Structure Based Molecular Fingerprint Prediction through Spec2Vec Embedding of GC-EI-MS Spectra

Aleksander Piciga
aleksander.piciga@gmail.com
Jožef Stefan Institute
Ljubljana, Slovenia

Milka Ljoncheva
milka.ljoncheva@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Tina Kosjek
tina.kosjek@ijs.is
Jožef Stefan Institute
Ljubljana, Slovenia

Sašo Džeroski
saso.dzeroski@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

ABSTRACT

Identifying the molecular structure of unknown organic compounds is a major challenge when dealing with mass spectrometry (MS) data. Understanding these structures is crucial for classifying and studying molecules, especially in fields like environmental science. Research efforts in the recent two decades have resulted in generation of rich MS data, both liquid chromatography (LC)-MS and gas chromatography (GC)-MS data, that can be exploited in exploring the possibilities of machine learning approaches in compound identification.

Our approach aims to predict molecular fingerprints directly from mass spectra. Fingerprint bits correspond to molecular structures and consequently, prediction of these will directly reveal the underlying features of the molecule. Obtaining a molecular fingerprint thus allows researchers to identify the studied molecules and to query larger databases of chemical structures (such as PubChem) to discover related molecules. Ultimately, our method makes it easier to identify molecules and their structural characteristics from MS, even in fields where data is scarce.

KEYWORDS

mass spectra, multi-label, Spec2Vec, prediction, Word2Vec, machine learning, embedding, molecular fingerprint, structure

1 DATA

1.1 Overview

The dataset we study [7] is composed of GC-MS, along with metadata information about the molecules. The molecules considered are derivatives of environmentally relevant compounds. Metadata contains the molecule name, formula, exact mass, PubChem ID, InChI, InChI Key, and SMILES of the trimethysilyl (TMS), derivative along with identical data for the parent compound [9]. PubChem ID is included for the PubChem database, which is one of the largest repositories of molecular entities. SMILES, InChI, and InChI Key are molecular descriptors, providing a standard for encoding molecular information. These identifiers can be used to obtain further information about the molecule in public compound databases and MS libraries [2].

GC-MS spectra show mass to charge ratios (m/z). Each GC-MS spectrum exhibits identifiable spikes called peaks, which hold significant value for compound structure classification, but also correlate to structural information [3].

Mass spectrometry has many different methods which can be employed. The data used in this study (GC-MS spectra) are obtained using electron impact ionization (EI). Gas chromatography involves heating the sample, which must possess volatility and thermal stability. The ionization process, on the other hand, occurs through electron emission. [5].

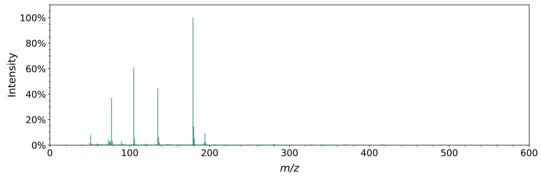

Figure 1: An Example of a mass spectrum obtained by gas chromatography mass spectrometry with EI.

1.2 Dataset

We used spectra produced by the authors (Milka Ljoncheva), which have been made publicly available [7]. These are spectra of TMS derivatives [9]. TMS derivatives are produced by replacing the active hydrogen atom of alcohols, acids, amines, and thiols by a trimethylsilyl group. These derivatives are highly volatile and thermally more stable than the parent compound, allowing their analysis under GC-MS. Fragmentation of these derivatives is also hugely structurally informative [5] [8].

The dataset is available in different formats, including *.mgf*, which is a common format for spectrometry data. These *.mgf* files contain precursor mass, charge, and m/z abundance pairs. Additional metadata is available in Excel files. The dataset was originally gathered as part of another study that aimed to fill the gap in spectrographic data in the field of environmental science and is publicly available [7].

There are a total of 3144 distinct spectra in the dataset, covering 106 unique compounds. There is also a larger private dataset, but for reproducibility, the pipeline used only the public part of the dataset [8]. Each compound in our dataset contained all the required metadata information and was represented by approximately 30 independent spectra. The distribution of the number of spectra per molecule is shown in the Figure 2 (*mean 30, min 3, max 60, std 6.85*). On average molecules have 34.6 positive labels.

