SPEC2MOL: AN END-TO-END DEEP LEARNING FRAMEWORK FOR TRANSLATING MS/MS SPECTRA TO DE-NOVO MOLECULES

Eleni E. Litsa[1], **Vijil Chenthamarakshan**[2], **Payel Das**[2,3] [*], and **Lydia E. Kavraki**[1] [†]

[1]Department of Computer Science, Rice University, Houston, TX
[2]IBM Research, IBM Thomas J. Watson Research Center, Yorktown Heights, NY
[3]Applied Physics and Applied Mathematics, Columbia University, New York, NY

ABSTRACT

Elucidating the structure of a chemical compound is a fundamental task in chemistry with application in multiple domains including the emerging field of metabolomics, with promising applications in drug discovery, precision medicine, and biomarker discovery. The common practice for elucidating the structure of a chemical compound is to obtain a mass spectrum and subsequently retrieve its structure from spectral databases. However, database retrieval methods fail to identify novel molecules that are not present in the reference database. In this work, we propose Spec2Mol, a deep learning architecture for molecular structure recommendation given mass spectra alone. Spec2Mol is inspired by the Speech2Text deep learning architectures for translating audio signals into text. Our approach is based on an encoder-decoder architecture. The encoder learns the spectra embeddings, while the decoder, pre-trained on a massive dataset of chemical structures for translating between different molecular representations, reconstructs SMILES sequences of the recommended chemical structures. We have evaluated Spec2Mol by assessing the molecular similarity between the recommended structures and the original structure. Our analysis showed that Spec2Mol is able to identify the presence of key substructures in the molecule from its mass spectrum, and shows on par performance, when compared to existing fragmentation tree based methods, in recommending molecules for a given mass spectrum.

1 Introduction

The identification of the chemical compounds that are present in a sample of chemical matter is a fundamental task in chemical analysis with applications in multiple domains. The field of metabolomics, for example, seeks to identify the chemical molecules that are present in a biological sample. In humans, the metabolome, that is the set of all chemical molecules that can be found in human tissues, is a great source for biomarker discovery as it reflects changes at a genetic, proteomic or environmental level [1]. Additionally, mapping the human metabolome will lead to a better understanding of human physiology and disease etiology and pathology which is essential for the identification of new therapeutic targets for developing new treatments. There increasing interest in mapping the metabolome extends to other organisms as well, such as plants which have been a great source of bioactive compounds for multiple products including drugs and supplements [2]. The identification of chemical compounds is also critical in product development such as in the production of pharmaceuticals and agrochemicals. Structure elucidation practices are being used for quality control and detection of impurities as well as in safety studies for identifying potential metabolites that can be formed in the human body. Finally, structure elucidation practices are being employed in forensics analysis.

The identification of the structure of a chemical compound is perceived as one of the most time consuming and laborious task in chemical analysis. This is often performed through analytical techniques such as mass spectroscopy (MS) and nuclear magnetic resonance (NMR) [3, 4, 5] with MS being used more often due to its higher sensitivity and specificity [3]. In MS, the molecules that are present in a biological sample are first separated using a chromatographic technique, such as liquid chromatography (LC) and gas chromatography (GC), with the latter being used more commonly [1, 6]. After the separation, the molecule is fragmented into positive or negative charged ions using an ionization source such

[*]daspa@us.ibm.com
[†]kavraki@rice.edu

as electron ionization (EI), chemical ionization (CI) and electrospray ionization source (ESI) [1, 6]. What the instrument records is the mass-to-charge (m/z) ratios of the generated fragment ions. The information that is collected from this process is presented in the mass spectrum which is a graph with the m/z of each recorded fragment in the horizontal axis and the relative abundance in the vertical axis. In order to obtain more detailed information on the query structure, a sequential fragmentation process is often used called tandem mass spectrometry [5]. Once the molecule has been fragmented into ions, a set of them, called precursor ions, is selected and further fragmented to generate MS2 (also called MS/MS) spectra. These second-level ions can be fragmented even further giving MS3 spectra and so on. The peaks and their intensity in the resulting spectrum depend not only on the structure of the chemical molecule that is being fragmented but also on the experimental conditions, that is the instrument used, the collision energy, the selected precursor ion and the ionization mode, as it is illustrated in Figure 1.

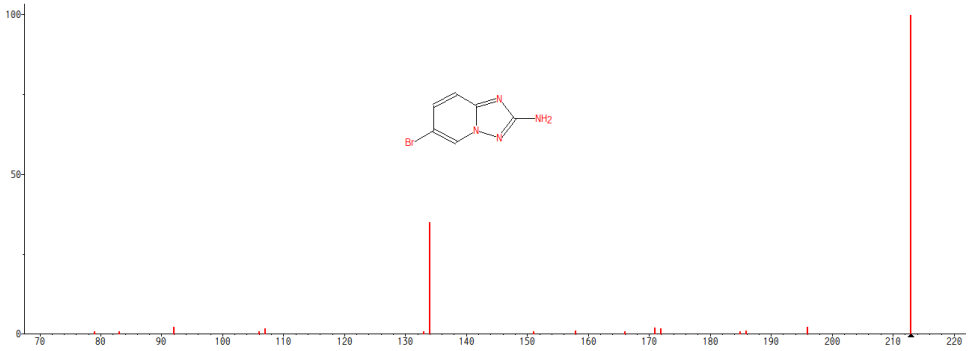
Once the mass spectrum is obtained, it is matched against the content of spectral databases of reference compounds in order to retrieve its structure. Various chemical databases provide spectra data of metabolites [7] such as Human Metabolome Database, METLIN, MassBank and mzCloud [7]. Certain databases are focused on the metabolites of specific organisms, such as the Human Metabolome Database, or on specific molecular classes, such as the LIPID MAPS Structure Database, while others have greater coverage such as METLIN. However, despite the intense ongoing efforts to map the metabolome of various organisms, existing databases cover only a small percentage of the actual metabolites that occur in organisms. Particularly for humans, it is estimated that less than 10% of metabolites have experimental reference mass spectra [8] which means that a large percentage of the current practice cannot identify a large percentage of the molecules that are found in human tissues. It is estimated that in untargeted metabolomics studies less than 2% of the detected spectral features are identified [8].

An approach that has been developed to address the problem of limited amount of experimental spectra data is *in silico* fragmentation which essentially attempts to solve the inverse problem. This approach aims at enhancing the content of existing spectra databases with computed spectra of known molecular structures which have no available experimental spectra. Essentially this approach seeks to close the gap between spectral and structural databases. *In silico* fragmentation tools predict the fragmentation process either relying on fragmentation rules or using combinatorial/optimization-based approaches or employing machine learning methodologies [6, 9, 10]. Fragment prediction methods have been especially successful for predicting spectra of peptides, however, fragmentation of small molecules into ions is a more stochastic process that is especially challenging to predict [6].

