UNIVERSITY OF CALIFORNIA
RIVERSIDE

HPC-Enabled Optimization and Scaling of QCxMS for Accelerated Mass Spectrum Prediction

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

Mastser's of Science

in

Computer Engineering

by

Yunshu Wang

March 2024

Thesis Committee:

Dr. Mingxun Wang, Chairperson
Dr. Tao Jiang
Dr. Wenxiu Ma

The Thesis of Yunshu Wang is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my advisor, without whose help, I would not have been here.

To my parents for all the support.

ABSTRACT OF THE THESIS

HPC-Enabled Optimization and Scaling of QCxMS for Accelerated Mass Spectrum Prediction

by

Yunshu Wang

Mastser's of Science, Graduate Program in Computer Engineering
University of California, Riverside, March 2024
Dr. Mingxun Wang, Chairperson

Mass spectrometry (MS) is a foundational element in contemporary analytical chemistry, facilitating biomolecule identification and playing a pivotal role in deciphering the intricacies of biological systems. Within the MS field, tandem mass spectra (MS/MS) are vital in the process of identifying molecules in untargeted approaches. MS/MS provides the fragmentation pattern of the molecules under study, providing more information about the molecule and enhancing the proper identification. Nevertheless, the availability of experimental MS/MS from compounds is limited due to the lack of time, money, and/or availability of reference standards. In those cases, the prediction of mass spectra enables the identification of molecules that have not been experimentally analyzed before, and they are especially important for de-novo identification. However, the current prediction methods still have limitations. QCxMS (Quantum Chemical Mass Spectrometry), the sole tool for predicting mass spectra using dynamic molecular methods, is a promising alternative for improving the predictions of fragmentation patterns of molecules. The current experiments using QCxMS have shown limited results. QCxMS has long grappled with issues of time inefficiency and accuracy. This thesis introduces an innovative approach to mitigate these challenges by employing a Nextflow workflow to parallelize computations on clusters. The workflow reduces the computation time and facilitates the execution of large-scale experiments and evaluations, thus enabling the possibility of improving the precision and recall predictions of QCxMS by tuning the setup of the simulations. Beyond efficiency enhancements, extensive experiments were conducted to assess the predictive capabilities of QCxMS, identifying the parameters that outperform the default settings of QCxMS.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Metabolomics

Metabolomics, a pivotal facet of systems biology, delves into the study of small molecules ($< 1500 Da$) within biological systems, offering profound insights into cellular processes, biomarker identification, and disease understanding [1], [15]. Metabolites serve as intermediate or end products of metabolism, influenced by the action of multiple enzymes. The metabolome undergoes constant changes, influenced by internal factors such as the proteome and genome, as well as external factors like the environment, lifestyle, medications, and diseases. It is regarded as the most accurate reflection of the phenotypic response, allowing researchers to monitor subtle organismal changes [17].

Metabolomics finds applications in diverse fields, including plant biology, nutrition, animal breeding, drug discovery, and the study of human diseases. One common application is biomarker discovery, identifying metabolomic differences between groups in response to specific conditions [16].

Metabolomic experiments follow either a targeted or untargeted approach. Untargeted studies involve the simultaneous measurement of all metabolites in a sample, aiming to identify and potentially discover new ones without a prior hypothesis. Targeted experiments, on the other hand, are hypothesis-driven, focusing on specific predefined groups of known metabolites related to a particular metabolic reaction. Targeted metabolomics offers high precision and accuracy, using internal standards and controlled conditions, often serving to validate untargeted analysis.

What sets metabolomics apart is its ability to conduct studies comparing two groups of metabolomics without establishing a prior hypothesis. Untargeted metabolomic analyses allow scientists to collect data to analyze global metabolic changes. As of now, two main measurement techniques for metabolomics data acquisition are Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS), coupled with various separation techniques such as Liquid Chromatography (LC), Capillary Electrophoresis (CE), Ion Mobility Spectrometry (IMS), among others [1].

1.2 Mass Spectrometry

MS is a widely employed analytical technique that detects the mass-to-charge ratio (m/z) and abundance of ions to identify and quantify molecules in simple or complex mixtures [14]. The advent of high-resolution mass spectrometers has significantly enhanced our understanding of metabolites in cellular and biological pathways, catalyzing advancements in metabolomics [12]. The process involves ionizing molecules in the gas phase, followed by the separation and detection of ions based on their m/z [14].

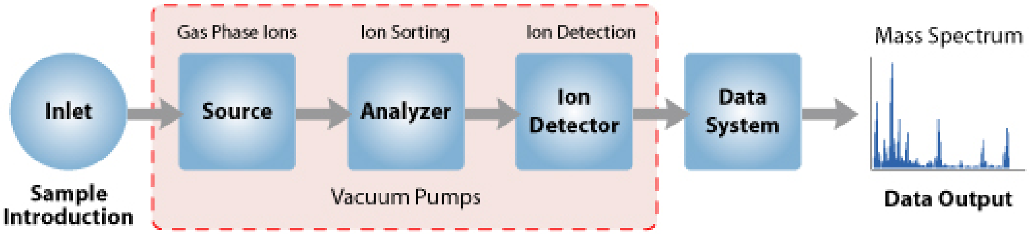


Figure 1.1: Mass Spectrometer Components

A Mass Spectrometer comprises the ion source, mass analyzer, and detector, as shown in Figure 1.1. Samples, introduced in liquid or gas form, undergo vaporization and ionization in the ion source, as the mass spectrometer exclusively measures charged particles. The formed ions gain kinetic energy, moving towards the mass analyzer. Electric and/or magnetic fields from the mass analyzer deflect individual ions' paths based on their m/z and propel them toward the detector. The detector records induced charge or produced current when ions pass by or hit a surface. The mass spectrometer is linked to a computer with specialized software for data analysis, generating a mass spectrum—a plot of intensity vs. m/z representing the outcomes [14].

Several ionization sources, such as electron ionization (EI), electrospray ionization (ESI), and matrix-assisted laser desorption/ionization (MALDI) exist [16]. Various Mass Analyzers, including Time-Of-Flight (TOF), orbitraps, quadrupoles, and ion traps are commonly used, each with its own characteristics [16].

Time-of-flight mass spectrometry (TOF-MS) utilizes a TOF mass analyzer to determine an ion's m/z based on its TOF to the detector. This method relies on the principle that ions with the same energy will travel at different velocities based on their mass. After ionization, an electrostatic field accelerates ions so that those with the same charge acquire the same kinetic energy. Consequently, the velocity of each ion, and thus the time it takes to reach the detector through the flight tube, can be used to determine its m/z [7].

However, MS has a limitation in that it can only identify the mass of a molecule and not its structure. This means that there can be multiple molecules with the same mass, making it difficult to determine the exact identity of a molecule based solely on its mass. Therefore, there is a need for an approach that can accurately identify the structure of molecules, going beyond just their mass.

1.3 Tandem Mass Spectrometry

Tandem Mass Spectrometry (MS/MS) is a powerful analytical technique that extends the capabilities of traditional mass spectrometry [11]. As shown in Figure 1.2, MS/MS involves the sequential use of two mass analyzers to obtain additional information about the structure and composition of molecules [4]. The first mass analyzer selects a specific m/z window and optionally a separation time unit window and collects all ions within those two windows from the sample in a collision cell. Those ions are then fragmented into smaller fragments by applying a voltage and a gas, usually dinitrogen [4]. Those fragments are then released under the MS detector that provides the m/z and abundance tuples of the ions coming from the collision cell, providing a more detailed understanding of the molecular structure [4].

MS/MS is particularly valuable for identifying and characterizing small molecules, providing insights into their fragmentation patterns and structures. When identifying molecules with the same m/z, MS/MS can provide information about fragmentation pat-

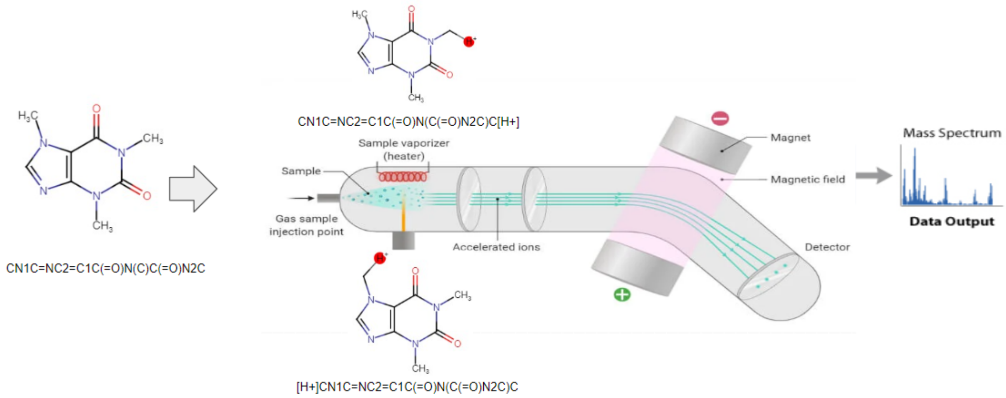


Figure 1.2: Tandem Mass Spectrometry

terns that aid in identification. In proteomics, MS/MS is widely used for peptide sequencing and post-translational modification analysis. This technique enhances the specificity and sensitivity of mass spectrometric analysis, enabling researchers to elucidate the intricate details of molecular structures [4].

1.4 Molecular Dynamics

Molecular Dynamics (MD) is a computational technique used to simulate the movements and interactions of atoms and molecules over time [10]. By solving Newton's equations of motion, MD simulations provide insights into the dynamic behavior of biological macromolecules, such as proteins and nucleic acids. This allows researchers to study how these molecules evolve and interact under different conditions [10].

MD simulations are crucial for understanding the structural dynamics of biomolecules, including their folding, unfolding, and conformational changes [10]. The simulations take into account forces between atoms, allowing for the observation of molecular behavior at the atomic level [9]. This computational approach has become an integral tool in structural biology, drug discovery, and materials science [10].

1.5 QCxMS

Current databases do not contain MS/MS fragmentation of most of the molecules due to the time and money limitation to experimentally analyze the complete set of molecules, especially for those molecules that have not been discovered yet [6]. One alternative to overcome the lack of MS/MS fragmentation data in databases is the prediction of mass spectra. Two prominent approaches have emerged to address this gap: machine learning and QCxMS.

While machine learning techniques have demonstrated efficacy in predicting mass spectra for compounds closely resembling those in the training set, they face limitations. These include challenges in identifying novel peptides and the considerable effort required to generate project-specific libraries [3]. In response, the computational modeling for electron ionization MS (QCxMS) has arisen. QCxMS is an advanced program building upon the Quantum-Chemical Electron Ionization Mass Spectra (QCEIMS) specifically tailored for electron ionization (EI) and collision-induced dissociation (CID) modes [8]. It utilizes Born-Oppenheimer ab initio molecular dynamics (BO-AIMD) simulations [2] to automatically compute EI mass spectra and introduces a groundbreaking CID module. In CID, QCxMS achieves chemical activation through collisions of precursor ions with neutral gas atoms, providing the capability to simulate controllable fragmentation rates for detailed structural characterization. This innovation positions QCxMS as a key tool in MS fragmentation prediction, enabling the routine calculation of CID spectra based solely on molecular structures as input, thereby advancing substance identification and chemical research. This unique approach provides detailed insights into collision kinetics, fragmentation pathways, and temperature-induced decomposition reactions. Notably, QCxMS is the first standalone Molecular Dynamics (MD)-based program capable of predicting mass spectra solely based on molecular structures [8].

However, the computational demands of QCxMS present significant challenges due to prolonged processing times. Recognizing the necessity to address this constraint, our research embarks on a strategic solution: we introduce a meticulously crafted Nextflow workflow by systematically exploring methods to overcome the computational hurdles. This innovative workflow leverages parallelization techniques, targeting the most time-consuming processes of QCxMS, thereby substantially reducing computation times and enabling systematic experiments to tune the parameters of QCxMS. The subsequent evaluation of QCxMS performance and the formulation of parameter tuning recommendations become pivotal components of our research, aiming to enhance the efficiency of this promising analytical approach. This thesis shows a dedication to improving how we use QCxMS workflow to predict mass spectrometry data. The results suggest a future where we tackle computational challenges, making accurate mass spectrometry predictions a reality.

1.6 Goals of the Thesis

The objectives of this thesis encompass a comprehensive examination of the existing state of Quality Control by Mass Spectrometry (QCxMS). The primary aims are as follows:

1. **Evaluation of Current QCxMS Practices:**

   Conduct survey and an assessment to analyze the primary impediment within QCxMS, focusing on the computational costs involved.

2. **Workflow Development:**

   Propose and establish a robust architectural workflow tailored to facilitate systematic and efficient analyses within the QCxMS domain. This workflow will serve as a structured foundation for the subsequent phases of the research.

3. **Evaluation of QCxMS Performance:**

   Use the workflow to perform systematic QCxMS analysis to find its current performance.

4. **Refinement of QCxMS Parameters:**

   Engage in a meticulous process of fine-tuning QCxMS parameters. This involves optimizing the various aspects of the methodology to enhance precision, accuracy, and overall efficacy.

Chapter 2

Methodology

2.1 The Current Stage of QCxMS Practices

The current phase of QCxMS computation time assessment involves addressing a significant computational challenge inherent in QCxMS, which remains relatively unexplored in existing research. To elucidate the extent of this challenge, our initial focus is on quantifying the computational demands imposed by QCxMS. To achieve this, we have systematically selected molecules spanning a range of 3 to 77 atoms. Subsequently, our workflow is employed to compute the comprehensive calculation time required for QCxMS predictions across these molecules.

