frontiers | Frontiers in Pharmacology

TYPE Original Research
PUBLISHED 18 June 2024
DOI 10.3389/fphar.2024.1400136

OPEN ACCESS

EDITED BY
Vinícius Gonçalves Maltarollo,
Federal University of Minas Gerais, Brazil

REVIEWED BY
Júlia Galvez Bulhões Pedreira,
University of Tübingen, Germany
José L. Medina-Franco,
National Autonomous University of Mexico,
Mexico
Ricardo A. M. Serafim,
University of Tübingen, Germany
Cleber C. Melo-Filho,
University of North Carolina at Chapel Hill,
United States

*CORRESPONDENCE
Zuojian Zhou,
✉ zhouzj@njucm.edu.cn
Ye Yang,
✉ yangye876@sina.com
Shanliang Sun,
✉ ssun@njucm.edu.cn

RECEIVED 14 March 2024
ACCEPTED 28 May 2024
PUBLISHED 18 June 2024

CITATION
Yuan Y, Tang X, Li H, Lang X, Li C, Song Y, Sun S, Yang Y and Zhou Z (2024), KLSD: a kinase database focused on ligand similarity and diversity.
Front. Pharmacol. 15:1400136.
doi: 10.3389/fphar.2024.1400136

KLSD: a kinase database focused on ligand similarity and diversity

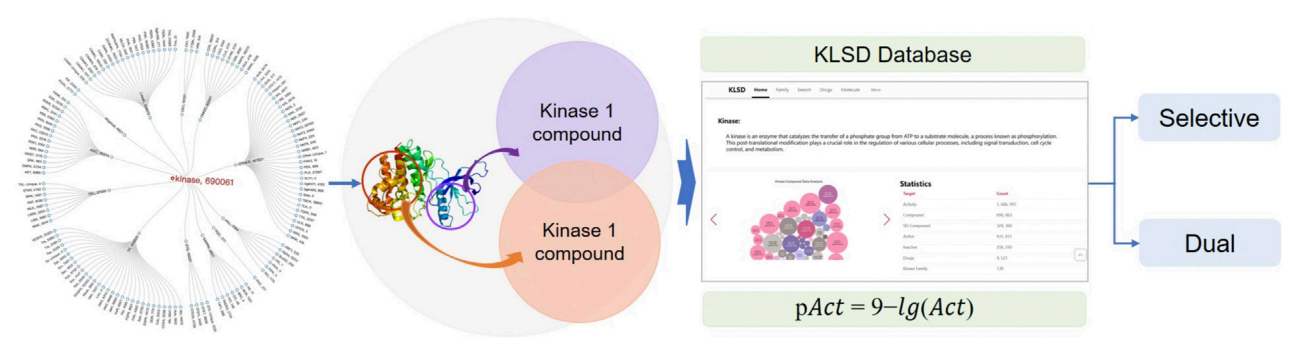
Yuqian Yuan[1], Xiaozhu Tang[2], Hongyan Li[1], Xufeng Lang[1], Can Li[1], Yihua Song[1], Shanliang Sun[3]*, Ye Yang[2]* and Zuojian Zhou[1]*

[1]School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing, China, [2]School of Medicine and Holistic Integrative Medicine, Nanjing University of Chinese Medicine, Nanjing, China, [3]National and Local Collaborative Engineering Center of Chinese Medicinal Resources Industrialization and Formulae Innovative Medicine, Jiangsu Collaborative Innovation Center of Chinese Medicinal Resources Industrialization, Jiangsu Key Laboratory for High Technology Research of TCM Formulae, Nanjing University of Chinese Medicine, Nanjing, China

