Hierarchical Deep Learning for De Novo Molecular Structure Prediction from EI-MS Spectra

Mohammad Falah, Anna Wilbik, Marcin Pietrasik
Department of Advanced Computing Sciences
Maastricht University
Maastricht, The Netherlands

John Mommers
Engineering Materials
Envalior
Geleen, The Netherlands

***Abstract*—Inferring molecular structures directly from electron-ionization mass spectrometry (EI-MS) spectra remains a challenging problem in chemoinformatics. While recent deep learning approaches have shown promise, most operate over the entire chemical space using a single model, limiting scalability and interpretability. This paper proposes a hierarchical deep learning framework for de novo molecular structure prediction from EI-MS spectra using SELFIES representations. The framework progressively narrows the chemical search space through a sequence of stages comprising group classification, subgroup classification, clustering, and molecular structure generation. The approach is evaluated on 183,516 compounds from the NIST EI-MS dataset using an end-to-end pipeline. Results show that the hierarchical model achieves a Tanimoto similarity of 1.0 for 39% of test molecules overall, with 41% accuracy for non-aromatic and 37% for aromatic compounds, outperforming non-hierarchical baselines on several subsets. Comparative analysis against state-of-the-art methods demonstrates that the proposed framework offers competitive accuracy while enhancing interpretability and robustness for de novo molecular identification. These findings highlight the effectiveness of hierarchical modeling combined with SELFIES encoding for molecular structure elucidation from EI-MS data.**

***Index Terms*—Molecular structure prediction, deep learning, mass spectral analysis, SELFIES**

I. Introduction

The ability to determine the molecular structure of unknown compounds is a core challenge of modern chemistry, with applications in fields including environmental monitoring [1], food science [2], and forensic toxicology [3]. Electron impact mass spectrometry (EI-MS) is one of the most commonly adopted methods for this purpose, in large part due to its capacity for high-resolution chromatographic separation combined with sensitive mass-spectrometric detection of complex mixtures [4]. The spectra obtained from EI-MS serve as molecular fingerprints, but interpreting them to recover full molecular structures remains challenging since EI-MS often produces complex fragmentation patterns, making it harder to determine the full molecular structure directly [5]. To support identification, spectral libraries are frequently used [6, 7]. However, these libraries are incomplete in their coverage, and when a molecule is absent, identification must rely on manual spectral interpretation, which is time-consuming, expert-dependent, and frequently inconclusive. Structural identification from EI-MS data is often composed of two related tasks: the forward problem and the inverse problem. The forward problem involves predicting the mass spectrum produced by a given molecular structure under electron-ionization conditions, whereas the inverse problem consists of inferring the underlying molecular structure from an observed mass spectrum. While the forward task has seen notable progress in recent years [8], the inverse problem remains an open challenge and is the primary focus of this work. The challenge in the inverse problem arises from the many-to-one nature of fragmentation, where structurally distinct molecules can produce similar spectra [9]. Recent advances in deep learning (DL) have created new opportunities to address this challenge. DL models can learn the relationship between mass spectra and molecular structures [10], while molecular representation schemes such as SELFIES [11] guarantee the chemical validity of generated molecules. In this work, we investigate a hierarchical DL framework for molecular structure prediction from EI-MS spectra. Specifically, our proposed approach progressively narrows the chemical search space by decomposing the problem into staged sub-problems involving classification, clustering, and generation. Such an approach contrasts with current state-of-the-art approaches, which employ a single prediction model over the entire chemical space. We demonstrate that by adopting a hierarchical DL approach, we are able to outperform existing approaches.