The results show that in the context of Collision-Induced Dissociation (CID) in QCxMS, the CPU-hours usage is modeled using a quadratic function, as depicted in Figure 2.1. The high R2 score of the fitting line (0.987) indicates a strong correlation, suggesting that the CPU-Time growth follows a polynomial pattern.

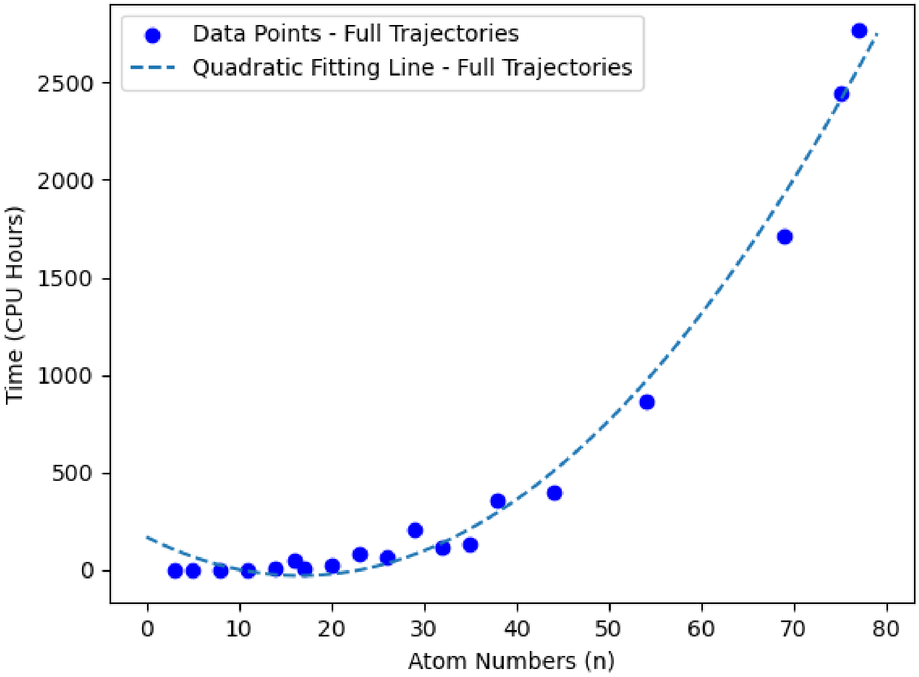
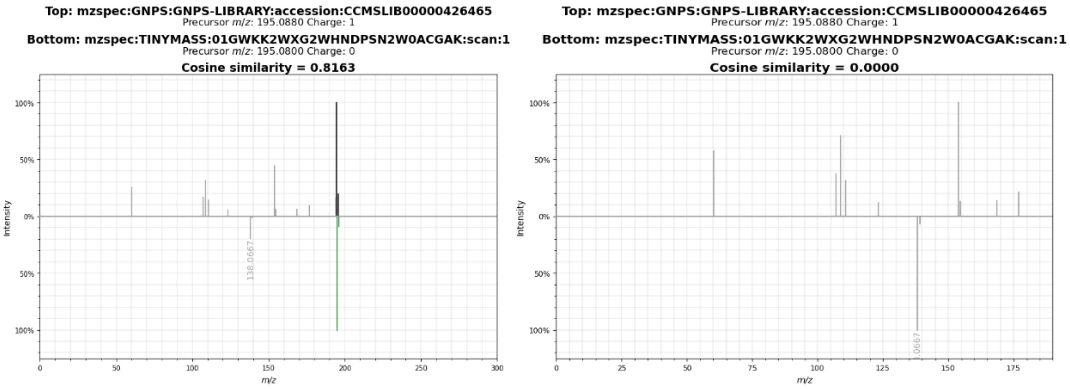


Figure 2.1: Original QCxMS computation time

During our investigation into the accuracy of QCxMS, we found only one paper that evaluates this aspect, reporting an accuracy of 0.68. However, we identified a methodological issue with this study. Firstly, the paper's analysis is limited to small molecules with fewer than 15 atoms, which may not provide a comprehensive benchmark as molecule size increases. Furthermore, the paper fails to address the exclusion of m/z values close to the precursor or those less than 50 Da. This omission is noteworthy as these m/z values can significantly impact the cosine score, a key measure of similarity.

To illustrate, in Figure 2.2, we use caffeine as an example. The bottom mass spectrum represents the simulated result, while the upper mass spectrum represents the experimental result. As shown in Figure **??**. a, the high cosine similarity observed between these spectra is primarily due to the matching precursor m/z, leading to an artificially inflated cosine score. However, the presence of a precursor m/z indicates that the molecule has not undergone fragmentation and therefore lacks MS/MS (tandem mass spectrometry) data. As mentioned earlier, the Orbitrap method cannot provide much information when the m/z is less than 50 Da. Therefore, to ensure the best setting for all experimental conditions, it is essential to remove m/z values less than 50 Da. Additionally, removing m/z values greater than m/z - 17 Da. is crucial for an accurate assessment.

In this thesis, we will address this by incorporating the removal of these m/z values in our methodology to ensure a more accurate evaluation of QCxMS results.



a. Simulated caffeine result vs. experimental result

b. Simulated caffeine result vs. experimental result
after removing the removing noise peaks

Figure 2.2: Current stage of QCxMS accuracy

2.2 QCxMS Workflow

In addressing the computational challenges posed by the time-intensive nature of QCxMS, the project is focused on optimizing the simulation of collision molecular dynamics (MD) for each trajectory. Traditionally, this involves employing a single core for trajectory simulation, leading to significant time overhead due to limitations in available cores on standard computers.

To overcome this challenge, we have developed a strategic approach leveraging a computer cluster for expedited calculations. This entails distributing the simulation tasks across multiple cores within a specialized computer cluster, mitigating the time constraints associated with single-core simulations. This innovative methodology significantly enhances the computational speed and overall efficiency of QCxMS analysis, making it more feasible for large-scale experiments and real-time applications.

Our meticulously designed workflow provides a user-friendly solution for MS/MS data analysis. Users can input a file containing molecules along with the specific conditions (parameters) they wish to apply to each molecule. The workflow parses these molecules and initiates QCxMS simulations, taking into account their conditions. Subsequently, the workflow invokes the respective QCxMS execution programs for each molecule to calculate the MS/MS.

A key feature of our workflow is the automatic deployment of each QCxMS simulation on a computer cluster. This strategic use of parallel processing significantly enhances the efficiency of the simulations, making it feasible for large-scale experiments. After the experiment, the results are consolidated into a single Mascot Generic Format (mgf) file and a .tsv file. The MGF file contains the predicted MS peaks and intensities and the .tsv file contains the substructures related to each MS signal predicted. Each molecule is represented as a scan within this file, providing a well-organized and comprehensive output for further analysis. The overall architecture of the workflow is shown in Figure 2.3 Below we show each process of our workflow.

Molecule Preparation

On one hand, the workflow parses an input file containing molecular structures in SMILES notation and their corresponding experimental conditions. Through a systematic parsing mechanism, the workflow organizes these molecules into distinct folders, facilitating subsequent processes. This transformation converts the molecular structures into .mol files, a widely recognized format for chemical structures representing the actual elements of the

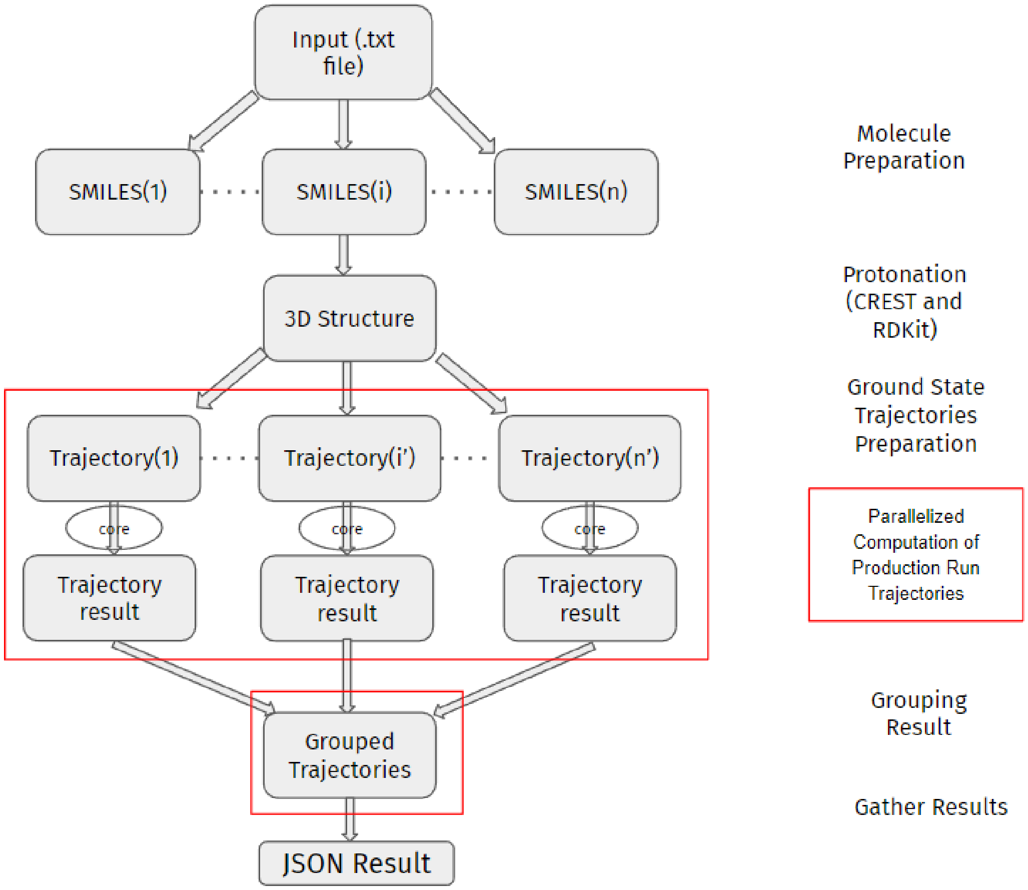


Figure 2.3: Workflow architecture

molecule as a graph. On the other hand, the experimental parameters are documented in a separate file, enabling QCxMS to recognize and utilize them as experiment conditions. The workflow adeptly identifies the parameters shown in Table 2.1.

This process ensures that each molecule is well-prepared and associated with the necessary experimental parameters for further analysis and experimentation.

Protonation

Protonation is a crucial step in predicting mass spectra, we have implemented two protonation methods to enhance our prediction accuracy:

- **Original Protonation Method (CREST):** This method begins by identifying the optimal protonation isomer for each molecule. We achieve this by utilizing the protonation tool CREST. The results of this process are saved in .xyz files, which are text-based files containing the three-dimensional coordinates of atoms within the molecule. This format facilitates the representation and analysis of molecular structures that can be recognized by the QCxMS software.

- **Alternative Protonation Method using RDKit:** In our quest for improved mass spectrum prediction, we've developed an alternative protonation approach with the assistance of RDKit, an open-source cheminformatics toolkit widely used in computational chemistry and molecular analysis. With RDKit, we identify all the potential heteroatoms (N, O, S) in the molecule and assign charges to them. Subsequently, RDKit is used to generate the structure of the protonated form, and this information is stored in .tmol files, which are employed in subsequent processes.

Table 2.1: Parameters recognized by the workflow

Parameter Name	**Default Value**	**Required Type**	**Purpose**
SMILES	N/A	String	This is a notation used to describe the structure of a chemical compound in a simple line format.
Energy	30	Integer	This parameter indicates the collision energy or the center of mass energy intended for use in the experiment.
Trajectories	25	Integer	This parameter controls the number of trajectories, where a trajectory represents the path that ions follow as they are accelerated, separated based on their mass-to-charge ratio, and detected within the instrument.
Gas	n2	String	This parameter specifies the neutral gas atom to be used, with options including argon (ar), neon (ne), krypton (kr), xenon (xe), and dinitrogen (n2).
Temperature	500	Integer	This parameter sets the simulation temperature in Kelvin (K).
Energy type	elab	String	This parameter specifies the type of energy for the experiment, with options for Collision Energy (CE) and Energy of Center of Mass (ECOM).

These protonation methods aim to enhance the accuracy of our mass spectrum predictions by considering different approaches and leveraging the capabilities of specialized tools and software like CREST and RDKit.

Preparing for Production Runs

Following the preparation of the protonated ion structure, the workflow is initiated to generate ground state (GS) trajectories using the QCxMS software. These GS trajectories, representing the initial state of the simulation, furnish fundamental information for subsequent stages. Subsequently, a second iteration of QCxMS is executed to create individual production run trajectories, each organized within its designated folder. As shown in Figure 2.4, these production trajectories offer a comprehensive exploration of varied parameters or configurations, contributing to a more in-depth analysis and experimentation of the molecular simulations. 2.3 Below we show each process of our workflow.

4. Parallelized Computation of Production Run Trajectories The methodical storage of trajectory data from each production run within dedicated folders is a fundamental aspect of our computational strategy. This organizational framework streamlines the implementation of a computer cluster for concurrent trajectory computations. Each trajectory undergoes processing by an individual core, and our workflow employs QCxMS for Born–Oppenheimer molecular dynamics (MD) simulations. As illustrated in Figure 2.5, we employ the molecule “C12=C(C=CC(=C1)OC(F)(F)F)N=C(S2)N” to exemplify how a cluster can enhance calculation speed. The MD simulations process contains a simulation of hundreds of trajectories and each trajectory is calculated by a core.

