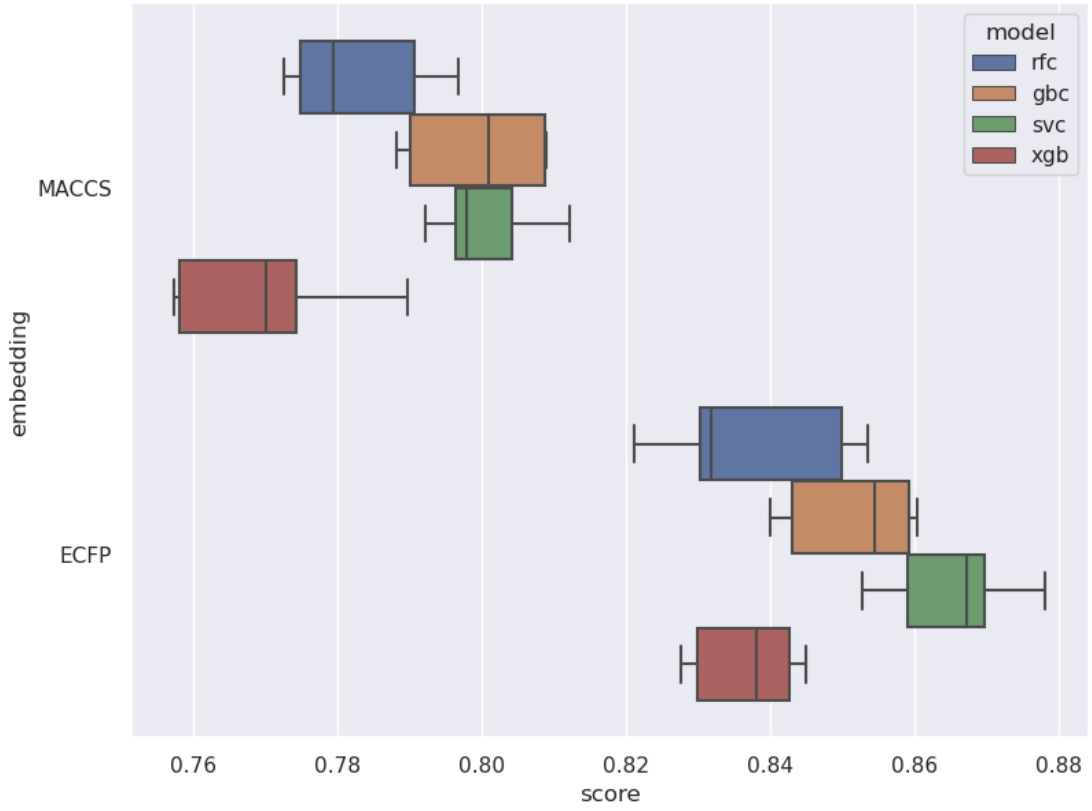

Fig. 1. Results for traditional models

For the traditional model we have used graph2vec, ECFP and MACCS embeddings and for models we choosedRandomForest, XGBoost, GradientBoosting and SVC. As we expected, ordinary graph2vec embedding from the networkx package with RandomForest showed

really bad performance, about 0.61 f1 score. Ordinary embeddings don't catch specific molecular properties, so we go further with ECFP, we choose length 1024 and radius 2. These settings worked quite well for us. We grid search parameters for every model and then count the result score with five fold cross validation. All models with ECFP show quite good performance, the best one was SVC with 0.87 average f1 score. Next approach was using MACCS keys. Models with this embedding work worth, best performance show GradientBoosting model with 0.8 average f1 score. You can see a graphical comparison in Fig. 1.

After MACCS we decided to try to add some structural properties to embeddings. We trained the same set of models and found out that performance became too good, 100% f1-score. When we looked to feature importance it showed that Inorganic Flag is the only property that takes effect in this kind of models. After plotting the confusion matrix (fig. 2) everything becomes clear. Drugs are mainly organic and we have quite a lot of non organic compounds, so our dataset is mainly an organic classification task.

