Development of Deep Learning approaches to predict relationships between chemical structures and sweetness

João Capela
Centre of Biological Engineering
University of Minho
Braga, Portugal
LABBELS –
Associate Laboratory,
Braga/Guimarães, Portugal
joao.capela@ceb.uminho.pt

João Correia
Centre of Biological Engineering
University of Minho
Braga, Portugal
LABBELS –
Associate Laboratory,
Braga/Guimarães, Portugal
jfscorreia95@gmail.com

Vítor Pereira
Centre of Biological Engineering
University of Minho
Braga, Portugal
LABBELS –
Associate Laboratory,
Braga/Guimarães, Portugal
vpereira@ceb.uminho.pt

Miguel Rocha
Centre of Biological Engineering
University of Minho
Braga, Portugal
LABBELS –
Associate Laboratory,
Braga/Guimarães, Portugal
mrocha@di.uminho.pt

***Abstract*—The non-caloric sweeteners market is catching up with the market of conventionally used sugars due to the benefits of preventing obesity, tooth decay and other health problems. Developing strategies for designing easier-to-produce novel molecules with a sweet taste and less toxicity are up-to-date motivations for the food industry. In this sense, Machine Learning (ML) approaches have been reported as cutting-edge technologies to guide the design of new molecules towards specific objectives, including sweet taste.**

The largest known dataset of sweet molecules is here provided. The dataset contains fully integrated 9541 sweeteners and 1141 bitterants from FooDB, FlavorDB and literature. This robust dataset allowed the development of standard Machine and Deep Learning pipelines towards conceiving Structure-Activity Relationships (SAR) between molecules and sweetness.

In this work, we showcase that Textual Convolutional Neural Networks (TextCNN), Graph Convolutional Networks (GCN), and Deep Neural Networks (DNNs) outperformed most of traditional "shallow" learning approaches. These Deep Learning (DL) models produced platforms to guide the design of new sweeteners and repurposing existing compounds.

Sixty million compounds from PubChem were evaluated using these models. Herein, we deliver a dataset of 67724 compounds that present high probabilities of being sweet. Quick searches in literature allowed us to find 13 molecules reported as potent sweetening agents, revealing that our approach is suitable for finding new sweeteners, valuable to expand food chemistry databases, repurposing existing chemicals and designing novel molecules with a sweet taste.

***Index Terms*—Machine Learning, Deep Learning, Computational Chemistry, Sweeteners**

I. Introduction

Although the demand for caloric sweeteners is predicted to increase around 1.5% until 2027 (OECD-FAO, 2018), this represents a low growth rate compared to previous periods [1]. The undissociated health concerns are the primary cause for this stagnation and for the race for non-caloric sweeteners in the last 10-15 years and, expectedly, in the years to come [1].

The sweet taste, unequivocally and innately attractive to humans, is the product of the interaction of sweet molecules with T1R2 and T1R3 subunits that compose the heterodimer belonging to the class C of the G-protein coupled receptor family [2]. In spite of its impact in the food industry, the sweet signal transduction response triggered by the T1R2–T1R3 complex remains largely unknown [2].

On the other hand, many authors broadly explored the molecular basis of sweetness in carbohydrates, amino acids, and artificial sweeteners [3]. The first theory postulated that there would be a relation between multiple hydroxyl groups and chlorine atoms. Later on, a more robust hypothesis theorised that a distance greater than 2.5 ångströms (Å) but not inferior to 4 Å separating two electronegative atoms of a molecule, let them be *A* and *B*, would elicit a sweet taste [4]. According to this theory, *AH* was usually an oxygen or a nitrogen atom attached to a hydrogen, while *B* was a Lewis base. The proposed *AH-B* (2.5-4 Å) group was defined as the "saporific group". A hydrophobic *X* group was later added to the theory since most amino acid L-enantiomers were sweet and D-enantiomers were not, even though satisfying the *AH-B* requirement [5].

The sensation of sweetness relies on the structure of both the receptor and the compounds in interaction. In this sense, the changes in compound structures can lead to shifts in taste potency and elicitation [6]. While solid lines of evidence point to the enumerated hypotheses on the molecular basis of sweetness as grounded theories, they seem insufficient to cover tweaks in structures or completely explain the phenomena [7]. Other molecular features also seem essential to confer the sweet taste, hindering the capability of these theories to predict taste on a large scale [8]. Moreover, as structures of sensory receptors are not entirely resolved, and although ligand-based methods have found some success in predicting taste, they are mainly restricted to specialised families of compounds

[6]. Correspondingly, determining molecular sweetness is still a challenging task. However, the advent of online resources and food chemistry databases provides means for data-driven approaches toward implementing computational models to predict sweetness, which are paramount in this context.

Insights into sweeteners' molecular structures may undoubtedly help sustain and strengthen the molecular basis of sweetness, providing a valuable platform to design novel molecules with a sweet taste [9]. In this sense, the last decades were marked with the development of (Quantitative) Structure-Activity Relationship (QSAR) models for sweetness prediction. Since the early 1980s, several authors have focused on training predictive models to discriminate sweet, bitter and tasteless compounds. From 1980 to 2009, the most used (Q)SAR models were Local Linear Approximation (LLA), k-Nearest Neighbours (kNN), Classification and Regression Trees (CART) algorithms, and Quadratic Discriminant Analysis (QDA). These models were trained with small datasets and tested against even smaller ones, mostly performing poorly [10]. The datasets size ranged from 20 to 132 compounds, including sweet and bitter aldoxime derivatives [5], perillartine derivatives, aspartyl dipeptides, carbosulfamates [11], and sulfamate derivatives [12]–[14]. More recent studies resorted to larger datasets to train kNNs [10], Support Vector Machines (SVM) [15], Random Forests (RF) [6], [15], [16], and partial least squares (PLS) regression analysis [17], achieving better results. Moreover, one of these last approaches enumerated structures and properties likely associated with sweetness [17], while other delivered platforms to assess the flavour of molecules in target datasets [6], [16].