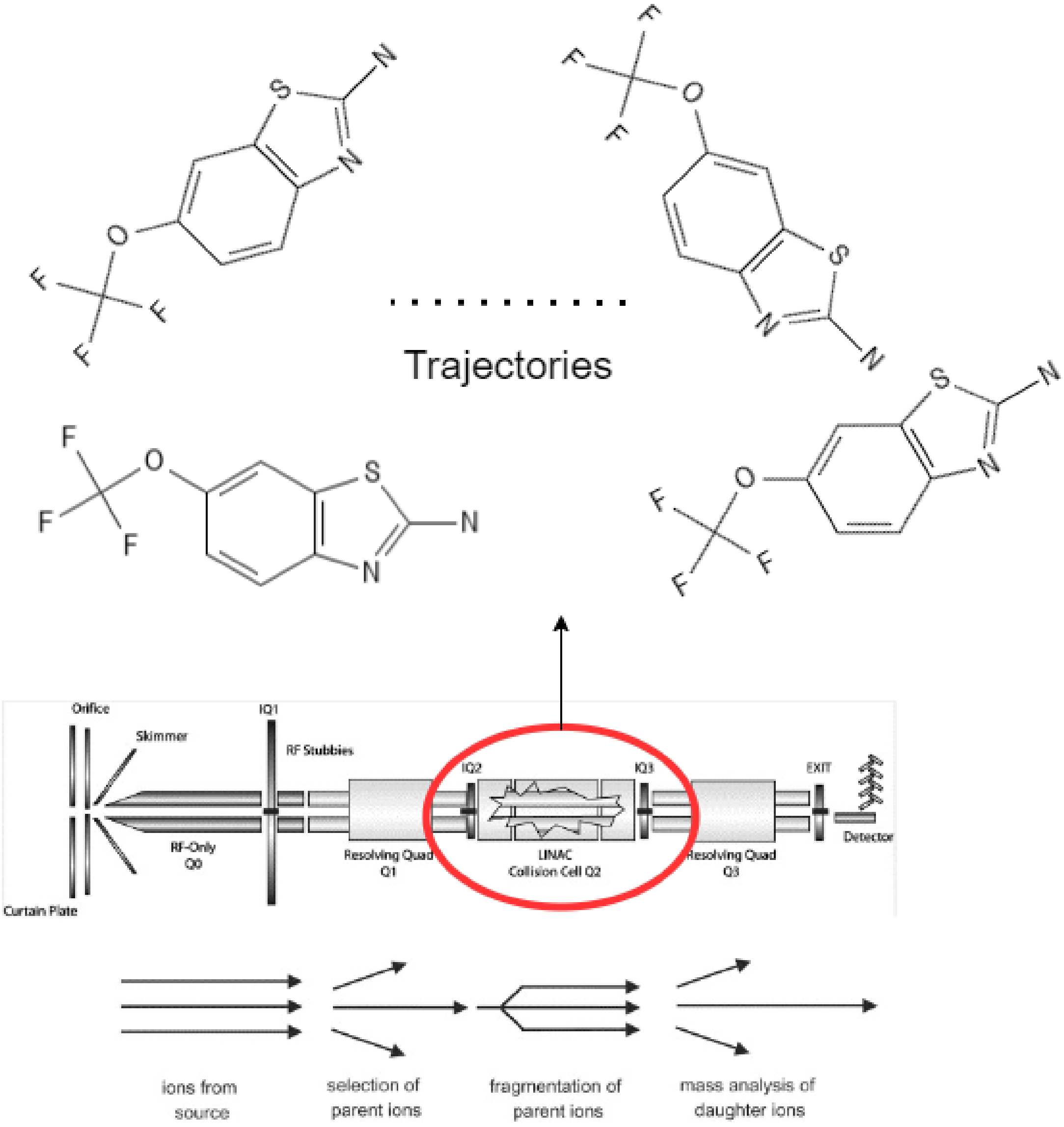


Figure 2.4: Preparation of trajectories

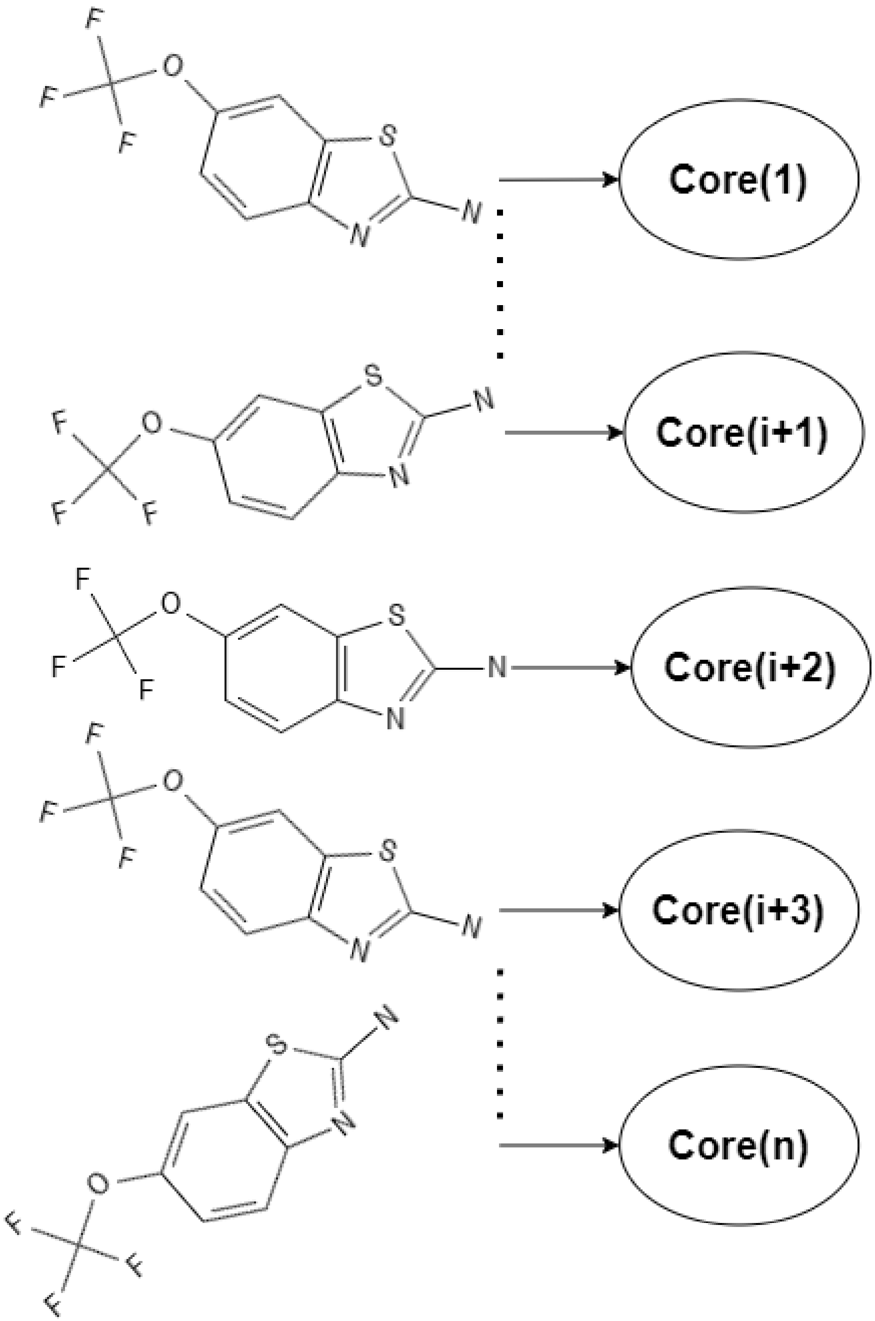


Figure 2.5: Trajectories in cores

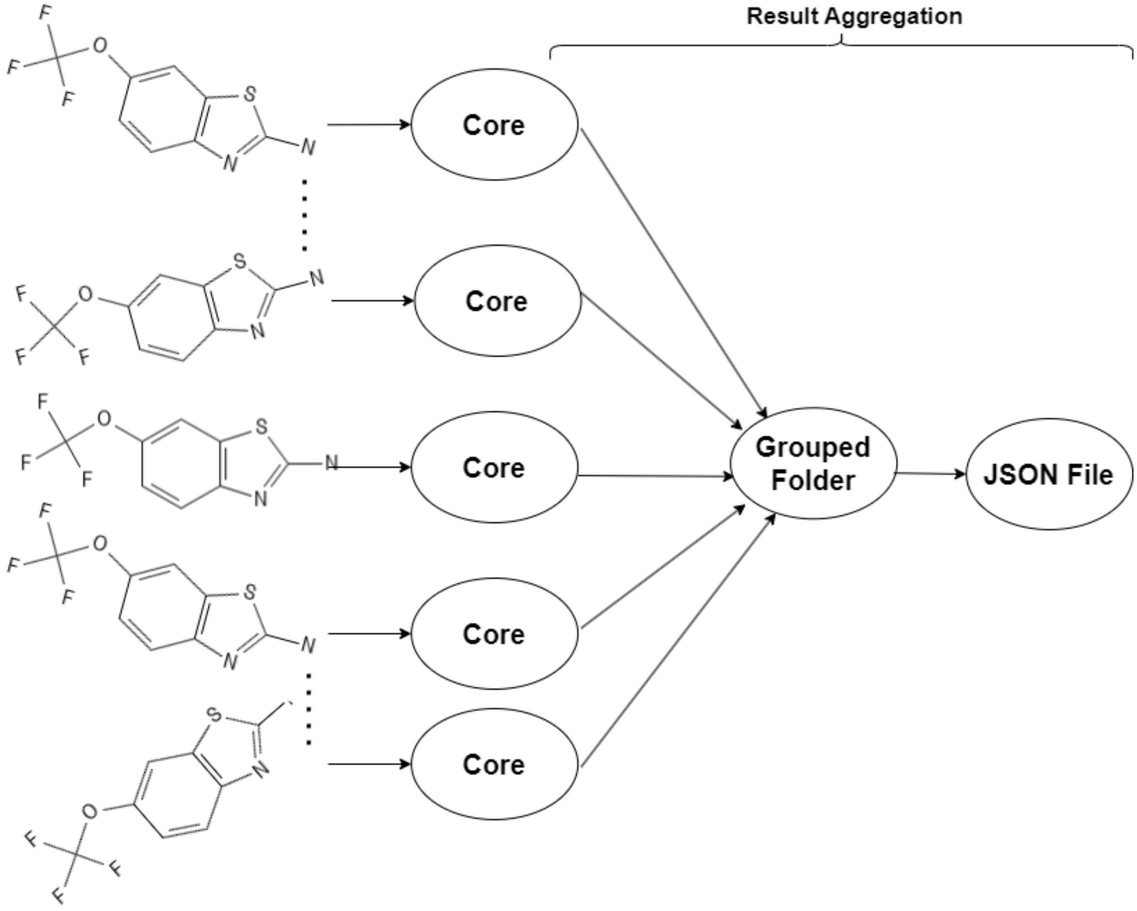


Figure 2.6: Result aggregation

Result Aggregation and Merged Mass Spectra

After the conclusion of collision Molecular Dynamics (MD) simulations, the workflow smoothly transitions to result aggregation. The specific outcomes for each molecule are consolidated within the corresponding folder. Subsequently, we compute the average of all the predicted mass spectra for each molecule and encapsulate it in a JSON file, facilitating organized and efficient data collection. The aggregation process is shown in Figure 2.6.

The JSON results subsequently undergo integration into a unified MGF file, thereby establishing a consolidated repository of mass spectra data derived from the analysis of various molecules. To enhance user readability and comprehension of the fragmentation process, a supplementary .tsv file is provided, documenting all fragment structures associated with identified peaks where substructures are discernible. Additionally, a JSON file is included, detailing the original chemical SMILE, the adduct for the experiment, the charge of the ion, the predicted mass spectrum, and substructures SMILES for each m/z that can be matched to a substructure, offering comprehensive insights into the corresponding substructures. This deliberate structuring of the workflow serves to alleviate the time-intensive nature of QCXMS, thereby facilitating the execution of large-scale experiments and enabling real-time data analysis.

By structuring the workflow in this manner, we have effectively addressed the time-consuming aspect of QCXMS, enabling large-scale experiments and real-time data analysis. The use of a computer cluster, coupled with the parallelization of tasks, has significantly expedited the simulation of collision MD, making it a practical solution for high-throughput MS data analysis.

2.3 Evaluation Data Set

We randomly selected a set of 20 molecules from the GNPS-SELLECKCHEM-FDA-PART2 library, deliberately chosen to exhibit a gradation in the number of atoms. The atom count spans from 20 to 39, with each successive molecule containing one additional atom compared to its predecessor. For clarity and reference, we formally designate this set as ”set 1.” A detailed listing of these molecules and their respective atom counts is provided in Table 2.2

Table 2.2: Set 1

Molecule	**Atoms**
C12=C(C=CC(=C1)OC(F)(F)F)N=C(S2)N	20
C1=CC=C2C(=C1)C(C(=CC2=O)C)=O	21
C1N(C(N(C=1)C)=S)C(=O)OCC	22
C1(CC(N(N=1)C2=CC=CC=C2)=O)C	23
C1C(OC2=C(C=1)C=C3C(=C2OC)OC=C3)=O	24
C1=CC=C(C(=C1F)CN2N=NC(=C2)C(=O)N)F	25
C1=C(C=CC(=C1)[C@@H](CC(O)=O)CN)Cl	26
C1=C(C=CC(=C1)S(NC2SC(=NN=2)C)(=O)=O)N	27
C1=CC(=C(C=C1)OCC(CO)O)OC	28
C1(N(C(C2C=CC=CC1=2)=O)C3C(NC(CC3)=O)=O)=O	29
C1(=CC2=C(C(=C1)O)C(C(=CO2)C3=CC=C(C=C3)O)=O)O	30
N1(C2=C(C(CC3=C1C=CC=C3)=O)C=CC=C2)C(N)=O	31
C1=CC(=C2C(=C1)C(N(C2)C3C(NC(CC3)=O)=O)=O)N	32
C1(=CC=C2C(=C1)NC(=N2)NC(OC)=O)SCCC	33
N1(C2CC2)C3=C(NC(C4=C1N=CC=C4)=O)C(=CC=N3)C	34
C1(C(=O)C2=CC3=C(C=C2)N=C(N3)NC(=O)OC)=CC=C(C=C1)F	35
C1=C(C=C2C(=C1)N3C(CN(C2=O)C)=C(N=C3)C(=O)OCC)F	36
C(/C(=C/C1C=C(C(=C(C=1)[N+](=O)[O-])O)O)CN)(=O)N(CC)CC	37
C1=C(C=CC(=C1)C(=O)O)S(N(CCC)CCC)(=O)=O	38
C(C[C@@H](C(N[C@H](C(OC)=O)CC1=CC=CC=C1)=O)N)(O)=O	39

In our pursuit of identifying the optimal trajectory, we selected five molecules of varying sizes from the GNPS-SELLECKCHEM-FDA-PART2 library. We call these molecules “set 2” and they are presented in Table 2.3.

