UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Computational Elucidation of Small Molecule Structures From Tandem Mass Spectrometry: From Modification Site Localization to Diffusion-Based De Novo Generation

Permalink

https://escholarship.org/uc/item/57j8j38k

ISBN

9798263308803

Author

Zare Shahneh, Mohammad Reza

Publication Date

2025-07-28

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-ShareAlike License, available at https://creativecommons.org/licenses/by-nc-sa/4.0/

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Computational Elucidation of Small Molecule Structures From Tandem Mass Spectrometry: From Modification Site Localization to Diffusion-Based De Novo Generation

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

by

Mohammad Reza Zare Shahneh

September 2025

Dissertation Committee:

Dr. Mingxun Wang, Chairperson
Dr. Stefano Lonardi
Dr. Tao Jiang
Dr. Daniel Petras

Copyright by
Mohammad Reza Zare Shahneh
2025

The Dissertation of Mohammad Reza Zare Shahneh is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to express my deepest gratitude to my advisor, Mingxun Wang, whose mentorship extended far beyond academic guidance. His trust, insightful direction, and personal support have profoundly affected my career and personality.

I am deeply grateful to my family, my parents, Moein, and Sanaz, whose unwavering love and faith in me have been a constant source of strength. Their support never wavered, and their belief in me gave me the courage to take risks and push forward without fear of failure. Without their encouragement, this journey would not have been possible.

I also wish to acknowledge my close friends Mohammadreza, Emad, Hesam, Mehrdad, Bahar, Mohsen, and Zakieh. Their presence and support during my most difficult moments meant more than words can convey.

I extend my thanks to my labmates Xianghu, Gwen, Youray, and Michael for the collaborative and stimulating environment they helped foster.

I am also thankful to my co-authors, Michael Strobel, Giovanni Andrea Vitale, Christian Geibel, Yasin El Abiead, Allegra Aron, Vanessa V Phelan, Daniel Petras, Amr Magdy, Samet Oymak and others, whose insights and contributions were integral to this work.

Lastly, I would like to acknowledge two of my early teachers, Mr. Pishnamaz and Mr. Forooghmand. Though they may not know it, their influence played a critical role in shaping the direction of my academic and professional life.

To my parents, Saeedeh and Mahdi.

ABSTRACT OF THE DISSERTATION

Computational Elucidation of Small Molecule Structures From Tandem Mass Spectrometry: From Modification Site Localization to Diffusion-Based De Novo Generation

by

Mohammad Reza Zare Shahneh

Doctor of Philosophy, Graduate Program in Computer Science
University of California, Riverside, September 2025
Dr. Mingxun Wang, Chairperson

The identification of unknown small molecules from tandem mass spectrometry (MS/MS) data is a fundamental challenge in the chemical and biological sciences, as a significant portion of spectra from untargeted experiments remain unidentified. While computational techniques like analog search can suggest structurally similar compounds, they fail to specify the exact location of chemical modifications, leaving a critical gap in the annotation process. To surmount this issue, this dissertation investigates the problem of structural annotation of small molecules using tandem mass spectrometry (MS/MS) data, with a particular focus on leveraging information from known similar molecules known as analogs and presents a cohesive suite of computational tools developed to address this challenge, progressing from targeted localization to full structure generation.

First, ModiFinder is introduced, a novel algorithm that localizes structural modification sites by aligning the MS/MS spectra of a known molecule and its unknown analog. By systematically analyzing shifted and unshifted fragment peaks, ModiFinder generates a likelihood score for each atom in the known structure being the modification site, successfully

outperforming baseline methods. Then, the robustness and applicability of this approach are significantly enhanced by incorporating richer data from multiple collision energies and mass spectrometry adducts. This multispectrum approach expands ModiFinder's utility to a wider range of compounds by providing complementary fragmentation information.

Finally, the more ambitious challenge of de novo structure elucidation is addressed by developing ModiStruct, a state-of-the-art generative model. This model reframes the problem as a diffusion-driven molecular inpainting task, where it reconstructs the complete molecular graph of an unknown compound conditioned on a known substructure, its molecular formula, and their respective MS/MS spectra. ModiStruct achieves state-of-the-art accuracy on public benchmarks, validating the power of using structural priors within a generative framework. Collectively, this work provides a powerful set of tools that advance the automated identification of small molecules, from pinpointing specific chemical changes to generating entire novel structures. These contributions help to bridge the gap between MS/MS data and chemical knowledge, with broad applications in metabolomics, drug discovery, and natural products research.

Contents

List of Figures **xi**

List of Tables **xiii**

1 Introduction **1**
1.1 The Challenge of Small Molecule Identification . . . 1
1.2 Tandem Mass Spectrometry as an Analytical Tool . . . 2
1.3 Related Work . . . 3
1.3.1 Spectral Library Matching and Computational Interpretation . . . 3
1.3.2 Forward Spectrum Prediction . . . 4
1.3.3 *De Novo* Structure Elucidation . . . 5
1.4 Research Questions and Contributions . . . 7
1.5 Notation & Definitions . . . 9
1.6 Dissertation Outline . . . 11

2 ModiFinder – Algorithm & Benchmarking **12**
2.1 Motivation . . . 12
2.2 ModiFinder Algorithm . . . 15
2.2.1 MS/MS Alignment . . . 17
2.2.2 Combinatorial Fragmentation . . . 18
2.2.3 Substructure Refinement by Formula . . . 19
2.2.4 Refinement by Helpers . . . 20
2.2.5 Calculating the Site Localization . . . 22
2.2.6 MS/MS Spectral Library Data Preparation . . . 24
2.2.7 LC-MS/MS Data Acquisition Methods . . . 26
2.3 Experimental Evaluation . . . 27
2.3.1 Evaluation Score . . . 27
2.3.2 Site Localization baselines and alternative approaches . . . 30
2.3.3 Performance . . . 31
2.4 Analysis of Key Performance Factors . . . 35
2.4.1 Shifted Peaks Matter in Site Localization . . . 35
2.4.2 Increasing the Fragmentation Depth Increases Ambiguity . . . 37

2.4.3 Improving the Quality of Annotation and reducing the Ambiguity increases the Evaluation Score . . . 39
2.4.4 additional factors . . . 40
2.5 Case Studies and User Interaction . . . 41

3 Enhancing Site Localization with Multispectrum Data – Exploiting Collision Energies & Adducts 48
3.1 Motivation . . . 48
3.1.1 Methods for Data Integration . . . 49
3.1.2 Data Acquisition Methods . . . 52
3.2 Strategies for Data Integration . . . 53
3.2.1 Merged Peaks . . . 53
3.2.2 Ensemble Learning, Merging the Predictions . . . 54
3.3 Experimental Evaluation . . . 54
3.3.1 Effect of collision energy on ModiFinder performance . . . 54
3.3.2 Effect of Merging the multiple collision energies . . . 55
3.3.3 Ensemble Learning by Merging the Predictions . . . 56
3.3.4 Polarity Adducts Effect on ModiFinder Performance . . . 58
3.4 Performance Across Different Chemical Classes . . . 59
3.5 Conclusion . . . 59

4 ModiStruct – Diffusion-Driven Molecular Inpainting with Substructure Constraints 61
4.1 Motivation . . . 61
4.2 Discrete Denoising Diffusion for Graphs . . . 64
4.3 The ModiStruct Architecture . . . 64
4.3.1 Problem statement . . . 65
4.3.2 Overview of ModiStruct . . . 65
4.3.3 Encoder . . . 67
4.3.4 Discrete graph diffusion decoder . . . 68
4.3.5 Pre-training and fine-tuning . . . 70
4.3.6 Inference and sampling with inpainting . . . 70
4.3.7 Oracle . . . 71
4.4 Experimental Evaluation . . . 72
4.4.1 Baselines . . . 72
4.4.2 Overall De Novo Structure–Prediction Accuracy . . . 72
4.4.3 ModiStruct Modification-Site Localization . . . 73
4.4.4 Fingerprint Prediction analysis . . . 74

5 Conclusions and Discussions 76
5.1 Summary of Contributions . . . 76
5.2 Synthesis: The Power of Structural Priors in Mass Spectrometry . . . 78
5.3 Limitations and Future Directions . . . 78
5.3.1 Improving Substructure Retrieval and Verification . . . 79
5.3.2 Handling More Complex Molecular Scenarios . . . 79

5.3.3 Advancing Generative Modeling Techniques 79
5.3.4 The Critical Need for High-Quality Data 80
5.3.5 Using the network 80
5.4 Concluding Remarks 80

Bibliography **83**

Appendix **92**

List of Figures

1.1 MS/MS workflow . . . 3

2.1 Unknown Spectra and its known Analog . . . 13
2.2 Illustration of the intuition behind ModiFinder . . . 14
2.3 The Overview of ModiFinder algorithm . . . 16
2.4 ModiFinder with helpers . . . 19
2.5 Toy example for refinement by helpers . . . 21
2.6 Example of ModiFinder method. . . . 25
2.7 Overview of the CFM-ID alternative approach . . . 32
2.8 Performance Results . . . 34
2.9 Comparative Analysis of Evaluation Scores for each library . . . 36
2.10 Correlation of ambiguity and performance . . . 38
2.11 Illustration of Negative Impact of Unshifted Peaks in ModiFinder . . . 42
2.12 Using ModiFinder to annotate unknown compounds . . . 44
2.13 Combination of domain knowledge and ModiFinder for Kirromycin . . . 46
2.14 Combination of domain knowledge and ModiFinder for Naphthomycin B . . . 47

3.1 Datasets Comparison . . . 50
3.2 Distribution of the selected compounds across different pathways . . . 51
3.3 The distribution of compound masses for each library . . . 52
3.4 ModiFinder Result for Different Collision energies across different datasets . . . 55
3.5 Performance of ModiFinder on each adduct individually and as an ensemble . . . 57
3.6 Distribution of peaks and annotated peaks with respect to Adduct . . . 58

4.1 ModiStruct Encoder . . . 66
4.2 Overview of the decoder . . . 69
4.3 Repaint on adjacency matrix . . . 71

5.1 ModiFinder in a network . . . 81
A-1 Illustration of imbalance symmetry on a toy example . . . 93
A-2 Pairwise comparison of MF-R and MF-O with RC . . . 93
A-3 The Evaluation score across different examples . . . 94
A-4 Key Factors Contributing to Efficacy of ModiFinder . . . 96
A-5 Contributing factors in ModiFinder's performance . . . 97

A-6 Example of the permutation algorithm on Deoxyadenosine and Deoxyadenosine Monophosphate 99
A-7 Example of the permutation algorithm on Tyramine and 3-MethoxyTyramine. 100

List of Tables

4.1 ModiStruct Overall Performance Comparison 73
4.2 ModiStruct ModiFinder Performance Comparison 74
4.3 ModiStruct Encoder Fingerprint Tanimoto Comparison 75

Chapter 1

Introduction

1.1 The Challenge of Small Molecule Identification

The identification of unknown small molecules in complex chemical mixtures is a primary challenge in many areas of chemical and biological science[23]. These molecules, including metabolites and natural products, orchestrate a vast array of biological processes and hold immense potential for applications ranging from drug discovery to environmental toxicology and personalized medicine. Unlocking the structural identity of these compounds is paramount to understanding their function, mechanism of action, and potential for therapeutic intervention. However, the sheer diversity and complexity of the molecular world present a formidable analytical hurdle, demanding high-throughput and accurate methods for structural elucidation.

1.2 Tandem Mass Spectrometry as an Analytical Tool

Mass spectrometry (MS), and in particular, tandem mass spectrometry (MS/MS), is a cornerstone analytical technique used to for this task and has enabled the discovery of many new compounds[49]. The mass spectrometry begins by ionizing the sample molecules to produce charged species known as precursor ions. These ions are then subjected to controlled fragmentation. For each fragment generated, the instrument records its mass-to-charge (m/z) ratio and corresponding signal intensity, which are shown as peaks in the two-dimensional spectrum. Figure1.1-A illustrates this process.

Despite the powerful insights offered by MS/MS data, translating the MS/MS spectra to 2D chemical structures poses a significant challenge in the field[37]. A single experimental run can yield thousands of spectra, yet only a small fraction are successfully annotated with known molecular structures[4]. This challenge stems from both computational and chemical complexities. *(i)* Computationally, each measurement can potentially arise from a vast number of substructures, leading to an exponential growth in candidate structures, a problem that is NP-hard in nature[36]. *(ii)* Chemically, the information in a standalone mass spectrum may be insufficient to pinpoint a unique molecular structure, particularly in cases such as isomers (molecules with the same chemical formula but different atomic arrangements) and isobars (molecules with the same nominal mass but different elemental compositions)[50]. For instance, the amino acids leucine and isoleucine, despite their structural differences, produce virtually indistinguishable MS/MS spectra under common fragmentation conditions. Figure1.1-B illustrates this challenge.

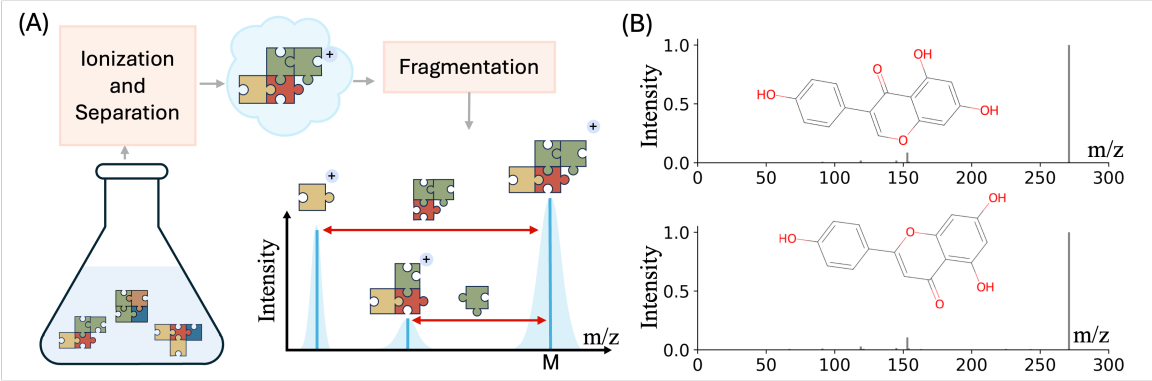


Figure 1.1: (A) The workflow of tandem mass spectrometry. A chemical mixture is ionized and filtered to isolate precursor ions of $m/z = M$; these are then fragmented into product ions and neutral losses, and a detector yields a histogram of product ions indexed by m/z, with measurement error proportional to m/z. On the example, the fragmentation breaks bonds, cutting the molecular graph (shown as the connected puzzle pieces) into connected components. The component retaining the charge is the product ion (shown on top of their respective peak); its complement is the neutral loss (shown on top of the red arrows).
(B) The MS/MS of two compounds can be nearly identical while their structures differ.

1.3 Related Work

1.3.1 Spectral Library Matching and Computational Interpretation

Spectrum library matching is a key strategy within the field of metabolomics to annotate known compounds [57]. A query spectrum is compared against a reference library of spectra from authenticated chemical standards, and a match suggests the identity of the unknown molecule. However, the limited coverage of these libraries presents a major bottleneck. In untargeted mass spectrometry experiments, it is estimated that, on average, 87% of MS/MS spectra remain unidentified by spectral library search[4], a phenomenon often referred to as the "dark matter" of metabolomics[13].

To bridge this gap, modification-aware spectral matching tools such as analog library searching [12] and molecular networking [81, 73] were developed. These methods

leverage the concept of structural propagation, connecting spectra of known compounds to those of unknown but structurally related analogs. They operate on the principle that molecules with a conserved core structure but minor structural modifications will exhibit similar fragmentation patterns[79, 76, 14]. A key shortcoming of these approaches, however, is that while they determine which pairs of MS/MS spectra are putatively similar in structure, they do not explicitly describe the structural difference, leaving manual interpretation to expert chemists.

1.3.2 Forward Spectrum Prediction

The limitations of experimental libraries motivate the development of *in silico* spectrum prediction methods, which computationally generate a tandem mass spectrum from a given molecular structure. These predicted spectra can be used to augment reference libraries, enabling the identification of compounds that have never been experimentally measured.

Early computational methods often relied on rule-based systems derived from known fragmentation mechanisms or combinatorial algorithms that systematically break bonds to generate all possible fragments[69, 46, 48]. A prominent example is Competitive Fragmentation Modeling (CFM), which extends the combinatorial approach by using a probabilistic model to predict fragment intensities[1, 71]. While effective, these bond-breaking methods are computationally expensive, scaling poorly with molecule size, and can struggle to model complex fragmentation events involving bond rearrangements[40]. At the other extreme, first-principles methods based on quantum chemical simulations offer high interpretability but are too computationally intensive for high-throughput applications[83].

Recent advances in deep learning have introduced powerful, data-driven alternatives for forward prediction. These models learn fragmentation patterns directly from large datasets of paired spectra and molecules. A primary challenge has been the representation of the output space, a set of high-precision mass-to-charge (m/z) peaks. One common strategy is *mass-binning*, where the m/z axis is discretized into a fixed-length vector. Models like NEIMS[77] and MassFormer[83] adopt this approach, using feed-forward networks on molecular fingerprints or graph transformers on molecular graphs, respectively. While computationally efficient and scalable, mass-binning sacrifices the high-resolution mass information that is critical for determining a fragment's elemental composition [83].

To address this trade-off, GRAFF-MS introduced an alternative formulation, casting spectrum prediction as a mapping from a molecular graph to a probability distribution over a small, fixed vocabulary of common molecular formulas [40]. By discovering that a vast majority of spectral signal can be explained by a recurring set of fragment formulas, this approach retains full m/z resolution without the costly enumeration of bond-breaking methods, achieving state-of-the-art performance in both accuracy and speed.

ICEBERG[25] tackles this problem in a different way, creating the potential sub-structures and calculating the m/z peaks for these fragments directly from the molecular graph and then using a transformer-based model to predict the intensities of these peaks.

1.3.3 *De Novo* Structure Elucidation

The inverse problem, predicting a molecular structure directly from an MS/MS spectrum, is known as *de novo* structure elucidation. This task is exceptionally challenging due to the inherent ambiguity of the data and the vastness of chemical space. Nevertheless, it

holds the potential to identify entirely novel compounds not present in any existing database.

Many modern methods approach this as a multi-stage process. First, a model predicts an intermediate molecular representation, such as a molecular fingerprint, from the spectrum. Tools like CSI:FingerID [18] and the transformer-based MIST[26] excel at this spectrum-to-fingerprint task. Subsequently, a second generative model attempts to reconstruct the full molecular structure from this predicted fingerprint. For example, MSNovelist uses a recurrent neural network (RNN) to generate a SMILES string from a predicted fingerprint and an inferred molecular formula [59].

More recent efforts have focused on end-to-end models that generate structures directly from spectra. Several approaches frame the problem as a sequence-to-sequence translation task, using transformer-based language models like Spec2Mol[34] and MS2Mol[10] to convert tokenized spectral data into a SMILES string. While powerful, these autoregressive models do not inherently capture the permutation-invariant nature of mass spectra and molecules, nor do they strictly enforce chemical formula constraints, which can be accurately inferred from high-resolution mass data [6].

To overcome these limitations, generative models have emerged as a new state-of-the-art. These methods largely follow two architectural trends: sequence-to-sequence language models and, more recently, diffusion models[56, 28, 55].

Inspired by the success of the Transformer in natural language processing[67, 15], models like **MassGenie** and **TeFT** frame the problem as a l̈anguage translationẗask, converting a sequence of mass spectral peaks into a SMILES string[54, 82]. A key innovation in this approach is training on massive datasets of *in silico* generated fragment-SMILES

pairs, which effectively circumvents the bottleneck of limited experimental data.

Diffusion models have shown significant promise due to their ability to be conditioned on prior knowledge or partial structures, offering greater control over the generation process. One powerful strategy is **scaffold-based conditioning**, used by models like ControlMol[44], DiffHopp [63], MolSnapper[86],and **MADGEN**, which first predicts a likely molecular scaffold from the spectrum and then generates the full structure around it [75]. Alternative approaches going from spectrum to molecule, exemplified by DiffSpectra[72] or **DiffMS**, uses an end-to-end, **formula-constrained** framework [6]. It combines a transformer encoder with a discrete graph diffusion decoder, uniquely pretraining the decoder on vast fingerprint-structure datasets before finetuning on the much smaller set of available spectra.

Despite these advances, all *de novo* methods face significant hurdles due to the complexity of the spectra and the limited availability of large-scale, high-quality paired spectrum-molecule datasets for training [9].

1.4 Research Questions and Contributions

This dissertation addresses these fundamental gaps through the development and validation of novel computational tools. The research journey progresses from a targeted analytical tool to a state-of-the-art generative model, guided by the following questions:

Can we computationally localize the site of a structural modification between a known molecule and an unknown analog using their respective MS/MS spectra? This question is answered by the development of **ModiFinder**, a novel tool that leverages the alignment of shifted and unshifted peaks in MS/MS spectra to score the likelihood of

each atom in a known structure being the modification site.

Do additional spectra (different collision energies/adducts) meaningfully improve localization? To answer this, Chapter 3 investigate the performance of ModiFinder using MS/MS data from multiple collision energies and mass spectrometry adducts and demonstrates that integrating this complementary spectral information significantly improves localization performance and expands ModiFinder's utility to a wider range of compounds.

Can we advance from site localization to predicting the complete structure of an unknown molecule by framing the problem as a conditional generative task? This leads to **ModiStruct**, diffusion-driven generative model for molecular inpainting. ModiStruct takes a known substructure and its corresponding spectra as a conditional input and reconstructs the complete molecular graph, achieving state-of-the-art results in de novo structure prediction.

1.5 Notation & Definitions

We denote vectors $\mathbf{x}$ by bold lowercase letters and tensor $\mathbf{X}$ by bold uppercase letters. To maintain consistency, we follow the terminology of the DiffMS [6] paper wherever possible.

Molecule Graph A Molecule Graph $G = (V, E, \mathbf{X}, \mathbf{A})$ is the representation of the structure of a molecule as a *connected undirected* graph with node set V, edge set $E \subset V \times V$, node labels $\mathbf{X} \in \{0,1\}^{n \times d}$ one-hot representation of node labels (in this manuscript only the most common atom types are considered for node labels $d = 8$ and atom types: {C, O, P, N, S, Cl, F, H]}), and $\mathbf{A} \in \{0,1\}^{n \times n \times k}$ the one-hot adjacency tensor for bond types (with $k = 5$ possibilities: {no bond, single, double, triple, aromatic}). It is easy to see that $\mathbf{X}$ and the molecular formula can be easily calculated from each other.

Substructure Molecule Given molecule graph $G = (V, E, \mathbf{X}, \mathbf{A})$ and $H = (V_H, E_H, \boldsymbol{X}_H, \boldsymbol{A}_H)$, H is a substructure of G when $V_H \subseteq V$ and $E_H \subseteq E$. A substructure of G can also be shown as a mask to the adjacency matrix: $\boldsymbol{A'}_{G,H} \in \{0,1\}^{n \times n}$ where $\boldsymbol{A'}_{G,H}[i,j] = 1 \iff (i,j) \in E_H$. This is true because the $\boldsymbol{A}_G$ (and so E_H) can be calculated by masking $\boldsymbol{A}$ and since the Molecule Graphs are considered connected, the node set V_H (and so X_H) is the set of nodes in V with at least one edge in E_H.

Modification Sites Given a molecule graph G and its substructure H. The modification sites $C_{(G,H)}$ is defined as:

$$C_{(G,H)} = \{v \in V_H \mid \exists u \in (V \setminus V_H), (u,v) \in E_G\}$$

Molecular Fingerprint For a Molecular Graph G, its *fingerprint* is a binary vector $\mathbf{y}_G \in \{0,1\}^c$ that encodes its structural features, where different elements indicate the presence of different features in the molecule for example existence of a ring structure. In this manuscript, $\mathbf{y}_G$ is a 2048-bit Morgan [39] fingerprint ($|c| = 2048$) that is easily deterministically calculated from the molecule structure G.

Mass Spectrum A mass spectrum S is a variable-length set of peaks where each peak is a tuple $s_i = (m_i; y_i) \in \mathbb{R}^2_+$. We assume the spectrum contains only those peaks whose y values exceed a chosen threshold

MS2-known: The MS/MS spectra of known compound

S-known: The 2D chemical structure of the known compound

MS2-unknown: The MS/MS spectra of unknown compound

Proximity cover: assesses the distribution of likelihood scores relative to the true modification site and examines whether the high-scoring atoms are in close proximity to the actual modification site.

Ambiguity cover: evaluates the entropy of the prediction array and its informativeness. For instance, an array where most atoms have the same high-score exhibits high ambiguity and may not be helpful for localization.

1.6 Dissertation Outline

This dissertation is structured to narrate the progression of this research.

- Chapter 2 introduces the foundational ModiFinder algorithm for modification site localization, detailing its methodology and initial validation.
- Chapter 3 describes the enhancement of ModiFinder through the integration of multi-spectrum data, demonstrating improved robustness and applicability.
- Chapter 4 presents ModiStruct, a diffusion-based generative model that reframes de novo prediction as a molecular inpainting problem, achieving a new state of the art.
- Chapter 5 synthesizes the key findings, discusses the overarching contributions of this body of work, and outlines promising directions for future research.

Chapter 2

ModiFinder – Algorithm & Benchmarking

2.1 Motivation

While analog search and molecular networking are powerful tools for connecting unknown compounds to structurally similar, known molecules in spectral libraries, they stop short of providing a complete structural solution. These methods can suggest a relationship but do not automatically describe the nature or, crucially, the specific location of the chemical modification that differentiates the two compounds. This leaves a critical gap in the workflow, requiring time-consuming manual interpretation by mass spectrometry experts to hypothesize and validate the structural difference. To bridge this gap and automate a key bottleneck in the discovery of novel compounds, **ModiFinder** is introduced. ModiFinder utilizes the existing library data to find known analogs to the unknown query compound (Figure 2.1) to

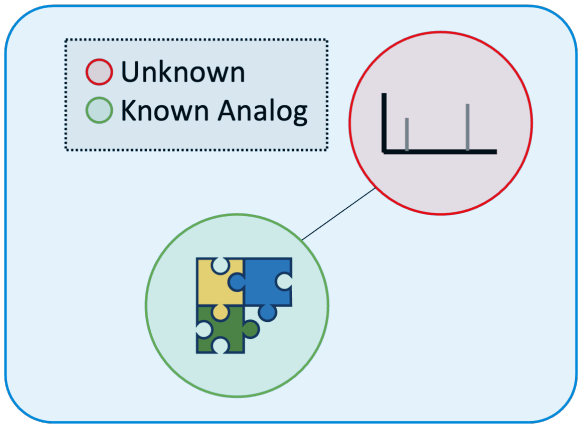


Figure 2.1: Unknown Spectra and its known Analog

provide insights into the structural differences between known compound and their unknown analog. This approach allows for more dynamic and comprehensive analysis.