The purpose of this work was to compile sweet, bitter and compounds with other flavours from available online resources and develop a relevant and robust SAR system based on both Machine Learning (ML) and Deep Learning (DL) models. Moreover, we aim at validating one of our models against the established theories of molecular basis by interpreting the developed SAR system from a molecular point of view. Ultimately, the main goal is to create a fast and optimised system to classify molecules in terms of sweetness. This system was used to conduct an optimised search over PubChem's molecular space towards repurposing existing chemicals, corroborating our approach's usefulness.

II. Material and Methods

A. DeepMol and overall pipeline

DeepMol is a python package that provides a smoother approach to ML/DL pipelines applied to chemoinformatics. The package covers the molecules preprocessing, generation of features, their selection, model construction, and hyperparameter optimisation. As for the model construction, DeepMol uses Tensorflow (https://www.tensorflow.org/), Keras (https://keras.io/), Scikit-learn (https://scikit-learn.org/) and DeepChem (https://deepchem.io/) to build custom ML and DL models or use pre-built ones. Moreover, it uses the *rdkit* "Mol" object as the data structure to represent molecules. It is worth noting that *rdkit* (https://www.rdkit.org/) is the *de facto* package for chemoinformatics and computational chemistry, providing compliance with a vast number of frameworks, including those specialised for ML/DL tasks.

The first phase of this work was to collect and integrate sweet, bitter and molecules with other flavours from various data sources (Blue line of Fig. 1), developing an Extract-Transform-Load (ETL) pipeline using the Django framework (https://www.djangoproject.com/). Afterwards, an ML pipeline was developed entirely using DeepMol (Red line of Fig. 1). This pipeline was divided into five steps: data standardisation, feature generation, feature selection, model construction and hyperparameter optimisation. These steps will be explained thoroughly in the following sections.

B. Data preparation

Before developing the ETL pipeline, we set up a staging area with molecular data from literature [6], [16], [18], FooDB (November 2021), and FlavorDB. FooDB's comma-separated values (CSV) files were downloaded directly from https://foodb.ca/downloads, whereas FlavorDB's information (November 2021) was retrieved with web-scraping and further converted into a CSV file. In the "Extract" step, the Simplified Molecular Input Line Entry System (SMILES) string, flavour properties, and external references were extracted from the CSVs. SMILES are computational representations of compounds' chemical structures, vastly used in ML [19]. In the "Transform" stage, molecule's SMILES suffered a simple standardisation and were converted into InChIKeys and *rdkit* "Mol" objects. InChIKeys are unique 27 character keys to hash the International Chemical Identifier (InChI) information, representing compounds' structures, providing concise notations to index molecules in databases. Finally, in the "Load" phase, compounds were stored in a relational database and integrated by InChIKeys to avoid redundancy.

SMILES strings were standardised according to the following operations: removing isotopes, neutralising charges, keeping the biggest fragment when disconnected; and, kekulizing molecules to simplify and expose double bonds in aromatic substructures. Finally, all duplicates were removed.

The dataset included sweet, bitter and compounds with other flavours. Since the latter set is bigger, undersampling was performed to balance the dataset. The Elkan algorithm was used to compute K-Means and find 10 clusters in the latent space of t-distributed Stochastic Neighbor Embedding (t-SNE) computed from molecular similarities. The centroid nearest compounds were chosen and kept for the final set of molecules. Finally, the bitter ones were included in the class "Non-sweet", along with compounds with other tastes. The final dataset was divided into train (50%) and test (50%) sets by a stratified split.

C. Molecular representations

There are several ways of representing molecules for supervised tasks. Molecular descriptors represent two-dimensional (2D) chemical properties and substructures, such as molecular weight and number of rings, for instance. On the other