Table 2.3: Set 2

Molecule	**Atoms**
N1=CSC=C1C2NC3=C(N=2)C=CC=C3	21
C1=CC=C(C(=C1F)CN2N=NC(=C2)C(=O)N)F	25
C1CCN(CC1)C2=CC(=[N+](C(=N2)N)[O-])N	30
C1(C(=O)C2=CC3=C(C=C2)N=C(N3)NC(=O)OC)=CC=C(C=C1)F	35
C(C[C@@H](C(N[C@H](C(OC)=O)CC1=CC=CC=C1)=O)N)(O)=O	39

Table 2.4 shows the SMILES representation of the molecules we used to evaluate the CPU time. We call these molecules “set 3”.

Table 2.4: Set 3

Molecule	**Atoms**
O	3
C	5
CC	8
CCC	11
CCCC	14
CCCCC	17
CCCCCC	20
CCCCCCC	23
CCCCCCCC	26
CCCCCCCCC	29
CCCCCCCCCC	32
CCCCCCCCCCC	35
CCCCCCCCCCCC	38
CCCCCCCCCCCCC	41
C1=CC(=C(C(=C1)Cl)Cl)C(=O)N	16
C(C(C1C(=C(C(=O)O1)O)O)(O)Cl)O	20
O=C(O)C[C@H](NC(=O)c1ccc([N+](=O)[O-])c(OCc2ccc(Cl)cc2)c1)C(=O)O	44
CCCCC[C@H](O)/C=C/[C@H]1C(=O)C=C[C@@H]1C/C=C(=O)O	54
CCC(=O)n1c(=O)n(C2CCN(CCC(CN)(c3ccccc3)c3ccccc3)CC2)c2ccccc21	69
C1(=C(C=C(C(=C1)C(NCC2OCCN(C2)CC3=CC=C(C=C3)F)=O)OCC)N)Cl.OC(CC(CC(O)=O)(C(=O)O)O)=O	75
C1(=C(C(C(=C(N1)C)C(=O)OC(C)C)C2C=C(C=CC=2)[N+](=O)[O-])C(=O)OC3CN(C3)C(C4C=CC=CC=4)C5C=CC=CC=5)N	77

2.4 Hardware Resources

The efficiency of QCxMS is contingent upon robust computational resources and substantial RAM allocation. The evaluation of QCxMS performance is facilitated through the streamlined workflow outlined earlier, optimizing execution on diverse computer clusters. Our experimental endeavors primarily unfolded within the following environments:

GCP (Google Cloud Platform):

- **Configuration and Resources:** In total 400 nodes, each equipped with 2 of the fastest cores, resulting in a total of 800 cores. Each node possesses a memory limit of approximately 15.12 GB.
- **Utilization:** To manage power consumption effectively, we have restricted the number of nodes to 248, providing access to around 500 cores for experimentation

HPCC (High-Performance Computing Cluster):

- **Configuration and Resources:** A robust and stable platform provided by UCR that is configured with a RAM limit of 8 GB per core, granting continuous access to 128 cores.
- **Utilization:** We can access up to a total of 128 cores on HPCC.

Own Slurm Cluster:

- **Configuration and Resources:** Independently designed and implemented Slurm cluster with a total of 128 cores.
- **Utilization:** Typically, we leverage 100 cores from this cluster for our experiments.

In aggregate, we have access to a computational infrastructure comprising over 700 cores, each with 8 GB of RAM, enabling the execution of comprehensive experiments. These resources are instrumental in advancing our understanding of QCxMS processes and their performance characteristics.

After designing, implementing and testing the platforms to run QCxMS, our goal now is to evaluate QCxMS's performance and identify and fine-tune parameters that can significantly enhance the fragmentation prediction, surpassing the default configurations.

2.5 Data Processing Techniques

After determining the most time-efficient trajectory, our next objective is to identify the key parameters that significantly influence the simulation results. To achieve this, we must compare the simulation results to the experimental data. This process necessitates thorough data preprocessing to ensure meaningful and accurate comparisons.

In our analysis of simulation results, we implement noise reduction techniques for both experimental and predicted data. These techniques are designed to enhance the quality and clarity of the data we work with, ultimately leading to more precise results.

- **For Experimental Results:** In the context of our experimental results, we implement a denoising method, filtering out peaks with intensities less than three times the smallest intensity in the spectrum. Additionally, we exclude peaks with mass-to-charge ratios (m/z) greater than the precursor mass minus 17 Da and those with m/z less than 50 Da. The decision to set the denoising threshold above 50 Da is rooted in careful consideration of instrument limitations. This strategy aims to eliminate noise or unreliable signals in regions where mass spectrometers, such as the Orbitrap, may experience reduced sensitivity or operational constraints. By doing so, we ensure the reliability and accuracy of the detected signals [9]. This meticulous process guarantees the utilization of clean and dependable experimental data.

- **For Prediction Spectrum:** Data processing for predicted spectra is more extensive. In addition to excluding peaks with mz values exceeding the precursor mass minus 17 Da, we further eliminate peaks with mz values less than 50 Da. These data processing steps are vital to ensure the quality and relevance of our evaluation datasets, which are essential for achieving accurate and meaningful mass spectrum analysis.

Additionally, when processing the predicted spectra, we adjust the peaks to match the lowest peak count (k) among all experimental conditions after noise reduction. This ensures that we select the best parameter based on the one that maximizes the highest-quality peaks in the spectrum, rather than merely increasing the number of peaks.

2.6 Tune-up Parameters

In real-world mass spectrometry experiments, various experimental parameters can be subject to manipulation, encompassing factors such as Collision Energy (CE), gas type, gas pressure, experiment temperature, chamber length, and other relevant variables [3]. The primary objective of our investigation is to assess the predictive capabilities of QCxMS under default parameter settings. Additionally, we aim to explore the potential for optimization by identifying alternative parameter configurations that yield superior results compared to the default settings. This exploration is integral to enhancing the precision and reliability of the QCxMS predictions, ultimately contributing to the refinement of mass spectrometry methodologies and analytical outcomes in practical applications.

Table 2.5: Experiment Parameters

Experiment	**Gas**	**Trajectories**	**Energy**
Default	Ar	25*# of Atoms	40 eV
Gas	Ar/N2	200	70 eV
Trajectory	N2	25*#of Atoms/[20 - 500]	70 eV
Collision Energy	N2	200	40/[30 - 100] eV
ECOM	N2	200	(N/A)/3-7
Multiple Proteomers	N2	200 * n proteomers	70 eV

Gas

In our experiments with QCxMS, we discovered that the default neutral gas used is Ar. However, as mentioned in the previous section, in real-world experiments, N2 is more commonly used. To investigate if using the more common gas can improve accuracy, we conducted the first experiment where we replaced the default gas with N2. In this experiment, we set the energy of the gas at 60 eV with 200 trajectories per molecule.

Optimal Trajectory

Optimizing trajectories is a crucial step in our quest to accelerate computation further. It is worth noting that, prior to our experiments, the default trajectory calculation relied on a fixed factor—multiplying the number of atoms in the ion by 25. For instance, a molecule with just 20 atoms would simulate a substantial 500 trajectories for computation. Our research aims to refine this approach and uncover parameters that outperform the default settings, contributing to a more efficient and streamlined computational process. We use the molecules in set 2 for this experiment. These molecules vary in the number of atoms they contain, spanning from 21 to 39. This variation provides an opportunity to explore how the size of a molecule affects its trajectory requirements. In each instance, we used the CREST software to predict a single proteomer. Subsequently, we employed QCxMS to predict the mass spectra of that same proteomer. We conducted ten experiments under consistent conditions, utilizing N2 as the neutral gas and setting the collision energy to 70 eV. The sole parameter manipulated in these experiments was the number of trajectories, ranging from 20 to 500 trajectories. Following each experiment, we computed the pairwise cosine score—a metric of similarity—for each of the ten runs.

Collision Energy

With the optimal trajectory and data processing method in place, our next experiment is dedicated to finding the ideal CE of its profound impact on results [4]. Recognizing that the QCxMS typically sets a default CE at 40 electron volts (eV), our aim is to determine whether we can pinpoint an optimal CE that performs effectively for all predictions. To find the optimal CE, we predicted the mass spectra under consistent conditions, using dinitrogen as the neutral gas and employing 200 trajectories. The sole variable we altered was the simulation CE, which ranged from 30 eV to 100 eV, incrementing by 10 eV each time.

ECOM

During the course of our experimentation, a noteworthy observation emerged in the context of QCxMS. Specifically, we identified a transformation of collision energy to the Energy of the Center of Mass (ECOM) as an initial step. However, we realized that the allocation of collision energy to different molecular masses led to variations in ECOM. This fluctuation in ecom raised concerns regarding its potential impact on prediction accuracy.

To address this concern, we embarked on a systematic investigation to determine whether maintaining a consistent ECOM energy level could yield improved accuracy. In conducting this investigation, a meticulous approach was applied to select five distinct energy levels within the ECOM spectrum, ranging from 3 to 7 with incremental steps of 1. The experimental parameters maintained a consistent framework, incorporating elements such as the utilization of dinitrogen gas and 200 trajectories.

Multiple Proteomers

In practice, the protonation of a molecule is not a unique process. [2] A single molecule may exhibit multiple protonation sites, leading to various proteomers following protonation. However, the default approach employed by the QCxMS method utilizes Crest software to identify the most probable proteomer, thereby imposing constraints on the predictability of the outcomes. Exploring the potential improvement in predictions by considering a broader range of possible proteomers becomes crucial. We systematically protonate the molecule considering all possible proteomers by introducing H+ ions in Atoms N, O, and S. This expands the proteomic landscape, and noteworthy, each proteomer requires an equivalent computation time for analysis.

Chapter 3

Results and Discussion

In this section, we present the experiment outcomes and engage in a comprehensive discussion of the discerned findings.

3.1 Gas Comparison

The experimental results comparing the use of dinitrogen and argon are depicted in Figure 3.1. The findings indicate that the use of dinitrogen yields results similar to those of the simulation. Given that dinitrogen is more commonly utilized in real-world experiments, we have chosen to set dinitrogen as the default gas for further analysis.

3.2 Optimal Trajectories

To visually present the results of the optimal number of trajectories, we utilized the molecules in set 1 to plot a line graph depicting the median cosine score with the increasing

