ISBN: 978-3-95450-249-3 ISSN: 2673-5350 doi: 10.18429/JACoW-IBIC2024-WEP69

# DESIGN AND IMPLEMENTATION OF MASS SPECTROMETER DATABASE

L. F. Liu, M. Li, R. S. Mao, Institute of Modern Physics, Lanzhou, China

## *Abstract*

Mass spectrometer, as a type of beam instrument, is capable of measuring and analyzing the mass and charge of different molecules and ions in a sample, thus identifying the type of particles. With the continuous development of mass spectrometry technology, the management, retrieval and analysis of mass spectrometry data have become crucial in many fields such as biomedicine, environmental sciences, and crime investigation. Mass spectrometer database software is an important part of mass spectrometer, which can realize the functions of storing, managing, sharing and analyzing mass spectrometer data. Therefore, the establishment and improvement of specialized mass spectrometry databases and library retrieval techniques can facilitate the rapid identification and confirmation of compounds, providing a more efficient and accurate solution for substance detection. In this paper, a comprehensive mass spectrometry database management system is designed and implemented to simplify the user operation process from the collection, storage and management of mass spectrometry data to the querying, matching and analyzing of the data, providing a fast and accurate solution to meet the needs of scientific research on mass spectrometry data. The software uses Python for the implementation of core algorithms, builds a database based on MySQL and collects mass spectrometry data to fill in the data-base, and finally uses PyQt to design and implement a friendly and beautiful graphical user interface. While this software improves the accuracy of matches, it also optimizes processing speed. With this software, un-known compounds in the samples can be identified and their possible structures and properties can be recognized, which provides a strong support for their application fields.

## INTRODUCTION

Since its introduction in the early 1970s, the mass spectrometry library search technique has occupied an important position in the field of mass spectrometry analysis by searching and analyzing mass spectrometry data by computer. The technique identifies chemical components in complex samples by converting chemicals into ions and separating, detecting, and quantitatively analyzing them. In recent years, the mass spectrometry library search technique has become a hot topic for researchers because of its high resolution, good separation and fast resolution speed. It is widely used as a fast and effective identification method in environmental monitoring, proteomics, life sciences, food testing, material analysis, petrochemical, and anti-terrorism and riot control. When analyzing unknown compounds, mass spectrometry search software provides qualitative references by comparing the mass spectra of samples with those of standards in the database and measuring the similarity by the match of mass spectra. Therefore, it is particularly important to establish an efficient and fully functional mass spectrometry data management system. Such a system not only improves the efficiency of data processing, but also enhances the ability of data integration and sharing, supports complex data analysis, and accelerates the process of scientific research and discovery.

Authoritative standard spectral libraries include the NIST Mass Spectral Library published b12/09/2024y the National Institute of Standards and Technology (NIST) [1], the NIST/EPA/NIH Library published by NIST together with the U.S. Environmental Protection Agency (EPA) and the U.S. National Institutes of Health (NIH), and the Wiley Library. In addition, there are a number of high-quality databases in China, such as the ICMSIS mass spectrometry database system of the Institute of Chemistry Chinese Academy of Sciences. The mass spectrometry database provided by the chemical specialty database system of the Shanghai Institute of Organic Chemistry [2]. There are also Sadtler databases applicable to specific fields such as criminal investigation, chemical product quality control, petroleum, plastics industry, mineral analysis, and other fields [3].

Currently, common mass spectrometry data processing systems are divided into two categories: One is the system provided by mass spectrometry instrument manufacturers, such as Agilent's Mass Hunter Chemical Workstation and Thermo Fisher Scientific's Xcalibur, which have user-friendly interfaces, high-quality technical support, and seamless connectivity to their branded mass spectrometers, but are costly and poorly customizable. The other category includes instrument-independent systems, such as NIST MS Search and AMDIS, which serve as standardized mass spectrometry databases and provide a wide range of data and tools for compound analysis and identification. However, these systems are not completely free, and many of the advanced features are hidden behind expensive licenses that increase the cost of research [4]. Given the limitations of existing systems, this study aims to develop an open source, customizable, and user-friendly mass spectrometry data management system.