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

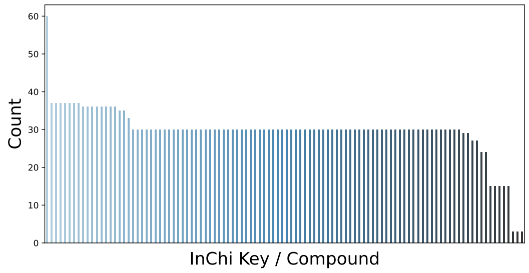

Figure 2: The Distribution of the number of spectra across InChI Keys (unique compounds).

2 PREPROCESSING

2.1 CG-MS Spectra

We used matchms package to refine the metadata and spectra representations. The matchms package is a publicly available Python package to import, process, clean, and compare mass spectrometry data. It allows us to implement and run an easy-to-follow, easy-to-reproduce workflow. There were two main phases in the preprocessing workflow [4]:

- metadata enrichment and
- spectrum standardization.

In the metadata prepossessing phase, we extracted valuable information like the InChI Key and molecule name from the *.mgf* files, which often contained both pieces of data. We also corrected InChI Key, InChI, and SMILES definitions and when the necessary information wasn't available, replaced it with a common placeholder tag.

On the data side, our efforts included adding parent mass, normalizing intensities, reducing the number of peaks to a range of 10 to 500, setting intensity thresholds between 0 and 1000, and deriving losses. We also required that each GC-MS spectrum contain not less than 10 peaks. These steps were crucial for getting the CG-MS spectral data ready for analysis and for removing any potentially corrupted spectra [4]. An example of the effects that processing the mass spectra peaks can have is shown in Figure 3.

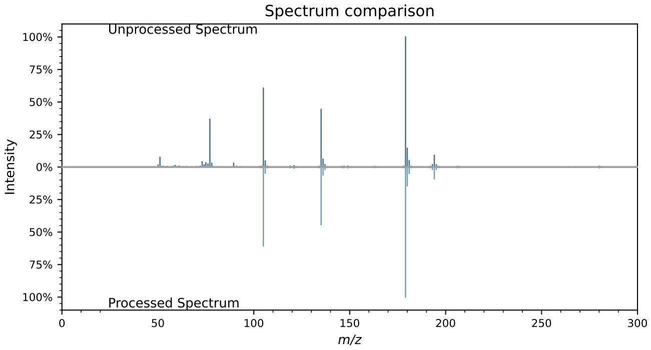

Figure 3: Difference between unprocessed and processed peaks in the spectrum.

2.2 Molecular fingerprints

Our pipeline enables the generation of common molecule fingerprints, given the molecule's InChI or InChI Keys by making queries to public APIs. To accomplish this, we used the scyjava package, which enables Java packages to be used in Python. This is convenient since our entire workflow is built in Python and we need to access the Chemistry Development Kit (CDK) written in Java. Within this framework, we've implemented a subset of molecular fingerprints which we tested in the study, that included the following molecular fingerprints: [11]:

- AtomPairs2D,
- Circular,
- EState,
- Extended,
- KlekotaRoth,
- Lingo,
- MACCS,
- Pubchem,

For our sample study, we selected the MACCS molecular fingerprint. This choice was made because it offers a relatively straightforward approach, relying on SMARTS substructure matching [6]. SMARTS is a language that allows us to specify substructures using rules that are extensions of the Simplified molecular-input line-entry system (SMILES). The Molecular fingerprint is then defined by a set of these SMARTS patterns. MACCS uses 166 patterns [6].