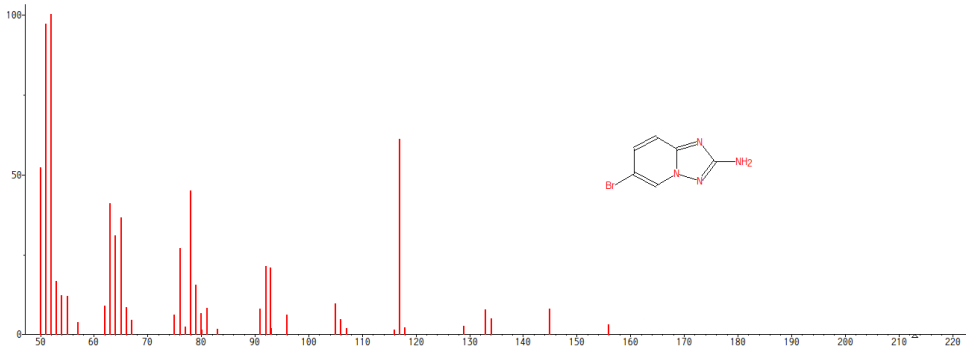
A more direct approach to the structure elucidation problem would be to reconstruct the underlying chemical structures given spectra features. Such an undertaking though is computationally challenging as it requires the generation of a molecular structure. Indeed, this approach is performed as a two step process to circumvent the need for generating molecular structures: A machine learning model is used to map the spectrum to an intermediate vector representation such as a molecular fingerprint. Once the fingerprint is obtained then it is matched against the content of structural databases in order to identify candidate molecular structures with similar fingerprints [11, 12]. This method though will also fail for molecules that are not present in the structural database and especially for novel molecules. A more direct association of spectra features with molecular structures through a rule-based approach has also been explored [13]. More specifically, this approach extracts rules, that associate spectra features with substructures, from spectra databases aiming at a partial structure identification.

An additional concept that has been introduced to facilitate the interpretation of mass spectra, and subsequently structure identification, is that of fragmentation trees [6, 14]. A fragmentation tree is derived computationally from tandem mass spectra using optimization algorithms such that its nodes correspond to fragments or precursor ions and the edges correspond to fragmentation reactions. Fragmentation trees have various uses such as identifying the molecular formula and clustering molecules by aligning fragmentation trees [15]. They have also been used for the prediction of molecular fingerprints that are subsequently used to search structural databases [16, 17]. The information in a mass spectrum is thought to be insufficient to explain the fragmentation process by itself while the fragmantation tree provides complementary information by elucidating the dependencies between the mass peaks [6]. However, fragmentation trees are expensive to compute and often approximations are preferred for practical applications.

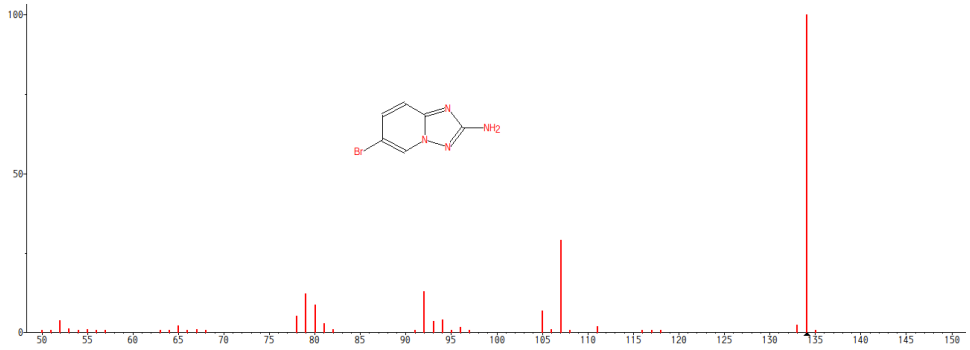
A more thorough review of existing methodologies for metabolite identification, including *in silico* fragmentation tools, fingerprint prediction and fragmentation trees, was recently presented by Nguyen et al. with a focus on machine learning (ML) approaches [6]. It should be noted here that early ML-based approaches were built on shallow ML models, such as Support Vector Machines (SVMs) and Random Forests (RFs), applied either on features extracted from the mass spectra or the fragmentation trees, and also kernel-based methods to judge similarity between either spectra or fragmentation trees. However, lately there is a growing interest in exploring Deep Learning (DL) architectures for the development of computation tools to support structure elucidation. There have been efforts to learn spectra embeddings that can be subsequently used to assess spectral similarity when searching in spectral databases [18, 12]. Additionally, there are DL-based methodologies for clustering spectra, either for identifying the compound class [19, 12] or for



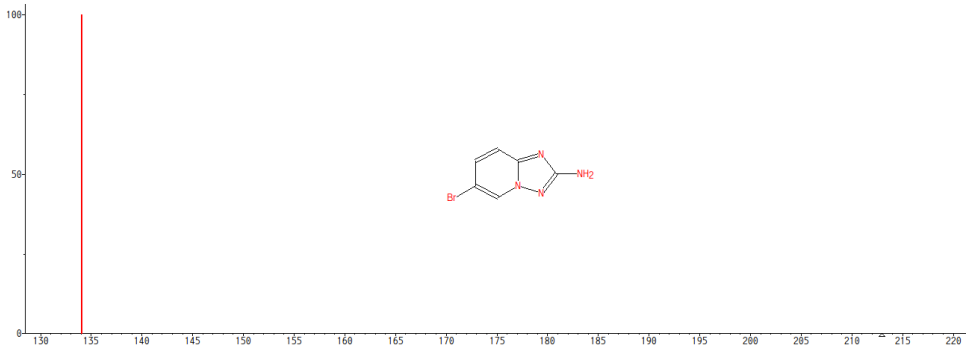
(a) Precursor ion: [M+H]+, NCE: 35%, Instrument: HCD



(b) Precursor ion: [M+H]+, NCE: 130%, Instrument: HCD



(c) Precursor ion: [M+H-Br]+, NCE: 35%, Instrument: HCD



(d) Precursor ion: [M+H+2i]+, NCE: 35%, Instrument: IT-FT

Figure 1: MS/MS spectra obtained through different experimental conditions (precursor ion, normalized collision energy (NCE), fragmentation method) from the same molecule. Spectra sourced from the NIST library.

medical diagnosis differentiating between healthy and cancerous tissues [20]. Most DL-based methodologies that operate directly on spectra data are based on Convolutional Neural Networks (CNNs) representing the spectrum as a vector that indicates the intensities of each fragment mass [20, 21, 22]. The CNN attempts to automatically identify spectra features replacing the need for manual featurization. Architectures that have adopted concepts from Natural Language Processing (NLP) have also emerged representing the mass spectrum as text and the mass peaks as words [18]. Due to the limited amount of mass spectra data different workarounds have been investigated including hybrid approaches [19], combining statistical ML models and DL architectures, and approaches based on transfer learning [20].

It should be noted that in parallel DL-based approaches are being investigated for identifying protein sequences from mass spectra in proteomics studies [23, 21, 22]. A noteworthy effort, DeepNovo, consists of an end-to-end DL architecture for de novo peptide sequencing from mass spectra [22], that is a direct reconstruction of the peptide sequence from the mass spectra data. Structure elucidation of small molecules though is perceived as a more challenging problem due to the stochastic nature of the fragmentation process. On top of that, the structure of small molecules has a graph-like representation as opposed to the linear nature of a peptide sequence. Existing approaches essentially attempt to retrieve molecules from structure databases that have a spectrum similar to the query spectrum. This method though cannot identify novel molecules, that is molecules whose structure currently remains unknown and therefore they do not exist in chemical databases.