Due to the similarity and diversity among kinases, small molecule kinase inhibitors (SMKIs) often display multi-target effects or selectivity, which have a strong correlation with the efficacy and safety of these inhibitors. However, due to the limited number of well-known popular databases and their restricted data mining capabilities, along with the significant scarcity of databases focusing on the pharmacological similarity and diversity of SMIKIs, researchers find it challenging to quickly access relevant information. The KLIFS database is representative of specialized application databases in the field, focusing on kinase structure and co-crystallised kinase-ligand interactions, whereas the KLSD database in this paper emphasizes the analysis of SMKIs among all reported kinase targets. To solve the current problem of the lack of professional application databases in kinase research and to provide centralized, standardized, reliable and efficient data resources for kinase researchers, this paper proposes a research program based on the ChEMBL database. It focuses on kinase ligands activities comparisons. This scheme extracts kinase data and standardizes and normalizes them, then performs kinase target difference analysis to achieve kinase activity threshold judgement. It then constructs a specialized and personalized kinase database platform, adopts the front-end and back-end separation technology of SpringBoot architecture, constructs an extensible WEB application, handles the storage, retrieval and analysis of the data, ultimately realizing data visualization and interaction. This study aims to develop a kinase database platform to collect, organize, and provide standardized data related to kinases. By offering essential resources and tools, it supports kinase research and drug development, thereby advancing scientific research and innovation in kinase-related fields. It is freely accessible at: http://ai.njucm.edu.cn:8080.

KEYWORDS

kinase, database, SMKIs, pharmacological similarity, ligand activity



Kinase:

A kinase is an enzyme that catalyzes the transfer of a phosphate group from ATP to a substrate molecule, a process known as phosphorylation. This post-translational modification plays a crucial role in the regulation of various cellular processes, including signal transduction, cell cycle control, and metabolism.

Statistics

Target	Count
Activity	1, 686, 997
Compound	690, 063
SD-Compound	324, 360
Active	431, 811
Inactive	258, 250
Drugs	4, 521
Kinase Family	138

$$pAct = 9 - lg(Act)$$

GRAPHICAL ABSTRACT

1 Introduction

Protein kinase is an important signal messenger that plays a crucial role in cellular activities. It functions as an enzyme that specifically transfers the γ-phosphate group from ATP to amino acid residues on substrate proteins. This phosphorylation enables the transmission of various signals that are necessary for numerous physiological processes (Fischer, 2016). Pharmacological and pathological studies have shown that kinases are an ideal and important drug targets for drug development in the treatment of diseases such as tumors, inflammatory diseases, central nervous system diseases, cardiovascular diseases and diabetes (Wu et al., 2015). Kinases typically contain an active center that plays a key role in catalyzing phosphorylation reactions. These enzymes are often regulated through the binding of other proteins, phosphorylation events, and the influence of other molecules. Given the significant biological functions of kinases in organisms and their potential as drug targets (Mukhtar et al., 2017), many researchers have been focused on studying SMKIs. These SMKIs can be used to modulated aberrant signaling pathways and treat a variety of diseases as mentioned. As science and technology advance, the proliferation of biomedical databases continues (Attwood et al., 2021; Cohen et al., 2021). However, language barriers, heavy workloads, and other obstacles often limit us to a few well-known databases. Heavy workloads often involve tasks such as conducting kinase target difference analysis, which may require importing kinase data into EXCEL, CSV, etc., for manual calculation, screening standard type and relationship manually, and calculating standard values. These tasks can significantly increase the workload of the study. Additionally, other obstacles, such as a cluttered website layout that makes it difficult to quickly locate research-related content, can also present challenges. This widespread reliance on popular databases restricts novel discovery opportunities and hinders innovation. In particular, research involving small molecule kinase inhibitors (SMKIs) critically requires a comprehensive pharmacological database. A kinase database that maintains strict integrity, provides high-quality data, and offers extensive customization is invaluable. It is imperative to venture beyond familiar grounds and delve into specialized, lesser-known research databases to discover and utilize new data resources, driving pioneering breakthroughs.

With the continuous development of artificial intelligence (AI) technology, significant progress has been made in kinase informatics research. In drug research, AI technology can leverage various molecular representation properties, such as molecular graph-based feature representation (Duvenaud et al., 2015; Kevin et al., 2019; Mengying et al., 2020), molecular string-based representation (Nathan et al., 2018; Popova et al., 2018; Karpov et al., 2020) image-based representation (Charles et al., 2018; Cortés-Ciriano and Bender, 2019), and knowledge-based molecular representation (Jan et al., 2019; Xiang et al., 2021). This allows for satisfactory performance in the research and analysis of various drug properties, including activity, pharmacokinetic properties, and toxicological properties. This development is conducive to facilitating early drug discovery. The advantages of AI technology in kinase informatics research are mainly reflected in the following aspects.

