

METHODOLOGY

Open Access



MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra

Florian Huber^{1*}, Sven van der Burg¹, Justin J. J. van der Hooft² and Lars Ridder¹

Abstract

Mass spectrometry data is one of the key sources of information in many workflows in medicine and across the life sciences. Mass fragmentation spectra are generally considered to be characteristic signatures of the chemical compound they originate from, yet the chemical structure itself usually cannot be easily deduced from the spectrum. Often, spectral similarity measures are used as a proxy for structural similarity but this approach is strongly limited by a generally poor correlation between both metrics. Here, we propose MS2DeepScore: a novel Siamese neural network to predict the structural similarity between two chemical structures solely based on their MS/MS fragmentation spectra. Using a cleaned dataset of > 100,000 mass spectra of about 15,000 unique known compounds, we trained MS2DeepScore to predict structural similarity scores for spectrum pairs with high accuracy. In addition, sampling different model varieties through Monte-Carlo Dropout is used to further improve the predictions and assess the model's prediction uncertainty. On 3600 spectra of 500 unseen compounds, MS2DeepScore is able to identify highly-reliable structural matches and to predict Tanimoto scores for pairs of molecules based on their fragment spectra with a root mean squared error of about 0.15. Furthermore, the prediction uncertainty estimate can be used to select a subset of predictions with a root mean squared error of about 0.1. Furthermore, we demonstrate that MS2DeepScore outperforms classical spectral similarity measures in retrieving chemically related compound pairs from large mass spectral datasets, thereby illustrating its potential for spectral library matching. Finally, MS2DeepScore can also be used to create chemically meaningful mass spectral embeddings that could be used to cluster large numbers of spectra. Added to the recently introduced unsupervised Spec2Vec metric, we believe that machine learning-supported mass spectral similarity measures have great potential for a range of metabolomics data processing pipelines.

Keywords: Mass spectrometry, Metabolomics, Spectral similarity measure, Supervised machine learning, Deep learning

Introduction

In the rapidly growing field of metabolomics, mass spectrometry fragmentation approaches are a key source of information to chemically characterize large numbers of detected molecules. Mass fragmentation (MS/MS or MS2) spectra are created through the fragmentation of

molecules in the mass spectrometer and consist of peaks that reflect the mass over charge (m/z) position of the resulting mass fragments. The peak intensities are reflective of the likelihood various fragmentation paths occur for the fragmented molecule. One of the core challenges in metabolomics is to link MS/MS spectra to chemical structures of the fragmented metabolites. Over the last years, many computational tools have been developed to help with annotating MS/MS data [1]. In many workflows for extracting chemical information from MS/MS

*Correspondence: florian.huber@hs-duesseldorf.de

¹ Netherlands eScience Center, 1098 XG Amsterdam, The Netherlands
Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

spectra, automated quantitative comparisons between pairs of spectra play a crucial role. Such comparisons are used to match unknown spectra to library spectra, i.e., spectra with known or reliably annotated structures [2], or to explore chemical relationships by means of networks (or graphs) built based on mass spectral similarity scores [3, 4].

One key limitation in many approaches using mass spectral similarities is that often the main interest is not the degree of similarity between two spectra, but the *structural similarity* between the fragmented chemical compounds [4–6]. There is no single absolute measure to determine such chemical structure relatedness. In practice, the structural similarity between molecules is frequently computed from molecular fingerprints: vectors that describe the presence/absence of many structural features in the molecule. Structural similarity calculations are central to many applications in cheminformatics including virtual screening [7, 8]. Molecular fingerprints, however, are computed from the chemical structure, which usually are only known for a tiny fraction of all mass spectra from complex mixtures [9]. The most established approach to infer molecular fingerprints without known chemical structure is through support vector machines in combination with the computation of fragmentation trees [10], but this is computationally expensive, in particular for larger compounds. Recently, first attempts have been made to also use deep neural networks for directly predicting molecular fingerprints from mass spectra [11, 12]. One of the main obstacles in predicting molecular fingerprints is that they are typically large, very sparse, binary vectors. With limited training data it becomes difficult to correctly predict less common structural features (i.e., the bits). As a consequence, previous deep learning approaches focused on predicting only frequently activated bits of the molecular fingerprints [11]. Despite missing less frequent bits, such predicted fingerprints can still be used by matching them with known library fingerprints [11, 12]. With current open spectral libraries growing to such sizes that machine learning approaches have sufficient data for training, validation, and testing; we recognize that there is an opportunity for the development of alternative mass spectral similarity scores. This is also in line with a recent review article sketching the current and future role of deep learning for metabolite annotation [13].

Here, we present MS2DeepScore, a deep learning approach that is trained to predict structural similarities (Tanimoto or Dice scores based on molecular fingerprints) directly from pairs of MS/MS spectra without first computing molecular fingerprints. This is similar in spirit to the approach by Ji et al. [14] but uses a conceptually simpler Siamese neural network architecture

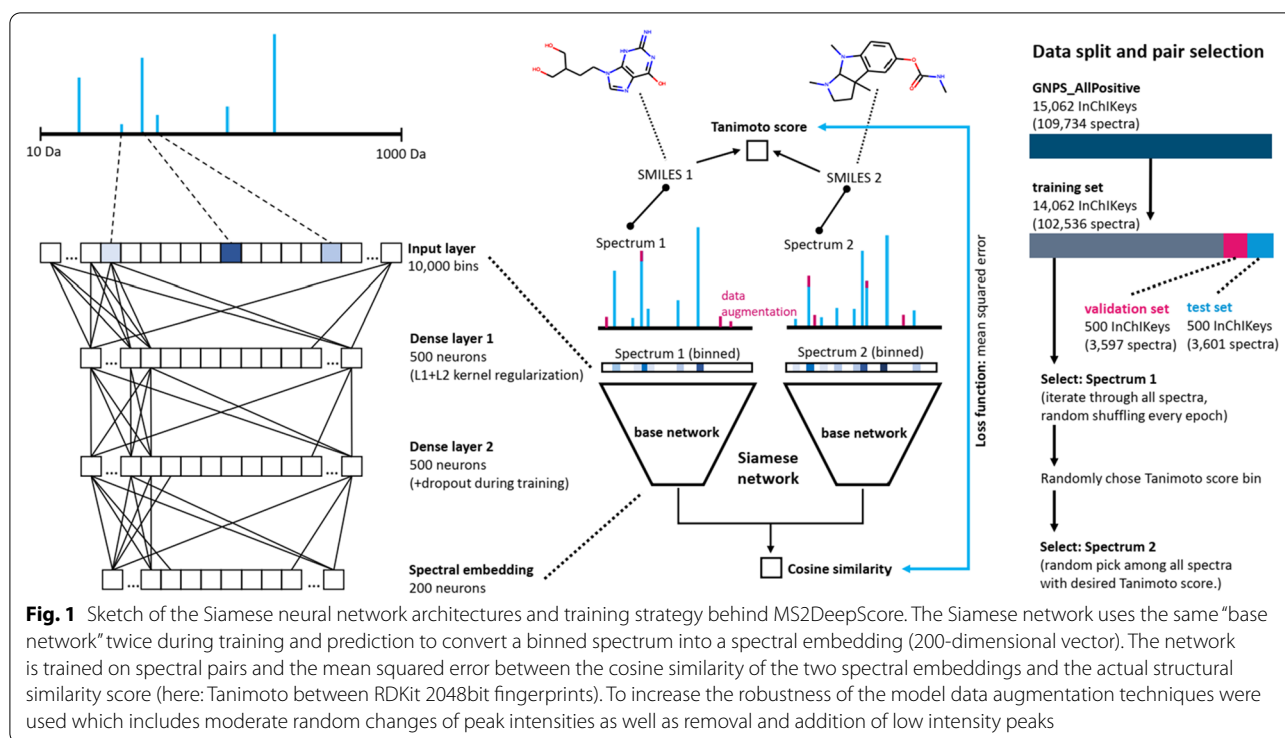
[15]. Furthermore, our approach only relies on peak m/z positions and intensities without requiring further spectrum information, in contrast to approaches such as DeepMass for which the mass and chemical formula are both necessary input [14]. The differences in input to these models also makes it difficult to directly compare these approaches quantitatively. Our proposed approach makes it possible to use MS2DeepScore for predicting structural similarities between spectra of various origins and with varying metadata quality. The model was trained using a dataset of 109,734 MS/MS spectra, which was built through curating and cleaning spectra obtained from GNPS [16] (see “Methods”). In addition to the prediction of a structural similarity, MS2DeepScore can also make use of Monte-Carlo dropout [17] to assess the model uncertainty.

