M.S. THESIS

Mass Spectra Prediction through Structural Motif-based Graph Neural Networks

구조 모티프 기반 그래프 신경망을 이용한 질량 스펙트럼 예측

BY

Jiwon Park

AUGUST 2023

INTERDISCIPLINARY PROGRAM IN
ARTIFICIAL INTELLIGENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

M.S. THESIS

Mass Spectra Prediction through Structural Motif-based Graph Neural Networks

구조 모티프 기반 그래프 신경망을 이용한 질량 스펙트럼 예측

BY

Jiwon Park

AUGUST 2023

INTERDISCIPLINARY PROGRAM IN
ARTIFICIAL INTELLIGENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Mass Spectra Prediction through Structural Motif-based Graph Neural Networks

구조 모티프 기반 그래프 신경망을 이용한 질량 스펙트럼 예측

지도교수 윤 성 로

이 논문을 공학석사 학위논문으로 제출함

2023년 8월

서울대학교 대학원

협동과정 인공지능 전공

박 지 원

박지원의 공학석사 학위 논문을 인준함

2023년 8월

위 원 장 : 조 정 효 (인)

부위원장: 윤 성 로 (인)

위 원: 김 승 (인)

Abstract

Mass spectrometry is widely used in various fields such as drug discovery, chemical synthesis, and environmental chemistry for identifying molecular structures. Mass spectra are collections of ionized fragments from a target molecule, and the fragmentation patterns within the spectra contain crucial information about the molecule. In the analysis of mass spectra to identify molecule structures, a common approach is to perform a spectral library search. This method involves matching the unknown spectra with a database of mass spectra from known materials. However, the effectiveness of search-based methods is limited by the availability of the mass spectra database.

In this work, we propose the Motif-based Mass Spectrum Prediction Networks (MoMS-Net) that incorporates structural motifs to predict mass spectra based on molecular structure. A motif refers to a frequently occurring subgraph or a related functional group in molecules. We leverage the information from structural motifs for applying GNNs because motifs are associated with fragmentation patterns and aid in mass spectra prediction. We evaluate our model on various types of mass spectra and demonstrate its superior performance compared to other deep learning models. MoMS-Net can consider substructure at the graph level, allowing it to incorporate long-range dependencies while requiring less memory than the graph transformer model.

keywords: mass spectra, GNNs, Motif, deep learning

student number: 2021-25101

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

Mass spectrometry (MS) [1, 2] is an essential analytical technique for identifying molecular structures in unknown samples [3, 4, 5]. In this technique, a molecule is ionized, and its fragment ions are detected by a mass analyzer, which records information about the mass-to-charge ratio (m/z). By analyzing the mass spectrum, which provides the m/z values and their relative abundances, it is possible to infer the molecular structure of the original chemical.

Modeling the fragmentation patterns for ionized molecules in order to analyze the mass spectrum is challenging. While some domain knowledge-based rules can be useful for certain types of molecules, it becomes difficult to apply them to smaller fragments with diverse functional groups.

The interpretation of mass spectra typically relies on library search, which compare the spectra with a large database of known molecules [6, 7]. While there are various extensive mass spectra libraries available, such as the National Institute of Standards and Technology (NIST) [8], Wiley [9], and Mass Bank of North America (MoNA), the search-based method is limited by its ability to access known materials and does not provide information on the mass spectra of new molecules. An alternative approach is to use *de novo* methods, which aim to directly predict the molecular structure from the input spectrum. However, these methods often have low accuracy and are challenging

to use effectively.

An approach to address the coverage issue in library search is to enhance existing libraries by incorporating predicted mass spectra generated by a model. Mass spectrum prediction models utilize either quantum mechanical calculations [10, 11, 12], or machine learning techniques [13]. These methods aim to predict the fragmentation patterns that occur after ionization. Quantum mechanical calculations require precise computation of orbital energies, but they are computationally inefficient. On the other hand, machine learning approaches can provide faster predictions, but they may lack the ability to simulate diverse and detailed fragmentation processes.

Recently, deep learning has been significantly developed in the fields of computer vision and natural language processing. Moreover, there has been a growing interest in applying deep learning to material science and drug discovery. Graph Neural Networks (GNNs) are widely used in material science and bioinformatics to predict chemical properties and generate new molecules, because molecules, which consist of atoms and bonds, can be represented as graphs with nodes and edges. Several studies have focused on predicting mass spectra using MLPs, GNNs, and graph transformer models [14, 15, 16, 17].

The properties of a molecule are highly dependent on its molecular structure, especially the functional groups. Even if two molecules have the same chemical composition, they can exhibit different properties if their functional groups differ. Motifs, which are important and frequently occurring subgraphs, can be used to model molecular functional groups.

In this work, we propose the Heterogeneous Motif Graph Neural Network (MoMS-Net) for predicting mass spectra, as shown in Fig. 1.1. We utilize motifs [18] because they are related to the stability of fragment ions and the fragmentation pattern in mass spectra. The MoMS-Net model consists of two GNNs: one for the molecule graph and the other for the heterogeneous motif graph. The molecule graph is generated based on the molecule itself, where the nodes and edges correspond to the atoms and bonds in

the molecule, respectively. The heterogeneous motif graph consists of all molecules in the dataset and the motifs in the motif vocabulary, with nodes representing the motifs and edges representing the relationship between the molecule and motif. In general, GNNs struggle to consider long-range dependencies as node information is updated by pooling neighboring nodes. While deep layers are typically required to incorporate long-range dependencies in GNNs, this can lead to oversmoothing problems where all nodes become similar, resulting in decreased performance. However, our model can consider the relationship with subgraphs at the graph level, allowing it to effectively incorporate long-range dependency effects. The graph transformer model has demonstrated good performance in predicting mass spectra but requires a significant amount of memory during training [17]. In contrast, our model requires less memory than the graph transformer. Ultimately, our model achieves the state-of-the-art performance in predicting mass spectra.

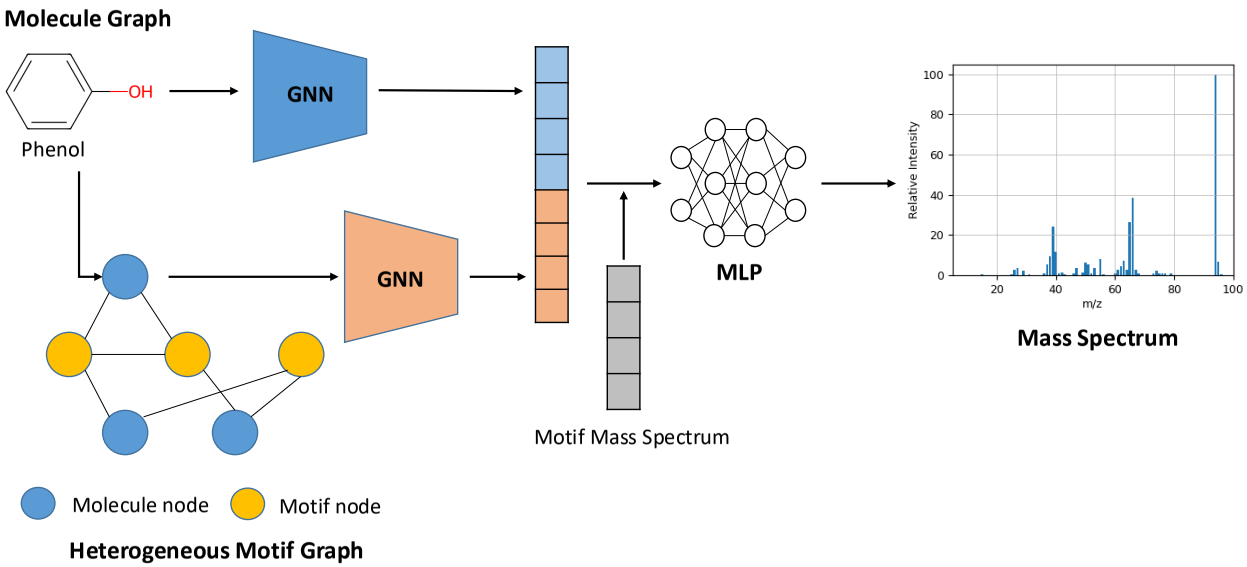


Figure 1.1: Overall architecture of MoMS-Net. The model consists of two GNNs for molecule graph and heterogeneous motif graph. We concatenate graph embeddings from two GNNs and apply FCL to predict mass spectra.

