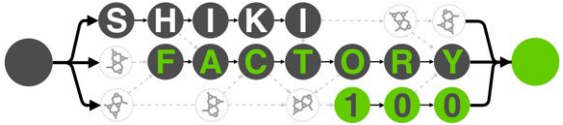


SHIKIFACTORY100

WP3
CHEMICAL SYNTHESIS OF NEW PRODUCTS

João Correia (UMINHO)

21-Nov-2019

DSPT
DATA SCIENCE PORTUGAL

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

About Me

João Correia

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

Universidade do Minho
Doctor of Philosophy - PhD, Informatics
2019 – 2023

DeepRetro: a computational framework for retrosynthesis and pathway design towards optimizing compound bioproduction

Universidade do Minho
Master's degree, Bioinformatics
2016 – 2018

SPINET: Syndemic Protein Interaction NETworks

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

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

WP3 Objectives

1. To design new sweeteners using computational approaches

   Task 3.1 Computational identification of new sweeteners

   Task 3.2 Generative models for new compounds

2. To synthesize the new structures using organic synthesis techniques

   Task 3.3 Synthesis of the new structures using organic synthesis techniques

3. To assess the functional potency and possible side effects/toxicity of the new compounds

   Task 3.4 Potency assay of the new compounds

   Task 3.5 Side effect assay of the new compounds

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

Pipeline Architecture

- The developed machine learning pipeline is divided into multiple parts:
  - The database;
  - Compound featurization, feature selection and hyperparameter optimization;
  - ML/DL models (Random Forest, Support Vector Machine and Deep Neural Network);
  - Virtual screening and ranking of new molecules.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

DATA

- Data set 1: 2.294 compounds (1.224 Sweet / 1.070 Non-sweet)
- Data set 2: 23.103 compounds (1.231 Sweet / 21.872 Non-sweet)

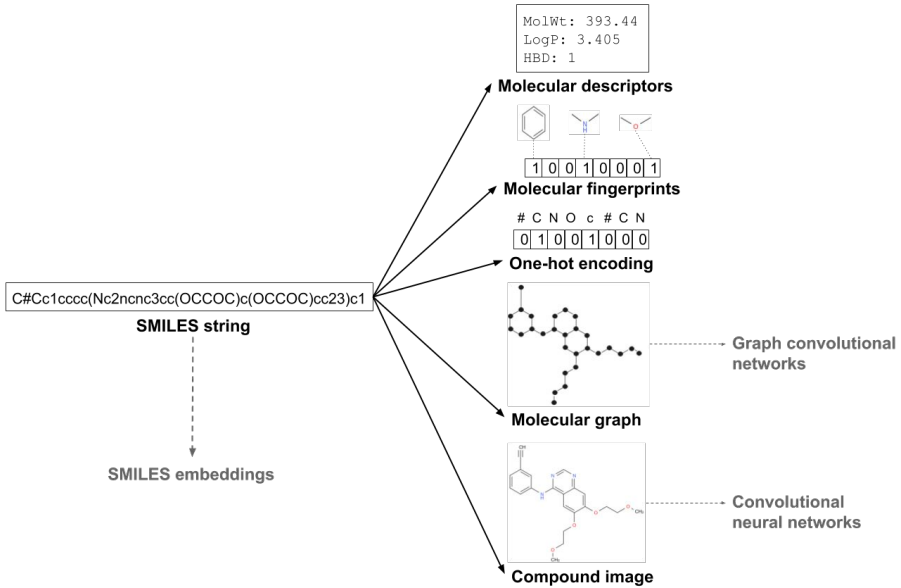
  Data set 1 + Compounds from FoodDB

- *Sources of information:*
  - *Belitz et al. (Food Taste Chemistry, ACS Symposium Series Vol. 115, 1979, pp 93-131)*
  - *foodDB (http://foodb.ca/)*
  - *FlavorDB (https://iiitd.ac.in/FlavorDB)*
  - *Tuwani et al. (Scientific Reports, 9, 7155 (2019))*
  - *other*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

