Reliable cross-ion mode chemical similarity prediction between MS^2 spectra

Authors:

Niek de Jonge[1], David Joas[2], Lem-Joe Truong[2], Justin J.J. van der Hooft[1,3*], Florian Huber[2,*]

[1] Bioinformatics Group, Wageningen University & Research, 6708 PB, Wageningen, the Netherlands
[2] Centre for Digitalisation and Digitality (ZDD), University of Applied Sciences Düsseldorf, Düsseldorf, Germany
[3] Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg, 2006, South Africa

* These authors jointly supervised this work.

Abstract

Mass spectrometry is commonly used to characterize metabolites in untargeted metabolomics. This can be done in positive and negative ionization mode, a choice typically guided by the fraction of metabolites a researcher is interested in. During analysis, mass spectral comparisons are widely used to enable annotation through reference libraries and to facilitate data organization through networking. However, until now, such comparisons between mass spectra were restricted to mass spectra of the same ionization mode, as the two modes generally result in very distinct fragmentation spectra. To overcome this barrier, here, we have implemented a machine learning model that can predict chemical similarity between spectra of different ionization modes. Hence, our new MS2DeepScore 2.0 model facilitates the seamless integration of positive and negative ionization mode mass spectra into one analysis pipeline. This creates entirely new options for data exploration, such as mass spectral library searching of negative ion mode spectra in positive ion mode libraries or cross-ionization mode molecular networking. Furthermore, to improve the reliability of predictions and better cope with unseen data, we have implemented a method to estimate the quality of prediction. This will help to avoid false predictions on spectra with low information content or spectra that substantially differ from the training data. We anticipate that the MS2DeepScore 2.0 model will extend our current capabilities in organizing and annotating untargeted metabolomics profiles.

Introduction

Mass spectrometry is one of the main methods used to map the chemical contents of natural extracts and other biological mixtures during metabolomics workflows. In untargeted metabolomics, tandem mass spectrometry (or mass spectrometry fragmentation, MS/MS, MS^2) is typically used to support structural annotation of metabolite features detected in untargeted metabolomics profiles. Interpretation of tandem mass spectra is increasingly done with the help of computational tools that assist with structurally annotating mass spectra, such as SIRIUS[1] and MS-Finder[1]. Mass spectral similarity scores, like the cosine score, modified cosine score[2], Spec2Vec[3], MS2DeepScore[4], MS2Query[5], and others[6, 7], play a crucial role in common methods like library matching, analogue searching, and organizing spectra by molecular networking.

Mass spectrometry can be performed in two ionization (ion) modes: positive and negative. How suitable a particular ion mode is for detecting a metabolite, largely depends on the metabolite's structure[8, 9]. Consequently, samples are often measured with runs in both ion modes to cover a larger fraction of the metabolome. Mass fragmentation spectra recorded in positive or negative mode typically result in very different fragmentation patterns for the same molecule. Therefore, common comparison metrics like the cosine score are not suitable for predicting chemical similarity between spectra of two different ion modes. As a result, positive and negative ion mode mass spectra are mostly analyzed separately, for instance, by searching in separate reference libraries and creating two separate molecular networks[10, 11, 12]. Where approaches like MolNotator[13] and Ion Identity Networking[14] can merge positive and negative mode spectra into one network, they require adduct identification based on well-aligned retention times and the recognition of specific mass differences between mass features. Achieving retention time alignment can be cumbersome and necessitates using the same chromatography column for both positive and negative ion modes. If these analyses could be integrated by using a similarity metric that works cross-ion mode, this would enable streamlined computational metabolomics workflows that seamlessly integrate both ion modes into one molecular network. Furthermore, a cross-ion mode mass spectral similarity score would allow us to use the larger positive ion mode reference spectral library as a source for annotations in negative ion mode data.

Here, we developed a mass spectral similarity metric that can reliably predict chemical similarity between mass spectra of different ion modes. The approach for this new similarity metric is based on the Siamese neural network architecture proposed in the previous version of MS2Deepscore[4]. The original MS2Deepscore model was trained to predict the chemical similarity between two compounds solely based on the mass fragment information of two respective MS^2 spectra, using mass spectra from a single ion mode.

The original MS2Deepscore model was able to predict chemical similarities with good overall accuracy, but it has several shortcomings. First, separate models had to be trained for positive and negative ion mode data which meant less training data for each model and no cross-ion mode applications. Secondly, the former MS2DeepScore models were trained on MS^2 fragments only; however, spectral metadata like precursor *m/z*, ion mode, or the way of acquisition could be valuable information for improving prediction quality. In the present work, we explore and evaluate the addition of metadata to the model input and show that using ion mode and precursor *m/z* as input improves model performance. Thirdly, we introduce a method that can estimate the mass spectral embedding quality for each input spectrum as well as an estimate of the expected absolute prediction error for each Tanimoto score prediction. This allows users to filter out spectra, or pairs of spectra, for which the MS2Deepscore predictions are unreliable, e.g., due to low spectrum quality or when spectra differ substantially from the training data. This further improves the reliability of MS2Deepscore results.

In addition to the above-mentioned key aspects, this work also contains many technical improvements on the MS2DeepScore code and hyperparameters which lead to better predictions and much shorter runtimes both for model training and chemical similarity prediction. We also added a training pipeline that makes training new models easier, more streamlined, and more robust.