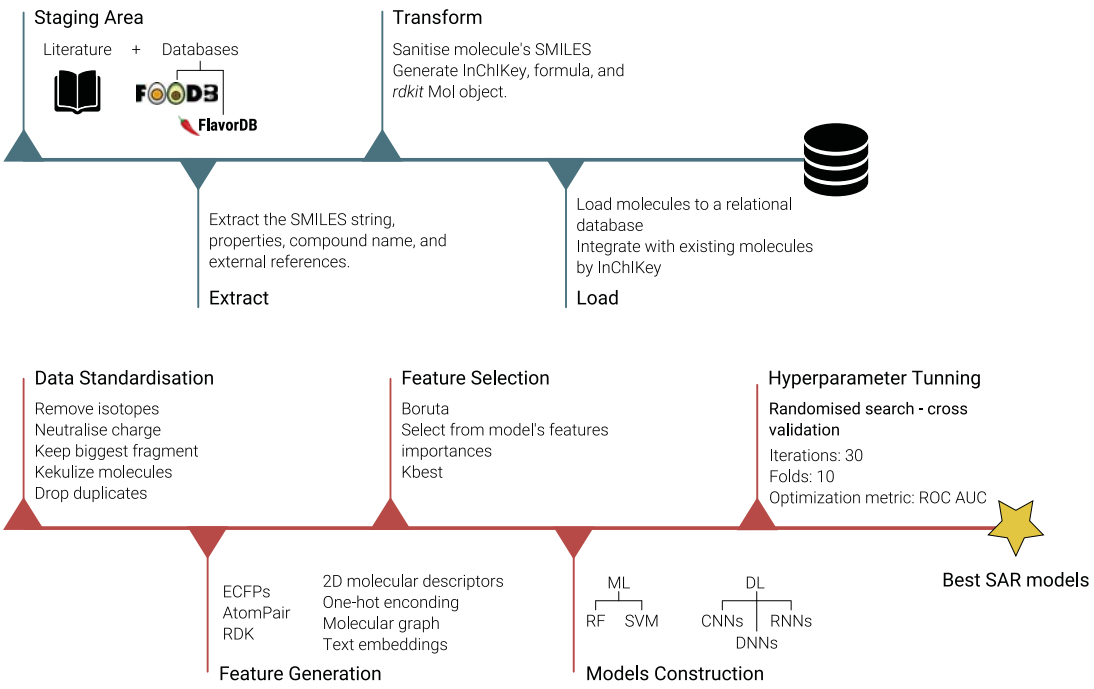

Fig. 1. General pipeline. Blue line - Extract-Transform-Load (ETL) pipeline. Red line - Machine Learning pipeline

hand, molecular fingerprints (FP)s are one-dimensional binary bit-vectors encoding the presence or absence of specific substructures. These approaches are widely used and established in developing (Q)SAR systems for sweetness [6], [10], [15]–[17]. Molecular FPs are mostly divided into atom pairs, and substructure FPs [20]. Accordingly, the former hashes the molecular distance between pairs of atoms and their chemical element, being documented as suitable for large molecules [20]. In contrast, the latter captures circular molecular fragments, which are suitable for predicting the biochemical activity of small organic molecules, especially the Extended Connectivity Fingerprints (ECFP) [20].

A recent approach is to use molecular graphs to represent molecules, being the most straightforward way to map the atoms and bonds into nodes and edges, adding atoms' spatial information as nodes or/and edges' attributes. Even though a graph is a 2D data structure, this representation can capture 3D information, not as spatial coordinates but as pairwise relationships [19]. Such an approach provides a suitable method to include chirality, bond length, and other relevant 3D information in (Q)SAR [19]. Alternatively, one can encode the SMILES sequence directly and use the resulting matrix to train ML/DL models. The encoding process is preceded by SMILES tokenisation to generate relevant tokens from a chemical point of view [21].

In this work, all these representations were covered. We applied *rdkit* molecular descriptors, circular FPs of two radius sizes (ECFP4 and ECFP8), atom pair FPs (AtomPairFP), RDK FPs to enumerate all possible substructures with atom count from 2 to 4, molecular graphs, and SMILES encodings. As for molecular graphs, we used two implementations from DeepChem: Duvenaud graph convolutions [22] to construct a vector of descriptors for each atom in each molecule with default parameters, and a graph representation that accounts for the characterisation of each node and edge. The information contained for each node was the following: the atom type, formal charge, hybridisation, hydrogen bonding, aromaticity, degree, and the number of hydrogens. The information for each edge included the bond type, whether the pair of atoms are in the same ring, and the conjugation. Finally, one-hot encodings were generated using Smiles Pair Encoding (SmilesPE) [21] as a tokeniser to train Long Short-Term Memory (LSTM)s and Bidirectional LSTMs (BiLSTM)s. The maximum size of tokens was set to 138 to cover 99.9% of the molecules of the dataset.

D. Feature Selection

The feature selection was the third step of the pipeline, where three methods were employed:

- Boruta;
- Train a model of our choice and select the features above a feature importance threshold (threshold of 1e-5) - this approach will be referred to as "SelectFromModel";
- Selection of the k best features using a chi-squared test (500 for FPs and 100 for 2D molecular descriptors);

Boruta is an algorithm that uses RFs to identify the most relevant features to the model. This algorithm starts by duplicating the dataset's features and shuffling the values in each

duplicated column. These values are referred to as ”shadow” features. The RF classifier is trained with the ”real” and the ”shadow” features. Ultimately, the algorithm evaluates whether the ”real” features are more important than ”shadow” features. This process is repeated several times until all the necessary and unimportant features have been identified.

The model used for the first and second approach was an RF with 900 estimators, with a maximum depth of 80 and 2 minimum samples per leaf. The number of iterations for Boruta was 100.

E. Model construction

One of the aims of this work was to evaluate the performance of established ML algorithms and alternative DL approaches. Accordingly, default RFs and SVMs from the Python package Scikit-learn were implemented for each set of generated FPs and 2D molecular descriptors. Alternatively, Deep Neural Networks (DNN)s, LSTMs, BiLSTMs, Convolutional Neural Networks (CNN)s, and Graph Neural Networks (GNN)s were implemented.

DNNs were composed of an input layer with as many neurons as the number of features, a variable number of dense hidden layers and neurons in each layer (values that may be optimised), and a dense output layer with one neuron for predicting the class. The activation functions of the hidden and output layers were Rectified Linear Unit (ReLU) and sigmoid, respectively. Also, a dropout mask was added to each hidden layer (the dropout rate being a hyperparameter to optimise) to avoid overfitting. We defined the binary cross-entropy as the loss function and the training algorithm selected as another hyperparameter, from Adaptive Moment Estimation (Adam), Adamax, Adaptive Gradient Algorithm (AdaGrad), or AdaDelta. DNNs were trained with FPs and 2D molecular descriptors.