The proposed approach borrows a concept from the computational challenge of site localization of post-translational modifications (PTM) of peptides in bottom-up proteomics[16, 84, 11]. In PTM site localization, b/y ions that flank the modification site are used to localize the putative PTM on a linear peptide. In this Dissertation, this concept is translated to localize structural modifications of small molecule graphs - representing 2D molecular structures. In contrast to PTM site localization, the ability to explain the MS/MS fragmentation, while simpler in peptides, is significantly more difficult in small molecules[47]. This complexity is underscored by the plethora of methodologies, including MetFrag[48], MAGMa[46], MIDAS[74], and MS-Finder[64] developed to tackle the small molecule fragmentation analysis. Despite the ongoing challenge of explaining MS/MS fragmentation[65], This dissertation shows that these in silico approaches are still useful in addressing the problem of localizing structural modifications on small molecules.

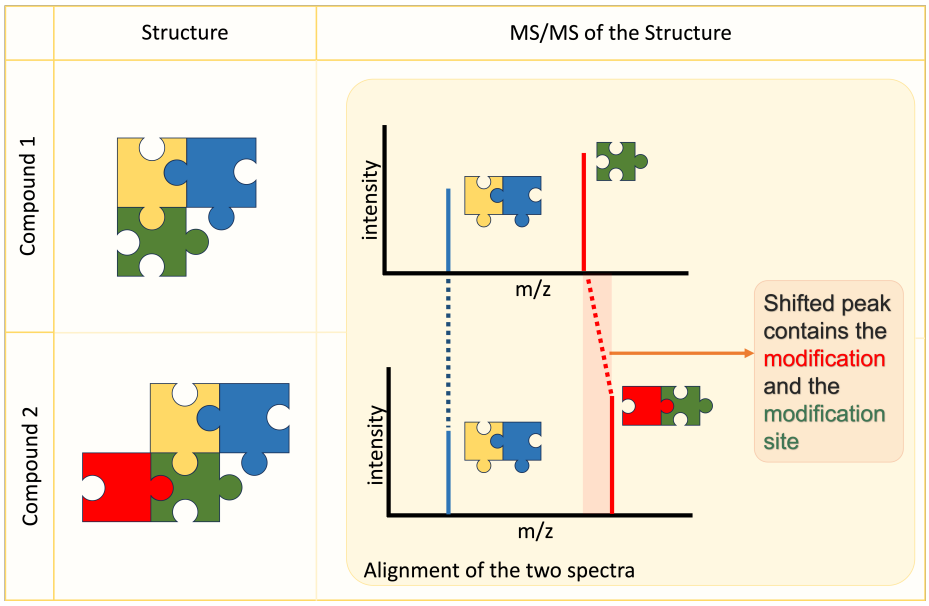


Figure 2.2: The MS/MS of Compound 1 and Compound 2 are aligned and matched peaks along with the substructures assigned to them are visualized. The matched peaks are shown in blue for the unshift peak and red for the shift peak. The matched shift peaks differ by the mass of the modification (the red puzzle piece) and contain the modification site (green puzzle piece).

ModiFinder leverages the insight that flanking masses for small molecule modifications can be determined by comparing the MS/MS spectrum of an unknown structure with a modification (MS2-unknown) with the MS/MS spectrum of the unmodified known structure (MS2-known). Specifically, the peaks that are shifted by the mass of the modification between the MS2-known and MS2-unknown putatively represent substructures that contain the modification site. Conversely, peaks that do not shift in mass between MS2-known and MS2-unknown, are less likely to include the modification (Fig. 2.2). Combining this information, ModiFinder computes a likelihood score for the specific site of modification across all atoms in the known compound (S-known). To accomplish this, each peak is

assigned a set of possible substructures using combinatorial fragmentation[46]. Then, for each peak of MS2-known that has a corresponding shifted peak in MS2-unknown (signifying the substructure includes site of the modification), ModiFinder increases the likelihood scores of atoms in the assigned substructures for the shifted peaks. If the peak is unshifted, the likelihood is decreased. Finally, ModiFinder is able to map out the likelihood landscape for where the modification may occur across the S-known using the likelihood scores (Fig. 2.3).

2.2 ModiFinder Algorithm

Given the MS/MS of known (MS2-known) and unknown (MS2-unknown) compounds and the structure of the known compound (S-known) as input (Fig. 2.3-A), ModiFinder produces a likelihood distribution of the modification site location (Fig. 2.3-D). The process begins with ModiFinder assigning a set of potential substructures to the peaks in the MS2-known spectrum through in-silico fragmentation of S-known (Fig. 2.3-B). Subsequently, the MS2-known and MS2-unknown are aligned to find the matching peaks in each respective spectrum, producing peaks that have shifted in mass (shifted) and those that remain unchanged (unshifted). To predict the site localization, a likelihood score is assigned to each atom, where atoms in the substructures of the shifted peaks are rewarded while the atoms in the substructures in unshifted peaks are penalized (Fig. 2.3-C). Finally, The likelihood score of each atom is calculated. The following sections provide a detailed explanation of each step in this process.

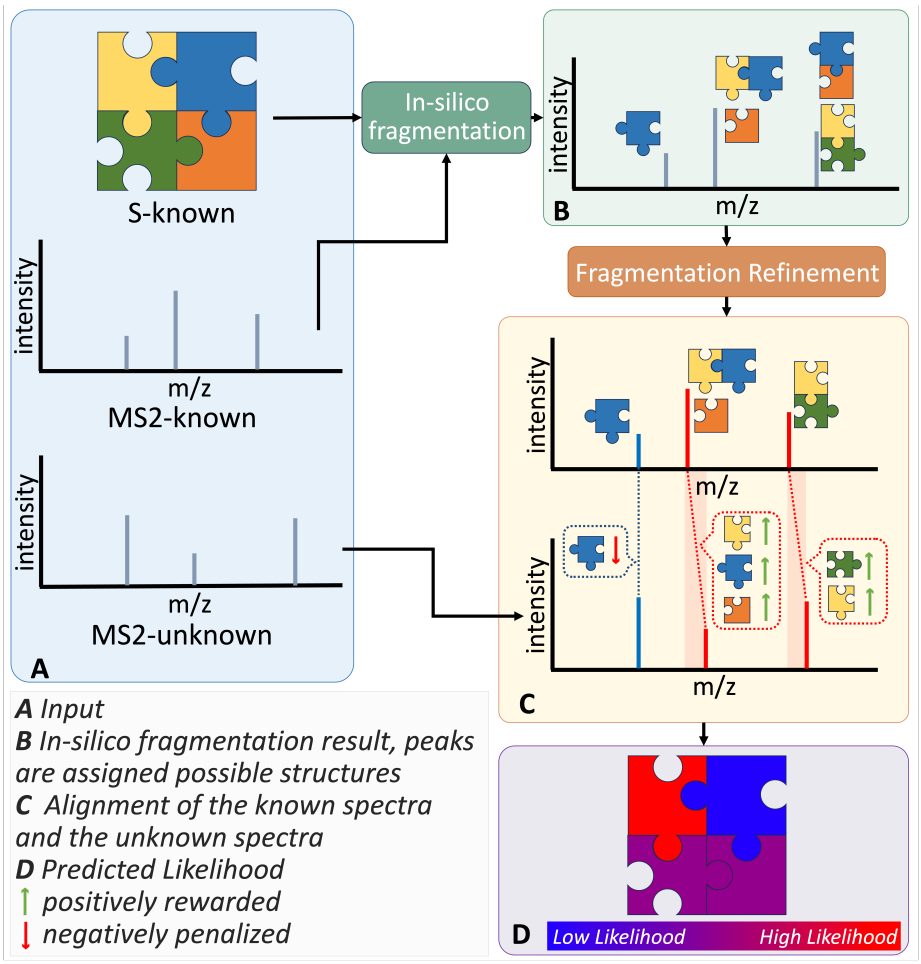


Figure 2.3: (A) The input of ModiFinder includes the known structure (S-known), spectra of the known compound (MS2-known), and spectra of the unknown compound (MS2-unknown). B) First, in-silico fragmentation methods compute potential substructure annotations for each MS/MS peak. Then, these substructures are refined by molecular formula and with helper MS/MS with a similar structure. C) Spectral alignment is performed to identify corresponding peaks (shifted and unshifted). The atoms in the substructures assigned to the shifted peaks are positively rewarded (increase in score) and atoms in the substructures assigned to the unshifted peaks are negatively penalized (decrease in score). D) Finally, a likelihood score is calculated proportionally to each atom's score.

2.2.1 MS/MS Alignment

First, as a preprocessing step, all the peaks with intensities less than 1% of the base peak are removed, and the peaks are normalized to sum to a Euclidean norm of 1 to reduce noise4. Then, the GNPS7 alignment method is utilized to identify matched peaks between the known and unknown spectrum, by accounting for the mass delta of their respective precursors[73, 3]. In the alignment process, the GNPS alignment method considers two types of matches: one where peaks have the same mass (non-shift), and another where peaks are offset by the difference in their precursor masses (shift). For each peak in the known compound's spectrum, the availability of both non-shift and shift peaks are examined and all possible matched candidates of each peak are considered; Out of all these possibilities, the GNPS alignment method efficiently approximates the best-scoring match. Specifically, A bipartite graph is created where the nodes represent the peaks of MS2-known and MS2-unknown. An edge is drawn between an MS2-known peak and an MS2-unknown peak under two conditions: if their difference is less than a predefined threshold, indicating an unshifted match, or if it lies within the threshold range relative to the difference in precursor masses of the known and unknown compound, indicating an shifted peak. The weight of each edge is the product of the intensities of the corresponding peaks. The goal is to find the maximum-scoring match. A greedy algorithm is used to approximate this matching. At each step, the edge with the maximum weight out of all the remaining edges is selected and added to the result. Then the chosen edge, both of its ends, and all edges connected to these ends are eliminated from the graph. In our experiments, a tolerance of 40 (ppm) is adopted as the difference threshold used to calculate the existence of the edges.

2.2.2 Combinatorial Fragmentation

The peaks within the MS2-known spectra are annotated by attributing a series of possible substructures to each peak. These substructures are derived from the known structure (S-known) utilizing the MAGMa method[46]. In short, the fragmentation of S-known goes as follows: First, each of S-known's heavy (non-hydrogen) atoms are removed once, each time yielding one or more substructures. Then the same process is repeated for each of the resulting substructures. The full fragmentation of the S-known is performed in a breadth-first search traversal. The generated substructures are stored as a bitstring where each bit represents one of the heavy atoms in S-known. Although MAGMAa lacks the functionality to manage rearrangements involving heavy atoms, it can to some degree track hydrogen rearrangements. This is achieved by recording each fragment with various weights, accommodating potential hydrogen loss or gain. ModiFinder begins with the initial structure (S-known) and continues the aforementioned breadth-first fragmentation approach up to a predetermined depth; a maximum depth of 2 is chosen here for the experimental setup.

Once the fragmentation step is done, for each substructure, the theoretical charged-m/z is calculated and compared to each peak's m/z in the MS2-known. The maximum charge is assumed to be 1. If the theoretical m/z of a substructure falls within a specified m/z tolerance to the empirical m/z of a peak, then that substructure is assigned to that peak. Here, 40 ppm was chosen as the default m/z tolerance.

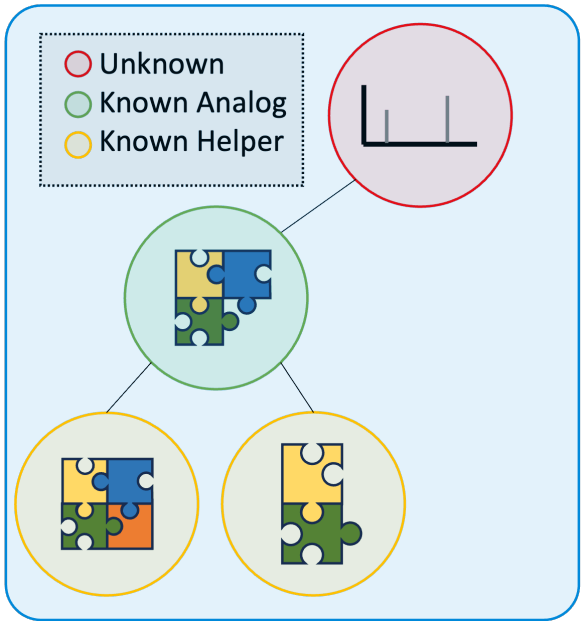


Figure 2.4: ModiFinder with helpers

2.2.3 Substructure Refinement by Formula

Predicted formulas provided by [19] (MF-S) or Buddy[80] were used to filter the possible substructures for each MS/MS peak. SIRIUS, given the spectra of a compound, generates a pool of potential candidates using the information of the MS1. Next, it evaluates the interpretability of MS/MS spectra for each candidate by constructing a fragmentation tree[45]. The ModiFinder algorithm leverages this information by parsing the fragmentation tree and retrieving the formula assigned to each peak. This formula is then used to remove any substructure assigned to that peak with a different formula. Due to the performance complexity, SIRIUS is only computed for compounds with a precursor mass of 500 Da or less.

2.2.4 Refinement by Helpers

ModiFinder leverages additional compounds in MS/MS libraries that exhibit structural similarities to the known compound S-known, referring as helper compounds, to refine the substructure annotations. Specifically, the compounds within the same MS/MS library as S-known that share identical adducts and instruments but differ from S-known at precisely one modification site are identified as potential helper compounds. To ensure there is no information leakage and the unknown compound is not among the helpers, any compound that possesses a precursor mass within a 0.5 range of the precursor mass of the unknown compound is eliminated. Suppose $H_{S-known} = \{h_1, \ldots, h_n\}$ as the set of selected helper compounds for S-known. Then, the main idea of ModiFinder is again applied. That is, if a peak is shifted, then the modification should appear in the substructures assigned to the peak, since both S-known and h_i's structure is known, the modification site is easily calculated and can be used to filter the substructures assigned to the peaks of MS2-known. For each helper compound, denoted as h_i, the same in-silico fragmentation process is performed on h_i's structure to annotate h_i's peaks. Next, h_i's spectra are aligned to the MS2-known to find matched peaks (shift and unshift). Then, the modification site is calculated based on S-known and h_i's structure. Finally, for every peak that's shifted, any substructure assigned to that peak in MS2-known that does not include the modification site between S-known and h_i is eliminated. Fig. 2.5 visually demonstrates this concept through a toy example.

The intuition behind this Refinement is that although the modification site of the S-known relative to the unknown analog might differ from that of the S-known and the helper, there may still be overlap in the shifted peaks between the helper and MS2-known,

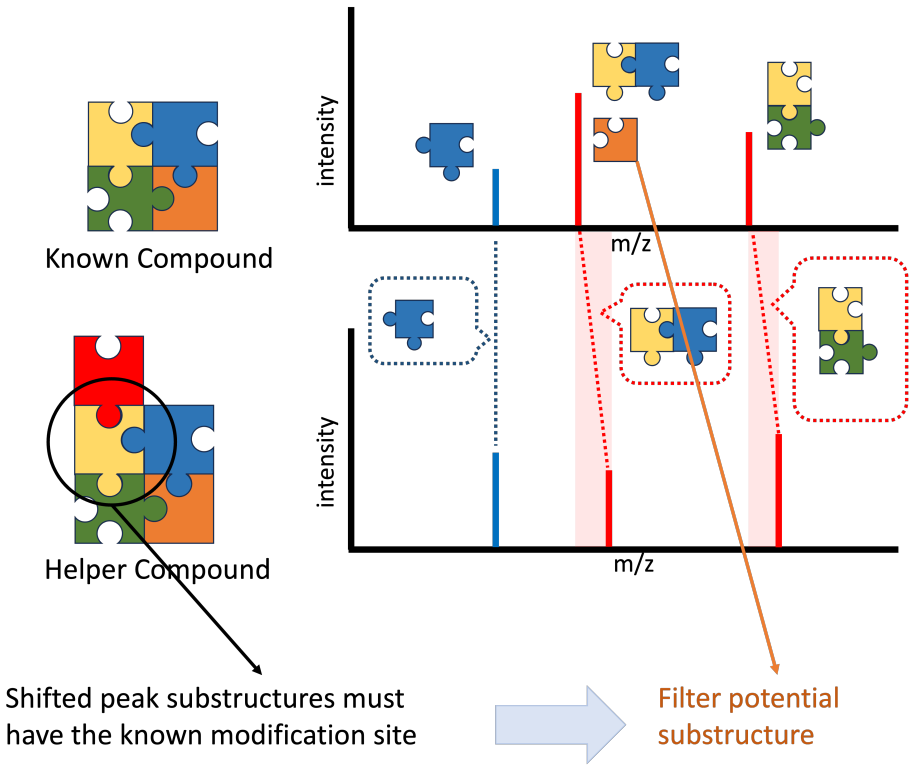


Figure 2.5: Toy example for refinement by helpers

and those observed between MS2-known and MS2-unknown. This overlap in shifted peaks can help refine the substructures of S-known that shift even if the modification sites are different.

2.2.5 Calculating the Site Localization

A score for each atom of the S-Known is computed, indicative of the likelihood of being the modification site. This score, termed as the "likelihood score" and shown by Ω (Ω_j for atom with index j in the graph), aims to serve as a score that measures the amount of evidence of an atom's candidacy for being the site of modification. This scoring is performed under the assumption that there is only one modification site. Under this assumption, shifted peaks are probable hosts of the modification site. In Contrast, the atoms presenting in matched but unshifted peaks are penalized.

Each matched peak assigns a contribution score to each atom. $\Theta_{i,j}$ shows the contribution score assigned to atom j by peak i. For i-th matched peak, the scores for the atoms are calculated as follows: initially, the contribution scores for all atoms are assigned a value of 0. Then, assuming S_i is the set of all the substructures assigned peak i:

$$\Theta_(i,j) = \sum_{\sigma \in S_i} \frac{1_{\{j \in \sigma\}}}{|\sigma| \times |S_i|}$$

Where $|\sigma|$ is the number of atoms in substructure σ, S_i is the number of structures assigned to peak i, and the $1_{j \in s}$ is the indicator function that is 1 if j-th atom exists in substructure σ and 0 otherwise.

Each matched peak itself receives a clarity score that represents how informative its substructures are. For example, if a peak has one substructure assigned to it but the substructure

contains all the atoms in the compound or if the peak has multiple substructures assigned to it and overall each atom appears the same number of times, then the peak is not informative and must receive a low clarity score. Similarly, if the peak has few structures and they all focus on a specific and small part of the atom, then the peak is considered informative and will receive a high clarity score. To compute this clarity score, the Shannon entropy[53] is calculated. The clarity score of the peak i, C_i, is proportional to this entropy score:

$$C_i = 1 - \frac{\sum_{i=0}^{n} \Theta_{i,j} log(\Theta_{i,j})}{log(n)}$$

Where n is the total number of atoms. Finally, Ω_j is calculated as follows

$$\Omega_j = \sum_{i \in \text{shifted}} C_i \Theta_{i,j} - \sum_{i \in \text{unshifted}} C_i \Theta_{i,j}$$

Where shifted and unshifted are the set of shifted peaks and unshifted peaks respectively. Figure 2.6A. displays the alignment of two spectra. Figure 2.6B. shows the annotations assigned to each peak. The annotation are highlighted on the structure of the known compound (L-glutamic). Shifted peaks are marked in a red-boxed outline, whereas the unshifted peaks (e.g. m/z 85.029) is outlined in blue. Figure 2.6C. Shows the contribution score $\theta_{(i,j)}$ that is calculated for each peak i and each atom j. This score correlates with the frequency of each atom j appearing in annotations of peak i, and is visually represented on the structure: red indicates the highest contribution and blue the lowest. In addition, the clarity score C_i is calculated for each peak i using the contribution scores. For each peak, the clarity score is shown beside the structure. This score is depicted alongside the structure, decreasing as the annotation distribution becomes broader (e.g., peak 4 with m/z 130.05 has a clarity score of 0.01) and increasing as it becomes more focused (e.g., peak 0 with m/z

56.05 has a clarity score of 0.32). These scores are then combined (shifted peaks as positive reward and unshifted peaks as negative penalties). Figure 2.6D. This is the true structure of the modified analog, with the structure change highlighted in red. E. ModiFinder prediction is created based on the combination of the scores. We note that this highest scoring atom (N) is the true modification site highlighted in (D) but due to drawing differences in drawing software the molecule is rotated.

2.2.6 MS/MS Spectral Library Data Preparation

The creation of the database used for ModiFinder evaluation involved a multi-step process. Initially, compounds with known 2D structures were selected from MS/MS libraries containing compounds with known structure[43, 62, 41, 22, 24, 66]. The following public MS/MS spectral libraries were used to retrieve the MS/MS and structure pairs:

1. GNPS Library
2. NIH Natural Products Library
3. GNPS-NIH-SMALLMOLECULEPHARMACOLOGICALLYACTIVE
4. Berkeley Lab

In addition, data from the TUEBINGEN-NATURAL-PRODUCT-COLLECTION was used for the web tool performance demonstration. For each library, every possible pair in that library are analyzed to verify their eligibility. For each pair with known structure, (i) their precursor mass is checked to be less than 2000 Da, (ii) the difference in precursor masses is less than 50% of the precursor mass of the smaller compound, (iii) they share the same M+H

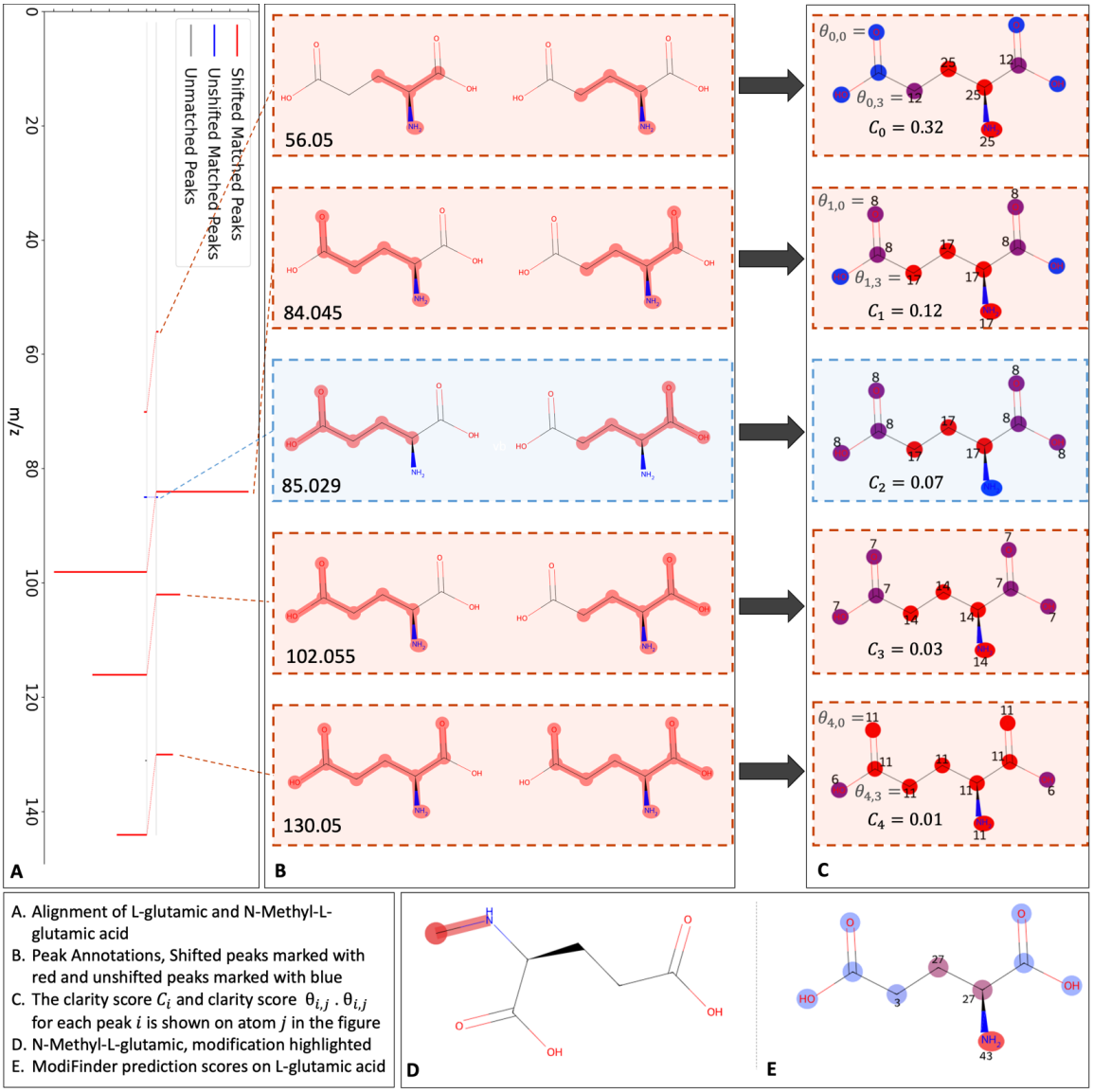


Figure 2.6: Example of ModiFinder method.

adduct, and finally, (iv) the structures are examined to differ in exactly once modification site. In the final verification step, both SMILES structures are converted to an RDKit[27] molecule object, then the "GetSubstructMatch" function is called on the heavier compound's object with the smaller compound's object as input. If the smaller compound is a substructure of the larger compound, the number of edges between the atoms in the substructure set and the atoms not in the substructure set is calculated as the number of modification sites. Any pair with more than one edge is discarded.