Chapter 2

Backgroud

2.1 Mass Spectrometry

Mass spectrometry [1] is an essential tool in analytical chemistry and drug discovery because it can give the information of molecular structure. Mass spectrometry analyzes the mass-to-charge (m/z) ratio after ionization and fragmentation of molecules. A mass spectrum is expressed as a plot of ion abundance versus m/z. Mass spectrometry analyzes gas-phase ions and is composed of three main components such as an ionization source, a mass analyser and a detector. Mass Spectrum is also categorized according to the resolution. Low resolution MS usually detect ion in the unit of integer m/z but high resolution MS detect ions as unit of 0.0001 Dalton. A peak in high resolution MS can be determined as its chemical composition because all atoms have different atomic weight and composition can be separated in 0.0001 Dalton unit. But high resolution mass spectrometry needs more expensive and complex equipment. *De novo* method to predict the molecular structure from mass spectrum is very challenging [19]. It refers to the process of determining the chemical structure of a molecule solely from experimental data, without relying on any prior knowledge or reference databases. While *de novo* sequencing methods have been developed for proteins and peptides [20, 21], *de novo* structure prediction for small organic molecules based solely on mass spec-

tra is still an active area of research with limited success [22]. Alternative method for mass spectrum analysis is done by comparing cosine similarity with database. There are many mass spectra libraries from NIST, MoNA and Wiley. But these library are constructed by getting mass spectrum using known material so it has coverage issue due to lack of spectra for unknown and new chemicals. It needs to augment database with generation of mass spectra with a model.

2.2 Motif Generation

We denote graph as $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ where $\mathcal{N}$ is a set of nodes and $\mathcal{E}$ is a set of edges. Subgraph has some nodes and corresponding edges, $\mathcal{G}^S = (\mathcal{N}^S, \mathcal{E}^S)$ where $\mathcal{N}^S \subseteq \mathcal{N}$ and $\mathcal{E}^S \subseteq \mathcal{E}$. Motif is common subgraph in molecule graph and has important role such as functional group. Edges and cycles in molecular graphs can be seen as bonds and rings, respectively. Many GNNs models utilize motif to improve the capability for property prediction, drug-gene interaction prediction and molecular generation [23, 24, 25, 26, 27, 28]. In the prediction task, motif is helpful to improve expressivity of the model and reflect chemical properties. In the generation task, motif can regularize atomic combinations and makes training and inference efficient because of larger building block than atoms. Motif generation can be done by rule-based method and data-based method.

2.2.1 Rule-based Method

The vocabulary of motifs is made by simple hand-crafted rules or utilize external chemical fragment libraries, which could be different from fragmentation pattern in the dataset [29, 30, 31]. This method defines several rules to preserve rings and conjugated bonds, and to break down bonds between specific groups. The goal is to decompose molecules into non-overlapping fragments that are meaningful within the domain area. But rule-based method is insufficient and defective because hand-crafted

rules cannot cover all fragmentation patterns and some molecules with new functional groups cannot follow pre-defined rules.

2.2.2 Data-based Method

Data-based method construct motif vocabulary by learning from the dataset. The motif mining algorithm is motivated by byte-pair encoding (BPE) [32], which is widely used in natural language processing (NLP) for subword tokenization. Compared with NLP, molecules have more complex structures due to different bond type and ring. Vocabulary starts from all distinct atoms of molecules in the dataset and then merges as a new one to update vocabulary. Motif constructed such a merge-and-update method can represent the largest and frequent patterns of molecules. The distribution of motif size also is wider than previous rule-based method and vocabulary can cover more diverse patterns. Data-based method is universal to apply any dataset such as molecules and proteins without domain knowledge for fragmentation rules. Z. Geng et al. [33] proposed a data-driven method to generate motif automatically by merging subgraphs based on their frequency. X. Kong et al. [34] define principal subgraph as the largest frequency in the data and proposed similar vocabulary generation process by merging two subgraphs.

2.3 Motif-based GNNs

Motif is used to increase expressivity of GNNs. It is widely used in the tasks node and graph prediction task, drug-gene interaction, molecule generation and self-supervised learning. Many research uses subgraph to increase the representation ability of GNNs. Nested Graph Neural Network (NGNN) [24] uses another GNN to embed node and show that subgraphs have higher expressivity than subtree. Graph Substructure Network (GSN) [25] uses message passing scheme with substructure-encoding and demonstrate higher expressivity and generalization. Heterogeneous Motif Graph Neural Net-

work (MoMS-Net) [23] uses both atomic-level and motif-level embeddings. They first build a motif vocabulary by searching all molecule graphs and extracting important subgraphs. Vocabulary is initialized by all bonds and rings. They remove duplicate and sort vocabulary by TF-IDF value to keep the most important motifs. They construct heterogeneous motif graph by connecting molecules through motifs. These models show that incorporating motifs increase graph classification performances. Motifs can be used as building blocks to generate molecules faster and more realistic [26]. Motif-based Graph Self-Supervised Learning (MGSSL) [35] generate motifs by BRICS fragmentation rules and then perform multi-level SSL in atom and motif levels. It shows better performance on molecular property prediction tasks than different pre-training strategies. Subgraph-level contrastive learning uses subgraph from itself and others as positive and negative sample respectively [36]. There are also researches regarding drug-gene interaction prediction through subgraph patterns to predict unseen data and multi-relational interactions [28].

2.4 Neural Networks for Mass Spectra Prediction

2.4.1 Multi-Layer Perceptron

Multi-layer perceptron (MLP) is a sequence of many perceptron and called as feedforward deep neural network. J. Wei et al. [14] applied MLP to predict mass spectra for molecules. Molecules are mapped to Extended Circular Fingerprints (ECPFs) [37]. Input vector uses a fingerprint length of 4096 with a radius of 2. They designed a bidirectional prediction combined with forward and reverse prediction. Forward prediction is an affine transformation of input features. In reverse prediction, they defined ion peaks as a function of the fragment groups which were detached from the original molecule. Forward and reverse predictions are combined to final prediction by gate function.

$$p_i^f(x) = w_i^{fT} f(x) + b_i^f \tag{2.1}$$

$$p^r_{M(x)+\tau-i}(x) = w_i^{rT} f(x) + b_i^r \tag{2.2}$$

$$p_i(x) = \sigma(\texttt{gate}_i) p_i^f(x) + (1 - \sigma(\texttt{gate}_i)) p_i^r(x) \tag{2.3}$$

w and b are the model's weights and biases. $\texttt{gate}_i$ is an affine transformation of f(x) and σ is a sigmoid fuction.

2.4.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are widely used to process and analyze grid-like data, such as image and video. CNNs apply local receptive fields (kernels or filters) to input image and can extract features and local patterns, such as shapes, edges and textures. After feature extraction with convolution operation, fully-connected layers are applied for downstream tasks, such as image classification. K. Liu et al. [15] apply CNN to predict mass spectra from peptide. They use encoded peptide with size of 27×23. Other CNN uses SELFIES (self-referencing embedded strings) from molecular structures as input to predict mass spectra [17].

