Journal of Cheminformatics

DATABASE

Open Access

Clc-db: an open-source online database of chiral ligands and catalysts

Gufeng Yu[1,2†], Kaiwen Yu[1†], Xi Wang[1,3†], Chenxi Zhang[1], Yicong Luo[1], Xiaohong Huo[1*] and Yang Yang[2*]

Abstract

The design and optimization of chiral ligands and catalysts are fundamental to advancing asymmetric catalysis, a critical area in organic chemistry with wide-ranging impacts across scientific disciplines. Traditional experimental approaches, while essential, are often hindered by their slow pace and complexity. Recent advancements have demonstrated that computational methods, particularly machine learning, offer transformative potential by significantly accelerating these processes through enhanced prediction and modeling capabilities. However, limitations such as data scarcity and model inaccuracies continue to challenge their broader application. To address these issues, we present the Chiral Ligand and Catalyst Database (CLC-DB), the first open-source, comprehensive database specifically designed for chiral ligands and catalysts. CLC-DB contains 1,861 molecules spanning 32 distinctive chiral ligand and catalyst categories, with each entry annotated with 34 types of curated information, validated by chemical experts and linked to authoritative chemical databases. The database features a user-friendly interface that supports efficient single and batch searches, as well as an integrated, high-performance online molecular clustering tool to facilitate computational analyses. CLC-DB is freely accessible at https://compbio.sjtu.edu.cn/services/clc-db, where all data are available for download.

Scientific Contribution

This work introduces CLC-DB, an innovative open-source database dedicated to chiral ligands and catalysts, encompassing 1,861 molecules across 32 distinct types and offering 34 types of annotated data per molecule. This resource significantly enhances data accessibility and quality for asymmetric catalysis. With its integrated molecular clustering tools and user-friendly platform, CLC-DB serves as a valuable resource for advancing the design and optimization of novel chiral ligands and catalysts.

Keywords Chiral ligand, Chiral catalyst, Open-source database, Chiral catalyst design

†Gufeng Yu, Kaiwen Yu and Xi Wang have equally contributed to this work.

*Correspondence:
Xiaohong Huo
huoxiaohong@sjtu.edu.cn
Yang Yang
yangyang@cs.sjtu.edu.cn
[1] Shanghai Key Laboratory for Molecular Engineering of Chiral Drugs, Frontiers Science Center for Transformative Molecules, School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2] Department of Computer Science and Engineering, and Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[3] Department of Computer Science, New York University, New York 10012, United States of America

Introduction

The development of new chiral catalysts is essential to advancing organic synthesis, medicinal chemistry, and the creation of novel materials and agrochemical compounds. It is the evolution of innovative, accurate, and practical methodologies for generating structurally diverse chiral compounds, particularly through asymmetric catalysis using chiral catalysts, that underpins advancements and fosters innovation in this field. Over recent decades, asymmetric catalysis has achieved significant progress. Thousands of chiral ligands [1, 2], chiral metal catalysts [3–5], and small-molecule organocatalysts [6, 7] have been designed and synthesized [8–11]. These developments have led to their successful implementation in the asymmetric synthesis of natural products and functional materials [12, 13], as well as the industrial preparation of chiral drugs and pesticides [14, 15].

The study of chiral ligands and catalysts is central to the research in asymmetric catalytic synthesis [16–18]. Nearly every major advance in this field correlates with the introduction of a novel chiral ligand or catalyst. However, current strategies for chiral ligand and catalyst development still have some limitations [19]. Although numerous chiral ligands and chiral catalysts have been reported in past decades, only a small subset based on limited core structures have actually proven highly effective across varied, mechanistically distinct reactions. Many important catalysts are discovered by chance or through trial-and-error processes extending over several years and discovering truly novel catalysts remains a formidable challenge [20]. The ongoing revolution in data science is anticipated to significantly expedite catalyst development. Data-driven methods provide several advantages, such as handling complex and incomplete datasets, automating feature extraction, and identifying new patterns without relying on explicit prior theories [21–25]. Machine learning (ML) is poised to play a central role in this paradigm shift. The use of ML techniques in catalysis [26], as well as in the wider realms of chemistry [27] and materials science [28, 29], is drawing notable interest. Although ML offers powerful tools for the design of chiral ligands and catalysts, the guiding principles derived from ML are currently inadequate due to the necessity for robust and expansive datasets [30]. Presently, the databases pertaining to the design of chiral ligands and catalysts are incomplete and inaccurate [31]. While comprehensive databases like PubChem [32] and ChEMBL [33] compile data relevant to these molecules, the specific details on chiral ligands and catalysts remain insufficiently categorized and challenging to retrieve. Additionally, high computational costs associated with employing accurate computational theories for such intricate systems further contribute to a scarcity of detailed structural data and molecular descriptions. In contrast, function-driven databases, which are typically smaller in scale (ranging from a few hundreds to tens of thousands of entries), focus on particular chemical domains or research questions. Recent efforts to develop specialized datasets for catalysts have shown promise. Notable examples include the KRAKEN database [31], which includes both chiral and achiral monophosphine ligands, and OSCAR [34], which focuses on a diverse range of organocatalysts. Nguyen *et al.* constructed a bias-free dataset and performed methane oxidation coupling catalyst design [35]. Olen *et al.* compiled a collection of BOX ligands and conducted computational analyses [36]. Ferraz-Caetano *et al.* proposed a framework to establish an ML-ready database for epoxidation reactions, with a focus on vanadium catalysts [37]. While these efforts are commendable, they often target specific ligands or catalysts and include a substantial proportion of achiral molecules. Additionally, these databases frequently lack detailed annotations specifying the chiral properties of the molecules and often fail to offer user-friendly interaction capabilities. Furthermore, the application of computational techniques such as ML for rapid and direct studies of chiral ligands and catalysts presents challenges to chemists. For instance, Hueffel *et al.* successfully utilized unsupervised ML methods to accelerate the identification of dinuclear palladium catalysts, accurately predicting several phosphine ligands [38]. Betinol *et al.* proposed a data-driven workflow based on unsupervised ML methods that can rapidly assess and quantify the versatility of catalysts [39]. However, the provided codes have limited reusability, which may present challenges for researchers less familiar with programming. Molecular modeling and analysis in this field remain complex and are hindered by the absence of user-friendly algorithms and tools. As a result, there is a clear need for a specialized and comprehensive database dedicated to chiral ligands and catalysts.