In this paper, we present Spec2Mol, an end-to-end DL architecture for translating MS/MS spectra to molecular structures. Spec2Mol is intended for recommending molecular structures that can explain observed MS/MS spectra. We represent molecular structures as sequences using the SMILES notation [24] and MS/MS spectra as vectors of fragment intensities. Spec2Mol consists of an encoder, that learns an embedding for the MS/MS spectrum, and a decoder that generates the SMILES sequences of the recommended chemical molecules. Due to the limited amount of available spectra data our approach is based on unsupervised pre-training on a large dataset of unlabeled molecules. In particular, we pre-trained the decoder as part of an auto-encoder (AE) architecture which is trained to reconstruct a molecule through its SMILES sequence. The encoder is subsequently trained such that the spectra embeddings match the embeddings that the AE has learnt. The data used to develop and evaluate the model, the architecture of Spec2Mol, as well as, the evaluation of the model are described in the following sections.

The main contributions of this work are as follows:

- To our knowledge this is the first approach for generating potential molecular structures from mass spectrometry data that is not based solely on database retrieval.
- Our method can facilitate database retrieval and additionally de novo molecular structure recommendation.
- Our approach takes advantage of large datasets of unlabeled molecules using unsupervised pre-training.
- We introduce metrics to assess the similarity of the generated molecules with the reference ones and we perform a comparative evaluation with a widely accepted method that makes use of additional information, that is fragmentation trees.

2 Datasets and data pre-processing

Spec2Mol consists of an encoder that learns spectra embeddings and a pre-trained decoder, which has been trained as part of an autoencoder architecture. The autoencoder has been trained on a large set of molecules (molecule dataset discussed in section 2.2) while the encoder has been trained on a set of molecules for which MS/MS data are available (spectral dataset discussed in 2.1).

2.1 Spectral dataset

The mass spectra data for training the encoder has been derived from the NIST Tandem Mass Spectral Library 2020 which is a commercial dataset of more than 1M spectra obtained from more than 30K compounds [25, 26]. The largest percentage of the NIST dataset (60%) corresponds to metabolites (6K human metabolites and 8K plant metabolites) while a significant amount of the data is drugs (20%). The rest corresponds to peptides, lipids, forensics, surfactants/contaminants and sugars/glycans. The dataset contains low and high resolution MS/MS spectra, obtained through different fragmentation techniques. For a small number of molecules in the dataset there are available up to MS4 spectra. Each molecule in the dataset may be associated with more than one spectrum which may be obtained through different experimental conditions, that is, different fragmentation instrument, precursor ion, ionization mode, collision energy or fragmentation level (MS2, MS3 or MS4).

2.2 Molecule dataset

The autoencoder, from which the Spec2Mol decoder had derived, was pre-trained on about 110 million molecules which were sourced from the PubChem dataset [27]. The structures of the molecules in the PubChem dataset are represented using the SMILES notation [24]. Stereochemistry information was not indicated in the SMILES representation. The reason for not accounting for stereochemistry is that in the subsequent task of spectra translation recovering stereochemistry information from the mass spectra is especially challenging or possibly even impossible and therefore it is out of the scope of this work.

2.2.1 Data filtering

In order to minimize variations in the spectra data, due to differences in the experimental conditions, we chose to keep certain variables in the dataset fixed. In particular, we used only the high resolution MS/MS spectra and more specifically we used the spectra that are obtained through higher-energy collision dissociation (HCD) which was the most common fragmentation method in the NIST dataset and additionally it is known to have high sensitivity and produce more fragments [7, 22]. We did not use MS3 and MS4 spectra as these were provided only for a small percentage of the data. Regarding the precursor ions, we retained only the most common ones, that is [M+H]+ and [M-H]-. For each precursor ion, we used two spectra, one obtained using low collision energy and one with high collision energy. The level for characterizing low collision energy was set to 35% NCE (Normalized Collision Energy) and for high energy it was set to 130% NCE. These values were selected because they were the most common energy levels in the NIST dataset for characterizing low and high energy, respectively. In the cases where a spectrum with energy 35% or 130% NCE was not available, we selected the spectrum that was obtained using collision energy that was closest to that level. Therefore, each instance in the dataset we constructed is characterized by four MS/MS spectra derived from two different precursor ions and two energy levels. It should be highlighted though, that not all molecules in the NIST dataset have experimental data for the specific precursors and energy levels. However, we have allowed cases with missing data in the dataset and the missing spectra are represented as empty spectra, that is spectra with no peaks, in an attempt to develop a model that is robust to missing data. Therefore, the model is being trained and evaluated on cases that may not have available all four spectra.

As part of the data filtering process, we additionally removed molecules with rare atom species, that is species that appeared in less than 30 molecules. Specifically, we excluded molecules with the following atoms: Co, Fe, Se, As, Si, B, Sn, Au, Cu. We also did not make use of the data corresponding to peptides since the goal of this work is to identify structures of small molecules. Finally, we filtered out all molecules for which the retained spectra, for the selected precursor ions and energies, did not have peaks with $m/z > 500$. The reason for this final constraint is explained in the following paragraph (Data representation).

2.2.2 Data representation

We represent each MS/MS spectrum as a vector in which each bit corresponds to a specific mass-over-charge (m/z) value, representing the m/z value of the recorded fragments, while the value of each bit corresponds to the intensity, or otherwise frequency, of the fragments that have been recorded with that specific mass-over-charge value. For that representation, we need to specify the resolution of the mass as well as the minimum and maximum allowed mass values. More specifically, the minimum mass is set to 50 Da while the maximum mass is set to 500 Da. The resolution for the mass values is 0.01 Da. Given that our dataset is of higher resolution, that is more than 4 decimal points are available, the intensity of each bit corresponds to an aggregation of all fragments that have been recorded and have the same mass when considering two decimal points. Finally, the intensity values are normalized by dividing with the maximum intensity over all the vector bits of a given spectrum. The minimum and maximum allowed mass values were selected based on the statistics of the dataset. More specifically, the minimum allowed mass corresponds to the minimum fragment mass that has been recorded over all data. Regarding the maximum allowed mass, although there are molecules in the dataset with larger recorded fragments, the percentage of molecules with fragments larger than 500 Da is very small. In general, a smaller maximum allowed mass, as well as a lower resolution, will result in a more compact and less sparse vector representation which is essential for preventing over-fitting when training the DL model.

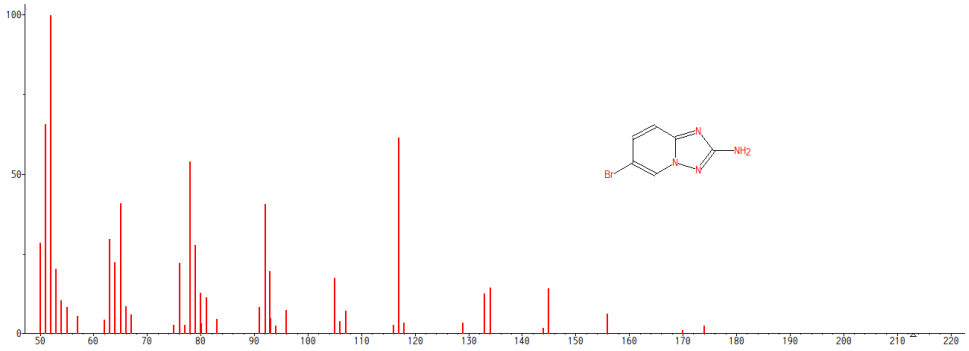
Regarding the molecular structures, we represent them using canonical SMILES without indicating stereochemistry information.

