Refining EI-MS library search results through atomic-level insights

Umit V. Ucak[1], Islambek Ashyrmamatov[2], and Juyong Lee[1,2]

[1]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Republic of Korea
[2]Research Institute of Pharmaceutical Sciences, College of Pharmacy, Seoul National University, Republic of Korea
nicole23@.ac.kr

Abstract

Mass spectral reference libraries are fundamental tools for compound identification in electron-ionization mass spectrometry (EI-MS). However, the inherent complexity of mass spectra and the lack of direct correlation between spectral and structural similarities present significant challenges in structure elucidation and accurate peak annotation. To address these challenges, we have introduced an approach combining CFM-EI, a fragmentation likelihood modeling tool in EI-MS data, with a multi-step complexity reduction strategy for mass-to-fragment mapping. Our methodology involves employing modified atomic environments to represent fragment ions of super small organic molecules and training a transformer model to predict the structural content of compounds based on mass and intensity data. This holistic solution not only aids in interpreting EI-MS data by providing insights into atom types but also refines cosine similarity rankings by suggesting inclusion or exclusion of specific atom types. Tests conducted on EI-MS data from the NIST database demonstrated that our approach complements conventional methods by improving spectra matching through an in-depth atomic-level analysis.

1 Introduction

Chemical analysis fundamentally hinges on structural elucidation, a process in which mass spectrometry (MS) plays a critical role. Despite a plethora of literature, identifying small molecules from their mass spectra remains an unsolved problem due to its complicated nature [1]. Electron Ionization Mass Spectrometry (EI-MS) is particularly notable for small-molecule investigation, often used in tandem with Gas Chromatography-Mass Spectrometry (GC-MS) setups. This technique ionizes molecular samples at a standardized ionization energy, typically set at 70eV, resulting in a mass spectrum that represents a frequency distribution of ions based on their mass-to-charge (m/z) ratio. While the key feature of EI-MS is its ability to generate fragment-rich mass spectra, a limitation is the low abundance or absence of the molecular ion, which is essential for accurate calculation of elemental compositions [2].

With the advancements in artificial intelligence (AI) and machine learning (ML), there is a growing interest in leveraging AI techniques to predict molecular structures directly from spectra [3–6] and vice versa [7–9]. There are also methods predicting molecular fingerprints from mass spectra followed by database search using fingerprint similarity [10–13]. Spectra-to-structure prediction faces challenges due to the limited size of available training sets, as only thousands of small molecule MS data are publicly available [14]. These end-to-end approaches currently have low accuracy and are difficult for practitioners to incorporate into their existing workflows. A widely adopted strategy for molecular identification involves juxtaposing a sample's mass spectrum against experimental or in-silico libraries. Spectra searching aims to match compounds with library entries for exact identification or to provide structural clues for similar compounds. Consistent and fragment-rich spectra obtained by EI-MS enhance the effectiveness of spectra searching, making it the most widely used method for molecular identification.

Although spectra searching is the preferred approach for EI-MS interpretation, it suffers from a coverage problem: If the query spectrum is outside of the library domain, the correct identification can be overly challenging [15]. This is an issue in practice, since existing mass spectral reference libraries, such

as the NIST/NIH/EPA MS database [16], Wiley Registry of Mass Spectral Data [17], and MassBank [18], only contain hundreds of thousands of reference spectra. One strategy to alleviate the coverage problem is augmenting existing libraries with model-generated synthetic spectra; however, the high computational cost of current prediction methods has limited their use in practical applications [19, 20] Recording additional spectra for more molecules can help mitigate this issue; for instance, NIST updates its library every three years, adding approximately 20K new spectra each time. The inclusion of new molecules in these libraries is often restricted to those of widespread interest, leaving out many newly synthesized compounds.

The accuracy of library matching largely depends on how well the metric reflects the true similarity between the query and reference spectra, assuming reasonable measurement noise (0.4-0.005 Da for low to high resolution measurements) in obtaining the query spectrum [21]. Various search algorithms and similarity metrics have been developed over the years to improve this process. Initially, algorithms like the dot-product [22], and probability-based matching [23] were introduced. Further, metrics such as normalized and absolute euclidean distance [24, 25], Hertz similarity index [26] representing the weighted average ratio, and neutral-loss matching [27] were also developed. Other significant advancements feature Fourier- wavelet- transform-based and partial correlation-based measures, introduced by Koo et al. [28] and Kim et al. [29], respectively. Among all, the weighted cosine similarity (defined by the standard dot-product), along with its variants such as the simple match factor and identity match factor, is widely adopted in mass spectrometry for molecular identification [30]

High-quality spectrum libraries, effective matching algorithms, and accurate peak assignments (annotations) are prerequisites for achieving reliable results in structural elucidation. As EI-MS libraries expand, it becomes increasingly difficult to accurately match query spectra with the vast number of reference spectra. With the high complexity of mass spectra in mind, there are instances where spectra matching may not yield insightful results, as spectral similarity does not necessarily imply structural similarity. We propose a holistic solution for this problem by using the fragment ions of super small organic molecules generated by CFM-EI fragmentation modelling tool [31] for electron ionization. We first form a multi-step one-to-many (mass-to-fragment) complexity reduction plan then train a transformer model to predict structural content of the compound given mass and intensity information. Model outcomes provide insight into atom-types to interpret EI-MS data more accurately by suggesting corrections to the library search results.

2 Result

Spectral library search stands as the most widely employed technique for structural elucidation. If the database does not contain the sample compound, we look for spectra that are similar to deduce the structural features of the unknown compound. Library searches typically link spectral to structural similarity—a trend supported by Demuth et al. [32]. We evaluated the spectra similarity searches' interpretative power, using structures and mass spectra of "unknown" compounds, and yielding hit lists (top-1 and top-10). The figure 1 presents the outcomes of a library search of 10K mass spectra from the NIST main library, showing disparities between spectral and structural similarities.