In this work, we develop the Chiral Ligand and Catalyst Database (CLC-DB) (Fig. 1). To our knowledge, CLC-DB is the first open-source and most comprehensive specialized online database of chiral ligands and catalysts, systematically compiling molecular data across diverse catalytic systems. CLC-DB comprises 1,861 molecules spanning 32 types of chiral ligands and catalysts, containing fundamental chiral classes such as point chirality, axial chirality, and planar chirality. Each molecule entry includes 34 data fields, such as 2D and 3D chemical structures, ligand or catalyst categories, chiral classifications, chemical and physical properties, and AI-generated descriptions. High-precision computational methods are employed to simulate

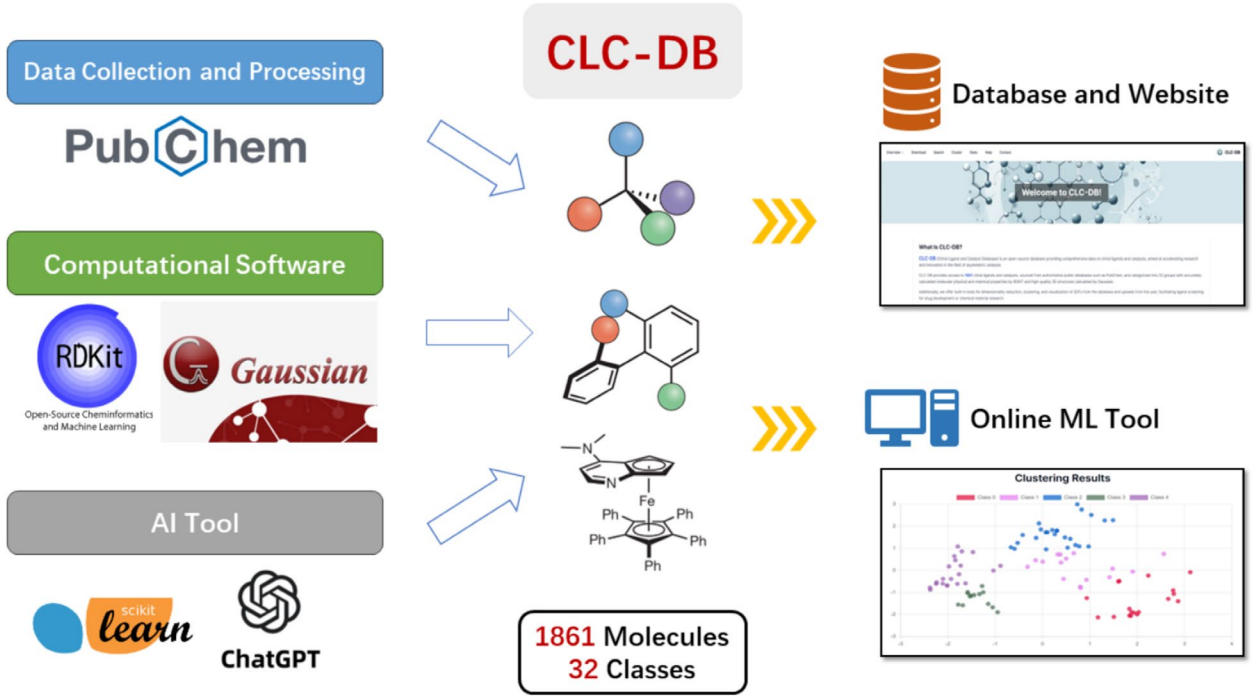

Fig. 1 An overview of CLC-DB

and optimize 3D structures, ensuring accuracy. All molecular data are cross-referenced with authoritative chemical databases and manually validated by chemical experts to ensure reliability and quality.

CLC-DB is a user-friendly database. It supports two quick search methods (text-based and structure-based search) and batch search. Users can perform fuzzy searches using CAS ID and the name of molecules, or structure matching by SMILES and drawing structure. It also facilitates high-speed vector searches based on descriptors. All the data in CLC-DB can be downloaded as CSV/SDF file formats and used for quantitative chemical analysis and molecular design. In addition, CLC-DB provides an online molecular clustering tool for ML computational analyses, improving research efficiency. For chemical scientists, CLC-DB provides intuitive search and retrieval capabilities for identifying target ligands, complemented by an online ML tool for molecular screening and mechanism analysis. For information scientists, CLC-DB serves as a valuable repository of high-quality raw data on chiral ligands and catalysts, ideal for training ML models, such as developing chirality classification models to function as discriminators in generative models.

Materials and methods

Data collection and processing

The chiral ligands and catalysts were primarily collected from various chemical reagent websites (e.g., Bidepharm (https://www.bidepharm.com/), Sigma-Aldrich (https://www.sigmaaldrich.cn/), Thermo Fisher Scientific (https://www.thermofisher.cn/)) and cross-referenced with public databases such as PubChem using CAS IDs. These chemical reagent websites typically feature dedicated sections for ligands and catalysts, where the compounds are categorically organized. These chiral ligands are listed with essential information on these websites, including their names, canonical indices, and 2D structures (presented as 2D images). To ascertain the correctness and completeness of the data, we verified and amended the information using the CAS ID to search in PubChem, and augmented it with additional details like the PubChem CID and SMILES notation. Upon searching for the collected molecules in PubChem, we found that 172 of them lacked an exact match in the database. This discrepancy primarily arises from the presence of coordinated metals and other ions in certain chiral catalysts, making it difficult to precisely identify their corresponding compound data entries in PubChem. To ensure the accuracy of the data, our team collaborated with experienced chemistry researchers, who conducted manual checks on these specific entries.

Table 1 The prompt to generate the description of molecules by GPT-4

Role of GPT-4	Content
System	You are an excellent chemist, particularly adept at organic chemistry and especially skilled in chiral analysis.
User (Name)	Please describe the chiral information of a molecule, the Name of which is {}. Generate results that are as concise and accurate as possible, ideally condensed into a single paragraph. Do not include the Name in the generated results.
User (SMILES)	Please describe the chiral information of a molecule, the SMILES of which is {}. Generate results that are as concise and accurate as possible, ideally condensed into a single paragraph. Do not include the SMILES in the generated results.

Computed molecular properties

In addition to basic information, we calculated the physical and chemical properties of the molecules using software such as Gaussian and RDKit (https://www.rdkit.org/). To get high-quality 3D coordinates, we used Gaussian 09 (Revision D.01) [40] and density functional theory (DFT) to optimize the 3D conformation of each molecule. All structures were optimized in the gas phase using the M062X [41] hybrid functional, which is well-suited for accurately modeling weak interactions. During the optimization process, all atoms were described with def2-SVP [42] (a double-ζ basis set). The M062X/def2-SVP method is notable for its balanced computational efficiency. Its generalized formulation is particularly optimized for capturing dispersion forces, which can more accurately describe the properties of these molecules. These structures are at a local minimum on the potential energy surface (no imaginary frequencies). All structures are stored in SDF format and can be downloaded directly. Quantitative analysis of various molecular electronic and thermodynamic properties was performed using the Multiwfn software (revision 3.7) [43, 44]. The analyzed properties include zero-point correction, thermal correction to energy, thermal correction to enthalpy, thermal correction to Gibbs free energy, the sum of electronic and zero-point energies, the sum of electronic and thermal energies, the sum of electronic and thermal enthalpies, the sum of electronic and thermal free energies, HOMO energy, LUMO energy, HOMO-LUMO gap, and atomic charges (Mulliken/Hirshfeld) for each atom. We also calculated some important physical and chemical properties related to catalytic capacity based on SMILES by using RDKit. These properties include IUPAC name, InChI, InChIKey, molecular formula, molecular weight, heavy atom count, ring count, H-bond acceptor count, H-bond donor count, and rotatable bond count. The SMILES of all molecules for properties calculation are from PubChem, and RDKit is utilized to standardize them. Based on the 3D coordinates optimized by Gaussian software, we used RDKit to generate 2D images of the molecules, which to some extent also reflect the 3D information of the molecules.

