Artificial Intelligence in Biological Activity Prediction

João Correia(✉), Tiago Resende, Delora Baptista, and Miguel Rocha

CEB - Centre of Biological Engineering, University of Minho,
Campus of Gualtar, Braga, Portugal
jfscorreia95@gmail.com, tiagofcresende@gmail.com, dlr.baptista@gmail.com,
mrocha@di.uminho.pt

Abstract. Artificial intelligence has become an indispensable resource in chemoinformatics. Numerous machine learning algorithms for activity prediction recently emerged, becoming an indispensable approach to mine chemical information from large compound datasets. These approaches enable the automation of compound discovery to find biologically active molecules with important properties. Here, we present a review of some of the main machine learning studies in biological activity prediction of compounds, in particular for sweetness prediction. We discuss some of the most used compound featurization techniques and the major databases of chemical compounds relevant to these tasks.

Keywords: Machine learning · Deep learning · Biological activity prediction · Sweetness prediction · Compound featurization

1 Introduction

For centuries, humans have been manually searching and documenting different compounds, assessing their interaction with biological systems to find suitable products that solve problems and enhance quality of life. Despite the broad amount of data collected on compounds capable of curing illnesses, fighting infections or satisfying our food sensory system, the search for compounds with improved biological capabilities is still in high demand. With the modernization of the pharmaceutical and food industries, there is a growing need for more sustainable compounds, with improved biological activities. By taking advantage of the enormous quantity of categorized data on compound biological activity existent, and still being generated, new approaches using artificial intelligence (AI) are continuously being developed. With datasets getting larger and more detailed and algorithms increasing their scope and accuracy, new tools for predicting biological activity arise, accelerating the generation of new products.

F. Fdez-Riverola et al. (Eds.): PACBB 2019, AISC 1005, pp. 164–172, 2020.
https://doi.org/10.1007/978-3-030-23873-5_20

2 Artificial Intelligence in Biological Activity Prediction

Machine learning (ML) is a field of AI where systems learn from data, identify patterns and make decisions without being explicitly programmed [1]. Although ML algorithms were created in the 1950s [2], ML only started to thrive in the 1990s and is becoming the most popular sub-field of AI. ML techniques are classified as supervised or unsupervised. In the former, given input-output pairs, a function to map the input to the output is learned so the model can predict future cases. In the latter, patterns are learned directly from unlabeled data. For biological activity prediction, it is common to use supervised methods.

In linear regression (LR) and logistic regression (LgR), linear relationships between independent and dependent variables are learned. LgR is used for linear classification when the dependent variable is categorical. Naive Bayess (NBs) is a probabilistic classification algorithm based on the Bayes theorem and the assumption of feature independence. Random forests (RFs) are ensembles of decision trees (DTs), tree-like models of decision rules where each node represents a feature, each branch represents a decision and each leaf an outcome. RFs apply bagging to generate distinct training sets and create different models, and predictions are obtained by majority voting. The objective of support-vector machines (SVMs) is to map the data into a high-dimensional space by identifying a lower dimensional hyperplane that separates the data using nonlinear kernels. K-nearest neighborss (KNNs) is an instance-based algorithm where data is classified by its similarity with its k-nearest neighbors. Partial least squares (PLS) regression is mostly used to predict a set of dependent variables from a large set of independent variables. PLS decomposes the original set of variables into a set of components that explain the most covariance between the independent and dependent variables, and uses these components to predict the outputs. Neural networks (NNs) are biologically-inspired algorithms designed to automatically recognize patterns from input labeled data in order to be able to predict the output of unlabeled data according to similarities with the example inputs.

Due to high accuracy and cost-effectiveness, ML is extensively used in many fields including chemoinformatics. Recent algorithmic advances, as well as the development of databases for the storage of molecule structures and their properties, accelerated the pace at which the field has evolved. Researchers have used combinations and different approaches of traditional ML, as well as complex deep learning (DL) architectures. A common approach is the use of these models for the optimization of quantitative structure-activity relationship (QSAR) models to improve the biological activity prediction of multiple compounds.