II. Related Work

The task of predicting molecular structures from mass spectra has received considerable research attention in recent years. One of the first models proposed, DeepEI [6], first predicts molecular fingerprints from EI-MS spectra, before retrieving candidate molecules from structure libraries, combining DL with database matching. Similarly, msFineAnalysis AI by JEOL[1] leverages database matching on EI-MS spectra for ranking structural candidates from PubChem[2], a large publicly-available molecular knowledge base. DeepCDM [7] is a DL model designed to predict spectra for derivatized molecules, which are often missing from reference libraries. As these approaches rely upon a large existing reference database, they are therefore incapable of de novo molecular prediction. De novo molecular prediction approaches

[1]https://www.jeol.com/
[2]https://pubchem.ncbi.nlm.nih.gov/

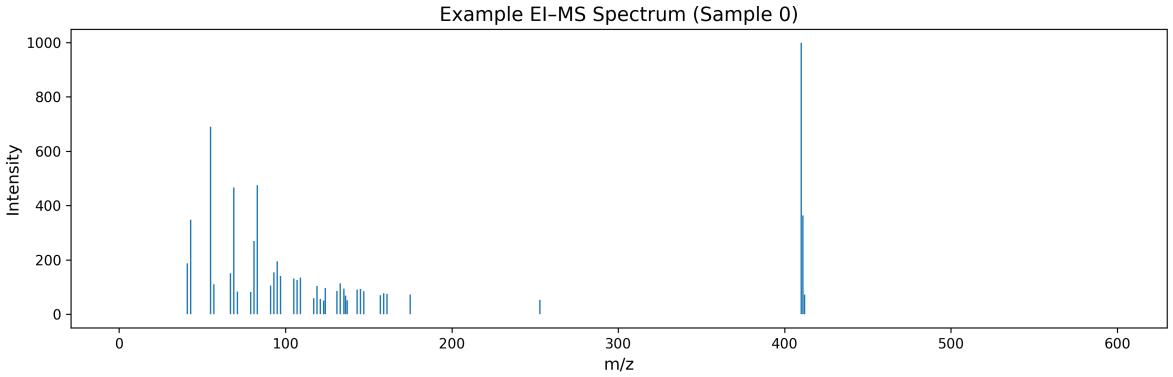

Fig. 1: Example EI–MS spectrum extracted from the dataset. This figure illustrates the characteristic peak intensity distribution across 600 mass-to-charge (m/z) channels, demonstrating the typical fragmentation pattern observed in the dataset used in this study.

Spec2Mol [10] and Mass2SMILES [12] propose DL models operating on MS-MS spectra to generate molecules using SMILES [13], a molecular encoding scheme that represents chemical structures as linear strings of ASCII characters encoding atoms, bonds, branching, and ring closures. Despite promising results, the models struggle with complex structures and lower-quality spectra. Similarly, SpecTUS [14] leverages a transformer-based approach for EI-MS spectra, also using SMILES as the output encoding. One of the issues with using a SMILES-based encoding scheme is the potential for generating chemically invalid structures, as chemical validity is not guaranteed by the encoding. MASSISTANT [15] resolves this issue by leveraging a competing encoding scheme, SELFIES [11], which inherently ensures structural validity. The approach utilizes a simple feed forward neural network architecture to predict one-hot encoded SELFIES tokens to obtain state-of-the-art results on the widely used NIST dataset [16].

III. Methodology

In this section, we outline our molecular prediction methodology by first describing the dataset used before defining our full model architecture. Our Python implementation has been made publicly available on GitHub[3] to ensure reproducibility.

A. Dataset

The NIST08 EI-MS dataset [16] is the primary data source used in our work, with 183,516 records used to train and evaluate our framework. Each record corresponds to a known compound and includes an EI-MS spectrum, along with metadata such as name, formula, and molecular weight as well as a corresponding SMILES representation. Each mass spectrum is represented as peak intensity values over discrete mass-to-charge (m/z) channels from 1 to 600, providing a fixed length vector for all compounds. A visualization of this is provided in Figure 1. The dataset covers a broad range of chemical diversity while remaining within the analytical range of standard GC-MS with electron ionization, ensuring that fragmentation patterns are information-dense and free from sparsity introduced by high-mass fragments. Compounds were assigned to aromatic and non-aromatic groups and then split by atomic composition into 43 chemical subgroups, including 17 non-aromatic subgroups (68,613 molecules) and 26 aromatic subgroups (111,553 molecules). The subgroups represent different classes of aromatic and non-aromatic compounds, defined by their elemental composition (e.g., presence of N, O, S, halogens) and, in some cases, heteroatom-containing functional groups.