Furthermore, we employed GPT-4 as an extensive chemical knowledge base to generate detailed descriptions of chiral molecules through specific prompts. These descriptions were primarily derived using SMILES notations or molecular names (as shown in Table 1). The generated content includes information about molecular components (such as atoms and functional groups), practical functions, and, importantly, specific chiral details, such as the type of chirality and the positions of chiral centers. All generated descriptions were reviewed and verified by expert chemists to ensure accuracy.

Dataset construction

Upon gathering all pertinent data, including basic information and computational properties, we organized and developed the CLC-DB database. Chiral ligands and catalysts were initially categorized into 32 distinct classes based on their molecular characteristics. The primary classification criteria included the functional groups and structural backbones of the molecules. Different classes of asymmetric catalytic reactions typically exhibit preferences for specific chiral ligands and catalysts. This classification allows chemical researchers to select appropriate categories, facilitating the use of CLC-DB for targeted reaction types. Each molecule was subsequently assigned a chiral type, such as point chirality, axial chirality, or planar chirality. The identification and analysis of chiral types are crucial for the study of chiral ligands and catalysts. Historically, this task was achievable solely by skilled chemists, and there remains a lack of accessible methods and data for researchers in other disciplines. Current chemical databases, such as PubChem, do not contain this key information. Researchers often rely on a basic method to assess chirality by searching for the "@" symbol in SMILES strings. In contrast, CLC-DB offers a novel resource that catalogues chiral molecules by their specific chiral types. To ensure data accuracy, the labeling process was performed independently by multiple experienced chemists, followed by final review and aggregation. Consequently, users can directly access batches of data, enabling the development of new ligands based on particular chiral types.

Additionally, some molecules are complex systems containing metals, which are difficult to represent in SMILES and for which there are no corresponding 2D images. For these molecules, we verified the data and reconstructed their 2D images and associated information. Overall, we refined and validated the data to enhance accuracy and user experience. Detailed molecular information is presented in CSV files, and SDF files, which include 3D structural data, are indexed by CAS ID. All files in CLC-DB are freely available for download.

Molecule clustering

In addition to the rich molecular information, we provide a convenient ML clustering tool. Molecular clustering helps to quickly identify molecules with similar catalytic properties, and optimize the screening and design process of chiral ligands and catalysts. It can deepen researchers' understanding of the catalytic reaction mechanisms by revealing structure-activity relationships, thereby guiding more effective catalyst development. For instance, following the method described in [38], researchers aiming to design ligands or catalysts for a specific reaction can begin by compiling a set of molecules known (or likely) to work effectively for that reaction (either downloaded from CLC-DB or independently collected). Candidate molecules can then be clustered alongside these known molecules using unsupervised clustering techniques. Typically, the catalytic performance of candidate molecules tends to strongly align with that of known molecules within the same cluster, especially those with high similarity. This strategy enables the rapid screening of candidate molecules through computational methods. To use the molecular clustering tool, researchers upload a batch of molecules (in SDF format), which are then represented as vectors using molecular descriptors. The server performs dimensionality reduction on these high-dimensional vectors and visualizes the resulting clusters. This online tool removes the need for researchers to write custom code, providing a user-friendly platform suitable for those with diverse technical backgrounds.

In CLC-DB, we use the extended 3D fingerprint (E3FP) [45] and Morgan fingerprint [46] descriptors in RDKit for molecular representation. These descriptors are widely used in cheminformatics and molecular simulations. They encode substructural patterns (e.g., rings, functional groups) as well as physical and chemical properties of molecules. Importantly, both descriptors capture stereochemical information, making them particularly useful for studying molecular spatial configurations [47, 48].

We use various dimensionality reduction methods (t-SNE [49] and PCA) and clustering methods (*K*-means and DBSCAN [50]). t-SNE is good at preserving local structures and is often used for visualizing high-dimensional data. PCA is a statistical method that reduces the vector dimension by transforming it into a set of orthogonal (uncorrelated) variables called principal components, which capture the most variance in the data. As a well-known clustering method, *K*-means is a centroid-based algorithm known for its simplicity and efficiency, making it a popular choice for partitioning data into *K* clusters based on their centroids. In contrast, DBSCAN, a density-based clustering method, is acknowledged for its effectiveness in detecting clusters of arbitrary shapes and identifying outliers, without the need for a pre-specified number of clusters. (Refer to Supplementary Information Section 3 for more details)

Database implementation

The online platform for CLC-DB is built using the Django framework (version 5.0.3) alongside Python 3.10, with all database information stored in SQLite. To ensure security, parameterized queries are employed to protect against SQL injection. Access to static files is strictly controlled to prevent unauthorized access. Data transmission is encrypted via HTTPS to protect against interception. File uploads are rigorously validated and stored in isolated directories to mitigate the risk of malicious file execution. Additionally, a strict cross-origin resource sharing policy is implemented to regulate data access across domains. Text-based searches can be performed by querying CAS IDs or molecule names. For batch searches, users can upload a CSV file containing multiple CAS IDs for rapid processing. Structure-based searches are enabled through SMILES string input or by drawing molecules directly using the interface powered by Kekule.js [51]. The vector similarity search leverages the high-performance Faiss library [52], facilitating efficient descriptor-based similarity computations. Molecular 3D structures are visualized using 3Dmol.js [53], an object-oriented, WebGL-based JavaScript library designed for interactive online molecular visualization. The clustering interface is implemented using Streamlit (https://streamlit.io/) and employs Pyecharts (https://pyecharts.org/) for the graphical representation of clustering results. Database statistics are visualized through Charts.js, which is embedded within the platform for seamless integration. All technical operations within CLC-DB adhere to established specifications and standards. The data storage framework of CLC-DB is designed for scalability. Regular updates are planned to integrate newly available commercial chiral ligands and catalysts, as well as cutting-edge academic discoveries, ensuring the database remains comprehensive and up-to-date.

