Data-Mining Contribution for the Information Retrieval in Cheminformatics

Gopala Krishna Murthy H.R.
GFGC, Nanjangud, Mysuru, India

Crossref
Open Access

Manuscript ID:
BIJ-SPL5-MAR26-MD-002

Subject:Chemistry
Received : 26.12.2025
Accepted : 28.02.2026
Published : 14.03.2026

DOI: 10.64938/bijsi.v10si5.26.Mar002

Abstract

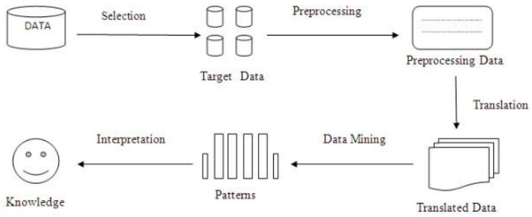
The rapid growth of chemical and biological data generated from high-throughput experiments, databases, and computational studies has created a critical need for efficient information retrieval systems in cheminformatics. Traditional search and retrieval methods are often insufficient to handle the volume, complexity, and heterogeneity of modern chemical data. Data-mining techniques play a significant role in improving information retrieval by enabling the extraction of meaningful patterns, relationships, and knowledge from large chemical datasets. This paper explores the contribution of data-mining approaches—such as classification, clustering, association rule mining, searching, and predictive modeling—in enhancing information retrieval within cheminformatics applications. Emphasis is placed on their use in chemical structure analysis, compound similarity assessment, property prediction, and bioactivity profiling. The integration of data-mining methods with cheminformatics tools facilitates faster, more accurate retrieval of relevant chemical information, supporting tasks such as drug discovery, virtual screening, and molecular database management. The study highlights current challenges, including data quality, scalability, and interpretability, and discusses future perspectives for advancing intelligent information retrieval systems in cheminformatics through data-driven methodologies.

Keywords: Compound, Database, Data-mining, software, cheminformatics.

1. Introduction

A database is a collection of interrelated information stored in the form of names, formulae, or patterns, and users can create, manipulate, and retrieve data using suitable query languages [1]. Data mining is considered a data retrieval and analysis process, and an efficient database framework is particularly important in cheminformatics due to the existence of a large number of chemical elements and compounds [2]. Data mining is a major step in the process of Knowledge Discovery in Databases (KDD), which originated in computer science and now supports a wide range of interdisciplinary applications [3]. Data mining is defined as the process that leads to the discovery of new patterns and relationships in large datasets using methods from database systems, statistics, artificial intelligence, and machine learning [4]. The primary goal of data mining is to extract meaningful knowledge from existing datasets and transform it into a human-understandable form for further use, involving database management, information processing, data collection, extraction, and data warehouse analysis. Although the term “data mining” was introduced largely for marketing purposes, it is widely used today as a practical machine learning and analytical tool; more general terms such as large-scale data analysis, artificial intelligence, and machine learning are often considered more appropriate [4]. MOLFEA (Molecular Feature Miner) is a cheminformatics tool designed to mine linear molecular fragments of interest from the two-dimensional structures of chemical compounds [5].

Overview of KDD



The Beilstein database is the largest database in the field of organic compounds, in which the database covers the scientific literature from 1771 to the present and contains millions of chemical reactions and compounds. Data mining in the Reactivity nature of chemical compounds is the classic data mining process of extracting information from distributed data sources, with a minimum of interaction between data sites. Data mining is an exploration of large amounts of data in search for consistent patterns, correlations and other systematic relationships. It can be a helpful tool to evaluate "hidden" information in a set of molecules. Finding the adequate descriptors for the representation of chemical structures is one of the most important problems in chemical data mining [6].