## DATABASE DESIGN

Database is a data storage warehouse, which refers to a collection of organized and shareable data stored in a computer over a long period of time.

Database management system is a computer system software specialized for database management, which can provide functions such as data definition, creation, query and operation for databases, and realize the control of data integrity and security. The advantages of managing mass spectrometry library data by database management system are that the library is easy to design, the data is easy to manage, and the database management system itself provides a powerful data search function that is easy to use. However, the disadvantages are that it occupies too much space and the cost is high [5]. This system uses the MySQL database management system, which is an open source relational database management system widely used for data storage and management. For mass spectrometry data, using MySQL can effectively manage a large amount of structured data, support efficient data query and multi-user access, ensure data integrity, and have good scalability and maintainability.

### *Database Entity-Relationship (E-R) Model*

The Entity-Relationship (E-R) diagram provides a clear visual overview for database design, aiding in understanding the data flow and relationships between different tables. By displaying the E-R diagram, one can clearly see how entities are connected through foreign keys and other relational methods. This structured information design helps ensure that all key data can be effectively linked and queried when actually establishing the database [6]. Figure 1 shows the E-R diagram of this system.

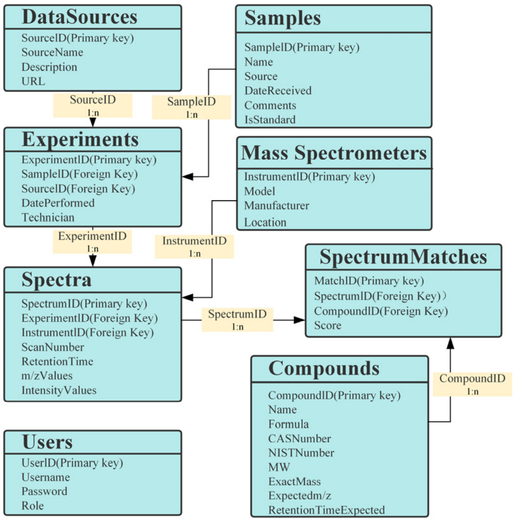


Figure 1: E-R diagram.

In data relationships, there is a one-to-many relationship between DataSources and Experiments, indicating that one data source may correspond to multiple experiments. Similarly, the relationship between Samples and Experiments is also one-to-many, suggesting that one sample can be used for multiple experiments. There is a one-to-many relationship between Experiments and Spectra, showing that one experiment can produce multiple spectra. Likewise, the one-to-many relationship between Mass Spectrometers and Spectra indicates that one mass spectrometer can be used to generate multiple spectra. Additionally, the relationship between Spectra and Spectrum is also one-to-many , reveals that one spectrum can correspond to multiple matching results, and the one-to-many relationship between Compounds and Spectrum Matches shows that one compound can be matched in multiple spectra. From sample collection and experimental operations to detailed analysis in chromatography and mass spectrometry, each level of data depends on the previous level, ensuring the logical consistency and data integrity of the entire database. This carefully designed database architecture facilitates the effective storage and management of mass spectrometry data.

### *Data Sources*

When building a mass spectrometry database, the source of the data is critical because it directly affects the quality, coverage, and availability of the data. The mass spectrometry data in this database come from two sources: first, the collection and organization of open source mass spectrometry data, and second, the integration of actual measurement data by performing tests on standards [7].

Open-source mass spectrometry data are obtained from recognized databases, such as the NIST. These data sources provide a wide range of compounds and their corresponding mass spectra, typically including: compound name and structure, mass spectra (including key information such as m/z, peak intensity, etc.). Data acquired from sources such as NIST are first validated and screened to ensure the accuracy and completeness of the information. The cleaned and formatted data is then imported into a database to update existing records or add new ones.

Another important source for the database is the mass spectrometry data obtained from tests on standards using the BRUKER maXs Plus mass spectrometer. Acquired through a series of sample preparation, experimental setup and data acquisition, the processed data are finally formatted and imported into the database. BRUKER maXs Plus mass spectrometer is a high-end mass spectrometry analyzer designed to achieve extremely high sensitivity and resolution. This instrument is widely used in chemistry, biomedicine and materials science and is particularly suited for in-depth analysis of complex samples.