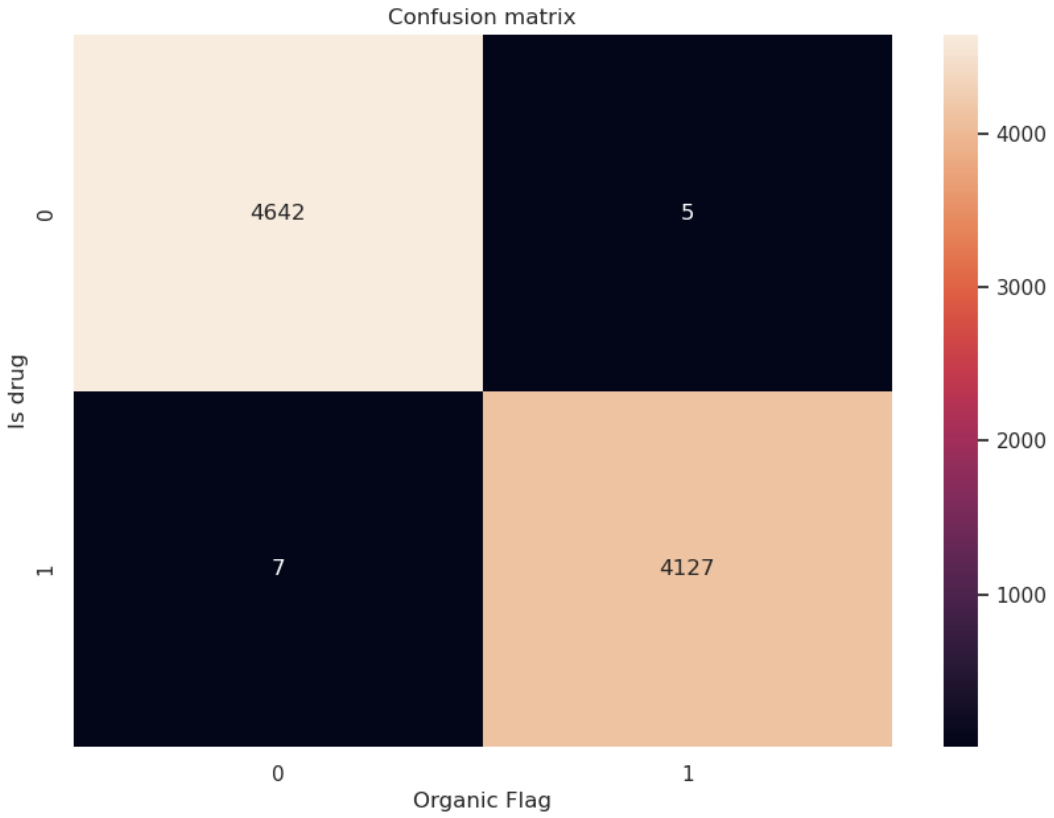

Fig. 2 Organic flag with drug flag confusion matrix

Graph neural models

There are many algorithms developed in this area, but we mainly focused on models presented in MolGraph[11] and Moleculenet[12]. These packages are based on popular libraries Tensor flow and Keras and can be easily added in popular ML pipelines. They are also open source and contribution friendly. This makes them the perfect choice for us.

As we discussed earlier there are several papers using GCM (Graph Convolutional Networks) as a preferred model for drug detecting problems. There are also several approaches for improving convolutional networks[13] with usage knowledge from chemical

domains such as drug solubility and toxicity. This can give a significant boost to model performance. This is why we choose GCM as our main model for tests. In addition we would like to compare the MPNN model (Message Passing Neural Network). This model is well known too and well presented in both packages that make it interesting for benchmarking and testing. There are also SVM and RF models which we are interested in, but without additional ECFPs[4] these models won't give expected performance.
GATConv, GMMConv, AFPConv and DMPNN were used for the analysis of molecules by applying it to graph representations of molecular structures. Molecules are represented as graphs, where atoms are nodes and chemical bonds are edges connecting the nodes. Convolutional layers stacked on top of each other and followed by readout operations and fully connected layers. By using convolutional layers within a graph neural network architecture, it is possible to leverage attention mechanisms to capture the dependencies and relationships between atoms in molecules. The layers were used from the molgraph python library[11]. Graph neural networks with various configurations were trained GAT (graph attention), AFP (attention features) and GMM (Gaussian mixture). Also DMPNN (directed message passing) was trained. Two different losses were used for models with convolutional layers. The lowest value of 0.66 f1-metric has model A, and the highest value among neural models of 0.77 f1-metric has a learning model using GAT convolutional layer and mae as a loss.