We demonstrate that MS2DeepScore can predict structural similarities with high reliability. When comparing commonly used molecular fingerprints, we achieve a root mean squared error for predicted Tanimoto scores of about 0.15 when run without uncertainty restrictions, and down to 0.1 with stronger restrictions on model uncertainty. MS2DeepScore is very well suited to detect compounds of high structural similarity and furthermore can create mass spectral embeddings that can be used for additional spectral clustering. We hence expect MS2DeepScore to become a key asset in building future MS/MS analysis pipelines. Depending on the desired application, MS2DeepScores could also be combined with other mass spectral metrics to make full use of their complementary aspects.

Results

A large set of MS/MS spectra was retrieved from GNPS [16] and subsequently curated and cleaned using matchms [18] (see “Methods”). The resulting training data set contains chemical structure annotations for 109,734 spectra, in the form of SMILES [19] and/or InChI [20], which allowed us to create molecular fingerprints to quantify structural similarities of spectral pairs. The dataset contains 15,062 different molecules (disregarding stereoisomerism—as represented by InChIKeys unique in the first 14 characters).

We randomly took out 500 of the 15,062 InChIKeys to form a validation set and again 500 to form a test set (see also “Methods”). The validation set (3597 spectra of 500 unique InChIKeys) is used to monitor the model training process and explore the key hyperparameters while the test set (3601 spectra of 500 unique InChIKeys) is used for a final unbiased evaluation of our model. Drawing pairs of spectra from the training set (102,536 spectra of 14,062 unique molecules), we trained a Siamese neural network to predict Tanimoto scores as depicted in Fig. 1.

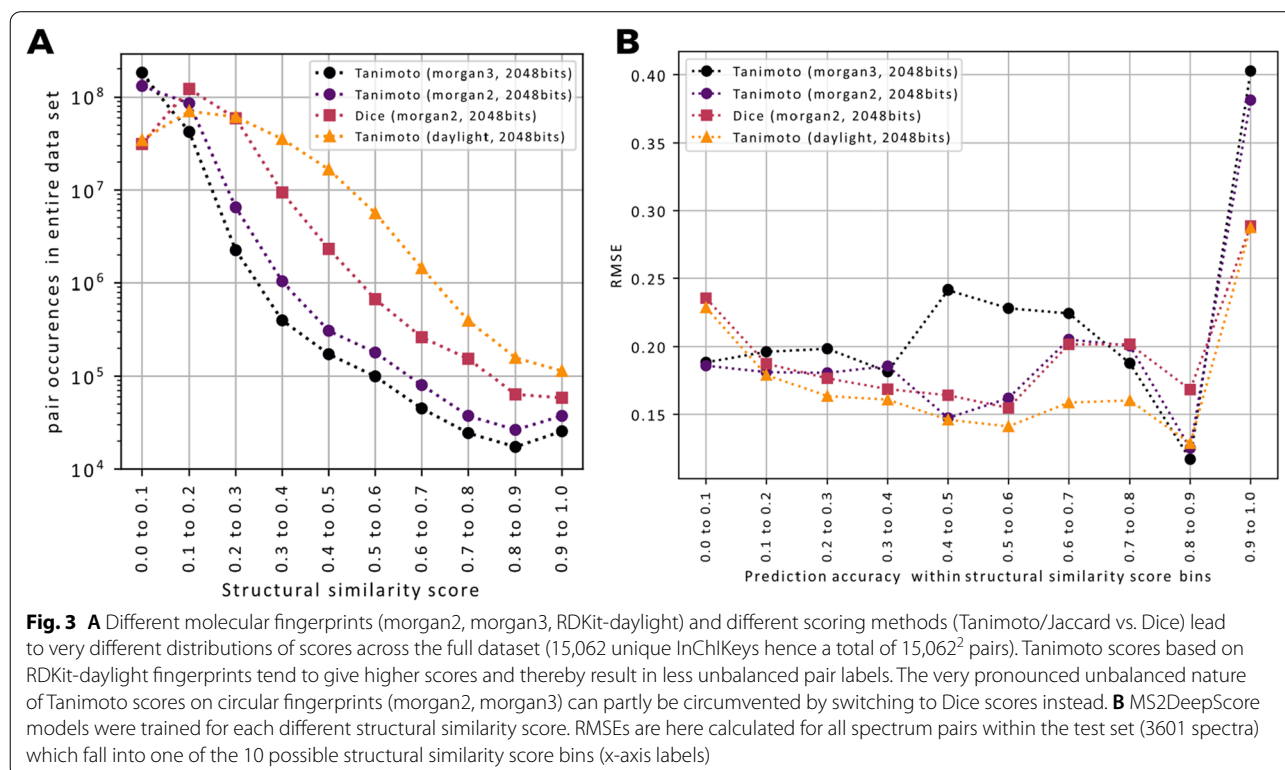
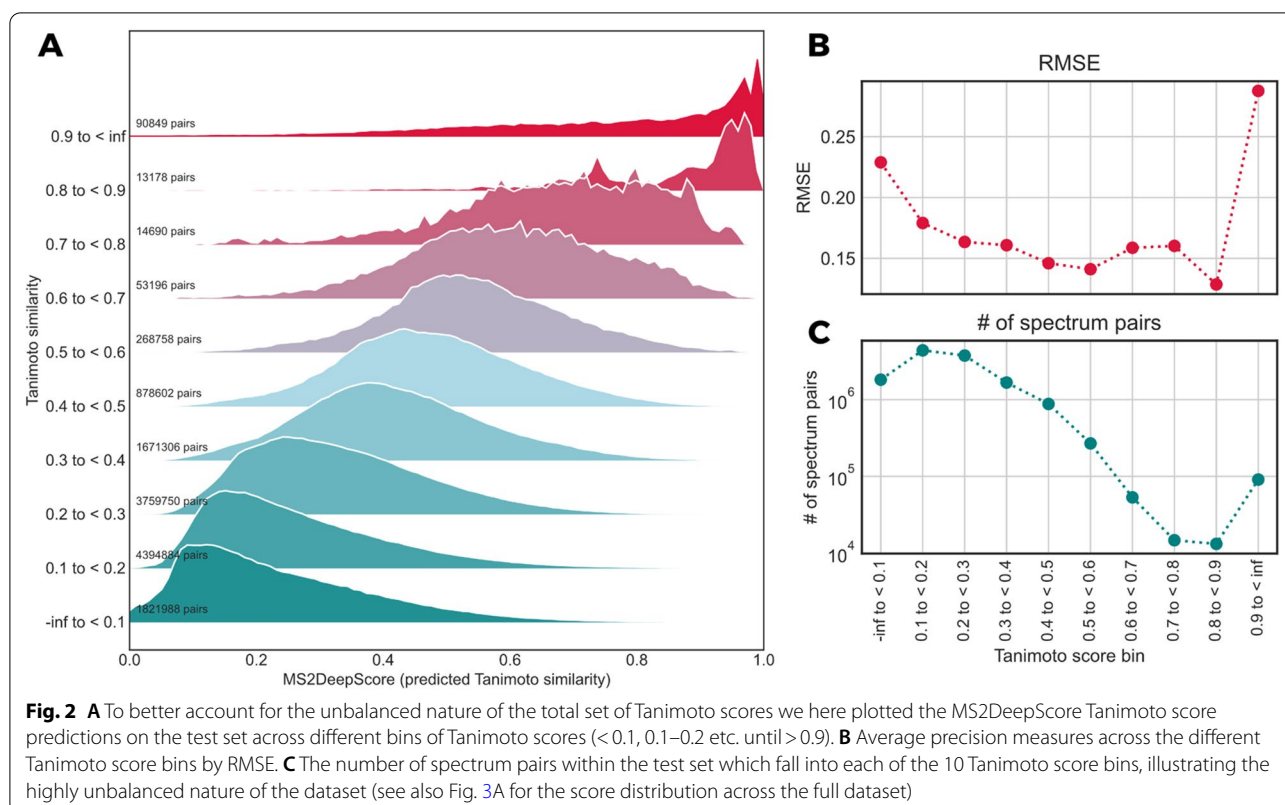


A key challenge when training a neural network to predict Tanimoto scores is that the total set of possible spectrum pairs shows a highly unbalanced distribution in structural similarities, with most pairs displaying low structural similarity. Our procedure for drawing spectral pairs compensates for the unbalanced nature of the data, by selecting pairs with probabilities that are weighted according to their structural similarity, as described in detail in the Methods section. We further applied L1, L2 and dropout regularization, as well as data augmentation techniques to ensure the generalization of the model to unseen data.