We show that our model trained on both ion modes still works equally well at predicting chemical similarity between spectra measured in the same ion modes when compared to single-ion mode models. In addition, however, the dual-ion mode model is also capable of predicting chemical similarities between spectra of the two different ion modes with high accuracy. We anticipate that this will create entirely new ways of doing molecular networking by creating similarity-based graphs that can contain both positive and negative ion mode spectra. Furthermore, MS2DeepScore 2.0 can also be used to compare negative ion mode spectra with a positive ion mode library (or vice versa), thereby greatly expanding the reference library spectra available for negative ion mode. We therefore expect that our new cross-ion mode model will help to enhance what we can learn from untargeted mass spectrometry data.

Methods

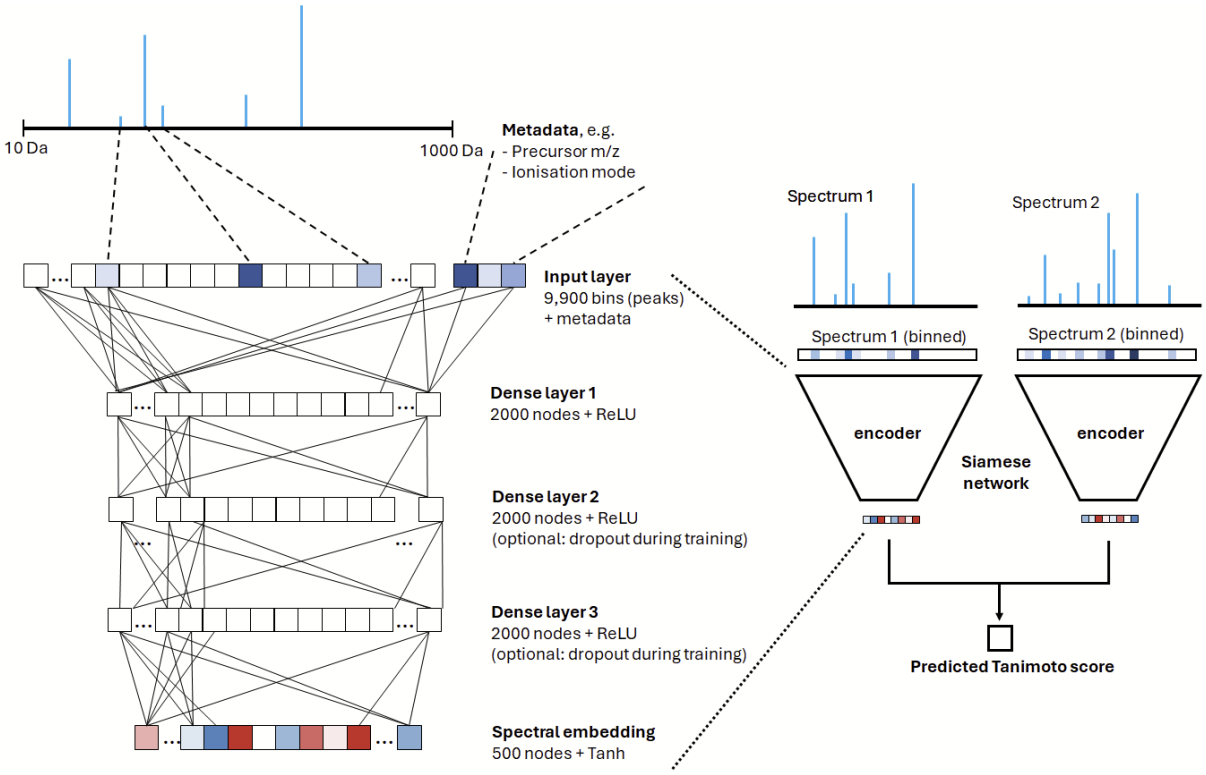

Figure 1: The model architecture for encoding mass spectra. *The output is a numerical vector (embedding) of length 500. As input the intensities of binned fragments are used and metadata is converted into numbers by one hot encoding or normalization. The model is trained to create embeddings in a way that the cosine similarity between two embeddings correlates well with chemical similarity (Tanimoto score).*

Metadata as input

MS2DeepScore 1.0 uses mass fragments as an input to predict chemical similarity between mass spectra. In the current work, MS2Deepscore 2.0 allows for additional use of recorded metadata of the fragmentation spectra. This is implemented in a flexible way which allows adding any type of metadata as an input into the model. Numerical data, e.g., precursor *m/z* or ionization energy, is transformed to have numbers in the range of 0 to 1, to optimize the learning. Textual inputs, like ion mode or instrument type, are one hot encoded. We note that for the dual-ion mode model used in the main text, precursor m/z and ion mode were used as additional metadata input. As a part of the current study, other experiments were run with instrument type and/or adduct type as additional input(s), but these did not notably improve performance. An overview of the model architecture can be found in Figure 1.

Tanimoto score

As a metric for chemical similarity between two molecules the Tanimoto score is used. An rdkit[15] daylight fingerprint (2048 bits) is generated for each unique 2D structure. A Tanimoto score[16] is calculated between two daylight fingerprints and used as a metric for chemical similarity. This Tanimoto score is used for training and benchmarking and will be referred to as Tanimoto score.

Spectrum pair selection for training

One of the main difficulties in training a model on predicting chemical similarities, here represented as Tanimoto scores, is the highly non-uniform distribution of Tanimoto scores across possible spectrum pairs, as can be seen in Figure 2.