LSTMs and BiLSTMs were composed of a variable number of (Bi)LSTM, dense layers and number of neurons in each layer (subject to optimisation). The output layer, the activation functions, dropout, loss function, and training algorithms were implemented as in the DNNs. LSTMs and BiLSTMs were trained with one-hot encoding vectors.

Three GNNs were implemented: Graph Attention Networks (GAT) [23], Graph Convolution Networks (GCN) [24] trained with DeepChem’s graph representation, and a GNN trained with Duvenaud graph convolutions’ features (GraphConv) [22]. Finally, the textual CNN described in [25] was implemented using DeepChem. The model pads and splits the SMILES strings into characters that are used to generate one-hot vectors and applies them to multiple 1D convolutional filters. This process is followed by a max-over-time pooling of the filters, extracting one feature per filter. The extracted features are inputted into hidden dense layers to obtain predictions.

After implementing the models, a hyperparameter optimisation was conducted with a randomised search. Hence, thirty combinations of hyperparameters were randomly chosen for each model and tested on 10-fold cross-validation. Finally, the best model was selected based on the higher Receiver Operating Curve-Area Under Curve (ROC-AUC).

Model construction, training and optimisation as well as all the operations here reported were performed in a computer with a Graphics Processing Unit (GPU) NVIDIA Tesla T4 with 16GB of memory.

F. Models evaluation - metrics

The SAR system will predict the probability of a compound being sweet. The models’ performance will be assessed through several metrics: sensitivity/recall, specificity, precision, Non-Error Rate (NER) (NER or balanced accuracy is the average Sensitivity and specificity), and ROC-AUC (the Area Under the Curve of a Receiver Operating Characteristic curve).

G. Feature explainability

One of the aims of this work is to interpret the developed SAR system from a molecular point of view. Accordingly, the best ML/DL models were evaluated using SHapley Additive exPlanations (SHAP) values. SHAP is a method used in game theory to determine each player’s contribution in a collaborative game towards success. When applied to ML tasks, SHAP values help to measure each feature’s contribution to the predictions made. The SHAP value for each feature corresponds to its mean marginal contribution to the model’s output on all the possible feature combinations.

H. PubChem’s search

The model with best balance in terms of the metrics above and classification speed was used to perform an optimised search over a large sample of PubChem’s molecules. These compounds were classified by the chosen model and filtered using a threshold of 0.8 (Filter 1 in Fig. 2). Then, the resulting compounds are classified by an ensemble of the best models for each molecular representation (Filter 2 in Fig. 2).

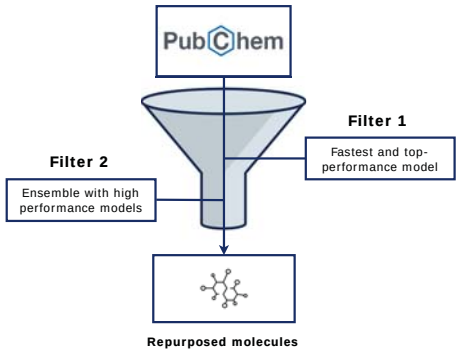

Fig. 2. PubChem search pipeline. PubChem compounds were classified by the fastest and more accurate models and filtered by a threshold of 0.80 (Filter 1). Then, the filtered compounds were classified by an ensemble of the best models in terms of ROC AUC and Precision for each molecular representation. In the end, a set of repurposed molecules is obtained.

III. Results

A. Data integration results

After the initial data integration, the database comprised 9541 sweet, 1141 bitter, and 69307 molecules with other flavours. After standardisation, dropping duplicates, the numbers decreased to 8008, 1065 and 67789, respectively. The primary source of exclusive sweeteners (unique to a given source) was FlavorDB, followed by Tuwani et al. 2018, as highlighted in Fig.3. Likewise, the major sources of unique bitter compounds were Tuwani et al. 2018 and FlavorDB. FooDB was, for a large margin, the primary source of compounds with other flavours.

The dataset's molecular space can be visualised in Fig. 4, depicting a t-SNE plot of molecular similarities, showing three prominent regions. The dark rounded region represents a "blend" region, where similar molecules of distinct classes occupy the same chemical space. Herein, 2539 are sweeteners, whereas 7833 are not. On the other hand, the green and red regions are spots dominated by sweeteners. 66,7% (2174/3259) of the sweeteners in the green region are compounds with amino-acid moieties. On the other hand, 93,2% (2059/2206) of the sweeteners in the red region possess at least one ring containing one or more oxygen atoms. The "blend" region will represent a set of more challenging molecules for classification. On the contrary, the red and green regions will likely be easier to classify, as these are dominated by molecules belonging to only one class. We will consider a test set with all regions and another with molecules exclusively in the "blend" region.

B. General appreciation on models performance

Four different types of FPs and a set of molecular descriptors were generated to train DNNs, RFs, and SVMs. Three modes of feature selection can be considered as stated above. Sixty models were implemented, 20 of each type (DNN, SVM, RF). Along with the other DL models (GAT, GCN, TextCNN, Graph convolutions, BiLSTMs and LSTMs), the number of implemented and optimised models was 66.

Table I shows the performance of the best 12 models per featurisation method (e.g. ECFP4, graph) and type of model (e.g. SVM, RF, TextCNN). The results are sorted by the optimised metric (ROC AUC).

2D-SelectFromModel-RF provided the best ROC AUC for the test set (0.929), although only by a margin of around 0.001 from the DNN trained with RDK FPs. Similarly, 2D-SelectFromModel-RF obtained a ROC AUC of 0.877 for the "blend" test set, representing around 0,019 more than the second-best model (TextCNN). Three of the five best SAR systems were DL models. The fifth best model was an SVM trained with ECFP4 FPs (0.925), providing an excellent platform to explain features, as ECFPs are broadly used and understandable from a molecular point of view.