MS2DeepScore predicts Tanimoto scores with high accuracy

In real-world applications of the model, it is acceptable that there is a small error in estimation of the structural similarity between two spectra, while outliers with large errors should be avoided. Therefore, the root mean squared error (RMSE) was used as an overall evaluation metric, since it penalizes large errors on individual samples. In addition, the model should ideally perform well across the full range of possible pair similarities, which for the here used Tanimoto and Dice scores lies between 0 and 1. However, the datasets are highly unbalanced in that respect, since most spectrum pairs have low Tanimoto scores (Fig. 3A). We hence decided to inspect the model accuracy not as a global average since that

would strongly bias the outcome to the performance on low Tanimoto pairs. Instead, we split all possible spectral pairs into 10 equally spaced Tanimoto score bins. In Fig. 2A we display the distributions of the predicted Tanimoto scores for each bin, which reveals that the individual distributions show a high overlap with the correct Tanimoto scores. As expected, the prediction is not perfect. The distributions show long tails of predictions that differ from the true structural similarities. Looking at root mean squared errors (RMSE) across all Tanimoto score bins, it can be noted that MS2DeepScore generally performs very well and can predict Tanimoto scores between 0.1 and 0.9 with a RMSE between 0.13 and 0.2. Accuracy is lower for the highest and lowest Tanimoto scores, which may partly be attributed to the regression to the mean effect (the training loss makes it unattractive to approach the upper and lower score limit). Another reason for the false high Tanimoto score predictions can arise from spectra that differ too much from the training data the model has seen (see also the later described effect of assessing the model uncertainty). The highest Tanimoto scores show a relatively long tail indicating more frequent wrong predictions (Fig. 2A). Predictions are also slightly more spread out for Tanimoto scores around 0.6–0.8, a range with relatively few occurrences and hence less training data in the dataset. Other underlying reasons cannot be ruled out at this point, such as poorer correlation between fragmentation information



and actual structural similarity scores. Taken together, MS2DeepScore is highly reliable in separating high, mid, and low structural similarity pairs (see also Additional file 1: Figure S1), but it might be more error prone when it comes to smaller nuances.

Accurate prediction of different structural similarity measures

Molecular fingerprints come in many different types and flavors which typically do not work equally well for all compound classes [7, 8, 21], and there is no general consensus on which molecular fingerprint to use (although MAP4 was recently said to “rule them all” [21]). 2D fingerprints can be characterized as belonging to one of four classes [7, 22], two of which are represented in the present work. Those were circular fingerprints (we used Morgan-2 and Morgan-3, the RDKit [23] pendants of ECFP-4 and ECFP-6) as well as topological fingerprints (we used Daylight fingerprints in RDKit). All fingerprints were computed with 2048 bits. Then there are also different metrics to compute the similarity between two molecular fingerprints [24]. Arguably, Tanimoto (=Jaccard) is the most common way to compare molecular fingerprints, other metrics such as Dice scores are occasionally used. To show that our approach generalizes to a variety of structural similarity scores, we trained and tested the MS2DeepScore model for the three different fingerprints (Morgan-2, Morgan-3 and Daylight) as well as 2 different similarity measures (Tanimoto and Dice). Overall, MS2DeepScore can make accurate predictions for all tested structural similarity measures (Fig. 3A). In addition, we observed that the precise distributions of all occurring structural similarity scores vary considerably when different structural similarity measures are used (Fig. 3B). Structural similarity measures by Tanimoto scores from circular fingerprints (Morgan-2 and 3) have a much higher tendency towards low scores when compared to Tanimoto scores from RDKit Daylight fingerprints. This can partly be adjusted by switching from Tanimoto to other metrics such as using a Dice score. Overall, we found that MS2DeepScore performs slightly better when predicting structural similarity scores with a less skewed distribution. This is to be expected since such scores display far more instances of moderate to high scores in the entire training dataset, for instance Tanimoto scores on daylight fingerprints result in 10–100 times more pairs across scores in the range from 0.2 to 0.9 when compared to Tanimoto scores on Morgan-3 fingerprints (Fig. 3A). Due to its less skewed distribution, we decided to mainly use Tanimoto scores on RDKit Daylight fingerprints (2048bits) for evaluating structural similarity (unless noted otherwise). It is important to note, though, that linear correlation coefficients between

predicted and actual scores were generally high for the pairs with the highest structural similarities (see Additional file 1: Table S1).

Detecting chemically related pairs: comparison to common mass spectral similarity measures

Due to the low number of available correctly or reliably annotated mass spectra, many analysis pipelines must rely on mass spectral similarity measures. A classical way to compare MS/MS mass spectra is to quantify the fraction of shared peaks as done by using variations of cosine-based similarity scores. They come in many types and flavors [6, 25], but typically rely on multiplying intensities of matching peaks. Those pairs of matching peaks between two spectra are usually computed as so-called assignment problems based on set m/z tolerances. One particular variant of such scores is the ‘modified cosine score’ which also allows matching peaks that were shifted by the difference in precursor m/z [26]. Those measures tend to work well for very similar spectra, i.e. with many identical peaks. We recently introduced Spec2Vec, an unsupervised machine learning approach for computing spectrum similarities based on learned relationships between peaks across large training datasets [6]. Spec2Vec based similarity scores were observed to correlate more strongly than classical cosine-like scores with structural similarities between the underlying compounds. An additional advantage is its fast computation, which allows to compare spectra against very large libraries. While trained on spectral data, Spec2Vec used an unsupervised method, meaning that it was trained on non-annotated data and did not make use of the structural information.

With MS2DeepScore, we now make use of the structural information that we have for a large fraction of the training data. Unlike Spec2Vec, which is trained to learn relationships between peaks from peak co-occurrences, and unlike modified Cosine, which computes the maximum overlap of matching peaks, MS2DeepScore is specifically trained to predict structural similarity scores. The ability of those different scores to identify chemical relatedness can thus not simply be compared by measuring their ability to predict Tanimoto score. In practice, however, all such scores are all used to identify chemically closely related compounds. Modified Cosine and Spec2Vec scores are for instance used to generate molecular networks in GNPS [4, 6]. We therefore tested the scores’ ability to detect chemically related compounds by counting identified chemically related pairs within the test set (3601 spectra). Since “chemically related” is a hard to define concept, we simply operated with a fixed Tanimoto score threshold of 0.6 above which we call two compounds related. We then computed the

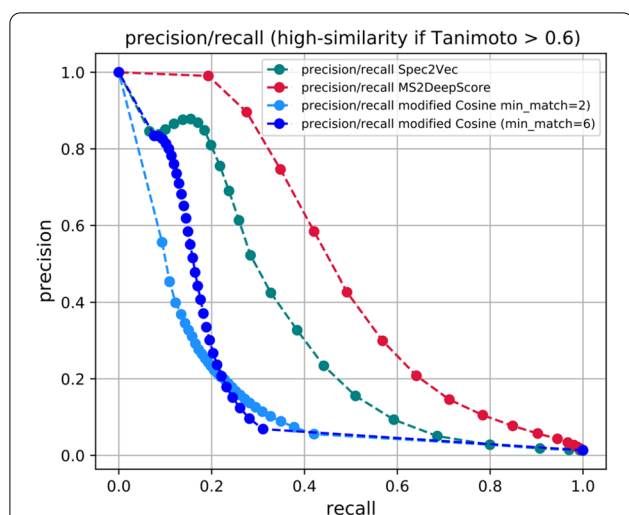


Fig. 4 We here define a high structural similarity as Tanimoto > 0.6 and explore how well high structural similarity pairs can be retrieved using various spectral similarities. Collecting all spectrum pairs from the test set (3601 spectra) with mass spectral similarity > X with X increasing from 0 to 1.0, we compute precision and recall for the different mass spectral similarity measures (MS2DeepScore, Spec2Vec, modified Cosine). The curves illustrate the tradeoff between higher recall (towards the right) and higher precision (towards the left). They also reveal that MS2DeepScore gives notably better precision/recall combination over the entire range, followed by Spec2Vec and only then modified Cosine

precision and recall (see “Methods”) for finding structurally related compounds for all spectrum pairs above a threshold for the spectral similarity score which could be either MS2DeepScore, Spec2Vec, or modified Cosine (Fig. 4). This reveals that MS2DeepScore clearly outperforms both classical measures (two forms of the modified Cosine) as well as the unsupervised spectral similarity measure Spec2Vec, with respect to identifying high Tanimoto pairs, which can also be seen in the overall distribution of scores (Additional file 1: Figure S1). This makes MS2DeepScore a very promising approach for searching analogues in large datasets.

Combining different mass spectra for the same compound decreases Tanimoto score prediction error further

In many applications, such as library matching or analogue searching, datasets will frequently contain multiple mass spectra for a given compound. This is also the case for the data retrieved from GNPS (see “Methods”). The test set, for instance, contains 3601 spectra of 500 unique compounds (ignoring stereoisomerism). We hence tested whether structural similarity score predictions can be improved by taking the median of the scores calculated for different pairs of spectra corresponding to the same compound pairs. This can indeed be seen on the test set (Fig. 5, compare red and dark blue lines).