## SYSTEM IMPLEMENTATION

The system software platform of the mass spectrometer is a software family and is divided into two parts according to the interdependence of the analytical processes: instrument-related software and mass spectrometry data analysis and processing software. The mass spectrometry data analysis and processing software is the downstream part of the software family. The goal of this system is to complete the design and implementation of the entire database processing software.

### *Technical Implementation*

Development efficiency, cost, and ease of use for end users were taken into consideration. The system uses MySQL database management system for database construction, Python programming language for development, and PyQt for graphical user interface design to create a full-featured, user-friendly, and high-performance mass spectrometry database management system [8].

MySQL is a widely used relational database management system that is open source and supports cross-platform use. In our mass spectrometry data management system, MySQL is responsible for storing, querying, and managing a large amount of sample data and related metadata. Its good compatibility with Python further improves development efficiency.

Python is a high-level programming language known for its clear syntax and rich library resources that is suitable for a variety of programming tasks, especially for rapid development. The NumPy and Pandas libraries provided by the Python community greatly optimize the efficiency of data processing and scientific computing. In addition, we use the matplotlib graphing library to visualize mass spectrometry data, facilitating the creation of detailed spectral graphs.

PyQt, a toolkit for creating graphical user interface (GUI) applications, provides excellent cross-platform capabilities and rich control support, allowing us to create professional and modern user interfaces quickly and easily.

In summary, by integrating MySQL, Python and PyQt, we are not only able to build an efficient and reliable mass spectrometry database management system, but also ensure the system's ease of use and extensibility to meet the needs of scientific research and practical applications.

## System Architecture

Observing the central business flow chart of the software, it can be seen that the software is centered on the functions of data storage, management, query and display, and the user sends various commands to the software through the view, and the software reacts according to the commands, and finally reflects the results to one or more views to deliver the information to the user. Figure 2 shows The business process diagram.

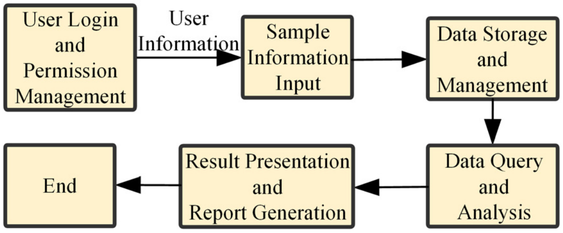

User Login and Permission Management
User Information
Sample Information Input
Data Storage and Management
Data Query and Analysis
Result Presentation and Report Generation
End

Figure 2: Business process diagram.

Through analysis and comparison, MVC (Model-View-Controller) architecture is chosen as the overall architecture for designing this program, which helps to separate data processing, user interface and business logic, and helps to structure the code to improve the maintainability and extensibility of the code [9]. The MVC architecture divides the system into three parts: Model, View and Controller.

Model is the core of the system, which is used to abstract system logic, encapsulate system state, and manage system data, including database connection, query, and processing.

The view is the interface that the user sees and interacts with and is a representation of the data for the user. The main task of the controller is to accept user input and invoke the appropriate model and view to satisfy the user's needs and control the application process.

Using the MVC architecture, it is possible to create and manipulate multiple views for a single model at runtime. Through the change propagation mechanism, as model data changes, all related views can be updated in a timely manner to ensure that all related views are synchronized with the behavior of the controller. This approach is ideal for scenarios where a model corresponds to multiple views in a program. Figure 3 shows a simple MVC architecture diagram.

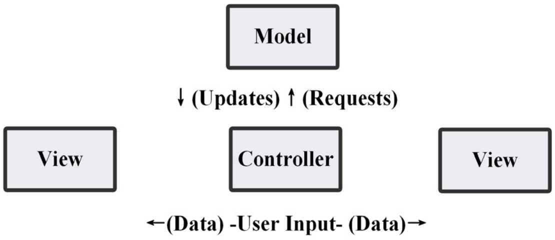


Figure 3: MVC architecture diagram.

