TO BIN OR NOT TO BIN: ALTERNATIVE REPRESENTATIONS OF MASS SPECTRA

Niek de Jonge, Justin J. J. van der Hooft,* Daniel Probst
Bioinformatics Group
University & Research Wageningen
The Netherlands
{niek.dejonge,justin.vanderhooft,daniel.probst}@wur.nl

1 INTRODUCTION

Mass spectrometry, especially so-called tandem mass spectrometry, is commonly used to assess the chemical diversity of samples. The resulting mass fragmentation spectra are representations of molecules of which the structure may have not been determined. This poses the challenge of experimentally determining or computationally predicting molecular structures from mass spectra. An alternative option is to predict molecular properties or molecular similarity directly from spectra. Various methodologies have been proposed to embed mass spectra for further use in machine learning tasks. However, these methodologies require preprocessing of the spectra, which often includes binning or sub-sampling peaks with the main reasoning of creating uniform vector sizes and removing noise. Here, we investigate two alternatives to the binning of mass spectra before downstream machine learning tasks, namely, set-based and graph-based representations. Comparing the two proposed representations to train a set transformer and a graph neural network on a regression task, respectively, we show that they both perform substantially better than a multilayer perceptron trained on binned data.

2 RELATED WORK

Machine learning models commonly take binned mass spectra as input, discretise peaks as part of a vocabulary through tokenisation, or select peaks with top-n intentsities. Examples are MS2DeepScore by Huber et al. (2021b), Spec2Vec by Huber et al. (2021a), MS2Mol by Butler et al. (2023), MSBert by Zhang et al. (2024), and DreaMS by Bushuiev et al. (2025). Meanwhile, models targeting the reverse problem, predicting mass spectra from molecular graphs, often apply graph neural networks or graph transformers to the input molecular graphs. Recent examples include MassFormer by Young et al. and others (Zhang et al., 2022; Murphy et al., 2023; Park et al., 2024). Recently, Nallapareddy et al. have represented sequential mRNA codons as graphs, inspiring us to view a mass spectrum as a sequence of intensities along the mass-to-charge ratio dimension. Furthermore, applying graph neural networks to time series has become a commonly used approach, hinting at the potential of GNNs applied to mass spectra (Jin et al., 2024).

The representation of mass spectra as sets is based on a recent publication on set representation learning for molecules by Boulougouri et al. (2024) and the work by Goldman et al. (2023). As peaks in mass spectra are pairs of intensities and m/z values of unequal numbers across spectra, a mass spectrum can be viewed as a set of intensity-m/z value pairs.

3 METHODOLOGY

Data and splits provided by de Jonge et al. (2025) were used for the experiments. The quantitative estimate of drug-likeness (QED) values were calculated using the RDKit implementation of the metric (Landrum). While the data set provides additional metadata, only the m/z (including the precursor) and intensity values were used for all subsequent steps. Further processing depended on the downstream architecture. For the multilayer perceptron (MLP), the m/z peaks were binned following the protocol provided by Huber et al. (2021b), changing the range of bins from 10 to 10,000

*Department of Biochemistry, University of Johannesburg, South Africa

to 0 to 10,000) to have consistency across experiments. For the SetTransformer, m/z–intensity pairs were encoded as a PyTorch TensorDataset. Finally, the data was encoded as set of graphs for the GNN: Each peak in a mass spectrum is represented by a vertex and connected to neighboring peaks through edges; the intensities were used as vertex attributes and the differences in m/z as edge attributes. An additional vertex with m/z and intensity of 0 was created to encode the m/z delta of the initial peak. For all experiments, intensities were normalized and the precursor m/z was treated as a normal peak with an intensity of 2.0.

To evaluate the binned data, an MLP with 2 hidden layers (1,024 and 512 neurons, respectively) with ReLU activation and a dropout of 0.5 was used. The SetTransformer, trained on the sets of m/z–intensity pairs, was configured with two hidden layers (32 and 16 neurons, respectively). Finally, the graph neural network, a graph attention network (GAT), was trained with 8 message passing layers, 1,024 hidden channels, and global mean pooling. The hyperparameters of the GAT were chosen to result in a similar number of parameters to the MLP and not further optimized.

4 Preliminary Results

We compare the three different methods using binned, set, and graph representations of mass spectra as input on the regression task of predicting the QED of a molecule from its mass spectra (Bickerton et al., 2012). As QED is a broad composite descriptor capturing both molecular structure and biological function, it serves as an effective surrogate for assessing the proficiency of a methodology in predicting molecular properties from mass spectrometry data through regression analysis.

Table 1: Performances of models trained on different mass spectrum representations. The MLP was trained on binned mass spectra, the SetTransformer on sets of m/z-intensity pairs, and the GNN (GAT) on graphs where the vertices represent peaks (intensity), and the edges distances (m/z) between peaks. Over all metrics, the GNN-based approach performs best.

Model	Params	MAE (↓)	RMSE (↓)	Pearson's r (↑)	R^2 (↑)
MLP	11.0 M	0.145 ± 0.008	0.200 ± 0.008	0.736 ± 0.011	0.437 ± 0.043
SetTransformer	0.4 M	0.134 ± 0.002	0.174 ± 0.001	0.758 ± 0.004	0.572 ± 0.006
GNN (GAT)	12.6 M	$\mathbf{0.110 \pm 0.006}$	$\mathbf{0.144 \pm 0.006}$	$\mathbf{0.843 \pm 0.015}$	$\mathbf{0.709 \pm 0.025}$

Our experiments, with results shown in Table 1, yielded three primary results: (1) Both the set representation and graph representation with their related architectures performed better than the binned representation-trained MLP. (2) The set representation model is highly efficient with only 400 k parameters, compared to 11.0 M and 12.6 M parameters of the other two models. (3) Representing mass spectra as graphs results in better performance than both binning and set representation. Given the performance increase of the GNN compared to the set-based approach, both of which use all available data without binning or sub-sampling, hints at the possibility that graphs are a favorable representation of mass spectra, as meaningful information may be propagated along the graph. As we chose a relative simple regression task, further research is needed to assess the proposed architectures on other tasks that are common in metabolomics research, including similarity prediction or molecular structure prediction.

