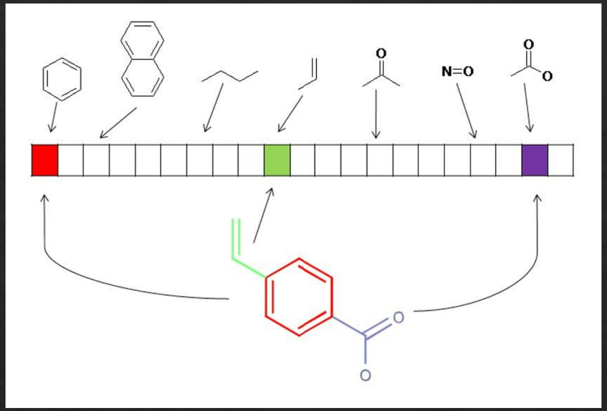
Structure Based Molecular Fingerprint Prediction through Spec2Vec Embedding of GC-EI-MS Spectra

Aleksander Piciga, Milka Ljoncheva, Tina Kosjek, Sašo Džeroski @ IJS

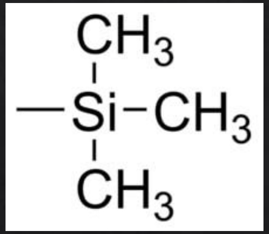
Jožef
Stefan
Institute

Terminology

- Molecular Fingerprint (MACCS – 166 patterns)



- TMS - Trimethysilyl



- SMARTS

  [R]1@*@*@1 = 3 ring

- InChI Key

  InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3

- SMILES

  OCC

- Gas chromatography (GC) mass spectrometry (MS) obtained by electron impact ionization (EI)

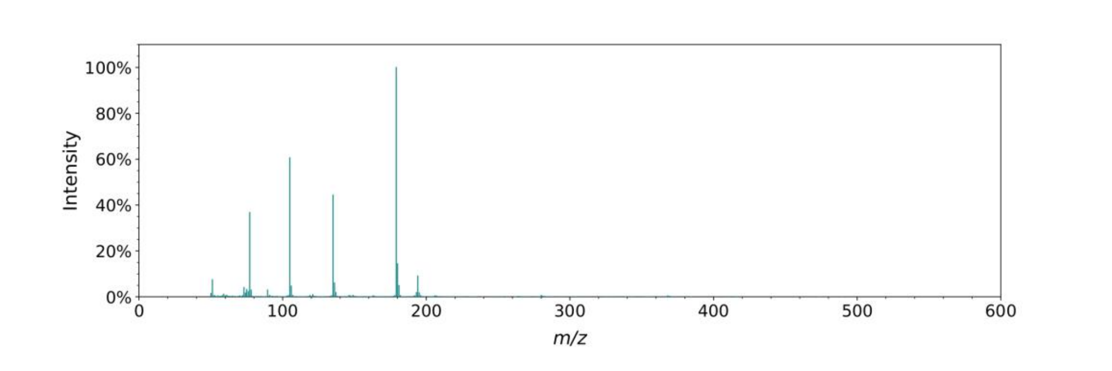
Structure

Spectrum peaks are significant for structure classification

Correlate to structural information

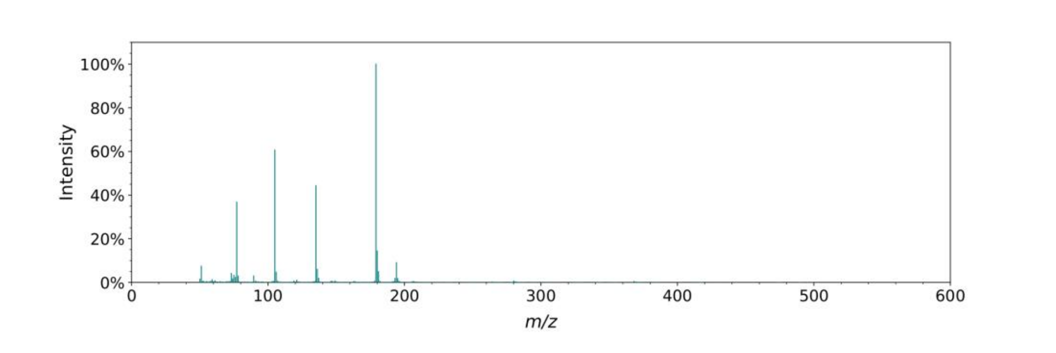
Important for:

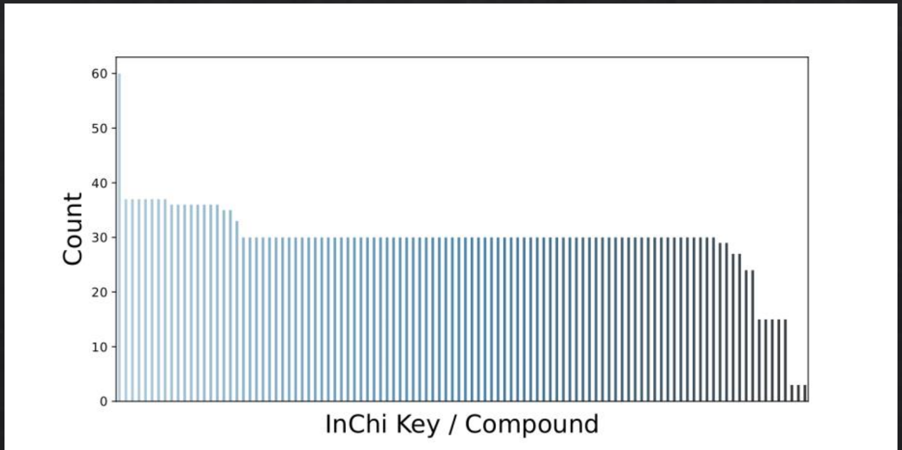
- Identification of Compounds
- Environmental Analysis
- Forensic Science
- Database Querying
- Molecular Properties



Dataset

- Mass Spectra
- TMS Molecules
- Publicly available
- 3144 distinct spectra
- 106 unique compounds