2.4.3 Graph Neural Networks

Graph Neural Networks (GNNs) are operated on graph-structure data which consists of nodes and egdes. GNNs propagete information through graph structure to learn representation nodes and graph. Each node aggregates information from its neighboring nodes and update its representation.

$$H_i^{(l+1)} = \sigma(\sum_{j \in N(i)} H_j^{(l)} W^{(l)} + b^{(l)}) \tag{2.4}$$

$$G = Pool(H) \tag{2.5}$$

H, G is hidden representation of node and graph and σ is activation function. W and b are weight and bias which are learnable parameters. GNNs are widely used in

chemical property prediction tasks because a molecule can be represented to graph whose nodes and edges are correspond to atoms and bonds in the molecule. B. Zhang [16] apply graph convolutional network (GCN) to predict mass spectra. They initialize nodes' features as concatenated one-hot vectors of several atoms' properties such as atom symbol, degree, valence, formal charge, radical charge and so on. Initial features of edges are also represented by one-hot vectors using type of bond, ring, conjugation and chirality. They apply several GCN layers and pool nodes' representations to graph representation and apply MLP layer to predict mass spectra.

$$
\begin{aligned}
a_i &= \text{concatenate}(v_i, \sum_{j=1}^{d} n_j, \sum_{j=1}^{d} e_{ij}) \\
h_{\text{conv}}(v_i) &= \sigma(w^{deg(v_i)} a_i + b^{deg(v_i)}) \\
h_{\text{conv}}(G) &= [h_{\text{conv}}(v_1), h_{\text{conv}}(v_2), h_{\text{conv}}(v_3), \ldots]
\end{aligned}
\tag{2.6}
$$

where d is degree of node. n_j is and neighbor node and e_{ij} is corresponding edge. w and b are weight and bias term. Zhu et al. [38] used GCN to predict mass spectra of liquid chromatography-mass spectrometry (LC-MS).

2.4.4 Graph Transformer

Transformers are neural networks with use of attention. It is originally developed for natural language processing, but is used for many area including computer vision and time-series data. Transformers are specifically designed to capture long-range dependencies in sequences. Unlike CNNs, which rely on local receptive fields and hierarchical feature extraction, Transformers use self-attention mechanisms to consider all positions or tokens in a sequence simultaneously. This makes Transformers more suitable for tasks that require understanding global context and dependencies, such as machine translation, text generation, and image processing. graph transformer is a kind of transformer which make graph information to input sequences. Graph transformer can model global interactions between all nodes in the graph, but GNNs can consider local interaction updating neighborhood information. A. Young et al. [17] proposed

Massformer model to predict tandem mass spectrum prediction with graph transformers. The attention mechanism a_{ij} consider pairwise attention between nodes, shortest path distance between nodes and edge embeddings as Eq. 2.7.

$$a_{ij} = \mathrm{softmax}\left(\frac{(W_Q h_i)^T (W_k h_j)}{\sqrt{d}} + b_{ij} + c_{ij}\right) \tag{2.7}$$

$$c_{ij} = \frac{1}{N} \sum {w_p}^T e_p \tag{2.8}$$

a_{ij} is attention value between node i, j and b_{ij} is a learnable value indexed by the shortest path distance between node i, j. c_{ij} is the edge embedding, averaged by embedding e_p in the shortest path between i and j, and w_p is a learnable weight parameter. Massformer reported better performance in mass spectra prediction compared to CNNs and GNNs.

Chapter 3

Method

Our model, MoMS-Net, consists of two GNNs: one for the molecule graph and another for the heterogeneous motif graph. The molecule GNN utilizes fingerprint models, specifically Morgan, MACCS, and RDKit fingerprints, as inputs. On the other hand, the heterogeneous motif GNN takes into account the molecule-motif relations and molecular weights as inputs. We concatenated the hidden representations from both GNNs with a specific relative ratio. To further fine-tune the hidden representation, we utilize the molecular weight distribution of the molecular ion and a few fragments obtained from RDKit fragmentation.

3.1 Dataset

We used the NIST 2020 MS/MS dataset for both training and evaluation purposes. The NIST dataset is widely employed due to its extensive coverage and convenience in the mass analysis process. It is important to note that mass spectra can vary depending on the acquisition conditions. In our study, we specifically focused on spectra obtained from Fourier Transform (FT) instruments, considering the large amount of available data. Additionally, we took into account the collision cell type, which are collision-induced dissociation (CID) and High-energy C-trap dissociation (HCD). A summary

of the dataset can be found in Table 3.1.

Table 3.1: NIST 2020 MS dataset

Collision Type	# Spectra	# Compounds
FT-CID	27,026	18,257
FT-HCD	322,372	19,620

3.2 Generation of Motif Vocabulary

A motif refers to the most frequent substructure, and some motifs are correspond to functional groups of molecules. To construct a motif vocabulary, we apply the byte-pair encoding (BPE) method introduced by A. Young et al. [33] to identify common patterns from a given dataset D. The goal is to learn the top K most frequent subgraphs from dataset D, where K is a hyperparameter. Each molecule in D is represented as a graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where atoms and bonds correspond to nodes $(\mathcal{V})$ and edges$(\mathcal{E})$. Initially, we consider each atom from the molecules as a single fragment.

We merge two fragments, $\mathcal{F}_i$ and $\mathcal{F}_j$, to create a new fragment, $\mathcal{F}_{ij} = \mathcal{F}_i \oplus \mathcal{F}_i$, using a defined operation "$\oplus$". The merging process involves iteratively updating the merging graphs, $\mathcal{G}_M^{(k)}(\mathcal{V}_M^{(k)}, \mathcal{E}_M^{(k)})$, where the edges come from fragments $\mathcal{F}_i$, $\mathcal{F}_j$ and the connections between the two fragments. The most frequent merged fragment, $\mathcal{F}_{ij}$, is added to the motif vocabulary $\{\mathcal{M}\}$. This process is repeated for K iterations to obtain the motif vocabulary.

To represent molecules, we utilize the Simplified Molecular Input Line Entry System (SMILES). However, it's important to note that invalid fragments may be created, such as "cc" as two carbons cannot form a ring. To ensure the validity of the fragments, we use the RDKit package to check if the valence of the molecule is incorrect, and abnormal fragments are removed from consideration.

By applying this approach, we can generate a motif vocabulary that captures the frequent substructures in the dataset, enabling further analysis and interpretation of the molecular structures.

3.3 Construction of Heterogeneous Motif Graph

The heterogeneous motif graph is constructed by combining molecule nodes from the molecular dataset and motif nodes from the motif vocabulary. This graph consists of two types of edges connecting the nodes. The first type is the molecule-motif edge, which is created when a molecule contains that motif. The second type is the motif-motif edge, which is established when two motifs share at least one atom. To differentiate the importance of these edges, different weights are assigned based on their types according to Z. Yu et al. [23]. For the molecule-motif edge, the weight is calculated using the TF-IDF (Term Frequency-Inverse Document Frequency) value. For the motif-motif edges, the weight is calculated as the co-occurrence information point-wise mutual information (PMI). So the edge weight A_{ij} between two nodes (i, j) is represented as

$$A_{ij} = \begin{cases} \text{PMI}_{ij}, & \text{if i, j are motifs} \\ \text{TF-IDF}_{ij}, & \text{if i or j is a motif} \\ 0, & \text{Otherwise} \end{cases} \tag{3.1}$$

The PMI value is calculated as

$$\begin{aligned} &\text{PMI}_{ij} = \log \frac{p(i,j)}{p(i)p(j)} \\ &p(i,j) = \frac{N(i,j)}{M}, p(i) = \frac{N(i)}{M}, p(j) = \frac{N(j)}{M}, \end{aligned} \tag{3.2}$$

where $N(i, j)$ is the number of molecules that have motif i and motif j. M is the number of molecules, and $N(i)$ is the number of molecules with motif i.

$$\text{TF-IDF}_{ij} = C(i)_j \left(\log \frac{1+M}{1+N(i)} + 1 \right), \tag{3.3}$$

where $C(i)_j$ is the number of frequency that the motif occurs in the molecule j.

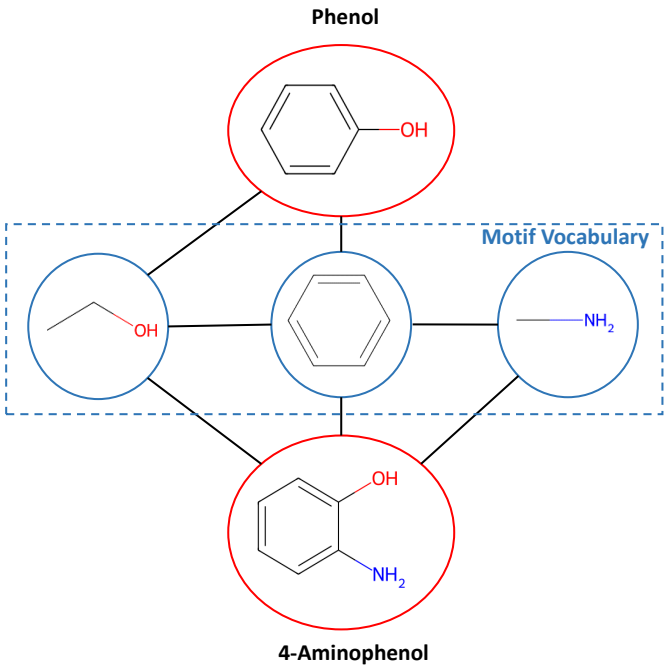


Figure 3.1: Example of heterogeneous motif graph. It consists of molecular nodes and motif nodes. There is two types of edges. Molecule-motif edges exist if the molecule contains that motif. Motif-motif edges exist if two motif share at least one atom.