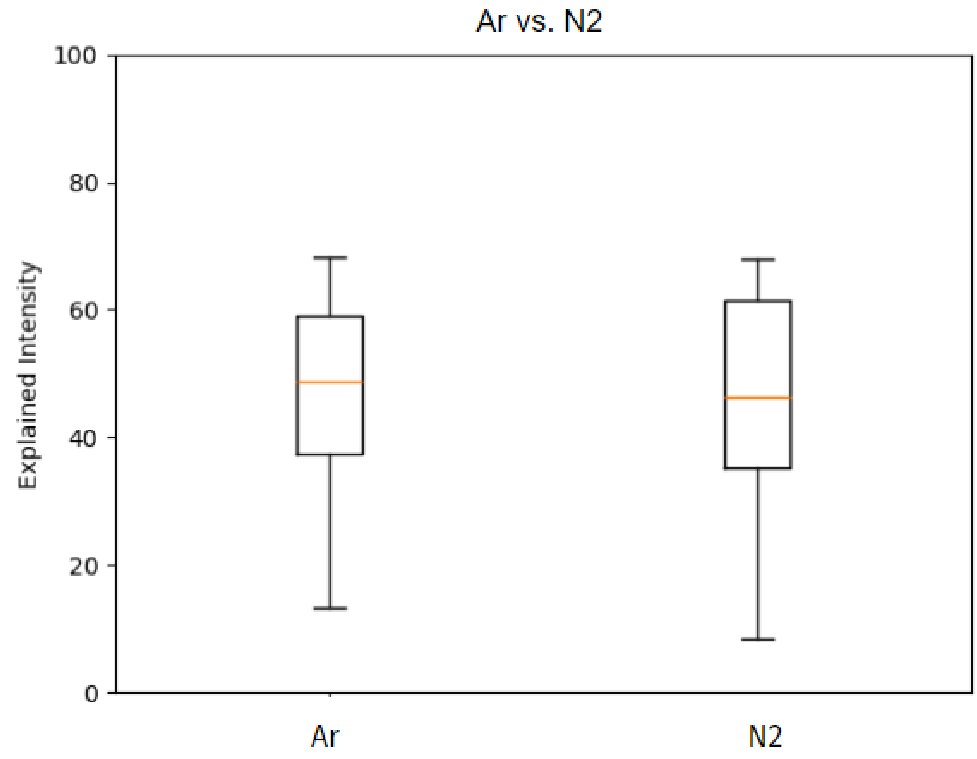


Figure 3.1: Result aggregation

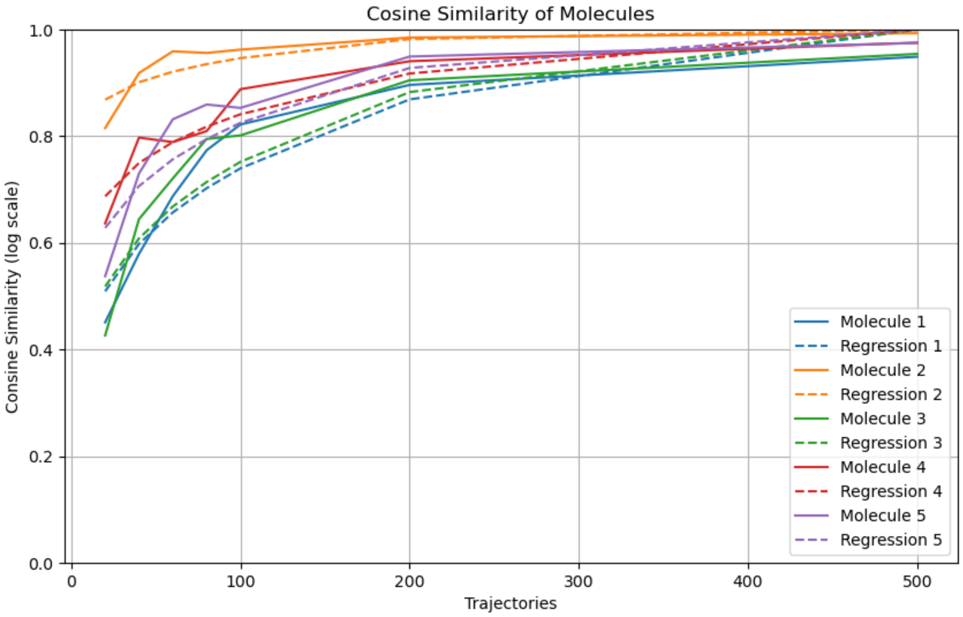


Figure 3.2: Reproducibility

number of trajectories. This analytical approach offers valuable insights into the influence of trajectories on reproducibility.

Figure 3.2 represents the reproducibility of mass spectra predictions. When there are only 20 trajectories, the result shows a median cosine score for 3 molecules that fall below 0.5. This score is considered low in the context of mass spectrum analysis [15]. As we increase trajectories from 20 to 500, we observe a rapid improvement in the cosine similarity between all pair-wised molecules. By the time we reach a trajectory count of 100, all of the molecules achieve a similarity score exceeding 80%, and when the trajectory count reaches 200, all five molecules exhibit a median cosine similarity above 0.9. This indicates a high degree of similarity and consistency in the predictions [15]. Based on this finding, we decided to use 200 trajectories for the remaining experiments, as they provide stable predictions for the molecules.

3.3 Computation Time Result

We compare the computation time between the QCxMS's default parameter and our proposed parameters in Figure 3.3. As mentioned above, our analysis indicates that the current computational time of QCxMS adheres to a cubic function or a quadratic function. This noteworthy observation serves as a foundation for our ongoing efforts to enhance the efficiency of QCxMS predictions. Notably, through our investigation of 200 trajectories, we have identified a point of convergence in QCxMS predictions. This pivotal finding motivates our continued exploration, as illustrated in Figure 3.3, where we present the CPU hours associated with the utilization of 200 trajectories, demonstrating a quadratic growth pattern. This strategic approach underscores our commitment to comprehending existing computational patterns and implementing targeted enhancements, ensuring a more streamlined and effective QCxMS.

3.4 Tune-up Parameters

3.4.1 Collision Energy

To evaluate the spectra predictions, we conducted a comparative analysis between the predicted spectra and the experimentally obtained spectra, utilizing cosine scores as the basis for comparison. The results were visualized in a heat map, illustrated in Figure 3.4.

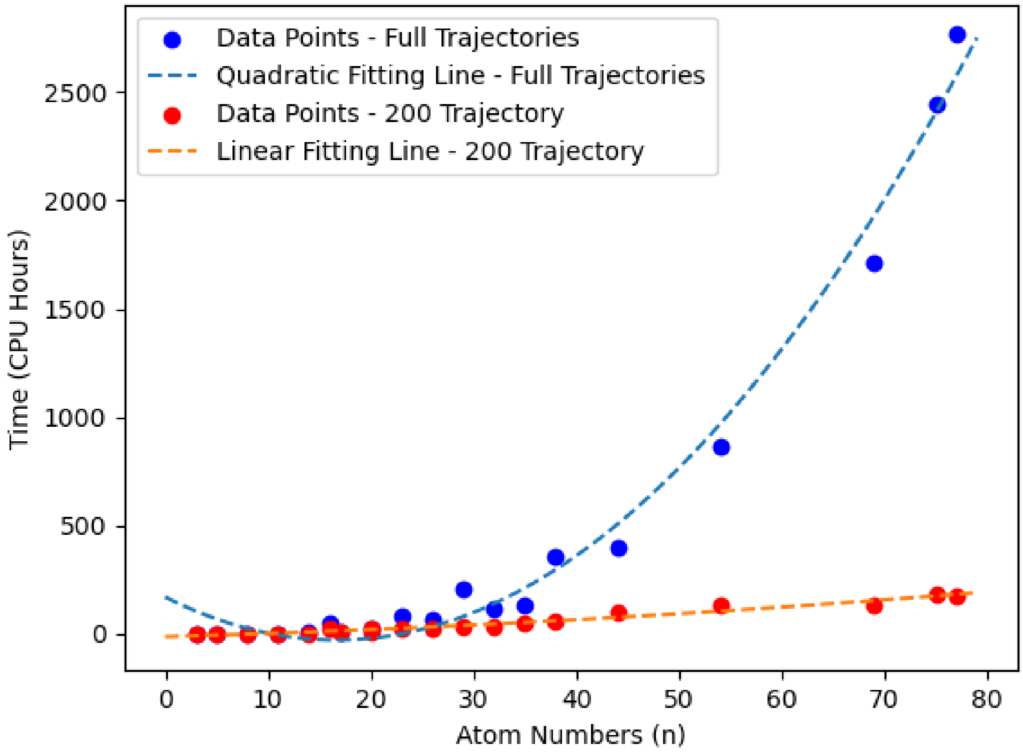


Figure 3.3: CPU hours comparison

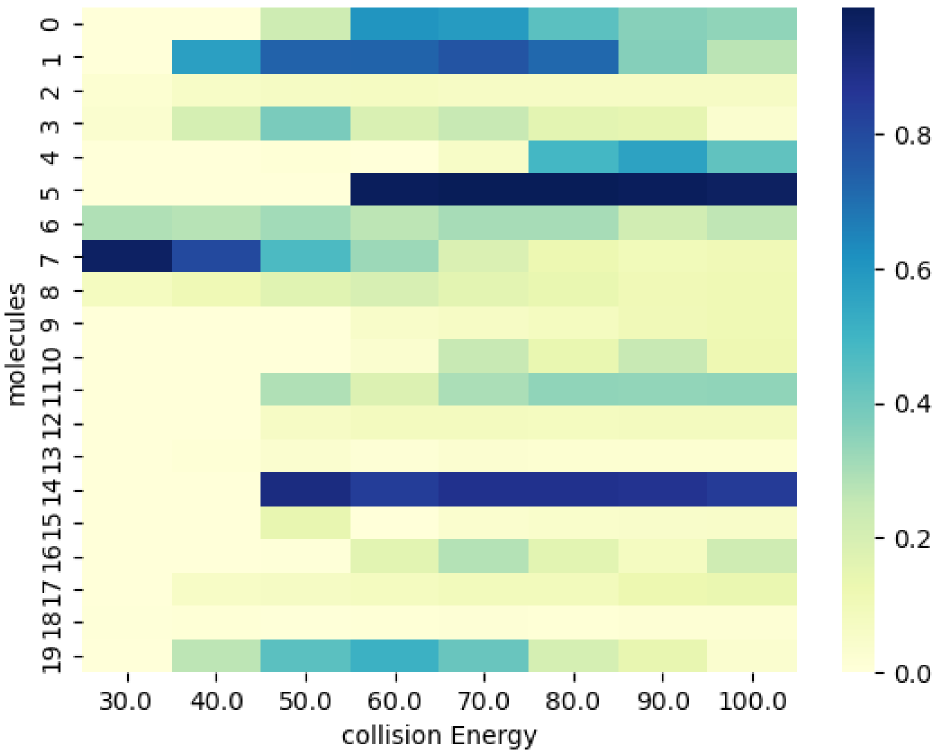


Figure 3.4: Cosine similarity of CE test

This heat map provides a comprehensive view of how varying Collision Energy (CE) values influence the accuracy of our predictions, enabling us to pinpoint an optimal CE setting that ensures reliable results across all molecules.

Upon reviewing the results, we observed a significant issue with the cosine similarity scores. There is not an absolute ”best” score, and most of the cosine scores fall below 0.7, some even registering as completely white. This prompted us to investigate the factors contributing to this low similarity [15].

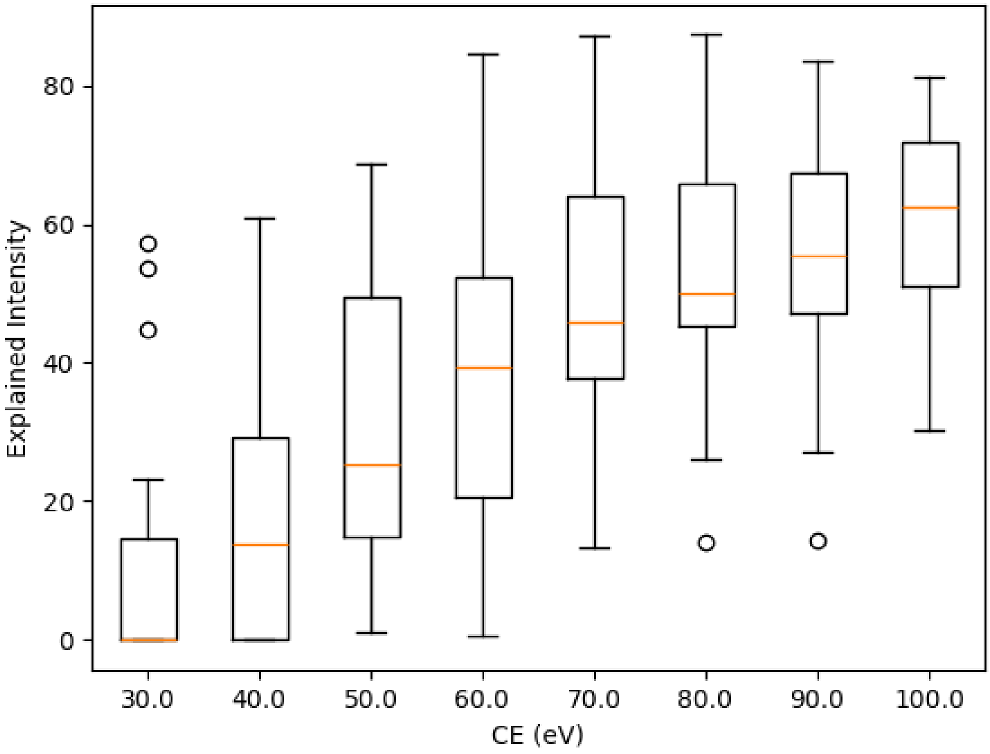


Figure 3.5: Raw EI

We sought to understand whether the issue arose from QCxMS's ability to predict the correct mz (mass-to-charge ratio), or if it was primarily due to poor intensity prediction. To address this question, we used a new metric called ”Explained Intensity” (EI). We calculated EI by summarizing the intensities of the peaks in the experimental data that were successfully predicted by QCxMS and then dividing this sum by the total intensities in the experimental data. This metric serves as an additional means to assess the accuracy of the simulation results by the ability to predict high-intensity peaks.

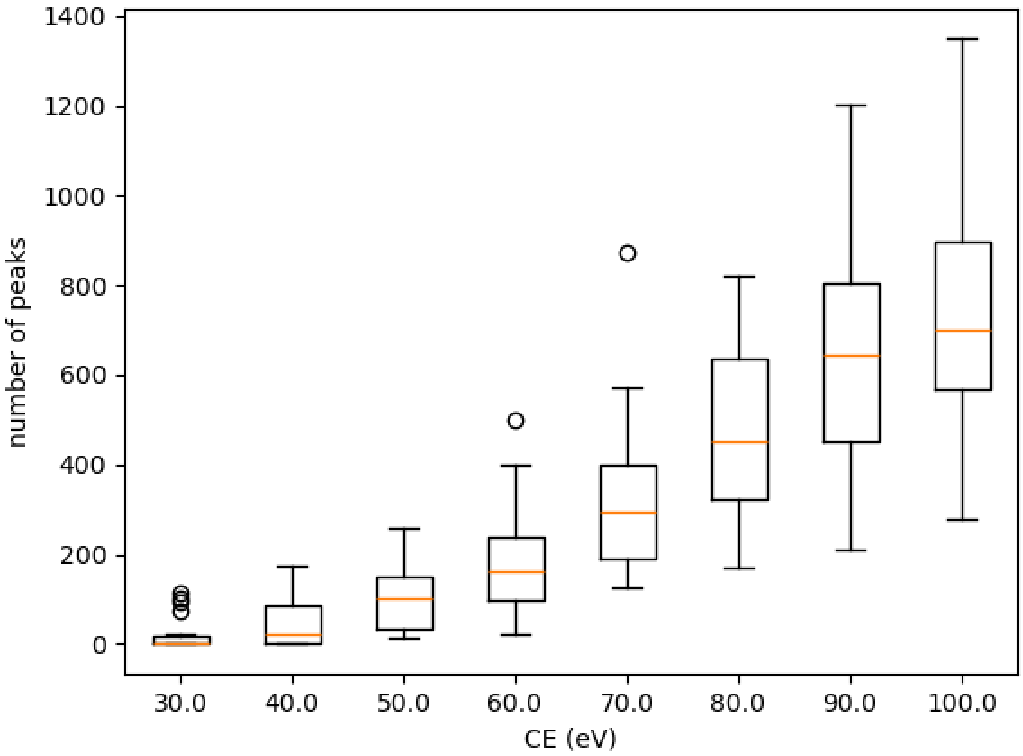


Figure 3.6: Number of peaks

The analysis of CE testing reveals a direct correlation between the increase in CE and the corresponding elevation in EI results. Specifically, the median EI value surpasses 60% across a set of 20 molecules when the CE attains 100 eV as shown in Figure 3.5. Subsequently, a detailed inquiry focuses on the number of peaks predicted by QCxMS, revealing an augmentation that correlates with the increased CE, as illustrated in Figure 3.6.