i) Rich data resources: Kinase data consolidates a vast amount of information about genes, protein structures, phosphorylation sites, regulatory pathways. This provides researchers with abundant data resources to explore the functions and roles of kinases.
ii) Diverse functional annotations: Kinase data not only offers basic information about kinases, but also encompasses functional annotations relating to cell signaling, disease development, and drug discovery. This comprehensive data can aid in a more thorough understanding of the biological roles of kinases.
iii) Bioinformatics tools: Numerous biological databases supply bioinformatics tools for data analysis, structure prediction, and protein interactions. These tools enable researchers to interpret and utilize data more effectively.
iv) Data integration and cross-linking: Large-scale biological databases integra data from various sources, providing cross-linkages to help researchers conduct more in-depth studies using integrated data.

Although kinase informatics research has achieved significant breakthroughs, there are still deficiencies in data quality consistency, data standardization, data visualization, personalized functional analysis, and prediction. Currently, there is a lack of specialized databases for kinases in the field of kinase informatics research. The KLIFS (Kanev et al., 2021) (Kinase-Ligand Interaction Fingerprints

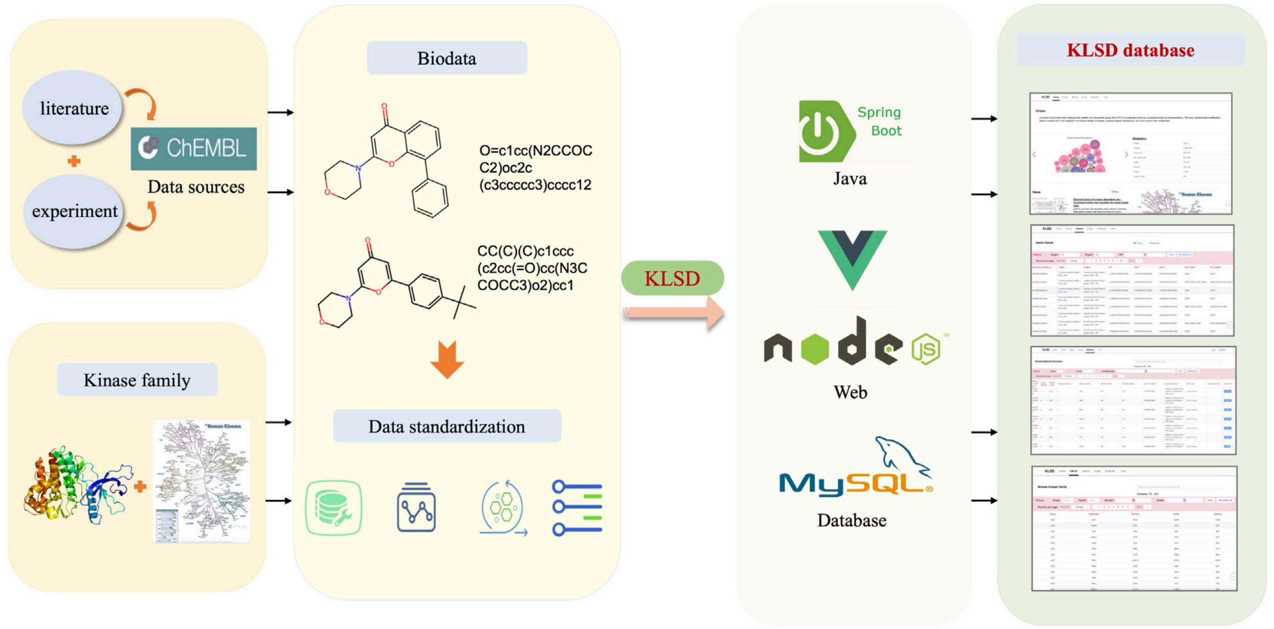


FIGURE 1
Comprehensive Architecture of the KLSD Database Platform. The KLSD database emphasizes the analysis of SMKIs among all reported kinase targets.

and Structures) database is one such specialized database. Obviously, there are differences in focus and application areas between the KLIFS database and the KLSD database discussed in this paper. KLIFS primarily focuses on kinase structures and co-crystalized kinase-ligand interactions, facilitating kinase drug discovery and design. On the other hand, KLSD emphasizes all reported SMKIs analysis between kinase targets. Therefore, based on the current foundations of kinase data, further research is being conducted to construct a professional and personalized kinase database. This database not only provides comprehensive data on the structure, function, regulation and biological roles of kinase-related proteins, but also helps to understand interactions between kinases and other proteins, molecules, or biological systems through the analysis of target difference analysis. This comprehensive understanding of kinase systems better support both scientific research and design, as well as signaling applications.