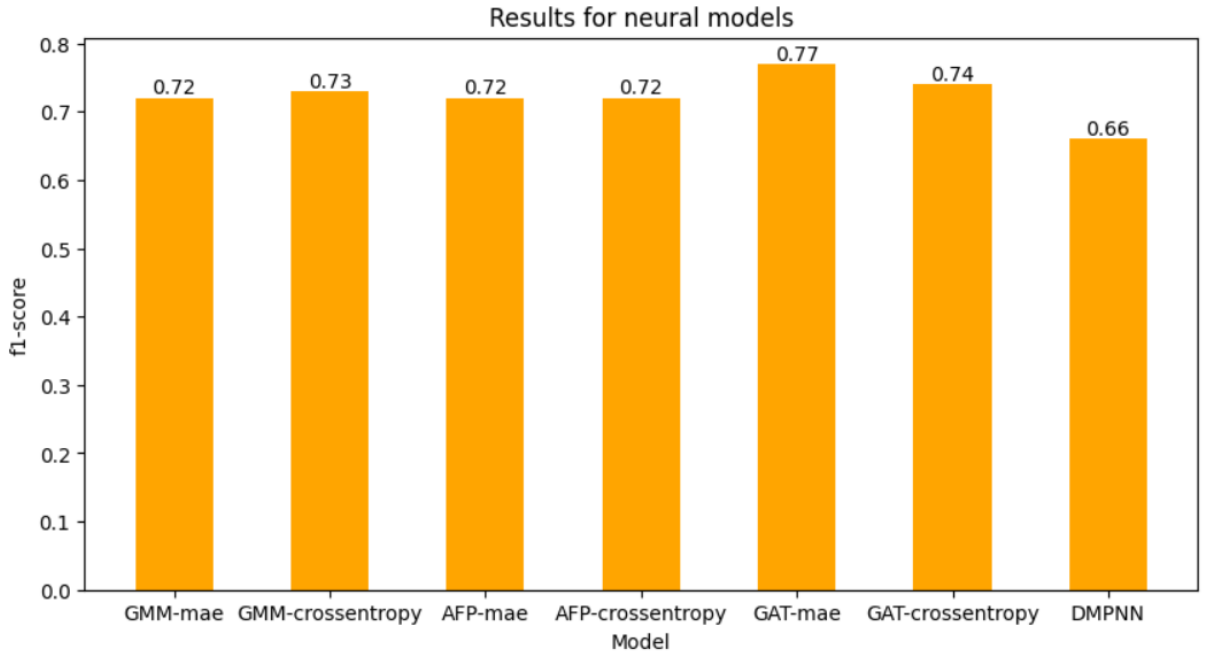

Fig. 3. Results for neural models

T-SNE and PCA plots

We have also tried to visualize our data with T-SNE and PCA algorithms. All embedding approaches showed quite bad results. Sets are very mixed and you can't really say something about them (fig. 4). But using only structural properties we got quite nice results. This can mean that high dimensionality of embedding data doesn't give the opportunity to have nice 2d graphics.