2.2.7 LC-MS/MS Data Acquisition Methods

LC-MS/MS Data was acquired on a Vanquish ultrahigh-performance liquid chromatography (UHPLC) coupled to a Q Exactive HF apparatus (Thermo Fisher Scientific) equipped with an heated electrospray ionization (HESI) source. The chromatographic separation was performed with a constant flow rate of 0.15 mL/min (micro-flow setup) or 0.5 mL/min with mobile phase A (H2O +0.1% formic acid) and mobile phase B (ACN +0.1% formic acid). The separation gradient started with 5% B as initial conditions, which was linearly increased to 50% B at 8 minutes, and then to 99% B at 10 minutes, followed by a washing phase with 99% B, and a re-equilibration phase at the initial conditions. The microflow gradient was run from 5-99% B over 5 min with 2 min washout at 99% and 3 min requilibration at 5% B. Both methods used a Kinetex 1.7 µm EVO C18 RP, 100 Å pore size, reversed phase UHPLC column (Phenomenex) with the dimension 50mm x 2.0mm and 50 x 1 mm (microflow). The MS/MS method was previously optimized[58], the HESI source parameters were set as follows: auxiliary gas flow and temperature were respectively 12 AU and 400 °C, sweep gas flow was 1 L/min, sheath gas flow rate was set to 50 AU. The MS

was operated in positive mode, The spray voltage was set to 3.5 kV while an inlet capillary temperature of 250 °C was adopted. The scan range was set to 150-1500 m/z or 120-1500 m/.z (microflwo setup) and the resolution to 30,000. The fragmentation was performed in Data Dependent Analysis (DDA) mode, the 5 most abundant ions were fragmented per MS survey scan with a resolution of 15,000, or 45,000 for the microflwo setup (both MS1 and MS2). An isolation window of 1 m/z, and with the following stepped normalized collision energy: 25, 35, 45 eV.

2.3 Experimental Evaluation

2.3.1 Evaluation Score

To measure the performance of ModiFinder and compare it to alternative approaches and baselines, an evaluation function is needed. This evaluation function takes in the predicted likelihood array together along with the true modification site, i.e. the 2D graph structure of the known compound and the actual modification site location. The evaluation produces a score between zero and one. Scores approaching one signify more accurate predictions, while those closer to zero indicate less accurate predictions.

Average-Distance The Average-distance evaluation method is proposed for the evaluation.

$$S_{Average-Distance} = \sum_{i=0} n\Omega_i e^{\frac{-d_{i,\gamma}}{\Phi - d_{i,\gamma}}}$$

Where $d_{i,\gamma}$ denotes the distance between the atom with index i and true modification site γ on the 2D graph structure and Φ denotes the diameter (greatest shortest distance between

any two nodes in the graph) of the 2D graph structure. Using the diameter helps normalize the distances based on the size and structure of the compound. Normalization ensures uniformity in the evaluation metric across molecules of varying sizes. In Average-distance, the impact of each atom on the total score decreases exponentially with its distance from the actual modification site. Atoms with high predicted likelihood situated far away from the true site contribute less to the evaluation score, whereas those in closer proximity contribute more. This aspect addresses the Average-distance's capacity to account for the proximity cover. In addition, since the scores are normalized to have a sum of 1, the likelihood scores are directly proportional to each atom's relative influence. Consequently, in ambiguous scenarios where many atoms have high predicted likelihood, the relative likelihood of a single atom diminished. This reduction, in turn, lessens their overall effect on the evaluation score, encapsulates the method's ambiguity cover.

Alternative Evaluation Metrics for Localization Performance

$S_{Is-max-unadjusted}$ The first evaluation method that comes to mind is a binary classifier where it assigns 1 to the predicted array if the prediction likelihood assigned to the true modification site is the maximum prediction likelihood score within all atoms. This can be formulated as follows:

$$S_{IS-Max-unadjusted} = \begin{cases} 1 & if\ \Omega_\gamma \in max(\Omega) \\ 0 & otherwise. \end{cases} \tag{2.1}$$

Where Ω is the likelihood scores for the atoms, γ is the true modification site, and max(Ω) is the set of all high-scoring atoms. This evaluation score falls short as an effective metric as

it returns 1 in scenarios when every atom is assigned the same value, despite the array of predictions providing no valuable insights, highlighting its inadequacies for the ambiguity cover properties. Additionally, this method disregards the distribution of high-likelihood atoms in the structure, indicating it also falls short in the proximity cover properties as well. To adjust this metric for ambiguity, we propose the is-max metric.

Is-max evaluation metric is defined as follows

$$S_{Is-Max} = \begin{cases} \frac{1}{|max(\Omega)|} & if\ \Omega_\gamma \in max(\Omega) \\ 0 & otherwise. \end{cases} \tag{2.2}$$

proximity-unadjusted The next evaluation method is based on the shortest distance from the true modification site to an atom with the maximum prediction score. This can be formulated as:

$$S_{Proximity-unadjusted} = \frac{1}{min_{i \in max(\Omega)} d(i,\gamma) + 1} \tag{2.3}$$

Where d(i,γ) is the shortest distance from atom i to the true modification site on the graph. This method disregards the size of the atom completely and the absence of adjustment for the atom's overall size leads to a scoring bias, favoring smaller compounds with higher evaluation scores. A prediction indicating the furthest distance from the actual modification site in a small molecule may still be closer than a prediction targeting an atom within 10% of the true modification site, resulting in a higher score while providing less useful information.

Proximity

$$S_{Proximity} = \frac{\Omega - min_{i \in max(\Omega)} d(i,\gamma) + 1}{\Omega} \tag{2.4}$$

This method does not have a good ambiguity cover as the same example provided for is-max-unadjusted applies here as well.

Sorted-rank Along with these evaluation functions, a Sorted-rank evaluation is also reported as it has been used for similar applications before. The Sorted-rank is correlated to the relative location of the true modification site's likelihood score in comparison to other atoms

$$S_{Sorted-rank} = \frac{index\ of\ \gamma\ after\ sorting\Omega}{n-1} \tag{2.5}$$

Where n is the number of atoms. If the true modification site has the highest score, it will be moved to the $(n-1)^{th}$ position after sort (index n) resulting in the evaluation score of 1. In contrast, if it has the lowest score, it will be moved to the first location after sorting (index 0) resulting in a score of 0. It is worth noting that the Sorted-rank also lacks proximity cover as it completely ignores the 2D structure graph.

2.3.2 Site Localization baselines and alternative approaches

Random Choice (RC) adopts a random selection approach for designating one of the atoms as the modification site. In contrast, our second baseline, termed "Random Distribution" assigns a likelihood score $|S_i|$ to atom i, where $S_i \sim N(0,1)$.

The Oracle approach is built on top of ModiFinder and uses the extra information of the true modification site. After ModiFinder has annotated the peaks with putative substructures using the combinatorial fragmentation and formula and helper refinements, Oracle applies an extra elimination step. Specifically, for every shifted peak of MS2-known, Oracle filters out any substructure assigned to that peak that does not contain the true

modification site. Once this step is completed, all the substructures assigned to the shifted peaks are guaranteed to contain the modification site.

In addition, utilizing the MS/MS fragmentation prediction property of CFM-ID[71] a different modification site method is developed as an alternative approach and a baseline to compare against. CFM-ID is a tool to predict spectra based on a given molecular structure. This alternative approach is only developed for evaluation as it uses the structure of the "unknown compound" which is paradoxical and impractical in real-world scenarios. It serves merely as a reference for comparison, and the ability of ModiFinder to surpass its performance, despite the latter's theoretical omniscience, further emphasizes the effectiveness of ModiFinder. This alternative method is designed to use CFM-ID as a black box to find the modification site (Refer to Fig 2.7 for a visual illustration of this method). First, using the extra information provided by the structure of the "unknown" compound, the modification substructure is calculated. Then, this modification substructure is permuted across the known structure (S-known). With each permutation, the modification is attached to an atom in S-known, creating an analog to S-known and a possible candidate for the unknown compound. To attach the modified part to an atom, first the same original bond type is tried, if that does not produce a valid structure, other bond types are tried. After this step, CFM-ID tool is used to predict spectra for each structure.

2.3.3 Performance

We introduce three versions of the ModiFinder method. First, is the basic version of ModiFinder (MF-N). Second, is the refined version of ModiFinder (MF-R) that utilizes

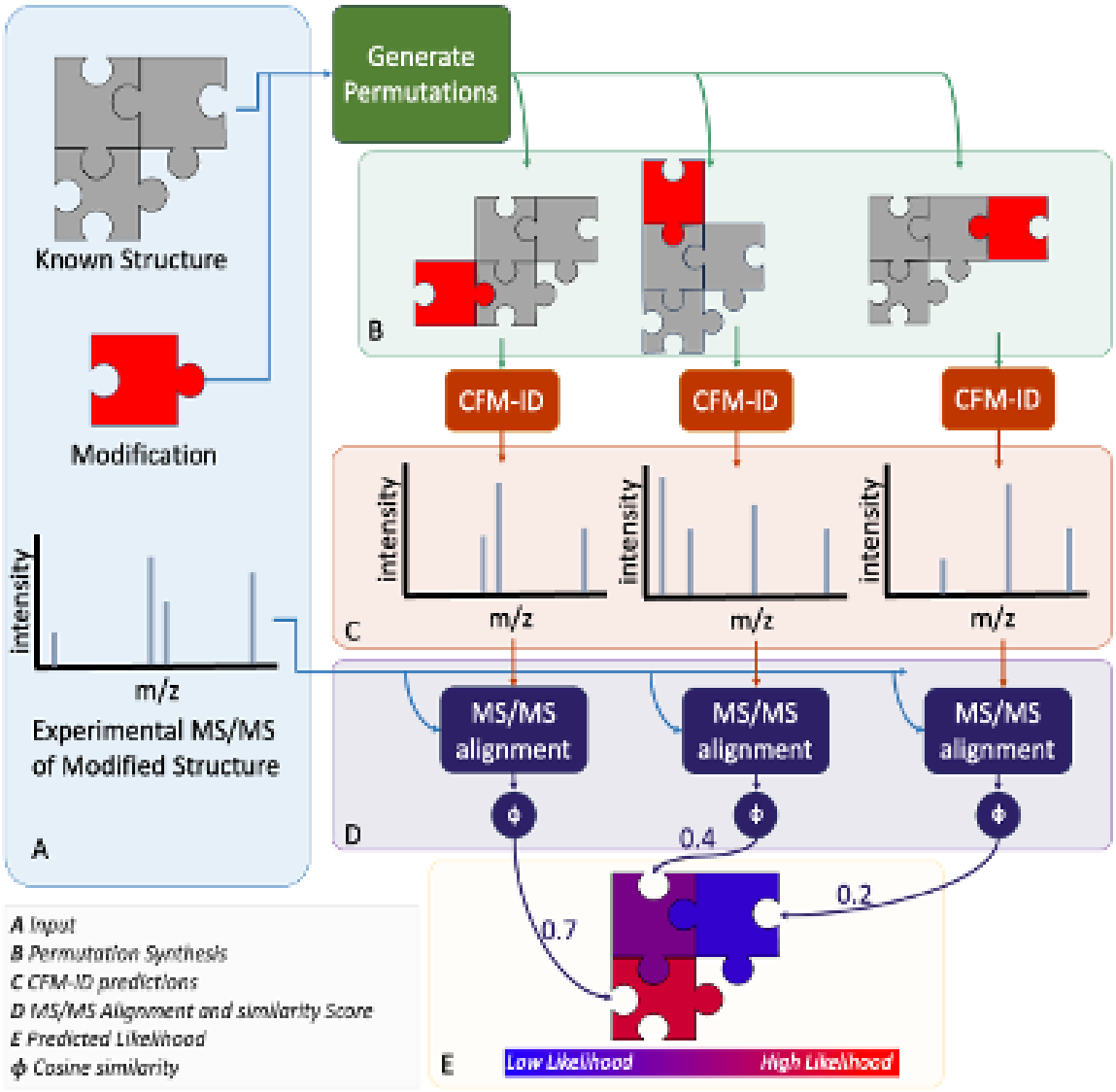


Figure 2.7: A. the input section with two structures and the spectra of the bigger compound. B. permutation of the difference around the smaller compound to create possible analogs. C. prediction of the spectra using CFM-ID. D. Aligning the spectra of the larger molecule with each of the spectra predictions. E. assigning a likelihood score to each atom based on the cosine score of the alignment.

molecular formula filtration and substructure ambiguity refinement utilizing structurally related helper MS/MS spectra. Third, an Oracle Method (MF-O) is introduced that has knowledge of the true modification site to provide an approximate upper bound of ModiFinder performance. This is achieved by simulating the ability to reduce the ambiguity of substructure assignments to the MS/MS peaks by eliminating substructures that do not contain the modification site in shifted MS/MS peaks. Additionally, the site localization performance is evaluated on two random baselines: Random choice (RC) and Random Distribution (RD). Finally, an alternative in silico prediction benchmarking approach is introduced which utilizes MS/MS fragmentation prediction of CFM-ID[71].

ModiFinder's performance (MF-R in Fig. 2.8-A) lies above the random baselines (RC and RD) and the CFM-ID approach and below the Oracle (MF-O). Specifically, MF-R when compared to RC and RD baselines shows an average Evaluation Score increase of 0.181 and 0.180, respectively, across all MS/MS pairs (Fig. 2.8-A). Moreover, MF-R outperforms these baselines in 81% and 80% percent of benchmark pairs. In comparison, the MF-O version of ModiFinder outperforms the RC and RD in 85% of benchmark pairs and shows an average increase of 0.208 and 0.207, respectively (Fig. 2.8-A). MF-R exhibits enhanced performance not only relative to random baseline comparisons but also over the CFM-ID based alternative approach evidenced by an increment of 0.266 in the average evaluation score (Fig. 2.8-A). Surprisingly, the CFM-ID based approach is found to be worse than RC and RD. This is because when simulating all regio-isomers, the resulting simulated MS/MS spectra were highly similar. This resulted in a nearly uniform likelihood distribution across all atoms, which was penalized in the "Ambiguity Cover" evaluation dimension. It scored relatively

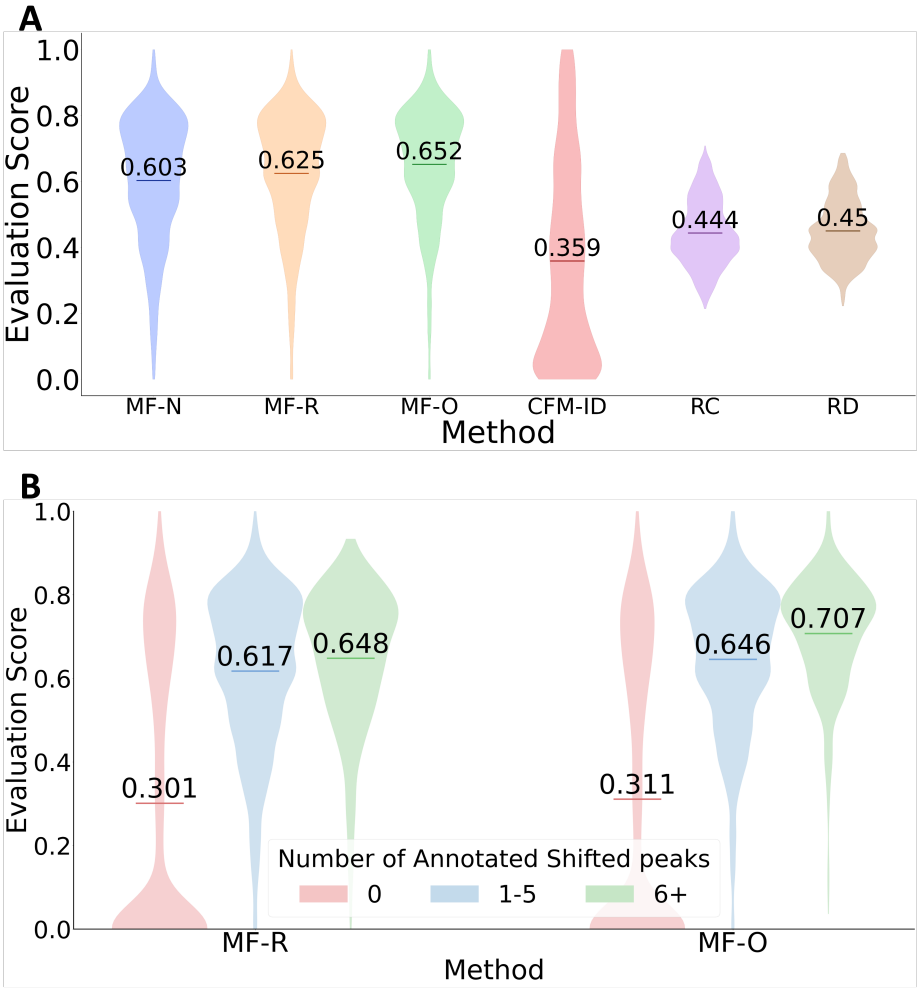


Figure 2.8: A) Evaluation scores across pairs in all the libraries for different methods where there is at least one shifted peak. ModiFinder outperforms the Random and CFM-ID baselines. MF-R, which utilizes helpers and formula constraints, improves upon MF-N, closing the gap to our upper bound performance of MF-O. B) Evaluation score across pairs in all libraries for MF-O and MF-R based on the number of annotated shifted peaks. By increasing the number of shifted peaks, the performance of ModiFinder increases. This performance increase is consistent across different datasets and even for the MF-O demonstrating the utility of shifted peaks in finding modification sites.

higher in other benchmark metrics that deemphasize “Ambiguity Cover”, e.g. “Is-max”

We note that the gap in performance between MF-R and other baselines is greatest in lib4, which constitutes the majority of our database and might introduce a bias in the result. Nevertheless, MF-R maintains a clear advantage over the baseline across all the libraries (Fig 2.9). Manual analysis of the specific cases where MF-R does not outperform the RC and RD indicates that in the majority of cases, this is because of a lack of shifted fragmentation peaks and or incomplete in silico substructure explanation of the MS/MS peaks.

With decreasing ambiguity of substructure assignments to MS/MS peaks, an increase in site localization performance is observed with MF-R improving upon MF-N, 0.625 vs 0.603 evaluation score respectively and MF-O, with the lowest ambiguity, further increases performance to 0.652 (Fig. 2.8-A, Results Section Improving the Quality of Annotation and reducing the Ambiguity increases the Evaluation Score for more details).

2.4 Analysis of Key Performance Factors

The evaluation revealed several key factors that influence ModiFinder’s performance

2.4.1 Shifted Peaks Matter in Site Localization

The results highlight the significance of annotated shifted peaks in the identification of the modification site where the MS/MS spectrum pairs featuring at least one shifted peak exhibit higher evaluation scores compared to those without any shifted peaks (Fig.

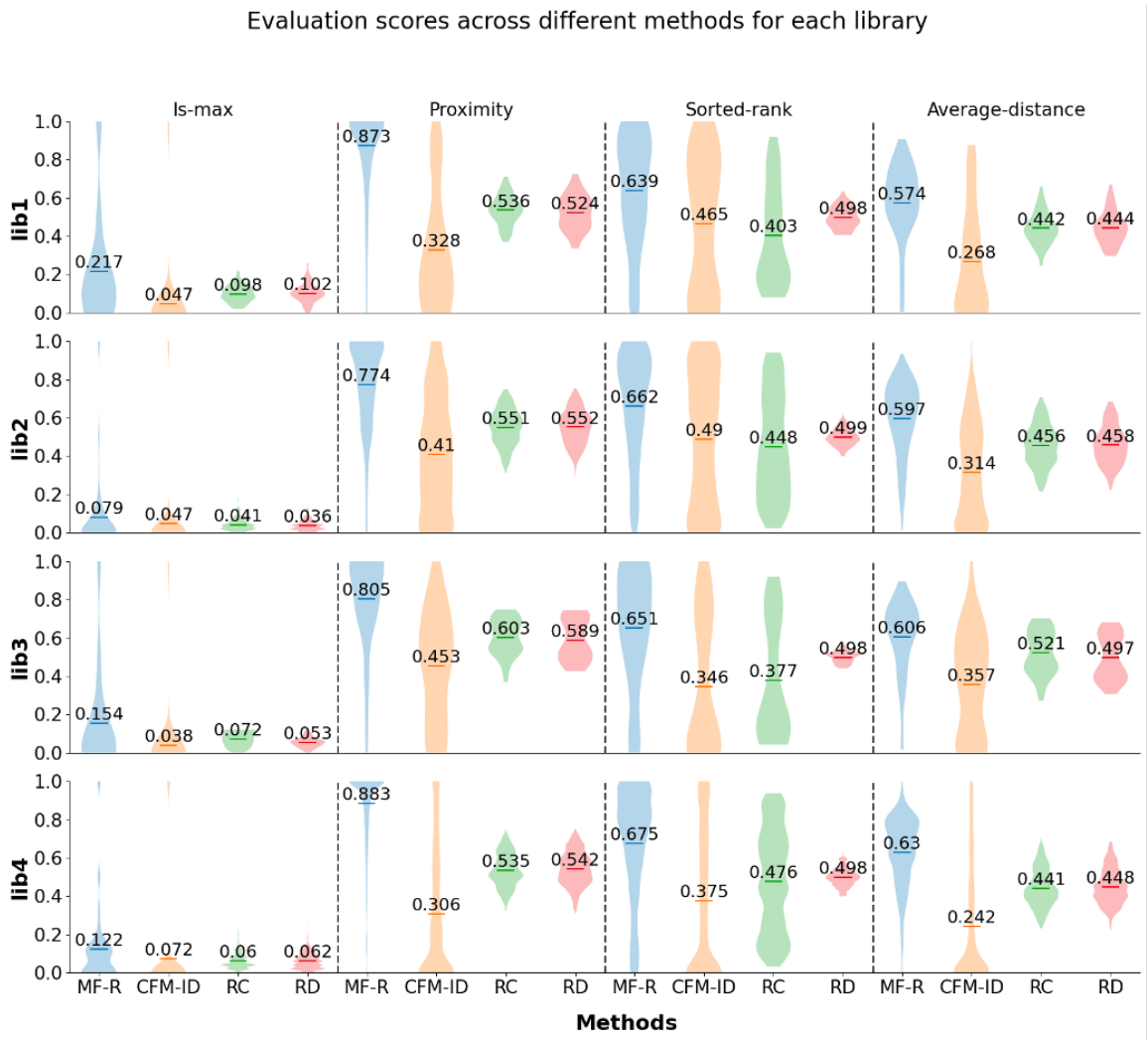


Figure 2.9: Demonstrating the Consistent Superiority of ModiFinder Over Other Baselines Across Is-Max, proximity, Sorted-rank, and Average-distance evaluation method within each library.

2.8-B). Furthermore, when there is an increase in the number of shifted peaks, the measured performance of MF-R and MF-O increases (Fig. 2.8-B). We hypothesize that the increase in shifted peaks potentially enriches the diversity of potential substructures, each MS/MS peak focusing on different segments of the compound. This diversity may reduce the overlap among substructures, thereby improving the precision in pinpointing the modification's location. MF-O utilizes the shifted peaks more efficiently than the MF-R as it can remove the ambiguity introduced by more annotations on the extra shifted peaks.

2.4.2 Increasing the Fragmentation Depth Increases Ambiguity

One of the key steps of ModiFinder is substructure annotation of MS/MS fragment peaks. ModiFinder utilizes combinatorial fragmentation, i.e. MAGMa[46] (one of the state of the art approaches [5, 69]), to generate a set of potential fragmentation substructures for every MS/MS peak. Varying the fragmentation depth from two to four modulates the site localization performance. An increase in performance with MF-O is observed with fragmentation depth 3 or 4 in comparison to depth 2 (Fig. 2.10-A). This increase in performance can be attributed to the introduction of more substructures to the peaks, revealing the true substructure, especially for the peaks that were previously unannotated. Benchmark datasets show a 19% increase in the number of annotated shifted peaks when increasing fragmentation depth from two to four. While this increased explanation benefited MF-O, at higher fragmentation depths MF-R performance decreases. This is due to increased substructure ambiguity (average number of annotations assigned to each shifted peak). MF-O can counteract the increased ambiguity by utilizing the true fragmentation site to filter out

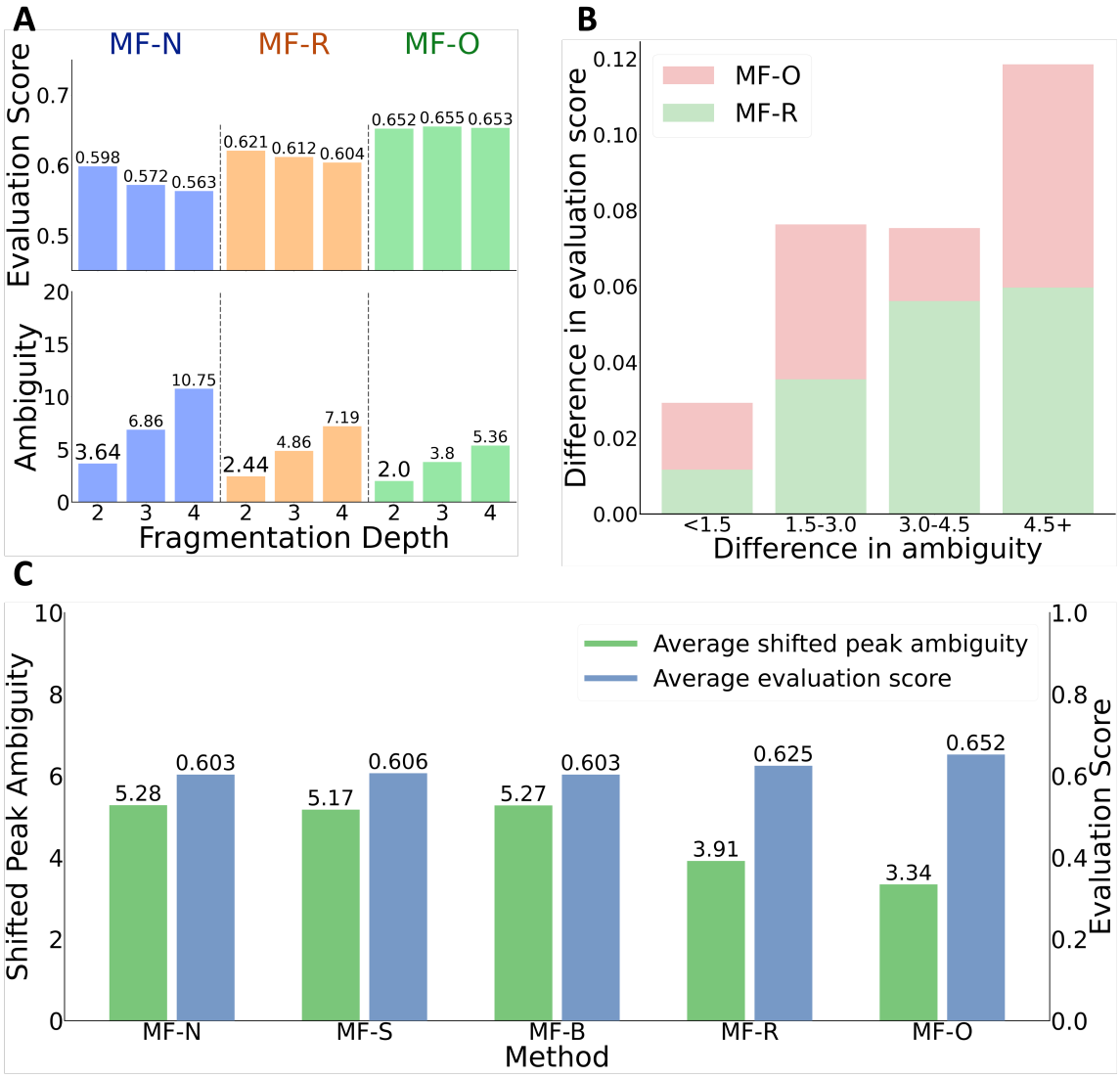


Figure 2.10: **Correlation of ambiguity and performance.** (A) Impact of Fragmentation Depth on MF-N, MF-R, and MF-O Performance over pairs in all libraries. By increasing the fragmentation depth, the Oracle method (MF-O) attains higher scores, benefiting from more explanatory fragmentation. However, due to ambiguity introduced with more in-depth fragmentation, MF-N and MF-R do not benefit in site localization performance (B) Correlation of ambiguity and evaluation scores at the pairwise level over pairs in all libraries. The evaluation score improvement from the unrefined ModiFinder (MF-N) for MF-O and MF-R based on the ambiguity reduction (difference in ambiguity). As the ambiguity difference increases, i.e., structural annotation becomes increasingly less ambiguous, the evaluation score difference increases. (C) Comparison of Average Ambiguity and Average Evaluation score across different settings of ModiFinder and Oracle for pairs in all datasets. Methods that yield lower ambiguity correspondingly achieve higher evaluation scores.

substructures that do not benefit the localization. On the other hand, both the MF-R and MF-N are unable to filter incorrect substructures, leading to a loss of focus on the true modification site. Given the performance differences, fragmentation depth of 2 is chosen as the default value for MF-R. However, we anticipate that the optimal fragmentation depth might increase with enhancements in substructure assignments to MS/MS peaks.