C. Deep Learning models' performance and architectures

The first two best DL models are DNNs trained with RDK FPs and 2D descriptors (0.928). The former is composed of only one hidden layer with 64 neurons, a dropout mask of 0.5, and a layer for normalising batches. Moreover, the training algorithm used was AdaGrad, with a learning rate of 0.01. This DNN achieves high ROC AUC with only one hidden layer. The latter took as input only the features selected by the 100 k-

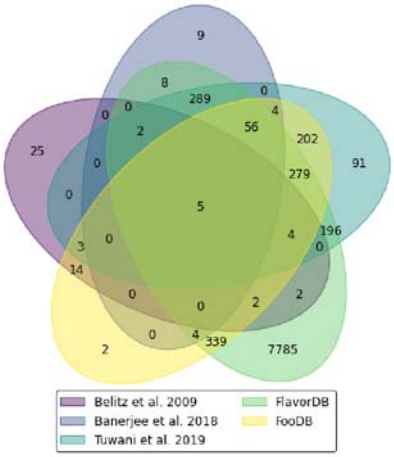

Fig. 3. Main data sources of sweet compounds.

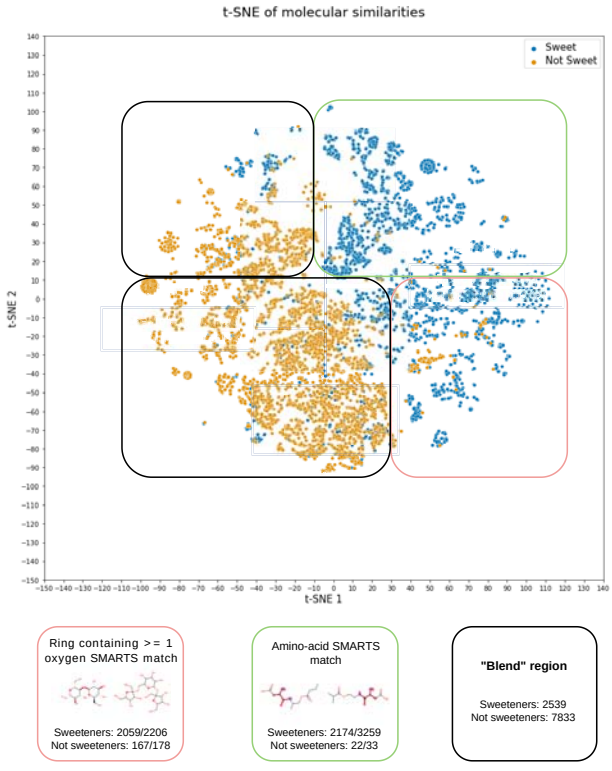

Fig. 4. t-SNE of molecular similarities. Three regions are prominent: 1) green rounded - mostly dominated by sweeteners with amino-acids in their molecular structure; 2) red rounded - mostly dominated by sweeteners with a ring containing one or more oxygens; 3) black rounded - a mixed region with similar sweeteners and non-sweeteners

best features evaluated by a chi-squared test. The architecture of this model comprised two hidden layers with 512 and 256 neurons and a layer for batch normalisation. Adam algorithm was employed to train the network at a learning rate of 0.001. These models might be avoiding overfitting and achieving high ROC AUC due to their simple architecture [26].

The fourth best model was the GCN (test ROC AUC of 0.925), which takes as input the molecular graph. The optimised architecture comprises four graph convolution layers with 128-width channels each. The weighted sum and max-pooling are applied to the node representations, and the results are concatenated. The resultant matrix is inputted to a DNN with a single 64-node hidden layer, and a dropout mask of 0.25. The architecture of this model provided better results than others that used the graph, as is the case of GraphConv (0.920 for the test set and 0.844 for the "blend" set) and GAT (0.914 and 0.812). Finally, the GCN also accomplished better performances than other CNNs (TextCNN) and RNNs (BiLSTM and LSTM).

The GAT outperformed all other models, considering the precision in the test set (0.954). BiLSTM also provided a high precision score for both the test and "blend" set, especially for the latter, as it achieved the best performance (0.949), while maintaining higher Recall and ROC AUC than GAT. GAT's architecture is composed of three convolutional layers (depth of 32). Similarly to GCN, max pooling and the weighted sum of the node representations are concatenated and inputted to a DNN, in this case with a single hidden layer with 265 neurons and a dropout mask of 0.5. The optimised architecture of the BiLSTM compiled three bidirectional LSTM layers with 256 neurons each, with a dropout mask of 0.25. The training algorithm used was Adamax at a learning rate of 0.001. A final hidden layer with eight neurons is added before estimating the output label. A high precision means that a model is not predicting many false positives compared with true positives. Such a metric can measure how conservative is a model for predicting sweetness. A highly precise model delivers more secure predictions for filtering molecules according to their sweetness. Hence, GAT and BiLSTM were the most conservative and precise models for classifying sweeteners, while maintaining satisfactory ROC AUC values.

Models that considered the sequence, namely the described BiLSTM and TextCNN, still achieved good performances, especially the latter. Although having been outperformed by other models in test ROC AUC, TextCNN surpassed all other DL models in the "blend" ROC AUC (0.858). This CNN comprises six kernels with variable sizes, 12 filters and a dropout mask of 0.5. DL algorithms have been achieving top-level performance in a wide range of fields, including (Q)SAR, especially those using larger datasets of molecules [27]. Notably, DL models developed in this work outperformed most of the other employed "shallow" learning models.