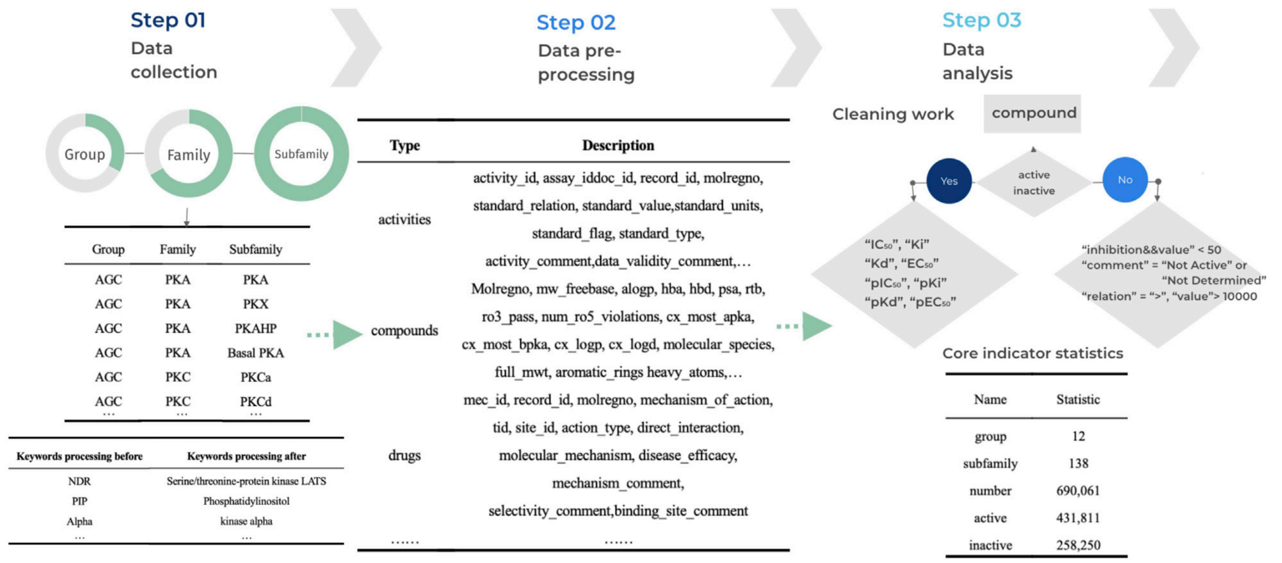
2 Materials and methods

2.1 Data collection and processing

The KLSD database platform is mainly based on the data in the ChEMBL (Chemical Database of Bioactive Molecules) database (Mendez et al., 2019). The data in the ChEMBL database is primarily derived from scientific literature, patent information, public databases, high-throughput screening data, clinical trials and drug registration information. These multiple data sources and channels ensure the accuracy and comprehensiveness of the data. Therefore, for our experiments, we used data from the CHEMBL32 version of the ChEMBL database. ChEMBL currently provides multiple publicly available web services for accessing compound and bioactivity data in their database. Through their web service, we obtained a total of 78 data tables with 57,424,581 records. These tables cover various of aspects such as genomics, protein expression, small molecule, system, ontologies and scientific literature. with 15,398 kinds of targets of all kinds and more than 2,390,000 different compounds. Since the focus of this study is mainly related to kinases, we selectively extract kinase-related data to build a kinase dataset. By conducting a keyword search, we initially obtained 1,131,037 kinase-related records. The data extraction process and platform construction is shown in Figure 1.