Results and discussion

Statistics in CLC-DB

The CLC-DB database contains 1,861 data records collected from public databases and categorized into 32 classes. Each record comprises physical and chemical properties calculated with RDKit and Multiwfn, while 3D structures are generated using Gaussian software, yielding 32 distinct information items per record. These comprehensive properties provide valuable insights for the screening and clustering analyses of chiral ligands and catalysts.

Figure 2A illustrates the distribution of the 32 ligand types in the database, revealing that Mono-oxazoline and BOX ligands are the most prevalent, with 294 and 274 occurrences, respectively, accounting for a significant portion of the dataset. These are followed by chiral bis-nitrogen ligands, which appear 143 times. The frequencies of other ligand types decrease progressively, including chiral diphenols (100) and phosphoric acids (94).

Figure 2B presents the proportion of chirality types in the database. Ligands with point chirality dominate, with 1,168 instances constituting 63.5% of the dataset. Axial chirality ligands are the second most abundant, with 416 occurrences, representing 22.6%. Ligands that exhibit both point chirality and axial chirality, as well as combinations of point chirality and planar chirality, account for smaller proportions at 8.3% and 5.4%, respectively, while planar chirality ligands make up only 0.3%. Despite their smaller representation, these ligands play critical roles within the database.

Figure 2C depicts the distribution of molecular weights in the database. The molecular weights predominantly fall within the range of 100 to 800 g/mol, peaking at the 300-400 g/mol interval with 368 occurrences. As molecular weight increases, the number of compounds decreases, particularly for those exceeding 800 g/mol.

Figure 2D and 2E depict the distributions of molecular HOMO and LUMO energies across the database. The HOMO energy distribution (Fig. 2D) is concentrated within the range of -9 to -6 eV, peaking around -7 eV. This prevalence of HOMO energies suggests that a substantial portion of the molecules in the database exhibit high chemical stability. In contrast, the LUMO energy distribution (Fig. 2E) peaks at approximately -2 eV and spans a wider range of higher energies. This broader distribution indicates varying electron-accepting capabilities among the molecules, with some displaying lower LUMO energies that imply greater reactivity and propensity for involvement in chemical reactions.

Database search and download procedures

CLC-DB provides users with precise and rapid search modules (Fig. 3). The database integrates two searching methodologies: text-based and structure-based searches. The text-based approach predominantly utilizes CAS ID and molecule names. Users can locate chiral ligands and catalysts by entering either the CAS ID or the molecule's name. For batch searches, users can upload a CSV file with CAS IDs in the designated format, and the database will generate all corresponding search results. Alternatively, the structure-based search offers two methods of entry: users can input SMILES codes or graphically depict the molecule using Kekule.js.

Upon completing the search, the main webpage displays the results as small molecule cards that include the CAS ID, 2D images, and succinct molecular information. Users can access detailed information by clicking the "Detail" button, which redirects to the molecule's main page. Notably, text-based searches in CLC-DB support fuzzy search capabilities, allowing users to retrieve a broad array of results by entering partial information.

CLC-DB also provides users with a convenient data download module (Fig. 4). In the middle of the "Download" page, all molecules are listed in 32 classes, e.g., BIDIME and BIBOP Ligands. After selecting a class, users can click "Download ALL" or "Download Page" to download all molecules in that class or website page. There is also a separate download button on the molecule card. Additionally, users can click the "Detail" button to jump to the main page of the molecule to view more information and download data. All molecule data can be downloaded in CSV/SDF file formats for free.

Ligand and catalyst information

CLC-DB offers users two convenient browsing methods ("Card" and "Table") along with categorized options that enhance navigation efficiency (Fig. 5A). Users can first select their desired browsing style and then choose a category relevant to the chiral ligand or catalyst of interest. In the "Card" view, the interface displays the molecule's image, name, and CAS ID, offering a visually detailed browsing experience. In contrast, the "Table" view adopts a minimalist approach, showing only the molecule's name and CAS ID for a more streamlined and efficient display. Detailed molecular information can be accessed by clicking the "Detail" button, which directs users to the molecule's main page.

On the website of CLC-DB, the information regarding chiral ligands and catalysts is organized into two primary sections: basic information and computational properties. Within the webpage layout, there are five distinct sections: basic information, image, 3D structure, computational properties, and description (Fig. 5B, C, D). This