3.4 Heterogeneous Motif Graph Neural Networks

We apply two different GNNs for molecule graphs and heterogeneous motif graph.

3.4.1 Molecule Graph

The molecule graph represents each atom and bond as nodes and edges, respectively. We utilize a 3-layer Graph Convolutional Network (GCN) to update the atom-level representations. To encode the atom and bond features, we employ the Deep Graph Library (DGL) package, which supports embedding them as either one-hot encoding or numerical values.

Table 3.2: Atom and Bond Features

Types	Features
Atom	mass, type, bond type, degree, total degree, explicit valence, implicit valence, hybridization, total number of H, formal charge, number of radical electrons, is aromatic, is in ring, is chiral
Bond	bond type, is conjugated, is in a ring of any size, stereo configuration

3.4.2 Heterogeneous Motif Graph

For the heterogeneous motif graph, we employ the other 3-layer Graph Isomorphism Network (GIN). The total number of nodes in the heterogeneous graph is the sum of the number of molecules ($|N|$) and the size of the motif vocabulary ($|V|$). The node feature in the heterogeneous motif graph is represented by the occurrence of motifs and their molecule weights. To represent the occurrence of motifs in molecules and other motifs, we create a vector of size $|V|$, where the values indicate motif occurrences. We then reduce the dimension of this vector by applying a linear layer, as the embedding

is sparse, and concatenate it with the molecule weight.

A heterogeneous motif consists of all molecule nodes and motif nodes, as shown in Fig. 3.1. Since the number of molecules can be large (e.g., 27K for CID and 232K for HCD), computational resource limitations may arise. To address this issue, we use an edge sampler to reduce the size of the heterogeneous motif graph. We employ a breadth-first algorithm for hop-by-hop sampling from a starting node. The first-hop neighbors of molecule nodes are motif nodes only. We use a 3-hop sampler, denoted as $[s_1, s_2, s_3]$, where s_i represents the number of nodes to be sampled. Before applying GINs, we first utilize a 2-layer MLP for input embedding.

3.5 Mass Spectra of Motif

After obtaining the graph embeddings for the heterogeneous motif graphs, we incorporate additional information from the mass spectra of motif. This is because the fragmentation patterns in mass spectra are associated with the motif structure. We construct the mass spectra of motifs, taking into account the isotope effect of the molecular ion. Additionally, we incorporate a few fragments generated from RDKit software into the motif mass spectra.

3.6 Objective Function

It is common to use cosine similarity to compare mass spectra after normalizing spectrum to make those invariant to scaling. So we choose cosine distance as loss function as Eq. 3.4, where $\hat{y}$ is the predicted spectrum and y is the target spectrum.

$$CD(y, \hat{y}) = 1 - \frac{y^T \hat{y}}{\|y\|_2 \|\hat{y}\|_2} \tag{3.4}$$

Chapter 4

Experiments and Discussion

4.1 Performance

4.1.1 Evaluation metrics

Spectrum similarity is calculated as cosine similarity score between target and predicted spectrum after normalization.

$$\text{Similarity}(\mathbf{I}, \hat{\mathbf{I}}) = \frac{\sum_{k=1}^{M_{max}} I_k \cdot \hat{I}_k}{\| \sum_{k=1}^{M_{max}} {I_k}^2 \| \cdot \| \sum_{k=1}^{M_{max}} \hat{I}_k^2 \|} \tag{4.1}$$

Here, $\mathbf{I}$ and $\hat{\mathbf{I}}$ are vectors of intensities versus m/z for reference and predicted spectrum.

4.1.2 Results

Each result has been obtained by conducting the experiments five times, with different random seeds for each run. The results for the NIST dataset are presented in table 4.1. Our proposed model, MoMS-Net, demonstrates the best performance compared to other models. Specifically, Massformer outperforms CNN, WLN, and GCN models. Furthermore, we observe that the performance on the FT-HCD dataset is higher compared to the FT-CID dataset. This can be attributed to the larger amount of data available in the FT-HCD dataset. It is commonly known that transformer-based models

can achieve better performance when trained on larger datasets. However, it is noteworthy that MoMS-Net surpasses the performance of Massformer even in the larger FT-HCD dataset.

Table 4.1: Cosine Similarity

	FT-CID	FT-HCD
CNN	0.356 ± 0.002	0.535 ± 0.002
Massformer	0.385 ± 0.005	0.573 ± 0.003
WLN	0.357 ± 0.001	0.569 ± 0.001
GCN	0.356 ± 0.001	0.565 ± 0.001
MoMS-Net	$\mathbf{0.389 \pm 0.001}$	$\mathbf{0.578 \pm 0.001}$

4.1.3 Molecule Identification

To address the coverage issue in spectral library searches, predicting mass spectra is a essential step to augment the existing database. By predicting mass spectra, we can expand the range of compounds and their corresponding spectra available in the spectral library. However, assessing the accuracy of a model in matching predicted spectra with unknown queries is challenging because confirming the identification of the compound requires experimental analysis. To simplify the evaluation process, we can employ a candidate ranking experiment inspired by [14, 17]. In this experiment, the objective is to accurately associate a query spectrum with the corresponding molecule from a set of candidate spectra. The query set comprises authentic spectra from the test set, which are heldout partitions. The reference set consists of spectra collected from distinct origins: predicted spectra in the heldout partition, and real spectra from the training and validation partitions. By evaluating the similarity between spectra in the query and reference sets, we calculate a ranking of spectra in the reference set for each query. This ranking, based on the degree of similarity, effectively induces a ranking of

candidate structures since each spectrum corresponds to a specific molecule.

Table 4.2 provides a summary of the results obtained from this experiment on the metric, Top-5%. This metric evaluates whether the true matched candidate is ranked within the top 5% of all candidates. As the number of candidates per query may vary, the Top-5% metric is normalized to ensure fair comparison. This metric provides insight into the model's ability to accurately identify the correct candidate among a larger set of options. The results indicate that our model demonstrates comparable performance with MassFormer and higher than other models. This consistent strong performance of our model suggests that it is one of the best performing models in terms of accurately matching query spectra with the correct molecule. Our model can be utilized for augmenting spectral libraries holds promise to address the coverage issue.

Table 4.2: Top-5% scores on the ranking task

	FT-CID	FT-HCD
CNN	0.802 ± 0.008	0.778 ± 0.004
MassFormer	$\mathbf{0.850} \pm 0.016$	0.830 ± 0.007
WLN	0.736 ± 0.011	0.812 ± 0.008
GCN	$0.728 \pm 0.0.016$	0.802 ± 0.008
MoMS-Net	0.824 ± 0.002	$\mathbf{0.840} \pm \mathbf{0.010}$

4.1.4 Analysis of Predicted Mass Spectra

Our model demonstrates the capability to accurately predict mass spectra for complex molecules, as shown in Fig. 4.1. Molecules containing conjugated aromatic rings are known to be highly stable, resulting in a smaller number of peaks in their mass spectra. On the other hand, molecules without aromatic rings tend to exhibit a greater number of peaks. Our model is effective in predicting both aromatic compounds and other cyclic compounds accurately. However, it should be noted that there is a restriction in

terms of the intensities of the main peaks in the predicted mass spectra. Our model tends to generate more smaller peaks, which can result in a reduction in the intensity of the main peak after normalization.