Figure 1a highlights a specific instance where a high weighted cosine similarity (0.86) contrasts with a low Tanimoto similarity (0.22), suggesting that high spectral similarity does not necessarily equate to structural similarity. In fact, we observed this as a visible trend in the scatter plot (Figure 1b), where the correlation between weighted cosine similarity and structural similarity (ECFP2 [33]) for top-1 hits is weak with Spearman and Pearson correlation coefficients of 0.58 and 0.54. We also conducted a hypothetical re-ranking of the top-10 hits for 10K query molecules by retrospectively applying structural information. This adjustment revealed an average of 2.8 swaps and a maximum of 6.3 swaps in the cosine rankings, indicating that, had the structures been known and structural similarity been the basis for ranking, it would have been preferable to re-rank the spectral similarity hit lists. Structural considerations thus necessitate careful hit list inspection and potential re-ranking for more accurate structural interpretation.

The deduction of key aspects of a given compound such as molecular formula, structural features like side groups or substructures, and overall molecular structure, heavily relies on the accurate annotation of mass spectra [35, 36]. Incorrect annotations are likely to amplify structural dissimilarities, prompting a reevaluation and correction of the molecular formula annotations. For example, identifying the most abundant fragment, typically the base peak that signifies the most stable fragment, is crucial for ascertaining a compound's molecular formula. However, precise molecular formula determination based solely on mass is arduous, as multiple molecular formulae can be assigned within a 2 millidalton

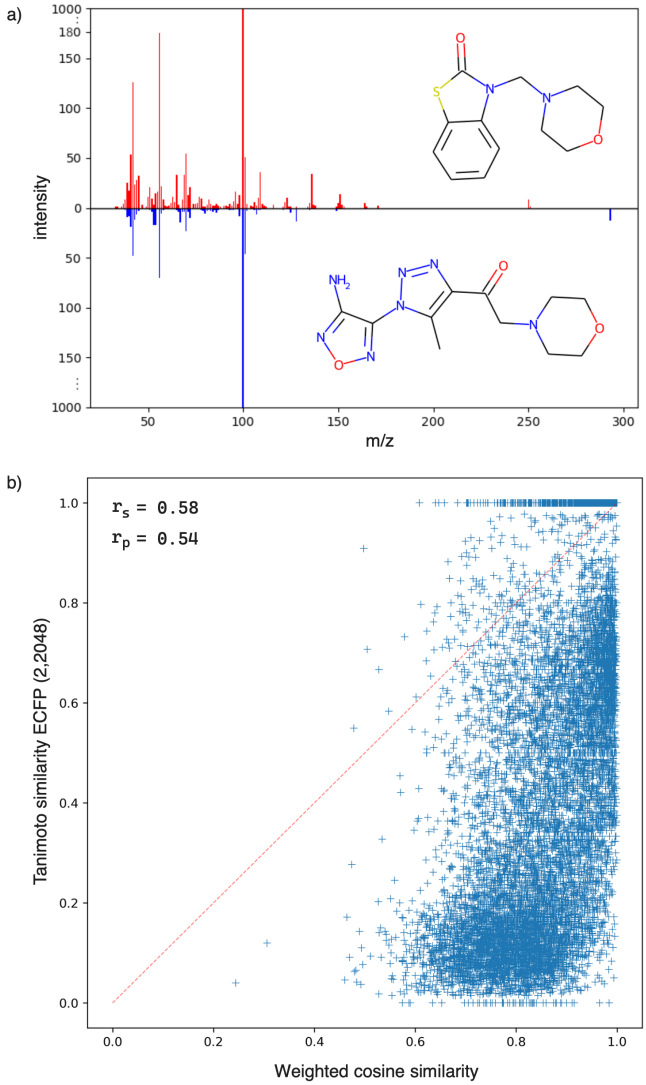
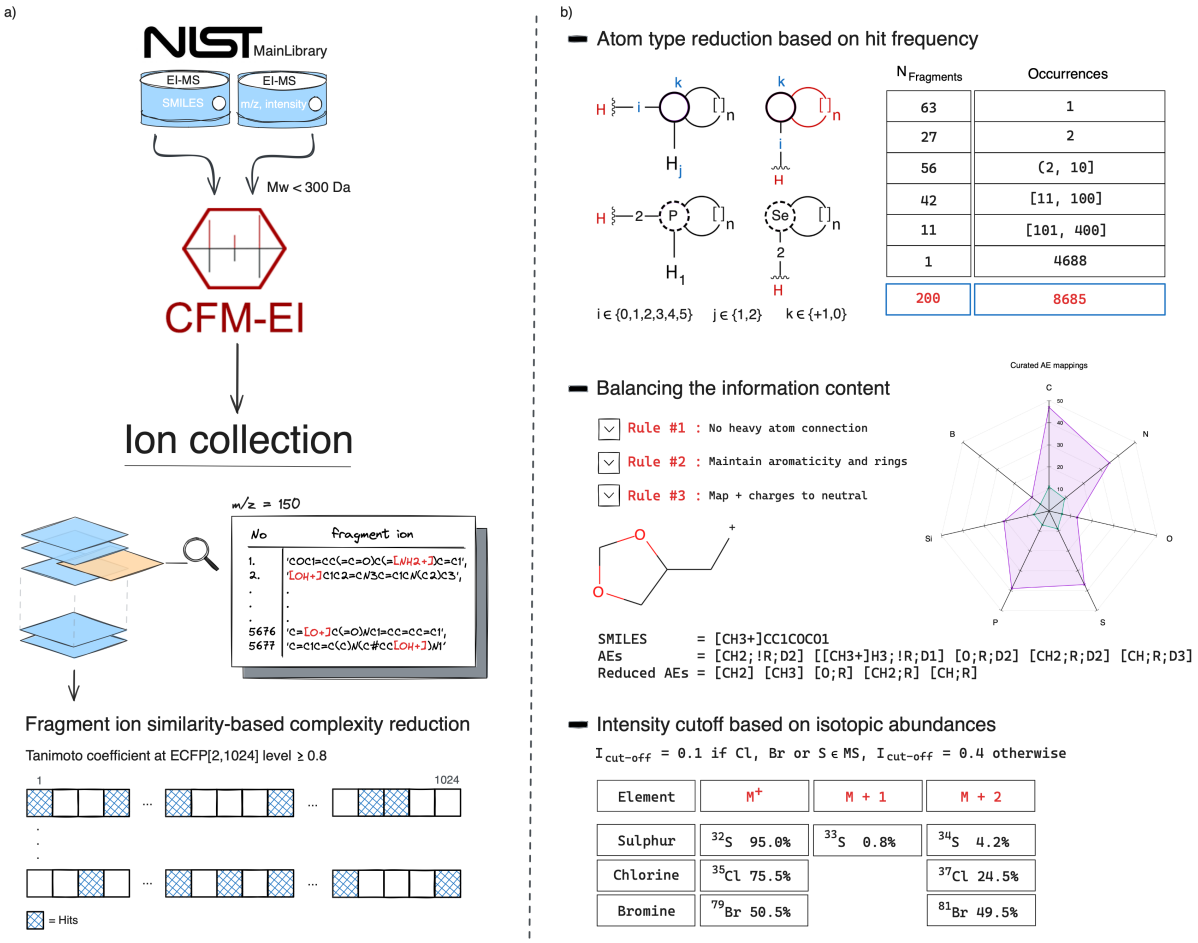