2.2.3 Data augmentation

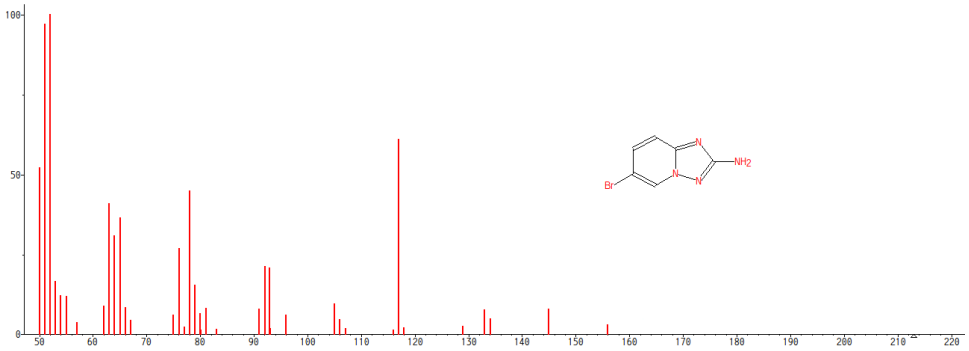
The variability in the spectra for a given molecule opens up the possibility for data augmentation. In particular, although some spectra from the same molecule may differ significantly, as shown in Figure 1, in many cases the obtained spectra are closely related. One such case is when the collision energies that are being used are relatively close. Such an

example is illustrated in Figure 2 where all experimental conditions are the same except the collision energy which however does not differ significantly between the two spectra.

In order to augment the dataset, for each instance in the training set we are creating an additional training instance by slightly perturbing the collision energy in all four spectra. In particular, each spectrum, out of the four spectra that are used to represent an instance in the dataset, is replaced with a spectrum that has the closest collision energy in the dataset with the spectrum to be replaced. In order to avoid large deviations from the preset energy levels (35% for low energy and 130% for high energy) we perturbed only the spectra that had exactly the pre-set energy levels (we recall here that in cases where a spectrum of 35% or 130% NCE was not available, it was already replaced with the closest available in the original dataset).



(a) NCE 115%



(b) NCE 130%

Figure 2: MS/MS spectra obtained using nearby collision energies (Normalized Collision Energy). Spectra sourced from the NIST library.

2.2.4 Data partition

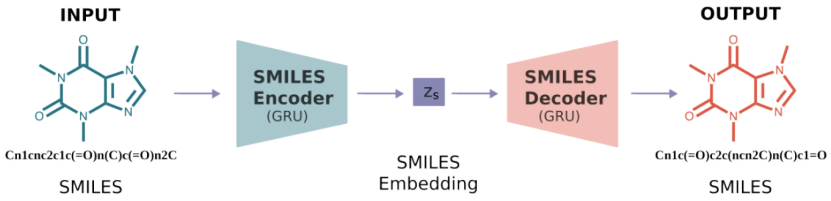
After the data filtering process, the acquired dataset consists of 23K molecules, each one of them is associated with four MS/MS spectra or more precisely up to four MS/MS spectra given that there are cases with missing spectra. This dataset was partitioned into a training, a validation and a test set with the validation and test set having about 1K molecules each. For the test set specifically, we used fingerprint similarity, based on the Tanimoto coefficient [28], in order to ensure that no test molecule is either in the train or in the validation set. The validation set was used to select the model hyper-parameters and the test set was used to evaluate the performance of the model.

3 Spec2Mol architecture

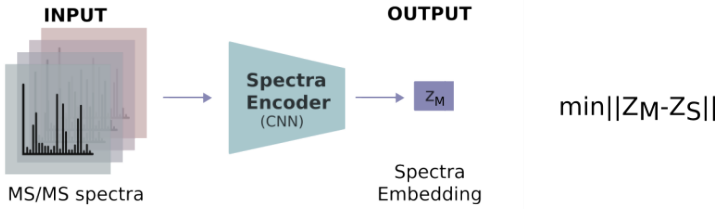
Spec2Mol uses an encoder-decoder architecture for recommending molecular structures from MS/MS spectra. The Spec2Mol encoder generates spectra embeddings while the decoder reconstructs the SMILES sequence from a spectra embedding. The encoder and the decoder have been trained separately as it is shown in figure 3. First, the decoder is trained as part of an autoencoder architecture for reconstructing the SMILES sequence from a SMILES embedding. Next, the spectra encoder is trained such that the learnt spectra embeddings match the corresponding SMILES embeddings. Finally, for making inference on unseen cases, Spec2Mol uses the spectra encoder to obtain the spectra embedding

which is subsequently used in order to decode potentially novel molecules and also to retrieve molecules from the pre-training dataset.

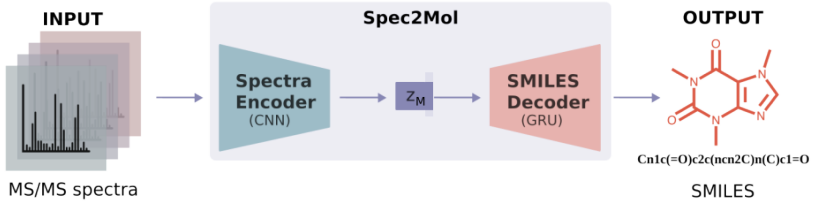
The specifications for training each model are given in the following paragraphs.



(a) The AE is pre-trained to translate from a random SMILES to the canonical SMILES string.



(b) The spectra encoder is trained to learn the same embedding as the SMILES encoder.



(c) During inference, the spectra encoder and the SMILES decoder of the pre-trained model are used to translate spectra into molecular structures.

Figure 3: The Spec2Mol model consists of a spectra encoder and a SMILES decoder which have been trained separately but share the same embedding space.

3.1 Pre-training the AE on chemical structures

The autoencoder is trained on a translation task where a randomized input SMILES is translated into its corresponding canonical SMILES, similar to the work of Winter et al [29]. The encoder and the decoder of the AE are both based on gated recurrent units (GRU) which is a variation of the standard long short term memory (LSTM) models, that are commonly used for learning sequence representations, with fewer parameters.

3.2 Training the spectra encoder

The spectra encoder is trained in a supervised manner such that the learnt spectra embeddings are the same as the SMILES embeddings that the AE has learnt. More specifically, the input of the spectra encoder consists of the four spectra that have been pre-selected to represent each molecule. The spectra encoder is based on 1-D CNNs and in particular consists of two 1-D CNN layers and two fully connected layers. The four spectra are represented as 4 discrete vectors which are fed into the 1-D CNN as data from four different channels. Each channel corresponds to a specific precursor ([M+H]+ or [M-H]-) and energy level (low or high). If any of the required four spectra is not available, then the input to the respective channel is an all zeros vector. The output of the spectra encoder is a 1-D vector which is the latent representation of the spectra in the embedding space. The model is trained such that the distance (root mean square error) between the latent representation that is learnt by the spectra encoder and the latent representation that is obtained from the pre-trained SMILES encoder is minimized.