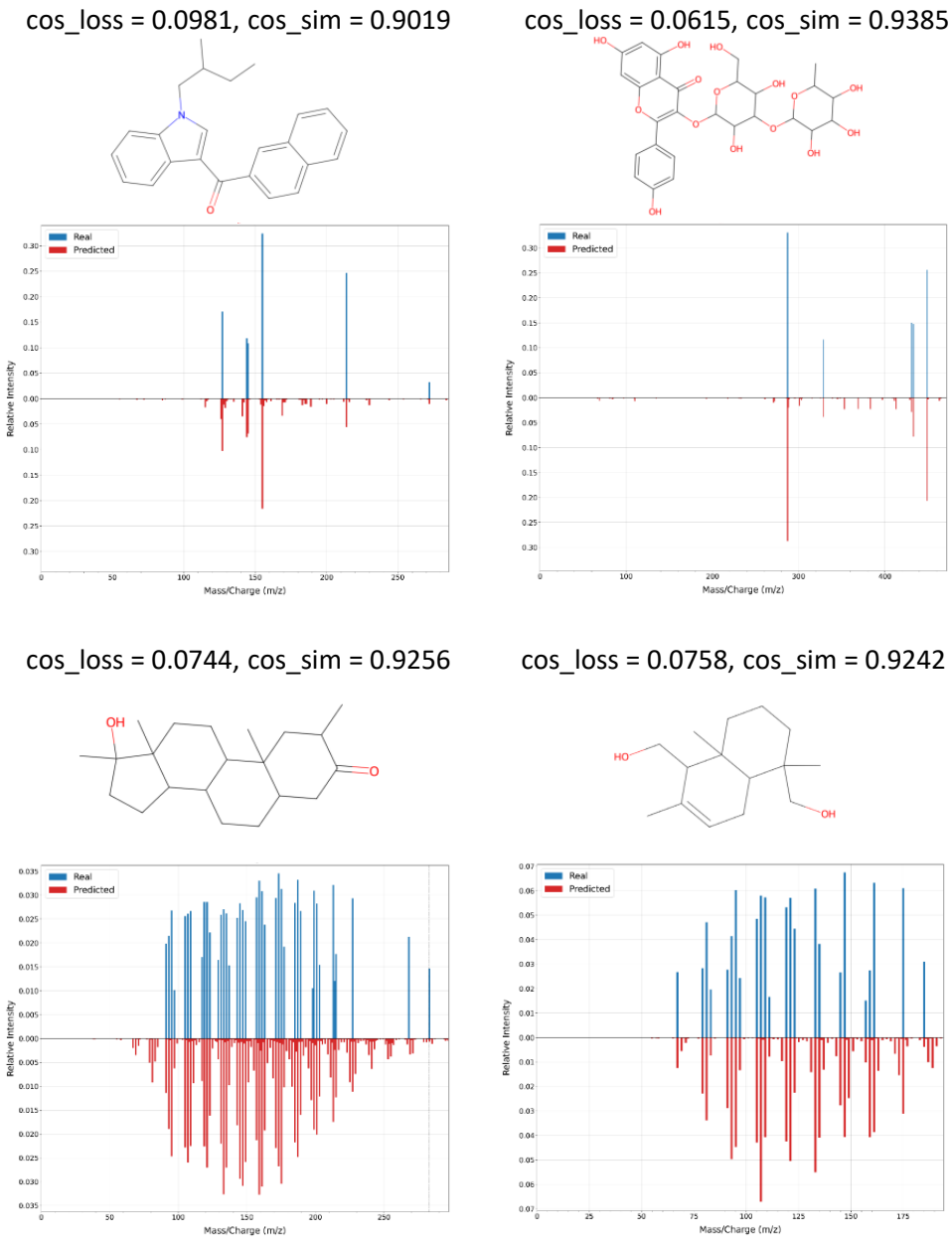


Figure 4.1: Real and predicted spectra for four molecules. Predicted spectra have similar patterns for aromatic and cyclic molecules but have lower intensity because of many smaller false peaks.

4.1.5 Generation of Motif

We employed the byte-pair encoding method to generate subgraphs from the dataset. The K most frequent subgraphs were selected as motifs, with some motifs occurring more than 10,000 times. The most frequent motif in our dataset was "CC" with a frequency of 109,000. The frequency count decreases exponentially as the number of motifs increases, as shown in Fig. 4.3. Unlike rule-based generation methods, our approach allows for the generation of various types and sizes of motifs. In Fig. 4.4, we provide examples of molecules that are large and contain various functional groups. We conducted tests with different sizes of motif vocabularies, as shown in Fig. 4.2. As the motif size exceeds 1,000, the cosine similarity begins to decrease. This decrease is attributed to the consideration of trivial motifs in the heterogeneous motif graph as the motif size increases. Therefore, in this study, we set the size of the motif vocabulary to 300.

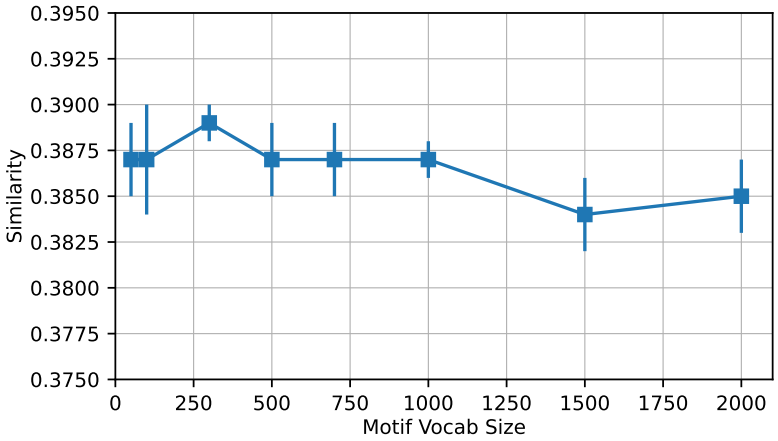
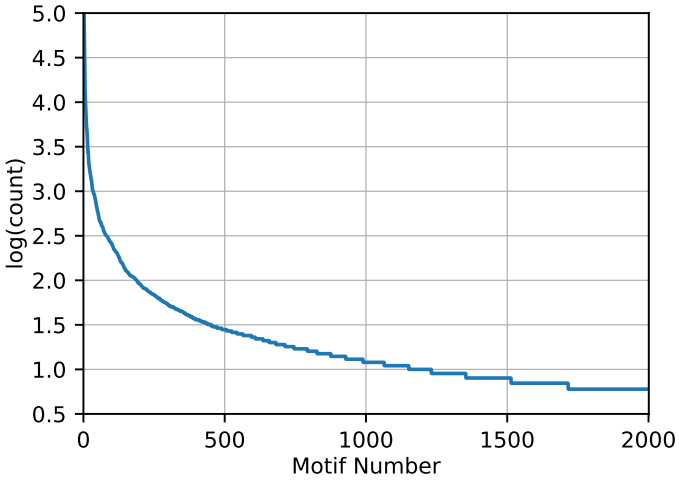
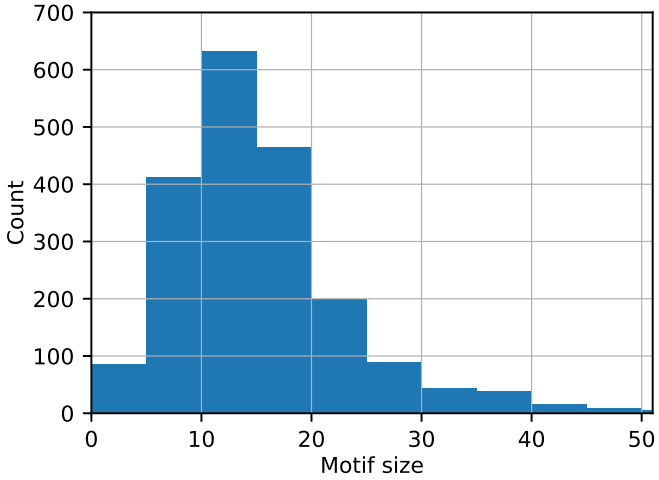


Figure 4.2: Cosine Similarity according to Motif Size. The model achieves its best performance when the motif size is set to 300. However, as the motif size surpasses 1000, the performance starts to decline.



(a)



(b)

Figure 4.3: Frequency of generated motifs (a) Motif number - count (b) Motif size - count. The frequency of motif is decreased exponentially as motif number and most motif has size of 5 to 20 atoms.