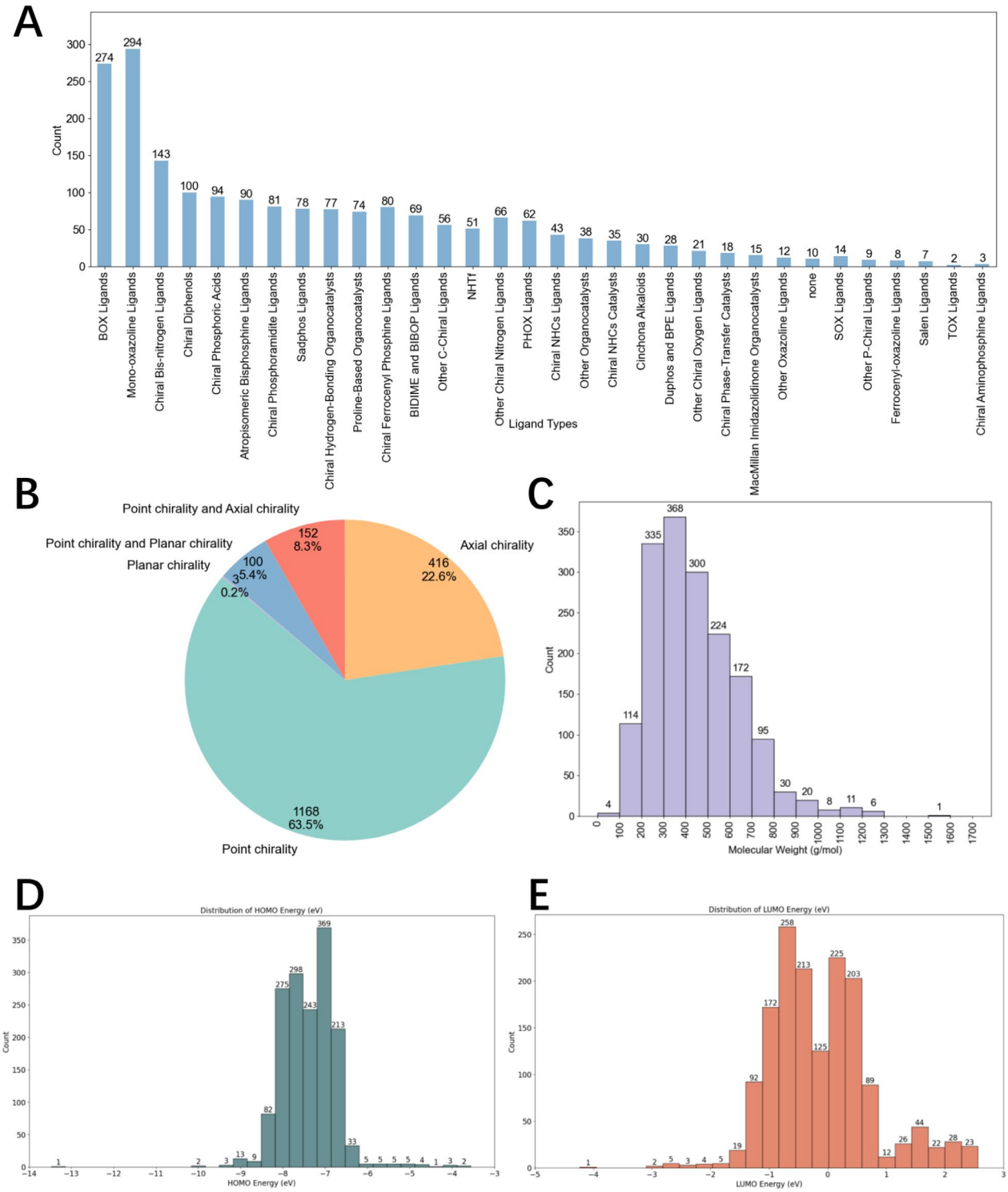

Fig. 2 Data statistics of CLC-DB. **A** The distribution of ligand types in the database. **B** The proportion of chirality types in the database. **C** The molecular weight distribution. **D** The molecular HUMO energy distribution. **E** The molecular LUMO energy distribution

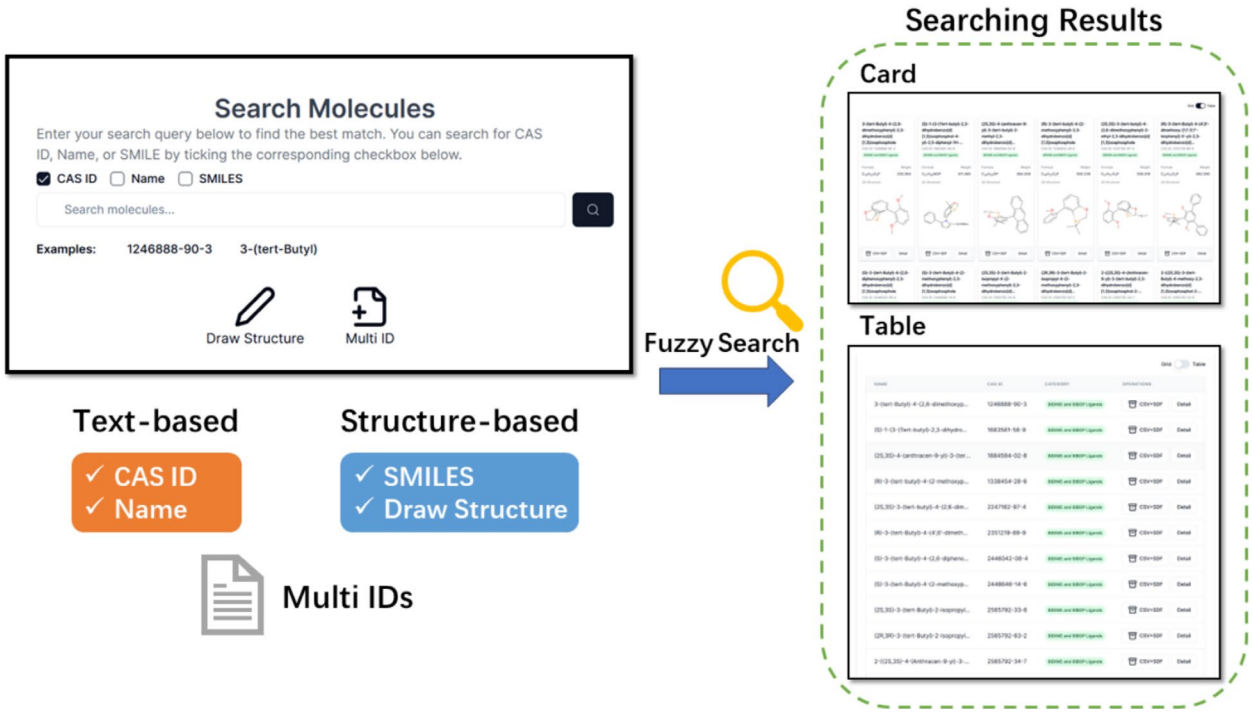

Fig. 3 The search page and result page of CLC-DB

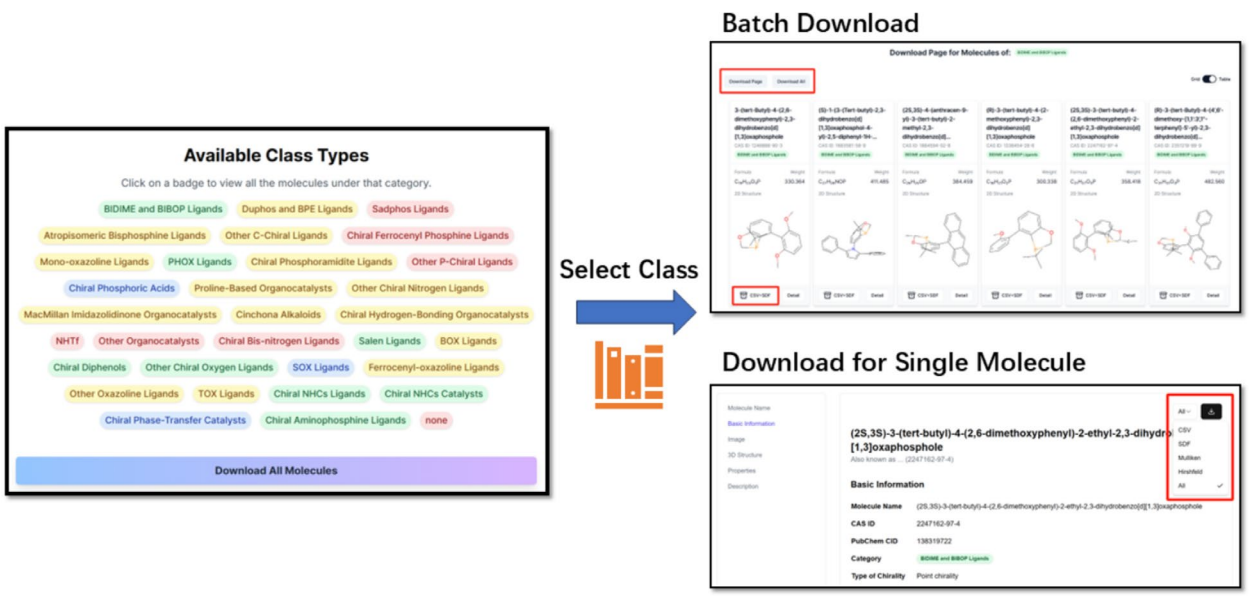

Fig. 4 The data download page of CLC-DB

structured presentation aids in providing a comprehensive overview of each molecule.