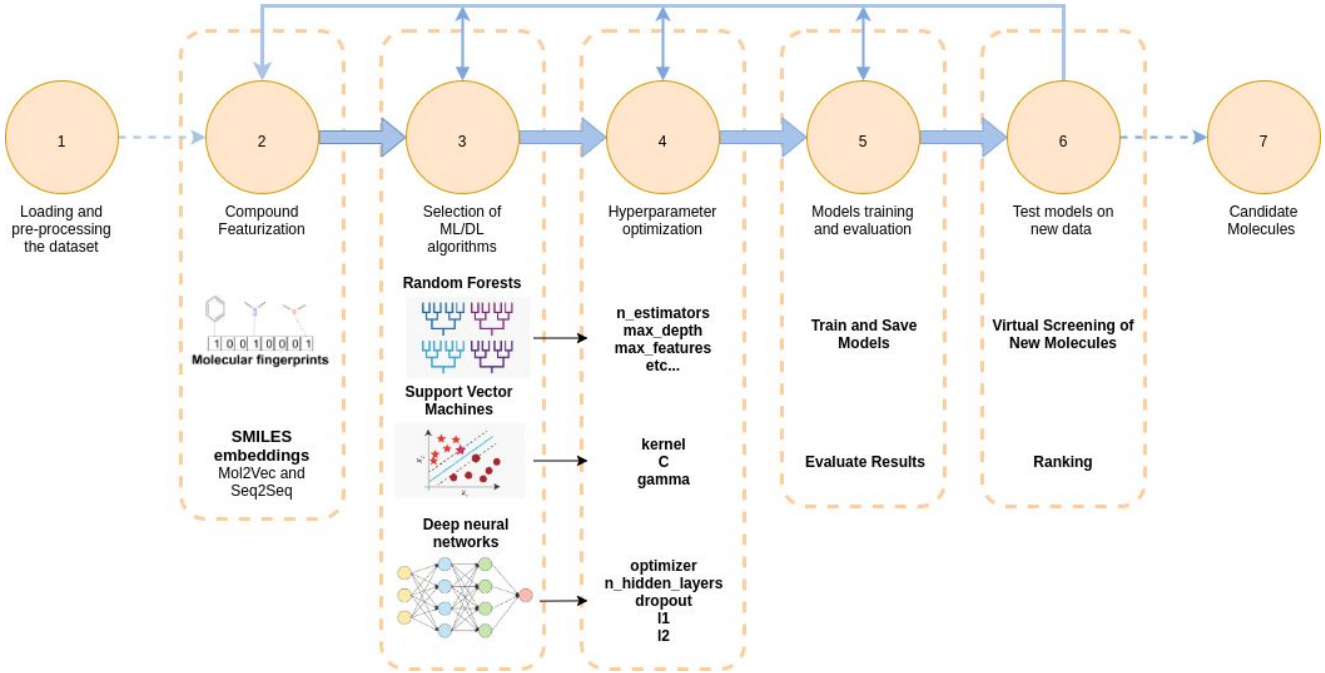
Compound Featurization

- How to efficiently represent the compounds so that the models can better learn and generalize the properties/substructures shared among the molecules?



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

Machine Learning Pipeline



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

RESULTS

- In general, RFs showed better performances than SVMs or even DNNs.