In biology, hierarchical classification aids in better understanding the similarities and differences among kinases, as well as their roles in cell signaling and biological processes. Thus, we employed the classification framework provided by Manning et al. (Manning et al., 2002), which is based on sequence homology and functional characteristics to categorize kinases into groups, families, and subfamilies. Kinases are first grouped based on their overall structural domains and sequence similarities of their kinase domains. Within each group, kinases are further classified into families using more specific sequence similarities and shared functional characteristics. Subsequently, within each family, subfamilies are established, which may consist of different forms of a single gene product or closely related genes with similar functions. This classification method is commonly used in biological research and drug development, aiding in the identification of relationships among different kinases and improving our understanding of their intricate roles in various signaling pathways and diseases. This process resulted in a total of 14 clusters, 166 families, and 314 subfamilies.

Data processing refers to the process of performing a series of operations and transformations on raw data to obtain valuable



Group	Family	Subfamily
AGC	PKA	PKA
AGC	PKA	PKX
AGC	PKA	PKAHP
AGC	PKA	Basal PKA
AGC	PKC	PKCa
AGC	PKC	PKCd
…	…	…

Keywords processing before	Keywords processing after
NDR	Serine/threonine-protein kinase LATS
PIP	Phosphatidylinositol
Alpha	kinase alpha
…	…

Type	Description
activities	activity_id, assay_iddoc_id, record_id, molregno, standard_relation, standard_value,standard_units, standard_flag, standard_type, activity_comment,data_validity_comment,…
compounds	Molregno, mw_freebase, alogp, hba, hbd, psa, rtb, ro3_pass, num_ro5_violations, cx_most_apka, cx_most_bpka, cx_logp, cx_logd, molecular_species, full_mwt, aromatic_rings heavy_atoms,…
drugs	mec_id, record_id, molregno, mechanism_of_action, tid, site_id, action_type, direct_interaction, molecular_mechanism, disease_efficacy, mechanism_comment, selectivity_comment,binding_site_comment
……	……

Name	Statistic
group	12
subfamily	138
number	690,061
active	431,811
inactive	258,250

FIGURE 2
Workflow for KLSD Data Processing and Standardization. It describles the key aspects of data processing—including cleaning, standardization, and normalization.

information. These operations can include extracting specific features, cleaning the data, eliminating noise, and converting the formats, among others (Han et al., 2022). The purpose of data processing is to make the data more useable, comprehensible, and applicable for further analyses, applications, and decision-making. To effectively improve the data quality of kinase data, reduce errors and bias in the analysis, and facilitate later data analysis, mining and modelling for more accurate and meaningful conclusions, we have implemented standardization measures for the data. This ensures data consistency, comparability and accuracy. The initial step involves processing data outlier, where we eliminate any potential outliers, such as invalid entries and missing values, to maintain data integrity. We consider an outlier when the data value is null or empty, or when the data value has an abnormal symbol such as at, #, etc. If the length of the data value falls outside the normal range of values for that property, we consider that data is invalid. Following this, we standardize the data by aligning it across different sources, including the Standard Type of compounds and Standard Relation. This standardization enables consistent querying and analysis within the database. The final step is data normalization, where we establish rules to standardize common abbreviations or variants by replacing them with their canonical nomenclature. For instance, the abbreviation 'pip' is replaced with 'phosphatidylinositol' to maintain uniformity throughout the data. The detailed process of kinase data processing is illustrated in Figure 2.

When working with kinase data, it is important to note that our focus should be on kinase target information. Targets typically refer to molecules, proteins, or other biomolecules that are either the object of a drug's action or a key component of a signaling pathway. These targets are crucial in drug discovery as they serve as possible sites of drug action and are the protein molecules that interact with a drug to produce a specific biological effect. Target data consists of information such as substrate name, identifier, phosphorylation site information, phosphorylation specificity information, phosphorylation effect, substrate regulatory mechanism and biological function. The target site information of kinases plays a crucial role in understanding kinase function, cell signaling, and disease mechanisms. In this paper, we have meticulously documented and performed differential data analyses on kinase target information as part of our efforts to construct a kinase database platform. This paper aims to support scientific research and drug development related to kinase.

