Establishment of PyUmami model

This study established the PyUmami model to predict the candidate umami molecules, complementing the BitterSweet model[1]. The following flow diagram (Figure 1) demonstrates that the construction of the PyUmami model is divided into three key steps: (1) conducted a comprehensive sweet/bitter compounds compilation and curation for training and applying the PyUmami model; (2) calculated the compounds' Mordred descriptor and extracted the bitter-feature or sweet-feature; (3) established the Sweet MLP model and the Bitter MLP model and applied them in searching candidate umami molecules. The development of these three steps will be detailed below.

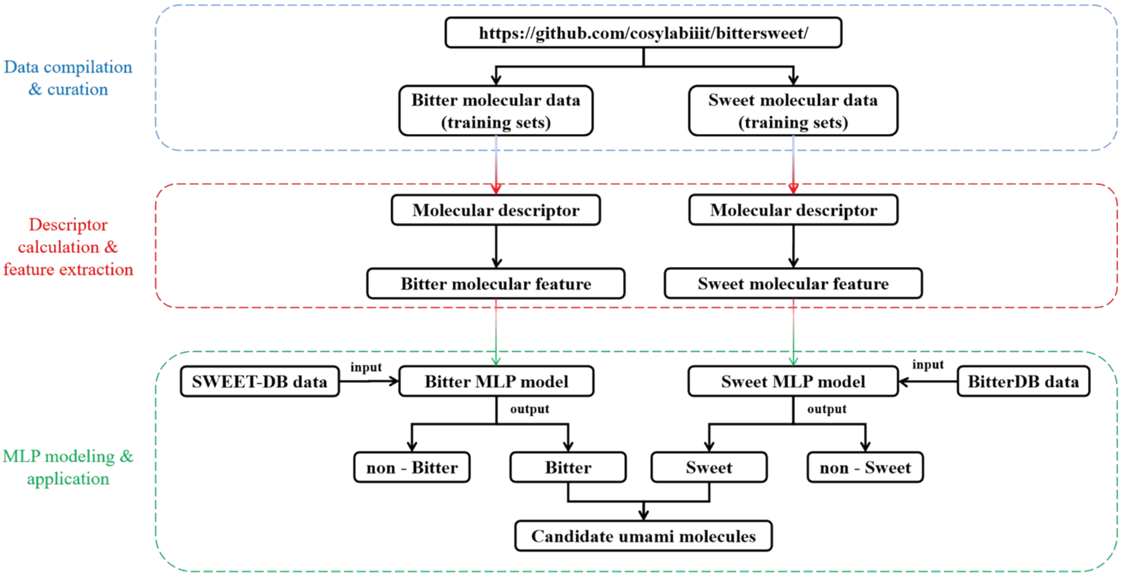
Data compilation & curation
https://github.com/cosylabiiit/bittersweet/
Bitter molecular data (training sets)
Sweet molecular data (training sets)
Descriptor calculation & feature extraction
Molecular descriptor
Molecular descriptor
Bitter molecular feature
Sweet molecular feature
MLP modeling & application
SWEET-DB data
input
Bitter MLP model
Sweet MLP model
input
BitterDB data
output
output
non - Bitter
Bitter
Sweet
non - Sweet
Candidate umami molecules

Figure 1. Flow diagram illustrating the running steps of the PyUmami model.

1. Data compilation and curation

The data of training the PyUmami model were the bitter and sweet molecular data from https://github.com/cosylabiiit/bittersweet/. The deduplication step needs to be executed on the bitter molecular data (training sets) after comparing it with the SWEET-DB data to ensure the generalization ability of the PyUmami model[2]. At the same time, the deduplication step also needs to be performed on the sweet molecular data (training sets) after comparing it with the BitterDB data.

2. Descriptor calculation and feature extraction

The umami taste prediction model was trained and evaluated using Mordred 2D descriptors, calculated by the Mordred software when Canonical SMILES as input. Mordred, a developed descriptor-calculation software application that can calculate more than 1800 descriptors and can be utilized for cheminformatics studies, such as those on quantitative structure-property relationships[3]. Moreover, the molecules with missing values in the calculation results were screened out.

The critical features for the sweet or bitter molecules were extracted using the boruta algorithm, a feature selection algorithm based on a random forest algorithm[4]. The importance of all the features of the training sets was acquired through the boruta algorithm. Then, the essential features were screened, and the redundant feature variables were deleted. Finally, 317 bitter-feature and 287 sweet-feature descriptors predicting sweet and bitter characteristics were extracted.

3. MLP modeling and application

A multilayer perception model (MLP) for sweetness or bitterness was established with the obtained bitter-feature or sweet-feature as the training sets (80%) and testing sets (20%). The training set was used to develop the classification models. As an extra independent sample, the test set was used to validate the reliability and stability of the classification models. Then, the Sweet MLP model and the Bitter MLP model were obtained. The performance of the PyUmami model is described in detail in the result part of our manuscript. The accuracy of the Bitter - and the Sweet - MLP models reached 0.83 and 0.81, respectively, and they all had an AUROC of 0.90. It can be concluded that the PyUmami model performs better than the BitterSweet model of Tuwani et al. with accuracy and AUROC of 0.801 and 0.852, respectively [1]. Therefore, these results suggest that the Bitter - and the Sweet - MLP models performed almost flawlessly in predicting compounds with bitter and sweet characteristics.

Next, the Mordred descriptor information of bitter compounds from the BitterDB database was input into the Sweet MLP model. And the compound would be classified as sweet/ non-sweet by the Sweet MLP model. In the same way, the descriptor information of sweet compounds, which comes from the SWEET-DB database, was input into the Bitter MLP model. And the compound would be classified as bitter/ non-bitter by the Bitter MLP model. Unlike standard regression, logistic regression predicts the ‘probability’ of given input (SWEET-DB/BitterDB data) belonging to a particular class (sweet or non-sweet, bitter or non-bitter)[5]. The output always lies between [0,1]. When the output lies [0.5,1], the Sweet - and the Bitter - MLP models classify the input as ‘sweet’ and ‘bitter,’ respectively. Finally, the candidate umami molecules were screened out when the BitterDB data was classified as ‘sweet,’ and the SWEET-DB was classified as ‘bitter,’ respectively.

As a result, 765 candidate umami molecules were screened out by the PyUmami model. More importantly, an evidence-based conclusion was reached that the PyUmami model would advance our understanding of the molecular correlates of bitter-sweet taste and provide a theoretical framework searching for umami compounds with bitter and sweet characteristics.

Data Availability

The datasets and models are publicly available at https://github.com/wzhanglab/pyumami.

References

1. Tuwani, R., Wadhwa, S. & Bagler, G. BitterSweet: Building machine learning models for

predicting the bitter and sweet taste of small molecules. *Sci. Rep.-UK*. **9**, (2019).

2. Oneto, L., Cipollini, F., Ridella, S. & Anguita, D. Randomized Learning: Generalization Performance of Old and New Theoretically Grounded Algorithms. *Neurocomputing*. **-**, (2018).
3. Moriwaki, H., Tian, Y., Kawashita, N. & Takagi, T. Mordred: A molecular descriptor calculator. *J. Cheminformatics*. **10**, (2018).
4. Degenhardt, F., Seifert, S. & Szymczak, S. Evaluation of variable selection methods for random forests and omics data sets. *Brief. Bioinform.* **20**, (2017).
5. Meurer, W. J. & Tolles, J. Logistic Regression Diagnostics Understanding How Well a Model Predicts Outcomes. *JAMA-J. Am. Med. Assoc.* **317**, 1068-1069 (2017).