Figure 1: Discrepancy between spectral and structural similarities in EI-MS data analysis. (a) A representative molecule querying against the reference spectra in the NIST main library, the candidate was retrieved at rank 1 exhibited a spectral similarity of 0.89. This is in contrast to the structural similarity score of 0.20 if compared to the structure of the query molecule. (b) Correlation scatter plot for the top-1 hits from a 10K mass spectra dataset. The data emphasizes the need for reevaluation of library search hits when structural similarity is a basis in the elucidation process.

window, which falls within the instrument's inherent measurement error [37]. Therefore, additional data should be integrated into the analysis like fragmentation patterns. This process involves generating all possible fragments for a given mass that could explain the observed data. Here, we adopted the CFM-EI fragmentation to explore all candidate fragment ions.

In Figure 2a, the workflow of fragment ions collection is depicted. In our analysis, we exclusively focused on experimental EI-MS data extracted from the The National Institute of Standards and Technology (NIST) commercial mass spectral library (version 20), consisting of 350,643 spectra for 306,869 compounds. Employing CFM-EI [31, 38], we cataloged an extensive fragmentation of super small ($\leq$300 Da) 93,324 NIST molecules. Developed as an extension of Competitive Fragmentation Modeling (CFM), CFM-EI has been tailored to predict EI-MS spectra, and its superior performance over other established tools like MetFrag [39] and Mass Frontier [40] in compound identification tasks has been demonstrated by the developers. The CFM-EI estimates the likelihood of any given fragmentation event occurring, thereby predicting those peaks that are most likely to be observed.



N Fragments	Occurrences
63	1
27	2
56	(2, 10]
42	[11, 100]
11	[101, 400]
1	4688
200	8685

No	Fragment ion
1.	'COC1=CC(=C=O)C(=[NH2+])C=C1',
2.	'[OH+]C1C2=CN3C=C1CN(C2)C3',
.	.
.	.
.	.
5676	'C=[O+]C(=O)NC1=CC=CC=C1',
5677	'C=C1C=C(C)NC#CC[OH+]N'

Element	M^+	M + 1	M + 2
Sulphur	^{32}S 95.0%	^{33}S 0.8%	^{34}S 4.2%
Chlorine	^{35}Cl 75.5%		^{37}Cl 24.5%
Bromine	^{79}Br 50.5%		^{81}Br 49.5%

Figure 2: Fragmentation and multi-step complexity reduction plan for EI-MS data interpretation (a) Schematic representation of the data processing workflow, beginning with EI-MS data selection from the NIST Main Library, focusing on compounds with Mw $\leq$ 300 Da, followed by fragmentation prediction using CFM-EI, and subsequent ion collection. The bottom panel illustrates the initial reduction applied to the pool of fragment ions via similarity thresholding using the Tanimoto coefficient at ECFP2 level. (b) Frequency-based filtering of atom types (depicted as SMARTS [34]), followed by the process of customizing atomic environment representations to suit analytical needs. The spider chart and adjacent table detail the modifications to AE mappings and the criteria for isotopic abundance-based intensity cutoffs, essential for elements such as S, Cl, and Br.

At its core, it utilizes a probabilistic generative model to simulate the fragmentation process in the mass spectrometer, scrutinizing and breaking down each bond in a molecule in a breadth-first manner to explore all possible fragment states. It assigns a 'break tendency' value to the transition from one fragment state to another, calculated based on the chemical characteristics of the bond undergoing fragmentation. The algorithm then focuses on fragments with a high likelihood, allowing it to generate further derivatives and, ultimately, the predicted spectrum. It is important to note that CFM-EI is not an actual simulation of the fragmentation process but rather an annotated interpretation of the spectrum, where each peak is labeled with corresponding molecular fragments. Considering there are more than a hundred peaks on average in experimental EI mass spectra, in-silico fragmentation represent the same spectra data with 27 in-silico peaks on average. In our case, CFM-EI process yielded a staggering 2,524,662 fragments, of which 858,499 were distinct, indicating a dataset expansion by more than 9 times. For instance, at an m/z value of 150, our dictionary contained 5,677 potential ion fragments expressed as SMILES.