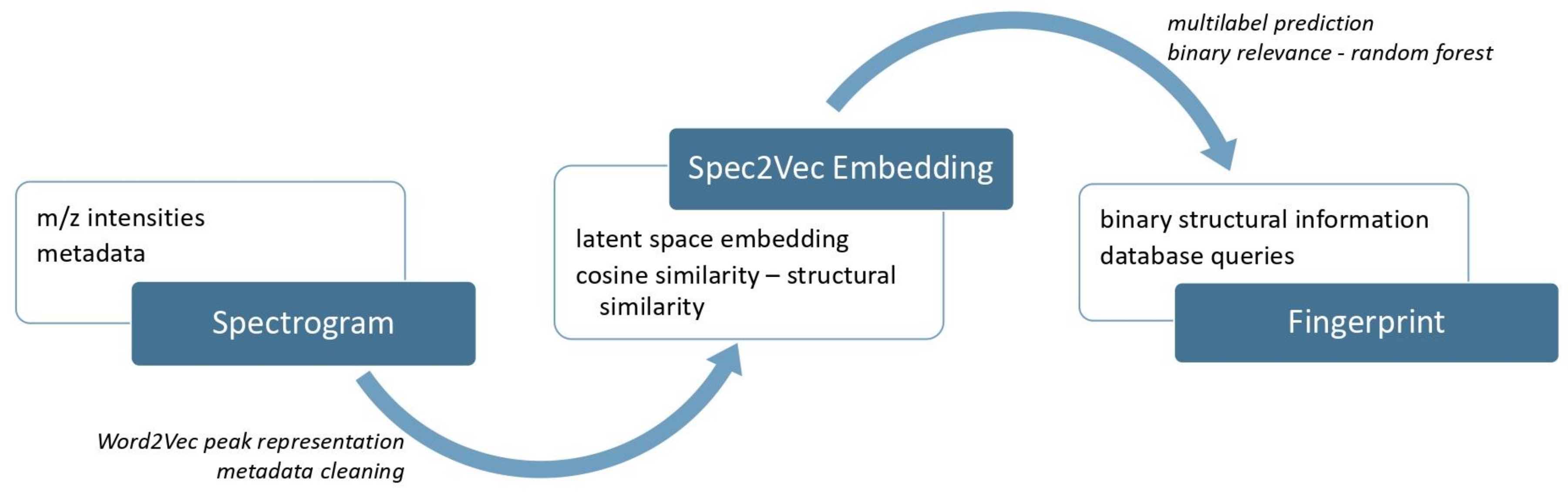




Spec2Vec

- Inspired by Word2Vec
- Peaks = “Words”
- Learns relationships among spectrum peaks
- Easy and inexpensive to train
- 300 dimensional embedding with locally trained model
- Embeddings do not directly reveal structure

Pipeline



Spectrogram

m/z intensities
metadata

Word2Vec peak representation
metadata cleaning

Spec2Vec Embedding

latent space embedding
cosine similarity – structural similarity

multilabel prediction
binary relevance - random forest

Fingerprint

binary structural information
database queries

Preprocessing

Metadata enrichment and correction

- InChI Key
- Molecule names
- SMILES definitions
- placeholders

Spectrum standardization

- Normalization
- Constrained number of peaks
- Remove spectra with too little significant peaks

Spec2Vec Embdedings

- Train local model on the data
- Embed the data
- Latent Space embeddings (300 dimensional)
  - cosine similarity ~ structural similarity

Multilabel classification (MLC)

- 300 attributes (Spec2Vec embedding)
- 106 targets (MACCS fingerprint) ~ structure
- Binary Relevance (n binary classifiers) / Power Set (2^n classes)
- Random Forest (with OVR)

Evaluations and Baseline

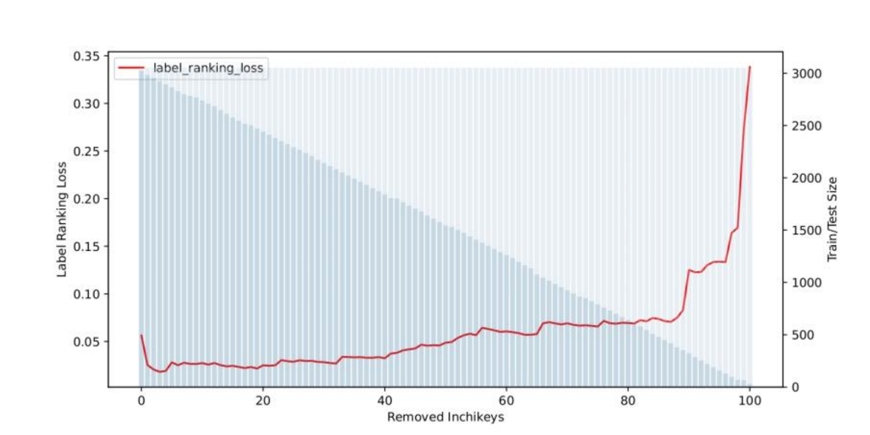
- Baseline
  - Default Classifier
  - Similarity Voting
- Evaluation
  - Tracked many metrics
  - Hamming loss, label ranking loss, weighted F1 score, coverage error
  - Prediction of unknown/unseen molecules (!)

Results

2 times repeated 5-fold on all data (compounds in test set can also appear in train set)

	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**

10-fold by removing 10% of compounds (evaluation on unseen compounds)



	Similarity Voting	Random Forest
Hamming Loss	0.047	0.070
Weighted F1 Score	0.639	0.752
Label Ranking Loss	0.084	0.43
Coverage Error	75.153	81.966

Future Work

- Evaluating the approach on larger public databases
- Prediction of other fingerprints
- Prediction of arbitrary SMARTS patterns
- Evaluation of other (more complex) ML techniques
- Comparison of publicly available Spec2Vec models trained on larger datasets

Conclusion

- Importance of molecular structure
- Molecular Fingerprints ~ Molecular Structure
- Spec2Vec embeddings for better structural correlation
- ML approach for predicting molecular fingerprints

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

https://github.com/al-pi314/mass_spectra/tree/article

Aleksander Piciga, Milka Ljoncheva, Tina Kosjek, Sašo Džeroski @ IJS

Jožef
Stefan
Institute