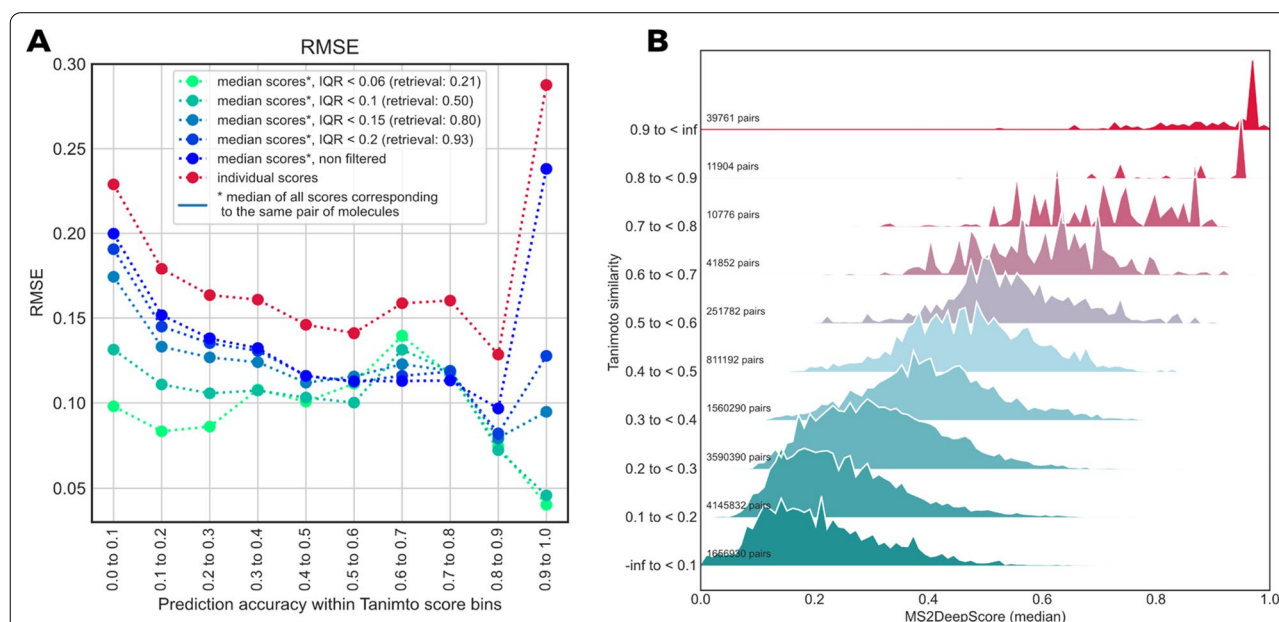


Fig. 5 Combining MS2DeepScore predictions for spectral pairs corresponding to the same pair of molecules leads to more reliable Tanimoto score predictions. **A** Individual predictions (red dots) show consistently higher RMSEs than the median of predictions for (pairs of spectra corresponding to) the same compound pair (“all”). We also computed the interquartile range (IQR) of predictions of the same molecule pairs, which can be used to remove high IQR outliers. Retrieval rates after each label indicate the total fraction of scores that fulfilled the given criterion (IQR < x). **B** When compared to Fig. 2 it is apparent that the high Tanimoto score predictions become notably more reliable when removing scores with large variations of same-InChIKey predictions (here: keep scores with IQR < 0.2)

The improvement in accuracy becomes even more pronounced when removing potential outliers based on the interquartile range (IQR) of all predictions for the same pair of molecules. In particular high and low Tanimoto score predictions become notably more reliable, even at comparably high threshold IQRs (Fig. 5A, e.g. threshold <0.2 which corresponds to 93% of all scores). Given the considerable improvement of the structural similarity prediction, we expect that this use of multiple predictions for mass spectra of the same compound pair can be applied successfully in practice, e.g. for library matching or analogue search, or when measuring multiple spectra of the same compound at various collision energies.

Using Monte-Carlo dropout ensemble models to estimate prediction uncertainty

Using ensembles of multiple machine learning models is a frequently used technique for improving machine learning results, but also to assess the model uncertainty (also referred to as epistemic uncertainty). Ensembles can be built in many ways, but one particularly efficient ensemble learning technique for neural networks is Monte-Carlo dropout [17]. It makes use of the dropout layers in a network to randomly silence a fraction of the nodes for each inference step. Traditionally, dropout is only activated during model training, but in Monte-Carlo dropout it stays in place when making actual predictions, which can be interpreted as a random sampling technique across a virtually unlimited set of model variations. Since the neural network architecture used for MS2DeepScore includes dropout layers (Fig. 1), it is straightforward to do such ensemble learning (Fig. 6).

For a given spectrum, we compute n different embeddings, each from a slightly different version of the base neural network where 20% of its nodes are silenced (dropout rate = 0.2). For a pair of spectra this results in $n \times n$ Tanimoto score predictions from which an ensemble score as well as a dispersion measure to assess the prediction certainty can be calculated. To be less sensitive to outliers, we chose to take the median score, rather than the mean score. The prediction uncertainty is measured by the interquartile range (25–75%) which is more suited than the median absolute deviation for non-symmetric distributions [27]. This is also very accessible computationally since only n embeddings need to be generated per spectrum to obtain a total of $n \times n$ independent Tanimoto score predictions. Inference will hence only take $10 \times$ longer for an ensemble of 100 predictions.

We tested the resulting uncertainty estimate on all possible pairs within the 3601 spectra of our test set. Taking the median of an ensemble of 100 scores already results in an overall drop in prediction error across nearly all Tanimoto score bins (Fig. 7B, red vs. blue line). We then

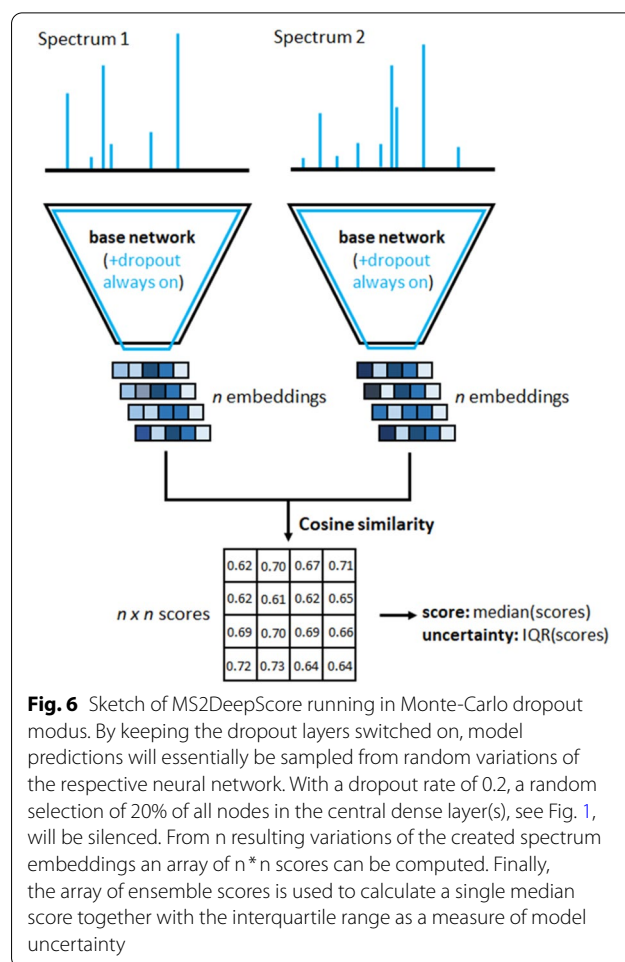
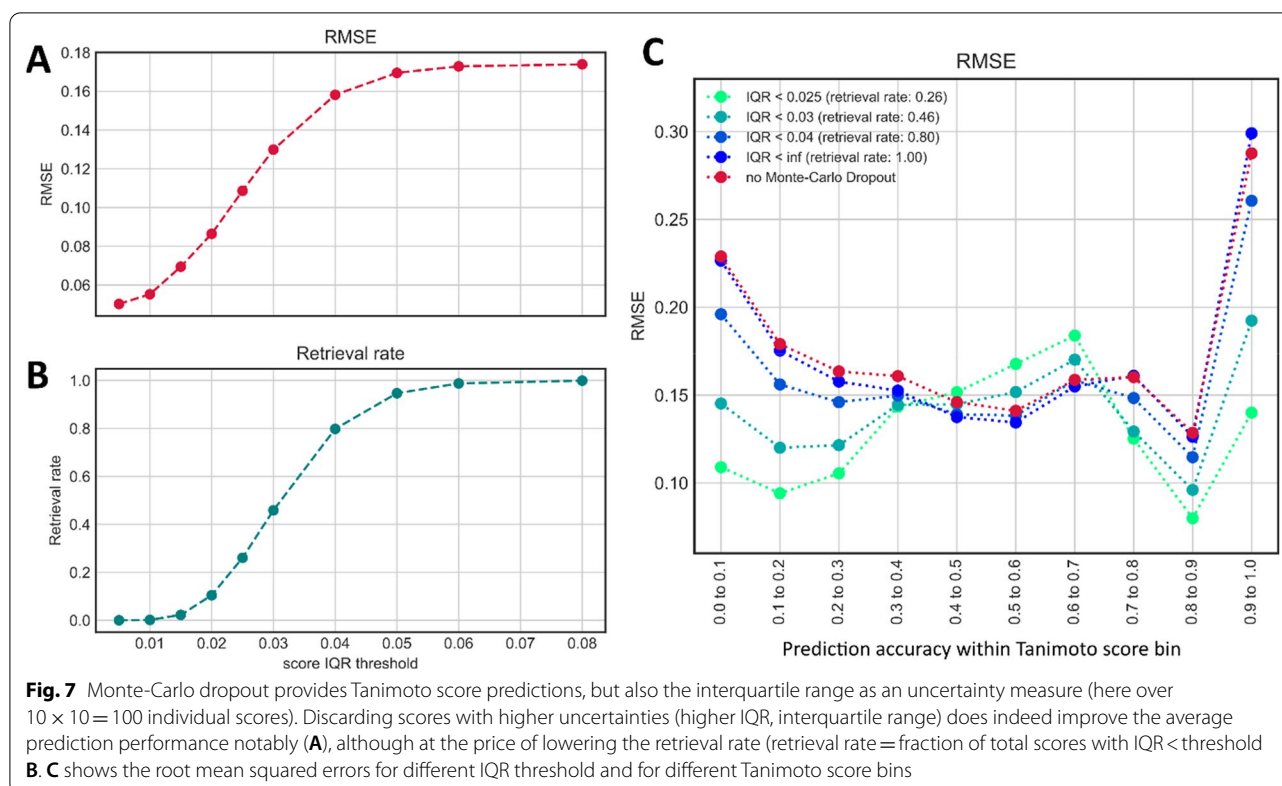