Each spectrum (intensities 0-999 over 1-600 m/z) was matched with a SMILES string. Incorrect SMILES were removed, and the remaining SMILES were transformed to SELFIES tokens and one-hot encoded as a multi-dimensional array encoding. To obtain labels for training the subgroup classification model, subgroup labels were derived directly from the dataset filenames. Each molecule was assigned the subgroup indicated by the file from which it originated. This rule-based labelling was applied consistently across the dataset, ensuring that all subgroup classes present in the data were represented.

B. Hierarchical Model Architecture

The framework follows a hierarchical DL architecture that progressively refines molecular structure prediction from EI-MS spectra. The chemical search space is narrowed iteratively by integrating classification and clustering models before molecular structure generation. This hierarchical design ensures each stage operates on more specific and consistent data subsets, reducing ambiguity and enhancing model specialization. The process begins with a binary group classification model that determines whether a sample is aromatic or non-aromatic. Once the group is identified, the output is used to select the subgroup classification model. Separate subgroup models are trained for the aromatic and non-aromatic groups to predict atomic composition classes such as CHO, CHNO,

[3]https://github.com/Mohammadfalah2400/A-Hierarchical-Deep-Learning-Framework-

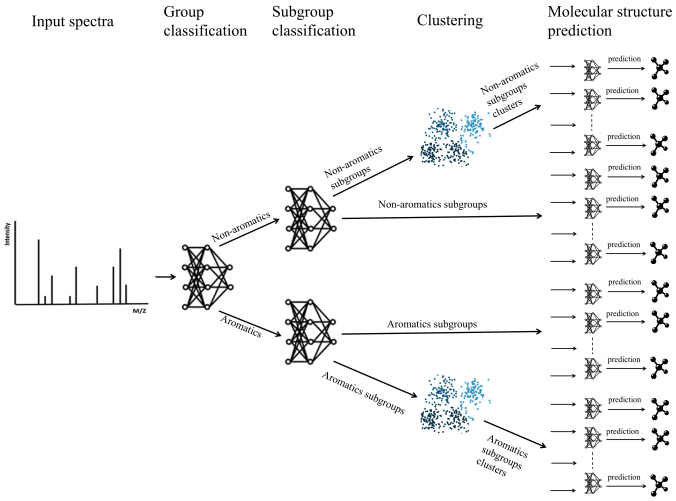

Fig. 2: Schematic overview of the hierarchical model architecture.

and CHOS. This provides finer chemical differentiation and further reduces the search space for structure prediction. After subgroup classification, each subgroup is processed independently. Within every subgroup (e.g., Aromatic CHO, Aromatic CHNO, Non-Aromatic CHO, etc.), a clustering algorithm splits molecules into smaller, structurally consistent clusters. Because each subgroup has its own distribution of molecules, each subgroup generates its own clusters. At the structure prediction stage, multiple models are trained for each subgroup. First, a subgroup model is trained on all molecules in the subgroups (without clustering) as a non-clustered baseline. Then, a dedicated cluster-specific model is trained for each cluster, so one subgroup can have several models corresponding to each of its clusters. Figure 2 provides a visualization of the hierarchical model architecture. What follows is a brief description of the model's components. The layer details of our classification models can be found in our implementation repository on GitHub.

1) Group Classification Model: To separate aromatic from non-aromatic groups given the input mass spectra, we leveraged a one-dimensional convolutional neural network (CNN) as the group classification model. A CNN was chosen because EI-MS spectra are one dimensional signals for which convolutional filters can capture local fragment patterns and hierarchical spectral features. CNNs have also been shown to be parameter efficient and robust to peak shifts, which commonly occur in spectral data, making them well suited for spectroscopy tasks. Moreover, prior studies have demonstrated their effectiveness for spectral-to-structure translation [17–19]. The model takes as input 600 intensity values corresponding to discrete m/z channels from 1 to 600. Training was performed using the Adam optimizer [20] (LR $= 0.001$) with sparse categorical cross-entropy for 100 epochs (batch size 2048). Hyperparameters were selected through a manual search, beginning with simpler CNN versions and increasing complexity, whilst keeping only models with stable validation results.