2.4.3 Improving the Quality of Annotation and reducing the Ambiguity increases the Evaluation Score

MF-N calculates the likelihood scores using annotations derived from MAGMa. Upon refining the molecular formula with SIRIUS[19] (MF-S) or Buddy[80] (MF-B) applied to MF-N, there was a decrease in ambiguity by 0.11 and 0.01, respectively (Fig. 2.10-C). This resulted in minimal improvements to the site localization evaluation score (Fig. 2.10-C). A larger magnitude reduction in MS/MS peak annotation ambiguity (reduction of 1.37) when helper compounds are utilized (MF-R) is observed. The most significant reduction in ambiguity (1.94) is observed, predictably, with the oracle (MF-O) (Fig. 2.10-C). There is a noticeable correlation between the decrease in annotation ambiguity and the improvement in evaluation scores; specifically, when the ambiguity of shifted peaks is reduced through refinements or the oracle's knowledge, there is a corresponding increase in the evaluation score (Fig. 2.10-C). Additionally, the impact of ambiguity reduction was analyzed on a pair-by-pair basis. By categorizing pairs of MS/MS spectra based on the extent of ambiguity reduction, it was found that larger decreases in ambiguity corresponded to more significant improvements in site localization evaluation scores (Fig. 2.10-B). Fig. 2.10-B further indicates that for a comparable reduction in ambiguity, the oracle, on average, achieves greater improvements.

This outcome is anticipated, as the oracle's role extends beyond merely reducing ambiguity and skews the distribution of unambiguous peaks towards the actual modification site by eliminating substructures that do not contain the modification site. While these substructure refinements improve the performance of ModiFinder, one key limitation is that our initial substructure annotation method (MAGMa) only supports hydrogen rearrangements. Therefore, peaks corresponding to fragments with heavy atom rearrangements may not be correctly annotated with substructures.

2.4.4 additional factors

In analyzing the instances where ModiFinder performs poorly, several factors stood out: low cosine score, low number of shifted peaks, or low number of annotated peaks. Additionally, as mentioned above, ambiguity of substructure annotations is also a contributing factor due to the true modification site being absent from the predicted substructures. This will cause ModiFinder to predict an incorrect region of the molecule as the modification site, resulting in a false positive. Alternatively, when the likelihood distribution is overly diffuse and fails to pinpoint the true modification site (false negatives) the presence of numerous unshifted peaks plays a detrimental role. Although penalizing atoms of substructures in unshifted peaks generally enhances localization performance, an excessive number can disrupt and degrade the localization.

Another key factor is the cosine score of the alignment. Although the number of matched peaks suggests a correlation between two compounds, it neglects the intensity of these peaks. The cosine score compensates for this oversight, and our experiments confirm that it plays a

crucial role in determining ModiFinder's effectiveness. Additionally, we observed that an excessive number of unshifted peaks, particularly in relation to shifted peaks, negatively impacts the results. Fig. 2.11 shows an example of this. While the modification site exists in all the substructure annotations of the shifted peaks and therefore rewarded, two other locations, referred to as interest points, are also rewarded as they appear in all but one substructure annotation. Conversely, the modification site is consistently found in the substructure annotations of the unshifted peak (with m/z 96.0447) and is therefore heavily penalized. In contrast, only one of the interest points is present in just one substructure, resulting in minimal penalties. Consequently, the final score for the modification site is lower than those for the interest points. The scoring discrepancy occurs because while we assume that the shifted peaks include the modification site, we make no definitive claims about the presence of the modification site in the unshifted peaks, which may also contain it. Nevertheless, we penalize the unshifted peaks to mitigate the bias caused by structural symmetry in the annotations of the shifted peaks.

2.5 Case Studies and User Interaction

To demonstrate its practical utility, ModiFinder was applied to localize modifications on complex natural products, including the N-methylation of Kirromycin and the methylation of Naphthomycin B.

We demonstrate ModiFinder's capabilities on exploring the chemical space of the metabolome of Synechococcus elongatus. After solid phase extraction and non-targeted LC-MS/MS analysis, we created a molecular network. We discovered a molecular family of acyl amides

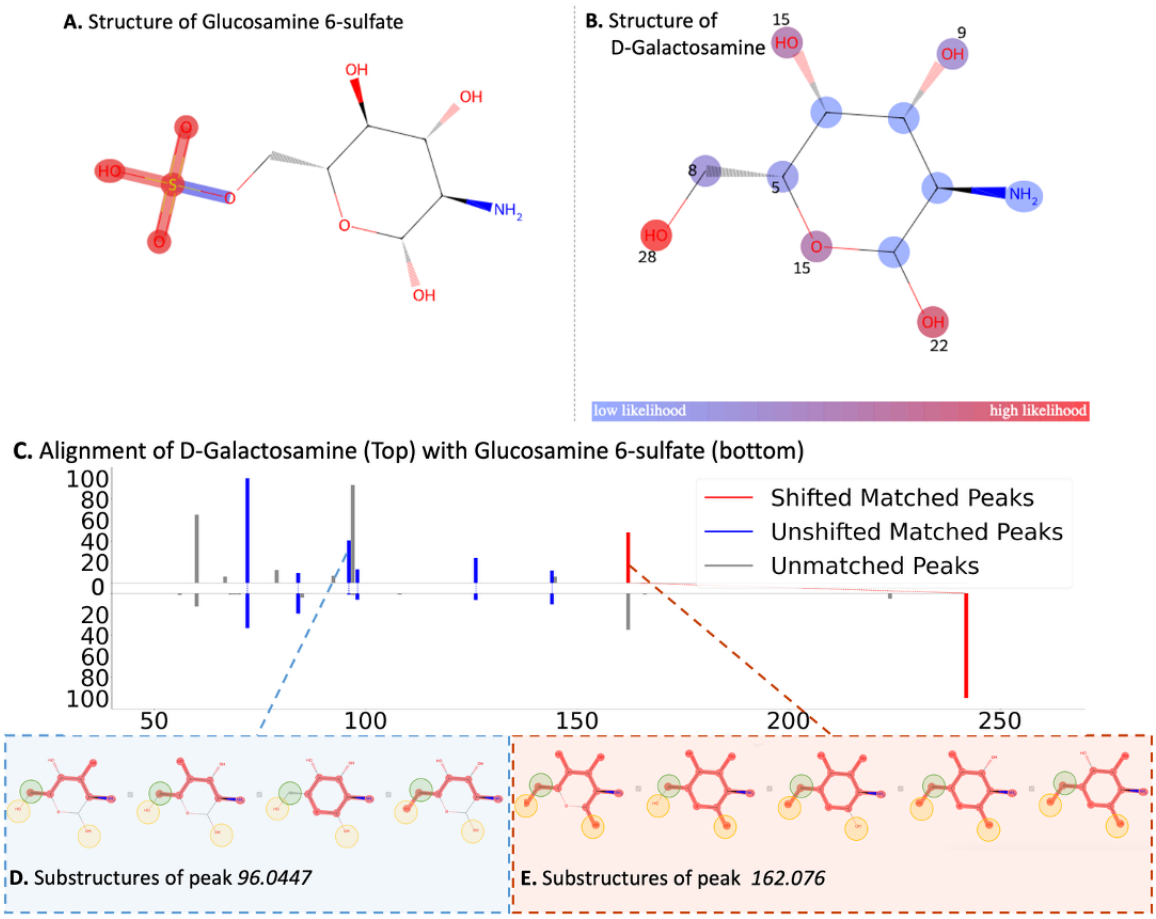


Figure 2.11: A. Structure of Glucosamine 6-sulfate modification to D-Galactosamine is highlighted. B. Structure of D-Galactosamine. C. Alignment of D-Galactosamine (Top) with Glucosamine 6-sulfate (bottom). D. Substructures of peak 96.0447 Interest points highlighted in yellow and true modification site highlighted in green E. Substructures of peak 162.076 m/z Interest points highlighted in yellow, True modification site highlighted in green.

identified by MS/MS library search (Fig. 2.12). An unknown compound (328.211 m/z) was prioritized as a putative analog through the neighboring network nodes Thr-C12:1 and Glu-C16:1. ModiFinder indicated that 328.211 m/z included a structure modification that reduced the lipid tail of Glu-C16:1. Corroborating this, ModiFinder additionally indicated that 328.211 m/z contained a modified amine headgroup compared to Thr-C12:1. Further manual inspection indicated the true structure was Glu-C12:1 – consistent with the ModiFinder localization predictions.

While the above example demonstrates the efficacy of ModiFinder on a real-world example, we recognize even with the integration of helper compounds and MS/MS peak formula refinement substructure annotation ambiguity may still remain. Therefore, we have enhanced our web interface to enable expert users to apply their domain knowledge to eliminate incorrect substructures for each MS/MS peak. The web interface then synthesizes this user input (or multiple user refinements) with ModiFinder to produce a refined likelihood distribution. As proof of principle, the web interface of ModiFinder and domain expertise were employed to solve the location of structural modifications of two natural products: Kirromycin and Naphthomycin B, two structurally complex natural product antibiotics[32, 78]. These compounds were selected as they exhibited high structural complexity and there existed a large diversity of structural analogs that remain unidentified. The first example demonstrates the ability to localize the N-methylation of Kirromycin that leads to its derivative Goldinodox (Fig. 2.13). In ModiFinder's initial prediction, the true modification site was among the highest scoring. However, these existed two regions on the structure with non-zero likelihood scores (Fig. 2.13-A). MF-R computationally assigned seven possible substructures to each

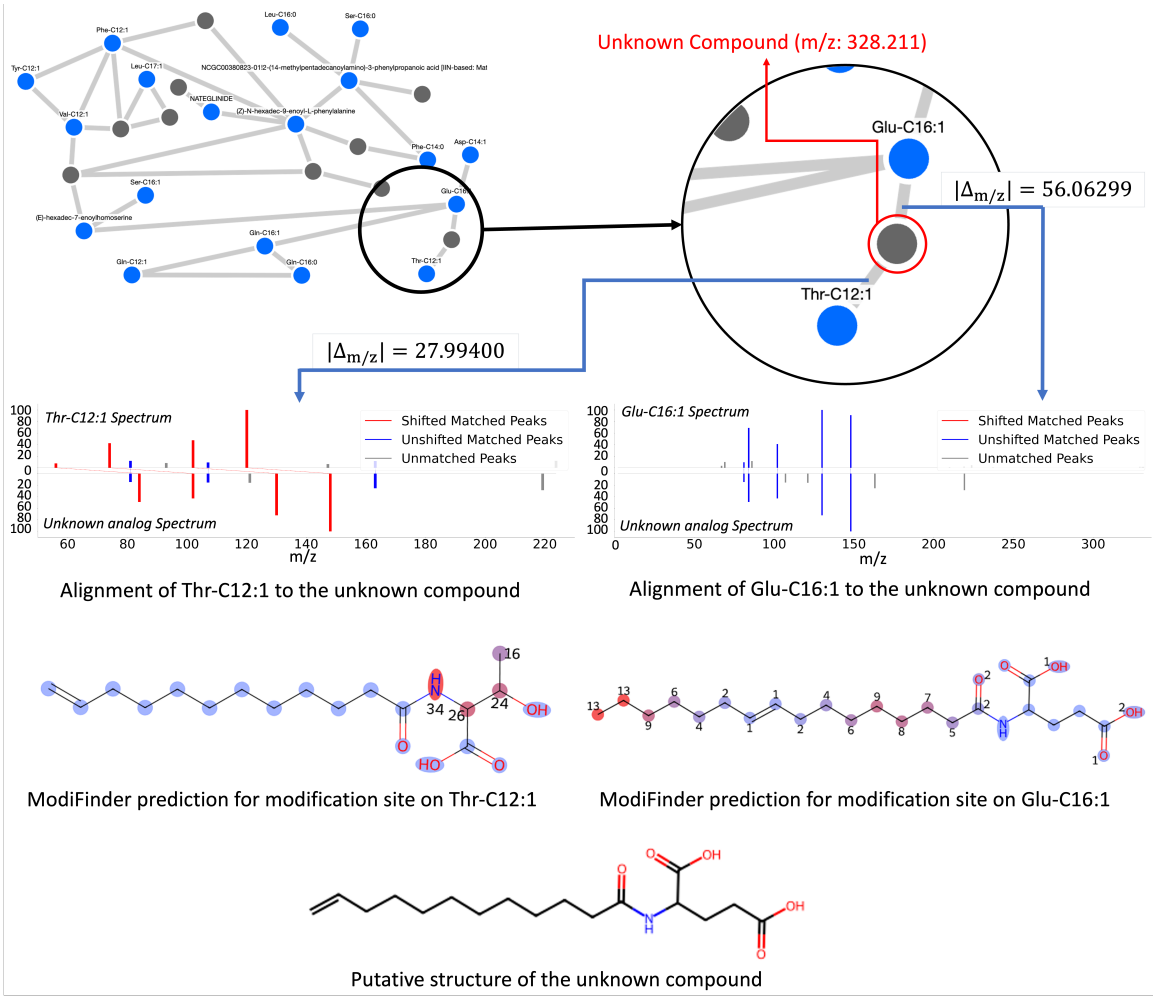


Figure 2.12: An unknown compound related to Thr-C12:1 and Glu-C16:1 is selected for annotation. The alignment of its MS/MS and ModiFinder's prediction are shown. ModiFinder assisted annotation led to the manual identification of 238.211 m/z as being Glu-C12:1, consistent with ModiFinder localization of modifications of Thr-C12:1 (amine headgroup modification) and Glu-C16:1 (lipid tail reduction).

of the 112.04 m/z and 178.05 m/z peaks. By taking the polarization of the neighboring bonds to the carbonyl as well as the numbers of (single) bonds to be broken into account, domain experts were able to limit the substructures assigned to the peak 112.04 m/z to one (visualized by the green dotted rectangle in Fig. 2.13). Similarly, the set of substructures assigned to the 178.05 m/z peak was also manually reduced to one, i.e. the unlikely events of double bond breaking, forming of terminal amides through alkene loss, or alkene side chain methyl cleavage were eliminated. This reduction in ambiguity improved the site localization of the methylation and increased the evaluation score from 0.86 to 0.93.

A second example demonstrates the methylation of Naphthomycin B (Fig. 2.14-A and B). The specific challenge of Naphthomycin B, is due to the cyclic 2D structure, which causes MS/MS fragments (133.07 m/z and 147.08 m/z) to be ambiguous between multiple substructures around the 2D cycle (Fig. 2.14-D). This ambiguity leads to 24 high scoring sites (atoms) across the compound. In the case of Naphthomycin B, taking into account the likely gas phase fragmentation site at the amide bond, the substructure ambiguity for 133.07 m/z and 147.08 m/z decreased from 27 and 40 substructures to 2 substructures each. This resulted in a decrease of 24 high scoring modification sites to two high scoring modification sites above the true modification site. Further, the site localization was narrowed to a single likelihood region for the potential methyl-carrying site. This manual refinement improved the evaluation score from 0.56 to 0.72. However, ModiFinder, even given this ambiguity reduction, reported the highest likelihood two atoms away from the true site.

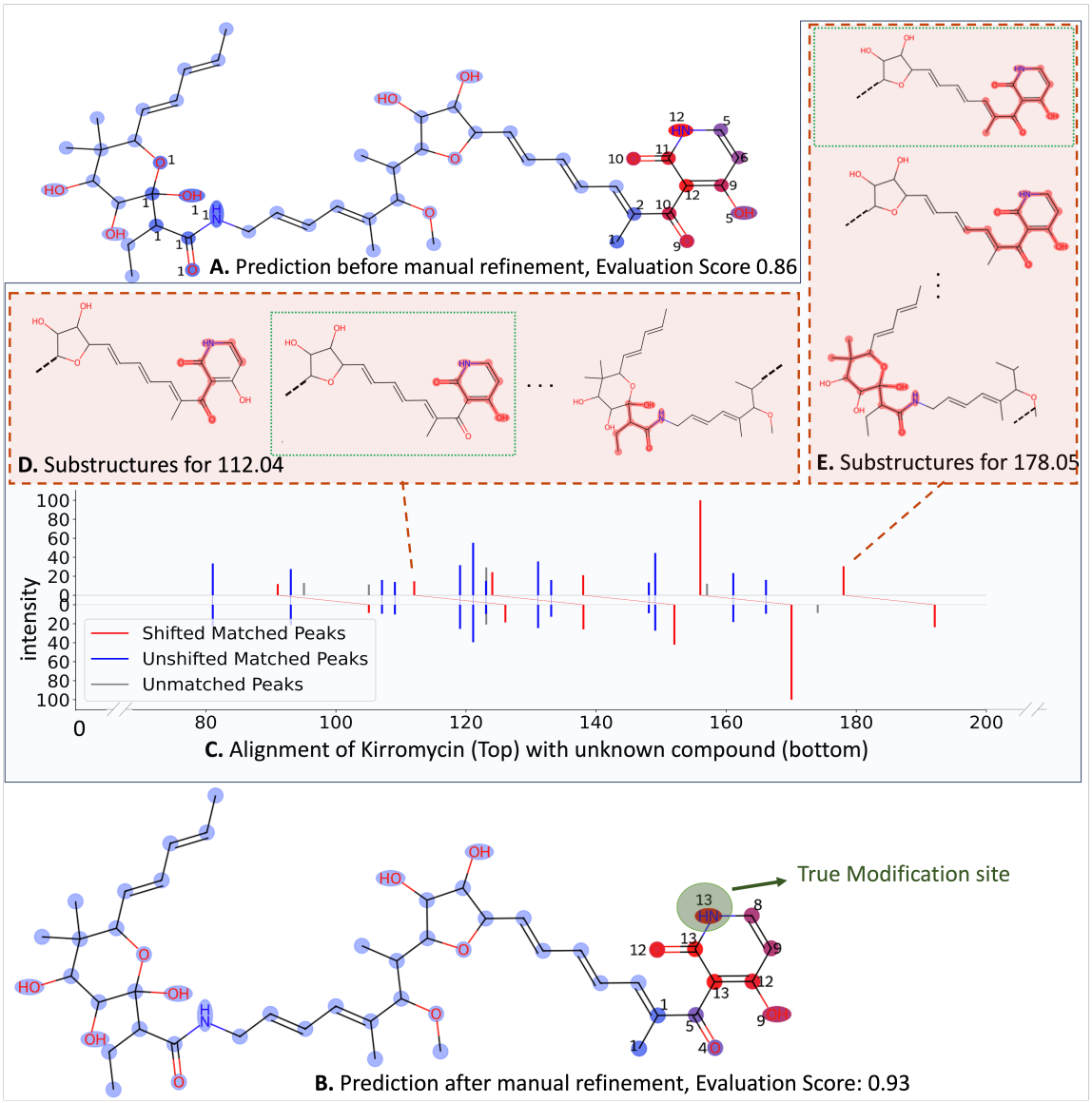


Figure 2.13: For the pair of Kirromycin and an unknown compound (Goldinodox), the initial prediction is improved by manually selecting the likely substructures for peak 112.04 m/z and 178.05 m/z improving the evaluation score from 0.86 to 0.93. A ModiFinder prediction before the refinement. B ModiFinder prediction after the refinement, the true modification site is highlighted by the green circle. C The alignment of Kirromycin and the unknown compound. The peaks of Kirromycin are shown at the top and the peaks of the unknown compound are shown at the bottom where shifted and unshifted peaks are highlighted with red and blue colors respectively. D, E shows the substructures assigned to peaks with 112.04 m/z and 178.05 m/z, due to the high number of substructures (ambiguity) only three substructures are shown. The green dotted box shows the substructure manually selected based on expert understanding of gas phase fragmentation.

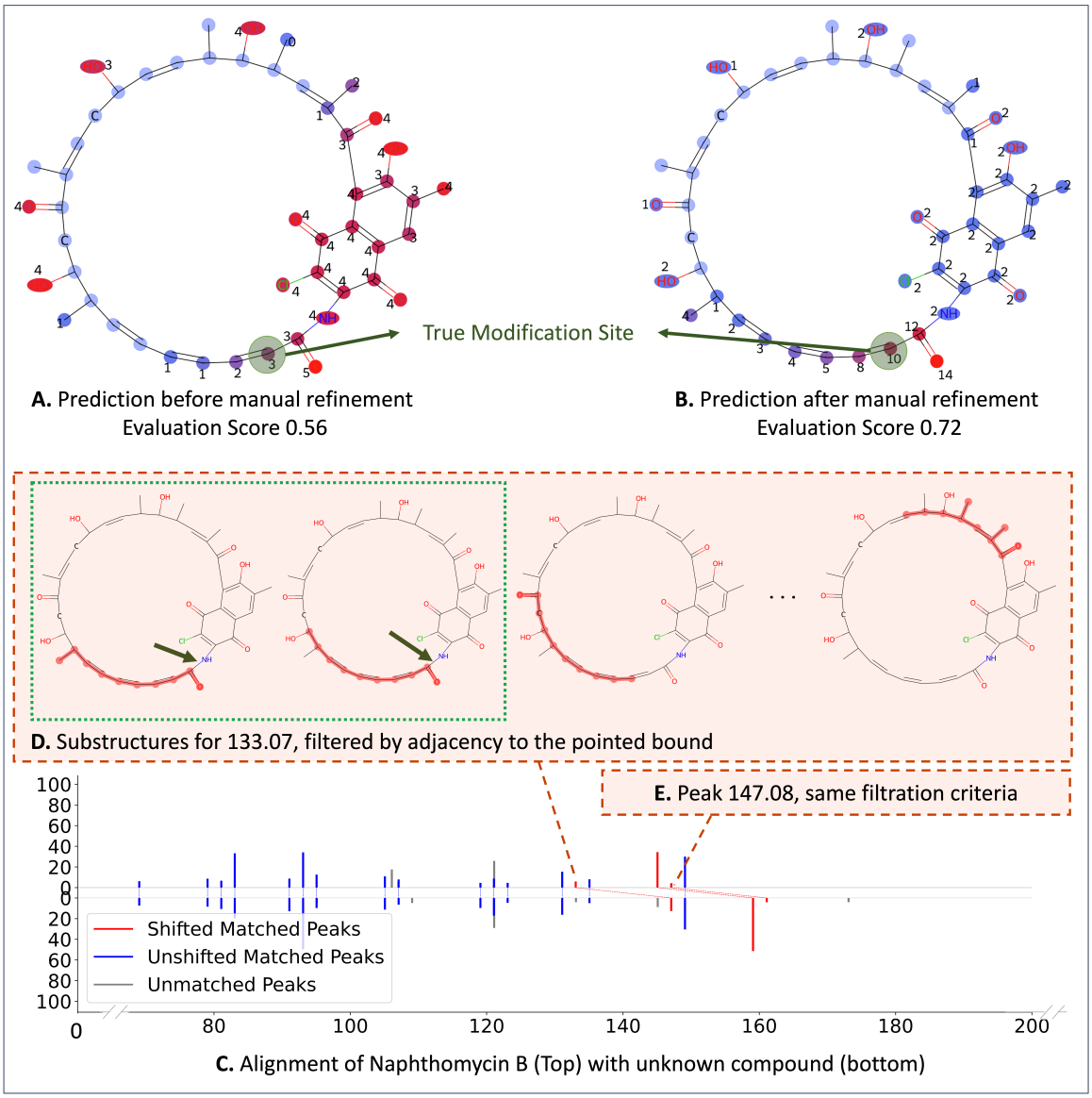


Figure 2.14: For the pair of Naphthomycin B and an unknown compound (Naphthomycin A), the initial prediction is improved by manually eliminating the unlikely substructures for peaks 133.07 m/z and 147.08 m/z, improving the evaluation score from 0.56 to 0.72. A, B ModiFinder prediction before and after the refinement, the modification site is highlighted by the green circle. C The alignment of Naphthomycin B and the unknown compound. The peaks of Naphthomycin B are shown at the top and the peaks of the unknown compound are shown at the bottom where shifted and unshifted peaks are highlighted with red and blue colors respectively. D shows the substructures assigned to peak with 133.07 m/z, due to the abundance of substructures (ambiguity) only four substructures are shown. The green dotted box shows the substructures manually selected based on expert understanding of gas phase fragmentation specific to amide bonds, which are indicated by green arrows. E. Same filtration applied to the peak with 147.08 m/z.

Chapter 3

Enhancing Site Localization with Multispectrum Data – Exploiting Collision Energies & Adducts

3.1 Motivation

The initial development of ModiFinder, as detailed in Chapter 2, established a novel and effective method for localizing chemical modifications. However, its performance was shown to be fundamentally limited by two key factors "insufficient peak data and fragmentation annotation ambiguities". In many cases, a single MS/MS spectrum acquired under one specific condition does not produce enough informative fragment peaks to confidently pinpoint a modification site. This chapter addresses this limitation directly by investigating how incorporating MS/MS spectra from multiple collision energies and mass spectrometry adducts

can provide complementary information to enhance ModiFinder's localization accuracy and expand its applicability. [51]

3.1.1 Methods for Data Integration

This investigation was made possible by leveraging a comprehensive dataset from Agilent Technologies, which, unlike many publicly available datasets, contained a large number of compounds acquired under consistent and wide-ranging experimental conditions.