Fig. 6 Sketch of MS2DeepScore running in Monte-Carlo dropout modus. By keeping the dropout layers switched on, model predictions will essentially be sampled from random variations of the respective neural network. With a dropout rate of 0.2, a random selection of 20% of all nodes in the central dense layer(s), see Fig. 1, will be silenced. From n resulting variations of the created spectrum embeddings an array of $n \times n$ scores can be computed. Finally, the array of ensemble scores is used to calculate a single median score together with the interquartile range as a measure of model uncertainty

filtered out scores, according to increasingly stringent interquartile range (IQR) thresholds. Over the entire dataset, this approach leads to a large decrease in prediction error (Fig. 7A) but comes at the cost of a lower retrieval rate which we here define as the fraction of total scores for which the IQR is below the set threshold. For instance, all predictions within $IQR < 0.025$ —which will discard about 75% of the scores—will result in a drop of the average RMSE from about 0.17 to about 0.11 (Fig. 7A). It is important to note, though, that this average gain in precision is not distributed equally across the full range of Tanimoto scores. The RMSE drops most significantly in the low (<0.4) and high (>0.8) Tanimoto score range (Fig. 7B), while the error slightly increases in the mid score range (0.5–0.7).

Embedding based mass spectral clustering

Unlike recent approaches to predict molecular fingerprints using deep learning [11, 12], we have chosen to train a Siamese neural network [15] to directly predict Tanimoto similarities. A key feature of our neural

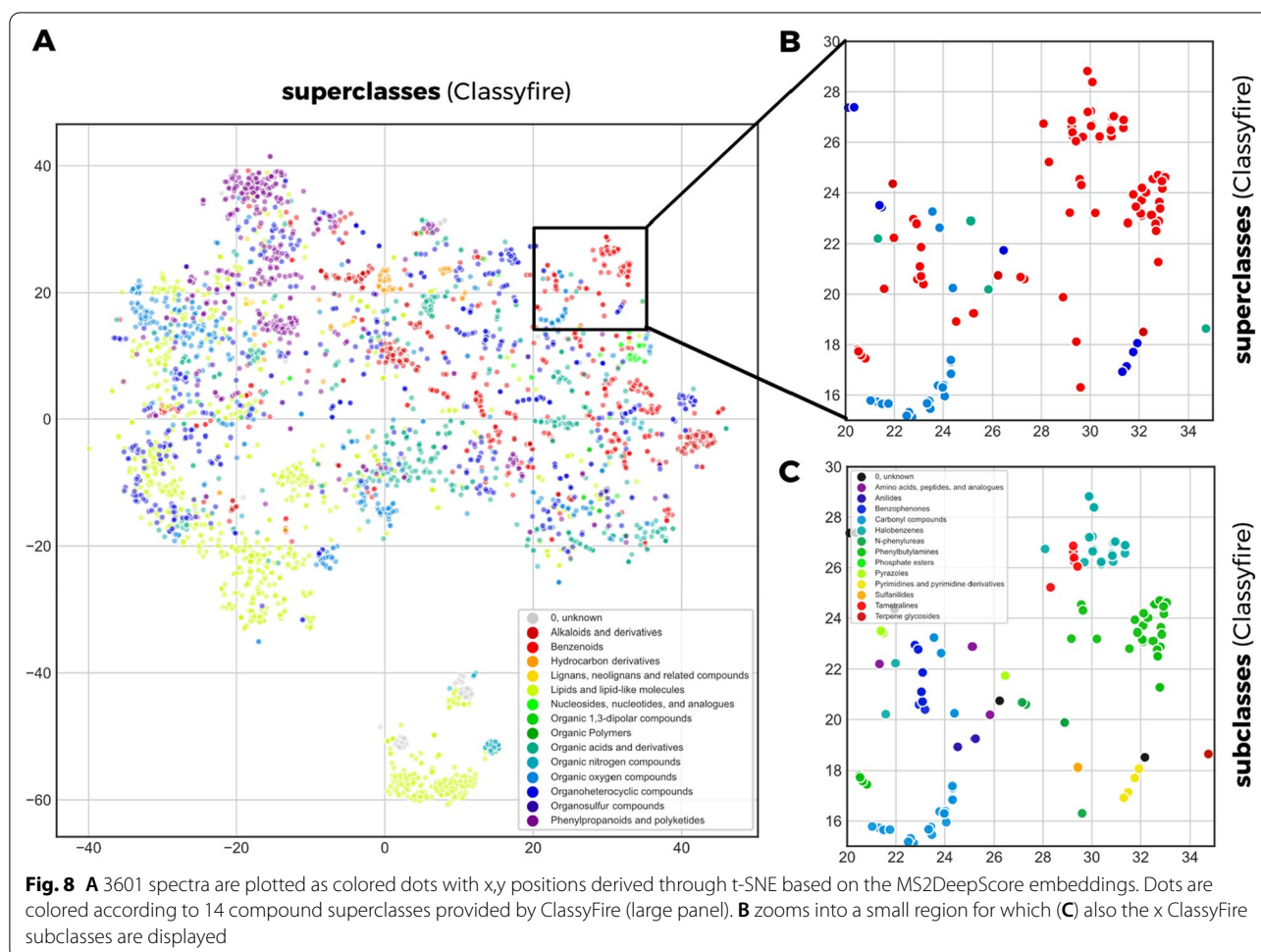


network design is the creation of abstract embedding vectors for each input spectrum (Fig. 1). This has two main benefits. First, it allows to scale similarity calculations much more efficiently to very large numbers of spectrum pair calculations by separating the mass spectrum embedding creation step from the actual similarity score calculation. The embedding creation includes the mass spectrum binning as well as the inference step with the ‘base’ neural network (Fig. 1) and is computationally far more expensive than the actual score calculation. Because embedding creation only needs to happen once for each spectrum instead of for each pair, this vastly reduces computational cost. As an example: predicting all possible similarity scores between the 3601 spectra in the test set (6,485,401 unique pairs) took 5–10 min on an Intel i7-8550U CPU. The second reason for choosing this network architecture design is that such embeddings can have additional value beyond the Tanimoto score prediction. Even though they do not directly correspond to any conventional molecular fingerprint, they are trained to support a prediction of a fingerprint-based similarity score, and therefore we hypothesize that they will contain features that reflect chemical properties. Embeddings also facilitate rapid large-scale comparisons beyond simple pair-wise similarity computations. To illustrate possible future use-cases, we ran t-SNE [28] as implemented

in scikit-learn [29] on the 200-dimensional embeddings of the test set (3601 spectra). This algorithm provides x,y-coordinates for every spectrum in the test set which we plotted and colored according to the 14 chemical superclasses provided by ClassyFire [30] (Fig. 8a). Molecules of the same chemical class tend to cluster together in the resulting t-SNE plot, confirming that the MS2DeepScore embeddings represent chemically meaningful molecular features. Figures 8b, c show that this conclusion also holds on a more detailed level, by zooming into a small region of the t-SNE plot (Fig. 8b) and coloring according to ClassyFire subclasses (Fig. 8c). To put those plots in perspective, side-by-side comparisons with t-SNE plots based on modified cosine scores are provided in Additional file 1: Figure S11.