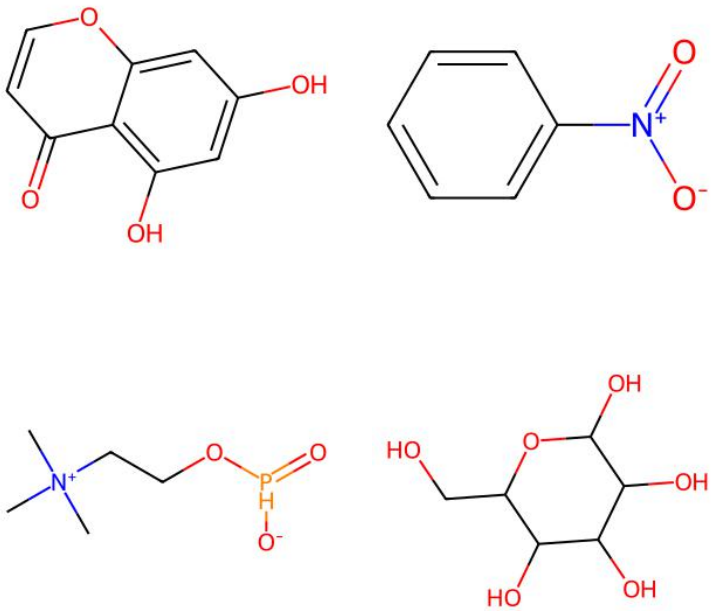


Figure 4.4: Example of large motifs. Data-driven motif generation method can generate large motifs which have various functional group such as aromatic ring, cycle, hydroxy group and ketone.

4.2 Ablation studies

4.2.1 Concatenation of Hidden Representations

The embedding vector of MoMS-Net is obtained by concatenating the hidden embeddings of the molecule GNN and the heterogeneous motif graph (HM-Graph). The concatenation is performed with a specific ratio, denoted as α.

$$e_{\mathrm{MoMS-Net}} = [\ (1-\alpha) * e_{\mathrm{molecule}} || \alpha * e_{\mathrm{HM-Graph}}\ ] \tag{4.2}$$

Through experimentation, we observed that ratios ranging from 0.1 to 0.8 exhibit similar performance in terms of cosine similarity. However, when α reaches 0.9, the performance decreases. This suggests that the structural information of molecules captured by the molecule GNN is more crucial for predicting mass spectra. Nevertheless, considering the relationships among molecules in the heterogeneous motif graph is still beneficial for accurate mass spectrum prediction.

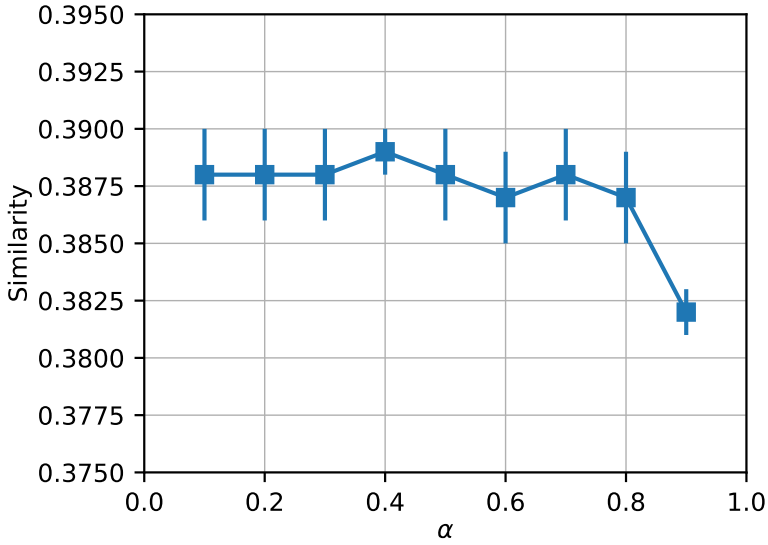


Figure 4.5: The ratio of heterogeneous motif graph to molecule graph. We tested five times for each condition. When α is less than 0.8, the similarity is similar, but decreased as α becomes 0.9.

4.2.2 Addition Method for Mass Spectrum of Motif

We incorporate the information of motif spectrum into the molecular representation using different methods. The motif spectrum embedding is created by applying a fully-connected layer to the generated motif spectrum. We explore different approaches for combining the motif spectrum embedding with the molecular representation. First, we consider adding the motif spectrum embedding to the molecular representation through summation or concatenation. Then, we apply a fully-connected layer with or without a residual connection to refine the combined representation. Among these variations, we find that the summation with a residual connection achieves the best performance.

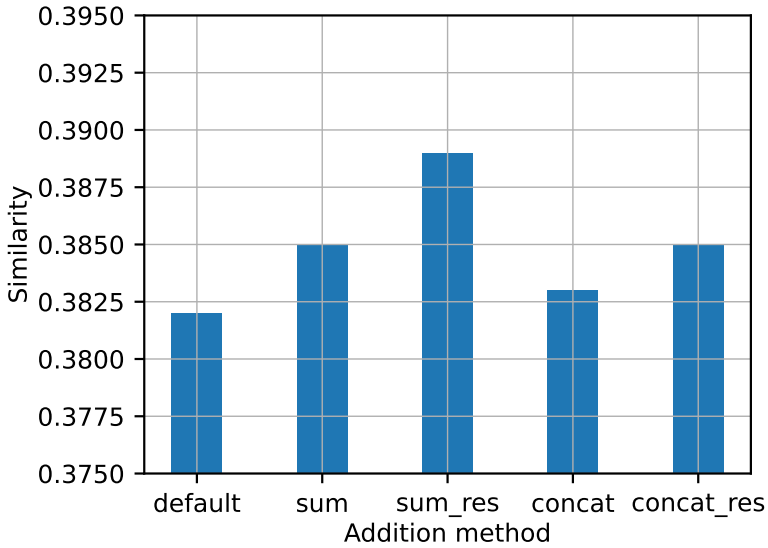


Figure 4.6: Various addition method for mass spectrum of motif. It shows the best performance when it uses summation with residual connection.

4.2.3 Mass Spectrum Generation for Motif

The Mass Spectrum for motif is generated by weight distribution of molecular ion considering isotope effect, and we utilize the RDKit package to add a few fragment. But some motifs are existing chemicals and have their own mass spectra, which are available in the MoNA dataset. Specifically, we use EI (electron impact) Mass Spectrum, as it does not contain adduct ions and provides a reasonable understanding of fragmentation mechanism from the molecular ion. Out of the 300 motifs in our vocabulary, there are 148 mass spectra. Comparing this method to the previous approach, incorporating the real mass spectra of motifs leads to improved performance, as shown in Table 4.3. We are unable to utilize this method for all motifs due to the unavailability of mass spectra. However, the results demonstrate that having precise information on mass spectra is beneficial for predicting mass spectra for molecules.

Table 4.3: Cosine Similarity according to Motif Spectrum

Motif MS	FT-CID
M.W.	0.388 ± 0.002
MoNA	$\mathbf{0.392 \pm 0.001}$

4.2.4 Model Parameters and Memory Allocation

Table 4.4 shows the information of the number of model parameters and memory allocation. We can see that all models have similar numbers of model parameters. However, we were unable to test batch size of 1024 for Massformer due to memory limitation. It should be noted that Massformer takes a large amount of memory. As a result, despite having a smaller batch size, Massformer requires a similar amount of memory allocation compared to MoMS-Net. We can see that MoMS-Net show better performance with less memory compared to Massformer.

Table 4.4: Number of Parameters and Memory Allocation

	# of Parameters	Memory Allocation(MB)	Batch Size
CNN	1.46E+07	717	512
Massformer	1.36E+07	1340	50
WLN	1.23E+07	1519	1024
GCN	1.31E+0.7	973	1024
MoMS-Net	1.82E+07	1519	1024

4.2.5 GNNs Architecture

Our model consists of GCN for molecule graphs and GIN for heterogeneous motif graphs. As shown in Table 4.5, we can see that GCN performs better than GIN. However, when the MoMS-Net model uses GIN instead of GCN, it shows similar performance.