We evaluated the results on two datasets provided by Agilent Technologies. namely Metlin, and Applied Markets comprising 10833, 6105 unique compounds from the Agilent MS/MS library, called Agilent library (Fig. 3.1). Our overarching goal was to identify a subset of compounds that enable fair comparisons across multiple collision energies and adducts. Although the Agilent MS/MS libraries provides spectra acquired at a range of collision energies, we focused specifically on 10, 20, and 40 V, since more than 98% of the data were collected under these three conditions (Fig. 3.1-C). Furthermore, although 18 unique adducts are present, we limited our analysis to [M+H] and [M-H], which account for more than 96% of the dataset (Fig. 3.1-B) (The distribution of compounds based on their mass is shown in figure 3.3). We call the compounds that have all the 6 scans - both adducts of interest ([M+H] and [M-H]) and the 3 collision energies (10, 20, and 40V) - the "Data-Rich" Compounds. There are 1203 and 947 Data-Rich compounds in Metlin and Applied Markets respectively. From these Data-Rich Compounds, we identified pairs of compounds differing by exactly one modification site (referred to as a "match" for ModiFinder evaluation), 539 (170) pairs were detected spanning 420 (248) compounds for the Metlin (Applied Markets) dataset.

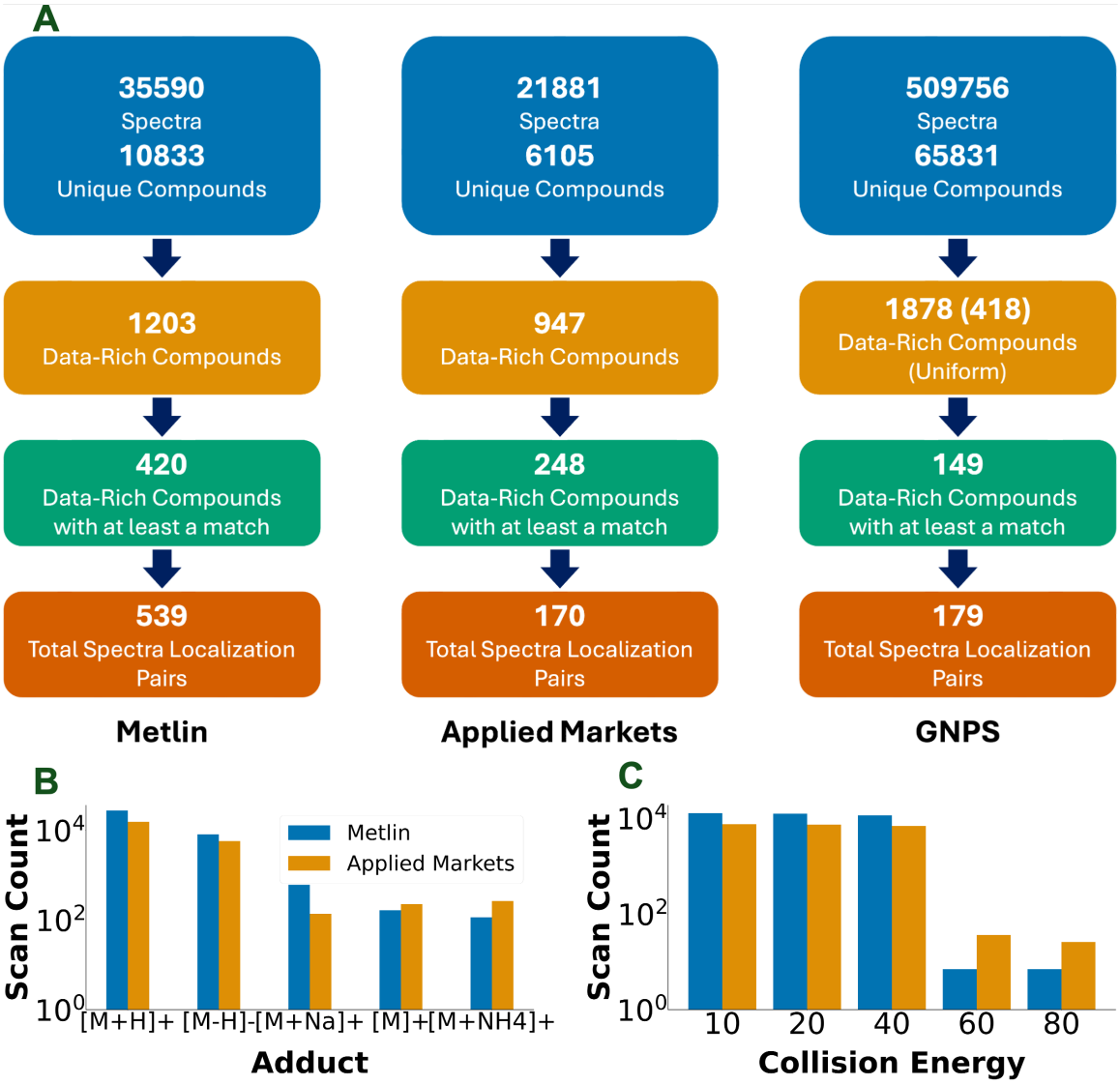


Figure 3.1: A. Showing the number of scans in each dataset, the number of unique compounds, and the number of Data-Rich unique compounds (Compounds with scans having both [M+H]+ and [M-H]- adduct and collision energy 10, 20, 40 V). A match for a compound is a compound that has one modification. B. The distribution of adducts in the datasets. C. The distribution of collision energies in the datasets.

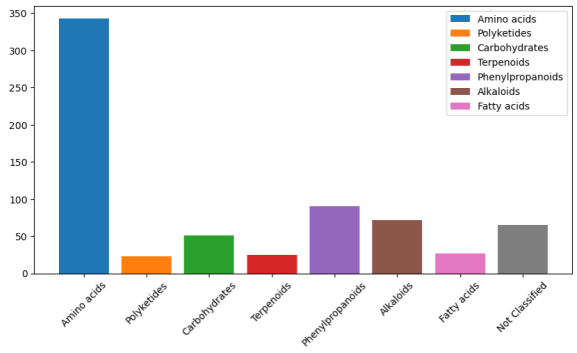


Figure 3.2: The compounds have a diverse range of classes from Amino acids to Carbohydrates and Fatty Acids.

In comparison, there are 179 pairs available in GNPS. Given the limited number of publicly available "Data-Rich" compounds, our analysis focuses exclusively on the Agilent dataset. Agilent not only provides significantly more data but also comes with detailed documentation of its collection and analysis methods, increasing our confidence in its accurate annotation. While the selected "Data-Rich" compounds that are selected to ensure a consistent basis for evaluating the impact of both collision energy and adduct polarity are a small set, to ensure diversity, the NPClassifier[33] was utilized to categorize the compounds by their primary class (e.g., alkaloids, carbohydrates, fatty acids, etc.). This analysis confirmed that our dataset spans a broad cross-section of chemical classes typically encountered in natural products (Fig. 3.2). In addition, we examined the precursor ion mass-to-charge ratios (m/z) and observed that the compounds range from approximately 100Da to more than 1200 Da, indicating a wide coverage of molecular weights (Fig. 3.3). As such, our selection captures both the structural variety and the physicochemical property space found in natural products, helping to ensure that our findings are broadly generalizable.

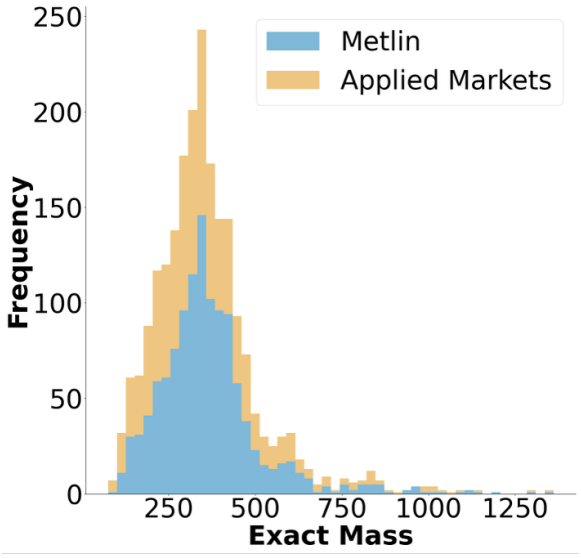


Figure 3.3: The compounds have a diverse range from less than 100Da to more than 1250Da.

Our findings indicate that increasing collision energies enhance ModiFinder's ability to analyze pairs previously unmanageable. However, the benefits plateau beyond a certain threshold due to increased ambiguity. Additionally, ModiFinder's performance on [M-H] ion pairs is generally inferior to that on [M+H] pairs, we attribute this to [M-H] ion pairs being more challenging to annotate. Lastly, employing an ensemble approach combining [M-H] and [M+H] substantially improves the pool of potential pairs ModiFinder can perform confidently.

3.1.2 Data Acquisition Methods

For GNPS, we used the cleaned data with available collision energy and adduct[60]. For Agilent, we used the raw data provided by Agilent Technologies. The data acquisition methods are described below. Accurate mass spectra were acquired using high purity single compound reference standards. All the data are acquired using Agilent LC/MS/MS Q-

TOFs with an Agilent HPLC using a strict Standard Operating Procedure (SOP), which includes instrument tuning and calibration as well as acquisition of quality control samples, ensuring that all data are acquired under the same experimental conditions. Compounds not available as single standards, such as fungal or bacterial metabolites, were synthesized and the compound of interest identified by chromatographic separation. LC/MS/MS accurate mass spectra were collected for the lowest m/z isotopic molecular ion in both positive and negative polarities, for multiple adduct ion species and collision energies. The data were acquired with flow injection or using a short column in Targeted MS/MS mode. To provide rich fragmentation spectra for compounds of varying stability, each compound was acquired at 10, 20, and 40 V collision energies. The spectra were curated using the MassHunter Qualitative Analysis Spectral Curation Workflow which utilizes the MassHunter Find by Formula Algorithm. [30, 21, 61, 20]

3.2 Strategies for Data Integration

3.2.1 Merged Peaks

To merge two spectra of the same compound acquired with the same adduct but different collision energies we first copy all the peaks from the smaller collision energy spectrum to the aggregate spectra. Then, we loop through the peaks of the data with the larger collision energy, for each peak pi, we first check if there is any peak within 10 ppm of this peak that exists in the aggregate spectra. If there is, we add the intensity to the current intensity. Otherwise, we add this peak to the aggregate spectrum with its original intensity. Finally, the aggregate spectrum is normalized.

3.2.2 Ensemble Learning, Merging the Predictions

Ensemble Learning, Merging the Predictions One effective approach for leveraging data from multiple experiments is ensemble learning. In this method, the likelihood predictions from different experiments are combined. To combine the predictions, first a weight is assigned to each prediction. The aggregate likelihood score for each atom is then calculated as the weighted average of its score across the experiments. The weights were chosen two different ways, the first was to assign the same weight to all the experiments, this is the most simple way of combining the predictions. The second method, which proved more effective, assigned weights based on the entropy of the predicted likelihood array. In this approach, higher entropy (indicating more dispersed and less focused predictions) results in lower weights, while lower entropy (indicating more focused predictions) leads to higher weights. The results presented in this manuscript use the entropy-based weighting method, as it generally produces better scores, though the equal-weight method shows a similar overall trend.

3.3 Experimental Evaluation

3.3.1 Effect of collision energy on ModiFinder performance

We evaluated the effect of collision energy on ModiFinder's localization performance. Broadly, as collision energy increases, the number of peaks rises, leading to an increased number of matched peaks and the number of annotated matched peaks. However, the number of annotated peaks alone is not enough of a measure for the performance of ModiFinder, our data shows that the Collision energy 20V scores higher than both 10V and 40V on both

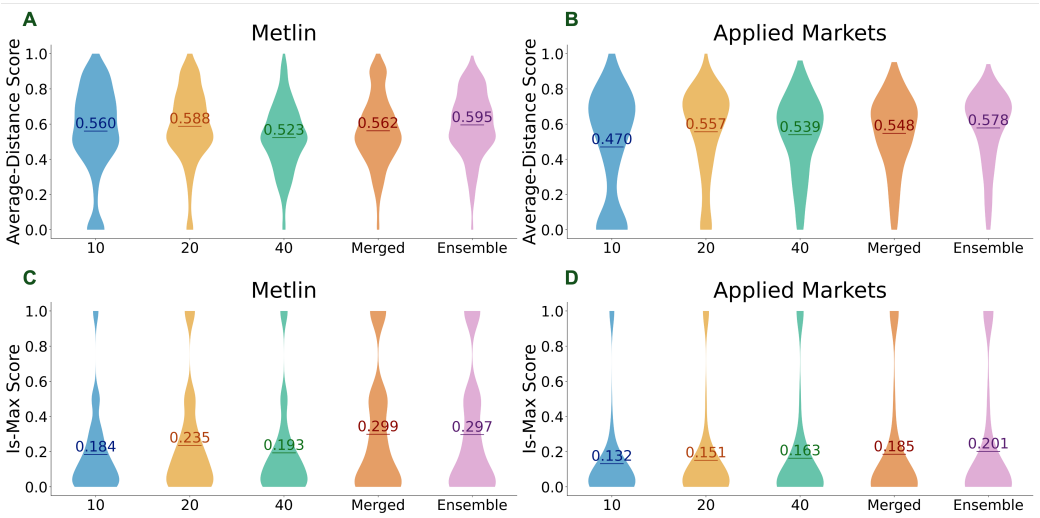


Figure 3.4: The result of the ModiFinder on different collision energies, Merged Peaks method and Ensemble. A, B The 20V performs better on average than 10V and 40V. C, D While the Average-Distance metric does not show significant improvements between 20V and Merged or Ensemble, the Is-Max metric shows significant improvement from Ensemble in comparison to other collision energies and the merged method.

Agilent datasets when using the Average-Distance metric [Fig. 3.4 - A & B]. By increasing the collision energy from 10V to 40V, we found that 165 compounds which could not localize a modification in 10V, were now localizing significantly better [Fig. 4]. Moreover, Figure 4 shows that no specific collision energy consistently yields better results with ModiFinder, though experiments show 20V performs the best on average [Fig. 3.4].

3.3.2 Effect of Merging the multiple collision energies

We hypothesized that multiple collision energies could be complementary in site localization performance. We merged MS/MS peaks from different collision energies of the same compound and adduct (see the methods section on "Merging the MS/MS Peaks for Different Collision Energies"). By naively merging 10, 20, and 40V into a single MS/MS spectrum, we did not observe a difference in localization performance using the average

distance metric [Fig. 3.4]. However, merging MS/MS spectra is shown to improve localization performance over the individual collision energy spectra when using the Is-Max evaluation method from 20V, up to 60 percent [Fig. 3.4]. This difference is rooted in the underlying construction of these methods. The Average-Distance method takes both ambiguity and proximity into account, while the Is-Max method focuses solely on the highest-scoring atoms in the localization—specifically, identifying the true modification site without considering the ambiguity from other closely scoring atoms scattered across the molecule. As more peaks are annotated, the probability of correctly identifying the modification site increases, leading to higher proximity scores. However, this also flattens the scoring distribution, increasing ambiguity and resulting in only modest changes to the average-distance score. However, the Is-Max method remains largely unaffected by this ambiguity as the additional annotated peaks typically include the true modification site. As a result, the Proximity score mostly increases, leading to a higher overall Is-Max score.

3.3.3 Ensemble Learning by Merging the Predictions

As an alternative approach to merging all energies into a single MS/MS spectrum, the ensemble approach aims to merge the localization predictions of several pairs of MS/MS spectra. Specifically, localization predictions were created for pairs of 10, 20 , and 40 V spectra separately, with the predictions merged. We observed that this ensemble approach improved the localization performance on the Is-Max method by up to 75% compared to its performance on individual collision energy spectra.

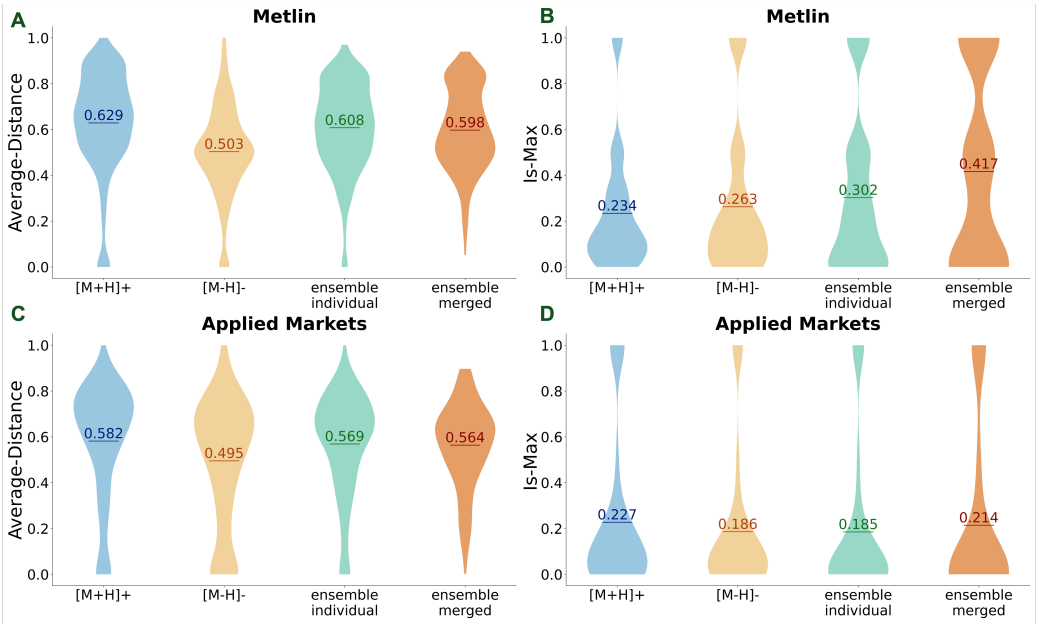


Figure 3.5: The result of ModiFinder for different adducts shows that ModiFinder performs better on positive mode. "Ensemble Individual" represents the predictions generated by merging results from the same collision energy but different polarities. The " Ensemble Merged" predictions are based on the combined MS/MS spectra from both polarities.

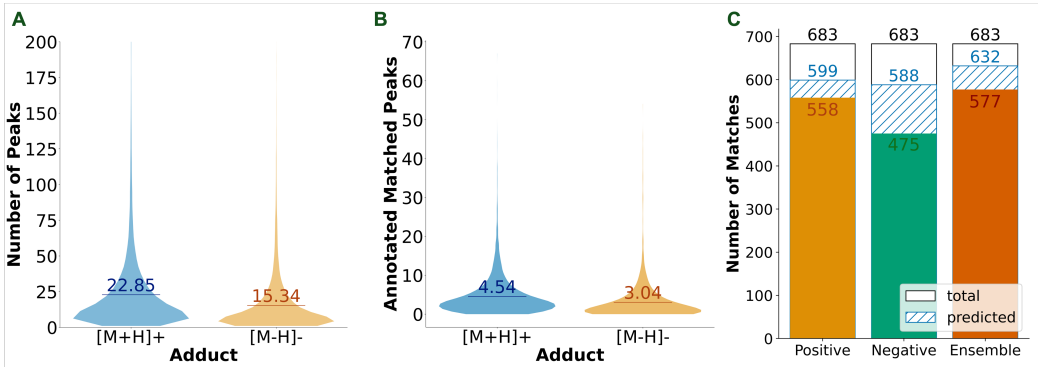


Figure 3.6: A The number of peaks B The number of annotated peaks based on the adduct. C Color Filled bars show the of the number of pairs where ModiFinder was able to outperform the random baseline. Out of all 683 pairs of compounds, Positive mode is able to outperform random localization more than Negative mode. Further, the Ensemble approach integrating Positive and Negative increases the number of MS/MS pairs that can improve over random localization.

3.3.4 Polarity Adducts Effect on ModiFinder Performance

We investigated the complementarity of different adducts in the localization of modification sites with ModiFinder. Specifically, we used [M+H] and [M-H] as the most prevalent adducts in the dataset. Initial evaluation of ModiFinder's performance given [M+H] and [M-H] adducts shows that ModiFinder performs considerably better in positive mode in Average-Distance Metric (Fig. 3.5). This difference in performance can be attributed in part to the negative mode having fewer peaks and fewer annotated matched peaks (Fig. 3.6). Although having more annotated matched peaks can sometimes lead to increased ambiguity - hence, not improving the evaluation score-, having fewer than a certain threshold of these peaks means there is less critical information for ModiFinder to utilize effectively.

3.4 Performance Across Different Chemical Classes

Using the NPClassifier tool to categorize compounds, we analyzed ModiFinder's performance across different chemical classes and uncovered insightful patterns.

Carbohydrates achieved the highest Average-Distance score. We hypothesize this is because their distinct ring structures and branching points allow ModiFinder to correctly identify the general region of modification, leading to a strong score. However, they ranked lower on the Is-Max metric, as pinpointing the exact atom within these complex, similar-looking regions remains difficult.

Fatty Acids, in contrast, achieved the highest Is-Max score. Their long, linear chains present fewer structurally distinct regions, so a modification is often either correctly identified (a perfect score) or is substantially distant, which can be heavily penalized in the Average-Distance score.

3.5 Conclusion

This investigation demonstrates that increasing the amount of spectral information by using data from different collision energies and adducts is a beneficial strategy for modification site localization. While this does not always lead to a higher evaluation score for every compound pair—as additional data can sometimes introduce more ambiguity—the primary benefit is a significant enhancement in the robustness and overall applicability of ModiFinder. The use of multispectrum data allows the tool to provide confident predictions for a wider range of compounds, particularly for those that are unanalyzable with a single spectrum. This work confirms that while more data of the same type is helpful, achieving the next

level of performance improvement will likely require the incorporation of drastically different fragmentation mechanisms or, as explored in the next chapter, entirely new computational paradigms to better resolve structural ambiguity.

Chapter 4

ModiStruct – Diffusion-Driven Molecular Inpainting with Substructure Constraints

4.1 Motivation

The methods presented in the previous chapters successfully established a framework for localizing chemical modifications on a known molecular scaffold. While this addresses a critical gap in annotating structural analogs, the ultimate goal in computational mass spectrometry remains the full de novo elucidation of a completely unknown molecular structure. Traditional de novo methods face immense computational hurdles due to the vastness of chemical space. This chapter introduces ModiStruct, a novel generative model that makes this ambitious goal more tractable by reframing the problem. Instead of generating

a molecule from scratch, ModiStruct treats the task as **molecular inpainting**. Given a known substructure—identified via a library search—and the MS/MS spectra of both the substructure and the full molecule, ModiStruct "inpaints" the missing atoms and bonds to reconstruct the complete molecular graph.

This approach requires a significant leap in modeling capability beyond the methods used for ModiFinder. To tackle this generative challenge, we turned to discrete denoising diffusion models, a powerful class of deep learning models that have achieved state-of-the-art results in generating complex, structured data like molecular graphs. ModiStruct exploits a key characteristic of molecular behavior: similar molecules tend to produce similar MS/MS spectra. By first identifying compounds in public libraries with spectra resembling those of the unknown samples, **ModiStruct** leverages the known structures of these similar compounds. This strategy allows ModiStruct to focus on learning the **Modi**fications needed to con**Struct** the complete molecule, rather than having to capture the complete underlying representation from sparse data. ModiStruct is a discrete graph inpainting diffusion model tailored for molecular structure prediction from tandem MS/MS data under known substructure constraints. Unlike prior spectrum-to-structure methods that generate complete molecules from a single spectrum, ModiStruct leverages two complementary spectra, one for the unknown molecule and another for a known substructure of the unknown molecule, along with the structure of this known substructure. The model operates in two key stages:

1. **Joint Spectrum–Structure Encoding:** A cross-modal encoder aligns spectral embeddings from both the known and unknown compounds with the structural embedding of the known substructure. This produces a molecular fingerprint prediction and

localizes probable modification sites.

2. **Conditioned Graph Diffusion Decoding:** A discrete graph diffusion process reconstructs the complete molecular graph, constrained by the given substructure scaffold and informed by modification-site probabilities, ensuring chemically consistent completions.

By integrating spectrum-aware conditioning, explicit substructure masking, and modification-site localization into the generative process, ModiStruct significantly reduces the search space and improves accuracy where substructures are available in public spectral libraries.

The main contributions of this work are summarized as follows:

- ModiStruct, The first MS/MS-based model to jointly encode spectra from a molecule and its known substructure, enabling scaffold-consistent inpainting of missing regions.
- We introduce a discrete graph diffusion process guided by predicted attachment points, improving chemical validity and localization accuracy. This is combined with a pretraining framework that trains the diffusion decoder on millions of synthetically generated molecules and substructures.
- State-of-the-art performance with significant improvements over prior de novo and scaffold-conditioned baselines on established benchmarks, achieving higher top-k accuracy and Tanimoto similarity. Additionally, our auxiliary modification-site prediction task outperforms state-of-the-art methods by up to 5× on the "Is-Max" and "Average-Distance" metrics.

4.2 Discrete Denoising Diffusion for Graphs

For the ModiStruct, diffusion models [56, 28, 55] are utilized as part of the model. Diffusion models are generative models that learn the distribution of the data p_data given a training set. In the forward process, a sample $\boldsymbol{z} \in p_{data}$ is progressively noised with noise model q such that it is transformed to a known prior p_{init} space in T steps. A neural network is trained to model the reverse process. At the inference time, the model samples $\boldsymbol{z}^T \in p_{init}$ and iteratively sample $p_\theta(\boldsymbol{z^{t-1}}|\boldsymbol{z^t})$ until reaching $\boldsymbol{z}^0$.

For discrete diffusion, [2] proposes that for a data point z that belongs to one of d classes and $\boldsymbol{z}$ is the one hot encoding for that, the noise at step t is defined as $q(z^t|z^{t-1}) = \boldsymbol{z}^{t-1}\boldsymbol{Q}^t$ where the $\boldsymbol{Q}^t_{i,j}$ is the probability of jumping from state i to state j.