MS2DeepScore python library

MS2DeepScore is available as an easily installable Python library running on Python 3.7 and 3.8. Source code and installation instructions can be found on GitHub (<https://github.com/matchms/ms2deepscore>). The presented results were obtained using version 0.2.0. MS2DeepScore is integrated in matchms, a recently developed Python library for mass spectra import, handling and comparisons [18].



Discussion

Modern deep learning techniques have quickly gained popularity in many research fields and in some cases even started to replace classical, more heuristic techniques (e.g., in computer vision and natural language processing). The application of deep learning on fragmentation mass spectrometry data though, has only just begun to enter the stage [13]. The first promising applications include the prediction of compound classes from MS/MS spectra [31] or from (predicted) molecular fingerprints [32, 33], the prediction of bioactivity signatures [34], the prediction of parts of molecular fingerprints [11, 12], as well as the prediction of the structural similarity from MS/MS spectra and chemical formula [14]. With MS2DeepScore, we show for the first time that neural networks can also be used to predict structural similarity scores, i.e., to obtain a chemical-driven measure, from MS/MS spectra without requiring a known molecular formula or other metadata. We found that predictions are generally accurate (MAE of about 0.12) and get the general tendency right with large outliers being rare

(RMSE of about 0.15). By constructing the test, validation, and training sets from separate sets of molecules we show that the presented MS2DeepScore models are predictive for novel molecules. Selecting all spectra for 500 randomly chosen compounds for our test set should reflect the overall diversity of the MS/MS dataset well enough. In addition, we observed that the distribution of Tanimoto scores within the test set shows a similar profile as the distribution for the training set (Fig. 2B vs. Figure 3A) and that the chemical diversity in the test set is high, with 14 chemical superclasses, 99 different chemical classes and 140 different chemical subclasses found via ClassyFire [30].

MS2DeepScore comes with two inherent downsides, when compared to conceptually simple, heuristic measures such as the Cosine spectral similarity score. One limitation that is common for all machine learning-based approaches is that a score itself is not deterministic but might change when a new model is trained (e.g., when using different parameters or different training data). In practice that can often largely be addressed by using

properly versioned, pre-trained models. The second limitation, which is typical for neural networks, is the lack of an intuitive explanation of *why* a certain score is given, a common problem in the field of deep learning [i.e., explainability or explainable artificial intelligence (AI)]. Here, however, we also provide the option to use Monte-Carlo Dropout, an ensemble learning technique which makes it possible to assess the model's own uncertainty. This should in practice help in better identifying and handling uncertain predictions, e.g. for data that are very different from the training data.

Another possible limitation of MS2DeepScore comes from the maximum achievable precision. Even after numerous experiments with training “deeper” neural networks (with more layers), wider neural networks (more nodes per layer), or less restrictive mass binning (more m/z bins), we could not significantly decrease the overall Tanimoto prediction error (Additional file 1). Obviously, that does not prove that there cannot be a better performing neural network for the given task. However, it strongly indicates that the achieved precision is already fairly high, given the many limitations of the used dataset. These limitations arise from the fact that the used spectra come from different instrumentation types, are of varying quality, and are likely to be very unbalanced with regard to represented compound classes and types. More general limitations lie in the possibility that some fingerprint features are simply not represented in any of the fragments. It was, however, possible to reduce prediction errors considerably by applying various ensemble learning techniques. Applying Monte-Carlo Dropout (Fig. 7) or using ensembles of different architecture models (Additional file 1: Figure S2) lead to more reliable results and provide means to assess the prediction uncertainty which allow users to further specify the desired level of precision. The same was seen when combining scores for spectra obtained for the same pair of molecules (Fig. 5). On our test set, the largest reduction of the prediction error is seen for the highest Tanimoto scores (0.9–1.0). Both combining scores for same molecules and applying a threshold for the prediction uncertainty as measured by Monte-Carlo Dropout, can remove a large part of the incorrect predictions (Figs. 5A and 7C). This indicates that many of those predictions stem from spectra which differ notably from the training data. Since such high Tanimoto scores are rare (Figs. 2C and 3B), a relatively small amount of such incorrect predictions could already explain the observed behavior.

It is important to note that the neural network was not trained on any spectrum metadata such as parent mass and elemental formula, like for DeepMASS [14]. Such metadata could include parent mass, precursor ion charge, adduct information, instrumentation type,

spectral quality, or more processed information such as the elemental formula or chemical compound class. We speculate that incorporating relevant metadata in the pairwise predictions would have increased the accuracy in our evaluation. However, here, we chose not to include such metadata. Having a way to predict structural similarities solely based on MS/MS peaks allows MS2DeepScore to be easily applied to a large number of spectra without costly and timely spectral processing or matching steps. If users have access to this metadata, they can anyway still use it for an independent selection step or to train an additional small model for removing likely outliers based on metadata pairs. Furthermore, the use of just mass fragments also makes MS2DeepScore applicable to GCMS data which generally lack precursor masses and to serve as an alternative metric to build mass spectral networks of volatile compounds [35].

We expect that another promising route to further improve the predicted scores lies in using complementary aspects of different spectral similarity scores. MS2DeepScore usually comes very close with its Tanimoto score predictions but might not always be precise enough to handle all nuances. It is—for instance—difficult to discriminate between high Tanimoto scores (say 0.8–0.9) and a near-complete chemical match. Reliable identification of exact compound matches hence requires additional algorithms, such as the successful use of machine learning in combination with computational fragmentation trees as well as with library data [36]. Key advantages of MS2DeepScore over such an approach are the very large gains in computation time which will allow to run very extensive screenings between many thousands of compounds, and its ability to predict structural similarities based on spectra of novel molecules without having to use any metadata or having to consult with library data. In practice, we expect hybrid approaches that combine multiple algorithms to be a promising route forward. MS2DeepScore could be applied for preselecting candidates prior to a computationally more expensive step (e.g. using fragmentation trees) or it could be combined with other similarity scores to improve the prediction reliability, e.g., by also consulting other scores such as cosine-based spectral similarity scores or Spec2Vec. The latter is frequently outperformed by MS2DeepScore (Fig. 5 and Additional file 1: Figure S1) but—as an unsupervised approach—has the advantage that it can be trained on unlabeled data which is not directly accessible for the presented supervised approach.

Being able to predict Tanimoto scores, or more precisely Tanimoto scores computed from Daylight2048bit fingerprints available in RDKit, can be interpreted as being able to infer chemical relatedness from MS/MS spectra. There is, however, no consensus on how to best

quantify chemical relatedness which resulted in a large variety of different molecular fingerprints [7, 8, 21] as well as fingerprint based scores [24, 37]. We showed that MS2DeepScore can be trained to predict various scores such as Tanimoto on Daylight fingerprints, Tanimoto on Morgan-2 (similar to ECFP-4), or Dice scores on Morgan-2. Given the huge variety of fingerprint types, their dimension (number of bits) as well as the used scoring metrics, our explorations should only be seen as a starting point, but our observations already suggest that MS2DeepScore will be able to cope reasonably well with a large variety of different fingerprints and metrics (Fig. 3, Additional file 1: Table S1). Since our neural network creates low dimensional embeddings for each spectrum it will also be possible to combine different structural similarity measures by stacking embeddings of different models that were trained on different scores. Another future path might be to modify the Siamese network architecture to predict various structural similarity scores at the same time. We do note that to fuel future improvements on structural similarity prediction of MS/MS spectrum pairs, it is vital that the field converges onto a standardized way of evaluating and comparing approaches. In that light, we have put effort to share processed input data, the used code, and the resulting models for future comparisons.

MS2DeepScore comes as an easy to install and easy to use Python library and the actual scores are fast to compute. In particular, the ability to split spectral embedding creation from similarity score calculation makes it very scalable to large-scale comparisons (many thousands of spectra). Our MS2DeepScore model which was trained on a public dataset of about 100,000 spectra of 15,000 compounds can be found online (see link in “Methods”). Even though we here showed that such a model performed well on spectral data of unseen compounds, it is to be expected that training on even larger, more diverse, or more curated datasets will further improve the model performance. We are aware that current spectral libraries remain inherently biased toward certain compound classes and this needs to be considered when applying MS2DeepScore to experimental data. In this light, training (and ideally: providing) new MS2DeepScore models on alternative spectral libraries such as METLIN [38] or NIST [39] could become an important step to improve neural network based predictions.

Finally, we speculate that MS2DeepScore-generated spectral embeddings can further be used for other fascinating tasks. Unlike alternative approaches that directly focused on using deep learning for creating low-dimensional embeddings [40], we trained the network on pairwise Tanimoto predictions. Yet we found that the spectral embeddings allow for chemically meaningful clustering

(Fig. 8). This also makes MS2DeepScore a promising complementary candidate to the mass spectral metrics used in established mass spectrometry based network analysis and clustering tools such as GNPS [16] or MetGem [41].