  - RFs with layered FPS (2.048 feat.) with no feature selection:

    CV (k=5) accuracy using Best values #1: 0.84 (+/- 0.01)

1 → Sweet
0 → Non-Sweet

Confusion Matrix

	0	1
0	80	14
1	11	89

	precision	recall	f1 score	support
0	0.88	0.85	0.86	94
1	0.86	0.89	0.88	100

		GridSearchCV (nfolds = 5)	
Parameters	**Values**	**Best values #1**	**Best values #2**
max_depth	10, 50, 80, None	50	80
max_features	auto, sqrt	auto	auto
min_samples_split	2, 5, 8, 11	8	11
bootstrap	True, False	True	False
criterion	gini, entropy	gini	entropy
n_estimators	range(100, 1000, 100)	400	300
		Accuracies	
		0.854 (+/- 0.009)	0.854 (+/- 0.01)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

Results with FooDB

- Because the dataset is highly imbalanced, different strategies were applied: SMOTE, class_weight, different levels of imbalance (1:1, 2:1, 3:1) and different thresholds.

- The results seemed a lot better, but…

RF with class_weights = {0: 1, 1: 10} :

Training accuracy: 0.90 (+/- 0.01)
Test accuracy: 0.90

Confusion Matrix

	0	1
0	2148	54
1	62	64

	precision	recall	f1 score	support
0	0.97	0.98	0.97	2202
1	0.54	0.51	0.51	126

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

Results with FooDB

- The results seemed a lot better, but…

SVM with SMOTEENN (over and under sampling):

Training accuracy: 0.99 (+/- 0.01)
Test accuracy: 0.87

Confusion Matrix

	0	1
0	1912	290
1	12	114

	precision	recall	f1 score	support
0	0.99	0.87	0.93	2202
1	0.28	0.90	0.43	126

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

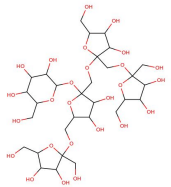
RESULTS

- Screen 25.000 compounds from PubChem:

  - 195 from the 25.000 molecules were predicted as sweeteners by the three models.

  - These molecules were ranked by the sum of the probabilities associated with the prediction of each model for each molecule.

  - Between these molecules we can found known sweeteners that are not present in our dataset. Molecules like Sinistrin - a naturally occurring sugar polymer or polysaccharide, also known as polyfructosane. It belongs to the fructan group. Sinistrin acts as an energy storage molecule in plants.



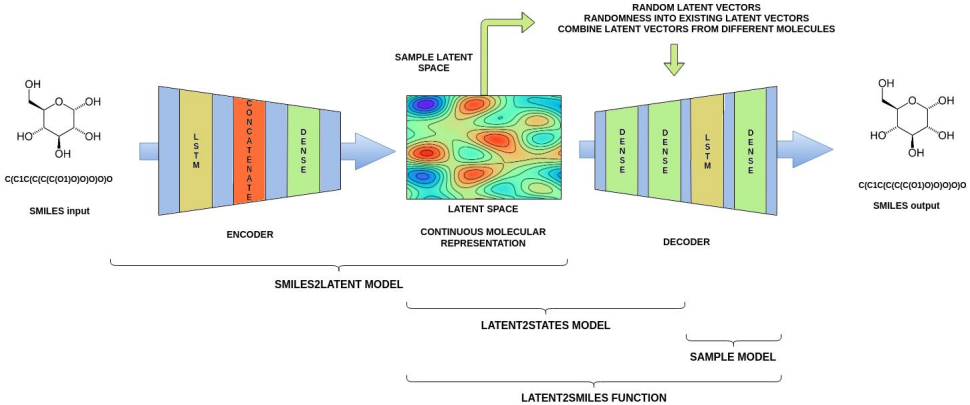
C(C1C(C(C(C(O1)OC2(C(C(C(O2)COC3(C(C(C(O3)CO)O)O)CO)O)O)COC4(C(C(C(O4)CO)O)O)COC5(C(C(C(O5)CO)O)O)CO)O)O)O)O

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

Model 1 - Generative Autoencoder

This model can be divided in three different parts:

- *smiles2latent model*: transforms the one hot encoded SMILES representation into a latent state vector representations (encoder);
- *latent2states model*: decodes the latent space vector representations into the h and c states that are used to decode these latent representations back to SMILES strings;
- *sample model*: predicts the SMILES strings based on the latent space vector representations and states character by character.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

Results - Model 1

- Dataset used for training: PubChem. FL Fragrance-like subset of PubChem downloaded from http://gdb.unibe.ch/downloads/ containing 568.200 molecules (426.150 train and 142.050 test).