For diffusion models over graphs, Digress [70] adapts this discrete setup and applies the discrete diffusion separately on each node ($\boldsymbol{X}$) and each edge ($\boldsymbol{A}$) feature. For any edge (e) and node(v), the transition probabilities are defined by the matrices $[Q^t_{edge}]_{i,j} = q(e^t = j|e^{t-1} = i)$ and $[Q^t_{node}]_{i,j} = q(v^t = j|v^{t-1} = i)$. Therefore $q(G^t|G^{t-1}) = (X^{t-1}Q^t_{node}, E^{t-1}Q^t_E)$.

4.3 The ModiStruct Architecture

This section first formalizes the prediction task, then details the ModiStruct architecture, an encoder–decoder model that combines a spectrum/formula encoder and a graph encoder with a discrete graph-diffusion decoder.

4.3.1 Problem statement

Assume an *unknown* molecule-spectrum pair $M = (G_M, S_M)$, where $G_M = (V, E, \mathbf{X}, \mathbf{A})$ is the Molecule Graph defined in Chapter 1.5, and S_M is its MS/MS spectrum. Given an *analog* compound $H = (G_H, S_H)$ retrieved from a spectral-library search such that $G_H \subseteq G_M$, the goal is to predict the full target graph G_M conditioned on the tuple (S_M, S_H, G_H). Formally, we learn $p_\theta(G_M \mid S_M, G_H, S_H)$.

4.3.2 Overview of MODISTRUCT

MODISTRUCT follows an **encoder–decoder** design:

1. **Encoder**: The encoder consist of multiple modules.
   (a) **Individual Encoders:** independent encoders embed the spectrum, fingerprint (if provided), and substructure graph into feature vectors.
   (b) **Global Context:** these features are concatenated and passed through a multi-layer perceptron (MLP) to produce a global context vector $\boldsymbol{r}$.
   (c) **Feature Modulation and Graph Refinement:** a Feature-wise Linear Modulation (FiLM) [42] layer conditions node embeddings of the substructure graph on $\boldsymbol{r}$. A Graph Attention Network (GAT) refines the modulated node embeddings using neighborhood structure.
   (d) **Prediction Heads:** two task-specific branches predict (i) node-level modification logits, and (ii) the 2048-bit delta fingerprint vector.
2. **Decoder**: a discrete denoising diffusion model that reconstructs the unknown bond

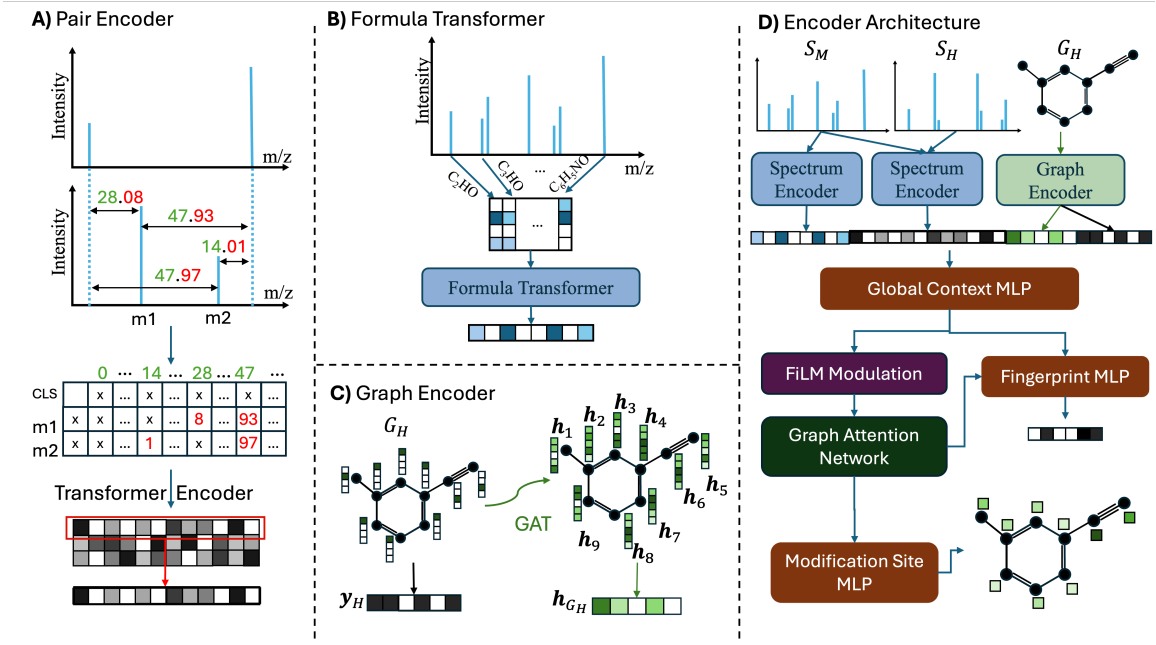


Figure 4.1: This figure provides an overview of the ModiStruct encoder architecture. (A) Pair Encoder: The differences between the pairs are calculated B): A formula transformer processes the input spectra to generate embeddings. (B) Graph Encoder: A Graph Attention Network (GAT)[68] is used to encode the topology of the substructure graph This produces initial node embeddings which are then aggregated to create a fixed-length vector representation of the substructure. (C) The Main Encoder: The encoder integrates multiple inputs to create a comprehensive representation. (D) The global context vector $\mathbf{r}$ is combined with each node embedding h_v from the substructure graph. This new embedding is then processed and refined by a GAT and then each embedding is passed through an MLP to determine the probability that each node is a modification site.

tensor $\mathbf{A}_M$ (and thus G_M). The decoder uses $\hat{\mathbf{y}}_M$, $\boldsymbol{X}_M$, the substructure mask $\boldsymbol{A}'_{G_M,G_H}$, and modification sites $C_{(G_M,G_H)}$ as conditional inputs.

The model is trained end-to-end with a multi-task loss combining diffusion, and modification-site objectives.

4.3.3 Encoder

Spectrum, Fingerprint, and Graph Encoders For each peak $s_i = (m_i, y_i) \in S$ an elemental-composition vector is computed via SIRIUS [19]. These peak-level embeddings are then passed to a *formula transformer* (identical to MIST [26]) to obtain spectrum representations $\boldsymbol{s}_M$. To capture the differences between the two spectrums, we also use a pair spectrum encoder following the TransExION [7] encoder architexture.

Graph encoder maps the substructure graph G_H to node embeddings $\{\boldsymbol{h}_v\}_{v \in V_H}$ and a pooled subgraph vector $\boldsymbol{h}_{G_H}$.

Global Context The embeddings are concatenated and mapped by an MLP to a fixed-dimensional global context vector $\boldsymbol{r}$.

$$\boldsymbol{r} = mlp([\, \boldsymbol{s} \,\|\, \boldsymbol{y}_H \,\|\, \boldsymbol{h}_{G_H} \,]),$$

Feature Modulation (FiLM) and Graph Refinement The global context $\boldsymbol{r}$ is linearly projected to produce scale $\boldsymbol{\gamma}$ and shift $\boldsymbol{\beta}$ parameters. These modulate each node embedding $\boldsymbol{h}_v$ of G_H as

$$\boldsymbol{h}'_v = \boldsymbol{\gamma}_{g(v)} \odot \boldsymbol{h}_v + \boldsymbol{\beta}_{g(v)},$$

where $g(v)$ denotes the graph index in the batch. The modulated embeddings $\boldsymbol{h}'_v$ are further processed by a multi-layer Graph Attention Network (GAT), producing refined node embeddings $\boldsymbol{h}''_v$.

Auxiliary Heads

1. **Modification-site head:** A node-wise MLP maps each $\boldsymbol{h}_v''$ to a scalar logit $\hat{m}_v \in \mathbb{R}$, indicating how likely node v is a modification site.

2. **Fingerprint head:** The refined graph embedding is obtained by pooling $\{\boldsymbol{h}_v''\}$, concatenated with $\boldsymbol{r}$, and mapped through an MLP to produce the predicted delta fingerprint vector $\hat{\boldsymbol{y}} \in [0,1]^{2048}$.

4.3.4 Discrete graph diffusion decoder

The decoder is a discrete denoising diffusion model that generates the full molecular graph G_M conditioned on the encoder's outputs. We build on the DiffMS framework [6], which adapts the discrete diffusion model DiGress [70] to molecule graphs. ModiStruct only diffuses the *edge* tensor. The corrupted adjacency tensor at timestep t can be calculated as $q(\mathbf{A}^t|A^{t-1}) = \mathbf{A}^{t-1}\mathbf{Q}^t_{\text{edge}}$, Since the noise is described by a markovian process, this can be rewritten as

$$q(A^t|A) = A\bar{\boldsymbol{Q}}^t, \bar{Q}^t = \boldsymbol{Q}^1\boldsymbol{Q}^2...\boldsymbol{Q}^t.$$

with class-transition matrices $\mathbf{Q}^t_{\text{edge}}$ fixed as in [70] with the following noise schedule:

$$\bar{Q}^t = \bar{\alpha}^t I + \bar{\beta}^t 1_k \boldsymbol{m}^t$$

where $\boldsymbol{m}$ is the marginal distribution of edge types in the training dataset and $\boldsymbol{m}^T$ is the transpose of $\boldsymbol{m}$.

Node features $\boldsymbol{X}_M$ remain fixed because the molecular formula, computed from S_M via SIRIUS, is assumed known at inference. The reverse diffusion is modeled by a graph neural

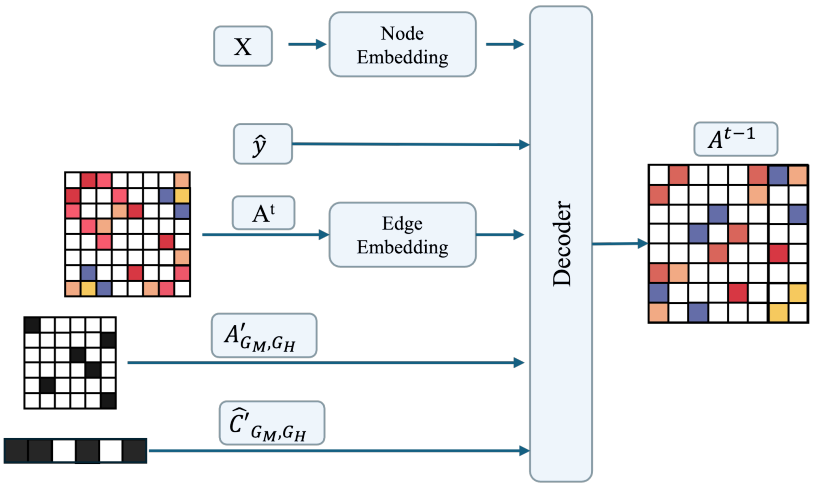


Figure 4.2: Overview of the decoder

network f_θ that attempts to iteratively denoise the adjacency. At each diffusion step t, the model f_θ takes as input

1. The current adjacency A^t,
2. The fixed node feature matrix $\mathbf{X}_M$,
3. The substructure mask $\boldsymbol{A'}_{G_M,G_H}$, and
4. The context vector $\hat{\boldsymbol{y}}_M$ and predicted modification sites $\hat{C}_{G_M,G_H}$ from the encoder

and outputs a prediction for the distribution of A^{t-1}. In essence, $\hat{\boldsymbol{y}}_M, \hat{C_{G_M,G_H}}$ serve as a conditioning signal that guides the generative process towards structures consistent with the spectrum S_M and substructure G_H. We pre-train the parameters of f_θ over the same dataset as Diffms (which was trained on a large corpus of molecules with ground-truth fingerprints as conditional inputs), allowing us to leverage its learned chemical priors and accelerate convergence. The diffusion loss $\mathcal{L}_{\text{diff}}$ is the KL term derived in [70].

4.3.5 Pre-training and fine-tuning

Encoder Trained on the training data with the target tasks of fingerprint prediction $\mathbf{y}_G$ and modification sites prediction with the loss

$$\mathcal{L} = \lambda_{\text{ms}}\mathcal{L}_{\text{ms}} + \lambda_{\text{fp}}\mathcal{L}_{\text{fp}}$$

Decoder: trained on publicly available dataset of molecule structures. For each input, a substructure is randomly generated. For the $\hat{y}$ and $\hat{C}$ and all the other required inputs for the decoder are calculated and passed to the diffusion model with loss

$$\mathcal{L} = \lambda_{\text{diff}}\mathcal{L}_{\text{diff}}$$

Fine-tuning: both modules are trained jointly with loss

$$\mathcal{L} = \lambda_{\text{ms}}\mathcal{L}_{\text{ms}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}}$$

4.3.6 Inference and sampling with inpainting

For sampling, we set $A^T \sim m$ where m is the marginal distribution of edges in training dataset and iteratively denoising to get to A^0. At each step of denoising, to apply the scaffold and make sure to maintain the substructure (inpainting), the RePaint [35] strategy is used: after every denoising steps, the current sample is blended with the noisy masked substructure. Figure 4.3 illustrates one step of the repainting procedure. Generation continues until $t = 0$, at which point the discrete bond types yield the final graph G_M.

$$\boldsymbol{A}_{G_H}^{t-1} = q(A_{G_H}^{t-1}|A_{G_H}) \tag{4.1}$$

$$\boldsymbol{A}_{G_M}^{t-1} = \underbrace{\boldsymbol{A}'_{G_M,G_H} \odot \boldsymbol{A}_{G_H}^{t-1}}_{\text{known}} + \underbrace{(1 - \boldsymbol{A}'_{G_M,G_H}) \odot f_\theta(\boldsymbol{A}_{G_M}^T, \boldsymbol{X}_{G_M}, \boldsymbol{y}_{G_M}, \boldsymbol{A}'_{G_M,G_H})}_{\text{unknown}} \tag{4.2}$$

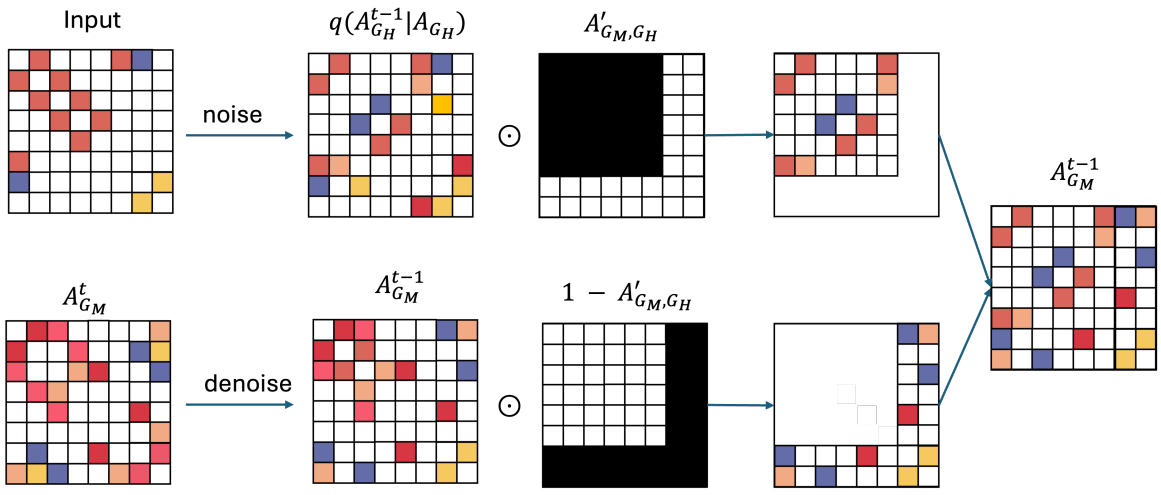


Figure 4.3: RePaint [35] strategy on the adjacency matrix during sampling

4.3.7 Oracle

Conventional de novo methods aim to reconstruct an entire molecular structure directly from a single MS/MS spectrum. The workflow presented here needs the substructure molecule and therefore, starts by retrieving candidate substructures through a library search, ranked by modified-cosine similarity. Then an oracle model, uses the true structures of the molecules to filter the ones that are not substructures. The substructure with highest tanimoto is passed to the model. To maintain a fair comparison, the same information is passed to the baselines.

The combination of spectrum-aware context, substructure masking, and discrete graph diffusion enables ModiStruct to reconstruct full molecular graphs from sparse MS/MS data while exploiting structural priors present in public libraries.

4.4 Experimental Evaluation

To train and evaluate our methods, we use both the MassSpecGym [8] and NPLIB1(CANOPUS) datasets [17]. The NPLIB1 dataset is the subset of GNPS data used to train the CANOPUS tool.

4.4.1 Baselines

To provide a comparison, the performance of MADGEN[75] and [6] is reported on the same datasets. Additionally, two other baselines are developed for comparison.

- Diffms (RePaint): This baseline follows the exact same training as DiffMS. Here, the reported weights from the paper are used as the weights of the model. However, during the sampling step, the RePaint step is applied. The advantage of this model in comparison to DiffMS is that it will use the substructure information and inpaint around it.

- MSE-DD: This uses **M**odi**S**truct's **Encoder** with **D**iffMS**D**ecoder. During training the same encoder as ModiStruct, generates the $\hat{y}$ value and is passed to the DiffMS decoder. During sampling, the same Repain step is applied. The advantage of this model is that it uses the information stored in the MS/MS and the graph of H. It will also use the substructure during the inpainting.

4.4.2 Overall De Novo Structure–Prediction Accuracy

For the De Novo task. we use the common defined metrics in the field:

- Top-k accuracy: measures whether the true molecule is in the top-k model predictions.

- Top-k maximum Tanimoto similarity: the structural similarity of the closest molecule to the true molecule in the top-k predictions.

We report metrics for k = 1. For the preliminary result reported here, we use a fraction of the test data.

Table 4.1 compares **ModiStruct** with strong baselines on the MassSpecGym and NPLIB1 test sets. Across both corpora, ModiStruct yields the highest accuracy and Tanimoto. It outperforms DiffMS by 3.0 percentage points in top-1 accuracy on MassSpecGym (5.3% vs. 2.3%) and by 7.76 percentage points on NPLIB1 (16.1% vs. 8.34%). When the substructure verifier is replaced with an oracle (ModiStruct-O), accuracy improves marginally by 0.1 percentage points on Gym and remains unchanged on NPLIB1. However, the average Tanimoto similarity improves to 0.45 and 0.46, respectively, indicating that substructure retrieval remains the primary bottleneck in the current pipeline.

Table 4.1: Overall performance (%).

Model	Accuracy ↑		Avg. Tan.↑	
	Gym	NPLIB1	Gym	NPLIB1
MADGEN	0.08%	1.0%	-	-
DiffMS	2.30%	8.34%	0.28	0.35
DiffMS (RePaint)	2.5%	11.21%	0.31	0.39
MSE-DD	2.6%	11.1%	0.35	0.4
ModiStruct	5.3%	16.1%	0.41	0.44
ModiStruct-O	5.4%	16.1%	0.45	0.46

4.4.3 ModiStruct Modification-Site Localization

For the Modification sites task, we use the same metrics $\{IsMax,\ Average - Distance$ defined in ModiFinder [52]. Since these losses are only designed for single mod-

Table 4.2: Single-site localization results (higher IsMax and AvgDist are better).

Model	IsMax ↑	AvgDist ↑
Random Guess	0.08	0.4
ModiFinder	0.13	0.61
ModiStruct	**0.48**	**0.67**

ification tasks, we select the pairs that have exactly one node as the modification site $|C_{G_M,G_H}| = 1$.

For spectra–substructure pairs containing a single modification site, Table 4.2 reports the $IsMax$ and $Average - Distance$ metrics. ModiStruct significantly outperforms the ModiFinder pipeline, achieving an $Is - Max$ of 0.48, representing a +0.35 absolute improvement over ModiFinder, and an $Average - Distance$ of 0.67, an increase of +0.06. These results highlight the benefit of joint structure generation and localization over sequential pipelines.

4.4.4 Fingerprint Prediction analysis

We evaluate our proposed dual encoder against the baseline model MIST by comparing the Tanimoto similarity of predicted molecular fingerprints. First, we trained MIST and directly compared its predicted fingerprints to those of our encoder. To isolate the contributions of different components, we also conducted ablation studies. In particular, we tested whether the auxiliary ModiFinder task improves performance, and whether predicting the *delta fingerprint* (the difference between the target fingerprint and the known analog's fingerprint) is advantageous compared to predicting the full fingerprint directly. Our encoder is trained to predict this delta representation; during inference, the predicted delta

is combined with the fingerprint of the known molecule via bitwise XOR to reconstruct the fingerprint of the unknown compound. Results are summarized in Table 4.3.

Table 4.3: Comparison of Tanimoto similarity scores between baseline and proposed models for fingerprint prediction.

Model	Tanimoto Similarity	
	Gym	NPLIB1
MIST	0.18	0.35
ModiStruct (No ModiFinder)	0.62	0.58
ModiStruct (No Delta)	0.23	0.47
ModiStruct (Full)	**0.63**	**0.60**

Chapter 5

Conclusions and Discussions

This dissertation has chronicled a research journey aimed at addressing fundamental challenges in the computational elucidation of small molecules from tandem mass spectrometry (MS/MS) data. The work progressed from developing a targeted tool for a specific analytical gap to creating a state-of-the-art generative model for the ambitious task of *de novo* structure prediction. This chapter synthesizes the key contributions, reflects on the overarching themes, and outlines promising directions for future research.

5.1 Summary of Contributions

The research provided here was guided by a series of questions aimed at automating and advancing the interpretation of MS/MS spectra for unknown compounds. We have provided robust answers to these questions through three primary contributions:

1. **Automated Modification Site Localization:** To address the inability of analog search methods to pinpoint specific structural differences, we developed **ModiFinder**,

a novel computational tool that localizes chemical modifications by aligning the MS/MS spectra of known and unknown analogs. By systematically rewarding atoms in fragments corresponding to shifted peaks and penalizing those in unshifted peaks, ModiFinder provides a quantitative likelihood map of the modification site, automating a previously manual and time-consuming task.

2. **Enhanced Robustness Through Multispectrum Data:** We demonstrated that the performance and applicability of ModiFinder could be significantly enhanced by incorporating richer spectral data. By evaluating the tool on a comprehensive dataset with spectra acquired at multiple collision energies and in different adduct polarities, we found that this complementary information expands ModiFinder's coverage to more compounds and improves its overall predictive confidence, making it a more robust and practical tool for real-world applications.

3. **De Novo Generation via Molecular Inpainting:** Finally, we took the significant leap from localization to full structure generation by developing **ModiStruct**, a diffusion-driven generative model. By framing *de novo* prediction as a molecular inpainting task, ModiStruct leverages a known substructure as a powerful conditional prior to reconstruct the complete molecular graph. This approach achieved state-of-the-art results, showcasing a new and highly effective paradigm for conditioned molecular generation from MS/MS data.

5.2 Synthesis: The Power of Structural Priors in Mass Spectrometry

A central, unifying theme throughout this dissertation is the power of leveraging known structural information to constrain an otherwise intractable search space. Both ModiFinder and ModiStruct are built on the premise that using a known, structurally related analog as a starting point is a far more effective strategy than attempting to solve for a structure from a single spectrum in isolation.

This body of work presents a methodological evolution in how such priors can be utilized. **ModiFinder** represents an explicit, alignment-based approach. It relies on the clear, interpretable signals of shifted and unshifted peaks to systematically deduce a modification's location. In contrast, **ModiStruct** represents an implicit, end-to-end learning approach. It uses a powerful cross-modal encoder and a generative diffusion model to learn the complex relationships between spectral features, substructure constraints, and the final molecular graph. The progression from the former to the latter mirrors the broader evolution in the field from rule-based systems to deep learning, demonstrating that as models become more powerful, they can leverage the same foundational priors in increasingly sophisticated ways.

5.3 Limitations and Future Directions

While the methods developed here represent significant advances, they also illuminate the path forward for future research.

5.3.1 Improving Substructure Retrieval and Verification

The oracle experiments conducted with ModiStruct revealed that the primary bottleneck in the end-to-end pipeline is not the generative model itself, but the initial task of retrieving a correct substructure from a spectral library. While our substructure verifier performs well, there is substantial room for improvement. Future work should focus on developing more accurate and robust models for identifying high-confidence substructure-query pairs from raw spectral comparisons, as this would yield the most significant improvement in overall performance.

5.3.2 Handling More Complex Molecular Scenarios

The methods presented in this dissertation were primarily developed and validated on compound pairs restricted to hard-to-validate conditions like having exactly one modification site or being substructure of the other. A key avenue for future work is to extend these frameworks to handle more relaxed conditions. For ModiFinder, this would require adapting the spectral alignment algorithm to search for combinations of mass shifts. For ModiStruct, it would involve training the model to be able to work with structurally similar but not necessarily substructure moleculess. Furthermore, developing strategies for molecules that lack a close analog in existing libraries remains an open and important challenge.

5.3.3 Advancing Generative Modeling Techniques

The ModiStruct framework is highly flexible and can benefit from future advances in generative modeling. Potential improvements include exploring classifier-free guidance[29] to further enhance the conditioning of the diffusion process, or integrating other data modalities

like ion mobility spectrometry to provide additional structural constraints[85]. Additionally, using the consensus of multiple pairs or building upon the Maximum Common Substructure of these might provide potential improvements[31].

5.3.4 The Critical Need for High-Quality Data

A recurring theme is that the performance of all these computational methods is intrinsically linked to the quality and comprehensiveness of the available data[38]. The development of the next generation of predictive and generative models will depend on the continued growth of large-scale, high-quality, and publicly accessible spectral libraries. In particular, libraries acquired under multiple, standardized experimental conditions are critical for training models that are robust and generalizable to real-world analytical data.

5.3.5 Using the network

Another potential future development of ModiFinder, is to enhance its analytical capacity by expanding the utilization of multiple known analogs within the spectral network (Fig. 5.1). This approach not only broadens the scope of potential structural identifications but also increases the accuracy of our analyses by incorporating a richer set of comparative data points.