To address the amplified complexity, we first applied a structural pairwise similarity cutoff to reduce redundancy among ion fragments. We utilized Tanimoto metric with Morgan fingerprint, configured with a radius of 1 and a bit vector length of 1024. Formally, letting $\mathcal{I}$ be the collection of ions, where each ion $i \in \mathcal{I}$ is associated with a mass-to-charge ratio (m/z) that belongs to the integer domain, we

define a function $T_c(i, j)$ to compute the Tanimoto coefficient between any two ions $i, j \in \mathcal{I}$. Then, for all pairs (i, j) where $i, j \in \mathcal{I}$ and $i \neq j$, if $\text{sim}(i, j) \geq 0.80$, we removed either i or j from $\mathcal{I}$. For m/z value of 150, from nearly 16 million pair comparisons, we identified 280 pairs above the threshold of 0.8 and eliminated 241 of them. Our observation reveals that highly similar ions were not as common as anticipated, particularly at lower mass-to-charge ratios. We were able to discard only two ions out of 377 fragments for m/z at 70.

Closer inspection of ion structures showed the presence of atom-type level intricacies on the fragment dataset. We filtered specific atom types and corresponding ions, guided by their frequency of occurrence. This elimination was absolutely necessary and, concomitant with the inspection, allowed us to scrutinize for further refinement possibilities. As depicted in generalized SMARTS notations [34] in Figure 2b, our attention was particularly on singletons, doubletons, and those with less common features such as aromatic phosphorus and selenium, heavy metals, and ions predisposed to negative charges, exemplified by `[[O-];!R;D1]`. Consequently, from the initial dataset of ≈ 900,000 ions, we purged 200 atom types that were present in 8,685 ions.

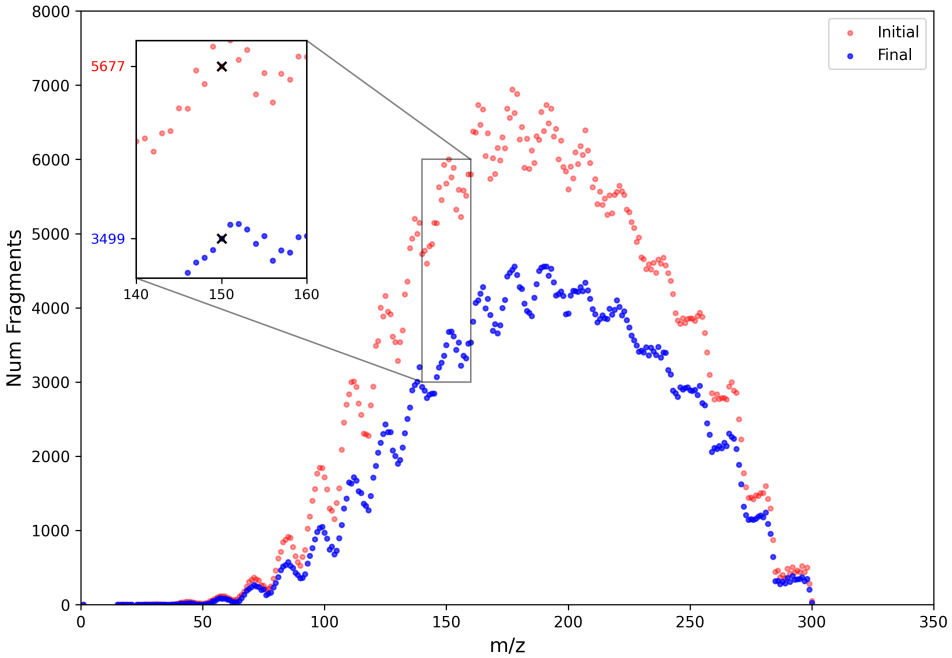


Figure 3: Quantitative analysis on mass-to-fragment mapping. The plot illustrates the number of fragments per m/z value before and after the application of our reduction procedure, showing an average complexity reduction of approximately 34.0 percent across the spectrum. For m/z at value 150, starting with 5,677 fragments, the complexity was reduced up to 38.3 percent.

We have represented fragment ions by their constituent atomic environments (AE), a method whose representational effectiveness and applicability in AI models were demonstrated in our previous studies [41–43]. Our dataset curation adhered to specific rules aimed at maintaining informational balance in AE representation, including avoiding heavy atom connections, preserving ring structures and aromaticity, and neutralizing positive charges. This curation is depicted in the spider graph (see Figure 2b) for the most common elements in the consolidated 217 reduced AE (RAE) mappings derived from an initial broader set. For instance, 19 nitrogen-centric AEs were collapsed into 9 RAEs through this protocol. The chemical elements covered by our fragment dataset are limited to C, N, O, S, P, Si, B and halogens, but enough to cover more than 94% of druglike molecules based on the ChEMBL database [44, 45].

Lastly, our isotopic abundance evaluation, especially for elements like chlorine, bromine, and sulfur, necessitated an adjustment in peak cutoff values regarding the peak density. We implemented an intensity cutoff of 0.1 if they were present in the mass spectra, compared to a standard cutoff of 0.4. While these elements' isotopic patterns aren't always observable, when available, they offer substantial insight. In Figure 2b, we summarize the abovementioned multi-step complexity reduction process. After the whole reduction procedure, we quantified the so-called complexity reduction, average number of fragments per m/z, as 34.0 percent (see Figure 3). In the subfigure of Figure 3, the initial count of 5,677 fragments at

m/z value of 150 was reduced by about 38.3 percent.