2) Subgroup Classification Models: Following binary group classification, two subgroup classification models were employed corresponding to the aromatic and non-aromatic chemical spaces. As before, a CNN was chosen for each of the subgroup classification models. As multi-class classifiers, both models took as input mass spectra and output the probability of belonging to each subgroup. There were 26 subgroups in the aromatics group and 17 subgroups in the non-aromatics group. The models were trained with Adam (LR $= 0.001$) using sparse categorical cross-entropy for 300 epochs (batch size 2048), following the same manual tuning strategy as the binary model.

3) Subgroup Clustering: Two approaches were considered following subgroup classification: with and without clustering. In the former, subgroup spectra were clustered into more similar chemical spaces and molecular prediction models were trained on each cluster. In the latter, molecular prediction models were trained on full subgroups. Initially, we considered hierarchical agglomerative clustering [21], DBSCAN [22], and k-means [23] clustering approaches, and assessing their performance using validation metrics and internal cluster analysis including distance/similarity checks with a combination of the elbow method and silhouette score [24] used to determine the number of clusters.

4) Molecular Structure Prediction Models: We employed the MASSISTANT model for molecular structure prediction due to its state-of-the-art performance and use of SELFIES as the output encoding. The model implemented through DeepChem's Multitask Classifier[4], a fully connected neural network for multitask classification. As with the previous classifiers, the model takes in a mass spectrum as input, predicting one-hot SELFIES components through shared hidden layers and segmented output layers for each SELFIES position, allowing multi label prediction and component level interpretability. The SELFIES output follows a fixed and predefined token order, allowing direct position-wise decoding. Training used stratified training and testing splits, and hyperparameters were selected through an extensive manual search. As mentioned earlier, the model was implemented in two settings: trained on full subgroups without clustering as a baseline, and trained on each subgroup cluster to specialize on more similar compounds, allowing subgroup versus cluster level comparison in the hierarchical pipeline.

IV. Evaluation

The evaluation of our hierarchical model was performed on the full, end-to-end pipeline as well as each of the pipeline's components. At the component level, the classification models were evaluated using a standard training (70%), validation (10%), and testing (20%) data split. Manual hyperparameter tuning was performed on the validation set. Five-fold cross-validation was employed and mean accuracy, precision, recall, and F_1 values were recorded to ensure the reliability of results.

[4]https://deepchem.readthedocs.io/

TABLE I: Mean component-level results of classification models as evaluation via cross-validation.

Group Type	Acc.	Prec.	Rec.	F_1
Binary Classification Model	0.95	0.95	0.95	0.95
Subgroup Classification Model (NA)	0.88	0.87	0.88	0.87
Subgroup Classification Model (A)	0.72	0.71	0.72	0.71

The silhouette score, Davies-Bouldin index [25], and Dunn index [26] were used to evaluate clustering at the component level. The results showed that k-means produced the best clustering results and was thus used in the remainder of our work. We omit these results from our paper due to space constraints. The molecular structure prediction models were evaluated using Tanimoto similarity (TS) [27], which provides a measure of similarity between two molecular fingerprints such that $0 \leq \text{Tanimoto} \leq 1$ with higher values indicating a higher degree of similarity. In order to calculate the TS, the models' output token probabilities were first converted into a SELFIES sequence (probability > 0.5 indicating a presence of that token) and then converted into a SMILES encoding for TS input. We note that two molecules may have a Tanimoto of 1 even if they do not have an exactly matching chemical structure as long as their molecular fingerprints are identical. The full, end-to-end hierarchical model was evaluated using the training, validation, and testing sets mentioned earlier such that outputs from previous stages served as inputs for subsequent stages, allowing for errors to accumulate in the process and thus provide a complete evaluation of our model.