Conclusions

MS2DeepScore is a deep learning technique to predict structural similarity scores between fragmentation mass spectral pairs. We show that MS2DeepScore can infer structural similarities between mass spectra with high overall precision, without requiring any additional metadata or library data. We demonstrate that the accuracy of the predictions can be improved notably by using various ensemble learning techniques, in particular by merging predicted scores of spectra belonging to the same compound pair or by applying Monte-Carlo Dropout to sample from random model variations.

MS2DeepScore is very fast and scalable. We conclude that this makes MS2DeepScore a powerful novel tool for running large scale comparisons and analyses, for instance on complex mixtures rich in spectra of unknown compounds. We expect that MS2DeepScore can generally be used to complement -or replace- common currently used spectral similarity measures in many metabolomic workflows, including mass spectral network analysis and clustering approaches.

Methods

Data and data preparation

Spectrum data preparation

We use annotated MS/MS spectra from GNPS, which underwent basic metadata cleaning as described in [6, 18]. The dataset was retrieved from GNPS (25/01/2021) and contains a total of 210,407 MS/MS spectra. Metadata was cleaned and checked using matchms [18] version 0.8.2, which included cleaning compound names, extracting adduct information from the given metadata, moving metadata to consistent fields and conversions between InChI and SMILES as well as to InChIKeys when missing and when possible. We then ran an automated search against PubChem [42] using pubchempy [43] for spectra which still missed InChI or SMILES annotations. The full cleaned dataset (210,407 spectra, 184,698 annotated with InChIKey and SMILES and/or InChI) can be found on zenodo: <https://zenodo.org/record/4699300>.

We here focus on spectra acquired in positive ionisation mode with proper InChIKey as well as a SMILES and/or InChI annotation, which in addition must contain ≥ 5 peaks between 10.0 and 1000.0 Da. This resulted in 109,734 spectra with 15,062 unique InChIKeys (considering only the first 14 characters).

The spectra underwent basic filtering to remove excessive amounts of peaks, by removing peaks with intensities < 0.1% of the maximum peak intensity and limiting the maximum number of peaks to the 1000 highest intensity peaks. This is mostly done to speed up the later binning and training steps and the hence removed peaks are most likely noise peaks. Peak intensities were square root transformed to avoid a too strong focus on the highest intensity peaks only. Spectrum peaks were binned in 10,000 equally-sized bins ranging from 10 to 1000 m/z. In case multiple peaks ended up in one bin the highest peak intensity was chosen as value for that bin. Bins that were not filled in any of the training-data spectrum representations were removed from the vector representation, here this meant that 9948 out of 10,000 possible bins are known to the model. The resulting vector-representation of the spectra served as input for the model.

Structural similarity label for spectrum pairs

Unless noted otherwise, we used Tanimoto scores on RDKit [23] Daylight fingerprints (2048 bits) to compute structural similarities. For every unique 14-character InChIKey the most common InChI was selected (if different InChI existed) and used to generate a molecular fingerprint (as implemented in matchms [18]). For each pair of molecular fingerprints Tanimoto scores were calculated, indicating the structural similarity of that pair. This resulted in a matrix of $15,062 \times 15,062$ Tanimoto scores to be used as labels for the model training.

Data generation

The set of 15,062 InChIKeys was split into a training ($n=14,062$), validation ($n=500$), and test set ($n=500$). To feed the data to the model effectively it was key to solve 2 challenges: (1) The structural similarity label distribution for all pairs is heavily left-skewed (most pairs are not similar, see Fig. 3A). (2) Per unique InChIKey multiple spectra could be used. Our MS2DeepScore Python library offers two types of data generators, one which iterates over all unique InChIKeys (*DataGeneratorAllInchikeys*) and one which iterates over all spectra and was used for the presented work (*DataGeneratorAllSpectrums*). The following algorithm was used to generate one cycle of training data, in each training epoch we used one cycle (i.e. we went through all spectrums in the training set once): each spectrum was then matched to a random other spectrum, with the condition that the resulting corresponding InChIKey pair had a structural similarity label falling into a randomly chosen bin, which in our case were 10 equally-sized bins between 0 and 1. In cases where the structural similarity label for none of the pairs fell into the selected bin, the bin was iteratively widened by 0.1 until a structural similarity label fell into the bin.

After every training epoch, the loss on the validation set was computed. As for the training data we here used *DataGeneratorAllSpectrums* on the validation set. To ensure dataset consistency across experiments we used a fixed random seed for the validation set. We also used 10 cycles for the validation set which means iterating 10 times over all 3597 spectra to monitor the training progress on a total of $n=35,970$ spectrum pairs. For the final evaluation on the reserved test set, we used all possible spectrum pairs between the 3601 for the test set ($n=6,485,401$ unique spectrum pairs).

Data augmentation

To ensure the model generalizes well to the test dataset and avoid overfitting we applied three forms of data augmentation on the binned spectra. (1) low-intensity peak removal: For a randomly chosen percentage (in the range of 0–20%) of non-zero bins with an intensity below 0.4 (actual intensity before transformation) the intensity was set to 0. (2) peak intensity jitter: Each non-zero bin intensity (after transformation) underwent changes between 0 and $\pm 40\%$. (3) new peak addition: For each of between 0–10 randomly selected zero-intensity bins that bin's intensity was set to random values between 0 and 0.01 (after transformation). Data augmentation was applied for every training example during training data generation.

Deep learning implementation

Network architecture

We train a deep learning network on pairs of MS/MS spectra to predict the respective structural similarity label. For this, a Siamese network is used [15] which has 2 components: (1) A base network that creates abstract embeddings from both input spectra, and (2) A “head” part of the Siamese network which consists of a cosine calculation between both embeddings (Fig. 1). In the base network, the binned spectrum vector is passed through a series of densely connected layers until an abstract embedding vector of desired dimension is created as output. Based on a screening of various key parameters (see Additional file 1) we settled on an architecture as depicted in Fig. 1: binning spectrum peaks between 10.0 and 1000.0 Da into maximum 10,000 same-width bins. This input vector is then followed by 2 densely connected layers, each with 500 nodes, followed by a final dense layer of 200 nodes for creating the spectral embedding. Two key measures are taken to prevent overfitting and improve generalization of the model to unknown data. Firstly, modern regularization techniques are applied [44]. The deep neural network is trained using L1 (10^{-6}) and L2 (10^{-6}) weight regularization in the first dense layer, as well as dropout in the subsequent layers

(dropout rate = 0.2). In addition, batch normalization is applied after each dense layer except the output layer.

Uncertainty quantification using Monte-Carlo Dropout ensembles

To estimate the uncertainty of a prediction we used Monte-Carlo Dropout ensembles [17]. At inference time, dropout was applied to all but the first layer of the base network. $N = 10$ embeddings were created from an ensemble of these networks with dropout enabled. This resulted in a distribution of structural similarity predictions, for which the median and interquartile range (IQR) were calculated.

Training details

Models were trained with the Adam optimizer [44, 45] that optimized the mean squared error (MSE) loss. We used a batch size of 32, and a learning rate of 0.001. Training continued until the validation loss did not decrease for 5 epochs (early stopping). Model training was done on GPU nodes from SURFsara with nvidia GTX 1080 Ti graphic cards (Lisa cluster).

Precision/recall analysis for selecting high Tanimoto score pairs

The precision/recall plot in Fig. 4 was created by measuring how many pairs with Tanimoto scores above a set threshold (= "high structural similarity pair") were among a subset of all pairs for which the spectral similarity score was $> \text{threshold_score}$. We varied the threshold score from 0 to close to 1 and recorded the precision and recall. By precision we here understand the number of high structural similarity pairs in the selection divided by the number of all selected pairs. Recall refers to the number of high structural similarity pairs in the selection divided by all high structural similarity pairs.

T-SNE on mass spectral embeddings from MS2DeepScore

For Fig. 8, we used the MS2DeepScore base network (Fig. 1) to compute the 200-dimensional spectral embeddings for all 3601 spectra in the test set. Using the t-SNE [28] implementation from scikit-learn [29] we computed two-dimensional coordinates for all spectra. Here we used the following settings: metric = 'cosine', perplexity = 100, learning-rate = 200 (default) and 1000 iterations (default).

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13321-021-00558-4>.

Additional file 1. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. **Figure S1.** Score comparisons with respect to precision-recall in highly unbalanced dataset. **Figure S2–S8.** Hyperparameter search. **Figure S9** Early stopping. **Figure S10** Assessing prediction uncertainty using Monte-Carlo Dropout. **Table S1.** Measuring the linear correlation between actual and predicted scores. **Figure S11** t-SNE using MS2DeepScore or modified cosine score.