Biological activity prediction of compounds is one of the main research areas in chemoinformatics [3]. The application of AI for this type of task is of critical importance for the identification of compounds with desired properties. The objective is to select a subset of compounds from all the compounds under consideration that have a higher probability of being bioactive when compared to a random sample. One important aspect for the success of ML in property prediction is the access to large datasets. Multiple large datasets from public-domain repositories are available and suited for activity prediction; such is the

Table 1. A selection of recent studies that use AI for biological activity prediction

ML methods	Study description
SVMs, KNN, RFs, NB, DNNs	Comparison of DL methods on a large-scale drug discovery dataset and other ML and target prediction methods [10]
RF, KNN, NB, DNN	Predicting kinase activities for around 200 different kinases using multiple ML methods [11]
NB, SVMs, LgR, RFs, DNN	Different ML methods were compared using a standardized dataset from ChEMBL [13] and standardized metrics [14]
NB, LgR, DTs, RFs, SVMs, DNNs	Comparison between DNNs and other ML algorithms for diverse endpoints (bioactivity, solubility and ADME properties) [15]
Multitask DNNs	Use of multitask DNNs as an improvement over single task learning [16]
DNNs, SVMs, RFs, NB, KNN	Shows that, when optimized, DNNs are capable of outperforming shallow methods across diverse activity classes [17]
DNNs, SVMs and RFs	Results from the Tox21 competition. DNNs show good predictivity on 12 different toxic endpoints [18]
Multitask DNNs	Multitask learning provided benefits over single task models. Smaller datasets tend to benefit more than larger datasets [19]
Multitask DNNs	Performance analysis of multitask DNNs (DeepChem implementation) and related DL models on pharmaceutical datasets [20]
DNNs, RFs, DTs	Comparison between multitask DNNs and alternative ML methods. Multitask DNNs outperformed alternative methods [21]
DNNs, RFs	Performance comparison between DNNs and RFs for QSARs using different datasets and descriptors [12]
DNNs	DL models to predict drug-induced liver injury [22]
Multitask DNNs	Multitask vs Single Task learning. Multitask DNNs showed better performance [23]
DNNs, SVMs, LgR, KNN, etc	Comparison of DL performance against multiple ML methods using data from ChEMBL [24]
NB and RFs	Comparison between NB and RFs to make accurate ADME-related activities predictions on 18 large QSAR datasets [25]
Shallow NNs	Prediction of biological activities of structurally diverse ligands using 3 types of fingerprints (ECFP6, FP2 & MACCS) [26]
Bayesian QSAR, PLS	Bayesian QSAR combines activities across the kinase family to predict affinity, selectivity, and cellular activity [27]

case of DrugBank [4], PubChem BioAssay [5], ChEBI [6], MoleculeNet datasets [7], ChemSpider [8] and T3DB: the toxic exposome database [9].

ML techniques, including SVMs, RFs and deep neural networks (DNNs) have been used to discover compounds with desired biological activities. Table 1 summarizes some of the most relevant studies in biological activity prediction using AI. In general, DNNs exhibit better results than classic ML methods [10,11],

where RF-based models are the most used and the ones showing better results, even outperforming DNNs in some situations [12].

2.1 Compound Featurization

Information for biological activity prediction comes primarily from the chemical structure of the compound. There has been a lot of research on how to transform molecules into a form suited for ML algorithms so that the model can learn and generalize the properties shared among the molecules. Some of the most useful molecular featurization methods include line notations, fingerprints, weave, graph convolutions and (NLP)-inspired embeddings.

Line Notations. Line notations express the 2D structure of compounds. These approaches represent structures and chemical properties such as atoms, bonds, aromaticity, chiral and isotopic information of compounds as compact strings of characters [28]. The most common line notations are SMILES [29] and InChI [30]. Most chemical databases like PubChem [5], ChEMBL [13] and DrugBank [4] provide line notations for the recorded compounds.

Fingerprints. Fingerprints are the most widely used molecular representations in chemoinformatics. They consist of binary arrays, where each dimension represents the presence or absence of a particular substructure or property. Fingerprints are used to encode multiple characteristics, including atomic attributes, atomic environments, bond properties and bond positions which enables it to be applied to tasks such as activity prediction. Extended-Connectivity Fingerprints (ECFPs), Functional-Class Fingerprints (FCFPs), and the 166-bit Molecular Access System (MACCS) are typical fingerprint-based featurization approaches.

Graph Convolutions. In this DL-based approach proposed by Duvenaud et al. [31], the chemical structure of the molecule is initially represented as a graph with atoms as nodes and bonds as edges, encoding the connectivity between atoms and each atom's local chemical environment. Then, different neighbor levels of the molecule graph representation are fed into a single layer convolutional NN to generate fixed-length vectors. The resulting vectors are transformed through a pooling-like operation using the softmax function, and then they are summed to form the final molecular-level vector representations.