5 Conclusion

We showed that representing mass spectra as sets or graphs is not only possible but performs better in our prospective study compared to the commonly used approach of representing mass spectra as fixed-length arrays by binning or sub-sampling peaks. As both established and recent machine learning approaches rely on binning or sub-sampling of mass spectra (Huber et al., 2021b; Bushuiev et al., 2025), we believe that our results are of immediate interest for machine learning researchers developing new architectures applied to mass spectra data. Furthermore, we provide ready-to-use encoders and models that can easily be integrated in existing machine learning architectures used in mass spectra-related tasks. Finally, the code and the scripts to get the data used in this study can be found in the following repository: `https://github.com/daenuprobst/mas2graph`.

MEANINGFULNESS STATEMENT

Hand-in-hand, the natural sciences and computer science have come a long way in finding ingenious and powerful ways to represent chemical and biological data. We believe that meaningful representations of life should serve as a bridge between experimental and computational research, supporting both fields in making new discoveries and developing new methodologies. This means that, in essence, meaningful representations of life should be generalizable enough to represent new experimental discoveries, but also simple enough to act as a basis for existing and new computational methodologies. We believe that the work we present in this article fits this definition well by reducing the amount of required preprocessing of experimental data, while also providing methods and basic implementations on which current and future machine learning architectures can build on.

REFERENCES

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, February 2012. ISSN 1755-4349. doi: 10.1038/nchem.1243.

Maria Boulougouri, Pierre Vandergheynst, and Daniel Probst. Molecular set representation learning. *Nature Machine Intelligence*, 6(7):754–763, July 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00856-0.

Roman Bushuiev, Anton Bushuiev, Raman Samusevich, Corinna Brungs, Josef Sivic, and Tomáš Pluskal. Emergence of molecular structures from repository-scale self-supervised learning on tandem mass spectra, January 2025.

Thomas Butler, Abraham Frandsen, Rose Lightheart, Brian Bargh, Thomas Kerby, Kiana West, Joseph Davison, James Taylor, Christoph Krettler, T. J. Bollerman, Gennady Voronov, Kevin Moon, Tobias Kind, Pieter Dorrestein, August Allen, Viswa Colluru, and David Healey. MS2Mol: A transformer model for illuminating dark chemical space from mass spectra, September 2023.

Niek F. de Jonge, Elena Chekmeneva, Robin Schmid, David Joas, Lem-Joe Truong, Justin J. J. van der Hooft, and Florian Huber. Bridging polarities in metabolomics: Cross-ionization mode chemical similarity prediction between tandem mass spectra, January 2025.

Samuel Goldman, Jeremy Wohlwend, Martin Stražar, Guy Haroush, Ramnik J. Xavier, and Connor W. Coley. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence*, 5(9):965–979, August 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00708-3. URL http://dx.doi.org/10.1038/s42256-023-00708-3.

Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. van der Hooft. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):e1008724, February 2021a. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008724.

Florian Huber, Sven van der Burg, Justin J. J. van der Hooft, and Lars Ridder. MS2DeepScore: A novel deep learning similarity measure to compare tandem mass spectra. *Journal of Cheminformatics*, 13(1):84, October 2021b. ISSN 1758-2946. doi: 10.1186/s13321-021-00558-4.

Ming Jin, Huan Yee Koh, Qingsong Wen, Daniele Zambon, Cesare Alippi, Geoffrey I. Webb, Irwin King, and Shirui Pan. A Survey on Graph Neural Networks for Time Series: Forecasting, Classification, Imputation, and Anomaly Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10466–10485, December 2024. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3443141.

Greg Landrum. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling.

Michael Murphy, Stefanie Jegelka, Ernest Fraenkel, Tobias Kind, David Healey, and Thomas Butler. Efficiently predicting high resolution mass spectra with graph neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pp. 25549–25562, Honolulu, Hawaii, USA, July 2023. JMLR.org.

Mohan Vamsi Nallapareddy, Francesco Craighero, Cédric Gobet, Felix Naef, and Pierre Vandergheynst. Towards improving full-length ribosome density prediction by bridging sequence and graph-based representations.

Jiwon Park, Jeonghee Jo, and Sungroh Yoon. Mass spectra prediction with structural motif-based graph neural networks. *Scientific Reports*, 14(1):1400, January 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-51760-x.

Adamo Young, Hannes Röst, and Bo Wang. Tandem mass spectrum prediction for small molecules using graph transformers. 6(4):404–416. ISSN 2522-5839. doi: 10.1038/s42256-024-00816-8. URL https://www.nature.com/articles/s42256-024-00816-8.

Baojie Zhang, Jun Zhang, Yi Xia, Peng Chen, and Bing Wang. Prediction of electron ionization mass spectra based on graph convolutional networks. *International Journal of Mass Spectrometry*, 475: 116817, May 2022. ISSN 1387-3806. doi: 10.1016/j.ijms.2022.116817.

Hailiang Zhang, Qiong Yang, Ting Xie, Yue Wang, Zhimin Zhang, and Hongmei Lu. MSBERT: Embedding Tandem Mass Spectra into Chemically Rational Space by Mask Learning and Contrastive Learning. *Analytical Chemistry*, 96(42):16599–16608, October 2024. ISSN 0003-2700. doi: 10.1021/acs.analchem.4c02426.