D. Models' speed performance

Table II shows a speed analysis of the best models. The classification of 1000, 10000, and 100000 molecules from PubChem was performed to assess the execution time of feature generation and prediction. The fastest model generating features and classifying datasets of 1000 and 10000 instances was BiLSTM. On the other hand, TextCNN was fastest for 100000 molecules, with an execution time of 109,59 seconds. Although 2D-RF is the slowest, it is faster to predict sweetness once features are computed.

TextCNN delivers acceptable ROC AUC, Recall and Precision scores for both test and "blend" sets, while being fast to generate features and predict sweetness. Hence, TextCNN is the most suitable system for optimising search over large databases (as PubChem).

E. Feature explainability

We calculated the SHAP values of ECFP4 bits in the best SVM, as ECFPs are circular FPs that allow us to easily understand structural features in a molecule. Moreover, we were only able to apply SHAP to descriptors/FPs; DL models would require other approaches. Specifically, we evaluated the impact of all bits in the prediction of three artificial sweeteners (cyclamate, alitame and 1-n-propoxy-2-amino-4-nitrobenzene) and one sweet carbohydrate (β-D-Glucopyranose).

Fig. 5 shows that structures reportedly associated with sweetness presented positive SHAP values, meaning that their presence is pushing up the prediction to the class "Sweet". Accordingly, we show here that the reported electronegative atoms separated by a distance between 2.5 and 4 Å and the *X* hydrophobic structure had a positive impact on classifying these molecules as sweet. In this sense, this model is in accordance with the *AH-B* [4] and the *AH-B-X* theories [5].

Notably, for the first time, a recent computational investigation on alitame structure hypothesised that the dihedral angles of -19.64º and 7.03º between two single NH groups and two adjacent C=O groups are associated with sweetness [9]. Our analysis shows that bit 1565 is associated with regions related to the two dihedral angles. The impact of bit 1565 in the sweetness prediction is of 0.016. Although not being the bit with the highest impact, it delivers a positive SHAP value, indicating that this structure contributes to predicting this molecule as sweet, supporting the theory.

F. PubChem search

Sixty million molecules from PubChem were evaluated with TextCNN. The molecules with predicted sweetness higher than 0.8 were selected. The resulting molecules were then subjected to class prediction using an ensemble of the models with highest ROC AUC and Precision. According to this criterion, 2D-SelectFromModel-RF, RDK-DNN, GCN, ECFP4-SVM, AtomPairFP-SelectFromModel-DNN, and BiLSTM were chosen to be part of the ensemble. Not only the models used were very different from each other, but also their molecular representations, providing a platform that covered different parts of the molecular space, including 1D (FPs and sequences) and 2D descriptors. In this sense, it is worth noting that this platform can classify isomers differently, providing predictions that take into account each molecule's stereochemistry.

TABLE I
MODELS' PERFORMANCE

Descriptor-FS Method -Algorithm	Test ROC AUC	Test Precision	Test Recall	Blend ROC AUC	Blend Precision	Blend Recall
2D-SelectFromModel-RF	**0.929**	0.925	**0.933**	**0.877**	0.929	**0.815**
RDK-DNN	0.928	0.947	0.906	0.849	0.942	0.745
2D-Kbest-DNN	0.928	0.941	0.912	0.857	0.945	0.757
GCN	0.925	0.946	0.901	0.844	0.940	0.736
ECFP4-SVM	0.925	0.937	0.911	0.855	0.942	0.756
AtomPairFP-SelectFromModel-DNN	0.925	0.945	0.902	0.842	0.939	0.732
ECFP8-SVM	0.920	0.930	0.908	0.847	0.936	0.746
GraphConv	0.920	0.931	0.906	0.844	0.926	0.748
TextCNN	0.920	0.915	0.925	0.858	0.907	0.798
GAT	0.914	**0.954**	0.870	0.812	0.949	0.659
BiLSTM	0.912	0.944	0.884	0.829	**0.949**	0.699
LSTM	0.729	0.698	0.918	0.687	0.687	0.779

TABLE II
BEST MODELS' SPEED PERFORMANCE IN THE CLASSIFICATION OF 1000, 10000, AND 100000 MOLECULES.

Model	Feature generation speed			Prediction speed		
	1000	10000	100000	1000	10000	100000
TextCNN	0.76s	7.19s	**58.6s**	3.28s	7.40s	**50.99s**
BiLSTM	**0.12s**	**2.53s**	114.12s	**1.27s**	**7.14s**	77.56s
ECFP4-SVM	5.47s	54.78s	552.51s	3.67s	36.39s	385.07s
RDK-DNN	8.44s	55.87s	563.90s	0.10s	0.80s	7.51s
GCN	7.97s	75.80s	844.81s	0.48s	5.40s	50.65s
GAT	7.97s	75.80s	844.81s	0.51s	6.23s	54.90s
2D-RF	16.89s	134.84s	1212.83s	0.10s	0.60s	6.98s

TABLE III
ASPARTAME DERIVATIVES FOUND TO BE REPORTEDLY USED AS SWEETENING AGENTS.