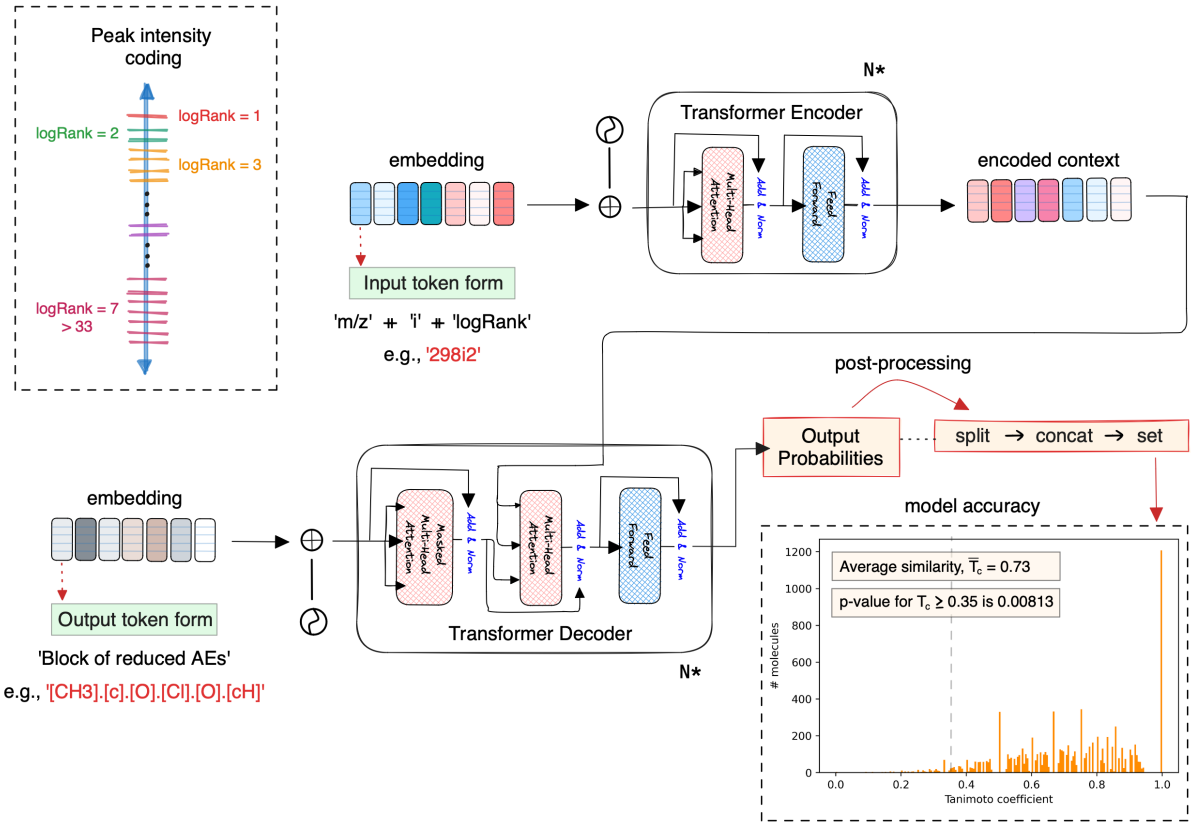


Figure 4: Schematic of the transformer model for converting EI-MS spectral data into structural information. Peak intensities are encoded as logRanks and combined with m/z values as inputs to the transformer encoder. The decoder then predicts structural content as blocks of reduced atomic environments (RAEs). The model accuracy is showcased in the histogram, where the average Tanimoto coefficient (Tc) of 0.73.

Our model operates on the principle of interpreting each peak not as a unique fragment, but as an assembly of RAEs. This allows for a granular breakdown of each fragment into its constituent atom-types. Figure 4 demonstrates our method of translating spectral data into structural information using a transformer model [46, 47]. Peaks are assigned a logRank, a logarithmic intensity measure ranging from 1 to 7. This system, adapted from the work of Cao and Guler [14], is designed to minimize parameter counts and prevent overfitting, offering a more refined approach than traditional intensity rankings. Inputs to the transformer model consist of m/z values, intensity indicators ('i'), and their corresponding logRanks. Target data is represented by blocks of RAEs, which are groupings of atom types that make up the observed fragment ions, as illustrated in the output token form. The PyTorch machine learning library was used for constructing and training the transformer model [48, 49].

The heart of the model is the transformer decoder, which takes the encoded context and predicts output probabilities for each input token, essentially 'decoding' the tokens that carry both mass and intensity information into the AE blocks. These generated blocks undergo post-processing—aggregated, decomposed, and organized into a set—to construct an atom-type fingerprint that predicts the molecular structure. We assessed the model's accuracy by computing the similarity between predicted content and the actual structure. The histogram reveals the predictive accuracy as an average Tc of 0.73, a robust indicator of similarity, considering that Tc of 0.35 was significant with a p-value less than 0.01. The results of this model provided the foundation for refining spectral similarity rankings through in-depth analysis at the atomic level.

Figure 5 illustrates the holistic nature of this model as it collects information from each peak and collectively elucidate the structural content. Since atom types are conserved across fragments (environments may change due to bond breaking), the model assesses the frequency of each RAE within the spectrum. For example, in Figure 5, predicted atom types that are recurrent across multiple fragments for the given