Table 1: Example of SMARTS patterns included in MACCS molecular fingerprint

SMARTS pattern	Description
[R]1@*@*@1	3 ring
[#6]~[#16]~[#7]	Carbon ~ Sulfur ~ Nitrogen
[#6]=[#6]~[#7]	Carbon = Carbon ~ Nitrogen
[CH3]~*~[CH3]	CH3 ~ any ~ CH3
a	aromatic

~ represents any bond type.
= represents a double bond.
definitions from [10]
more detailed definition of the language is available at
https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

2.3 Spec2Vec

Spec2Vec [3] is a spectral similarity score inspired by Word2Vec. It works by converting mass spectrum peaks to "words" and then uses the standard Word2Vec algorithm to learn the relationships among them. It is an unsupervised algorithm so the evaluation can be performed on the same data used to train Spec2Vec models. There are large pretrained models which are publicly available, but custom models can be quite inexpensive to train on local data. The model was trained specifically for TMS derivatives from the public dataset. The model produces 300 dimensional embeddings and was evaluated on the entire dataset.

Spec2Vec embeddings outperform traditional methods of comparing spectra, such as cosine similarity, and even modified versions that consider data noise. These embeddings also exhibit a much better correlation between high similarity scores and high structural similarity [3]. However, the structure cannot be directly derived from latent space embedding, which is why we employ machine learning to learn these structural characteristics [3].

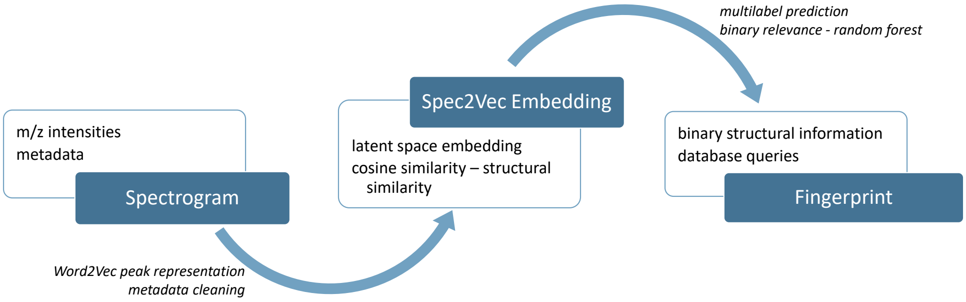

Figure 4: Overview of the prediction pipeline

3 PIPELINE

Our main goal is to predict molecular fingerprints that represent structural information based on the mass spectra embeddings following the workflow diagram presented in 4. Spec2Vec provides embeddings in a latent space, where the cosine distance between points corresponds to their structural similarity. The molecular fingerprint generation task is framed as a multi-label classification because each instance or example can exhibit multiple identifiable structural characteristics, and these correspond to multiple different bits in the fingerprint. These structural components have correlations among them, which is another reason to treat the problem as multi-label classification rather than just multi-class classification.

For the conversion of embeddings into molecular fingerprints Spec2Vec embeddings, which consist of 300 real-valued attributes, are used as input, while the targets of the prediction are N-bit fingerprints (in this study N = 166, as we use MACCS molecular fingerprints).

4 METHODS

Multi-label classification (MLC) can be approached in many different ways. The most straightforward approach involves treating each label independently and training a separate binary classifier for each label (Binary Relevance). Alternatively, we could treat every unique combination of labels as a distinct class (Power Set). However, given our 166 labels, the latter approach would create a large number of classes, especially if we extend our research to a broader range of molecules. We chose One Vs Rest classifier (OVR) from sklearn, which works like Binary Relevance when provided with an indicator matrix for the target (y) values. Binary Relevance trains a separate estimator for each of the target indicator labels [1].

We need to choose an approach for classification since we have reduced the MLC task into multiple binary classifications. Random Forests are used due to their empirically proven high accuracy [1], ability to handle imbalanced data, and good bias variance trade-off. Other models, such as Decision Trees and Logistic Regression were also quickly tested and proved worse in preliminary testing with double 5-fold validation as shown in the Table 2. Worse performance and efficiency of these models are known from the literature [1].

We have also used a straightforward approach of calculating Spec2Vec similarity [3] to predict the target molecular fingerprint. First, the Spec2Vec embedding is constructed for known molecules and is stored along with their fingerprints. When predicting for a new molecule its Spec2Vec embedding is calculated. The embedding of the new molecule is compared to known embeddings using built in function that calculates similarity score based on cosine similarity. Voting for fingerprint labels is then done proportionally based on similarity score. This approach, which corresponds to the weighted nearest neighbor, is further discussed in the section 5.