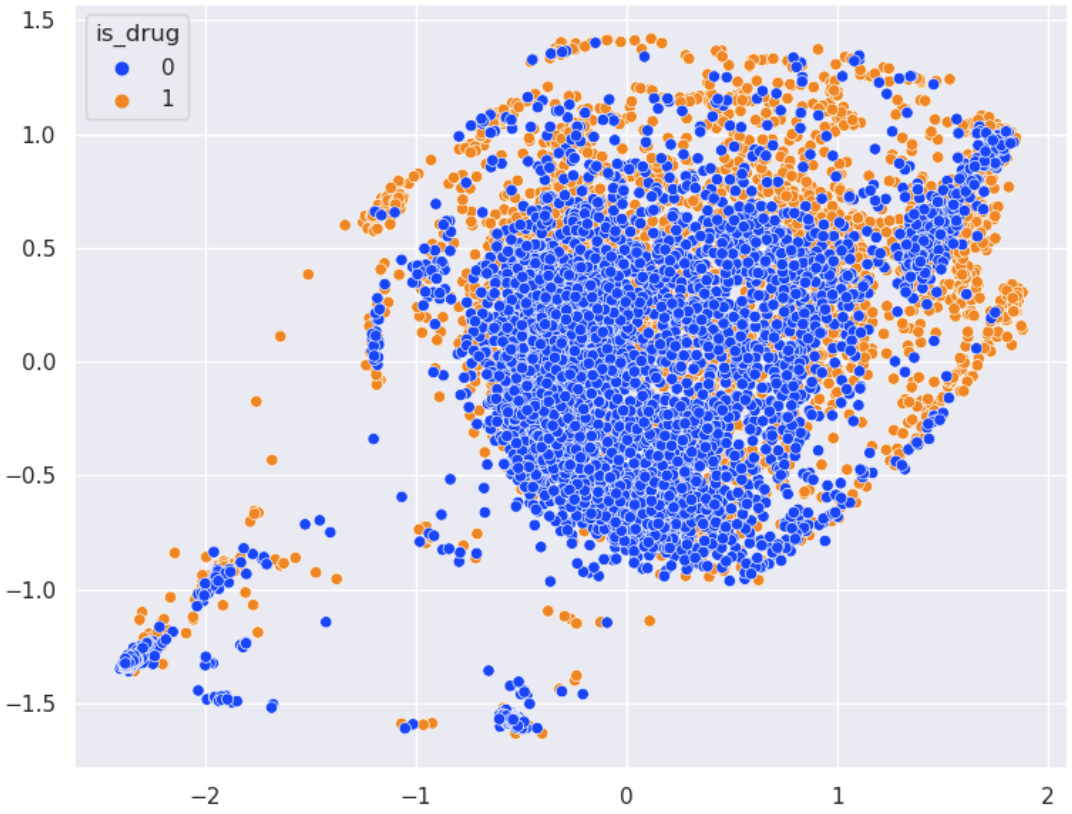

Fig. 4 Data dimensionality reduction using T-SNE on MACCS embedding

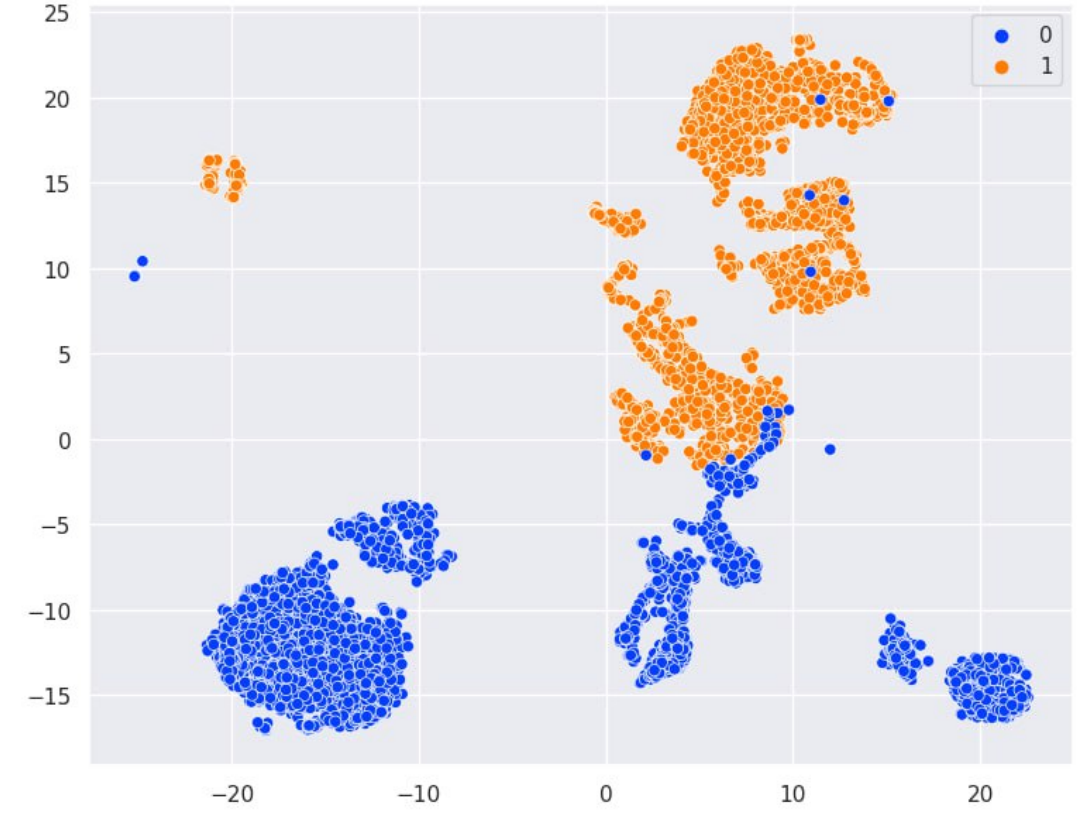

Fig. 5. Data dimensionality reduction using T-SNE on molecules structural properties