2.2 Database design and construction

KLSD platform serve as an open database for chemical biology and drug discovery, containing extensive compound and bioactivity data. It encompasses various data tables for the storage of different types of data. Table 1 illustrates some of the common data tables in the database and their key functions, such as assays, activities, compounds, targets, etc. Apart from these prevalent tables, KLSD also includes additional tables to accommodate a wider range of data types. Additional tables, such as the article table exists to contain kinase-related articles for researchers to keep up with progress on kinases, have been added accordingly in the text. The primary objective of KLSD is to provide researchers with open access information resources in the field of chemical biology and drug discovery, aiding their comprehension of the biological activities of kinase compounds. Special emphasis is placed on highlighting target information and target difference data of kinases as depicted in Figure 3.

In this paper, we explore data representation and storage for kinase compounds' molecular structures within the KLSD database. Compounds are primarily represented using SMILES (Simplified Molecular Input Line Entry System) strings and depicted through chemical molecular structure diagrams (Weininger, 1988). In kinase research, converting SMILES to chemical molecular structure

TABLE 1 List and functions of data tables in the KLSD database.

Number	Name	Description
1	assays	Stores information about a bioassay experiment, including the type of experiment, objectives, literature citations, assay results, etc. The records in this table correspond to a bioassay experiment
2	activities	Contains bioactivity data such as bioactivity values for compounds, units, assay conditions, etc. This table is associated with the assays table, which correlates the results of bioassay experiments to compounds and targets
3	compounds	Stores information about the compound, including structure, chemical properties, identifiers, etc. Records in this table correspond to compound entities
4	targets	Contains information about biological targets, such as protein identifiers, names, classes, etc. This table is used to describe targets in drug development
5	docs	Store document information related to bioassay experiments and literature citations, such as literature titles, authors, abstracts, etc.
6	compound_structures	Contains structural information about the compound, such as a description of the two- or three-dimensional structure
7	bioactivities	This is a combined table containing activity data that brings together compounds, bioassay experiments and biological activity data
8	cell_lines	Store information about the cell line, including the source, type, and characteristics of the cell line. This is important for drug screening and bioassay experiments
9	tissue	Contains information about the tissue sample, such as tissue source, tissue type, etc. This is also important for describing the tissue expression of the target.
10	drug_indication	Used to store information about the indication for a drug (the disease or condition that the drug treats)
11	drug_mechanism	Contains mechanistic information about the drug, describing how the drug affects the biological target.

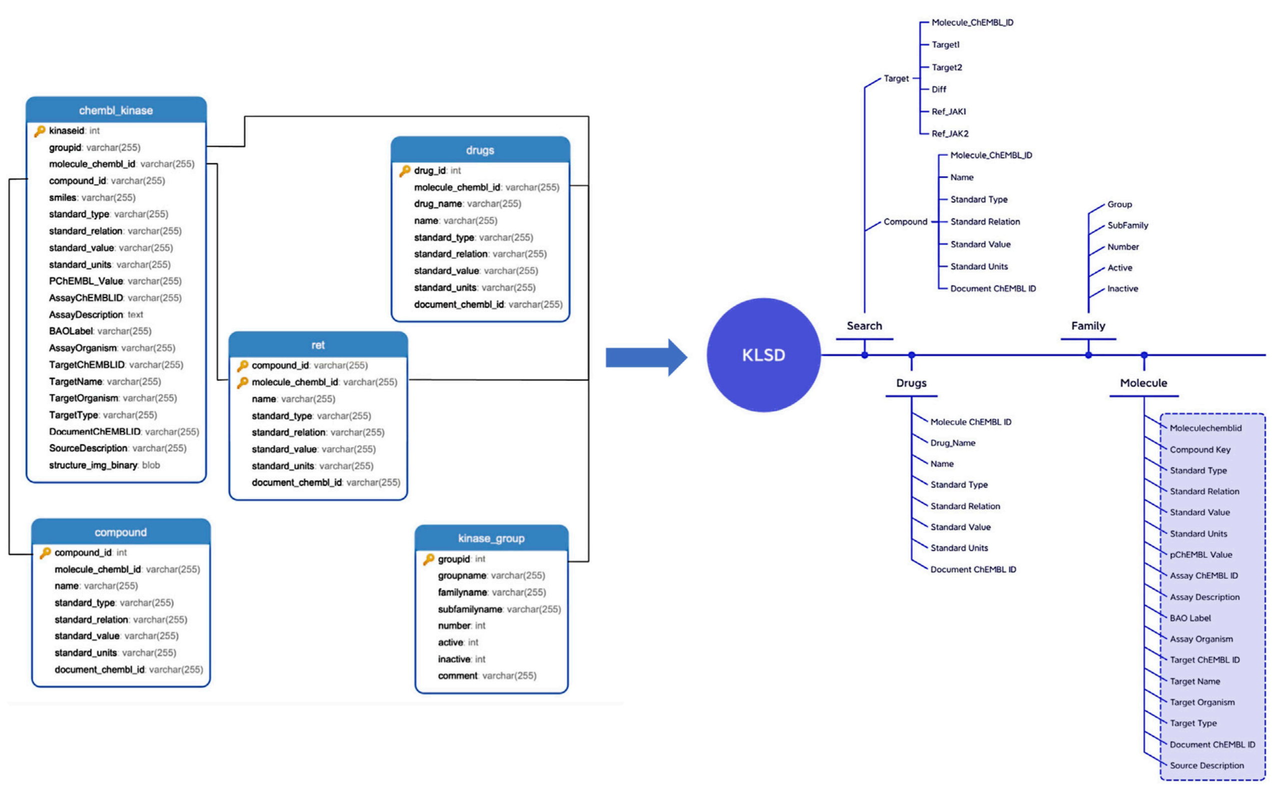
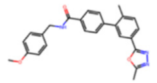
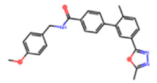