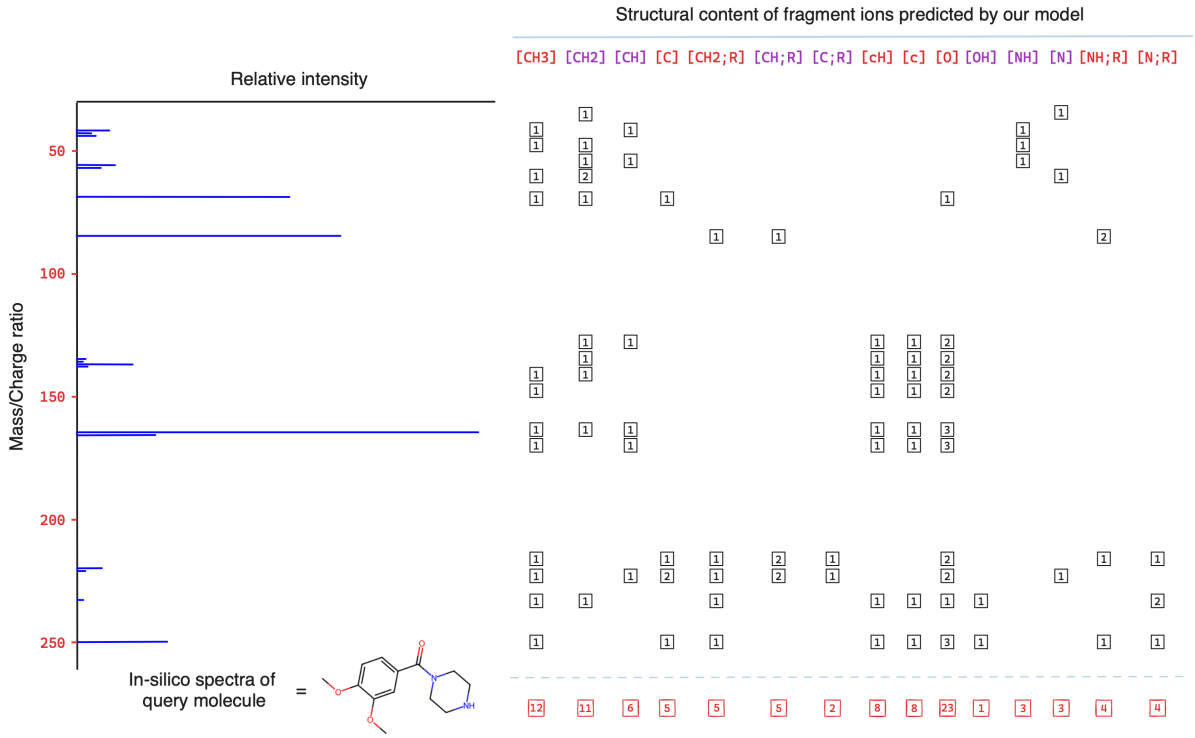


Figure 5: Visualization of how the model synthesizes information from individual peaks to predict structural content. The figure depicts 17-peak in-silico spectrum of a query molecule alongside predicted atom types with their occurrence. Color coding indicates RAEs common to both the query molecule and fragments (Red) and those unique to fragments (Purple).

in-silico spectrum with 17 peaks are shown. An atom type supported by a majority of fragments would be deemed highly probable to be part of the true structure. Conversely, an atom type supported by only a few fragments would raise doubts about its presence, potentially pointing to an incorrect annotation. By exploiting the interconnectedness and richness of fragment ions, the output of the model can be seen as a probabilistic representation of the molecular content, where the likelihood of each RAE's presence is quantified. The results allow us to fine-tune spectral similarity rankings, suggesting targeted additions or removals of atom-types for improved accuracy.

Three possible cases were identified through review of the results, as shown in Figure 6. In the first case, the query molecule `N#CCN=C=S` was matched with a top-1 hit that exhibited a spectral similarity score of 0.831. The identified RAE content of the top-1 hit, comprising `'[CH3]', '[S]', '[CH2]', '[C]', '[CH]'`, did not include nitrogen. This was in contrast to our model's prediction, which included nitrogen, corroborated by multiple peaks and specifically indicated by `'[N]': 4`. In the second case, the library search for the the molecule `CCCCCCCCC(C)NCCCC` yielded a top-1 candidate with a spectral similarity score of 0.869. The RAE content of the top-1 hit encompassed a variety of atom-types, including `'[CH3]', '[C;R]', '[CH2]', '[C]', '[O;R]', '[N]', '[CH2;R]'`, and `'[CH;R]'`, indicating the presence of an oxygen atom within a ring structure `'[O;R]'`. Our model, however, did not predict oxygen and instead assigned scores to other atom types: `'[CH3]': 21, '[CH2]': 24, '[NH2]': 6, '[C]': 2, '[CH]': 12, '[NH]': 1`. The absence of `'[O;R]'` in our prediction suggested that oxygen, especially within a ring, may not be a part of the molecule's true structure.

The third case examined was an organosulfur compound `CN(C)C(=S)NC(=S)N(C)C`. The top-1 hit included silicon and oxygen within the following content `'[CH3]', '[Si]', '[O]', '[N]'`. In contrast, our model featured sulfur at five peaks (`'[S]': 5`) and omitted silicon suggesting that sulfur is likely a part of the actual structure, while silicon is a result of an erroneous annotation. These examples highlighted the potential of our method to refine structural hypotheses by enabling the inclusion of atom-types that are supported by the model and the exclusion of those that are not. The content generated by our model introduces an additional layer to refine spectral similarity results to further increase the identification rates.

There are several limitations of this study. Our model was trained on spectra of super small molecules

(Mw $\leq$ 300 Da). This focus inherently introduces a limitation: relaxing the molecular weight cut-off beyond 300 Da tightens the trade-off, as it amplifies the complexity of mass-to-fragment mapping. Consequently, while the model exhibits robust performance within its specified domain, its applicability to larger molecules or to more complex scenarios like MS/MS tandem mass spectra is currently limited. Furthermore, our fragment dataset, confined to elements C, N, O, S, P, Si, B, and halogens, provides sufficient coverage only for drug-like small organic molecules. Despite these limitations, the model finds strength in the sufficient number of peaks present in idealized in-silico spectra and fragmentation patterns. Peaks below 300 Da provides adequate information to refine library search rankings once assessed collectively because there are several dozens of peaks contributing to the overall structure in EI-MS. In inference mode, especially with larger spectra, the model relies on the lower end of the spectrum for structural content prediction.

3 Conclusion

Mass spectral reference libraries provide a means of identification for compounds. In EI-MS data, where the molecular ion peak is typically missing, conducting a thorough library search becomes even more important. In addition to the absence of a direct correlation between spectral similarity and structural similarity, the inherent complexity of mass spectra introduces a significant challenge for structure elucidation, specifically for accurate peak annotation. In this work, we have introduced a follow-up analysis to library search, employing atomic environments, CFM-EI fragmentation tool, neural machine translation, and structural similarity concept. This approach aims to refine cosine similarity rankings of unknown EI-MS data, offering a holistic solution to the challenges of structure elucidation. To achieve this, we utilized reduced atomic environments (unconnected, stand-alone substructures) to represent fragment ions. Additionally, we used CFM-EI, an approach that models how a molecule might fragment within the collision cell of the mass spectrometer.