Weave. The weave featurization method is very similar to graph convolutions. It also encodes both the connectivity between atoms in a molecule and each atom's local chemical environment, but connectivity uses more detailed pair features instead of information for the neighbor's list. It also encodes both the connectivity between atoms in a molecule and each atom's local chemical environment, but uses more detailed pair features instead of information for the neighbors list. Weave modules combine and transform the atom-level and pair-level features by applying specific convolution operators [32].

NLP-Inspired Embeddings. Deep learning-based NLP techniques can be directly applied to SMILES strings to generate continuous feature vectors instead of learning from molecular graphs. The Seq2seq fingerprint [33] translates molecules represented as SMILES strings into continuous embeddings using a model based on the sequence to sequence [34] machine translation model. Mol2vec [35] is a method inspired by the Word2vec [36] word embedding algorithm that learns continuous embeddings of compound substructures.

2.2 Sweetness Prediction

Sweetness prediction is a particular application of biological activity prediction, very important for many disciplines, especially food chemistry. As sugars and saccharides are widely used in the food industry, their overconsumption can severely affect human health, leading to serious diseases, such as obesity, diabetes and cardiovascular diseases. It is, thus, of extreme importance to identify low-calorie sweeteners present in natural or chemically synthesized compounds, avoiding, this way, associated health risks while preserving the sweetness perception.

The high cost associated with compound sweetness determination in the laboratory remains a barrier, justifying the necessity to build computational models capable of learning the relationship between sweetness and the structure of known sweeteners. Therefore, these models would facilitate the identification and design of new sweeteners with different degrees of sweetness. Moreover, existing sweeteners have been the subject of controversies regarding health and food safety [37]. In this aspect, computational methods for biological activity prediction can offer additional value by combining sweetness prediction with other tasks such as toxicity and bitterness prediction, removing in this way compounds with undesirable properties.

In recent years, multiple ML based models to predict compound sweetness were developed. In 2011, Yang et al. [38], developed three quantitative models (linear regression, neural networks (ANN), SVM) for the prediction of the sweetness of 103 compounds. Zhong et al. [39], in 2013, developed two quantitative models (linear regression and SVM) to predict the sweetness of 320 compounds. In 2016 and 2017, Rojas et al. [40,41] used KNN to discriminate sweet from non-sweet molecules. In the same year, Chéron et al. [42] used RF to predict either sweetness, bitterness and toxicity properties. In 2018, Goel et al. [43] developed QSAR models based on Genetic Function Approximation and ANNs analysis to predict the sweetness of molecules. A RF-based binary classifier to predict the bitterness and sweetness of chemical compounds was implemented by Banerjee et al. [44]. Ojha et al. [45] proposed 13 new sweet molecules using a quantitative structure-property relationship model and PLS regression analysis. In 2019, Zheng et al. [46] implemented multiple ML methods (KNN, SVM, Gradient Boosting Machine, RF, and DNN) for the prediction of sweeteners and their corresponding relative sweetness. Comparing the results obtained in the above mentioned studies is not completely feasible, because different datasets, number and type of descriptors and validation methods were used. However, a simple comparison shows that nonlinear methods such as RFs, SVMs, PLSs and

Table 2. Available databases containing data on sweeteners/non-sweeteners.

Database	Description
SweetenersDB [42] (http://chemosim.unice.fr/SweetenersDB/)	316 compounds belonging to 17 chemical families with known sweetness values
SuperSweet [47] (http://bioinformatics.charite.de/sweet/)	More than 15,000 natural and artificial sweeteners. Information on origin, sweetness class, predicted toxicity, molecular targets, etc.
FooDB (http://foodb.ca/)	The largest and most comprehensive database on food constituents
BitterDB [48] (http://bitterdb.agri.huji.ac.il/dbbitter.php)	Information on over 1,000 bitter-tasting natural & synthetic compounds
FlavorDB [49] (https://cosylab.iiitd.edu.in/flavordb/)	Contains 25,595 flavor molecules (618 sweet-tasting, 253 bitter-tasting)
Super natural II [50] (http://bioinf-applied.charite.de/supernatural_new/index.php)	Database comprising 325,508 natural compounds. Includes information about 2D structures, physicochemical properties and predicted toxicity

ANNs exhibit slightly better results. These methods, in general, can more easily capture the sweetness chemical space and therefore the structural diversity of known sweeteners, generating better results. Nonetheless, more accurate models are still in high demand. The use of DNNs models and taking into account the complex interactions between different sweeteners and respective receptors can further improve the results in the field.