## Implementation Details

Mass spectral fingerprint matching is the identification of compounds by comparing the mass spectral fingerprint of an unknown sample with the spectra in a database [6]. Specifically, mass spectral fingerprint matching is a technique for identifying unknown compounds by comparing the characteristic peaks of a sample's mass spectrum with records in a database of known mass spectra. This method relies heavily on features such as accurate mass measurements, peak intensities, and peak patterns. Figure 4 shows the key steps and techniques for mass spectrometry fingerprint matching:

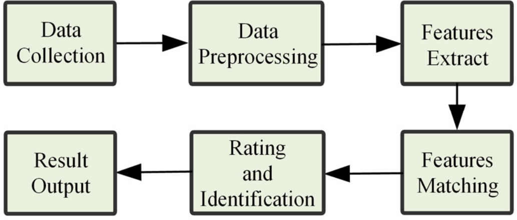


Figure 4: Fingerprint Matching Algorithm Flowchart.

Important steps in the data preprocessing phase of mass spectrometry data analysis include denoising, mass correction, and normalization. Denoising is typically performed using filters or wavelet transforms to remove background noise and ensure data clarity. Mass correction ensures the accuracy of the analysis by removing instrumental errors through calibration techniques. In addition, normalization is essential to make the data more consistent and comparable by standardizing the intensity of the peaks to eliminate the effects of sample volume differences.

Peak detection and feature extraction are the core steps in mass spectrometry. At this stage, the SciPy library for Python can be used to detect peaks in one-dimensional data, while algorithms such as CentWave and Matched Filter are used to identify characteristic peaks in the data. Each detected peak is further analyzed to extract its key features, such as mass/charge ratio (m/z), peak area, peak height, and retention time, which are critical for subsequent data interpretation.

The database matching phase involves the selection of a mass spectrum database and a matching algorithm. Computational methods such as cosine similarity are used to compare the sample mass spectra with the reference mass spectra in the database to determine the best matching term. This process is a critical part of identifying the structure of unknown compounds and relies on high quality databases and accurate algorithms.

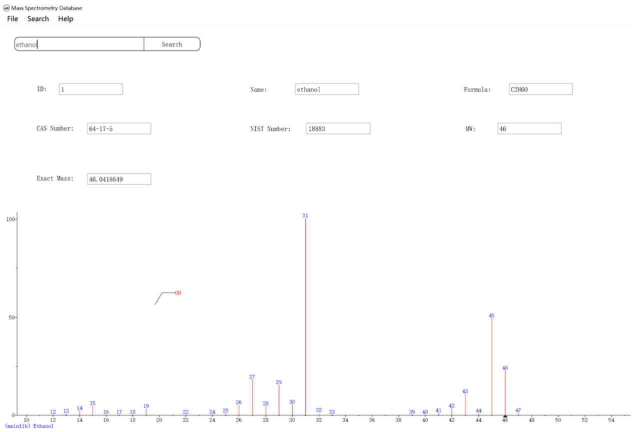
Finally, verification and validation are important steps in ensuring the accuracy of mass spectrometry results. After finding the matches with the highest similarity to the mass spectra in the database, these results must be further validated. This may involve performing secondary experiments or using different analytical methods for confirmation. In addition, statistical analyses must be used to assess the reliability and accuracy of the matches to ensure the data quality and credibility of the final report.

In mass spectrometry analysis, the use of high-resolution mass spectrometers allows for more accurate measurement of mass-to-charge ratios (m/z values), which improves the accuracy of data matching to databases. In addition, the quality of the database and its coverage have a direct impact on the accuracy of the matching results; databases that are frequently updated and contain multiple substances can greatly improve the reliability of the matching. Meanwhile, the selection of an appropriate similarity calculation method is also crucial to the accuracy of compound identification, as it can effectively assess the degree of similarity between the sample mass spectra and those in the database. Combining all these factors, the effectiveness and credibility of mass spectrometry analysis have been greatly enhanced.

## RESULTS AND DISCUSSION

In this paper, a mass spectrometry data management system has been designed and implemented, which provides a variety of data input and query options, such as by compound name, molecular formula, CAS number, and so on [10]. The interface also includes graphical display functions of the data, such as the dynamic display of the mass spectrometry diagram, which not only helps users to intuitively understand the data, but also facilitates in-depth analysis. The system also provides rich data management functions, including data import, export, editing, and deletion operations, as shown in Fig. 5; this ensures high availability and security of the database.