FIGURE 3
Detailed Data Mapping of the KLSD Database Platform. It describes the detailed mapping and the complex interconnections among various data types within the KLSD database platform.

diagrams is vital for understanding the topology, stereo configuration, and functional group positions of kinase molecules. This transformation aids in structure-activity relationship (SAR) analyses, which are critical for identifying structural features that influence kinase inhibitory activity. Thus, effectively translating SMILES into diagrams enhances drug design,

TABLE 2 Chemical structure of kinase ligands: Representative examples.

Number	Name	Description
1	Smiles	COc1ccc (CNC(=O)c2ccc (-c3cc (-c4nnc(C)o4)ccc3C)cc2)cc1
2	Molecular Structure Diagram	
3	Highlight specific parts (oxygen atoms)	
4	logP	logP (calculated by RDKit): 4.959040000000004 logP (calculated by Crippen): 4.959040000000004
5	MACCS keys fingerprint (166 bits)	0000000000000000000000000000000000000000000000000000100001000010010000000000000000101100 000000111110001000000000100100010011001111000000000100011111100111111111111111110

virtual screening, SAR analysis, and the study of kinase-related diseases, playing a key role in kinase research and drug discovery (Li and Fourches, 2021).

In this paper, RDKit (Landrum, 2013) is mainly utilized for converting compound's SMILES into molecular structure maps. RDKit is an open source cheminformatics tool that enables the manipulation of compounds' 2D and 3D molecular structures. It uses machine learning algorithms to generate compound descriptors and fingerprints, calculate compound structural similarities, and facilitate the visualization of 2D and 3D molecules, among other functions.

During the construction of the kinase molecular structure datasheet, the following steps are executed. Firstly, the SMILES for kinase data is obtained by connecting to the MySQL database (DuBois, 2014). Then, an SQL query is executed to extract the SMILES of the kinase data from the database. The query result is stored in a Python3.7 (Lutz, 2013) data structure. The RDKit library is then employed to transform the SMILES string into the chemical molecule structure diagram. By looping through the SMILES data, the Chem. MolFromSmiles () function is utilized to convert each SMILES into its corresponding molecular structure. Lastly, the resulting molecular structure diagrams are stored in binary format in the database. Due to MACCS keys (Taylor, 2024) simplicity and computational efficiency. The 166-bit structure of MACCS keys provides a sufficient level of detail for initial analyses, making them well-suited for broad feature-based screening and rapid processing of large datasets. While Extended Connectivity Fingerprints (ECFP) (Rogers and Hahn, 2010) or the Multilevel Atom Pair fingerprint (MAP4) (Capecchi et al., 2020) offer a higher resolution by capturing more detailed molecular structures and could potentially yield more nuanced insights, their complexity and computational demand are higher. Therefore, we have decided to stick with the simpler MACCS keys. MACCS keys fingerprints and the molecular descriptors are calculated and used to build models for predicting the logP (Partition Coefficient) oil-water partitioning (Wildman and Crippen, 1999). The molecular structure data for a particular kinase chemical is shown in Table 2.