- To generate new molecules using Model 1, we used 3 approaches:

  - Decode random latent vectors (5.000 SMILES):
    - 95.2% wrongly formatted SMILES
    - 88.4% new SMILES
    - 11.6% equal SMILES

  - Decode existing latent vectors with different degrees of randomness (5.000 SMILES):
    - 11.3% wrongly formatted SMILES
    - 97.6% new SMILES
    - 2.4% equal SMILES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

Results - Model 1 (cont.)

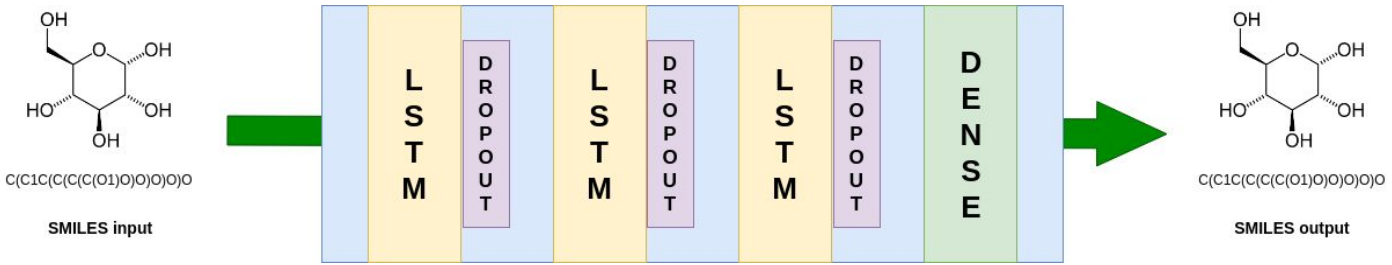
- To generate new molecules using Model 1, we used 3 approaches:

  - Decode latent vectors that are a mix of different ratios of two different molecules (5.000 SMILES):
    - 2.9% wrongly formatted SMILES
    - 83.4% new SMILES
    - 16.6% equal SMILES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

Model 2 - Recursive NN Model with memory - LSTM

- This model simply consists of three stacked LSTM layers followed by dropout and a final dense layer.



- Dataset used for training: molecules from PubChem similar to the sweet molecules present in our internal dataset.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

Results - Model 2

To generate new molecules using Model 2, we used different sampling temperatures*:

- Temperature = 0.5:
  - ~22% wrongly formatted SMILES
  - ~94% new SMILES
  - ~6% equal SMILES

- Temperature = 0.75:
  - % of wrongly formatted SMILES increase
  - % new SMILES increases
  - % of equal SMILES decreases

*Higher sampling temperatures lead to greater structural diversity of the generated molecular structures, but at the same time decrease the fraction of chemically valid SMILES, while lower temperatures lead to lower structural diversity but more conservative predictions.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408

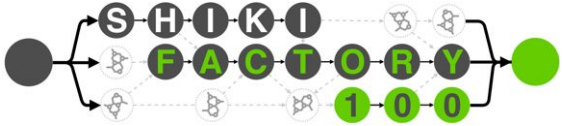
Candidate Molecules

Compounds are generated with the previous generative models

Compounds predicted as sweeteners by the three discriminative models are selected

Compounds predicted non-toxic in 5 QSAR models (mutagenicity, carcinogenicity, developmental toxicity, skin sensitization) are selected as potential sweeteners (Vega software - https://www.vegahub.eu)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408



THANK YOU FOR THE ATTENTION.

QUESTIONS?

AKNOWLEDGMENTS:

MIGUEL ROCHA
BISBII RESEARCH GROUP

University of Minho
School of Engineering

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408