Table 2: Initial Comparison of Internal Estimators

	Logistic Regression	Random Forest	Decision Tree
Hamming Loss	0.045	**0.043**	0.067
Weighted F1 Score	**0.895**	0.854	0.837
Label Ranking Loss	0.016	**0.010**	0.182
Coverage Error	54.601	**42.964**	151.832

5 EVALUATION

We evaluated the learning methods using various metrics, with a focus on the most informative ones, such as hamming loss, label ranking loss, weighted F1 score, and coverage error [1], results of these evaluations are shown in Table 3. To ensure robust evaluation, we employed a 5-fold cross-validation approach, which we repeated twice to obtain reliable performance measurements.

Table 3: Random Forest performance metrics

	Default Classifier	Similarity Voting	Random Forest
Hamming Loss	0.083	0.038	0.043
Weighted F1 Score	0.635	0.642	0.854
Label Ranking Loss	0.630	0.083	0.010
Coverage Error	166.000	64.794	42.964

The Default Classifier always predicts the majority class for each label.

Similarity Voting uses Spec2Vec similarity to proportionally vote for labels. This approach is presented as a stronger baseline from which we can measure improvements of our models.

Random Forests were trained for each label, using One Vs Rest (OVR) method. Each forest had 100 estimators with balanced class weights (inversely proportional). Impurity was measured using Gini Impurity measure and no other restricting parameters were set - the defaults of sklearn Random Forest Classifier apply.

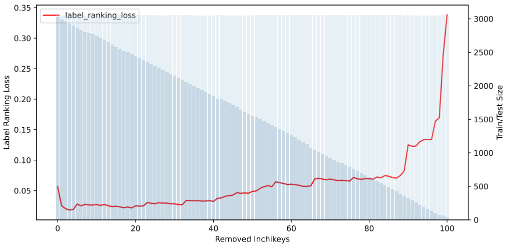

Figure 5: Models ability to generalize to unseen InChI Keys.

Our goal isn't predicting fingerprints for known molecules, but handling new ones effectively. To test this, we deliberately removed some InChI Keys from our dataset. By doing this, we checked how well our models perform in predicting the structures of these unfamiliar molecules. This real-world scenario testing helps us understand how practical and effective our approach is when dealing with novel compounds not present in our initial training data.

We have also performed 10-fold validation by removing 10 InChI Keys at a time from the training data. The model was trained on the remaining ~90 InChI Keys (~2700 samples of mass spectra) and evaluated on ~10 unseen ones (~300 samples of mass spectra). The results are shown in Table 5. The Random Forests' ability to predict larger amounts of unseen InChI Keys and effects of less training data and therefore less diverse embedding knowledge is shown in Figure 5. Even though the label ranking loss is increasing it is still well below the loss of the Default Classifier and even Similarity Voting, when a large amount of InChI Keys are missing and the training dataset is smaller.

Table 4: Similarity Voting on Unseen InChI Keys

	Hamming Loss	Weighted F1 Score	Label Ranking Loss	Coverage Error
average	**0.047**	**0.639**	**0.084**	**75.153**

Here only the average is shown to provide a reference point for the quality of Random Forests. More data was not included to not clutter the article. Unseen InChI Keys were simulated by keeping only the test rows (unseen InChI Keys) and train columns (other InChI Keys) in the similarity matrix.