To discern whether the observed increase in EI is attributable to an enhancement in prediction quality or simply a consequence of a greater number of predicted peaks, a further analysis is conducted. In this investigation, we introduce the concept of a minimum peak threshold, denoted as k, which represents the minimum number of peaks among all energies for all molecules. Subsequently, we select the top k peaks for instances where the predicted peaks exceed this threshold. This approach allows us to evaluate the increase in EI concerning the quality of predictions, thus providing a more nuanced understanding of the interplay between CE adjustments and the predictive capabilities of QCxMS.

In the pursuit of establishing the optimal minimum peak threshold (k), an extensive preprocessing procedure was executed employing the aforementioned techniques. Subsequent analysis revealed that, under a CE of 30 eV, 11 out of the 20 molecules considered exhibited an absence of informative peaks. This number reduced to 7 out of 20 molecules under a CE of 40 eV, where these molecules demonstrated a complete absence of informative peaks. Upon increasing the CE to 50 eV, a noteworthy shift occurred, with all molecules manifesting a minimum of 14 informative peaks. This observation led to the inference that the simulation energy should be selected from energies exceeding the 50 eV threshold, ensuring the availability of informative peaks across all molecules under investigation.

Our investigation revealed a consistent trend across all molecules, indicating that a higher collision energy (CE) consistently results in a greater number of predicted peaks. As a result of this observation, we have opted to select the number of peaks (k) based on simulations conducted at a collision energy of 50 eV for all molecules in our study. This decision ensures a uniform and standardized criterion for peak selection, facilitating a comprehensive and comparative analysis across the dataset.

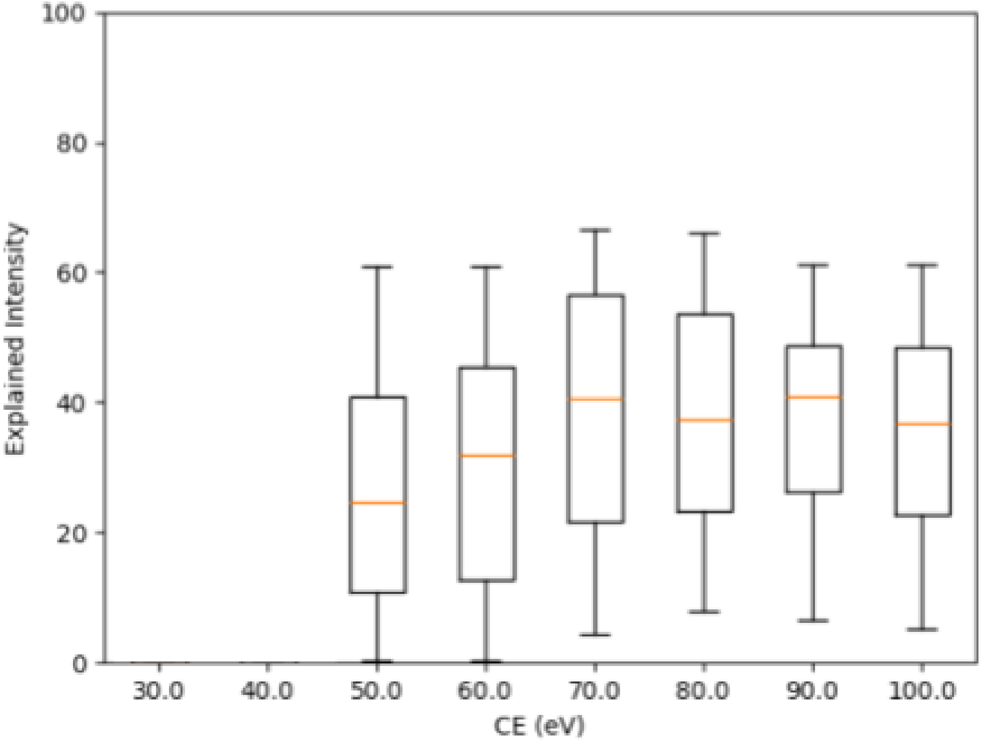
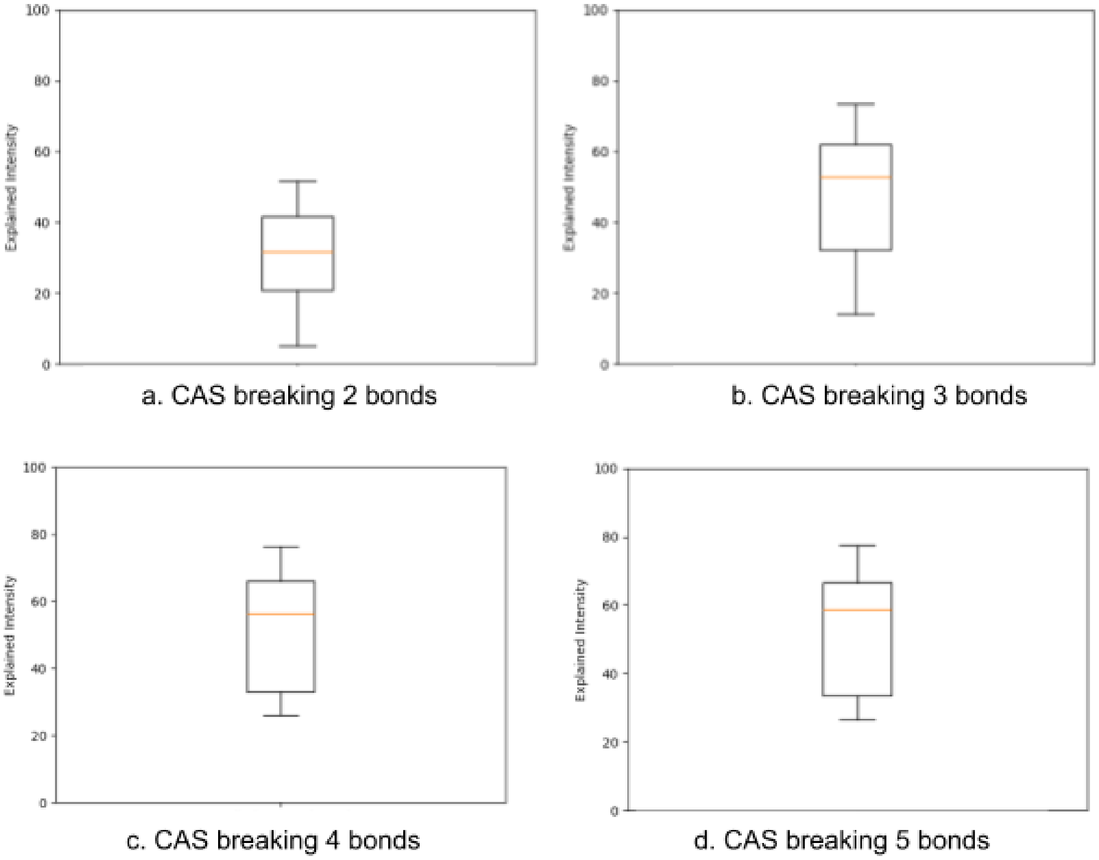


Figure 3.7: EI result after aligning to k peaks

Following the peaks alignment process utilizing an identical number of peaks, the EI results exhibit a more discernible and clarified outcome, as illustrated in Figure 3.7 The result shows there is no significant best collision energy.

We further compare the prediction result with a Compound Annotation Software(CAS) that can predict all the possible substructures by a given molecule's SMILE. We show the result of the CAS by breaking different bonds in Figure 3.8.



a. CAS breaking 2 bonds

b. CAS breaking 3 bonds

c. CAS breaking 4 bonds

d. CAS breaking 5 bonds

Figure 3.8: EI of CAS breaking bonds (before alignment)

However, the outcomes of the CAS and QCxMS currently exhibit disparities in the predicted peaks. Consequently, a direct assertion regarding the superior performance of QCxMS cannot be made. To establish a meaningful comparison between the predictions of QCxMS and CAS, a peak alignment technique is employed, which ensures both methods yield an equivalent number of predicted peaks. For each molecule under consideration, the smaller number of peaks (k') between the two methods is selected. Subsequently, the top k' peaks in the QCxMS prediction are chosen based on intensities, while for CAS, k' peaks are randomly selected, as CAS lacks intensity information. This method ensures a fair and standardized comparison, allowing for a comprehensive evaluation of the predictive capabilities of QCxMS and CAS in metabolite identification.

Figures 3.9 and 3.10 depict the EI result of QCxMS and CAS. The findings indicate that QCxMS can predict EI similar to CAS when the Collision Energy (CE) exceeds 50 eV. However, since CAS is chosen randomly, our analysis reveals that QCxMS does not significantly outperform random selection in terms of overall simulation quality.

3.4.2 Intensities Analysis

To gain deeper insights into predictive quality, we conducted a comprehensive analysis by segmenting the data based on the intensities of the predicted spectra, resulting in the creation of five groups with an equal number of peaks in each group. This division was instrumental in our efforts to evaluate whether the higher intensities of the predicted spectrum are more likely to be the highest intensity peaks in the experimental spectrum.

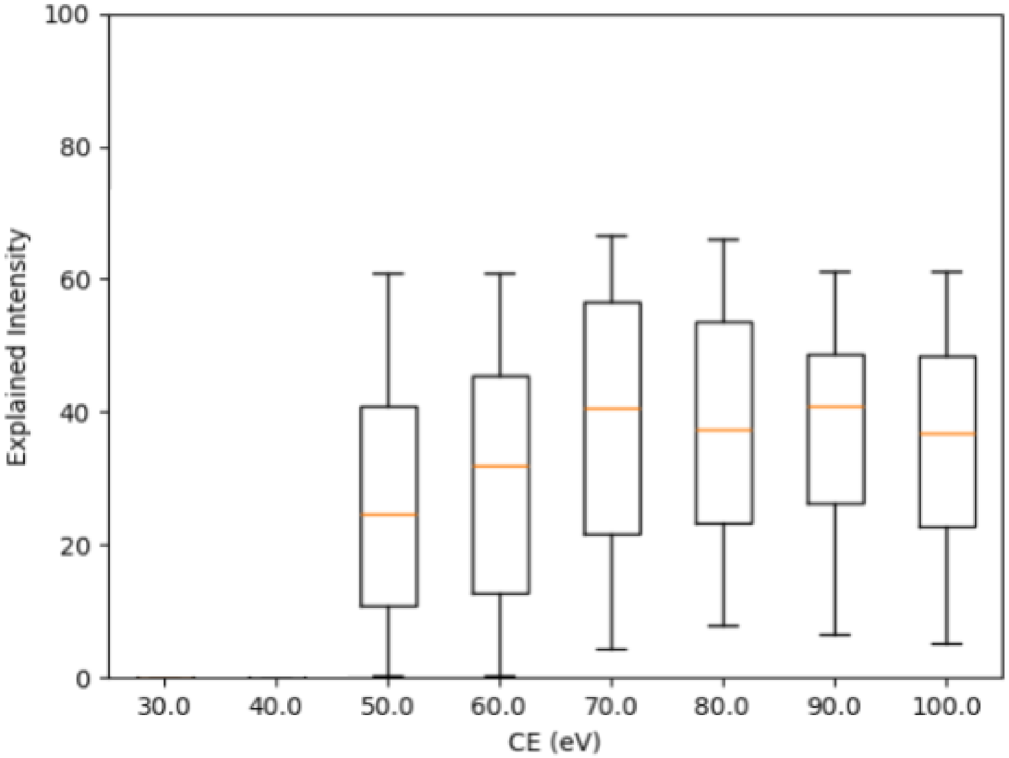
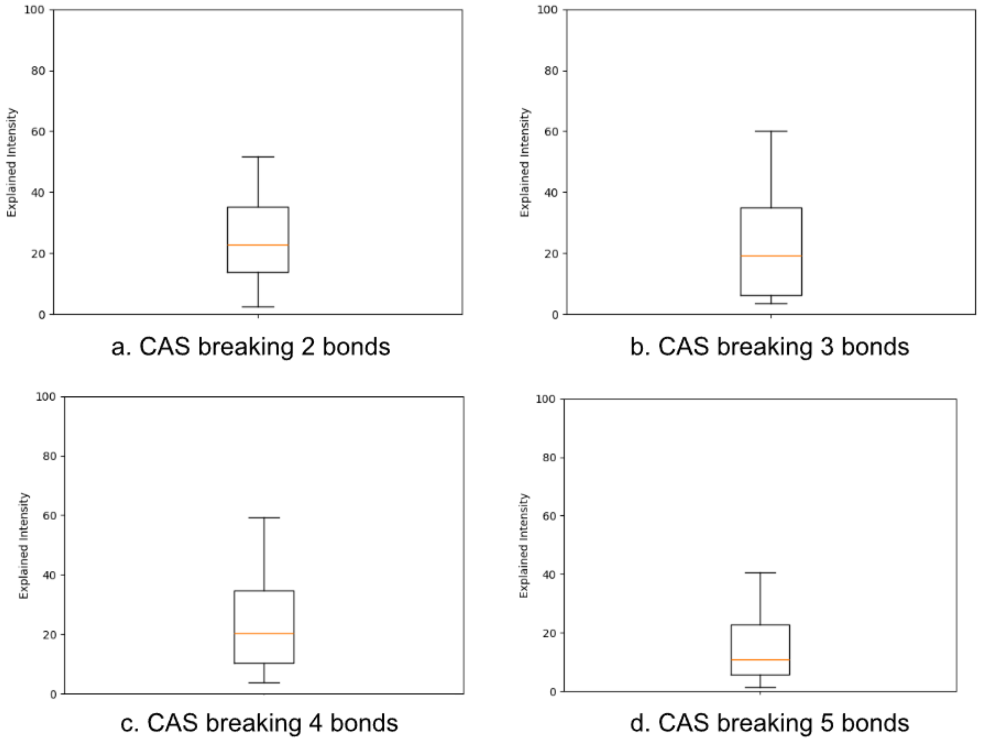