2. Materials and Methods

The explosive growth of data in the field of chemical science has created an urgent need for effective analysis and mining of useful knowledge from large chemical datasets [7]. This has motivated the development of system-level solutions for next-generation data analysis centers, including adaptive and dynamic data management strategies, storage services, disk-based sampling, and efficient data placement mechanisms. Such developing services aim to support the re-use of chemical data and previously mined information related to chemical compounds. The major objective of this work is to evaluate data mining tools for studying the reactive nature of organic compounds and to develop a framework that enables timely and accurate decision-making in chemical research [8]. Data mining techniques are extensively used in the study and analysis of cheminformatics. Cheminformatics is an interdisciplinary scientific domain that addresses complex chemical problems using computational techniques and methods, providing innovative solutions for scientific, educational, and industrial research activities of chemists and chemical engineers [9]. Various data mining and computational tools are employed, including decision trees for classification based on rule-based structures, WEKA software for data visualization and graphical user interface-based analysis, CRUISE for unbiased decision tree classification, and SPARC for automated reasoning in chemistry, enabling estimation of physicochemical properties and chemical reactivity parameters of organic compounds.

The goal of the present work is to design a software architecture for a data mining framework specifically aimed at analyzing the reactive nature of organic compounds and to develop additional tools as required for chemical data analysis. In this context, the Internet-based Data Mining Service Chemistry (DMSC) project has been proposed to provide centralized access to diverse data mining methods, including statistical processing, simulation, prediction of chemical properties, sequencing, encoding, classification, and information retrieval.

Several software tools are available for extracting chemical information from databases. The IUPAC International Chemical Identifier (InChI) is a textual identifier developed by IUPAC and NIST to provide a standard, human-readable encoding of molecular information and to facilitate chemical information retrieval on the web. It has been widely adopted by major chemical databases such as PubChem and ChemSpider and is distributed under an open-source LGPL-based license through the InChI Trust [10]. In addition, molecule editors are computer programs used to create and modify two-dimensional and three-dimensional representations of chemical structures, supporting both database querying and molecular modeling applications.

3. Interpretation and Discussion

Database molecular editors such as Leatherface, RECAP, and Molecular Slicer allow large numbers of molecules to be modified automatically according to user-defined rules, including operations such as deprotonation of carboxylic acids and cleavage of exocyclic bonds. Molecular editors generally support reading and writing one or more chemical file formats or line notations; common examples include Molfile for structure representation and SMILES for linear notation of molecular structures [11].

Table 1 The Files Generated by Molecule Editors [Molecular Graphics Tools]

Software Tool	Developer	License	Application
Accelrys Draw	Accelyrs	Proprietary	Freeware version: name to structure and structure to name. InChI naming and canonical SMILES.
ACD/ChemSketch	ACD/Labs	GPL*	3D molecule Editor and Visualizer.
Avogadro	OpenMolecules.net	GPL	3D molecule Editor and Visualizer.
Bkchem	Beda Kosata	GPL	2D molecule editor
ChemDoodle	iChemLabs	Proprietary	Open source editor
ChemDraw	CambridgeSoft	Proprietary	Molecular editor.
ChemTool	Martin Kroeker and Team	GPL	2D editor for chemical formulas
ChemWindow	Bio-Rad	Proprietary	Freeware for academic research and teaching
ICM-Chemist	MolSoft	Proprietary	GUI desktop chemistry editor.
JchemPaint	CDK Project	LGPL**	2D structural formula editor
XdrawChem	Randstand Technologies	GPL	Based of OpenBabel
Zem	Example	GPL	Based of OpenBabel

***** General Public License

**Library General Public License [licensed cheminformatics library]

Applets are small application programs designed to perform specific tasks and execute within the environment of a host program or platform using a programming language, rather than operating as standalone applications [12].

Table 2 Applets with Respective Applications

Applet	Developer	License	Application
Accelrys Draw	Accelyrs	Proprietary	Commercial and Freeware versions for non-profit use.
JchemPaint	CDK Project	LGPL	Editor and Viewer applet
MarvinSketch	ChemAxon	Proprietary	Freeware chemical editor
MarvinSpace	ChemAxon	Proprietary	Freeware and commercial versions:3D-macromolecular visualization and ligand editing
ChemWriter	metamolecular	Proprietary	Touch interface supported on ipad
Chemis3D	Didier Collomb	Proprietary	Mol3D

JME-Molecular Editor	Norbert Haider	Freeware	Draw, edit & display molecules.
Marvin molecular editor and viewer	ChemAxon	Proprietary	molecular editor and viewer (properties)
PubChem	NCBI	Public domain	Online Molecular editor supports SMILES, SMARTS and InChI.
OLN JSDraw	Scilligence	Proprietary	Chemical structure editor & viewer. Run on desktops, Laptops, iPad, iphone and tablets.