3.3 Recommending molecular structures for unseen spectra

Spec2Mol provides as output molecular structures that can potentially explain the observed spectra peaks. The recommended molecules for unseen spectra are obtained using two strategies: a direct and an indirect molecule generation strategy. The direct molecule generation strategy generates molecular structures using the SMILES decoder from the computed MS/MS embedding. Multiple SMILES are generated for each MS/MS embedding using a pure sampling strategy [30], and subsequently filtered in order to retain only the valid ones, i.e., the sequences that respect the SMILES syntax. The indirect strategy retrieves molecular structures from the dataset that was used for pre-training the AE based on the distance in the embedding space. More specifically, for each MS/MS embedding we find the closest embeddings from the pool of molecules used to pre-train the AE and decode those embeddings into SMILES sequences.

The predicted molecules obtained through these two strategies are combined and ranked based on their discrepancy from the expected molecular weight. The molecular weight of the underlying chemical structure is easily inferred from the mass spectrum and therefore in this work we consider it as known. The molecular structures that have molecular weight closer to the reference weight are highly ranked. The top 20 ranked predictions are returned to the user.

4 Method evaluation

4.1 Reconstruction accuracy of the AE on the NIST molecules

As a sanity check, we evaluated the ability of the pre-trained AE to reconstruct the SMILES of the molecules in the testing set of the spectra dataset. We recall that the AE has been pre-trained on molecular structures derived from the PubChem database while the molecules in the spectra dataset are from the NIST database. The molecules from the NIST dataset were not used for pre-training the AE, although it is expected that a portion of the NIST molecules is present in the PubChem database which is used for pre-training.

The AE was able to correctly reconstruct the SMILES sequence for about 95% of the NIST molecules. This demonstrates that the pre-trained model has been trained on a diverse set of molecules and therefore it is able to handle the large variability of the molecules in the NIST dataset.

4.2 Spec2Mol performance evaluation

Spec2Mol generates a set of recommended molecular structures given MS/MS spectra. Our evaluation focuses on assessing the similarity between the generated structures and the reference molecular structure from the NIST dataset. We recall here that the information in an MS/MS spectrum may not be sufficient to fully reconstruct the molecular structure. It is possible that more than one molecular structures may explain a given spectrum. For that reason our analysis has been focused on assessing whether the model has learnt to identify key features in the molecular structure from the mass spectra rather than identifying the exact same structure with the reference molecule from the NIST dataset.

For the evaluation of the model, we first perform a coarse-level comparison taking into account physicochemical properties and more specifically the molecular weight and the element composition of the molecule. Next, we assess molecular similarity at the substructure level. In particular, we compute the fingerprint similarity as well as the maximum common substructure between the generated structures and the reference structure. The specifications for each metric are given below, while the results are aggregated in Tables 1 and 3. We evaluate the overall performance in the entire test set as well as the performance of the model when not all four required spectra are available as input. Additionally, we assess the contribution of each of the two strategies for generating the recommended structures.

- **Physicochemical attributes**: A property of special interest is the molecular weight since it is directly reflected in the mass spectrum. In particular, the spectra indicates the mass of the fragments and therefore the mass of the original, non-fragmented, molecule can be approximated more easily given the mass spectra as opposed to determining the composition or the structure of the molecule. We record the difference between the molecular weight of the generated structures and the reference structure and we report the relative average minimum difference, that is, the average-minimum difference over all the predicted structures divided by the average molecular weight of the reference structures (DMW_{min}). We also report the average-average difference over all the predicted structures divided by the average molecular weight of the reference structures (DMW_{avg}). Additionally, we also evaluate whether the model is able to identify the element composition of the molecule. In particular, we assess whether the atom species that are present in the reference molecule have been identified in the predicted structures ignoring the numbers of atoms for each atom species. More specifically, for each

atom species we report sensitivity and specificity for detecting the presence of this species. In order to account for discrepancies in the number of atoms per atom species we also report the difference between the molecular formulas of the predicted structures and the reference structure (DMF). We define the distance between two molecular formulas as the number of atoms that differ between the formulas when accounting for the atom species (not including hydrogen atoms). We report the minimum distance over all predictions divided by the average number of heavy atoms (DMF_{min}) as well as the average distance over all predictions divided by the average number of heavy atoms (DMF_{avg}).

- **Fingerprint similarity:** Fingerprints are vector representations of chemical molecules, which indicate the presence of certain substructures in the molecule, and are widely used as an efficient way to judge similarity between molecules [28]. We extracted fingerprint representations based on the Morgan algorithm [31] using the RDKit toolkit [32] and used the cosine coefficient to assess similarity ($Fngp_{cosine}$). The Morgan fingerprints are computed for radius 2 and 32 bits.
- **Maximum common substructure (MCS)**: We computed the MCS between two molecular structures using the RDKit toolkit [32] with the following constraints: the substructure match respects the atom species, the bond orders, as well as the ring bonds, that is ring bonds are only matched to ring bonds. From the computed MCS we extracted the following three metrics: i) MCS ratio, ii) MCS Tanimoto, and iii) overlap coefficient, which are defined as follows, respectively: $MCS_{ratio} = \frac{a_{MCS}}{a_r}$, $MCS_{tan} = \frac{a_{MCS}}{a_r + a_p - a_{MCS}}$, $MCS_{ovrlp} = \frac{a_{MCS}}{min(a_r, a_p)}$, where a_{MCS} denotes the number of atoms in the MCS, a_r the number of atoms in the reference compound, and a_p the number of atoms in the predicted compound. For each metric we report the maximum value as well as the average value over all predictions.

Table 1 summarizes the evaluation of the effect of missing data in the predictions. More specifically, we present the evaluation metrics on four different partitions of the test-set depending on the number of the available spectra. We recall that the input to the model consists of four different spectra obtained through different specifications. However, not all molecules in the dataset have all four spectra available. Our results indicate that missing only one spectrum does not severely impact performance, but performance starts to degrade when less than three spectra are available.

Next, we evaluate the effect of the strategy that is used to generate the recommended molecules. The analysis is shown in Table 2. We recall that the recommended structures are obtained either directly through decoding the computed embeddings or indirectly by identifying the closest embeddings from the pre-trained dataset. In particular, we are comparing the top-20 predictions, as ranked using the molecular weight criterion, through i) only the direct strategy, ii) only the indirect strategy, and, iii) the two strategies combined. According to the results, the indirect approach, that generates molecules through decoding the closest embeddings from the pre-trained dataset appears to have a larger contribution on the effectiveness of the method to generate relevant structures. However, combining the two strategies appears to slightly improve performance.