A. Results

The binary classification model for predicting whether a compound is aromatic or non-aromatic achieved 95% accuracy on the test set. This indicates strong performance for binary molecular group prediction in this stage. The non-aromatic and aromatic subgroup classification models achieved accuracies of 88% and 72%, respectively, indicating that a non-trivial number of misclassified molecules will propagate to subsequent modelling stages. The full classification results are summarized in Table I. The molecular prediction models were evaluated under two variants: with and without clustering. The variant without clustering obtained a TS = 1.0 rate of 20% across all molecules (36% for NA and 12% for A) whereas the variant with clustering obtained a TS = 1.0 rate of 21% (41% for NA and 10% for A). We notice that on the aggregate, clustering improves the molecular prediction results although this is not true across groups and subgroups. Non-aromatic groups are easier under strict matching, while aromatic groups remain more challenging due to higher structural diversity and fragmentation behavior, which often yield fewer diagnostic fragments, resulting in less informative mass spectra. Nonetheless, for the evaluation of the full, end-to-end architecture, we employ the clustering variant. The final results as per the TS = 1.0 rate are summarized at the aggregate, group, and subgroup levels in Table II. We notice that our full hierarchical model achieves a TS = 1.0 rate of 39% (41% on NA and 37% on A) with subgroup models ranging from 45% to 87% on NA subgroups and 12% to 87% on A subgroups. A cursory examination of subgroup results does not indicate a relationship between the number of entries in a subgroup and its prediction quality. We provide illustrative examples of correct (TS = 1.0) and incorrect (TS < 1.0) model predictions in Fig. 3.

TABLE II: Hierarchical model molecular prediction results for 43 subgroups. NA and A indicate non-aromatic and aromatic groups, respectively

Nr.	Subgroup	Entries	Group	TS = 1.00
1	CHO	1671	NA	0.52
2	CHS	924	NA	0.84
3	CHOS	107	NA	0.53
4	CH	4401	NA	0.66
5	CCIHO	1769	NA	0.81
6	CHNO	851	NA	0.56
7	CHNOS	1326	NA	0.50
8	CHNS	491	NA	0.87
9	BrCHO	958	NA	0.84
10	CCIH	576	NA	0.80
11	CHOP	698	NA	0.71
12	CFHO	795	NA	0.77
13	CHN	2715	NA	0.86
14	CFHNO	594	NA	0.61
15	CHOSI	2648	NA	0.53
16	CCIHNO	581	NA	0.45
17	REST	322	NA	0.61
18	CCIHNO	274	A	0.78
19	CHO	604	A	0.68
20	BrCHNO	89	A	0.71
21	CHOS	87	A	0.74
22	CHNOS	521	A	0.63
23	CFHNO	254	A	0.67
24	CCIHNOS	103	A	0.74
25	CHOSI	50	A	0.75
26	CHNO	1840	A	0.12
27	CH	62	A	0.66
28	CHOP	24	A	0.66
29	CHN	201	A	0.71
30	CFHN	26	A	0.72
31	CFHO	105	A	0.66
32	CCIHO	118	A	0.59
33	CCIHINS	32	A	0.18
34	BrCHO	55	A	0.83
35	CFHNOS	44	A	0.71
36	CCIHN	57	A	0.87
37	CHINO	21	A	0.61
38	CCIFHNO	34	A	0.63
39	CHNOSI	32	A	0.47
40	CHNS	95	A	0.81
41	CCIH	28	A	0.57
42	CHS	31	A	0.75
43	REST	307	A	0.70
Avg.	**ALL NA**	**21427**	**NA**	**0.41**
Avg.	**ALL A**	**24993**	**A**	**0.37**
Avg.	**ALL**	**46420**	**NA+A**	**0.39**

B. Discussion

The proposed hierarchical framework integrates classification, clustering, and molecular prediction to reduce the set of possible molecules for mass spectrometric structure identification, providing greater prediction stability and generalizability than single stage models, particularly for unknown molecules not present in spectral libraries. Aromatic molecules were