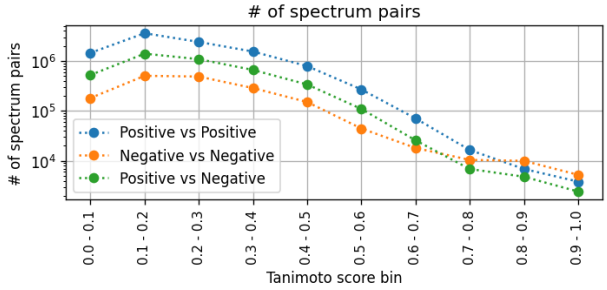

***Figure 2**: **Distribution of Tanimoto scores.** The distribution of Tanimoto scores is plotted for pairs of metabolites in the validation sets sampled from the GNPS library.*

In our previous work[4], this was partly mitigated by a data generator that selects a spectrum pair belonging to a random Tanimoto score bin for each pair selection step during model training. However, compounds in the used dataset often have no corresponding compound in high Tanimoto score ranges. Hence, even though the former data generator substantially reduced the bias, there still was a considerable shift towards lower Tanimoto scores.

Here, we re-designed the spectrum pair selection process during the model training. This resulted in a better compensation of the bias in chemical similarity and a better computational performance. Each unique compound is sampled equally, even though most compounds have multiple corresponding spectra. Unique compounds are selected by selecting unique InChIKeys, only considering the first 14 characters, since the rest of the characters are related to stereochemistry. When selecting pairs of compounds, we compensate for the Tanimoto score bias, by using 10 equally sized bins between 0 and 1. We sample compound pairs equally from each bin; however, for some compounds there is no match in one of the Tanimoto bins. These cases are compensated by selecting extra pairs for other compounds in that bin. A fully balanced distribution across all Tanimoto scores during training would mean that some compounds are selected more frequently because they have pairs in all score bins. Instead, we iterate through all unique compounds during one training iteration (epoch) and randomly pick from the pre-selected pairs for this compound. The resulting Tanimoto score distribution therefore is not perfectly balanced, but the remaining bias is very moderate. Furthermore, this does allow us to use the diversity of unique compounds in our dataset more equally.

Embedding and score uncertainty estimation

The prediction quality of the MS2Deepscore model is sensitive to the quality of input spectra and the similarity to the training data. To detect spectra that are hard to predict for MS2Deepscore, we designed a new pipeline using a convolutional neural network that predicts the quality of a spectrum embedding. We designed the "Embedding Evaluator" model by implementing an Inception Time architecture[17] using Pytorch[18], and trained it on the mean squared error (MSE) of all Tanimoto score predictions between the embedding in question and 999 randomly sampled other spectra from the training data. The conceptual idea here is that the Embedding Evaluator will learn to identify embeddings for low-quality or out-of-distribution input data. In later applications, the predicted embedding qualities can be used for uncertainty estimation.

For a per-score uncertainty estimation, we added a simple linear model. This will take six inputs derived from the predicted embedding quality $q_1$and q_2 of both input spectra: 1, q_1, q_2, q_1^2, q_2^2, $q_1 q_2$to predict the absolute prediction error for this pair of spectra, which we here consider as a proxy for a later uncertainty estimation. This model was implemented using Scikit-Learn[19]. A schematic overview of the embedding evaluator and linear model can be found in Figure 3.

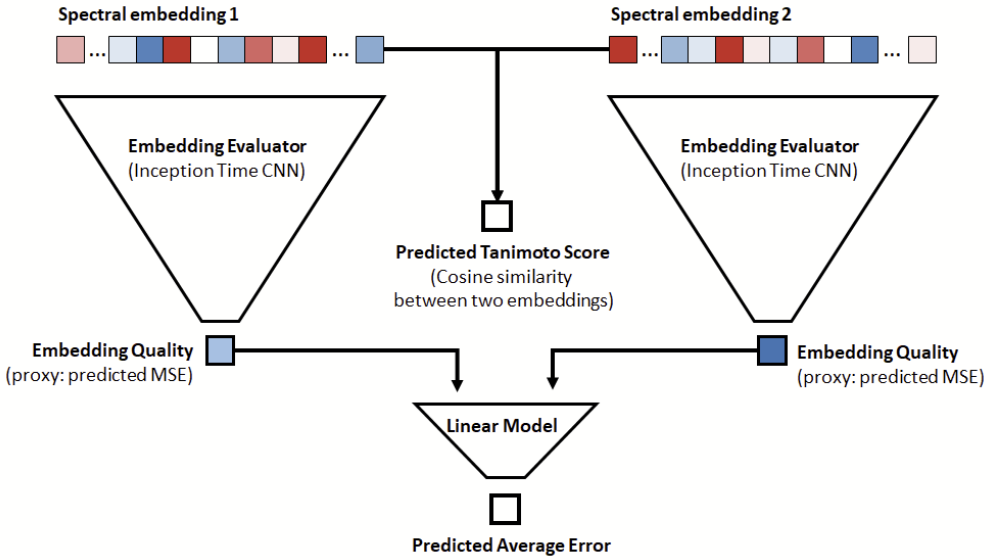

Figure 3: Schematic overview of embedding evaluator. *An embedding evaluator is implemented that predicts the reliability of an embedding. This embedding evaluator is trained by using the MSE for this embedding as a proxy. A linear model is implemented to predict the average error for a pair of spectra based on the prediction of the embedding evaluator for two spectra.*

Binning spectra

Before training the fragments are binned, to make them suitable as input for the neural network. Binning happens by making bins of 0.1 Da between 10 and 1000 Da, resulting in 9.900 bins. In the former MS2Deepscore work, bins were only included if they had at least 1 fragment in the training data. Instead, in MS2Deepscore 2.0, all bins are used, even if none of the training spectra have a fragment in this bin. This reduces code complexity and reduces the risk of accidental mismatch between the binning method and model versions.