In mass spectrometry analysis, the incorporation of Tandem Mass Spectrometry (MS/MS) data into mass spectrometry databases is critical because these data provide detailed information about the molecular structure of compounds, greatly improving the accuracy and reliability of compound identification [11]. In addition, mass spectrometry data can be used in a wider range of applications, such as supporting complex structure resolution and validation [12]. Therefore, it is planned to add secondary mass spectrometry-related content to this system in the future to provide more complex data analysis and interpretation functions.



ethanol

ID: 1

Name: ethanol

Formula: C2H6O

CAS Number: 64-17-5

NIST Number: 18883

MW: 46

Exact Mass: 46.0418649

Figure 5: Search results display page.

## REFERENCES

[1] L. Zhang, L. Yao, K. Zhang, G. Wei and Z. Yang, “Metabolomics databases for natural products research and development”, *Chem. Anal. and Meterage.*, vol. 28, no. 5, pp. 128-134, Sep. 2019. doi:10.3969/j.issn.1008-6145.2019.05.030

[2] Q. Hu, “Research and Application of Data Format of Analytical Instrument and Searching System of Mass Spectrum” , Ph.D. thesis, Sch. of Instrum. Sci. and Elect. Eng.， Jilin University, Jilin, China, 2006.

[3] Y. Zhou, “The research and application of the algorithm to Mass Spectral Data”, M.D. thesis, Fac. of Electr. Eng. and Comput. Sci., Ningbo University, Ningbo, China, 2017.

[4] C. Shou, G. Naren, S. Gao, J. Liu and F. Zhao, “Establishment and application of mass spectral database for natural dyes”, *J. Text. Res.*, vol. 44, no. 11, pp. 120-131, Nov. 2023. doi:10.13475/j.fzxb. 20220907501

[5] J. Li, “The Establishment of Mass Spectrometry Database include Interpretation based on Pesticide Residue Analysis”, M.D. thesis, Coll. of Chem., Chem. Eng. and Environ., Qingdao University of Science and Technology, Qingdao, China, 2012. doi:10.7666/d.y2179089

[6] W. Yu, “Research and Implementation of the Establishment Method of Serum standard reference material Database in Clinical metabolome”, M.D. thesis, Coll. of Software., South China University of Technology, Guangzhou, China, 2020.

[7] S. Kim, P. Thiessen, E. Bolton *et al.*, “PubChem Substance and Compound databases”, *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1202-D1213, Sept. 2015. doi:10.1093/nar/gkv951

[8] Y. Xu, H. Yang, T. Wu *et al.*, “BioM2MetDisease: a manually curated database for associations between microRNAs, metabolites, small molecules and metabolic diseases”, *Database*, vol. 2017, no. 2017, pp. bax037, May 2017. doi:10.1093/database/bax037

[9] B. Zhang, “Research and Development of MS Software Platform”, M.D. thesis, Coll. of Comp. Sci. and Tech., Jilin University, Jilin, China, 2006. doi:CNKI:CDMD:2.2006.092732

[10] Y. Luo and Z. Liu, “Pesticide Residue Fragmentation Spectrum library based on NIST Database Establish and Apply”, *China Food Saf. Mag.*, no. 21, pp. 161-162+164, Jul. 2019. doi:10.16043/j.cnki.cfs. 2019.21.124

[11] H. Hisayuki, A. Masanori, K. Shigehiko, et al., “MassBank: a public repository for sharing mass spectral data for life sciences”, *J. of mass Spectrom.*, vol. 45, no. 7, pp. 703-714, Jul. 2010. doi:10.1002/jms.1777

[12] S. Colin A, M. Grace O’, W. Elizabeth J *et al.*, “METLIN A Metabolite Mass Spectral Database”, Ther. Drug Monit., vol. 27, no. 6, pp. 747-751, Dec. 2005. doi:10.1097/01.ftd.0000179845.53213.39