Figure 3.9: EI of QCxMS with k' peaks



a. CAS breaking 2 bonds

b. CAS breaking 3 bonds

c. CAS breaking 4 bonds

d. CAS breaking 5 bonds

Figure 3.10: EI of CAS breaking bonds (Aligned to k' Peaks)

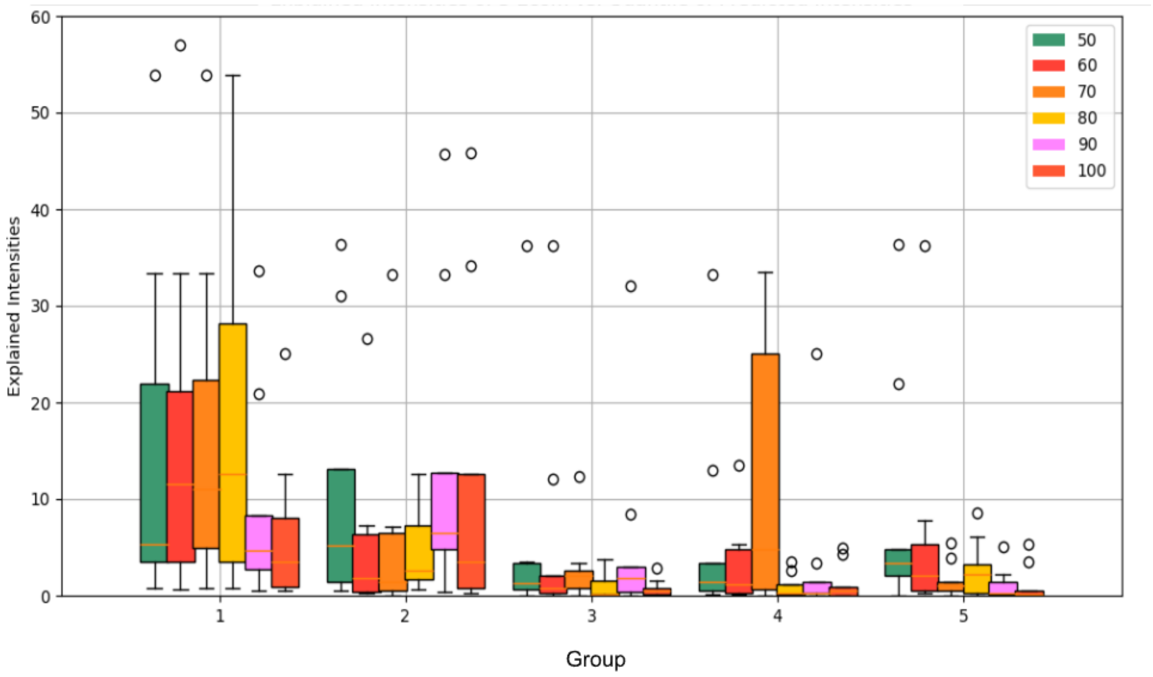
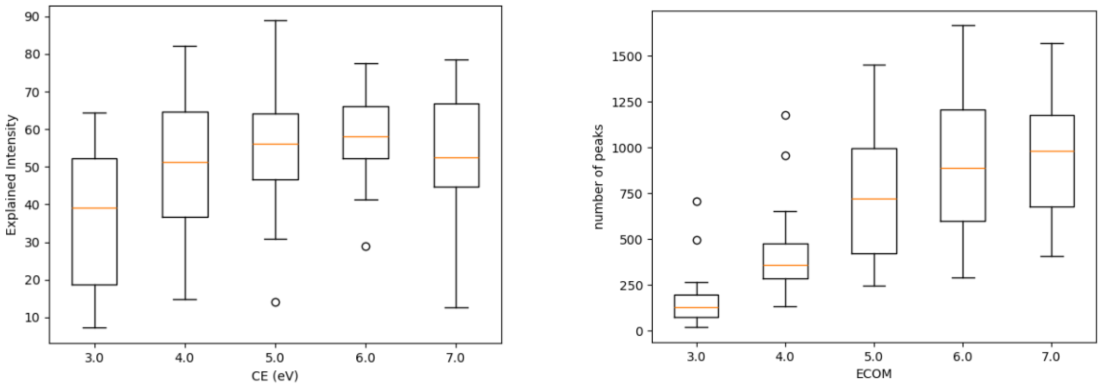


Figure 3.11: Intensities analysis with m = 5

Our initial procedure involves ranking the intensities of the predicted spectrum. Subsequently, we distributed an equal number of peaks among each group. The first group is allocated the highest "m" peaks, followed by an increment of "m" peaks for each subsequent group. This ensures a systematic arrangement with the highest peaks assigned to the initial group and a gradual decrease towards the final group. If a molecule does not have 5m peaks, we simply discard the molecule. With this approach, each group is methodically saved as an individual mgf file, facilitating a comprehensive comparison with the experimental spectrum.

When m = 5, the outcomes and insights derived from the group analysis are presented in Figure 3.11. As mentioned above, when CE is less than 40 eV, the resulting spectrum is not informative. Therefore, we removed the energies that are less than 40 V.



a. EI result without filtering

b. Number of Peaks

Figure 3.12:

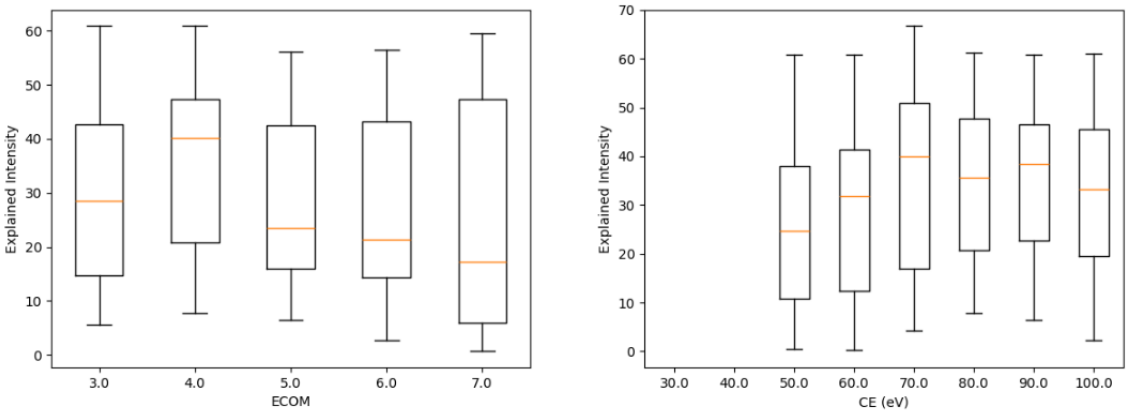
EI result of Ecom test without filtering

In Figure 3.11, the initial group exhibits a superior EI compared to the subsequent groups, indicating that higher intensity in the predicted MSMS spectra corresponds to higher EI. Moreover, no significant differences were observed in the energy range from 50 eV to 80 eV, suggesting a consistent trend in EI across this range.

3.4.3 ECOM

The comprehensive outcomes of this experiment, presented in Figure 3.12, offer valuable insights into the EI trends corresponding to the increasing number of peaks with ascending ECOM values.

When assessing the effectiveness of EI with ECOM as the energy type, we employed peak alignment to ensure that the number of peaks corresponded to the same peaks obtained with CE. This process ensured a standardized dataset for comparison. The aligned results,



a. ECOM result with k' peaks

b. CE result with k' peaks

Figure 3.13: EI result of ECOM vs.CE after peaks alignment k' peaks

as illustrated in Figure 3.13, were subsequently compared with those obtained using CE as the energy type. The comparison indicated no significant differences between the two approaches.

In examining the intensities analysis results, as depicted in Figure 3.14, the consistent trend reveals that the first group consistently exhibits superior EI compared to the other groups when the ECOM is equal to 3 and 4. Despite the absence of improvement in the results with using ECOM as the energy type, this reaffirms our earlier observation regarding the enhanced EI information in the first intensity group compared to the results of the other groups.

3.4.4 Multiple Proteomers

Based on the CE test results, it was observed that the range between 50 eV and 100 eV does not exhibit a significant variance in EI prediction accuracy. To further explore

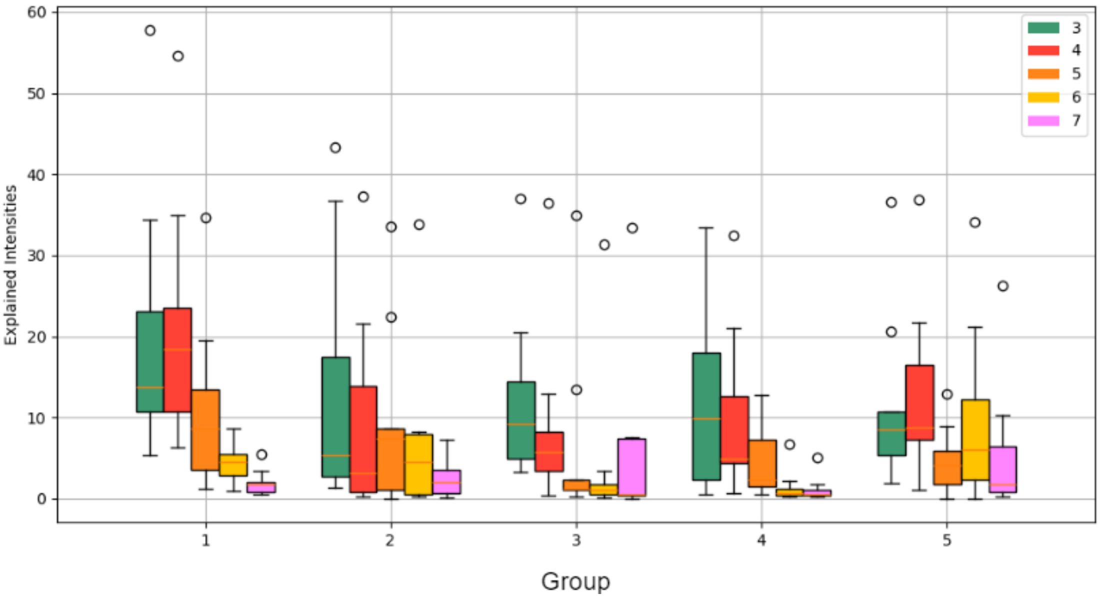


Figure 3.14: Intensities analysis of using ECOM

this finding, we opted for 70 eV as the testing energy while introducing the concept of employing multiple likely proteomers under constant conditions, with the exception being the protonation technique. Following the experimental phase, the mass spectra of the individual proteomers for a given molecule were amalgamated into a unified mass spectrum. The merging process entailed combining identical m/z values, aggregating their intensities, and subsequently dividing the total intensity by the number of potential proteomers. For peaks with sole existence, a straightforward addition of the peak intensity divided by the number of possible peaks was performed.

We then limit the number of peaks for each molecule to the same peak number k" as in the Collision Energy test. As the result shown in Figure 3.15, we observed that there is no significant improvement in EI with using multiple proteomers.