Architecture improvements

MS2Deepscore 1.0 was implemented in Tensorflow[20]. Here, the entire MS2Deepscore model was reimplemented using Pytorch[18], to not only improve compatibility with GPUs, Apple M1 chips, but also overall code readability. Combined with an entirely new implementation of the DataGenerators, this resulted in a substantial speed-up in the training of models.

A pipeline is now available that performs all steps necessary for training new MS2Deepscore models. The wrapper function only requires a file with annotated mass spectra and the settings for model training. In a single run mass spectra are split on ion mode, split in test, train, and validation splits, then models are trained, and benchmarking figures are created.

Changes to model settings

The original MS2Deepscore paper used 2 layers of 500 and an embedding size of 200. Since the training library increased and since we moved to dual-ion mode models, it was expected that larger models would result in better performance. The results of experiments to find a good model size can be found in Supplementary Figure 1. We settled for using 3 layers of size 2000 and an embedding size of 500. This was used for all models used in the main text. Compared to the former MS2Deepscore models, a few other changes were made. Tanh activation was used in the last layer instead of ReLU[21], dropout and batch normalization were not used anymore and the settings for data augmentation were changed: augment removal max was changed from 0.3 to 0.2, augment intensity was changed from

0.4 to 0.2, and augment noise intensity was changed from 0.01 to 0.02. A JSON with all settings can be found in Supplementary Settings 1.

Input data filtering and splitting

For training the models we used the GNPS library spectra[2], which were first cleaned using our matchms library cleaning pipeline[22, 23]. The settings for cleaning can be found in Supplementary Settings 2. A considerable amount of mass spectra in the public GNPS libraries contain few mass fragments, which makes it hard to differentiate between different metabolites. We therefore decided to only add spectra that were fragmented well enough for substantiated predictions of chemical similarity. We used a minimum of 5 peaks with an intensity of at least 5% of the highest peak intensity. Experiments that assessed the model performance for different minimum peak number and intensity thresholds can be found in Supplementary Figure 2.

The cleaned spectrum library is first split on ion mode, followed by creating a training, validation and test set. We selected 1/20th of unique compounds (identical 2D structure) for both the validation and test set, all corresponding spectra to these compounds were removed from the training set. The number of spectra per set is given in Table 1. For the positive and negative set, the selection of InChIKeys to use for the validation and test sets was different, since we would otherwise have a bias in the validation set for spectra that were available in both ionisation modes. The dual-ion mode library was trained by combining the positive ion mode training spectra and the negative ion mode training spectra. The validation spectra were used for all experiments for the optimization of our model, like changing the filtering of input spectra, or adjustments to the model size. The test set was not used during any experimentation or hyperparameter optimization and was not used for any figures in the current study.

Spectrum set	Number of spectra	Number of compounds
Positive ion mode training spectra	173966	18169
Positive ion mode validation spectra	9704	1009
Positive ion mode test spectra	8945	1009
Negative ion mode training spectra	35118	7445
Negative ion mode validation spectra	1754	413
Negative ion mode test spectra	2109	413

Table 1: *The size of the different training, validation and test sets. The number of unique compounds is determined by considering the first 14 characters of the* InChIKey*.*

Benchmarking

The GNPS mass spectral library[2] often contains multiple mass spectra for one compound, in some cases up to several hundred spectra for the same compound. To avoid judging performance mostly on the performance of few compounds with a high number of mass spectra, one spectrum was randomly sampled per unique compound. This sampling was repeated 10 times, to better reflect the diversity in the validation spectra.

MS2Deepscore was compared to the cosine score and the modified cosine score. These scores were calculated between the same pairs in the test set by using the implementation in matchms[22].

Results

An extensive and systematic overhaul of the MS2DeepScore model was necessary, largely due to new demands and insights from the work with growing dataset sizes. Based on numerous experiments, with model size, embedding size and library filtering (see Supplementary Figures 1-3), several model parameters were changed. In addition, to achieve much faster training and prediction times, the entire model was re-implemented using Pytorch[18]. A MS2Deepscore model can now be trained in 2 hours on a server with Intel Xeon gold 6342 2.8Ghz, Nvidia A40 GPU and 512 GB Memory. The resulting model can predict Tanimoto scores between spectra of (to the model) unknown compounds with considerable accuracy, see Figure 4.

Additional metadata input

Next, we explored the influence of adding key metadata information to the model input. The relevance, but also the abundance of certain metadata entries, is likely to vary notably between different research areas and sample types. We here decided to add metadata that we expect to have a direct link to the fragmentation pattern. At the same time, we excluded options that require prior data interpretation, such as chemical formula, or that are very specific to a particular instrument or protocol, such as retention time. We further limited our exploration to metadata that was readily available in the used harmonized GNPS public library dataset, which currently does not provide consistent information about collision energy. The remaining metadata fields were precursor *m/z*, ion mode, and instrument type. Adding precursor *m/z* and ion mode as input for the model notably improved the performance, while one-hot encoding of the instrument type and adduct type did not notably improve performance (see Supplementary Figure 4).

Cross-ion mode models

Training MS2Deepscore 2.0 models on both ion modes at the same time resulted in a model that performs well for predicting chemical similarity between the same ion modes, but also for predicting chemical similarity between two ion modes. Figures 4a and b show that MS2Deepscore 2.0 trained on both ion modes performs, for within ion mode predictions, similarly to models trained only on spectra from a single ion mode. In addition, Figure 4C shows that MS2Deepscore 2.0 can also predict chemical similarity between spectra of two different ion modes. More detailed benchmarking showing the distributions of predictions and true values per bin can be found in Supplementary Figures 6-9.
Figure 5 shows two examples from the test set, where dual-ion mode MS2Deepscore correctly predicts a high chemical similarity and illustrates why conventional heuristic approaches like the cosine score are not suitable for predicting chemical similarity between mass spectra of different ion modes.