In many instances of EI-MS data, peaks are not uniquely identifiable, leading to ambiguity in structural elucidation. We developed a multi-step plan reducing the one-to-many (mass-to-fragment) mapping complexity. Subsequently, we trained a transformer model to predict the structural content of unknown peaks using their mass and intensity information. Each peak, through its assigned AE blocks, contributes to a unified understanding of the compound's structure at the atomic level. The interconnectedness of fragment ions is crucial because it implies that the peaks corroborate each other's content. The outcomes of the model undergo post-processing to yield scores for each atom-type, thereby facilitating targeted corrections to the results of the library search. Tests on EI-MS data from the NIST database have demonstrated that predicting the structural content in conjunction with spectral hits helps to reduce this uncertainty, narrowing down the range of potential candidates for consideration.

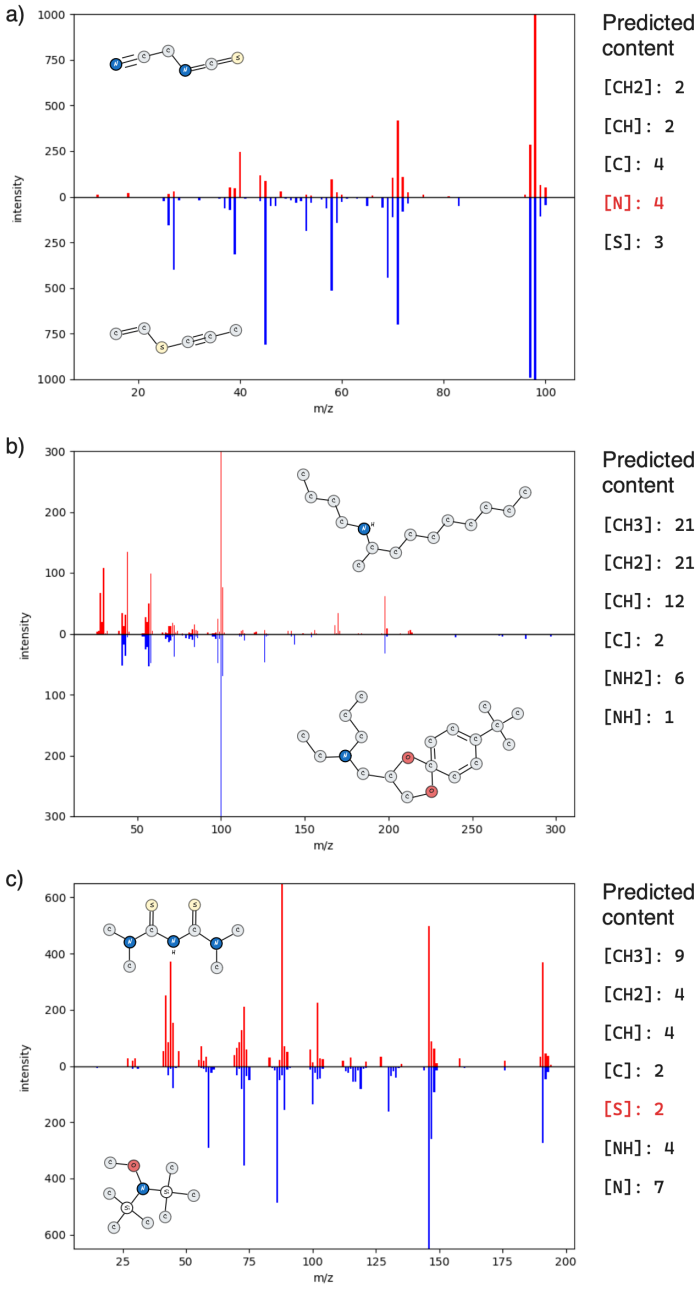


Figure 6: Case studies illustrating the refinement of hit lists using our model's predictions. (a) The top-1 library hit for a query molecule lacks nitrogen, whereas our model predicted its presence within four peaks. (b) In the case of a larger organic molecule, the model predicts a rich content of carbon and nitrogen atoms, unlike the top-1 hit which includes oxygen—a discrepancy in the analysis. (c) For an organosulfur compound, the spectrum reveals sulfur atoms as predicted by our model, whereas silicon, present in the top-1 hit, is absent, indicating a potential annotation error.

References

[1] Kind, T. & Fiehn, O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews* **2**, 23–60 (2010).

[2] Scheubert, K., Hufsky, F. & Böcker, S. Computational mass spectrometry for small molecules. *Journal of Cheminformatics* **5**, 12 (2013).

[3] Elser, D., Huber, F. & Gaquerel, E. Mass2SMILES: deep learning based fast prediction of structures and functional groups directly from high-resolution MS/MS spectra (2023). Preprint at `https://www.biorxiv.org/10.1101/2023.07.06.547963v1`.

[4] Litsa, E. E., Chenthamarakshan, V., Das, P. & Kavraki, L. E. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Communications Chemistry* **6**, 132 (2023).

[5] Stravs, M. A., Dührkop, K., Böcker, S. & Zamboni, N. MSNovelist: de novo structure generation from mass spectra. *Nature Methods* **19**, 865–870 (2022).

[6] Shrivastava, A. D. *et al.* MassGenie: A Transformer-Based Deep Learning Method for Identifying Small Molecules from Their Mass Spectra. *Biomolecules* **11**, 1793 (2021).

[7] Wang, F. *et al.* CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Analytical Chemistry* **93**, 11692–11700 (2021).

[8] Wei, J. N., Belanger, D., Adams, R. P. & Sculley, D. Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks. *ACS Central Science* **5**, 700–708 (2019). `1811.08545`.