The basic information includes molecule name, CAS ID, PubChem CID, category, type of chirality, URL, PubChem URL, and SMILES notation. Users can click on the URL and PubChem URL links to access corresponding pages for more details. Molecule images are provided to help users intuitively analyze molecules and better understand their 3D structures. The 3D structures of molecules are displayed within an interactive window,

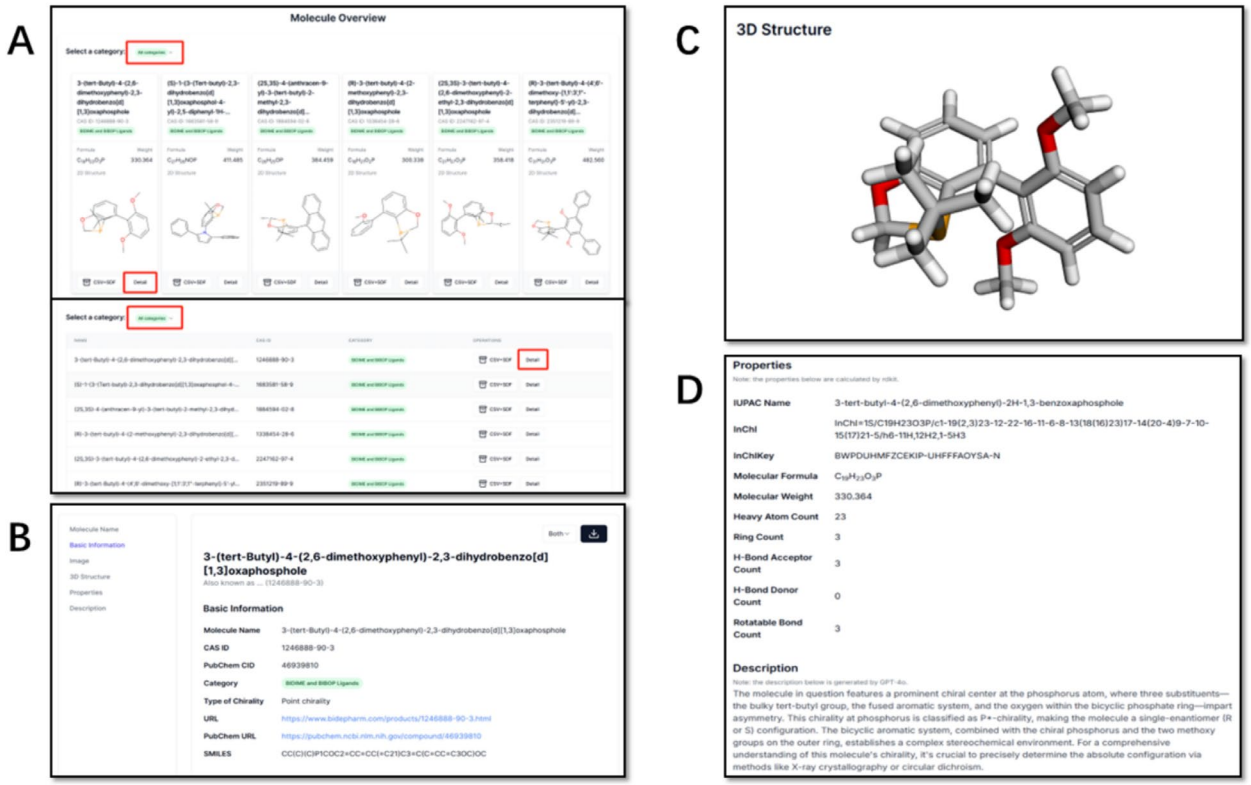

Fig. 5 Overview and detailed pages for molecules in CLC-DB. **A** Molecule overview page in CLC-DB, where users can choose between "Card" and "Table" browsing modes and select the molecule class. **B** Basic information about the molecule. **C** 3D molecular structure display window. **D** Calculated physical and chemical properties along with the molecule's description

allowing users to rotate and zoom in or out. In CLC-DB, 3D conformation optimization is performed using Gaussian software, incorporating carefully selected high-quality DFT methods. More accurate 3D coordinates are obtained for all molecules except for some metal systems with planner chiral ligands. In other databases, such as PubChem, 3D coordinates for these molecules are frequently missing or of suboptimal quality due to the limitations of computational methods. Thus, CLC-DB serves as a valuable resource by providing high-quality 3D information, complementing existing databases and supporting broader chemistry studies. For example, these 3D structures can be utilized as source data to generate descriptors and representations, which are essential for applications such as reaction prediction, retrosynthetic analysis, and molecule generation.

In addition, computational properties and detailed descriptions are displayed below the 3D structure window. RDKit and Multiwfn are leveraged to calculate physical and chemical properties, while GPT-4 is employed to generate detailed molecular descriptions, facilitating comprehensive analyses of chiral ligands and catalysts. These descriptions offer new insights into the study of these molecules, encouraging researchers to explore them from innovative perspectives. For instance, by integrating chirality types and molecular categories with computational properties and detailed descriptions of chiral centers and configurations, researchers can more effectively deduce potential reaction types and conditions. All molecular data is available for download in CSV or SDF file formats.

Molecular clustering tool

Beyond conventional functionalities, CLC-DB offers an integrated online ML tool that facilitates rapid molecular clustering. Users can upload a batch of molecules and receive an intuitive visualization of the clustering results (Fig. 6). The clustering process consists of several steps. First, users upload SDF files, which can be flexibly removed or replaced if needed. Next, users select the molecular descriptor to be used. CLC-DB supports two descriptor options: E3FP and Morgan, both of which capture molecular structure information while maintaining computational efficiency. Each descriptor includes various adjustable parameters, such as the number of bits in the vector (*n*-Bits), allowing users to fine-tune the

Fig. 6 The molecular clustering tool in CLC-DB

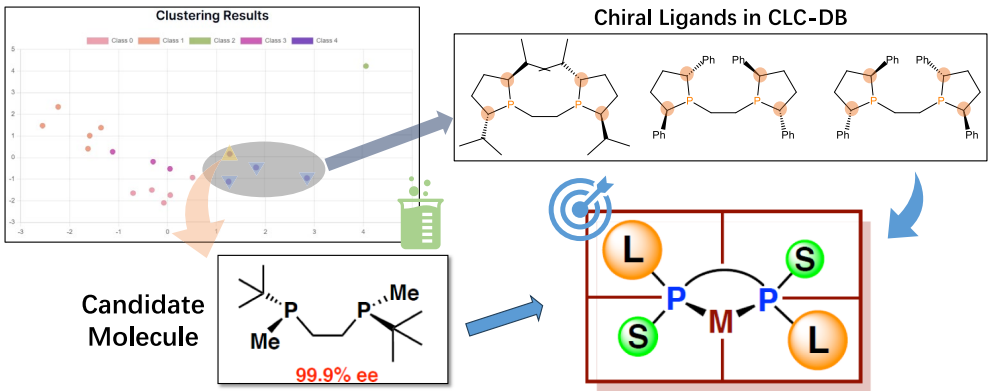

Fig. 7 Application of molecular clustering tool in asymmetric hydrogenation reactions

settings based on specific molecular properties. After configuring the dimension reduction and clustering methods, users can submit their data. Upon submission, the system promptly displays the clustering results.