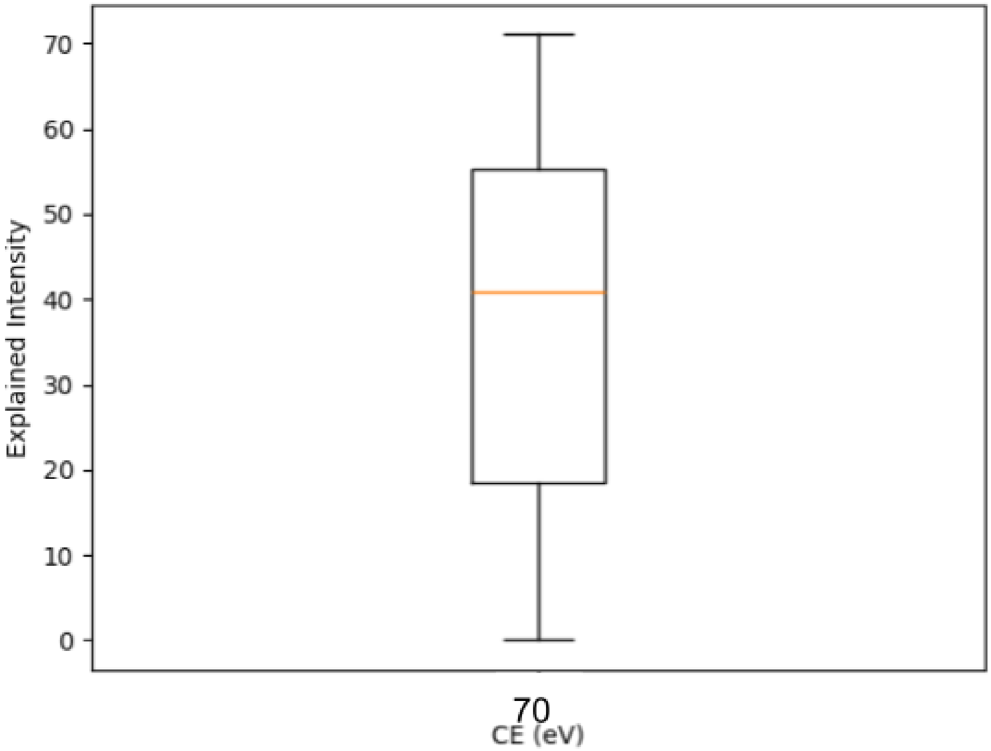


Figure 3.15: EI result of multiple proteomers after alignment of k” peaks.

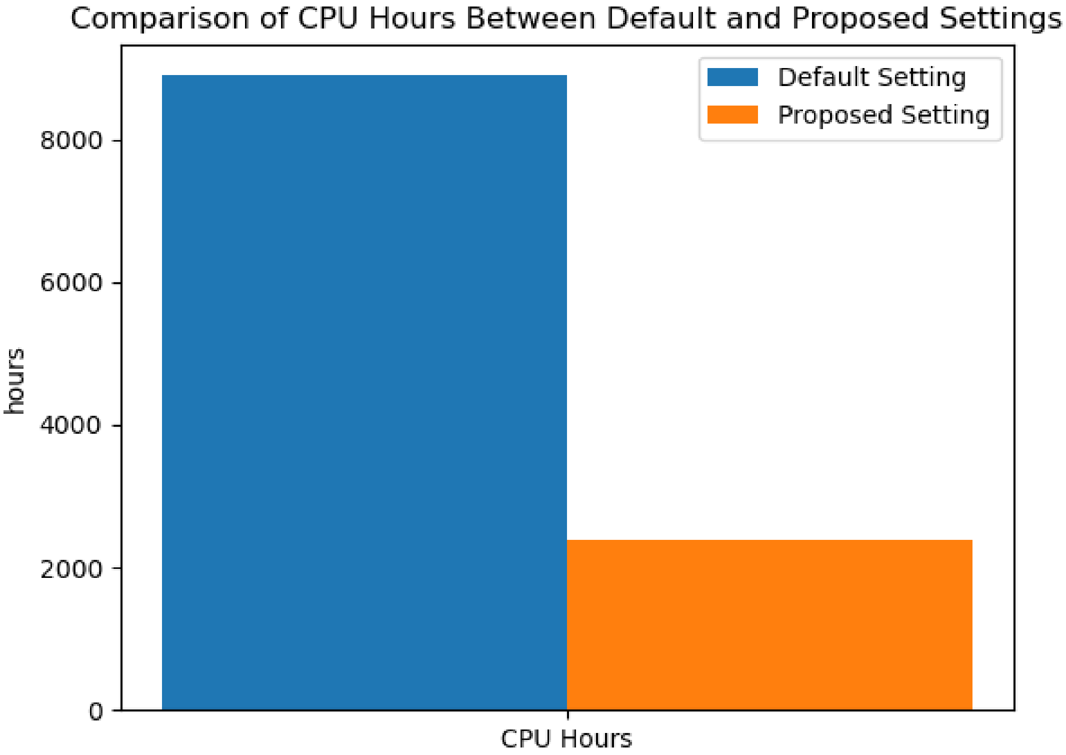


Figure 3.16: Overall CPU Hours improvement

3.5 Overall Improvement

Having finalized our parameter tuning, we proceeded to compare the overall performance of our tuned parameters with the default settings. When using set1, the CPU cost with the default setting was 8901 CPU hours, whereas the CPU cost with the proposed parameters was 2373 CPU hours, representing a 73% reduction in CPU cost, as illustrated in Figure 3.16.

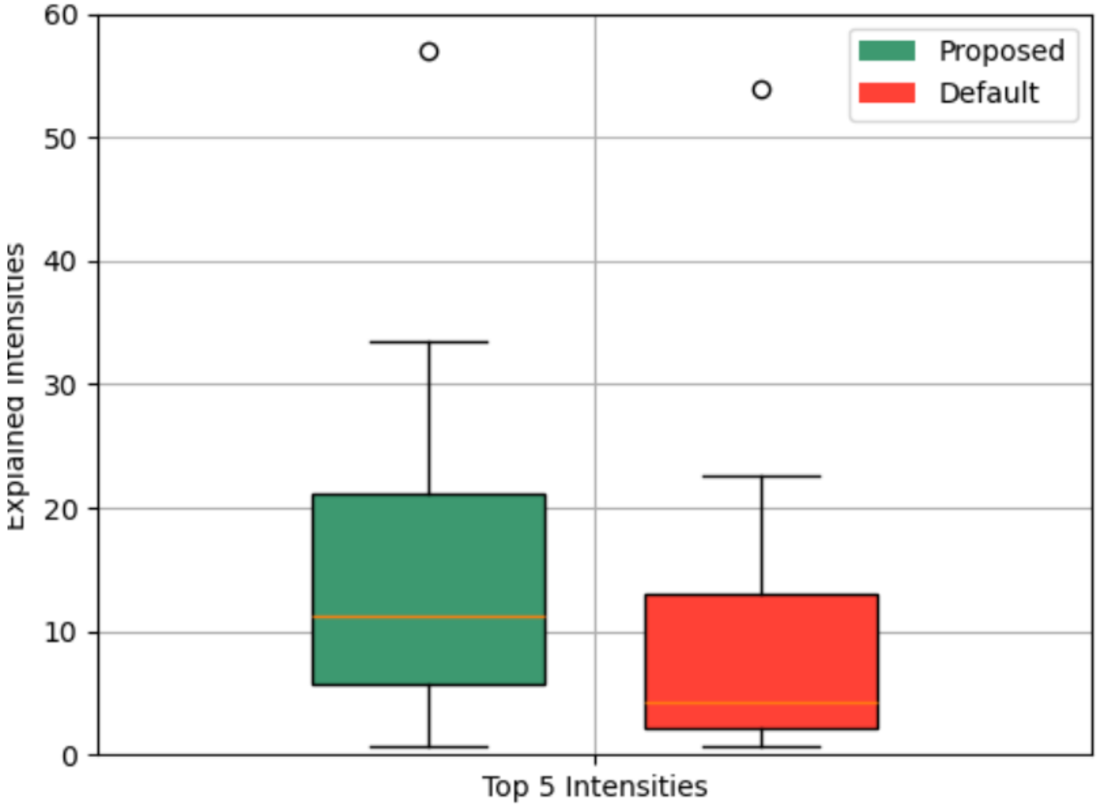


Figure 3.17: Overall CPU accuracy improvement

Additionally, as depicted in Figure 3.17, the simulation accuracy improvement is evident, with the median EI score of the top 5 intensities of the Simulated Spectrum increasing from 4.6 to 11.6, reflecting a 152% improvement. .

Chapter 4

Conclusions

The biggest contribution of this thesis is that we designed and implemented a workflow that could reduce the computation time of QCxMS. Utilizing the workflow designed for accelerating QCxMS, we have achieved a noteworthy reduction in mass spectrum analysis duration—from days to hours. Capitalizing on this computational efficiency, we conducted comprehensive experiments to assess the performance of QCxMS, identifying parameters that outperform the original defaults.

In our initial attempt to address the substantial computation time cost, we conducted an analysis of the reproducibility of QCxMS. Our findings indicate that selecting a trajectory count of 200 yields highly consistent results. Further escalation of the trajectory count would impose a considerable demand on computational resources without proportionate additional benefits. Consequently, we have chosen a trajectory count of 200 as it represents a balance between computational efficiency and prediction accuracy. For a molecule with 54 atoms, we found the default setting of QCxMS would use 864 CPU hours, which would use at least 3.5 days on a computer with 8 CPU cores. But by running the workflow with 200 trajectories, this molecule would only take 131 CPU hours and should take 16 - 24 hours to finish the entire calculation process. Furthermore, if a cluster is used for the simulation, the actual execution time can be reduced according to the number of cores used.

Upon thorough evaluation, we assert that while QCxMS may not excel in accurately predicting mass spectra due to a low cosine score in comparison to the actual experimental spectrum, it exhibits a commendable capability in predicting the highest intensity peaks. Notably, the accuracy of QCxMS's predictions surpasses that of CAS, underscoring that the results generated by QCxMS are not arbitrary.

Our observations also indicate that the intensities of the predicted spectrum by QCxMS can serve as reliable indicators of m/z predictions. Notably, the top five intensities demonstrate a higher EI, adding a valuable dimension to the assessment of the reliability of the predictions.

To enhance prediction accuracy, we propose the use of dinitrogen as the gas and recommend elevating the CE to the range of 50 eV to 100 eV. While we explored the utilization of ECOM as the energy type and tried using multiple possible proteomers to improve prediction accuracy, our findings indicated no significant enhancement through this approach.

Chapter 5

Future Directions

We have successfully built a QCxMS image that installs QCxMS from the source code, which fundamentally enables the modification of QCxMS. This opens a door to modifying the source code of QCxMS and solving the computation cost in a fundamental way.

Bibliography

[1] Arnald Alonso, Sara Marsal, and Antonio Julià. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology*, 3:23, 2015.

[2] Robert N Barnett and Uzi Landman. Born-oppenheimer molecular-dynamics simulations of finite systems: Structure and dynamics of (h 2 o) 2. *Physical review B*, 48(4):2081, 1993.

[3] Jürgen Cox. Prediction of peptide mass spectral libraries with machine learning. *Nature Biotechnology*, 41(1):33–43, 2023.

[4] Bruno Domon and Ruedi Aebersold. Mass spectrometry and protein analysis. *science*, 312(5771):212–217, 2006.

[5] GA Nagana Gowda and Danijel Djukovic. Overview of mass spectrometry-based metabolomics: opportunities and challenges. *Mass Spectrometry in Metabolomics: Methods and Protocols*, pages 3–12, 2014.

[6] Jian Guo, Huaxu Yu, Shipei Xing, and Tao Huan. Addressing big data challenges in mass spectrometry-based metabolomics. *Chemical Communications*, 58(72):9979–9990, 2022.

[7] George Kaklamanos, Eugenio Aprea, and Georgios Theodoridis. Mass spectrometry: principles and instrumentation. In *Chemical analysis of food*, pages 525–552. Elsevier, 2020.

[8] Jeroen Koopman and Stefan Grimme. From qceims to qcxms: A tool to routinely calculate cid mass spectra using molecular dynamics. *Journal of the American Society for Mass Spectrometry*, 32(7):1735–1751, 2021.

[9] Libretexts. 6.4: Mass analyzer orbitrap, November 2022.

[10] Raimund Mannhold, Hugo Kubinyi, and Hendrik Timmerman. *Molecular Modeling: Basic Principles and Applications*. John Wiley & Sons, 2008.

[11] Anca-Narcisa Neagu, Madhuri Jayathirtha, Emma Baxter, Mary Donnelly, Brindusa Alina Petre, and Costel C Darie. Applications of tandem mass spectrometry (ms/ms) in protein analysis for biomedical research. *Molecules*, 27(8):2411, 2022.

[12] Farhana R Pinu, David J Beale, Amy M Paten, Konstantinos Kouremenos, Sanjay Swarup, Horst J Schirra, and David Wishart. Systems biology and multi-omics integration: viewpoints from the metabolomics research community. *Metabolites*, 9(4):76, 2019.

[13] PPREMIER Biosoft. Mass spectrometry: Introduction, principle of mass spectrometry, components of mass spectrometer, applications. Retrieved June 24, 2021, from http://www.premierbiosoft.com/tech$_{n}otes/mass - spectrometry.html, n.d. Premierbiosoft.comwebsite.$

[14] Thermo Fisher Scientific. Overview of mass spectrometry for metabolomics. Retrieved from https://www.thermofisher.com/es/es/home/industrial/mass-spectrometry/mass-spectrometry-learning-center/mass-spectrometry-applications-area/metabolomics-mass-spectrometry/overview-mass-spectrometry-metabolomics.html. Accessed: Insert date accessed.

[15] Daniel GC Treen, Mingxun Wang, Shipei Xing, Katherine B Louie, Tao Huan, Pieter C Dorrestein, Trent R Northen, and Benjamin P Bowen. Simile enables alignment of tandem mass spectra with statistical significance. *Nature Communications*, 13(1):2510, 2022.

[16] Waters. What is ms and how does it work? Retrieved June 24, 2021, from https://www.waters.com/waters/en$_{C}A/What - is - MS - and - How - does - it - WorkWaters.comwebsite.$

[17] Jun Feng Xiao, Bin Zhou, and Habtom W Ressom. Metabolite identification and quantitation in lc-ms/ms-based metabolomics. *TrAC Trends in Analytical Chemistry*, 32:1–14, 2012.

[18] Naythan Yeo, Dillon Tay, and Shi Jun Ang. Benchmarking tandem mass spectra of small natural product molecules via ab initio molecular dynamics. 2023.