With the generation of vast amounts of data from experimental and computational screening experiments, the need for structured databases to store and publish the generated data in a well-organized way is increasing. As a result, several compound databases that store thousands of molecules and respective chemical attributes, molecular descriptors, activity measurements and other information are available through the web. In particular, databases containing data on sweet/bitter molecules are starting to become more common. Table 2 describes the main databases containing sweet/non-sweet compounds.

3 Concluding Remarks

Here, we provide a review of literature related to AI algorithms used for biological activity prediction and in particular for sweetness prediction. Over the last decades, ML witnessed rapid development, and multiple methods have been successfully applied in chemoinformatics. Both shallow and DL methods have been widely used in this task and they have an important role in its future.

With the increase in the complexity and the size of the available datasets, DL models seem to frequently outperform traditional shallow ML algorithms. It is also common to benefit from multitask learning, as it has been shown that the

prediction of related properties seems to be beneficial to the predictive performance of the models. The use of AI in chemoinformatics strongly benefits from open source implementations of different ML models and from the availability of extensive datasets allowing the implementation of fine-tuned complex NNs. With the progress of AI in chemoinformatics, an increase in the use of these approaches to automate compound discovery is expected.

With this review, we can conclude that improved methods are still in high demand. Combining state-of-the-art deep learning models with different data types and with approaches from different fields will be crucial for the discovery of added-value compounds. Following this research line, we are implementing in our group methods to improve the identification and generation of new sweeteners that can be produced using only biologically feasible reactions, replacing the chemical synthesis currently used.

Acknowledgments. This study was supported by the European Commission through project *SHIKIFACTORY100 - Modular cell factories for the production of 100 compounds from the shikimate pathway* (Reference 814408), and by the Portuguese FCT under the scope of the strategic funding of UID/BIO/04469/2019 unit and BioTec-Norte operation (NORTE-01-0145-FEDER-000004) funded by the European Regional Development Fund under the scope of Norte2020.

References

1. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge [u.a.] (2013)
2. Samuel, A.L.: Some studies in machine learning using the game of checkers. IBM J. Res. Dev. **3**(3), 210–229 (1959)
3. Toccaceli, P., et al.: Conformal prediction of biological activity of chemical compounds. Ann. Math. Artif. Intell. **81**(1–2), 105–123 (2017)
4. Wishart, D.S., et al.: DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. **46**(D1), D1074–D1082 (2017)
5. Kim, S., et al.: PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. **47**(D1), D1102–D1109 (2018)
6. Hastings, J., et al.: ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. **44**(D1), D1214–D1219 (2015)
7. Wu, Z., et al.: MoleculeNet: a benchmark for molecular machine learning. Chem. Sci. **9**(2), 513–530 (2018)
8. Pence, H.E., Williams, A.: ChemSpider: an online chemical information resource. J. Chem. Educ. **87**(11), 1123–1124 (2010)
9. Wishart, D., et al.: T3DB: the toxic exposome database. Nucleic Acids Res. **43**(D1), D928–D934 (2014)
10. Mayr, A., et al.: Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem. Sci. **9**(24), 5441–5451 (2018)
11. Merget, B., et al.: Profiling prediction of kinase inhibitors: toward the virtual assay. J. Med. Chem. **60**(1), 474–485 (2016)
12. Ma, J., et al.: Deep neural nets as a method for quantitative structure-activity relationships. J. Chem. Inf. Model. **55**(2), 263–274 (2015)