Building upon the previously described clustering functionality, we now illustrate its application with a case study. This study focuses on asymmetric hydrogenation reactions, employing molecular clustering tools to screen chiral ligands. Asymmetric hydrogenation is a pivotal class of catalytic reactions, essential for the synthesis of numerous drugs, fragrances, and other high-value small molecules [54–57]. It remains an area of active research, with ongoing efforts to explore and develop new systems.

P-Chiral Bis(trialkylphosphine) Ligands constitute a highly efficient family of chiral ligands [58]. Despite their potential, they are commercially underrepresented and not extensively cataloged in databases such as CLC-DB. This class of ligands belongs to the Duphos category. To begin our study, we compiled all available molecules in this class and uploaded them, along with the candidate P-Chiral Bis(trialkylphosphine) ligands, to the clustering tool for analysis. As illustrated in Fig. 7, the candidate

molecules belong to a class comprising four members, with the other three being well-established in hydrogenation reactions. Importantly, these ligands exhibit considerable mechanistic similarity to the candidates. A key feature of the candidate ligands is the presence of a bulky alkyl group and the smallest alkyl group (a methyl group) bonded to each phosphorus atom. This unique spatial structure introduces selectivity [58, 59]. The other three ligands within the same class achieve a comparable catalytic spatial structure via carbon chirality on their rings.

Conversely, other uploaded ligands demonstrate either lower catalytic activity or significantly different reaction mechanisms. This example highlights the tool's capability to identify similar reaction mechanisms, thereby facilitating the screening of ligands and catalysts.

Conclusion

We introduce CLC-DB, the first open-source database specifically for chiral ligands and catalysts, designed to provide free and valuable data resources for research and development in this field. CLC-DB offers accurate and comprehensive data on chiral ligands and catalysts, encompassing 1,861 molecules and their precisely calculated and validated molecular details. Additionally, CLC-DB features a user-friendly ML tool for efficient molecular clustering and analysis, enhancing the productivity of related research endeavors. Despite these advancements, room for further enhancement remains. The current collection of chiral ligands and catalysts in CLC-DB is not fully exhaustive, and the accuracy and scope of theoretical calculations for complex systems require additional improvement. Looking forward, our aim is to expand CLC-DB by incorporating a more extensive range of newly developed chiral ligands and catalysts, as well as advancing the online ML tool for enhanced molecular analysis and design. Future developments will include sophisticated virtual screening tools, highly accurate reaction outcome prediction models, and advanced algorithms for ligand optimization and design. We are confident that CLC-DB will become an indispensable resource and a powerful tool in the study of chiral ligands and catalysts, particularly for advancing the design of asymmetric catalytic syntheses through the integration of ML technologies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00991-9.

Additional file 1.

Author contributions

G.Y., K.Y., X.W, C.Z., and Y.L. collected and processed the datasets, and built the webserver. X.H. and Y.Y. supervised the project. G.Y., K.Y., X.W., X.H., and Y.Y. analyzed the results. G.Y., K.Y., and Y.Y. wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work is supported by the National Key R&D Program of China (No. 2023YFC2811500 and 2018YFE0126800) and the National Natural Science Foundation of China (Nos. 62272300, 21831005, 21991112, and 22171183).

Data availability

All information recorded in CLC-DB can be freely accessible and downloaded at https://compbio.sjtu.edu.cn/services/clc-db. The full chiral molecule dataset is freely available under a CC-BY-NC 4.0 license (https://creativecommons.org/licenses/by/4.0). Our code is available on GitHub (https://github.com/ukeSJTU/clc-db). The visualization tool is released under the MIT License (https:/opensource.org/licenses/MIT).

Declarations

Competing interests

The authors declare no competing interests.

Received: 29 August 2024 Accepted: 20 March 2025
Published online: 03 April 2025

References

1. Kagan HB (1985) Chiral ligands for asymmetric catalysis. Asymmetric Synthesis 5:1–39
2. Zhou QL (2011) Privileged chiral ligands and catalysts. John Wiley & Sons
3. Mallat T, Orglmeister E, Baiker A (2007) Asymmetric catalysis at chiral metal surfaces. Chem Rev 107(11):4863–4890
4. Bauer EB (2012) Chiral-at-metal complexes and their catalytic applications in organic synthesis. Chem Soc Rev 41(8):3153–3167
5. Cao ZY, Brittain WD, Fossey JS et al (2015) Recent advances in the use of chiral metal complexes with achiral ligands for application in asymmetric catalysis. Catalysis Sci Technol 5(7):3441–3451
6. Doyle AG, Jacobsen EN (2007) Small-molecule h-bond donors in asymmetric catalysis. Chem Rev 107(12):5713–5743
7. Brandt JR, Salerno F, Fuchter MJ (2017) The added value of small-molecule chirality in technological applications. Nat Rev Chem 1(6):0045
8. Gennari C, Piarulli U (2003) Combinatorial libraries of chiral ligands for enantioselective catalysis. Chem Rev 103(8):3071–3100
9. Kang QK, Wang L, Liu QJ et al (2015) Asymmetric h2o-nucleophilic ring opening of d-a cyclopropanes: catalyst serves as a source of water. J Am Chem Soc 137(46):14594–14597
10. Xia Y, Liu X, Feng X (2021) Asymmetric catalytic reactions of donor-acceptor cyclopropanes. Angew Chem 133(17):9276–9288
11. He YM, Cheng YZ, Duan Y et al (2023) Recent progress of asymmetric catalysis from a Chinese perspective. CCS Chem 5(12):2685–2716
12. Koskinen AM (2022) Asymmetric synthesis of natural products. John Wiley & Sons
13. Gladiali S, Alberico E (2006) Asymmetric transfer hydrogenation: chiral ligands and applications. Chem Soc Rev 35(3):226–236
14. Noyori R (2003) Asymmetric catalysis: science and opportunities (nobel lecture 2001). Adv Synth Catal 345(1–2):15–32
15. Farina V, Reeves JT, Senanayake CH et al (2006) Asymmetric synthesis of active pharmaceutical ingredients. Chem Rev 106(7):2734–2793
16. Sawamura M, Ito Y (1992) Catalytic asymmetric synthesis by means of secondary interaction between chiral ligands and substrates. Chem Rev 92(5):857–871
17. Zhang W, Chi Y, Zhang X (2007) Developing chiral ligands for asymmetric hydrogenation. Acc Chem Res 40(12):1278–1290
18. Gopalaiah K (2013) Chiral iron catalysts for asymmetric synthesis. Chem Rev 113(5):3248–3296

