PACBB 2019

Artificial Intelligence in Biological Activity Prediction

João Correia*, Delora Baptista, Tiago Resende, and Miguel Rocha

*jfscorreia95@gmail.com

University of Minho
School of Engineering
CENTRE OF BIOLOGICAL ENGINEERING

About Me

João Correia

Research Fellow at CEB - Centre of Biological Engineering - University of Minho

University of Minho
Master's degree, Bioinformatics
2016 – 2018

University of Trás-os-Montes e Alto Douro
Bachelor's degree, Bioengineering
2013 – 2016

University of Minho
School of Engineering
CENTRE OF BIOLOGICAL ENGINEERING

Introduction

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

- In this review, we discuss some of the most relevant machine learning studies for biological activity prediction and in particular for sweetness prediction.

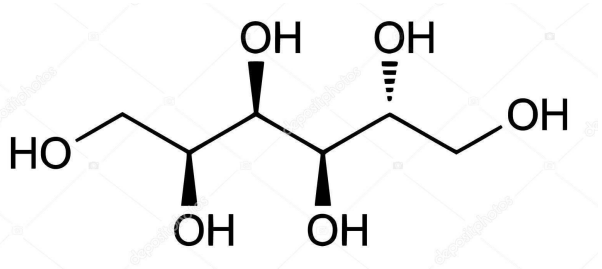
- We also address multiple compound featurization techniques and the major databases of chemical compounds.

University of Minho
School of Engineering
CENTRE OF BIOLOGICAL ENGINEERING

AI in Biological Activity Prediction

- 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.

- Due to high accuracy and cost-effectiveness, AI is extensively used in many fields including chemoinformatics.



University of Minho
School of Engineering
CENTRE OF BIOLOGICAL ENGINEERING

Studies

Large-scale comparison of machine learning methods for drug target prediction on ChEMBL[‡]

Andreas Mayr,[a,‡] Günter Klambauer,[a,‡] Thomas Unterthiner,[a,‡] Marvin Steijaert,[b] Jörg K. Wegner,[c] Hugo Ceulemans,[c] Djork-Arné Clevert,[d] and Sepp Hochreiter[a]

Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Datasets

Alexandru Korotcov,[†] Valery Tkachenko,[†*] Daniel P Russo,[‡§] and Sean Ekins[‡*]

Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions.

Myint KZ[1], Wang L, Tong Q, Xie XQ.

Is Multitask Deep Learning Practical for Pharma?

Ramsundar B[1], Liu B[2], Wu Z[2], Verras A[3], Tudor M[4], Sheridan RP[3], Pande V[2].

Databases

- One important aspect for the success of AI is the access to large and structured datasets. Multiple large chemical datasets from public domain repositories are available and suited for activity prediction:

DRUGBANK

ChEBI

PubChem

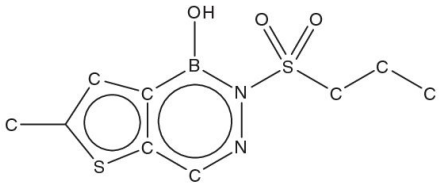
ChemSpider
Search and Share Chemistry

University of Minho
School of Engineering
CENTRE OF BIOLOGICAL ENGINEERING

Compound Featurization

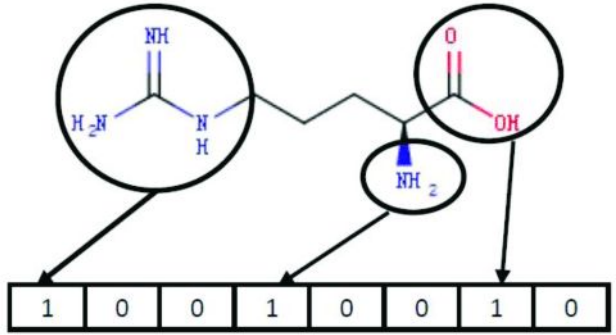
- Information for biological activity prediction comes from the chemical structure of the compounds. There has been a lot of research on how to transform molecules into forms suited for AI.