Table 4.5: Cosine Similarity according to GNN Architecture

	FT-CID	FT-HCD
GIN	0.352 ± 0.004	0.558 ± 0.001
GCN	0.356 ± 0.002	0.565 ± 0.001
MoMS-Net(GIN)	0.389 ± 0.002	0.575 ± 0.001
MoMS-Net(GCN)	0.388 ± 0.002	0.578 ± 0.001

4.2.6 Hidden Dimension Size

Heterogeneous motif graph has $|N + V|$ nodes and node input is represented as concatenation of connectivity with motifs in vocabulary as Eq. 3.1 and molecular weight. Before concatenation, 2-layer fully-connected layer is applied for input features of

connectivity as it is very sparse. We compared different size of hidden dimensions as Fig. 4.7. The dimension size of the hidden representations does not seem to have a crucial role in determining the results, as the performance is similar regardless of the hidden dimension size.

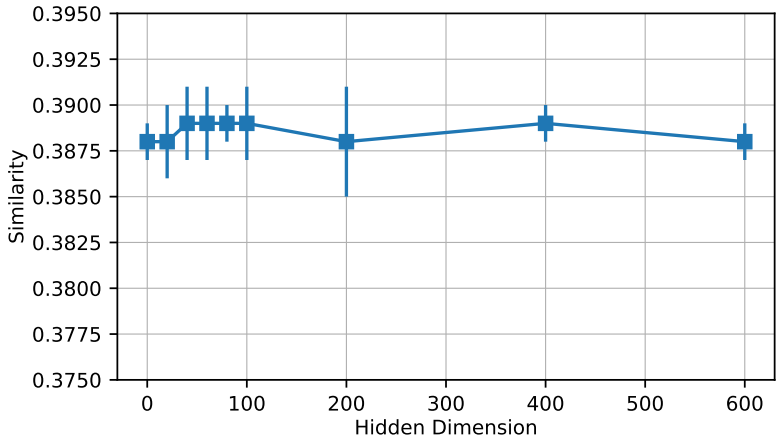


Figure 4.7: The dimension size of the hidden representation. A 2-layer fully connected layer is applied to the input features of connectivity in order to transform it into various dimension sizes. The similarity remains similar regardless of the hidden dimension size.

4.2.7 Loss Function

We tested various loss functions for training. Weighted cosine similarity (w_cos) loss is a similarity measure that takes into account the relative intensities for each peaks as weight. Mean squared error (MSE) loss quantifies the average of the squared differences between the predicted values and its corresponding real values. The Kullback-Leibler (KL) divergence loss a measure of dissimilarity between two probability distributions. The Jensen-Shannon (JS) loss is calculated by taking the average of the Kullback-Leibler (KL) divergences between two probability distributions (P, Q), and the KL divergence between Q and P. Wasserstein (WS) loss, also known as Earth

Mover's Distance (EMD) loss, is a loss function, which measures the dissimilarity between two probability distributions. As shown in Table 4.6, cosine similarity loss shows the best performance.

Table 4.6: Cosine Similarity according to Loss Function

	FT-CID
cos	$\mathbf{0.388 \pm 0.001}$
w_cos	0.378 ± 0.002
MSE	0.343 ± 0.001
KL	0.374 ± 0.013
JS	0.348 ± 0.003
wass	0.213 ± 0.004

4.2.8 Generalization

We tested prediction tasks using different ratio of training set. CNN did not perform well with the usual setting of the split ratio of 7/2/1 for train/valid/test sets. However, its performance did not decrease significantly as the ratio of the training size decreased. CNN demonstrates good generalizability due to its ability to capture patterns with receptive filters. It performs better by effectively capturing patterns even with a smaller training size. On the other hand, Massformer showed better performance compared to GNN models like MoMS-Net, WLN, and GCN, particularly with smaller ratios of the training set.

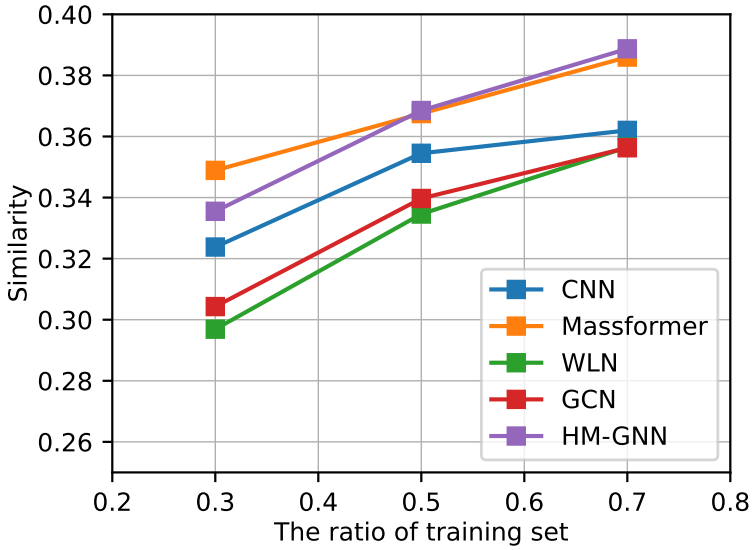


Figure 4.8: Generalizability. We tested various models with different split ratio. CNN showed poor performance at a training size of 0.7 but the decrease in similarity was small as the ratio decrease.

Chapter 5

Conclusion

The analysis of mass spectra plays a crucial role in identifying molecular structures in material chemistry and drug discovery. Search-based methods are widely employed for mass spectra analysis. However, They often suffer from a coverage issue. To address this problem, it is necessary to generate mass spectra using a model to augment the database. Numerous deep learning models have been employed for mass spectra prediction. Graph Neural Networks (GNNs) are particularly useful for predicting molecular properties since molecules can be represented as graphs. However, GNNs have limitations in considering long-range dependencies, thereby affecting their performance. The graph transformer has been reported to exhibit excellent performance in predicting mass spectra. However, it consumes excessive memory during training.

In this study, we proposed the MoMS-Net model, which incorporates motifs to predict mass spectra from molecular structures. Motifs play a crucial role in the molecular property prediction task as they are related to functional groups in the molecule and provide valuable information on the relationships between molecules. We applied the byte-pair encoding method to generate a motif vocabulary from the dataset. We constructed a heterogeneous motif graph consisting of molecules and motifs as nodes, with edges being formed if a molecule has a motif or if two motifs share any atoms. The MoMS-Net model consists of two GNNs, one for the molecule graph and the other

for the heterogeneous motif graph. We conducted tests with different sizes of motif vocabularies and varying model architectures.

MoMS-Net outperforms other deep learning models in predicting mass spectra from molecular structures. It effectively considers long-range dependencies by incorporating motifs at the graph level. Additionally, our model requires less memory compared to the graph transformer. We found that real mass spectra of motifs are useful in predicting the mass spectra of molecules, although the predicted mass spectra may contain more small and false peaks. In future work, we aim to improve the initialization method of mass spectra for motifs and incorporate regularization techniques to prevent false peaks. Furthermore, we plan to apply MoMS-Net to larger molecules and proteins.

Bibliography

[1] Gary L. Glish and Richard W. Vachet. The basics of mass spectrometry in the twenty-first century. *Nat Rev Drug Discov*, 2:140–150, 2003.

[2] A. T. Lebedev. Environmental mass spectrometry. *Annual Review of Analytical Chemistry*, 6(1):163–189, 2013.

[3] R. Aebersold and M. Mann. Mass-spectrometric exploration of proteome structure and function,. *Nature*, 537(7620):347–355, 2016.

[4] G. N. Gowda and D. Djukovic. Overview of mass spectrometry-based metabolomics: Opportunities and challenges,. *Methods in molecular biology (Clifton, N.J.)*, 1198:3–12, 2014.

[5] T. De Vijlder, D. Valkenborg, F. Lemiere, E. P. Romijn, K. Laukens, and F. Cuyckens. A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation,. *Mass Spectrometry Reviews*, 37(5):607–629, 2018.

[6] S. E. Stein. Chemical substructure identification by mass spectral library searching. *J. Am. Soc. Mass Spectrom.*, 6:644–655, 1995.