The SYBYL Line Notation (SLN) is a specification used to unambiguously describe the structure of chemical molecules using concise ASCII strings. SLN differs from SMILES in several important aspects: SLN can represent molecules, molecular queries, and chemical reactions within a single line notation, whereas SMILES supports such functionality mainly through language extensions. SLN provides explicit stereochemical support for distinguishing enantiomers from pure molecules, and while aromaticity in SMILES is treated as a property of both atoms and bonds, in SLN aromaticity is defined solely as a property of bonds [13].

In SLN, elemental atoms are specified using standard chemical abbreviations, and atom attributes follow the atom symbol enclosed in square brackets, such as ^{14}C for the carbon-14 isotope. Hydrogen atoms are typically expressed in shorthand form, for example, CH_4 for methane. In addition to elemental atoms, SLN supports wildcard atoms such as Any (matching any atom) and Hev (matching any heavy atom). It also includes an extensive Markush syntax for defining combinatorial libraries and R-group queries, along with specialized query atom types such as R for side chains and X for side chains and rings.

SMILES (Simplified Molecular Input Line Entry Specification) and SMARTS (SMILES Arbitrary Target Specification) are widely used notations for molecular representation and substructure searching, respectively. For more advanced and problem-specific search requirements, the Molecular Query Language (MQL) has been developed to allow the specification of spatial and physicochemical properties of atoms and bonds. Unlike SMARTS, MQL can be extended to support non-atom-based or “reduced feature” graphs. These query languages are commonly defined using Extended Backus–Naur Form (EBNF) grammars and implemented using parser generators such as JavaCC, with SMARTS and InChI being notable examples [14].

4. Conclusion

The explosive growth in data collection with respect to business and scientific fields has forced us to mine useful knowledge from it. Data mining refers to the process of extracting useful and novel patterns (models) from large datasets. Due to the large size of data and amount of computation involved in data mining, high-performance computing is an essential part for successful large-scale data mining application. Data Mining is one such possible technique that can increase the performance of the Data Analysis, which is very essential for future generation.

The information extracted will be useful to,

a) Drug synthesis: In the selection of suitable biologically important organic compound.

b) Chemist: In Research and Development (R & D) department.

c) Chemical waste Management: Selecting suitable chemical in the neutralization of chemical waste released from the Chemical Industry.

d) e-waste management: Few electronic wastes managed through R & D activity.

References

Date, C. J. *An Introduction to Database Systems*, 8th ed.; Pearson Education: New Delhi, 2004.

Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer: Dordrecht, 2007.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI Magazine* 1996, *17*(3), 37–54.

Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; Morgan Kaufmann: Elsevier, 2012.

Mayer, M.; Schneider, G. MOLFEA: A molecular feature miner for structure–activity relationship analysis. *Journal of Chemical Information and Modeling* 2006, *46*, 1817–1824.

Gasteiger, J.; Engel, T. *Chemoinformatics: A Textbook*; Wiley-VCH: Weinheim, 2003.

Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; Morgan Kaufmann: Elsevier, 2012.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI Magazine* 1996, *17*(3), 37–54.

Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer: Dordrecht, 2007.

Heller, S. R.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI—The worldwide chemical structure identifier standard. *Journal of Cheminformatics* 2013, *5*, 7.

Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer: Dordrecht, 2007.

Schildt, H. *Java: The Complete Reference*, 11th ed.; McGraw-Hill Education: New York, 2021.

Warr, W. A. Representation of chemical structures: SMILES, SLN, SMARTS, and InChI. *Journal of Chemical Information and Modeling* 2020, *60*, 3–15.

Landrum, G. RDKit: Open-source cheminformatics. *Journal of Cheminformatics* 2021, *13*, 36.