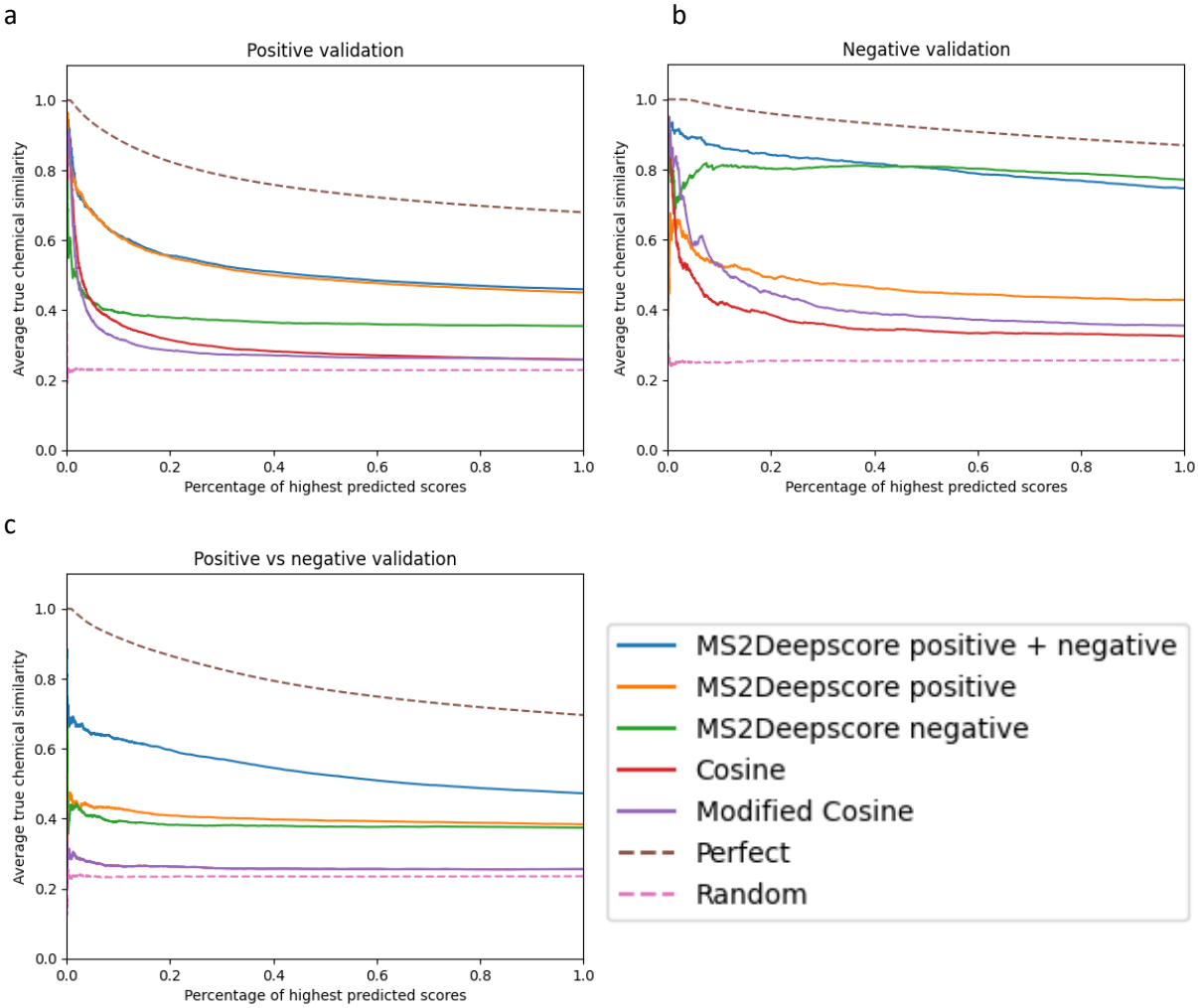

Figure 4: Dual-ion mode MS2Deepscore model predicts reliable results between ion modes and performs similar to single-ion mode MS2Deepscore models, when predicting between the same ion mode. *Performance is compared between an MS2Deepscore model trained on both positive and negative mode spectra and two models trained on only one of the ion modes, in addition the performance is compared to the predictions made by cosine or modified cosine scores. The validation set was used for the benchmarking. Predictions are made between single spectra from all unique compounds in the validation set. The predictions are sorted from high to low and the average for the real chemical similarity (Tanimoto score) is plotted on the y axis for the given 1% of highest predictions on the x axis. The full 100% of highest predictions can be found in Supplementary Figure 10.* ***a.*** *Benchmarking on the validation set with pairs of positive ion mode spectra.* ***b.*** *Benchmarking on the validation set with pairs of negative ion mode spectra.* ***c.*** *Benchmarking on the validation set with pairs between a positive and a negative ion mode spectrum. Only spectrum pairs compared between different ion modes are displayed in this panel.*