Table 5: 10-fold evaluation results for unseen InChI Keys, Results per Fold

	Hamming Loss	Weighted F1 Score	Label Ranking Loss	Coverage Error
0	0.068	0.749	0.043	63.432
1	0.064	0.806	0.039	85.369
2	0.061	0.775	0.045	94.405
3	0.066	0.757	0.031	70.266
4	0.060	0.759	0.033	79.687
5	0.101	0.676	0.066	97.522
6	0.124	0.596	0.077	115.793
7	0.036	0.864	0.019	63.857
8	0.047	0.818	0.017	64.828
9	0.077	0.721	0.063	84.503
average	**0.070**	**0.752**	**0.043**	**81.966**

6 REPRODUCIBILITY

The whole pipeline and evaluation were built with repeatability in mind to allow for future studies, model comparisons, and reevaluation of results. The dataset used is public, Spec2Vec models are built upon these data, and model training functions along with parameters are available in the repository github.com/alpi314/mass_spectra tagged *article*. Training of the models is done with fixed random seeds and stores models with training parameters, train and test data with the use of the pickle package. Metrics and evaluations are always stored along with the models.

7 CONCLUSION

Our results demonstrate that Spec2Vec embeddings of TMS can effectively be converted into molecular fingerprints using machine learning methods. These methods have proven to be reliable even when predicting molecular structures for molecules that have not been encountered before. This is significant because it allows processing new MS spectra to uncover their most likely structural components, which we can then match against databases. This structural information can be directly applied in various research studies. Our plans for future work involve expanding this approach to larger compound databases. Additionally, we plan to broaden our research to predict more SMARTS patterns as part of expanding our molecular fingerprint prediction capabilities. While we'll stay focused on fingerprints for database queries, we will be also looking into predicting arbitrary SMARTS patterns.

REFERENCES

[1] Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. 2022. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203, 117215. DOI: 10.1016/j.eswa.2022.117215.

[2] Juliane Glüge, Kristopher McNeill, and Martin Scheringer. 2023. Getting the SMILES right: identifying inconsistent chemical identities in the ECHA database, PubChem and the CompTox Chemicals Dashboard. *Environmental Science: Advances*, 2, 4, 614. DOI: 10.1039/D2VA00225F.

[3] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. van der Hooft. 2021. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*. DOI: 10.1371/journal.pcbi.1008724.

[4] Florian Huber, Stefan Verhoeven, Christiaan Meijer, and Hanno Spreeuw. 2020. matchms - processing and similarity evaluation of mass spectrometry data. *Journal of Open Source Software*, 5, 2411. DOI: 10.21105/joss.02411.

[5] Rontani Jean-Francois. 2022. Use of Gas Chromatography-Mass Spectrometry Techniques (GC-MS, GC-MS/MS and GC-QTOF) for the Characterization of Photooxidation and Autoxidation Products of Lipids of Autotrophic Organisms in Environmental Samples. *Molecules*, 27, 5. DOI: 10.3390/molecules27051629.

[6] Hiroyuki Kuwahara and Xin Gao. 2021. Analysis of the effects of related fingerprints on molecular similarity using an eigenvalue entropy approach. *Journal of Cheminformatics*, 13, 1, 27. DOI: 10.1186/s13321-021-00506-2.

[7] Milka Ljoncheva, Tina Kosjek, Sašo Džeroski, and Sintija Stevanoska. 2023. GC-EI-MS datasets of trimethylsilyl (TMS) and tert-butyl dimethylsilyl (TBDMS) derivatives. *Mendeley Data*. DOI: 10.17632/j3z5bmvmnd.6.

[8] Milka Ljoncheva, Tomaž Stepišnik, Tina Kosjek, and Sašo Džeroski. 2022. Machine learning for identification of silylated derivatives from mass spectra. *Journal of Cheminformatics*, 14, 1, 62. DOI: 10.1186/s13321-022-00636-1.

[9] Milka Ljoncheva, Sintija Stevanoska, Tina Kosjek, and Sašo Džeroski. 2023. GC-EI-MS datasets of trimethylsilyl (TMS) and tert-butyl dimethyl silyl (TBDMS) derivatives for development of machine learning-based compound identification approaches. *Data in Brief*, 48, 109138. DOI: 10.1016/j.dib.2023.109138.

[10] 2013. RDkit MACCS Keys. Accessed on 2023-08-31. (2013). https://github.com/rdkit/rdkit-orig/blob/master/rdkit/Chem/MACCSkeys.py.

[11] Egon L. Willighagen et al. 2017. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics*, 9, 1, 33. DOI: 10.1186/s13321-017-0220-4.