[9] Murphy, M. *et al.* Efficiently predicting high resolution mass spectra with graph neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23 (JMLR.org, 2023).

[10] Heinonen, M., Shen, H., Zamboni, N. & Rousu, J. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics* **28**, 2333–2341 (2012).

[11] Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* **16**, 299–302 (2019).

[12] Ji, H., Deng, H., Lu, H. & Zhang, Z. Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks. *Analytical Chemistry* **92**, 8649–8653 (2020).

[13] Fan, Z., Alley, A., Ghaffari, K. & Ressom, H. W. MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation. *Metabolomics* **16**, 104 (2020).

[14] Cao, L. *et al.* MolDiscovery: learning mass spectrometry fragmentation of small molecules. *Nature Communications* **12**, 3718 (2021).

[15] Yang, Q. *et al.* Ultra-fast and accurate electron ionization mass spectrum matching for compound identification with million-scale in-silico library. *Nature Communications* **14**, 3722 (2023).

[16] Stein, S. E. National Institute of Standards and Technology (NIST) Mass Spectral Library. `http://chemdata.nist.gov` (2020).

[17] McLafferty, F. W. Wiley Registry of Mass Spectral Data. `https://sciencesolutions.wiley.com` (2023).

[18] Horai, H. *et al.* MassBank a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* **45**, 703–714 (2010).

[19] Grimme, S. Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules. *Angewandte Chemie International Edition* **52**, 6306–6312 (2013).

[20] Wang, S., Kind, T., Tantillo, D. J. & Fiehn, O. Predicting in silico electron ionization mass spectra using quantum chemistry. *Journal of Cheminformatics* **12**, 63 (2020).

[21] Kind, T. *et al.* Identification of small molecules using accurate mass MS/MS search. *Mass Spectrometry Reviews* **37**, 513–532 (2018).

[22] Steve, S., Joel, K. & Phil, G. The Finnigan Library Search Program. Tech. Rep., Finnigan Corp., March (1978).

[23] McLafferty, F. W., Hertel, R. H. & Villwock, R. D. Probability based matching of mass spectra. rapid identification of specific compounds in mixtures. *Organic Mass Spectrometry* **9**, 690–702 (1974).

[24] Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry* **5**, 859–866 (1994).

[25] Rasmussen, G. T. & Isenhour, T. L. The Evaluation of Mass Spectral Search Algorithms. *Journal of Chemical Information and Computer Sciences* **19**, 179–186 (1979).

[26] Hertz, H. S., Hites, R. A. & Biemann, K. Identification of mass spectra by computer-searching a file of known spectra. *Analytical Chemistry* **43**, 681–691 (1971).

[27] Aisporna, A. *et al.* Neutral Loss Mass Spectral Data Enhances Molecular Similarity Analysis in METLIN. *Journal of the American Society for Mass Spectrometry* **33**, 530–534 (2022).

[28] Koo, I., Zhang, X. & Kim, S. Wavelet- and Fourier-Transform-Based Spectrum Similarity Approaches to Compound Identification in Gas Chromatography/Mass Spectrometry. *Analytical Chemistry* **83**, 5631–5638 (2011).

[29] Kim, S. *et al.* Compound Identification Using Partial and Semipartial Correlations for Gas Chromatography–Mass Spectrometry Data. *Analytical Chemistry* **84**, 6477–6487 (2012).

[30] Moorthy, A. & Anthony. *Pattern similarity measures applied to mass spectra* (SEMA SIMAI Springer Series, Milano,, 2021).

[31] Allen, F., Pon, A., Greiner, R. & Wishart, D. Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification. *Analytical Chemistry* **88**, 7689–7697 (2016).

[32] Demuth, W., Karlovits, M. & Varmuza, K. Spectral similarity versus structural similarity: mass spectrometry. *Analytica Chimica Acta* **516**, 75–85 (2004).

[33] Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

[34] Schomburg, K., Ehrlich, H. C., Stierand, K. & Rarey, M. Chemical pattern visualization in 2D - The SMARTSviewer. *J. Cheminformatics* **3**, 2–3 (2011).

[35] Hoffmann, M. A. *et al.* High-confidence structural annotation of metabolites absent from spectral libraries. *Nature Biotechnology* **40**, 411–421 (2022).

[36] Goldman, S. *et al.* Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence* **5**, 965–979 (2023).

[37] Ridder, L., Hooft, J. J. J. v. d. & Verhoeven, S. Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa. *Mass Spectrometry* **3**, S0033–S0033 (2014).

[38] Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **11**, 98–110 (2015). `1312.0264`.

[39] Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics* **8**, 3 (2016).

[40] Scientific, T. F. Mass frontier spectral interpretation software. Available on: `https://www.thermofisher.com/`.

[41] Ucak, U. V., Kang, T., Ko, J. & Lee, J. Substructure-based neural machine translation for retrosynthetic prediction. *J. Cheminformatics* **13**, 1–15 (2021).

[42] Ucak, U. V., Ashyrmamatov, I. & Lee, J. Reconstruction of lossless molecular representations from fingerprints. *Journal of Cheminformatics* **15**, 26 (2023).

[43] Ucak, U. V., Ashyrmamatov, I. & Lee, J. Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization. *Journal of Cheminformatics* **15**, 55 (2023).

[44] Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Research* **45**, D945–D954 (2016).

[45] Stevenson, J. M. *et al.* Schr\"odinger-ANI: An Eight-Element Neural Network Interaction Potential with Greatly Expanded Coverage of Druglike Chemical Space (2019). Preprint at `https://arxiv.org/abs/1912.05079`.

[46] Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems* **4**, 3104–3112 (2014).

[47] Bahdanau, D., Cho, K. H. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* 1–15 (2015).

[48] Ucak, U. V., Ashyrmamatov, I., Ko, J. & Lee, J. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nat. Commun.* **13**, 1186 (2022).

[49] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).