Acknowledgements

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

The authors thank the GNPS community for contributing annotated spectra to the public GNPS spectral library.

Authors' contributions

FH, JJJvdH, and LR worked on the conception of the work. FH and SvdB designed the deep learning method and wrote the software. FH ran the analysis. All authors contributed to the interpretation of the results and the writing of the manuscript. All authors read and approved the final manuscript.

Funding

JJJ.v.d.H. acknowledges funding from an ASDI eScience grant, ASDI.2017.030, from the Netherlands eScience Center.

Availability of data and materials

MS2DeepScore is available as an easily installable Python library running on Python 3.7 and 3.8. Source code and installation instructions can be found on GitHub (<https://github.com/matchms/ms2deepscore>). The presented results were obtained using version 0.2.0. MS2DeepScore is integrated in matchms, a recently developed Python library for mass spectra import, handling and comparisons [18].

The fully trained model used to create Fig. 2, 4, 5, 7, 8 can be downloaded from zenodo: <https://zenodo.org/record/4699356>.

The code used to run the presented experiments and create the presented figures can be found on GitHub as Jupyter notebooks [46]: <https://github.com/matchms/ms2deepscore/tree/main/notebooks>.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

¹Netherlands eScience Center, 1098 XG Amsterdam, The Netherlands.

²Bioinformatics Group, Wageningen University, 6708 PB Wageningen, The Netherlands.

Received: 9 July 2021 Accepted: 25 September 2021

Published online: 29 October 2021

References

1. Blaženović I, Kind T, Ji J, Fiehn O (2018) software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites* 8(2):31
2. Lai Z, Tsugawa H, Wohlgemuth G, Mehta S, Mueller M, Zheng Y et al (2018) Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat Methods* 15(1):53–56

3. Ramos AEF, Evanno L, Poupon E, Champy P, Benidrir MA (2019) Natural products targeting strategies involving molecular networking: different manners, one goal. *Nat Prod Rep* 36(7):960–980
4. Aron AT, Gentry EC, McPhail KL, Nothias L-F, Nothias-Esposito M, Bouslimani A et al (2020) Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat Protoc* 15(6):1954–1991
5. Schollée JE, Schymanski EL, Stravs MA, Gulde R, Thomaidis NS, Hollender J (2017) Similarity of high-resolution tandem mass spectrometry spectra of structurally related micropollutants and transformation products. *J Am Soc Mass Spectrom* 28(12):2692–2704
6. Huber F, Ridder L, Verhoeven S, Spaaks JH, Diblen F, Rogers S et al (2021) Spec2Vec: improved mass spectral similarity scoring through learning of structural relationships. *PLOS Comput Biol* 17(2):e1008724
7. Riniker S, Landrum GA (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J Cheminform* 5(1):26
8. Muegge I, Mukherjee P (2016) An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin Drug Discov* 11(2):137–148
9. da Silva RR, Dorrestein PC, Quinn RA (2015) Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci U S A* 112(41):12549–12550
10. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A* 112(41):12580–12585
11. Ji H, Deng H, Lu H, Zhang Z (2020) Predicting a molecular fingerprint from an electron ionization mass spectrum with deep neural networks. *Anal Chem* 92(13):8649–8653
12. Fan Z, Alley A, Ghaffari K, Ransom HW (2020) MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation. *Metabolomics* 16(10):104
13. Liu Y, Vijlder TD, Bittremieux W, Laukens K, Heyndrickx W (2021) Current and future deep learning algorithms for tandem mass spectrometry (MS/MS)-based small molecule structure elucidation. *Rapid Commun Mass Spectrom*. <https://doi.org/10.1002/rcm.9120>
14. Ji H, Xu Y, Lu H, Zhang Z (2019) Deep MS/MS-aided structural-similarity scoring for unknown metabolite identification. *Anal Chem* 91(9):5629–5637
15. Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R (1993) Signature verification using a “Siamese” time delay neural network. In: *Proceedings of the 6th international conference on neural information processing systems*. San Francisco: Morgan Kaufmann Publishers Inc., pp 737–44. (NIPS’93)
16. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y et al (2016) Sharing and community curation of mass spectrometry data with GNPS. *Nat Biotechnol* 34(8):828–837
17. Gal Y, Ghahramani Z (2016) Dropout as a bayesian approximation: representing model uncertainty in deep learning. *ArXiv150602142 Cs Stat*. <http://arxiv.org/abs/1506.02142>. Accessed 1 Apr 2021
18. Huber F, Verhoeven S, Meijer C, Spreeuw H, Castilla EMV, Geng C et al (2020) matchms—processing and similarity evaluation of mass spectrometry data. *J Open Source Softw* 5(52):2411
19. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28(1):31–36
20. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D (2015) InChI, the IUPAC international chemical identifier. *J Cheminform* 7(1):23
21. Capecchi A, Probst D, Reymond J-L (2020) One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J Cheminform* 12(1):43
22. Bender A, Jenkins JL, Scheiber J, Sukuru SCK, Glick M, Davies JW (2009) How similar are similarity searching methods? a principal component analysis of molecular descriptor space. *J Chem Inf Model* 49(1):108–119
23. Landrum G. RDKit: open-source cheminformatics. <https://zenodo.org/record/3732262#.YXFdhxCQ2w>
24. Bajusz D, Rácz A, Héberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* 7(1):20
25. Demuth W, Karlovits M, Varmuza K (2004) Spectral similarity versus structural similarity: mass spectrometry. *Anal Chim Acta* 516(1):75–85
26. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD et al (2012) Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci U S A* 109(26):E1743–1752
27. Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. *J Am Stat Assoc* 88(424):1273–1283
28. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(86):2579–2605
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12(85):2825–2830
30. Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G et al (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminform* 8(1):61
31. Xing S, Jiao Y, Salehzadeh M, Soma KK, Huan T (2021) SteroidXtract: deep learning-based pattern recognition enables comprehensive and rapid extraction of steroid-like metabolic features for automated biology-driven metabolomics. *Anal Chem*. <https://doi.org/10.1021/acs.analchem.0c04834>
32. Dührkop K, Nothias L-F, Fleischauer M, Reher R, Ludwig M, Hoffmann MA et al (2020) Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol* 39:1–10
33. Kim H, Wang M, Leber C, Nothias L-F, Reher R, Kang KB, et al (2020) NPClassifier: a deep neural network-based structural classification tool for natural products. https://www.articles/preprint/NPClassifier_A_Deep_Neural_Network-Based_Structural_Classification_Tool_for_Natural_Products/12885494/1. Accessed 3 Apr 2021
34. Bertoni M, Duran-Frigola M, Badia-i-Mompel P, Pauls E, Orozco-Ruiz M, Guitart-Pla O et al (2020) Bioactivity descriptors for uncharacterized compounds. *bioRxiv*. <https://doi.org/10.1101/2020.07.21.214197v2>
35. Aksenov AA, Laponogov I, Zhang Z, Doran SLF, Belluomo I, Veselkov D et al (2021) Auto-deconvolution and molecular networking of gas chromatography–mass spectrometry data. *Nat Biotechnol* 39(2):169–173
36. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M et al (2019) SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* 16(4):299–302
37. Rácz A, Bajusz D, Héberger K (2018) Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. *J Cheminform* 10(1):48
38. Guijas C, Montenegro-Burke JR, Domingo-Almenara X, Palermo A, Warth B, Hermann G et al (2018) METLIN: a technology platform for identifying knowns and unknowns. *Anal Chem* 90(5):3156–3164
39. Kramida A, Ralchenko Y, Reader J, NIST ASD Team (2020) NIST Atomic Spectra Database (ver. 5.8). <https://doi.org/10.18434/T4W30F>
40. May DH, Birmes J, Noble WS (2018) A learned embedding for efficient joint analysis of millions of mass spectra. *bioRxiv*. <https://doi.org/10.1101/483263v1>
41. Olivon F, Elie N, Grelier G, Roussi F, Litaudon M, Touboul D (2018) MetGem software for the generation of molecular networks based on the t-SNE algorithm. *Anal Chem* 90(23):13900–13908
42. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S et al (2019) PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 47(D1):D1102–D1109
43. Swain M, Sjögren R, Zachcp, Yi H, Lazzaro L, Dahlgren B (2017) mcs07/ PubChemPy: PubChemPy v1.0.4. Zenodo. https://zenodo.org/record/541438#.Xyq_ROexUUE. Accessed 5 Aug 2020
44. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. The MIT Press. <http://www.deeplearningbook.org>. Accessed 1 Apr 2021
45. Kingma DP, Adam BJ (2017) A method for stochastic optimization. *ArXiv1412.6980 Cs*. <http://arxiv.org/abs/1412.6980>. Accessed 2 Apr 2021
46. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M et al (2016) Jupyter Notebooks—a publishing format for reproducible computational workflows. In: Kluyver T (ed) *Position das*. IOS Press, Amsterdam, pp 87–90

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.