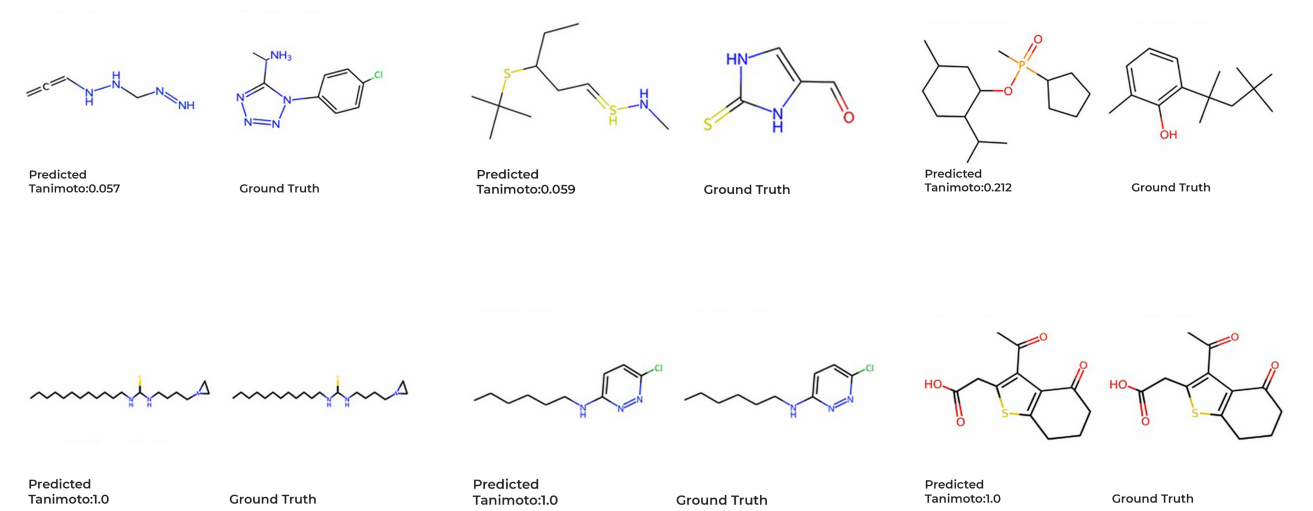

Fig. 3: Examples of correct (TS = 1.0) and incorrect predictions (TS $<$ 1) illustrating subgroup misassignment and prediction errors.

more difficult to predict, partly due to the larger number of aromatic subgroups, which led to smaller cluster sizes and reduced training data per model. In addition, aromatics are frequently more complex and diverse, and many provide EI-MS spectra dominated by persistent ring related fragments, which offer limited information for structure prediction. Furthermore, as noted earlier, dividing the aromatic group into many subgroups creates smaller clusters with less training data for each model, which makes learning more difficult. Finally, clustering exhibits large differences in cluster reliability: some aromatic clusters have noisy spectra or outliers and exhibit low overall similarity. This leads to low performance for those cluster-based models and reduces the average TS = 1.0 value to about 10%. Analysis with cosine similarity and distance matrices shows that clusters with low cosine similarity and high spectral distances are linked to the lowest prediction results, indicating that large intra-cluster differences reduce performance. The classification stage forms a reliable foundation, achieving 95% accuracy in separating aromatic from non-aromatic classes and subgroup accuracies of 72% and 88% for aromatic and non-aromatic molecules. The first stage shows higher accuracy because it distinguishes between only two classes, leading to a higher chance of correct classification, whereas the second stage considers a larger set of classes and therefore produces lower accuracy; nevertheless, it remains useful by reducing the search space for subsequent prediction. Moreover, prediction accuracy is influenced by how informative the spectra are, as noisy or less detailed fragmentation patterns often lead to poorer results. Fig. 3 shows this pattern: TS = 1.0 predictions are common when spectra are easy to read and the subgroup is chosen correctly. When TS $<$ 1, the model may return a close but incorrect structure, which indicates an incorrect subgroup choice or spectra that are noisy or not very informative. Even when TS $<$ 1, the prediction is often close in structure to the real molecule, so it can help scientists reduce the options and concentrate on the most promising structures. Lastly, the hierarchical layout makes it possible to examine how errors transfer between stages. Future extensions could limit early stage errors by using spectrum reconstruction based validation with models such as NEIMS [8] or by testing multiple subgroup models per sample, thereby enhancing robustness and scalability.