Overall the results illustrate that the predicted structures have a molecular weight that is significantly close to the molecular weight of the reference compound. This is not surprising as the generated molecules are ranked based on the molecular weight. The molecular formula though seems to also be considerably close to the reference one. The

Table 1: Effect of missing spectra in the model input. Evaluation metrics when considering the entire test set and the test-data partitions that have available all 4, only 3, only 2 and only 1 spectrum. The arrows show the desired trend for each metric.

metric		full dataset	4 spectra	3 spectra	2 spectra	1 spectrum
# test cases		1000	413	65	483	39
correct molecules (↑)	(%)	7.0	9.7	15.4	4.1	5.1
correct formulas (↑)	(%)	26.0	31.2	29.2	22.8	10.3
$DMW_{\%}$ (↓)	min	1.5	0.9	0.5	1.9	3.0
	avg	5.4	4.6	3.9	6.1	7.4
$DMF_{\%}$ (↓)	min	9.2	6.5	8.1	10.8	21.1
	avg	21.7	17.8	24.5	24.0	32.9
$Fngp_{cosine}$ (↑)	max	0.85	0.87	0.85	0.84	0.83
	avg	0.74	0.76	0.72	0.72	0.72
MCS_{ratio} (↑)	max	0.68	0.70	0.72	0.66	0.57
	avg	0.51	0.53	0.55	0.50	0.43
MCS_{tan} (↑)	max	0.55	0.58	0.60	0.53	0.44
	avg	0.38	0.39	0.41	0.36	0.30
MCS_{coef} (↑)	max	0.71	0.73	0.74	0.69	0.63
	avg	0.54	0.55	0.58	0.53	0.48

Table 2: Effect of the molecule generation strategy. Comparative evaluation of the top-20 predictions using the direct strategy, the indirect strategy and the two strategies combined. The arrows show the desired trend for each metric.

metric		direct	indirect	combined
correct molecules (↑)	(%)	0.8	6.9	**7.0**
correct formulas (↑)	(%)	13.5	20.3	**26.0**
$DMW_{\%}$ (↓)	min	1.9	2.9	**1.5**
	avg	10.3	7.7	**5.4**
$DMF_{\%}$ (↓)	min	10.4	11.9	**9.2**
	avg	24.2	22.4	**21.7**
$Fngp_{cosine}$ (↑)	max	0.84	0.84	**0.85**
	avg	0.73	0.73	**0.74**
MCS_{ratio} (↑)	max	0.65	0.66	**0.68**
	avg	0.50	**0.51**	**0.51**
MCS_{tan} (↑)	max	0.50	**0.55**	**0.55**
	avg	0.34	**0.38**	**0.38**
MCS_{coef} (↑)	max	0.68	**0.71**	**0.71**
	avg	0.53	**0.56**	0.54

Table 3: Sensitivity and specificity for detecting the presence of each atom species in the entire test set, having as reference the frequency of each species in the training spectra dataset.

	O	N	S	Cl	F	Br	P	I
Sensitivity	0.94	0.86	0.50	0.68	0.48	0.79	0.53	0.51
Specificity	0.50	0.76	0.96	0.91	0.92	0.98	0.99	0.99
Frequency (%)	85.4	71.5	18.4	15.2	11.5	7.5	2.5	1.4

model was able to retrieve the exact structure for a small percentage of the test cases (7%) while it identified the exact molecular formula for a considerably larger percentage (26%). The performance of the model was significantly better when at least 3 out of the 4 input spectra where available.

Regarding the structural similarity between the predicted structures and the reference structure, the obtained values for the respective metrics demonstrate that the structures share common substructures. More specifically, the metrics that are based on the MCS between the reference and the predicted structures indicate that the common substructure is, on average, nearly 70% of the size of the reference structure for the closest structure and more than 50% for the average prediction. This result is in agreement with the high correlation between the molecular fingerprints.

Regarding the ability of the model to identify the presence of each atom species in the molecular structure, it varies significantly and it correlates with the frequency of each atom species in the training dataset, as it is shown in Table 3. More specifically, the model has very high sensitivity for nitrogen (N) and oxygen (O) which are the most common atom species in the dataset (excluding carbon which is not included in this analysis as it is present in all molecules). However, the specificity for oxygen is significantly lower than that of nitrogen which means that there is a significant number of false positives for oxygen compared to nitrogen. Regarding the more rare atom species, the opposite phenomenon is observed: specificity is significantly high while sensitivity is low. This means that for the rare species there is a very small number of false positives which is expected as these atoms are under-represented in the training set. However, sensitivity is at least 0.5 for all atoms, which shows that the model is able to capture the presence of rare atoms quite well considering that some atom species are severely under-represented in the training set.

Figure 4 shows a few examples of successful cases with the model correctly identifying key substructures such as rings and long chains, and the presence of rare atoms and functional groups. Given the vast space of possible molecular structures, these cases demonstrate that the model has indeed learnt to associate spectra features with molecular structures.

We also identify two general scenarios where the model has a difficulty in predicting relevant structures: (1) Molecules with large rings and (2) Molecules that have poor quality spectra. An example of the first case is illustrated in Figure 5. We believe this is because molecules with large rings are significantly under-represented in the dataset that was used to pre-train the decoder. Also, it is hard to generate a valid SMILES sequence for molecules with very large rings. Regarding the second cases of poor quality input spectra, it includes cases where there is a very small number of peaks in the spectra and therefore not adequate information to reconstruct the SMILES sequence.