Conclusion

In our work we compared traditional machine learning models with ECFP and MACCS embeddings with neural network models. We have shown that in our drug classification task traditional models worked better. Best performance gives an SVC model with ECFP embedding 0.87. The other results were quite close to each other. More tests with a variety of datasets are needed for further conclusions, but we can already conclude that neural networks are not the only method worth using in the tasks of analysis of molecular graphs. All model results you can see in the table in Appendix.

Appendix

Table with models and f1-score values

Model	**f1-score**
GMM mae	0.72
GMM crossentropy	0.73
AFP mae	0.72
AFP crossentropy	0.72
GAT mae	0.77
GAT crossentropy	0.74
DMPNN	0.66
RandomForest MACCS	0.78
GradientBoosting MACCS	0.80
SVC MACCS	0.80
XGBoost MACCS	0.77
RandomForest ECFP	0.84
GradientBoosting ECFP	0.85
SVC ECFP	**0.87**
XGBoost ECFP	0.84

References

[1] ‘Exposing the Limitations of Molecular Machine Learning with Activity Cliffs | Journal of

Chemical Information and Modeling'. https://pubs.acs.org/doi/10.1021/acs.jcim.2c01073 (accessed May 01, 2023).
[2] 'AI in Drug Discovery 2022 - A Highly Opinionated Literature Review'. https://practicalcheminformatics.blogspot.com/2023/01/ai-in-drug-discovery-2022-highly.html (accessed Apr. 26, 2023).
[3] H. Askr, E. Elgeldawi, H. Aboul Ella, Y. A. M. M. Elshaier, M. M. Gomaa, and A. E. Hassanien, 'Deep learning in drug discovery: an integrative review and future challenges', *Artif. Intell. Rev.*, Nov. 2022, doi: 10.1007/s10462-022-10306-1.
[4] D. Rogers and M. Hahn, 'Extended-Connectivity Fingerprints', *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, May 2010, doi: 10.1021/ci100050t.
[5] 'Reoptimization of MDL Keys for Use in Drug Discovery | Journal of Chemical Information and Modeling'. https://pubs.acs.org/doi/abs/10.1021/ci010132r (accessed Jun. 21, 2023).
[6] 'Graph Convolutional Network with Attention Mechanism for Discovering the Brain's Abnormal Activity of Attention Deficit Hyperactivity Disorder | IEEE Conference Publication | IEEE Xplore'. https://ieeexplore.ieee.org/document/9979902 (accessed Jun. 21, 2023).
[7] J. Craig, 'OpenSMILES specification'. [Online]. Available: www.opensmiles.org
[8] 'RDKit'. RDKit, May 01, 2023. Accessed: May 01, 2023. [Online]. Available: https://github.com/rdkit/rdkit
[9] 'Index of /pub/databases/chembl/ChEMBLdb/latest'. https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/ (accessed May 01, 2023).
[10] 'ChEMBL webresource client'. The ChEMBL Group, Jun. 19, 2023. Accessed: Jun. 21, 2023. [Online]. Available: https://github.com/chembl/chembl_webresource_client
[11] A. Kensert, G. Desmet, and D. Cabooter, 'MolGraph: A Python package for the implementation of small molecular graphs and graph neural networks with TensorFlow and Keras', Sep. 25, 2022. http://arxiv.org/abs/2208.09944 (accessed Apr. 27, 2023).
[12] 'MoleculeNet'. https://moleculenet.org/ (accessed May 01, 2023).
[13] H. Xiao and X. Chen, 'Drug ADMET Prediction Method Based on Improved Graph Convolution Neural Network', in *2022 4th International Conference on Robotics and Computer Vision (ICRCV)*, Sep. 2022, pp. 266–271. doi: 10.1109/ICRCV55858.2022.9953254.