TABLE III: Comparison against results obtained from the literature.

Model	Data	Result
DeepEI [6]	Low-res EI-MS	90% fingerprint accuracy
Spec2Mol [10]	High-res EI-MS/MS	7% exact structure
Mass2SMILES [12]	High-res EI-MS/MS	1% TS = 1.0
MSNovelist [28]	High-res EI-MS/MS	25% top-1 accuracy
SpecTUS [14]	Low-res EI-MS	43% TS = 1.0
MASSISTANT [15]	Low-res EI-MS	44% TS = 1.0
Our model	Low-res EI-MS	39% TS = 1.0

We compare our model results against those obtained from competing models in the literature in Table III. We stress that due to the use of different evaluation datasets and metrics, a direct comparison is not possible. Nonetheless, these results contextualize those obtained in our work with respect to the current academic landscape. To perform a direct pairwise comparison, we leverage the implementation provided by the current state-of-the-art MASSISTANT model and evaluate its

TABLE IV: Pairwise comparison of our model against MASSISTANT on identical data subsets.

Data Subset	MASSISTANT	Our Model
CHO1.NO.AROMATIC	50%	48%
NA.CHO.0.1	46%	60%
NA.CHONS.0.1	44%	46%
NON-AROMATIC	36%	50%

performance against ours on identical data subsets. The results of this comparison are summarized in Table IV. The subsets in Table IV split non-aromatic molecules based on their elemental makeup. CHO subsets are usually easier and give better accuracy, whereas subsets that also include N and S tend to have lower accuracy because their spectra are more complex. In general, accuracy drops when chemical complexity becomes higher.

V. Conclusion

In this paper, we proposed a hierarchical deep learning framework for de novo molecular structure prediction from EI-MS spectra. Our approach progressively narrows the chemical search space through staged classification and clustering before molecular generation via SELFIES encoding. Evaluation on the NIST dataset demonstrates that the framework achieves strong end-to-end performance, with notable gains for non-aromatic compounds over existing state-of-the-art methods. Overall, this work underscores the value of combining existing DL approaches in a hierarchical framework to advance automated structure elucidation. While the proposed hierarchical framework shows strong performance compared to existing methods, several extensions could further improve it. Exploring alternative clustering strategies beyond k-means, including custom designed algorithms, may better capture structural patterns and increase cluster homogeneity. Improvements at the prediction stage, such as more advanced DL architectures, ensemble methods, or transfer learning, could enhance robustness, particularly for small or ambiguous chemical groups. Another option is Top-K or beam-search subgroup routing to check several candidates. Increasing the number of instances per cluster and incorporating encoder-decoder architectures are also identified as promising directions. Additionally, developing scalable pipelines and evaluating the framework on larger and more diverse datasets, such as NIST23, would help demonstrate generalization to broader chemical spaces.

References

[1] E. Barón, E. Eljarrat, and D. Barceló, "Gas chromatography/tandem mass spectrometry method for the simultaneous analysis of 19 brominated compounds in environmental and biological samples," *Analytical and bioanalytical chemistry*, vol. 406, no. 29, pp. 7667–7676, 2014.

[2] K. Pastor, M. Ačanski, D. Vujić, G. Bekavac, S. Milovac, and S. Kravić, "Rapid method for small grain and corn flour authentication using gc/ei–ms and multivariate analysis," *Food analytical methods*, vol. 9, no. 2, pp. 443–450, 2016.