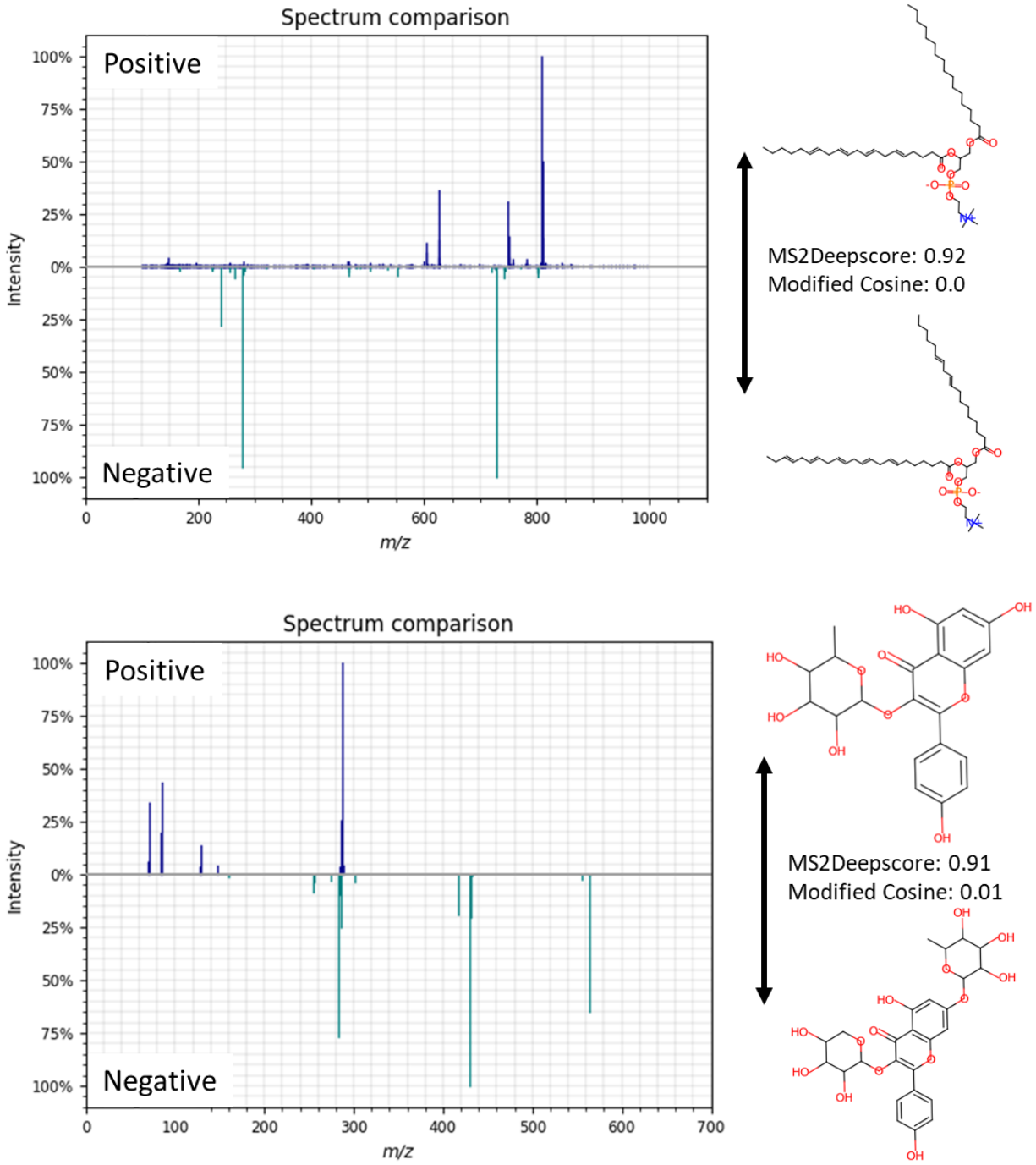

***Figure 5**: **Two example spectra of the validation test sets with high MS2Deepscore scores.** Conventional scores like modified cosine scores are by design unable to predict chemical similarity since almost no peaks overlap. However, MS2Deepscore 2.0 can correctly predict high similarity between a spectrum in positive and negative ion mode even though almost no peaks align. This highlights how conventional methods of predicting spectral similarity like the modified cosine score are unsuitable for predicting chemical similarity between different ion modes.*

Uncertainty evaluation

In some cases, MS2Deepscore is not able to make reliable predictions for a mass spectrum. For instance, because of bad fragmentation, fragments of multiple metabolites in one spectrum (i.e., "hybrid" spectra), or simply because there were no similar spectra in the training data.

The Embedding Evaluator model can identify spectra for which MS2Deepscore cannot predict reliable chemical similarities. By filtering out spectra that have a high predicted MSE, we can improve prediction reliability. The effect of removing spectra for which a high MSE is predicted is visualized in Figure 6. Additional analysis for the embedding evaluator can be found in Supplementary Figure 11-13.