Line Notations



SMILES	*InChI*
CCCS(=O)(=O)[N]1=NC=c2sc(C)cc2=B1O	InChI=1S/C9H13BN2O3S2/c1-3-4-17(14,15)12-10(13)8-5-7(2)16-9(8)6-11-12/h5-6,13H,3-4H2,1-2H3

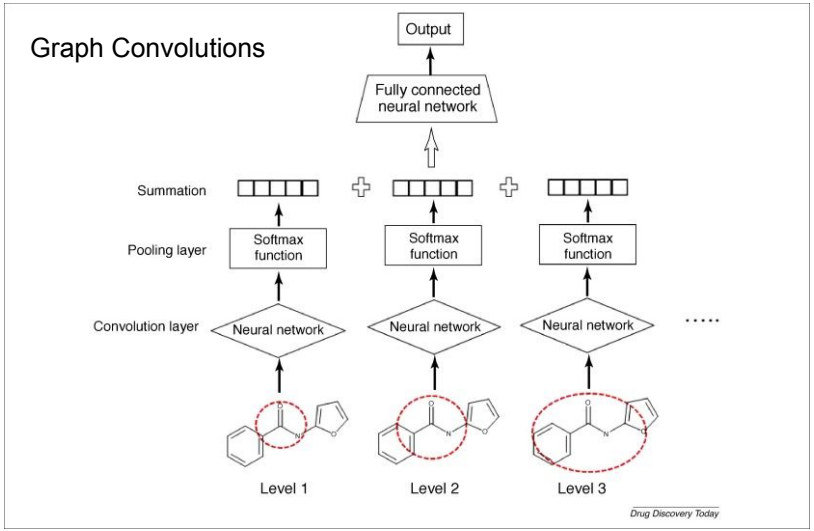
Fingerprints



University of Minho
School of Engineering
CENTRE OF BIOLOGICAL ENGINEERING

Compound Featurization

- Graph Convolutions
- Weave
- NLP-inspired Embeddings



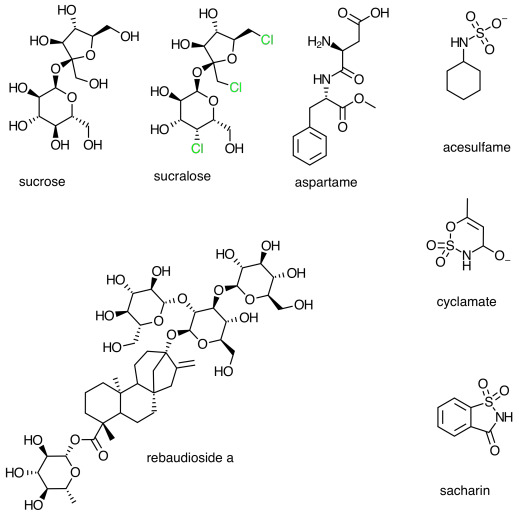
Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039

University of Minho
School of Engineering
CENTRE OF BIOLOGICAL ENGINEERING

Sweetness Prediction

- Sugars and saccharides are widely used in the food industry and their overconsumption can severely affect the human health, leading to serious diseases.

- The high cost associated with compound sweetness determination in the laboratory remains a barrier, justifying the necessity to build computational models.



University of Minho
School of Engineering
CENTRE OF BIOLOGICAL ENGINEERING

Studies

Prediction of sweetness by multilinear regression analysis and support vector machine.

Zhong M[1], Chong Y, Nie X, Yan A, Yuan Q.

BitterSweetForest: A Random Forest Based Binary Classifier to Predict Bitterness and Sweetness of Chemical Compounds.