[3] K. Synowiec, S. Rojek, M. Maciów-Głab, K. Kula, A. Romańczuk, and M. Kłys, "The role of gc-ei-ms and derivatization in the detection of new psychoactive substances exemplified by 49 synthetic cathinones," *Journal of Analytical Chemistry*, vol. 77, no. 10, pp. 1315–1324, 2022.

[4] O. D. Sparkman, Z. Penton, and F. Kitson, *Gas Chromatography and Mass Spectrometry: A Practical Guide*. Oxford, UK: Academic Press, 2011.

[5] T. Kind and O. Fiehn, "Advances in structure elucidation of small molecules using mass spectrometry," *Bioanalytical reviews*, vol. 2, no. 1, pp. 23–60, 2010.

[6] H. Ji, H. Deng, H. Lu, and Z. Zhang, "Predicting a molecular fingerprint from an electron ionization mass spectrum with deep neural networks," *Analytical chemistry*, vol. 92, no. 13, pp. 8649–8653, 2020.

[7] B. Chen, H. Li, R. Huang, Y. Tang, and F. Li, "Deep learning prediction of electrospray ionization tandem mass spectra of chemically derived molecules," *Nature Communications*, vol. 15, no. 1, p. 8396, 2024.

[8] J. N. Wei, D. Belanger, R. P. Adams, and D. Sculley, "Rapid prediction of electron–ionization mass spectrometry using neural networks," *ACS Central Science*, vol. 5, no. 4, pp. 700–708, 2019.

[9] J. H. Gross, *Mass Spectrometry: A Textbook*. Berlin, Germany: Springer, 2011.

[10] E. E. Litsa, V. Chenthamarakshan, P. Das, and L. E. Kavraki, "An end-to-end deep learning framework for translating mass spectra to de-novo molecules," *Communications Chemistry*, vol. 6, no. 1, p. 132, 2023.

[11] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-referencing embedded strings (selfies): A 100% robust molecular string representation," *Machine Learning: Science and Technology*, vol. 1, no. 4, p. 045024, 2020.

[12] D. Elser, F. Huber, and E. Gaquerel, "Mass2smiles: deep learning based fast prediction of structures and functional groups directly from high-resolution ms/ms spectra," *bioRxiv*, pp. 2023–07, 2023.

[13] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[14] A. Hájek *et al.*, "SpecTUS: Spectral translator for unknown structures," *arXiv preprint arXiv:2502.05114*, 2025.

[15] J. Mommers, L. Barta, M. Pietrasik, and A. Wilbik, "Massistant: A deep learning model for de novo molecular structure prediction from ei-ms spectra via selfies encoding," *Journal of Chromatography A*, vol. 1759, p. 466216, 2025.

[16] A. Mikaia, P. E. W. V. EI, V. Z. EI, D. Z. EI, O. D. S. EI, P. Neta, I. Z. RI, P. Linstrom, Y. Mirokhin, D. Tchekhovskoi *et al.*, "Nist standard reference database 1a," *Standard Reference Data*, 2014.

[17] N. H. Tran, X. Zhang, L. Xin, B. Shan, and M. Li, "De novo peptide sequencing by deep learning," *Proceedings of the National Academy of Sciences*, vol. 114, no. 31, pp. 8247–8252, 2017.

[18] Y.-M. Lin, C.-T. Chen, and J.-M. Chang, "Ms2cnn: predicting ms/ms spectrum based on protein sequence using deep convolutional neural networks," *BMC genomics*, vol. 20, no. Suppl 9, p. 906, 2019.

[19] K. Seddiki, P. Saudemont, F. Precioso, N. Ogrinc, M. Wisztorski, M. Salzet, I. Fournier, and A. Droit, "Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification," *Nature communications*, vol. 11, no. 1, p. 5595, 2020.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[22] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[23] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[24] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[25] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 2009.

[26] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[27] T. T. Tanimoto, "An elementary mathematical theory of classification and prediction," 1958.

[28] M. A. Stravs, K. Dührkop, S. Böcker, and N. Zamboni, "Msnovelist: de novo structure generation from mass spectra," *Nature Methods*, vol. 19, no. 7, pp. 865–870, 2022.