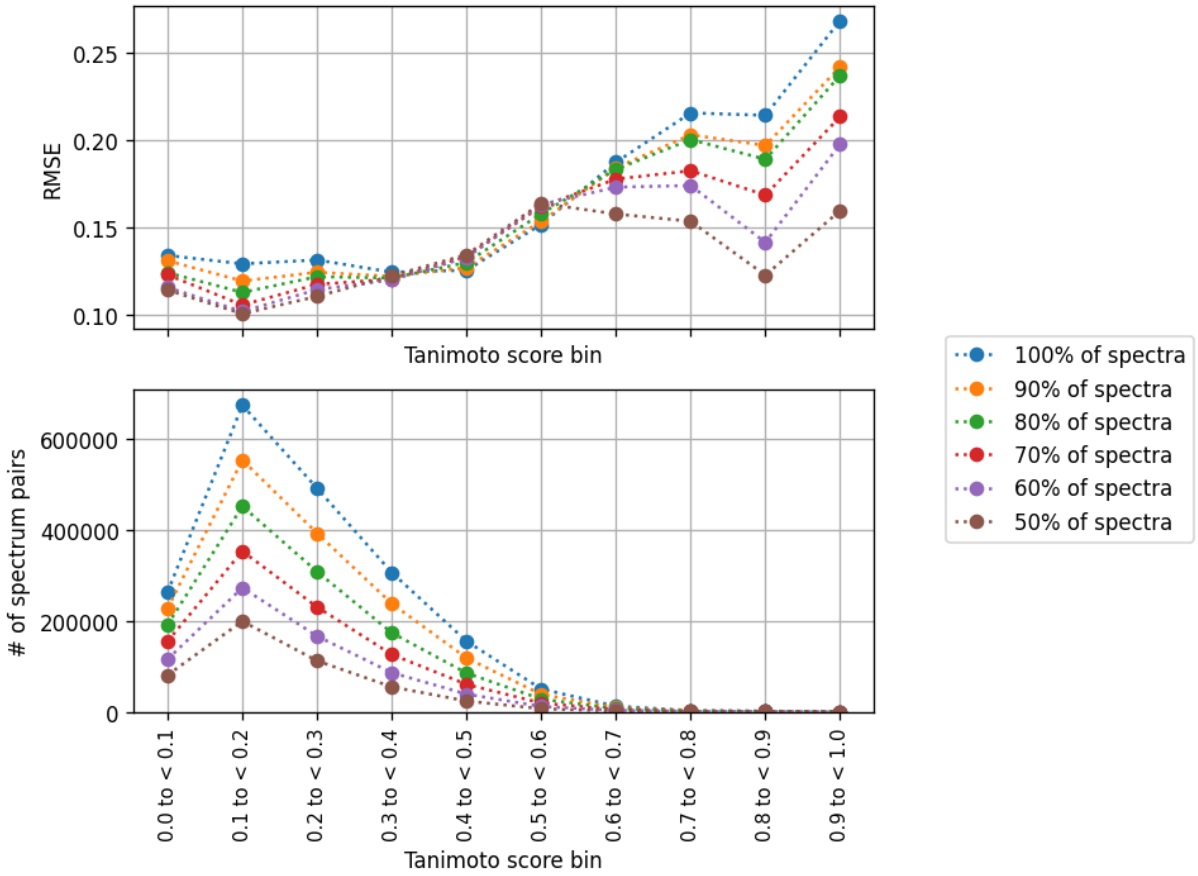

***Figure 6**: **Removing low-scoring embeddings improves prediction accuracy.** Predictions for accuracy are made using our Embedding Evaluator model. The spectra with the highest predicted MSE are removed, with the remaining percentage of mass spectra indicated in the label. The average RMSE per Tanimoto score bin is plotted for the resulting spectra.*

Discussion

The most common classical spectrum similarity measure, the cosine score, measures the similarity of the fragmentation pattern and is maximal for a maximal visual equivalence between two spectra. With sufficiently strict settings, very high cosine scores between two spectra can be a good indicator for (near-)identical compounds. However, even moderate chemical modifications can cause notable peak shifts, making the cosine score not well suited for searching chemically similar compounds. The so-called modified cosine score is a very common variation that can account for a single structural modification which makes it more suitable when searching for structurally similar compounds[2, 24, 25]. However, this typically no longer holds for more complex fragmentation relationships resulting from multiple structural modifications[3]. In addition, both cosine and modified cosine scores assume similar experimental conditions, since parameters such as instrument type or collision energy, but also the data cleaning pipeline, substantially influence the fragmentation pattern. As a result, both cosine score and modified cosine score usually only identify a tiny fraction of all high chemical similarity pairs[3, 26] and also suffer from a high false positive rate for larger datasets[3]. By design, both cosine and modified cosine scores should not be used to compare spectra across different ion modes.

A machine learning method like MS2DeepScore can largely compensate for those limitations. Trained on a very diverse set of MS^2 spectra, MS2Deepscore can still make reliable predictions between mass spectra measured under different conditions, even if hardly any of the fragments overlap (Figure 5). We show that even for spectra that are measured in different ion modes MS2Deepscore delivers very good estimates of the actual chemical similarity (Figure 4c). In addition, the dual-ion mode MS2Deepscore model performs similarly or even better to models that were only trained on one ion mode, when doing predictions between the same ion mode (Figure 4a and 4b). Therefore, the dual-ion mode model can be used without compromising on model performance for normal same ion mode comparisons.

Interestingly MS2Deepscore models trained on one of the ion modes show a better than random prediction performance when predicting mass spectral similarity of spectra obtained in the other ion mode (e.g., positive ion model to predict between neg-neg – see Fig. 4a). This suggests that there are some patterns that MS2Deepscore learns that already generalize between the two ion modes.

The prediction quality of the MS2Deepscore model is sensitive to the quality and type of the input spectra. Both low-quality spectra due to limited fragmentation or multiple metabolites in one MS^2 spectrum, or spectra with little similarity to our training data can result in bad predictions. In the original MS2Deepscore paper[4], the uncertainty was estimated using a Monte-Carlo dropout regularization[27]. In later real-world applications, however, we noted that this was not an adequate solution. For example, we noticed that spectra with little similarity to the training data as well as low-quality spectra often received very similar embeddings. This can be very detrimental, because similar embeddings will lead to -mostly false- predictions of high chemical similarities, see Supplementary Figure 13.

Our model, however, is now complemented by an uncertainty estimation for both individual input spectra and Tanimoto score predictions. We demonstrated that the overall accuracy can be raised by removing the more uncertain predictions (Figure 6). Still, given the enormous range of possible mass spectral datasets and applications, there remains a risk of our model not being well-suited for very specific tasks or compound types. In such cases, we recommend training a custom MS2DeepScore model. Training a new ms2deepscore model is now relatively easy since an automatic training pipeline is available. For smaller custom datasets we would recommend merging them with larger available datasets, such as the here-used GNPS data, before training a new model from scratch. We speculate that a promising alternative route could be to start with our pre-trained model and run additional training on the custom reference data (a common “fine tuning” strategy in deep learning).