PubChem identifier	Ensemble prediction probability	Reference
14151484	0.999	EP-0186292-A2 (patent)
14151460	0.999	EP-0186292-A2 (patent)
14151470	0.994	EP-0186292-A2 (patent)
14151451	0.965	EP-0186292-A2 (patent)
11213284	0.939	WO-2021076608-A1 (patent)
22798087	0.930	[28]

1444212 out of sixty million (2.4%) molecules were filtered using TextCNN (Filter 1). The filter 2 delivered 67724 repurposed chemicals (4.6% of the ones obtained from Filter 1). A substructure search on this set revealed the following number of derivatives of potent and artificial sweeteners:

- 199 of aspartame (200 times sweeter than sucrose);
- 8 of cyclamate (50x);
- 11 of acesulfame (120x);
- 7 of alitame (2000x);
- 231 of saccharin (500x);
- 3 of dulcin (250x);

A quick search on 30 aspartame derivatives returned 6 that were reportedly used as sweetening agents (Table III).

Notably, 28 molecules of this subset are compounds with guanidine moieties. Nofre and Tinti [29] reported that guanidine-derived molecules demonstrate high efficiency as sweeteners. A quick search on these 28 molecules revealed that at least 7 are highly sweet. Table IV provides the PubChem identifiers, the ensemble's prediction probability, and an article or patent reference where the molecules are registered as sweeteners. These findings reveal the effectiveness of our

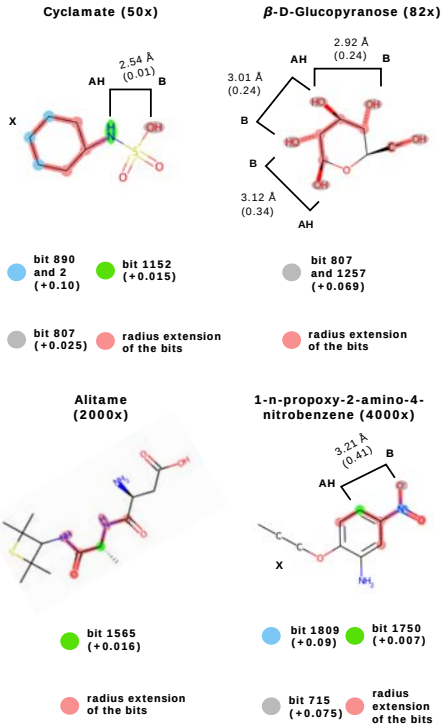

Fig. 5. The four structures of cyclamate, alitame, 1-n-propoxy-2-amino-4-nitrobenzene and β-D-Glucopyranose and their relative sweetness to sucrose are presented here. Moreover, the bits associated with the *AH-B-(X)* [4], [5] and the Altunayar-Unsalan and Unsalan [9] dihedral angles theories are highlighted and the impact in model prediction is shown between parenthesis. The distances between atoms are the average of distances calculated for 50 different optimised conformations of the molecule (distances between *AH* and *B* atoms), with the standard deviation between parenthesis. Notably, the highlighted substructure of alitame covers the adjacent NH and C=O groups, supporting Altunayar-Unsalan and Unsalan theory [9].

TABLE IV
GUANIDINES FOUND TO BE REPORTEDLY USED AS SWEETENING AGENTS.

PubChem identifier	Ensemble prediction probability	Reference
13748439	0.879	US-4673582-A (patent)
14230963	0.770	JP-H0655730-B2 (patent)
14230962	0.747	EP-0241395-A2 (patent)
13960823	0.739	EP-0241395-A2 (patent)
4447	0.649	[30]
13960822	0.628	EP-0241395-A2 (patent)
14230964	0.544	EP-0241395-A2 (patent)

approach in finding new sweetening agents.

IV. CONCLUSION

In this work, we implemented an ETL pipeline to integrate sweet, bitter and tasteless molecules, and 66 ML and DL models to distinguish molecules with a sweet taste. To the best of our knowledge, we provide the largest dataset of sweet molecules and the first DL models to predict sweetness. We have shown that our models are useful to expand the molecular basis of sweetness and suitable for large-scale predictions. Furthermore, we provide accurate and precise classifiers for the design of new molecules, e.g., using multi-objective evolutionary algorithms (EA), along with deep generative models [31]. However, EAs should include the absence of toxicity as one of the objective functions since we neglected that matter in the data integration and models' training.

The whole pipeline, models and results can be found in https://github.com/BioSystemsUM/DeepSweet.

ACKNOLEDGEMENTS

This research has been supported by the Portuguese Foundation for Science and Technology (FCT) through the DeepBio project - ref. NORTE-01-0247-FEDER- 039831, funded by Lisboa 2020, Norte 2020, Portugal 2020 and FEDER - Fundo Europeu de Desenvolvimento Regional. We also thank FCT for the PhD fellowships to J. Capela (DFA/BD/08789/2021) and J. Correia (SFRH/BD/144314/2019).

REFERENCES

[1] A. Hernández-Pérez, F. M. Jofre, S. De Souza Queiroz, P. Vaz De Arruda, A. K. Chandel, and M. D. G. D. A. Felipe, "Biotechnological production of sweeteners," *Biotechnological Production of Bioactive Compounds*, pp. 261–292, 2020.

[2] J. M. Perez-Aguilar, S. G. Kang, L. Zhang, and R. Zhou, "Modeling and Structural Characterization of the Sweet Taste Receptor Heterodimer," *ACS Chem. Neuroscience*, vol. 10, pp. 4579–4592, 2019.

[3] J. F. Robyt, *Sweetness*. Springer, New York, NY, 1998, pp. 142–156.

[4] R. S. Shallenberger and T. E. Aeree, "Molecular structure and sweet taste," *J. Agricultural and Food Chemistry*, vol. 17, pp. 701–703, 1969.