5.4 Concluding Remarks

The journey from localizing a chemical change with ModiFinder to generating a complete molecular structure with ModiStruct demonstrates a powerful new direction

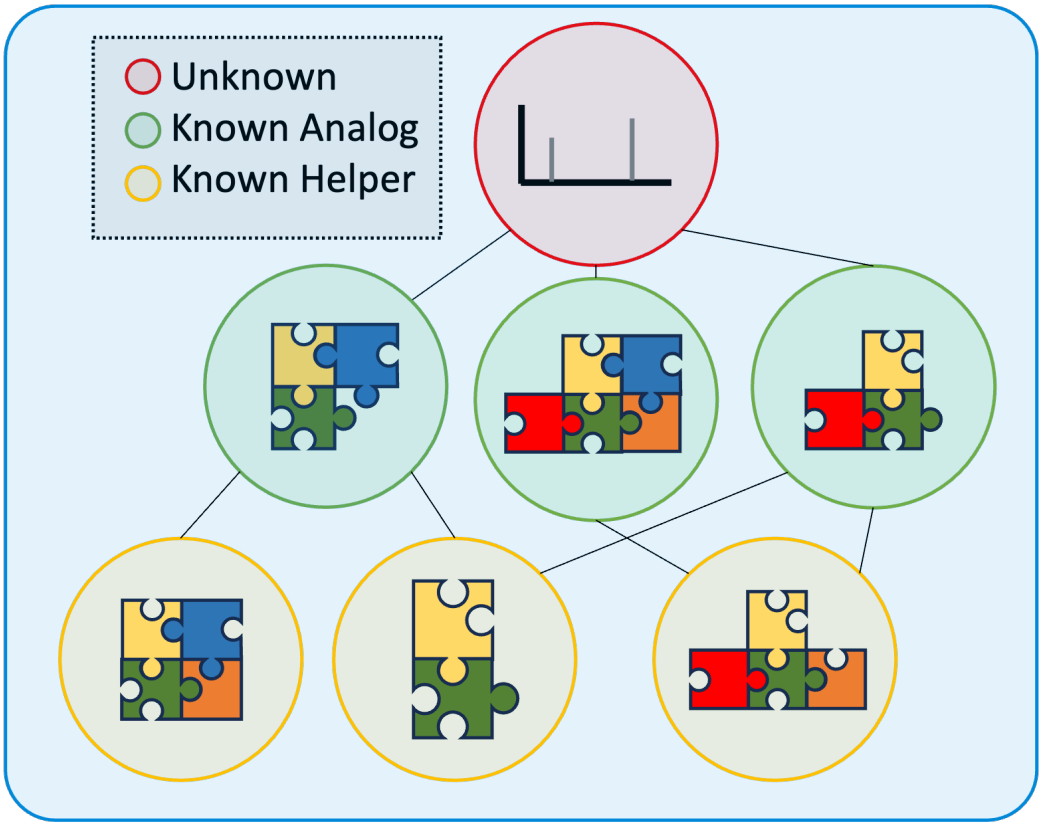


Figure 5.1: ModiFinder in a network

for computational mass spectrometry. By intelligently combining experimental data with structural priors, we have developed tools that automate complex analytical tasks and push the boundaries of what is possible in automated molecular discovery. It is our hope that these methods will be broadly applied in fields from toxicology and pharmacology to metabolomics and natural product discovery, accelerating the pace at which we can identify novel molecules and understand their role in the world around us.

Bibliography

[1] Felicity Allen, Russ Greiner, and David Wishart. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. 11(1):98–110.

[2] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17981–17993. Curran Associates, Inc., 2021.

[3] Wout Bittremieux, Robin Schmid, Florian Huber, Justin J. J. van der Hooft, Mingxun Wang, and Pieter C. Dorrestein. Comparison of cosine, modified cosine, and neutral loss based spectrum alignment for discovery of structurally related molecules. 33(9):1733–1744.

[4] Wout Bittremieux, Mingxun Wang, and Pieter C Dorrestein. The critical role that spectral libraries play in capturing the metabolomics community knowledge, 2022.

[5] Ivana Blaženović, Tobias Kind, Hrvoje Torbašinović, Slobodan Obrenović, Sajjan S. Mehta, Hiroshi Tsugawa, Tobias Wermuth, Nicolas Schauer, Martina Jahn, Rebekka Biedendieck, Dieter Jahn, and Oliver Fiehn. Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: database boosting is needed to achieve 93% accuracy. 9(1):32.

[6] Montgomery Bohde, Mrunali Manjrekar, Runzhong Wang, Shuiwang Ji, and Connor W. Coley. DiffMS: Diffusion generation of molecules conditioned on mass spectra.

[7] Danh Bui-Thi, Youzhong Liu, Jennifer L. Lippens, Kris Laukens, and Thomas De Vijlder. TransExION: a transformer based explainable similarity metric for comparing IONS in tandem mass spectrometry. 16(1):61.

[8] Roman Bushuiev, Anton Bushuiev, Niek F de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, Marcus Ludwig, Nils A Haupt, Apurva Kalia, Corinna Brungs, Robin Schmid, Russell Greiner, Bo Wang, David S Wishart, Li-Ping Liu, Juho Rousu, Wout Bittremieux, Hannes Rost, Tytus D Mak, Soha Hassoun, Florian Huber, Justin J J van der Hooft, Michael A Stravs, Sebastian Böcker, Josef Sivic, and Tomáš Pluskal. MassSpecGym: A benchmark for the discovery and identification of molecules. 2024.

[9] Roman Bushuiev, Anton Bushuiev, Niek F. de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, Marcus Ludwig, Nils A. Haupt, Apurva Kalia, Corinna Brungs, Robin Schmid, Russell Greiner, Bo Wang, David S. Wishart, Li-Ping Liu, Juho Rousu, Wout Bittremieux, Hannes Rost, Tytus D. Mak, Soha Hassoun, Florian Huber, Justin J. J. van der Hooft, Michael A. Stravs, Sebastian Böcker, Josef Sivic, and Tomáš Pluskal. Massspecgym: A benchmark for the discovery and identification of molecules, 2025.

[10] Thomas Butler, Abraham Frandsen, Rose Lightheart, Brian Bargh, James Taylor, T J Bollerman, Thomas Kerby, Kiana West, Gennady Voronov, Kevin Moon, Tobias Kind, Pieter Dorrestein, August Allen, Viswa Colluru, and David Healey. MS2Mol: A transformer model for illuminating dark chemical space from mass spectra. June 2023.

[11] Robert J. Chalkley and Karl R. Clauser. Modification site localization scoring: strategies and performance. 11(5):3–14.

[12] Brian T. Cooper, Xinjian Yan, Yamil Simón-Manso, Dmitrii V. Tchekhovskoi, Yuri A. Mirokhin, and Stephen E. Stein. Hybrid search: A method for identifying metabolites absent from tandem mass spectrometry libraries. 91(21):13924–13932.

[13] Ricardo R da Silva, Pieter C Dorrestein, and Robert A Quinn. Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. U. S. A.*, 112(41):12549–12550, October 2015.

[14] Ricardo R. da Silva, Mingxun Wang, Louis-Félix Nothias, Justin J. J. van der Hooft, Andrés Mauricio Caraballo-Rodríguez, Evan Fox, Marcy J. Balunas, Jonathan L. Klassen, Norberto Peporine Lopes, and Pieter C. Dorrestein. Propagating annotations of molecular networks using in silico fragmentation. 14(4):e1006089.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 2018.

[16] Sophia Doll and Alma L. Burlingame. Mass spectrometry-based detection and assignment of protein posttranslational modifications. 10(1):63–71.

[17] Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin A Hoffmann, Daniel Petras, William H Gerwick, Juho Rousu, Pieter C Dorrestein, and Sebastian Böcker. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.*, 39(4):462–471, April 2021.

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

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

[20] Elisabeth Varga, Franz Berthiller, Rudolf Krska, Michael Sulyok, Emma Rennie, and Thomas Glauner. Screening and verifying mycotoxins in food with q-TOF LC/MS and an accurate mass library. (5991).

[21] Emma E. Rennie, Robert H. Williams, Ruben Garnica, and Maria VanDamme. Improving confidence in compound identification using agilent curated databases and libraries. (5991).

[22] Alexander E. Fox Ramos, Pierre Le Pogam, Charlotte Fox Alcover, Elvis Otogo N'Nang, Gaëla Cauchie, Hazrina Hazni, Khalijah Awang, Dimitri Bréard, Antonio M. Echavarren, Michel Frédérich, Thomas Gaslonde, Marion Girardot, Raphaël Grougnet, Mariia S. Kirillova, Marina Kritsanida, Christelle Lémus, Anne-Marie Le Ray, Guy Lewin, Marc Litaudon, Lengo Mambu, Sylvie Michel, Fedor M. Miloserdov, Michael E. Muratore, Pascal Richomme-Peniguel, Fanny Roussi, Laurent Evanno, Erwan Poupon, Pierre Champy, and Mehdi A. Beniddir. Collected mass spectrometry data on monoterpene indole alkaloids from natural product chemistry research. 6(1):15.

[23] Amit Gangwal and Antonio Lavecchia. Artificial intelligence in natural product drug discovery: Current applications and future perspectives. *Journal of Medicinal Chemistry*, 68(4):3948–3969, 2025. PMID: 39916476.

[24] Julia M. Gauglitz, Christine M. Aceves, Alexander A. Aksenov, Gajender Aleti, Jehad Almaliti, Amina Bouslimani, Elizabeth A. Brown, Anaamika Campeau, Andrés Mauricio Caraballo-Rodríguez, Rama Chaar, Ricardo R. Da Silva, Alyssa M. Demko, Francesca Di Ottavio, Emmanuel Elijah, Madeleine Ernst, L. Paige Ferguson, Xavier Holmes, Alan K. Jarmusch, Lingjing Jiang, Kyo Bin Kang, Irina Koester, Brian Kwan, Jie Li, Yueying Li, Alexey V. Melnik, Carlos Molina-Santiago, Bohan Ni, Aaron L. Oom, Morgan W. Panitchpakdi, Daniel Petras, Robert Quinn, Nicole Sikora, Katharina Spengler, Bahar Teke, Anupriya Tripathi, Sabah Ul-Hasan, Justin J.J. Van Der Hooft, Fernando Vargas, Alison Vrbanac, Anthony Q. Vu, Steven C. Wang, Kelly Weldon, Kayla Wilson, Jacob M. Wozniak, Michael Yoon, Nuno Bandeira, and Pieter C. Dorrestein. Untargeted mass spectrometry-based metabolomics approach unveils molecular changes in raw and processed foods and beverages. 302:125290.

[25] Samuel Goldman, Janet Li, and Connor W. Coley. Generating molecular fragmentation graphs with autoregressive neural networks. 96(8):3419–3428.

[26] Samuel Goldman, Jeremy Wohlwend, Martin Stražar, Guy Haroush, Ramnik J Xavier, and Connor W Coley. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nat. Mach. Intell.*, 5(9):965–979, August 2023.

[27] Greg Landrum, Paolo Tosco. rdkit/rdkit.

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020.

[29] Jonathan Ho and Tim Salimans. Classifier-Free diffusion guidance. 2022.

[30] Jean Daniel Berset, Emma Rennie, and Thomas Glauner. Screening and identification of emerging contaminants in wastewater treatment plant effluents using UHPLC/q-TOF MS and an accurate mass database and library. (5991).

[31] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. 2018.

[32] Walter Keller-Schierlein, Michael Meyer, Axel Zeeck, Manfred Damberg, Reinhard Machinek, Hans Zähner, and Günther Lazar. Isolation and structural elucidation of naphthomycins b and c. 36(5):484–492.

[33] Hyun Woo Kim, Mingxun Wang, Christopher A. Leber, Louis-Félix Nothias, Raphael Reher, Kyo Bin Kang, Justin J. J. Van Der Hooft, Pieter C. Dorrestein, William H. Gerwick, and Garrison W. Cottrell. NPClassifier: A deep neural network-based structural classification tool for natural products. 84(11):2795–2807.

[34] Eleni Litsa, Vijil Chenthamarakshan, Payel Das, and Lydia Kavraki. Spec2Mol: An end-to-end deep learning framework for translating MS/MS spectra to de-novo molecules. September 2021.

[35] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. 2022.

[36] Bin ma. Challenges in computational analysis of mass spectrometry data for proteomics. *J. Comput. Sci. Technol.*, 25:107–123, 01 2009.

[37] S. A. McLuckey. Principles of collisional activation in analytical mass spectrometry. 3(6):599–614.

[38] Sedir Mohammed, Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. The effects of data quality on machine learning performance on tabular data. *Inf. Syst.*, 132(102549):102549, July 2025.

[39] H L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Doc.*, 5(2):107–113, May 1965.

[40] Michael Murphy, Stefanie Jegelka, Ernest Fraenkel, Tobias Kind, David Healey, and Thomas Butler. Efficiently predicting high resolution mass spectra with graph neural networks, 2023.

[41] Louis-Félix Nothias, Daniel Petras, Robin Schmid, Kai Dührkop, Johannes Rainer, Abinesh Sarvepalli, Ivan Protsyuk, Madeleine Ernst, Hiroshi Tsugawa, Markus Fleischauer, Fabian Aicheler, Alexander A. Aksenov, Oliver Alka, Pierre-Marie Allard, Aiko Barsch, Xavier Cachet, Andres Mauricio Caraballo-Rodriguez, Ricardo R. Da Silva, Tam Dang, Neha Garg, Julia M. Gauglitz, Alexey Gurevich, Giorgis Isaac, Alan K. Jarmusch, Zdeněk Kameník, Kyo Bin Kang, Nikolas Kessler, Irina Koester, Ansgar

Korf, Audrey Le Gouellec, Marcus Ludwig, Christian Martin H., Laura-Isobel McCall, Jonathan McSayles, Sven W. Meyer, Hosein Mohimani, Mustafa Morsy, Oriane Moyne, Steffen Neumann, Heiko Neuweger, Ngoc Hung Nguyen, Melissa Nothias-Esposito, Julien Paolini, Vanessa V. Phelan, Tomáš Pluskal, Robert A. Quinn, Simon Rogers, Bindesh Shrestha, Anupriya Tripathi, Justin J. J. Van Der Hooft, Fernando Vargas, Kelly C. Weldon, Michael Witting, Heejung Yang, Zheng Zhang, Florian Zubeil, Oliver Kohlbacher, Sebastian Böcker, Theodore Alexandrov, Nuno Bandeira, Mingxun Wang, and Pieter C. Dorrestein. Feature-based molecular networking in the GNPS analysis environment. 17(9):905–908.

[42] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. 2017.

[43] Daniel Petras, Vanessa V. Phelan, Deepa Acharya, Andrew E. Allen, Allegra T. Aron, Nuno Bandeira, Benjamin P. Bowen, Deirdre Belle-Oudry, Simon Boecker, Dale A. Cummings, Jessica M. Deutsch, Eoin Fahy, Neha Garg, Rachel Gregor, Jo Handelsman, Mirtha Navarro-Hoyos, Alan K. Jarmusch, Scott A. Jarmusch, Katherine Louie, Katherine N. Maloney, Michael T. Marty, Michael M. Meijler, Itzhak Mizrahi, Rachel L. Neve, Trent R. Northen, Carlos Molina-Santiago, Morgan Panitchpakdi, Benjamin Pullman, Aaron W. Puri, Robin Schmid, Shankar Subramaniam, Monica Thukral, Felipe Vasquez-Castro, Pieter C. Dorrestein, and Mingxun Wang. GNPS dashboard: collaborative exploration of mass spectrometry data in the web browser. 19(2):134–136.

[44] Zhengyang Qi, Zijing Liu, Jiying Zhang, He Cao, Xiaohua Xu, and Yu Li. ControlMol: Adding substructure control to molecule diffusion models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, April 2025.

[45] Florian Rasche, Ales Svatos, Ravi Kumar Maddula, Christoph Böttcher, and Sebastian Böcker. Computing fragmentation trees from tandem mass spectrometry data. 83(4):1243–1251.

[46] Lars Ridder, Justin J. J. Van Der Hooft, Stefan Verhoeven, Ric C. H. De Vos, René Van Schaik, and Jacques Vervoort. Substructure-based annotation of high-resolution multistage MS ^{n} spectral trees. 26(20):2461–2471.

[47] Christoph Ruttkies, Steffen Neumann, and Stefan Posch. Improving MetFrag with statistical learning of fragment annotations. 20(1):376.

[48] Christoph Ruttkies, Emma L. Schymanski, Sebastian Wolf, Juliane Hollender, and Steffen Neumann. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. 8(1):3.

[49] Alexandra C. Schrimpe-Rutledge, Simona G. Codreanu, Stacy D. Sherrod, and John A. McLean. Untargeted metabolomics strategies-challenges and emerging directions. 27(12):1897–1905.

[50] Emma L Schymanski, Christoph Ruttkies, Martin Krauss, Céline Brouard, Tobias Kind, Kai Dührkop, Felicity Allen, Arpana Vaniya, Dries Verdegem, Sebastian Böcker, Juho Rousu, Huibin Shen, Hiroshi Tsugawa, Tanvir Sajed, Oliver Fiehn, Bart Ghesquière, and Steffen Neumann. Critical assessment of small molecule identification 2016: automated methods. *J. Cheminform.*, 9(1):22, March 2017.

[51] Mohammad Reza Zare Shahneh, James Pyke, Emma E Rennie, and Mingxun Wang. Multispectrum ModiFinder site localization performance. *J. Am. Soc. Mass Spectrom.*, 36(5):999–1007, May 2025.

[52] Mohammad Reza Zare Shahneh, Michael Strobel, Giovanni Andrea Vitale, Christian Geibel, Yasin El Abiead, Neha Garg, Berenike Wagner, Karl Forchhammer, Allegra Aron, Vanessa V Phelan, Daniel Petras, and Mingxun Wang. ModiFinder: Tandem mass spectral alignment enables structural modification site localization. 35(11):2564–2578. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

[53] C. E. Shannon. A mathematical theory of communication. 27(3):379–423.

[54] Aditya Divyakant Shrivastava, Neil Swainston, Soumitra Samanta, Ivayla Roberts, Marina Wright Muelas, and Douglas B Kell. MassGenie: A transformer-based deep learning method for identifying small molecules from their mass spectra. *Biomolecules*, 11(12):1793, November 2021.

[55] Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. 2015.

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. 2020.

[57] S. E. Stein and D. R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. 5(9):859–866.

[58] Paolo Stincone, Abzer K. Pakkir Shah, Robin Schmid, Lana G. Graves, Stilianos P. Lambidis, Ralph R. Torres, Shu-Ning Xia, Vidit Minda, Allegra T. Aron, Mingxun Wang, Chambers C. Hughes, and Daniel Petras. Evaluation of data-dependent MS/MS acquisition parameters for non-targeted metabolomics and molecular networking of environmental samples: Focus on the q exactive platform. 95(34):12673–12682.

[59] Michael A. Stravs, Kai Dührkop, Sebastian Böcker, and Nicola Zamboni. MSNovelist: de novo structure generation from mass spectra. 19(7):865–870.

[60] Michael Strobel, Alberto Gil-de-la Fuente, Mohammad Reza Zare Shahneh, Yasin El Abiead, Roman Bushuiev, Anton Bushuiev, Tomáš Pluskal, and Mingxun Wang. An evaluation methodology for machine learning-based tandem mass spectra similarity prediction. *BMC Bioinformatics*, 26(1):174, July 2025.

[61] Syed Salman Lateef, David A. Weil, Mahsan Miladi, Dan-Hui-Dorothy Yang, Shi-Fen Xu, and Emma Rennie. Application of an accurate mass database and library for the screening and verification of extractables and leachables in pharmaceutical drug products. (5991).

[62] Nicholas J. Tobias, Hendrik Wolff, Bardya Djahanschiri, Florian Grundmann, Max Kronenwerth, Yi-Ming Shi, Svenja Simonyi, Peter Grün, David Shapiro-Ilan, Sacha J. Pidot, Timothy P. Stinear, Ingo Ebersberger, and Helge B. Bode. Natural product diversity associated with the nematode symbionts photorhabdus and xenorhabdus. 2(12):1676–1685.

[63] Jos Torge, Charles Harris, Simon V Mathis, and Pietro Lio. DiffHopp: A graph diffusion model for novel drug design via scaffold hopping. 2023.

[64] Hiroshi Tsugawa, Tobias Kind, Ryo Nakabayashi, Daichi Yukihira, Wataru Tanaka, Tomas Cajka, Kazuki Saito, Oliver Fiehn, and Masanori Arita. Hydrogen rearrangement rules: Computational MS/MS fragmentation and structure elucidation using MS-FINDER software. 88(16):7946–7958.

[65] Lara Van Tetering, Sylvia Spies, Quirine D. K. Wildeman, Kas J. Houthuijs, Rianne E. Van Outersterp, Jonathan Martens, Ron A. Wevers, David S. Wishart, Giel Berden, and Jos Oomens. A spectroscopic test suggests that fragment ion structure annotations in MS/MS libraries are frequently incorrect. 7(1):30.

[66] Fernando Vargas, Kelly C. Weldon, Nicole Sikora, Mingxun Wang, Zheng Zhang, Emily C. Gentry, Morgan W. Panitchpakdi, Andrés Mauricio Caraballo-Rodríguez, Pieter C. Dorrestein, and Alan K. Jarmusch. Protocol for community-created public MS/MS reference spectra within the global natural products social molecular networking infrastructure. 34(10):e8725.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

[68] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. 2017.

[69] Dries Verdegem, Diether Lambrechts, Peter Carmeliet, and Bart Ghesquière. Improved metabolite identification with MIDAS and MAGMa through MS/MS spectral dataset-driven parameter optimization. 12(6):98.

[70] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation, 2023.

[71] Fei Wang, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S. Wishart. CFM-ID 4.0: More accurate ESI-MS/MS spectral prediction and compound identification. 93(34):11692–11700.

[72] Liang Wang, Yu Rong, Tingyang Xu, Zhenyi Zhong, Zhiyuan Liu, Pengju Wang, Deli Zhao, Qiang Liu, Shu Wu, and Liang Wang. DiffSpectra: Molecular structure elucidation from spectra using diffusion models. 2025.

[73] Mingxun Wang, Jeremy J. Carver, Vanessa V. Phelan, Laura M. Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A. Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V. Melnik, Michael J. Meehan, Wei-Ting Liu, Max Crüsemann, Paul D. Boudreau, Eduardo Esquenazi, Mario Sandoval-Calderón, Roland D. Kersten, Laura A. Pace, Robert A. Quinn, Katherine R. Duncan, Cheng-Chih Hsu, Dimitrios J. Floros, Ronnie G. Gavilan, Karin Kleigrewe, Trent Northen, Rachel J. Dutton, Delphine Parrot, Erin E. Carlson, Bertrand Aigle, Charlotte F. Michelsen, Lars Jelsbak, Christian Sohlenkamp, Pavel Pevzner, Anna Edlund, Jeffrey McLean, Jörn Piel, Brian T. Murphy, Lena Gerwick, Chih-Chuang Liaw, Yu-Liang Yang, Hans-Ulrich Humpf, Maria Maansson, Robert A. Keyzers, Amy C. Sims, Andrew R. Johnson, Ashley M. Sidebottom, Brian E. Sedio, Andreas Klitgaard, Charles B. Larson, Cristopher A. Boya P, Daniel Torres-Mendoza, David J. Gonzalez, Denise B. Silva, Lucas M. Marques, Daniel P. Demarque, Egle Pociute, Ellis C. O'Neill, Enora Briand, Eric J. N. Helfrich, Eve A. Granatosky, Evgenia Glukhov, Florian Ryffel, Hailey Houson, Hosein Mohimani, Jenan J. Kharbush, Yi Zeng, Julia A. Vorholt, Kenji L. Kurita, Pep Charusanti, Kerry L. McPhail, Kristian Fog Nielsen, Lisa Vuong, Maryam Elfeki, Matthew F. Traxler, Niclas Engene, Nobuhiro Koyama, Oliver B. Vining, Ralph Baric, Ricardo R. Silva, Samantha J. Mascuch, Sophie Tomasi, Stefan Jenkins, Venkat Macherla, Thomas Hoffman, Vinayak Agarwal, Philip G. Williams, Jingqui Dai, Ram Neupane, Joshua Gurr, Andrés M. C. Rodríguez, Anne Lamsa, Chen Zhang, Kathleen Dorrestein, Brendan M. Duggan, Jehad Almaliti, Pierre-Marie Allard, Prasad Phapale, Louis-Felix Nothias, Theodore Alexandrov, Marc Litaudon, Jean-Luc Wolfender, Jennifer E. Kyle, Thomas O. Metz, Tyler Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav, Rolf Müller, Katrina M. Waters, Wenyuan Shi, Xueting Liu, Lixin Zhang, Rob Knight, Paul R. Jensen, Bernhard O. Palsson, Kit Pogliano, Roger G. Linington, Marcelino Gutiérrez, Norberto P. Lopes, William H. Gerwick, Bradley S. Moore, Pieter C. Dorrestein, and Nuno Bandeira. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. 34(8):828–837.

[74] Yingfeng Wang, Guruprasad Kora, Benjamin P. Bowen, and Chongle Pan. MIDAS: A database-searching algorithm for metabolite identification in metabolomics. 86(19):9496–9503.

[75] Yinkai Wang, Xiaohui Chen, Liping Liu, and Soha Hassoun. MADGEN: Mass-spec attends to de novo molecular generation.

[76] Jeramie Watrous, Patrick Roach, Theodore Alexandrov, Brandi S. Heath, Jane Y. Yang, Roland D. Kersten, Menno van der Voort, Kit Pogliano, Harald Gross, Jos M. Raaijmakers, Bradley S. Moore, Julia Laskin, Nuno Bandeira, and Pieter C. Dorrestein. Mass spectral molecular networking of living microbial colonies. 109(26):E1743–1752.

[77] Jennifer N Wei, David Belanger, Ryan P Adams, and D Sculley. Rapid prediction of

electron-ionization mass spectrometry using neural networks. *ACS Cent. Sci.*, 5(4):700–708, April 2019.

[78] Heinz Wolf, Gianni Chinali, and Andrea Parmeggiani. Kirromycin, an inhibitor of protein biosynthesis that acts on elongation factor tu. 71(12):4910–4914.

[79] Shipei Xing, Yan Hu, Zixuan Yin, Min Liu, Xiaoyu Tang, Mingliang Fang, and Tao Huan. Retrieving and utilizing hypothetical neutral losses from tandem mass spectra for spectral similarity analysis and unknown metabolite annotation. 92(21):14476–14483.

[80] Shipei Xing, Sam Shen, Banghua Xu, Xiaoxiao Li, and Tao Huan. BUDDY: molecular formula discovery via bottom-up MS/MS interrogation. 20(6):881–890.

[81] Jane Y. Yang, Laura M. Sanchez, Christopher M. Rath, Xueting Liu, Paul D. Boudreau, Nicole Bruns, Evgenia Glukhov, Anne Wodtke, Rafael de Felicio, Amanda Fenner, Weng Ruh Wong, Roger G. Linington, Lixin Zhang, Hosana M. Debonsi, William H. Gerwick, and Pieter C. Dorrestein. Molecular networking as a dereplication strategy. 76(9):1686–1699.

[82] Yiming Yang, Shuang Sun, Shuyuan Yang, Qin Yang, Xinqiong Lu, Xiaohao Wang, Quan Yu, Xinming Huo, and Xiang Qian. Structural annotation of unknown molecules in a miniaturized mass spectrometer based on a transformer enabled fragment tree method. *Commun. Chem.*, 7(1):109, May 2024.