13. Gaulton, A., et al.: The ChEMBL database in 2017. Nucleic Acids Res. **45**(D1), D945–D954 (2016)
14. Lenselink, E.B., et al.: Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. J. Cheminformatics **9**(1), 45 (2017)
15. Korotcov, A., et al.: Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. Mol. Pharm. **14**(12), 4462–4475 (2017)
16. Xu, Y., et al.: Demystifying multitask deep neural networks for quantitative structure-activity relationships. J. Chem. Inf. Model. **57**(10), 2490–2504 (2017)
17. Koutsoukas, A., et al.: Deep-learning: investigating deep neural networks hyperparameters and comparison of performance to shallow methods for modeling bioactivity data. J. Cheminformatics **9**(1), 42 (2017)
18. Mayr, A., et al.: DeepTox: toxicity prediction using deep learning. Front. Environ. Sci. **3**, 80 (2016)
19. Kearnes, S., et al.: Modeling industrial ADMET data with multitask networks, June 2016
20. Ramsundar, B., et al.: Is multitask deep learning practical for pharma? J. Chem. Inf. Model. **57**(8), 2068–2076 (2017)
21. Dahl, G., Jaitly, N., Salakhutdinov, R.: Multi-task neural networks for QSAR predictions. CoRR arXiv:1406.1231v1 (2014)
22. Xu, Y., et al.: Deep learning for drug-induced liver injury. J. Chem. Inf. Model. **55**(10), 2085–2093 (2015)
23. Ramsundar, B., et al.: Massively multitask networks for drug discovery. CoRR arXiv:1502.02072 (2015)
24. Unterthiner, T., et al.: Deep learning as an opportunity in virtual screening, January 2014
25. Chen, B., et al.: Comparison of random forest and pipeline pilot naïve bayes in prospective QSAR predictions. J. Chem. Inf. Model. **52**(3), 792–803 (2012)
26. Myint, K.Z., et al.: Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. Mol. Pharm. **9**(10), 2912–2923 (2012)
27. Martin, E., et al.: Profile-QSAR: a novel meta-QSAR method that combines activities across the kinase family to accurately predict affinity, selectivity, and cellular activity. J. Chem. Inf. Model. **51**(8), 1942–1956 (2011)
28. O'Boyle, N.M.: Towards a universal SMILES representation - a standard method to generate canonical SMILES based on the InChI. J. Cheminformatics **4**(1), 22 (2012)
29. Weininger, D.: SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. J. Chem. Inf. Model. **28**(1), 31–36 (1988)
30. Heller, S.R., et al.: InChI, the IUPAC international chemical identifier. J. Cheminformatics **7**(1), 23 (2015)
31. Duvenaud, D.K., et al.: Convolutional networks on graphs for learning molecular fingerprints. CoRR arXiv:1509.09292 (2015)
32. Kearnes, S., et al.: Molecular graph convolutions: moving beyond fingerprints. J. Comput.-Aided Mol. Des. **30**(8), 595–608 (2016)
33. Xu, Z., et al.: Seq2seq fingerprint. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM-BCB 2017, pp. 285–294. ACM Press, New York (2017)
34. Sutskever, I., et al.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)

35. Jaeger, S., et al.: Mol2vec: unsupervised machine learning approach with chemical intuition. J. Chem. Inf. Model. **58**(1), 27–35 (2018)
36. Mikolov, T., et al.: Efficient estimation of word representations in vector space, January 2013
37. Whitehouse, C.R., et al.: The potential toxicity of artificial sweeteners. AAOHN J. **56**(6), 251–259 (2008)
38. Yang, X., et al.: In-silico prediction of sweetness of sugars and sweeteners. Food Chem. **128**(3), 653–658 (2011)
39. Zhong, M., et al.: Prediction of sweetness by multilinear regression analysis and support vector machine. J. Food Sci. **78**(9), S1445–S1450 (2013)
40. Rojas, C., et al.: A new QSPR study on relative sweetness. Int. J. Quant. Struct.-Prop. Relat. **1**(1), 78–93 (2016)
41. Rojas, C., et al.: A QSTR-based expert system to predict sweetness of molecules. Front. Chem. **5**, 53 (2017)
42. Chéron, J.B., et al.: Sweetness prediction of natural compounds. Food Chem. **221**, 1421–1425 (2017)
43. Goel, A., et al.: In-silico prediction of sweetness using structure-activity relationship models. Food Chem. **253**, 127–131 (2018)
44. Banerjee, P., Preissner, R.: BitterSweetForest: a random forest based binary classifier to predict bitterness and sweetness of chemical compounds. Front. Chem. **6**, 93 (2018)
45. Ojha, P.K., Roy, K.: Development of a robust and validated 2D-QSPR model for sweetness potency of diverse functional organic molecules. Food Chem. Toxicol. **112**, 551–562 (2018)
46. Zheng, S., et al.: e-sweet: a machine-learning based platform for the prediction of sweetener and its relative sweetness. Front. Chem. **7**, 35 (2019)
47. Ahmed, J., et al.: SuperSweet-a resource on natural and artificial sweetening agents. Nucleic Acids Res. **39**(Database), D377–D382 (2010)
48. Dagan-Wiener, A., et al.: Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. Sci. Rep. **7**(1) (2017)
49. Garg, N., et al.: FlavorDB: a database of flavor molecules. Nucleic Acids Res. **46**(D1), D1210–D1216 (2017)
50. Banerjee, P., et al.: Super natural II–a database of natural products. Nucleic Acids Res. **43**(D1), D935–D939 (2014)