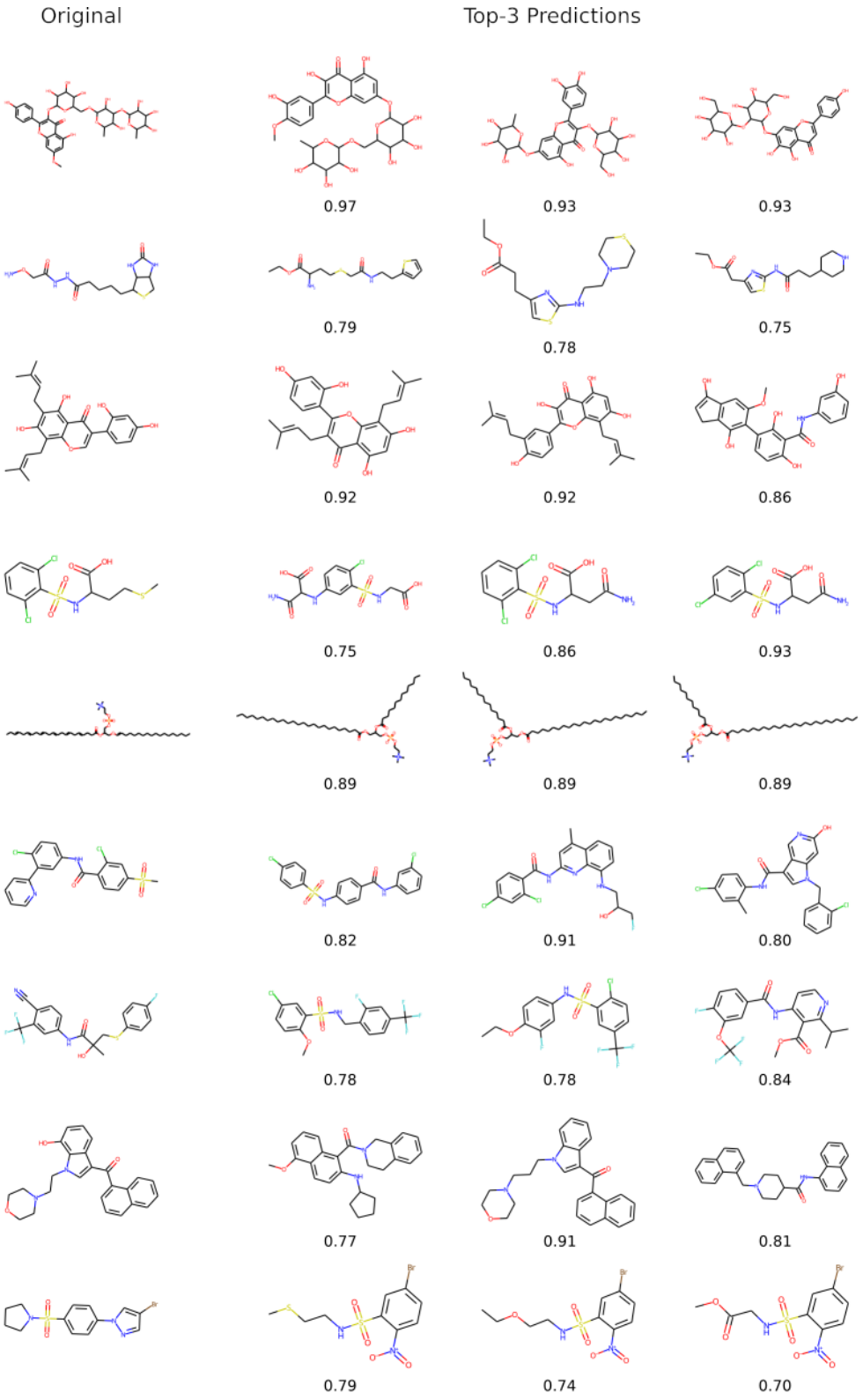


Figure 4: Examples of the most likely predicted structures from Spec2Mol along with the cosine similarity values with respect to the original reference structures.

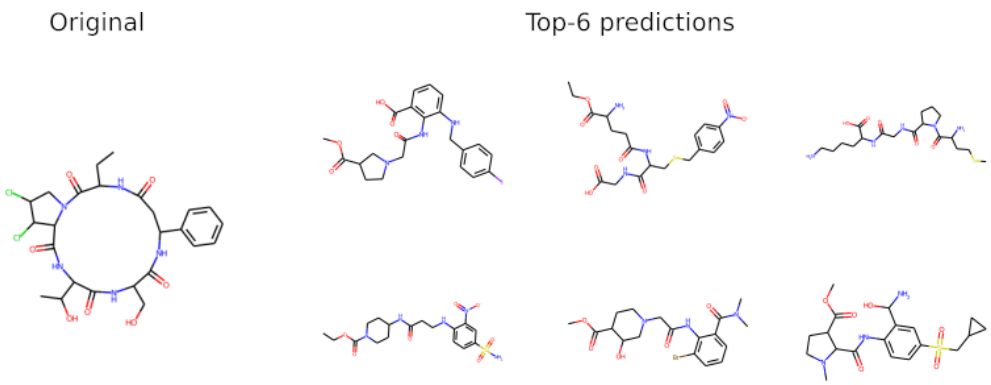


Figure 5: An example where Spec2Mol failed to identify a similar structure for a reference compound containing a large ring.

4.3 Comparative evaluation

In order to perform a comparative evaluation, we have used SIRIUS 4 [33], which offers multiple functions including chemical formula, as well as molecular structure, identification from mass spectra . SIRIUS' structure elucidation method, called CSI:FingerID, is a database retrieval method [16]. It relies on Support Vector Machines (SVMs) for predicting a molecular fingerprint and subsequently compares the predicted fingerprint against those of a reference database in order to identify candidate structures. The input to the SVM is the MS/MS spectrum along with the corresponding computed fragmentation tree. CSI:FingerID has shown superior performance when compared to other existing tools for automatic identification of molecular structures from spectra data. In particular, it was the best performing method in the Critical Assessment of Small Molecule Identification (CASMI) contest for 2016 and 2017 [33]. However, the performance of this method degrades significantly for cases that are not covered in the training set [33]. Additionally, the dependence of CSI:FingerID on fragmentation tree data adds significantly to the running time of this method.

We run SIRIUS on the same test set we developed for evaluating Spec2Mol. As input, we provided SIRIUS with the positive mode spectra (that is [M+H]+ at low and high energy) as they were selected for Spec2Mol. The spectra from negative ions were not used since a single run for SIRIUS accepts spectra from a single precursor which may be obtained through different energies. As 53 test cases out of the 1000 cases of the test set did not have any positive mode spectra and therefore the test set used for the comparison consists of 947 cases. As a side note, SIRIUS performs structure elucidation after identifying the molecular formula. The number of molecular formulas to be explored is one of the parameters of the tool which we set to 10. An additional parameter is the reference database which we set to PubChem, which is the largest available source offered by SIRIUS. Finally, SIRIUS allows the user to define the set of chemical elements to be considered when performing the search which we set to: C, H, O, N, S, Cl, F, Br, P and I. It should be noted that expanding the pre-defined set of atoms (C, H, N, O, P, S) to account for more rare atoms which were present in the NIST dataset significantly increased the running time.

On the test set of 947 cases, SIRIUS found the correct formula for about 98% of the test cases while it found the correct structure for about 67%. For 6 cases out of 947 SIRIUS did not return any structures. At this point, it should be highlighted that the CSI:FingerID method from SIRIUS for structure identification has been trained on the NIST dataset (NIST v17). As it is discussed in the original study for the SIRIUS tool, the presence of spectra for a given test structure in the training set can significantly boost performance even if these spectra which are used when testing are not the exact same spectra used in training [33].

The comparative evaluation between SIRIUS and Spec2Mol was performed on the cases where SIRIUS failed to find the exact molecular structure. Since Spec2Mol is intended for recommending potential molecular structures given mass spectra, our intention here is to evaluate how relevant are the recommendations, when compared to widely accepted and state-of-the-art method like SIRIUS. By focusing our comparison on the cases where SIRIUS did not find an exact match, we are essentially evaluating the relevance of the recommended structures when an exact match is not found, which points to the case of novel molecules. In particular, we compared SIRIUS and Spec2Mol on the 307 cases, for which SIRIUS failed to find an exact match, using the metrics based on fingerprint similarity and MCS. It should be noted here that failure to identify the exact structure includes cases where SIRIUS either did not return any structure and cases where the reference structure was not among the predicted structures. The results are summarized in Table 4. According to our analysis, the structures recommended by Spec2Mol are at least as relevant as the ones

recommended by SIRIUS. More specifically, Spec2Mol achieved better cosine similarity for the closest structure, while almost all metrics based on the MCS are improved in the case of Spec2Mol. This outcome is especially interesting and encouraging, given that Spec2Mol is an end-to-end approach that does not take into account any prior knowledge. Spec2Mol generates potential molecular structures by solely looking at raw MS/MS spectra. On the other hand, the combination of CSI:FingerID and SIRIUS attempts to retrieve the exact molecular structure from a reference database taking as input the computed fragmentation tree on top of the raw mass spectra. Although a direct comparison of the two methods is not possible, still the outcome of our comparative evaluation demonstrates that the molecular structures generated by Spec2Mol are at least as successful as the ones obtained by state-of-the-art tools when considering novel molecules despite the fact that Spec2Mol relies solely on raw MS/MS spectra.

Table 4: Comparative evaluation of structural similarity between the recommended structures and the reference structure between SIRIUS and Spec2Mol.