19. Hong X, Yang Q, Liao K et al (2024) Ai for organic and polymer synthesis. Sci China Chem 1:36
20. Du Y, Jamasb AR, Guo J et al (2024) Machine learning-aided generative molecular design. Nat Machine Intell 1:16
21. Bishop CM, Nasrabadi NM (2006) Pattern recognition and machine learning, vol 4. Springer
22. Clauset A, Larremore DB, Sinatra R (2017) Data-driven predictions in the science of science. Science 355(6324):477–480
23. Fortunato S, Bergstrom CT, Börner K et al (2018) Science of science. Science 359(6379):0185
24. Montáns FJ, Chinesta F, Gómez-Bombarelli R et al (2019) Data-driven modeling and learning in science and engineering. Comptes Rendus Mécanique 347(11):845–855
25. Hey T, Tansley S, Tolle KM et al (2009) The fourth paradigm: data-intensive scientific discovery, vol 1. Microsoft research Redmond, WA
26. Toyao T, Maeno Z, Takakusagi S et al (2019) Machine learning for catalysis informatics: recent applications and prospects. ACS Catal 10(3):2260–2297
27. Noé F, Tkatchenko A, Müller KR et al (2020) Machine learning for molecular simulation. Annu Rev Phys Chem 71(1):361–390
28. Pyzer-Knapp EO, Pitera JW, Staar PW et al (2022) Accelerating materials discovery using artificial intelligence, high performance computing and robotics. NPJ Computational Materials 8(1):84
29. Sun W, Zheng Y, Yang K et al (2019) Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. Sci Adv 5(11):4275
30. Xu LC, Zhang SQ, Li X et al (2021) Towards data-driven design of asymmetric hydrogenation of olefins: Database and hierarchical learning. Angew Chem Int Ed 60(42):22804–22811
31. Gensch T, dos Passos Gomes G, Friederich P et al (2022) A comprehensive discovery platform for organophosphorus ligands for catalysis. J Am Chem Soc 144(3):1205–1217
32. Kim S, Chen J, Cheng T et al (2023) Pubchem 2023 update. Nucleic Acids Res 51(D1):D1373–D1380
33. Hastings J, Owen G, Dekker A et al (2016) Chebi in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res 44(D1):D1214–D1219
34. Gallarati S, van Gerwen P, Laplaza R et al (2022) Oscar: an extensive repository of chemically and functionally diverse organocatalysts. Chem Sci 13(46):13782–13794
35. Nguyen TN, Nakanowatari S, Nhat Tran TP et al (2021) Learning catalyst design based on bias-free data set for oxidative coupling of methane. ACS Catal 11(3):1797–1809
36. Olen CL, Zahrt AF, Reilly SW et al (2024) Chemoinformatic catalyst selection methods for the optimization of copper-bis (Oxazoline)-mediated, asymmetric, vinylogous mukaiyama aldol reactions. ACS Catal 14(4):2642–2655
37. Ferraz-Caetano J, Teixeira F, Cordeiro MND (2024) Navigating epoxidation complexity: building a data science toolbox to design vanadium catalysts. New J Chem 48(12):5097–5100
38. Hueffel JA, Sperger T, Funes-Ardoiz I et al (2021) Accelerated Dinuclear palladium catalyst identification through unsupervised machine learning. Science 374(6571):1134–1140
39. Betinol IO, Lai J, Thakur S et al (2023) A data-driven workflow for assigning and predicting generality in asymmetric catalysis. J Am Chem Soc 145(23):12870–12883
40. Frisch MJ, Trucks GW, Schlegel HB, et al (2009) Gaussian˜09 Revision D.01. Gaussian Inc. Wallingford CT
41. Zhao Y, Schultz NE, Truhlar DG (2006) Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions. J Chem Theory Comput 2(2):364–382
42. Weigend F, Ahlrichs R (2005) Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to RN: Design and assessment of accuracy. Phys Chem Chem Phys 7:3297–3305
43. Lu T, Chen F (2012) Multiwfn: a multifunctional wavefunction analyzer. J Comput Chem 33(5):580–592
44. Lu T (2024) A comprehensive electron wavefunction analysis toolbox for chemists Multiwfn. J Chem Phys 161:8
45. Axen SD, Huang XP, Cáceres EL et al (2017) A simple representation of three-dimensional molecular structure. J Med Chem 60(17):7393–7409
46. Morgan HL (1965) The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J Chem Doc 5(2):107–113
47. Riniker S, Landrum GA (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. J Cheminform 5:1–17
48. Xu Y, Cai C, Wang S et al (2019) Efficient molecular encoders for virtual screening. Drug Discov Today Technol 32:19–27
49. Van der Maaten L, Hinton G (2008) Visualizing data using t-Sne. J Machine Learn Res 9:11
50. Ester M, Kriegel HP, Sander J, et al (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd, pp 226–231
51. Jiang C, Jin X, Dong Y et al (2016) Kekule js: an open source javascript chemoinformatics toolkit. J Chem Inform Model 56(6):1132–1138
52. Douze M, Guzhva A, Deng C, et al (2024) The faiss library. arXiv:2401.08281
53. Rego N, Koes D (2015) 3dmol. js: molecular visualization with webgl. Bioinformatics 31(8):1322–1324
54. Wen J, Wang F, Zhang X (2021) Asymmetric hydrogenation catalyzed by first-row transition metal complexes. Chem Soc Rev 50(5):3211–3237
55. Xie JH, Zhu SF, Zhou QL (2011) Transition metal-catalyzed enantioselective hydrogenation of enamines and imines. Chem Rev 111(3):1713–1760
56. Imamoto T (2024) P-stereogenic phosphorus ligands in asymmetric catalysis. Chem Rev 124(14):8657–8739
57. Zhang Z, Butt NA, Zhang W (2016) Asymmetric hydrogenation of nonaromatic cyclic substrates. Chem Rev 116(23):14769–14827
58. Imamoto T, Watanabe J, Wada Y et al (1998) P-chiral bis(trialkylphosphine) ligands and their use in highly enantioselective hydrogenation reactions. J Am Chem Soc 120(7):1635–1636
59. Gridnev ID, Imamoto T (2009) Mechanism of Enantioselection in rh-catalyzed asymmetric hydrogenation. the origin of utmost catalytic performance. Chem Commun 7447:7464

Publisher's Note