The ability to reliably predict chemical similarities across ion modes creates entirely new options for mass spectral data exploration by combining positive and negative ion mode data. Similarity-based graphs can now be generated independent of the ion mode, rendering cross-ion mode molecular networking feasible. It is now also possible to compare a negative ion mode spectrum with a large positive ion mode library, and vice versa. We expect that this will help researchers to make new discoveries and identify new compounds and links.

Code availability

All code is available on https://github.com/matchms/ms2deepscore. MS2Deepscore is pip installable. The version used for this manuscript is version 2.0.0. The notebooks used for creating the benchmarking figures can be found in the folder https://github.com/matchms/ms2deepscore/tree/main/notebooks/ms2deepscore_2

Data availability

The dual-ion mode ms2deepscore model used in this study can be downloaded from Zenodo, https://doi.org/10.5281/zenodo.10814307. The training spectra, validation spectra, test spectra, embedding evaluator, linear evaluator and model settings can also be downloaded from this DOI.

Author contributions

NdJ, JJJvdH, FH designed the research and wrote the manuscript. NdJ, DJ, LJT, FH contributed to the code. NdJ, FH designed and evaluated the code for the current version. All authors contributed to the data analysis and interpretation. JJJvdH and FH supervised this work.

Competing interests

JJJvdH is member of the Scientific Advisory Board of NAICONS Srl., Milano, Italy and consults for Corteva Agriscience, Indianapolis, IN, USA. All other authors declare to have no competing interests.

References

1. Tsugawa H*, et al.* Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Analytical chemistry* **88**, 7946-7958 (2016).

2. Wang M*, et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature biotechnology* **34**, 828-837 (2016).

3. Huber F*, et al.* Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS computational biology* **17**, e1008724 (2021).

4. Huber F, van der Burg S, van der Hooft JJ, Ridder L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of cheminformatics* **13**, 84 (2021).

5. de Jonge NF*, et al.* MS2Query: reliable and scalable MS2 mass spectra-based analogue search. *Nature Communications* **14**, 1752 (2023).

6. Treen DG*, et al.* SIMILE enables alignment of tandem mass spectra with statistical significance. *Nature communications* **13**, 2510 (2022).

7. Li Y, Kind T, Folz J, Vaniya A, Mehta SS, Fiehn O. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nature Methods* **18**, 1524-1531 (2021).

8. Liigand P*, et al.* Think negative: finding the best electrospray ionization/MS mode for your analyte. *Analytical chemistry* **89**, 5665-5668 (2017).

9. Yamamoto FY, Pérez-López C, Lopez-Antia A, Lacorte S, de Souza Abessa DM, Tauler R. Linking MS1 and MS2 signals in positive and negative modes of LC-HRMS in untargeted metabolomics using the ROIMCR approach. *Analytical and bioanalytical chemistry* **415**, 6213-6225 (2023).

10. Hegazi NM, Radwan RA, Bakry SM, Saad HH. Molecular networking aided metabolomic profiling of beet leaves using three extraction solvents and in relation to its anti-obesity effects. *Journal of Advanced Research* **24**, 545-555 (2020).

11. Neto FC, Raftery D. Expanding urinary metabolite annotation through integrated mass spectral similarity networking. *Analytical chemistry* **93**, 12001-12010 (2021).

12. Renai L, Ulaszewska M, Mattivi F, Bartoletti R, Del Bubba M, van der Hooft JJ. Combining feature-based molecular networking and contextual mass spectral libraries to decipher nutrimetabolomics profiles. *Metabolites* **12**, 1005 (2022).

13. Olivier-Jimenez D*, et al.* From mass spectral features to molecules in molecular networks: a novel workflow for untargeted metabolomics. *bioRxiv*, 2021.2012. 2021.473622 (2021).

14. Schmid R*, et al.* Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nature communications* **12**, 3832 (2021).

15. Landrum G. Rdkit documentation. *Release* **1**, 4 (2013).

16. Tanimoto TT. Elementary mathematical theory of classification and prediction. (1958).

17. Ismail Fawaz H*, et al.* Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery* **34**, 1936-1962 (2020).

18. Paszke A*, et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32**, (2019).

19. Pedregosa F*, et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).

20. Abadi M*, et al.* {TensorFlow}: a system for {Large-Scale} machine learning. In: *12th USENIX symposium on operating systems design and implementation (OSDI 16)*) (2016).

21. Sharma S, Sharma S, Athaiya A. Activation functions in neural networks. *Towards Data Sci* **6**, 310-316 (2017).

22. Huber F*, et al.* matchms-processing and similarity evaluation of mass spectrometry data. *bioRxiv*, 2020.2008. 2006.239244 (2020).

23. de Jonge NF, Hecht H, van der Hooft JJ, Huber F. Reproducible MS/MS library cleaning pipeline in matchms. (2023).

24. Beniddir MA, Kang KB, Genta-Jouve G, Huber F, Rogers S, Van Der Hooft JJ. Advances in decomposing complex metabolite mixtures using substructure-and network-based computational metabolomics approaches. *Natural product reports* **38**, 1967-1993 (2021).

25. de Jonge NF*, et al.* Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. *Metabolomics* **18**, 103 (2022).

26. Bittremieux W, Wang M, Dorrestein PC. The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics* **18**, 94 (2022).

27. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *international conference on machine learning*). PMLR (2016).