Summary of work

Artur van Bemmelen, period of June-October 2021

Contents

[Jupyter notebooks and output 1](#_Toc86911054)

[Data management 4](#_Toc86911055)

[Input files 4](#_Toc86911056)

[Intermediate files 4](#_Toc86911057)

[Presentations 4](#_Toc86911058)

# Jupyter notebooks and output

Notebooks are located on Github: <https://github.com/iomega/mass-spectral-embeddings-visualization>  
Output files are located on the server: /mnt/LTR\_userdata/hooft001/mass\_spectral\_embeddings/embeddings/ALL\_GNPS\_210409\_positive/

**t-SNE and UMAP comparison.ipynb**

A first glance at UMAP and t-SNE, in particular a t-SNE produced by Joris Louwen. The UMAPs were made with only 15 neighbours. Later work suggests that 50 neighbors is a much better starting point for a dataset of this size and complexity. The best feature of this notebook is the code for “coupled” interactive plots, where two or more interactive plots are controlled with the same legend.

**Instrument type comparison.ipynb**

This notebook takes a look at how spectra with the same planar inchikey that are measured by different instrument types, are embedded in t-SNE. I distinguish between HCD, CID, qToF, and Orbitrap. The embeddings of HCD Velos and CID Velos instruments are strikingly similar in my t-SNE embeddings, but differ in the 21/7 t-SNE embedding by Joris Louwen.

**Jaccard similarity per classification group.ipynb**

To gain an understanding of which groups may be expected to cluster together in embeddings, I computed the internal Jaccard similarity per class and classifier. As was to be expected, classification at subclass level resulted in classes with more internal similarity, whereas internal similarity decreased at superclass level. NPClassifier classes stood out as the classifier and class level where most spectra were classified into “high similarity” classes.

Internal jaccard similarity per classification group.csv

**Jaccard similarity between groups.ipynb**

To complement the notebook ‘Jaccard similarity per classification group’, I computed the similarity between classes of the same classifier to quantify the distinctness of each class. As was expected, the between-class similarity of Classyfire was lower than NPClassifier, as NPClassifier was designed specifically for natural products which would result less distinct classes. The similarity matrices this notebook output could elucidate confusion matrices of similarity-based classification algorithms.

Jaccard similarity between NPClassifier classes.csv  
Jaccard similarity between Classyfire classes.csv

**Spec2Vec similarity per classification group.ipynb**

To quantify how well the parameters for the Spec2Vec embedding process were chosen, and how much of the internal similarity the Spec2Vec embeddings capture, the Spec2Vec similarity of each class and classifier was also computed. The results show that there is much space for improvement, which, while disappointing, are not surprising considering Spec2Vec’s tendency to underestimate the similarity of all compounds with a Jaccard score lower than ~0.8.

Internal spec2vec similarity per classification group.csv

**Compare internal similarities to MS2DeepScore.ipynb**

After determining the internal similarities of various Classyfire classes and NPClassifier classes using Jaccard score and Spec2Vec similarity, the point was raised that it would be very interesting to see how these compare to the supervised embeddings produced by MS2DeepScore. Calculating the embeddings takes a while, so I wrote a script that uses parallel computation to speed up the process. Whereas Spec2Vec drastically underestimates the similarity of all but highly similar compounds, MS2DeepScore slightly overestimates the similarity. Setting the cut-off value for “high similarity” just slightly higher (0.65 instead of 0.6) however, created a nice correlation with the percentage of pairs exceeding 0.6 Jaccard similarity. This was true for both NPClassifier classes, and Classyfire classes, suggesting this threshold might generalize to other datasets.

Calculate\_MS2DS\_embeddings.py (available on Github)  
MS2DeepScore\_embedding\_annotated\_spectra\_210409\_joblib.pickle

**Unsupervised spec2vec UMAP exploration.ipynb**

Spec2Vec-based UMAP exploration of classes larger than 3000 spectra of each of the following classification groups: NPClassifier pathway, NPClassifier superclass, Classyfire superclass, and Classyfire class. It includes minimal working examples for static seaborn plots and interactively plots using plotly.express.

spec2vec\_npc\_pathway\_umap.pickle  
spec2vec\_npc\_superclass\_umap.pickle  
spec2vec\_cf\_superclass\_umap.pickle  
spec2vec\_cf\_class\_umap.pickle

**Unsupervised MS2DS UMAP exploration.ipynb**

MS2DeepScore-based UMAP exploration of classes larger than 3000 spectra of each of the following classification groups: NPClassifier pathway, NPClassifier superclass, Classyfire superclass, and Classyfire class.

ms2ds\_npc\_pathway\_umap.pickle  
ms2ds\_npc\_superclass\_umap.pickle  
ms2ds\_cf\_superclass\_umap.pickle  
ms2ds\_cf\_class\_umap.pickle

**Supervised UMAP – minimal working examples.ipynb**

This notebook contains minimal working examples for a few methods of supervised UMAP clustering. Please note that these results are not meant to be exhaustive, methods that appear to perform poorly may yet yield surprising results with different parameters. The methods outlined in this notebook are: supervised UMAP, semi-supervised UMAP, and tree-based UMAP using ExtraTrees. To illustrate how different tree-based UMAPs can look, even when using the same model, UMAPs of the best and worst ExtraTrees model are shown.

# Data management

## Input files

**Cleaned and processed spectra:** /mnt/LTR\_userdata/hooft001/mass\_spectral\_embeddings/datasets/ALL\_GNPS\_210409\_positive/ALL\_GNPS\_210409\_positive\_cleaned\_peaks\_processed\_s2v.pickle

**Spectra classifications:**/mnt/LTR\_userdata/hooft001/mass\_spectral\_embeddings/classifications/ALL\_GNPS\_210409\_positive/ALL\_GNPS\_210409\_positive\_processed\_annotated\_CF\_NPC\_classes.txt

**Spec2Vec model:**/mnt/LTR\_userdata/hooft001/mass\_spectral\_embeddings/embeddings/ALL\_GNPS\_210409\_positive/ALL\_GNPS\_210409\_positive\_cleaned\_spec2vec\_embedding\_iter\_15

**MS2DS model:**   
https://surfdrive.surf.nl/files/index.php/s/ECykxtumNdnzxcD

## Intermediate files

**GNPS\_210409 Metadata inchi annotated spectra.csv**Tab-seperated values describing the metadata of all spectra that were annotated with an Inchikey. Columns included are ID, inchikey, inchi, instrument, spectrum\_index (which refers to the index of each spectrum in the spectra pickle file), as well as every classification obtained from Classyfire and NPClassifier. The code for this dataframe occurs in multiple notebooks, the first of which is t-SNE and UMAP comparison.ipynb.

## Presentations

20-09-2021 Diving into compound and spectrum similarities  
01-10-2021 Preparing for MetaboNews  
21-10-2021 Spec2vec vs MS2DeepScore UMAPs

**Important:** The downsampled Spec2Vec UMAPs in the “Spec2vec vs MS2DeepScore UMAPs” presentation look very promising, but these turned out to be the **result of a bug**. This bug is explained in the notebook “Unsupervised spec2vec UMAP exploration”.