[83] Adamo Young, Bo Wang, and Hannes Röst. Massformer: Tandem mass spectrum prediction for small molecules using graph transformers, 2023.

[84] Fengchao Yu, Guo Ci Teo, Andy T. Kong, Sarah E. Haynes, Dmitry M. Avtonomov, Daniel J. Geiszler, and Alexey I. Nesvizhskii. Identification of modified peptides using localization-aware open search. 11(1):4065.

[85] Xing Zhang, Kevin Quinn, Charmion Cruickshank-Quinn, Richard Reisdorph, and Nichole Reisdorph. The application of ion mobility mass spectrometry to metabolomics. *Curr. Opin. Chem. Biol.*, 42:60–66, February 2018.

[86] Yael Ziv, Fergus Imrie, Brian Marsden, and Charlotte M Deane. MolSnapper: Conditioning diffusion for structure-based drug design. *J. Chem. Inf. Model.*, 65(9):4263–4273, May 2025.

Appendix

ModiFinder Imbalance Symmetry

One of the scenarios that ModiFinder can fail is when there is a symmetry combined with an imbalance in the compound's structure that leads to a bias. For example, assume the modification (M) in the toy example illustrated in Figure A-1 A happens at the green puzzle piece A. In addition, assume there is a shifted peak corresponding to a fragment that has a green and a blue puzzle piece. In this case, Puzzle Piece B achieves a higher likelihood score as it appears in more potential structures (Figure A-1 B).

Key Factors

We integrated these attributes to develop a filtering method that helps identify scenarios where ModiFinder underperforms. Figure A-4 illustrates how these factors affect the dataset, with filter thresholds set as follows: For the low cosine score, we established a threshold of 0.3 to identify critical cases. For the low annotation ratio, we focused on pairs where less than 20% of the shifted peaks were annotated. For the "Too Ambiguous"

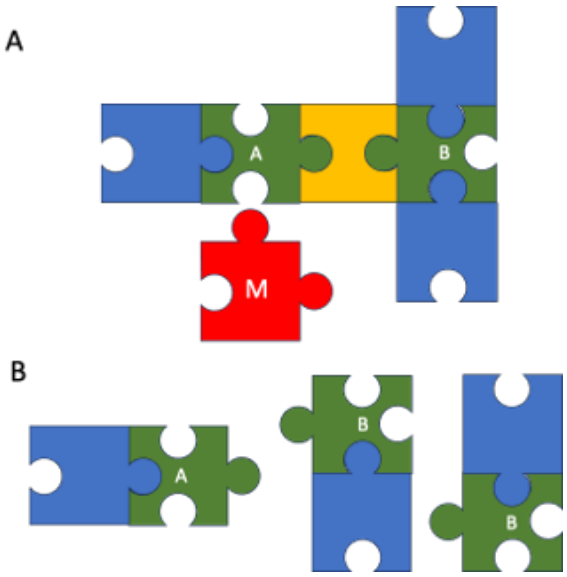


Figure A-1: A This figure shows a base known compound and its modification (red piece M) and the modification site (Green piece A). B This shows all the substructures that have a green and a blue piece. Even though the modification happens at A and there is a potential substructure containing A, B achieves a higher likelihood score.

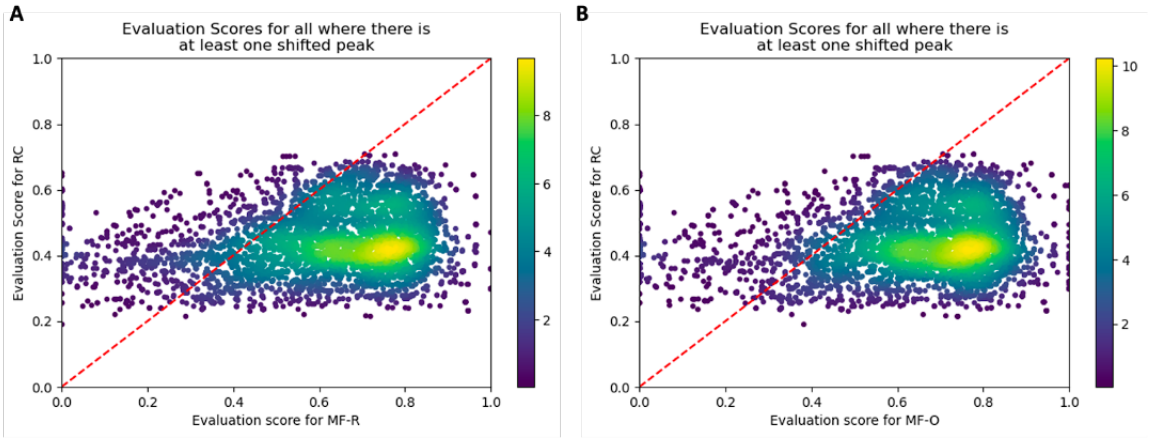


Figure A-2: A Evaluation scores for all the pairs in the dataset using ModiFinder-Refined (MF-R) and Random Choice. The points below the dashed line represent instances where MF-R achieves a better resul than RC. B: Evaluation scores for all the pairs in the dataset using ModiFinder-Oracle (MF-O) and Random Choice calculated and plotted. The points below the dashed line represent instances where MF-R achieves a better result than RC.

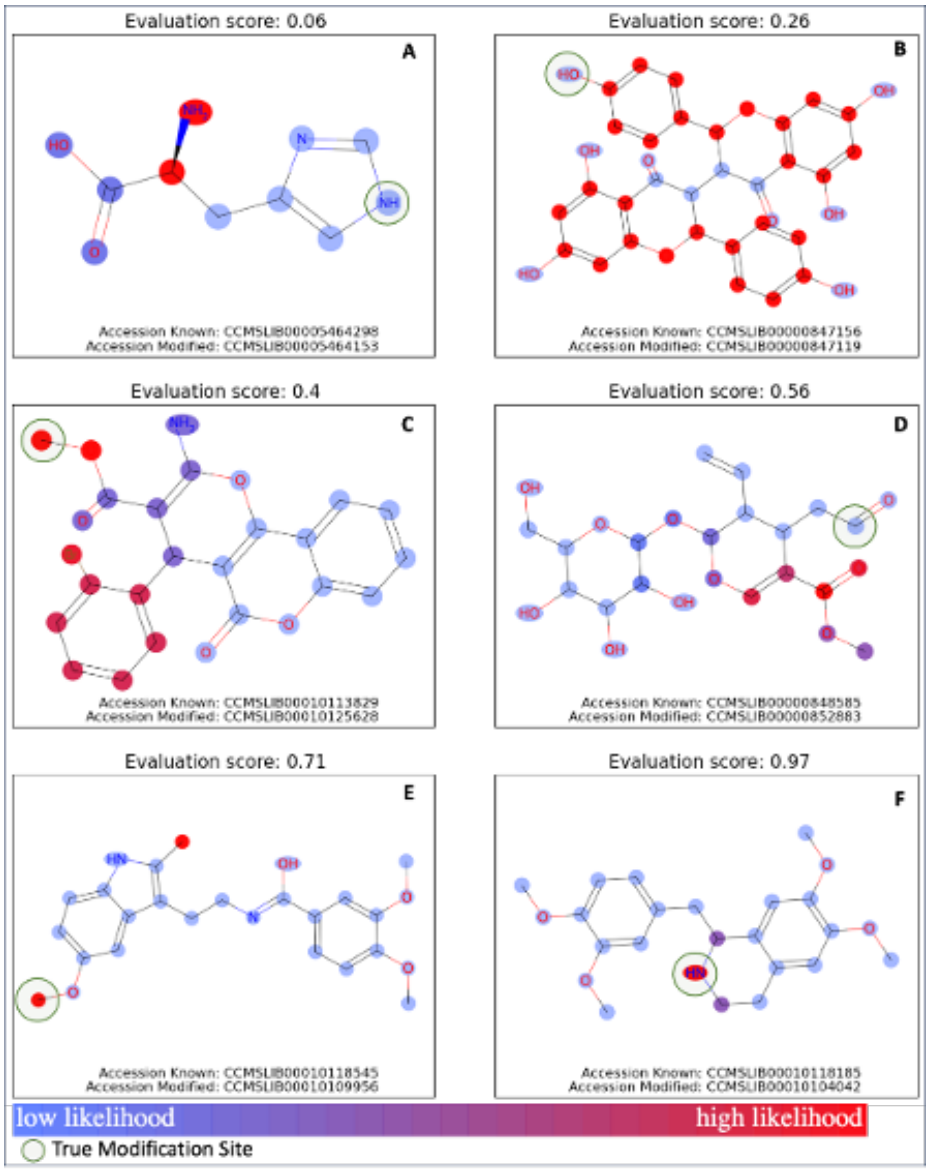


Figure A-3: For each example, the evaluation score along is specified. The GNPS accession code of the known and the modified unknown compound is provided for each example. The true modification site is highlighted with the green circle. A The prediction has a bad proximity cover and receives the low score of 0.06. B This prediction has a bad ambiguity cover and receives the low score of 0.26. C This prediction is moderately ambiguous and therefore receives an evaluation score of 0.4. D This prediction cannot detect the true modification site and the prediction is somewhat distant leading the a reasonable score of 0.56 corresponding to its moderate proximity cover. E while modifinder can find the true modification site with low ambiguity, it is still penalized for the distance and ambiguity introduced by the second option leading to a score of 0.71. F The prediction has great proximity and low ambiguity leading to the high score of 0.97.

filter, we selected instances with an average of 8 or more annotations per peak. Finally, for the presence of too many unshifted peaks, we identified pairs with three times as many unshifted peaks as shifted ones. For cases with excessive unshifted peaks, we focused on pairs with three times as many unshifted peaks compared to shifted ones. This approach eliminated 88% of problematic pairs and 45% of effective pairs, resulting in a precision of 90% and an accuracy of 67%. While the accuracy may not be exceptionally high, precision is crucial in this context. High precision ensures that chemists can rely on the results, leading to more accurate and effective outcomes. While individually these features provide insights into their significance, their stringent nature leads to the exclusion of many valid examples. Additionally, these features are more effective when used in conjunction. To better understand their interrelationships and improve our filtering approach, we developed a simple random forest classifier. This allowed us to assess the feature importance and predictive capabilities of our model. The analysis confirmed the Cosine Score as the most significant factor, alongside ambiguity, and the weight of the compounds. While our observations indicate that the molecular weight does not directly correlate with ModiFinder's performance or the average evaluation score, its significance in the random forest model is pronounced. This is because, due to the low atom count, any deviations in prediction lead to greater variability in the evaluation score. Consequently, this high variance elevates the importance of molecular weight in our model, as illustrated in Figure A-5. It's important to recognize that while these factors are critical, they are based solely on mass spectra data and do not consider the molecular structure. Factors like molecular symmetry and the relationship between structure and its MS/MS spectra are not accounted for in this analysis. Addressing

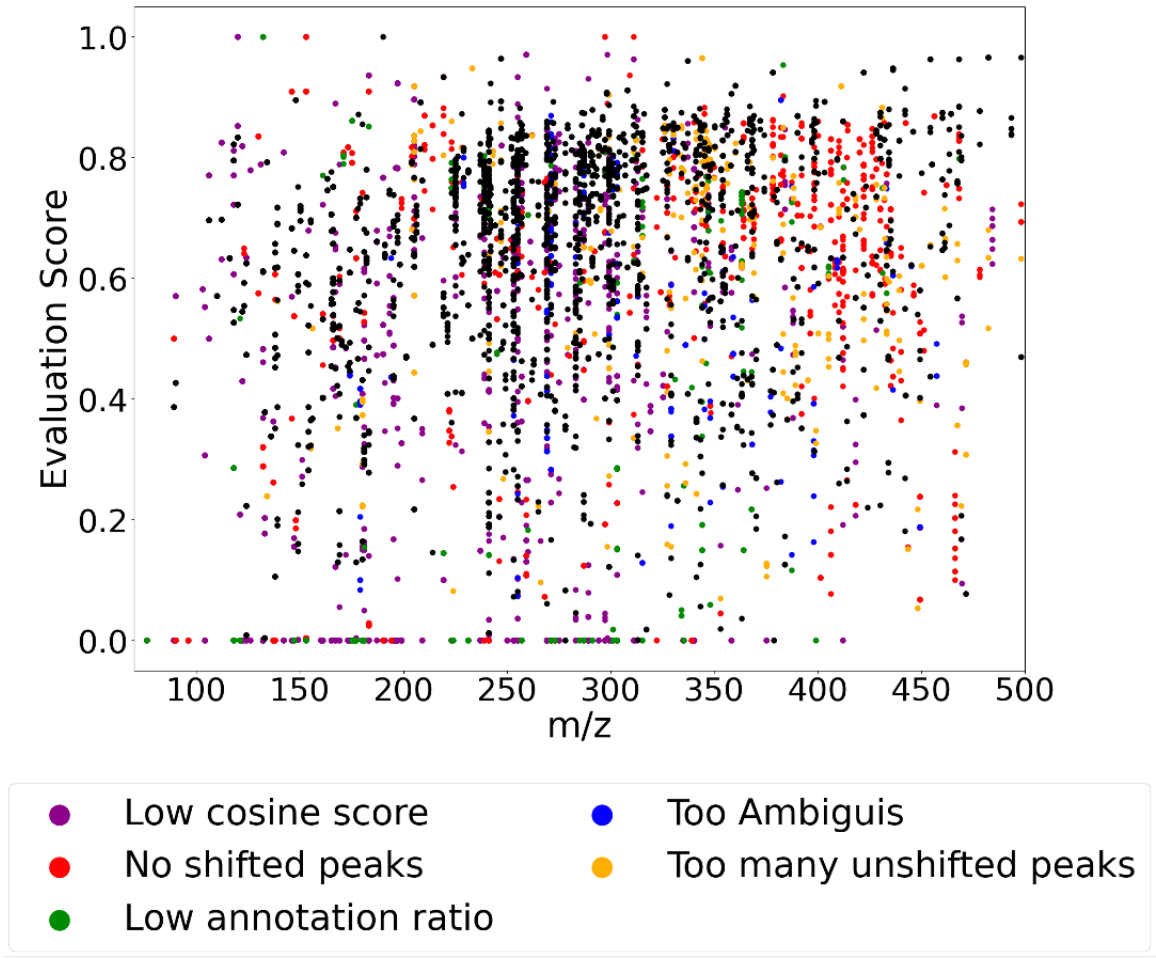


Figure A-4: Key Factors Contributing to Efficacy of ModiFinder

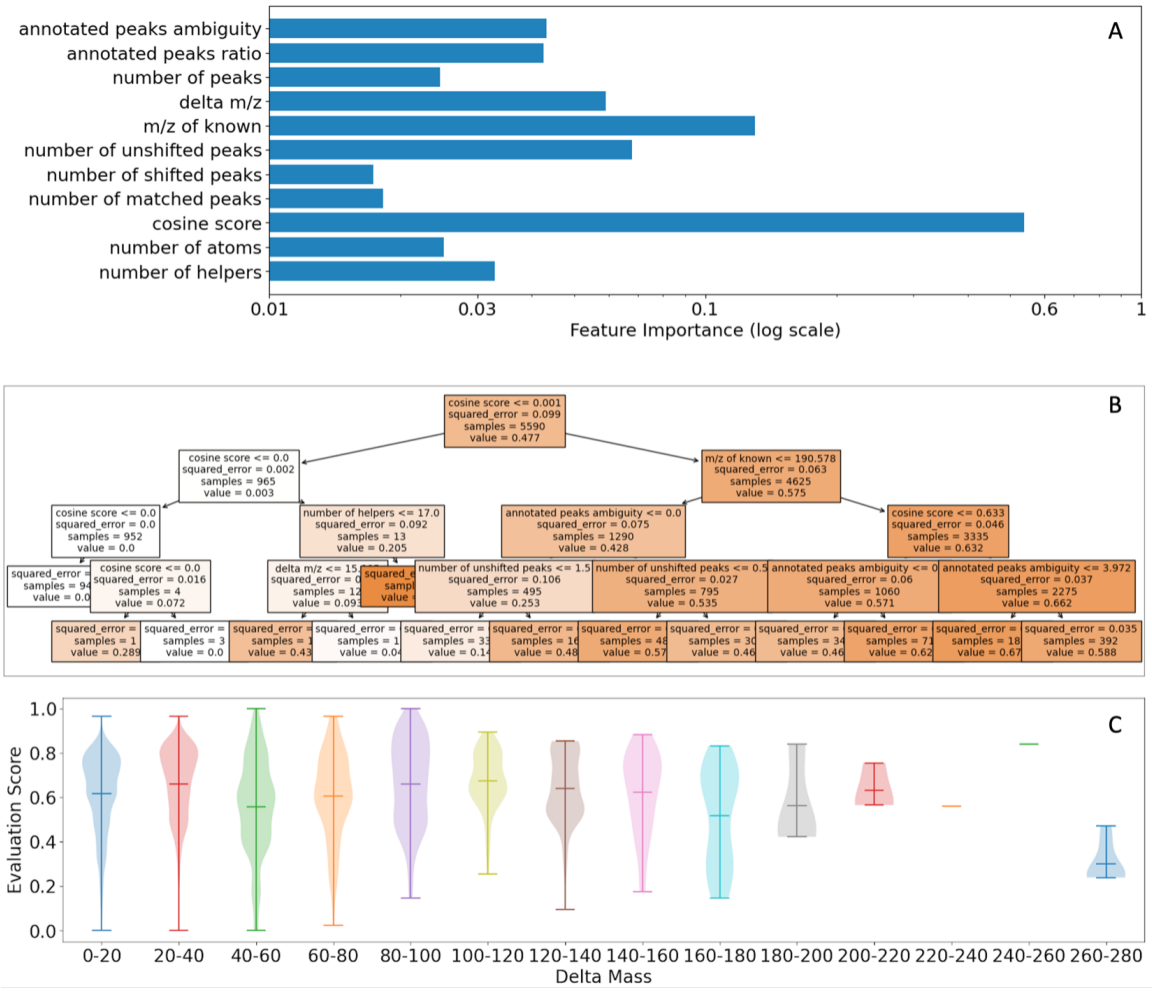


Figure A-5: A. shows the importance of features in the random forest regression model trained on the data. B. shows an illustration of one of the trees in the random forest model showing how different features perform in conjunction. C. shows the average evaluation score for different delta mass, this data shows no significant difference in ModiFinder's performance across different delta masses.

these aspects would require a separate dedicated project, as they represent a significant area of interest in their own right.

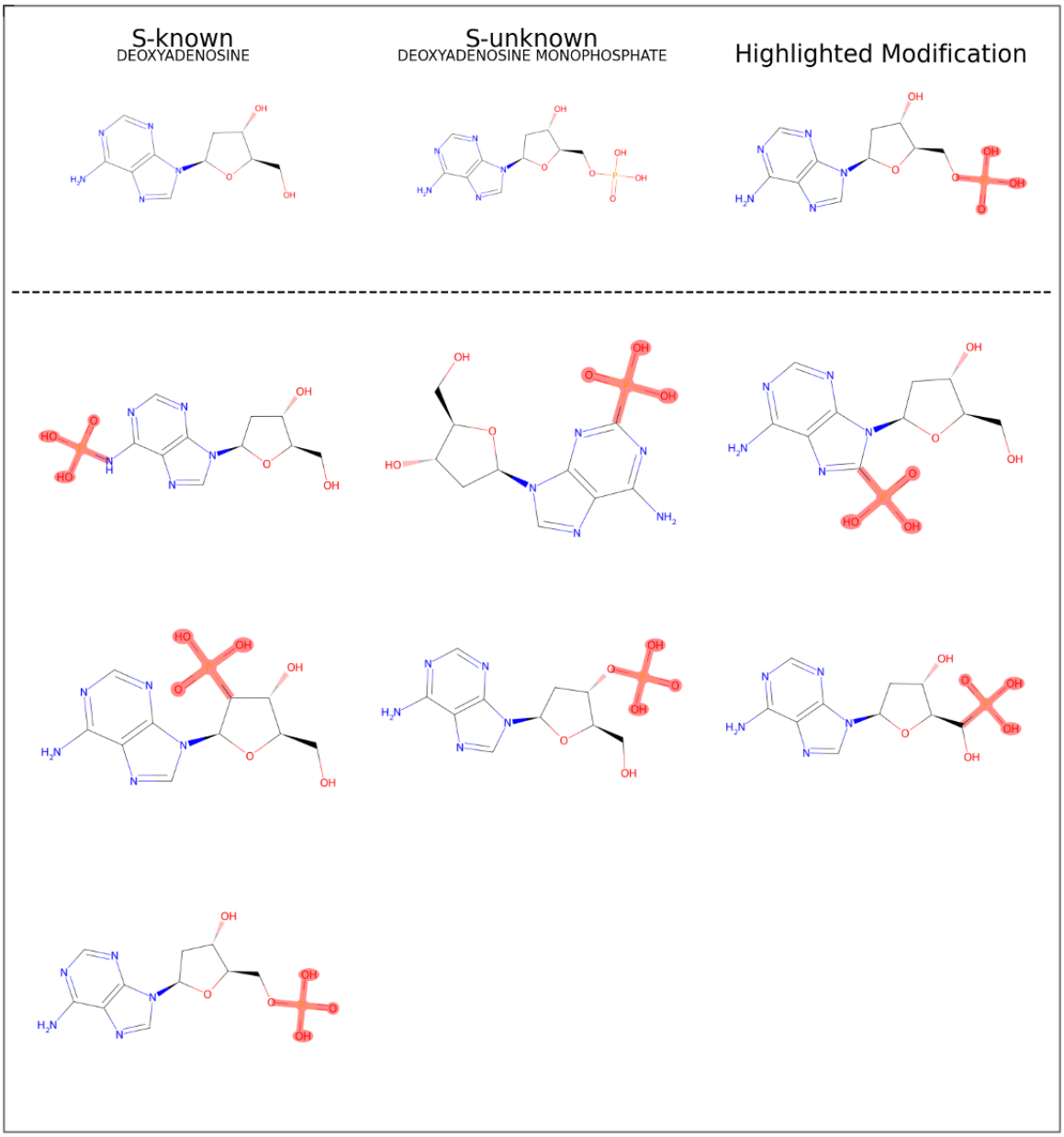


Figure A-6: Example of the permutation algorithm on Deoxyadenosine and Deoxyadenosine Monophosphate. The upper section displays both molecules with the modification part highlighted. The lower section exhibits various molecules synthesized by permutating the modification on Deoxyadenosine.

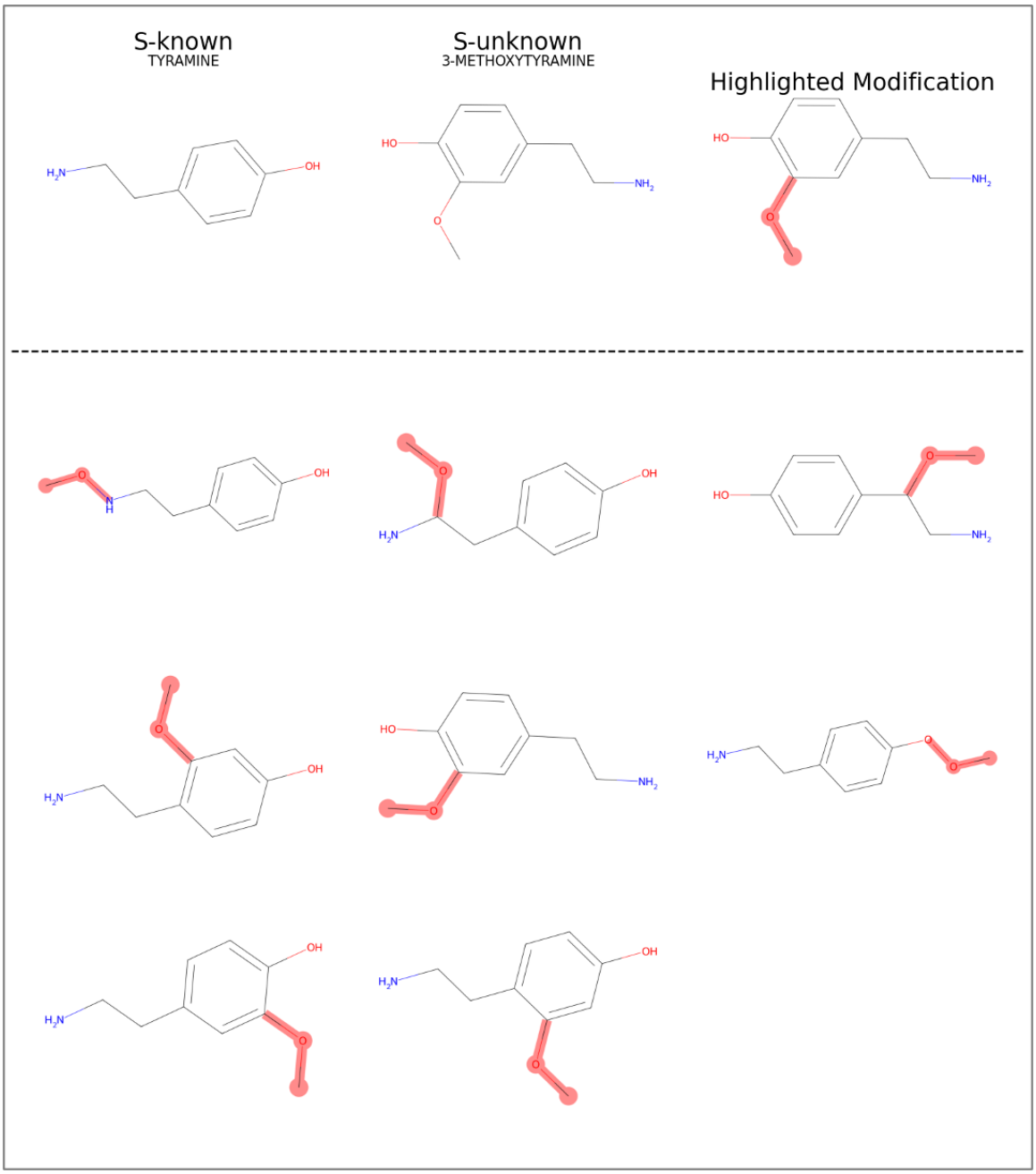


Figure A-7: Example of the permutation algorithm on Tyramine and 3-MethoxyTyramine. The upper section displays both molecules with the modification part highlighted. The lower section exhibits various molecules synthesized by permutating the modification on Tyramine.