[5] L. B. Kier, "Molecular structure influencing either a sweet or bitter taste among aldoximes," *J. Pharmaceutical Sciences*, vol. 69, 1980.

[6] R. Tuwani, S. Wadhwa, and G. Bagler, "BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules," *Scientific Reports*, vol. 9, 2019.

[7] W. J. Spillane, "Molecular structure and sweet taste," *Advances in Sweeteners*, 1996.

[8] J. M. DeMan, *Principles of Food Chemistry*, ser. Food Science Text Series. Boston, MA: Springer US, 1999.

[9] C. Altunayar-Unsalan and O. Unsalan, "Structural and anharmonic vibrational spectroscopic analysis of artificial sweetener alitame: A DFT study for molecular basis of sweet taste," *Journal of Molecular Structure*, vol. 1246, p. 131 157, 2021.

[10] C. Rojas, D. Ballabio, V. Consonni, P. Tripaldi, A. Mauri, and R. Todeschini, "Quantitative structure-activity relationships to predict sweet and non-sweet tastes," *Theor Chem Acc*, vol. 135, 2016.

[11] Y. Takahashi, H. Abe, Y. Miyashita, Y. Tanaka, H. Hayasaka, and S.-I. Sasaki, "Discriminative Structural Analysis Using Pattern Recognition Techniques in the Structure-Taste Problem of Perillartines," *J. Pharmaceutical Sciences*, vol. 73, pp. 737–741, 1984.

[12] W. J. Spillane and G. McGlinchey, "Structure-activity studies on sulfamate sweeteners II: semiquantitative structure-taste relationship for sulfamate (RNHSO-3) sweeteners-the role of R," *J Pharmaceutical sciences*, vol. 70, pp. 933–935, 1981.

[13] D. P. Kelly, W. J. Spillane, and J. Newell, "Development of structure-taste relationships for monosubstituted phenylsulfamate sweeteners using classification and regression tree (CART) analysis," *Journal of agricultural and food chemistry*, vol. 53, pp. 6750–6758, 2005.

[14] W. J. Spillane, C. M. Coyle, B. G. Feeney, and E. F. Thompson, "Development of structure-taste relationships for thiazolyl-, benzothiazolyl-, and thiadiazolylsulfamates," *Journal of agricultural and food chemistry*, vol. 57, pp. 5486–5493, 2009.

[15] J. B. Chéron, I. Casciuc, J. Golebiowski, S. Antonczak, and S. Fiorucci, "Sweetness prediction of natural compounds," *Food chemistry*, vol. 221, pp. 1421–1425, 2017.

[16] P. Banerjee and R. Preissner, "Bitter sweet forest: A Random Forest based binary classifier to predict bitterness and sweetness of chemical compounds," *Frontiers in Chemistry*, vol. 6, 2018.

[17] P. K. Ojha and K. Roy, "Development of a robust and validated 2D-QSPR model for sweetness potency of diverse functional organic molecules," *Food Chem. Toxicology*, vol. 112, pp. 551–562, 2018.

[18] H. D. Belitz, W. Grosch, and P. Schieberle, "Food chemistry," *Food Chemistry*, 2009.

[19] L. David, A. Thakkar, R. Mercado, and O. Engkvist, "Molecular representations in AI-driven drug discovery: a review and practical guide," vol. 12, 2020.

[20] A. Capecchi, D. Probst, and J. L. Reymond, "One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome," *Journal of Cheminformatics*, vol. 12, 2020.

[21] X. Li and D. Fourches, "SMILES Pair Encoding: A Data-Driven Substructure Tokenization Algorithm for Deep Learning," *Journal of chemical information and modeling*, vol. 61, pp. 1560–1569, 2021.

[22] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, *et al.*, "Convolutional Networks on Graphs for Learning Molecular Fingerprints," *Advances Neural Information Processing Systems 28 (NIPS)*, 2015.

[23] P. Veličković, A. Casanova, P. Liò, G. Cucurull, A. Romero, and Y. Bengio, "Graph attention networks," *6th International Conference on Learning Representations (ICLR) - Conference Track Proc.*, 2018.

[24] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2016.

[25] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *EMNLP - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1746–1751, 2014.

[26] M. Uzair and N. Jamil, "Effects of hidden layers on the efficiency of neural networks," *Proceedings - 2020 23rd IEEE International Multi-Topic Conference*, 2020.

[27] F. Ghasemi, A. Mehridehnavi, A. Pérez-Garrido, and H. Pérez-Sánchez, "Neural network and deep-learning algorithms used in qsar studies: Merits and drawbacks," *Drug Dis Today*, vol. 23, 10 2018.

[28] M. Kawai, M. Chorev, J. Marin-Rose, and M. Goodman, "Peptide sweeteners. 4. Hydroxy and methoxy substitution of the aromatic ring in L-aspartyl-L-phenylalanine methyl ester. Structure-taste relationships," *Journal of medicinal chemistry*, vol. 23, 1980.

[29] J.-M. Tinti and C. Nofre, "Design of Sweeteners," *Sweeteners*, vol. 450, pp. 88–99, 1991.

[30] Y. Wan, H. Wu, N. Ma, *et al.*, "De novo design and synthesis of dipyridopurinone derivatives as visible-light photocatalysts in productive guanylation reactions," *Chem Sci*, vol. 12, 2021.

[31] T. Sousa, J. Correia, V. Pereira, and M. Rocha, "Combining multi-objective evolutionary algorithms with deep generative models towards focused molecular design," *Lecture Notes in Computer Science*, vol. 12694, pp. 81–96, 2021.