[7] Stein, S. E. and Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.*, 5:859–866, 1994.

[8] Stein, S. E. Mass spectral database; national institute of standards and technology (nist). 2017.

[9] F. W. Mclafferty. Wiley registry of mass spectral data, 11th ed.; john wiley and sons. 2016.

[10] Bauer, C. A. and Grimme, S. How to compute electron ionization mass spectra from first principles. *J. Phys. Chem. A*, 120:3755–3766, 2016.

[11] S. Grimme. Towards first principles calculation of electron impact mass spectra of molecules. *Angew. Chem., Int. Ed.*, 52:6306–6312, 2013.

[12] Guerra, M., Parente, F., Indelicato, P., and Santos, J. P. Modified binary encounter bethe model for electron-impact ionization. *Int. J. Mass Spectrom.*, 313:1–7, 2012.

[13] Allen, F., Pon, A., Greiner, R., and Wishart, D. Computational prediction of electron ionization mass spectra to assist in gc/ms compound identification. *Anal. Chem.*, 88:7689–7697, 2016.

[14] Jennifer N. Wei, David Belanger, Ryan P. Adams, and D. Sculley. Rapid prediction of electronionization mass spectrometry using neural networks. *ACS Cent. Sci.*, 5:700–708, 2019.

[15] Kaiyuan Liu, Sujun Li, Lei Wang, Yuzhen Ye, and Haixu Tang. Full-spectrum prediction of peptides tandem mass spectra using deep neural network. *Anal. Chem.*, 92:4275–4283, 2020.

[16] Baojie Zhang, Jun Zhang, Yi Xia, Peng Chen, and Bing Wang. Prediction of electron ionization mass spectra based on graph convolutional networks. *International Journal of Mass Spectrometry,*, 475(116817), 2022.

[17] Adamo Young, Bo Wang, and Hannes Rost. Massformer: Tandem mass spectrum prediction with graph transformers. *arXiv*, 2021.

[18] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[19] Hongchao Ji, Hanzi Deng, Hongmei Lu, and Zhimin Zhang. Predicting a molecular fingerprint from an electron ionization mass spectrum with deep neural networks. *Anal. Chem.*, 92:8649–8653, 2020.

[20] Eng, J. K., McCormack, A. L., and Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, 5:976–989, 1994.

[21] Tran, N. H., Zhang, X., Xin, L., Shan, B., and Li, M. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. U. S. A.*, 114:8247–8252, 2017.

[22] Duhrkop, K., Shen, H., Meusel, M., Rousu, J., and Bocker, S. Searching molecular structure databases with tandem mass spectra using csi: Fingerid. *Proc. Natl. Acad. Sci. U. S. A.*, 112:12580–12585, 2015.

[23] Zhaoning Yu an Hongyang Gao. Molecular representation learning via heterogeneous motif graph neural networks. *ICML*, 2022.

[24] Muhan Zhang and Pan Li. Nested graph neural networks. *NeurIPS*, 2021.

[25] Giorgos Bouritsas, Fabrizio Frasca, Stefanos Zafeiriou, and Michael M. Bronstein. Improving graph neural network expressivity via subgraph isomorphism counting. *ICLR*, 2022.

[26] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. *ICML*, 2022.

[27] Dexion Chen, Leslie O'Bray, and Karsten Borgwardt. Structure-aware transformer for graph representation learning. *ICML*, 2022.

[28] Jiahua Rao, Shuangjia Zheng, Sijie Mai, and Yuedong Yang. Communicative subgraph representation learning for multi-relational inductive drug-gene interation prediction. *IJCAI*, 2022.

[29] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem: Chemistry Enabling Drug Discovery*, 3(10):1503–1507, 2013.

[30] Tairan Liu, Misagh Naderi, Chris Alvin, Supratik Mukhopadhyay, and Michal Brylinski. Break down in order to build up: Decomposing small molecules for fragment-based drug design with emolfrag. *J. Chem. Inf. Model*, 57:627–631, 2017.

[31] Ivanov, Nikita N., Shulga, Dmitry A., and Palyulin, Vladimir A. Decomposition of small molecules for fragment-based drug design. *Biophysica*, 3(2):362–372, 2023.

[32] Philip Gage. A new algorithm for data compression. *C Users Journal*, 12:23–38, 1994.

[33] Zijie Geng, Shufang Xie, Yingce Xia, Lijun Wu, Tao Qin, Jie Wang, Yongdong Zhang, FengWu, and Tie-Yan Liu. De novo molecular generation via connection-aware motif mining. *ICLR*, 2023.

[34] Xiangzhe Kong, Wenbing Huang, Zhixing Tan, and Yang Liu. Molecule generation by principal subgraph mining and assembling. *NeurIPS*, 2022.

[35] Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. *NeurIPS*, 2021.

[36] Shichang Zhang, Ziniu Hu, Arjun Subramonian, and Yizhou Sun. Motif-driven contrastive learning of graph representations. *AAAI*, 2020.

[37] Duhrkop, K., Shen, H., Meusel, M., Rousu, J., and Bocker, S. Searching molecular structure databases with tandem mass spectra using csi: Fingerid. *Proc. Natl. Acad. Sci. U. S. A.*, 112(12580-12585), 2015.

[38] H. Zhu, L. Liu, and S. Hassoun. Using graph neural networks for mass spectrometry prediction. *arXiv*, 2020.

초 록

질량 분석학은 재료 화학과 약물 합성 분야에서 분자 구조를 식별하는 데 중요한 역할을 한다. 검색 기반 방법은 일반적으로 질량 스펙트럼 분석에 널리 사용되지만, 가용 데이터의 부족으로 인한 한계가 있다. 이 문제를 해결하기 위해서는 모델을 사용하여 질량 스펙트럼을 생성하여 데이터베이스를 보강해야 할 필요가 있다. 다양한 딥러닝 모델이 질량 스펙트럼 예측에 사용되고 있다. 그래프 신경망(GNN)은 분자를 그래프로 표현할 수 있어 분자 속성 예측에 유용하다. 그러나 GNN은 장거리 의존성을 고려하는 데 한계가 있어 성능이 저하되게 된다. 그래프 트랜스포머는 질량 스펙트럼 예측에서 우수한 성능을 나타내지만 훈련 중에 과도한 메모리를 소비하게 된다.

본 연구에서는 분자 구조로부터 질량 스펙트럼을 예측하기 위해 구조 모티프를 포함하는 MoMS-Net 모델을 제안하였다. 모티프는 분자 내의 기능성 그룹과 관련이 있으며 분자 간의 관계에 대한 의미 있는 정보를 제공하여 분자 속성 예측 과제에서 중요한 역할을 한다. 우리는 데이터셋으로부터 모티프 집합을 생성하기 위해 병합 방법을 적용하였다. 분자가 모티프를 가지고 있거나 두 모티프가 어떤 원자를 공유하는 경우에는 연결성을 갖게되도록 분자와 모티프로 구성된 이종 모티프 그래프를 구성하였다. MoMS-Net 모델은 분자 그래프와 이종 모티프 그래프 각각에 대한 두 개의 GNN으로 구성된다. 우리는 다양한 크기의 모티프 집합과 다양한 모델 구조로 실험을 진행하였다. MoMS-Net은 분자 구조로부터 질량 스펙트럼을 예측하는 데 있어 다른 딥러닝 모델보다 우수한 성능을 발휘하였다. 그래프 수준에서 모티프를 정보를 활용함으로써 장거리 의존성을 효과적으로 고려하였다. 게다가, 우리의 모델은 graph transformer에 비해 더 적은 메모리를 요구하였다. 우리는 모티프의 실

제 질량 스펙트럼이 분자의 질량 스펙트럼 예측에 효과가 있다는 것을 발견하였다. 그러나, 예측된 질량 스펙트럼에는 더 작고 잘못된 피크가 많이 포함되어 있었다. 향후 연구에서는 모티프에 대한 질량 스펙트럼의 초기화 방법을 개선하고 잘못된 피크를 방지하기 위해 정규화 기법을 도입할 계획이다. 또한, MoMS-Net을 더 큰 분자와 단백질에 적용할 예정이다.

주요어: 질량 스펙트럼, 그래프 신경망, 모티프

학번: 2021-25101