Banerjee P[1], Preissner R[1].

e-Sweet: A Machine-Learning Based Platform for the Prediction of Sweetener and Its Relative Sweetness

Suqing Zheng,[1,2,*] Wenping Chang,[1] Wenxin Xu,[1] Yong Xu,[3] and Fu Lin[1,*]

Databases

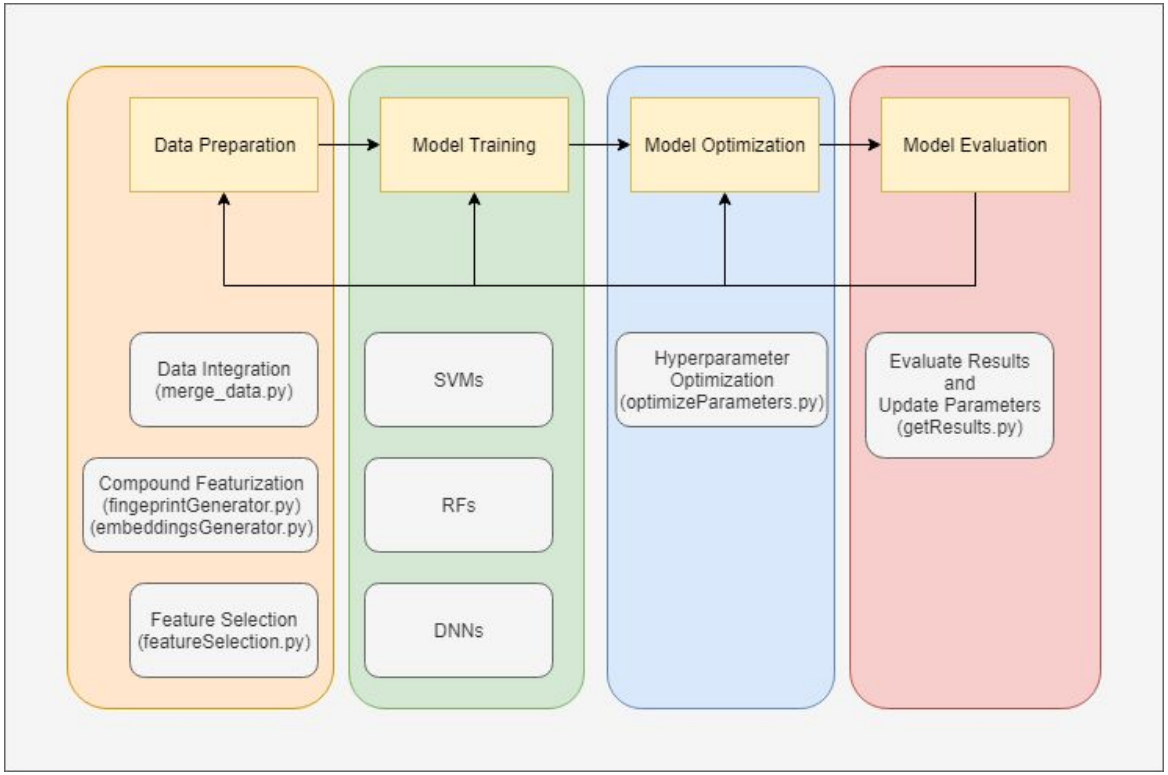
- With the generation of vast amounts of data from experimental and computational screening experiments, the need for structured databases to store and publish the generated data in a well-organized way is increasing.

- Databases containing data on sweet/non-sweet molecules are becoming more common.

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

University of Minho
School of Engineering
CENTRE OF BIOLOGICAL ENGINEERING

Following this research line...



Conclusions

- Over the last decades, ML witnessed rapid development, and multiple methods have been successfully applied in chemoinformatics.

- With the expected increase in the complexity and size of the available datasets it is expected that the use of DL could improve the results in the field, especially for sweetness prediction.

- 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.

Acknowledgements

Miguel Rocha, Delora Baptista, Tiago Resende
Biosystems Group

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 BioTecNorte operation (NORTE-01-0145-FEDER-000004) funded by the European Regional Development Fund under the scope of Norte2020.

Centre of Biological Engineering, Universidade do Minho
Campus de Gualtar 4710-057 Braga
ceb@ceb.uminho.pt
www.ceb.uminho.pt