Method		$Fngp_{cosine}$	MCS_{ratio}	MCS_{tan}	MCS_{coef}
SIRIUS	max	0.82	0.65	0.54	0.66
	avg	0.72	0.49	0.35	0.49
Spec2Mol	max	**0.84**	**0.66**	0.53	**0.69**
	avg	0.72	**0.50**	**0.36**	**0.53**

5 Conclusions

Elucidating the structure of chemical compounds is a fundamental, but cumbersome, task, in metabolomics studies, as well as in chemical analysis in various domains including drug development and forensics analysis. The available computational tools for aiding structure elucidation are based on fragment annotation and database retrieval methods. This approach fails to identify molecules that are not present in the reference database which, in practice, may correspond to a considerably large percentage of the query spectra. We have developed Spec2Mol, an end-to-end deep learning architecture for directly generating molecular structures (SMILES sequences) from the input MS/MS spectra. Spec2Mol is based on an encoder-decoder architecture that generates molecular SMILES sequences, given mass spectra. While the proposed architecture supports the retrieval of molecules from a database that best matches the input spectra, it can also generate new molecules that have not been seen before in any dataset. Our analysis demonstrates that the recommended molecules are structurally and physiochemically similar to the reference compounds, suggesting that the latent embeddings has indeed learnt informative associations between the spectra and the structural features. When compared to an existing method that depends on the fragmentation tree annotation on top of the raw spectra for molecule identification, Spec2Mol performed on par for the task of recommending potential molecular structures. Our results indicate that the proposed approach of recommending de-novo molecules directly from input MS spectra provides critical insights on the characteristics of the underlying molecular structure, and, can complement existing tools especially when the current tools fail to identify the right molecule from existing databases. We speculate that incorporating prior knowledge in the model, for example in the form of fragmentation trees, can further boost the performance of the proposed method.

References

[1] S. Nalbantoğlu. Metabolomics: Basic principles and strategies. In S. Nalbantoğlu and H. Amri, editors, *Molecular Medicine*. 2019.

[2] S. Lee, D.G. Oh, D. Singh, J.S. Lee, S. Lee, and C.H. Lee. Exploring the metabolomic diversity of plant species across spatial (leaf and stem) components and phylogenic groups. *BMC Plant Biology*, (1), 2020. PMID: 31992195; PMCID: PMC6986006.

[3] A.H. Emwas. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods in Molecular Biology*, pages 161–193, 2015.

[4] D.S. Wishart. Computational strategies for metabolite identification in metabolomics. *Bioanalysis*, (9):1579–1596–193, 2009.

[5] Diogo Ribeiro Demartini. A short overview of the components in mass spectrometry instrumentation for proteomics analyses. In Ana Varela Coelho and Catarina de Matos Ferraz Franco, editors, *Tandem Mass Spectrometry - Molecular Characterization*. IntechOpen, 2013.

[6] Nguyen DH, Nguyen CH, and Mamitsuka H. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Briefings in Bioinformatics*, 20(6):2028–2043, 2019.

[7] Maria Vinaixa, Emma L. Schymanski, Steffen Neumann, Miriam Navarro, Reza M. Salek, and Oscar Yanes. Mass spectral databases for lc/ms- and gc/ms-based metabolomics: State of the field and future prospects. *TrAC Trends in Analytical Chemistry*, 78:23–35, 2016.

[8] D.S. Wishart, Y.D. Feunang, A. Marcu, A.C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Liu Y., R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, and A. Scalbert. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, 46(D1):D608–D617, 2018.

[9] Yannick Djoumbou-Feunang, Allison Pon, Naama Karu, Jiamin Zheng, Carin Li, David Arndt, Maheswor Gautam, Felicity Allen, and David S. Wishart. Cfm-id 3.0: Significantly improved esi-ms/ms prediction and compound identification. *Metabolites*, 9(4), 2019.

[10] Jennifer N. Wei, David Belanger, Ryan P. Adams, and D. Sculley. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS Central Science*, 5(4):700–708, 2019.

[11] M. Heinonen, H. Shen, N. Zamboni, and J. Rousu. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*, 28:2333–2341, 2012. PMID: 22815355.

[12] Hongchao Ji, Hanzi Deng, Hongmei Lu, and Zhimin Zhang. Predicting a molecular fingerprint from an electron ionization mass spectrum with deep neural networks. *Analytical Chemistry*, 92(13):8649–8653, 2020. PMID: 32584545.

[13] Youzhong Liu, Aida Mrzic, Pieter Meysman, Thomas De Vijlder, Edwin P. Romijn, Dirk Valkenborg, Wout Bittremieux, and Kris Laukens. Messar: Automated recommendation of metabolite substructures from tandem mass spectra. *PLOS ONE*, 15(1):1–17, 01 2020.

[14] Arpana Vaniya and Oliver Fiehn. Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *TrAC Trends in Analytical Chemistry*, 69:52–61, 2015.

[15] Florian Rasche, Kerstin Scheubert, Franziska Hufsky, Thomas Zichner, Marco Kai, Aleš Svatoš, and Sebastian Böcker. Identifying the unknowns by aligning fragmentation trees. *Analytical Chemistry*, 84(7):3417–3426, 2012. PMID: 22390817.

[16] Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using csi:fingerid. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015.

[17] Huibin Shen, Kai Dührkop, Sebastian Böcker, and Juho Rousu. Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics*, 30(12):i157–i164, 06 2014.

[18] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. van der Hooft. Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):1–18, 02 2021.

[19] Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin A. Hoffmann, Daniel Petras, William H. Gerwick, Juho Rousu, Pieter C. Dorrestein, and Sebastian Böcker. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology*, 2020.

[20] Khawla Seddiki, Philippe Saudemont, Frédéric Precioso, Nina Ogrinc, Maxence Wisztorski, Michel Salzet, Isabelle Fournier, and Arnaud Droit. Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification. *Nature Communications*, 11:5595, 2020.

[21] Yang-Ming Lin, Ching-Tai Chen, and Jia-Ming Chang. Ms2cnn: predicting ms/ms spectrum based on protein sequence using deep convolutional neural networks. 20, 2019.

[22] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, 114(31):8247–8252, 2017.

[23] Fatema Tuz Zohora, M. Ziaur Rahman, Ngoc Hieu Tran, Lei Xin, Baozhen Shan, and Ming Li. DeepIso: A deep learning model for peptide feature detection from lc-ms map. *Scientific Reports*, 9, 2019.

[24] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[25] X. Yang, P. Neta, and S. Stein. Extending a tandem mass spectral library to include ms2 spectra of fragment ions produced in-source and msn spectra. *Journal of the American Society for Mass Spectrometry*, 28:2280–2287, 2017.

[26] NIST 20 dataset. `https://chemdata.nist.gov/dokuwiki/lib/exe/fetch.php?media=chemdata:asms2020:xiaoyu_yang_asms2020_presentation.pdf`. Accessed: 2021-04-04.

[27] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, 11 2020.

[28] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath. Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8):3186–3204, 2014. PMID: 24151987.

[29] Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, 10(6):1692–1701, 2019.

[30] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[31] H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965.

[32] Rdkit: Open-source cheminformatics software. `https://www.rdkit.org/`.

[33] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A. Aksenov, Alexey V. Melnik, Marvin Meusel, Pieter C. Dorrestein, Juho Rousu, and Sebastian Böcker. Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16:299–302, 2019.