2.3 Platform construction and functional implementation

To efficiently store, analyze and access kinase data, this paper proposes a database service framework based on spring Boot architecture. The front-end utilizes the modern Vue3 JavaScript framework to create a dynamic and high-performance user interface. The back-end utilized the spring Boot rapid development framework to build an efficient, reliable, and easily scalable back-end service. The MySQL relational database is chosen for data storage and management, with Redis caching utilized to improve system performance. The overall architecture follows a layered approach, comprising of an access layer, service layer and storage layer. The access layer primarily handles service invocation and user interaction, focus on tasks such as kinase data querying and kinase target calculation. The service layer is responsible for platform data management, intelligent analysis, and open queries to ensure the stable and effective operation of the service. The

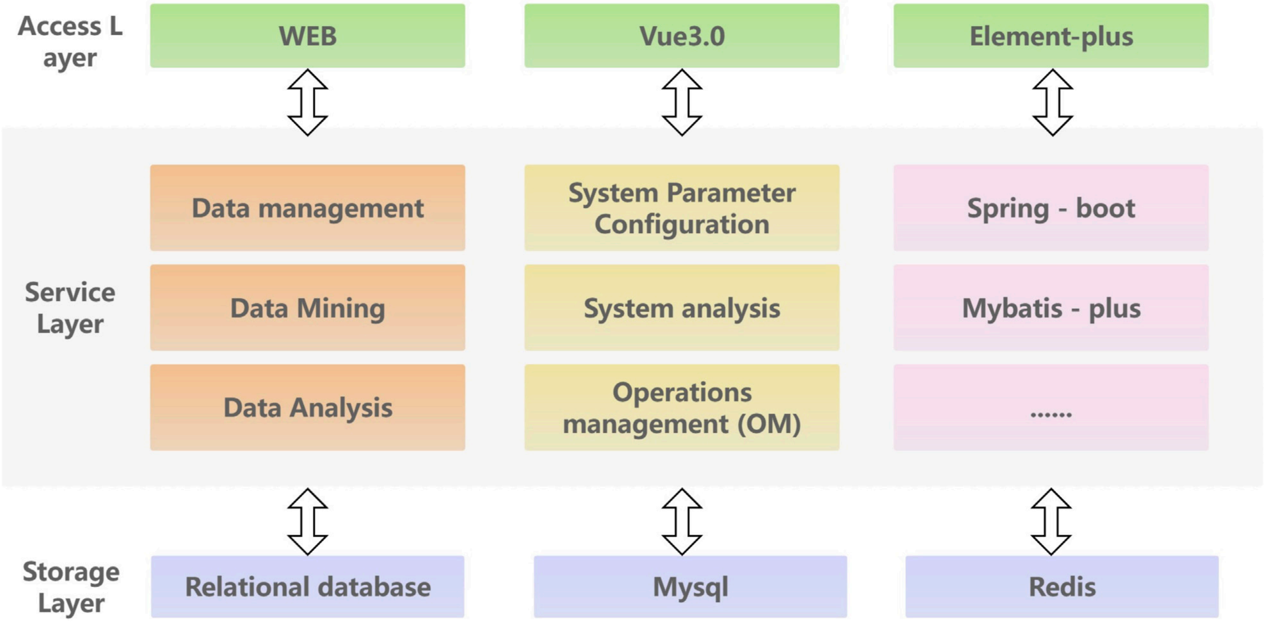


FIGURE 4
Detailed Architecture of the KLSD Database Platform. It describes a comprehensive architectural diagram of the KLSD database, highlighting key technologies and components.

storage layer is responsible for data storage, ensuring high reliability and efficiency. Additionally, an external layer is provided to facilitate access to the data access service. The detailed platform architecture information of KLSD is shown in Figure 4.

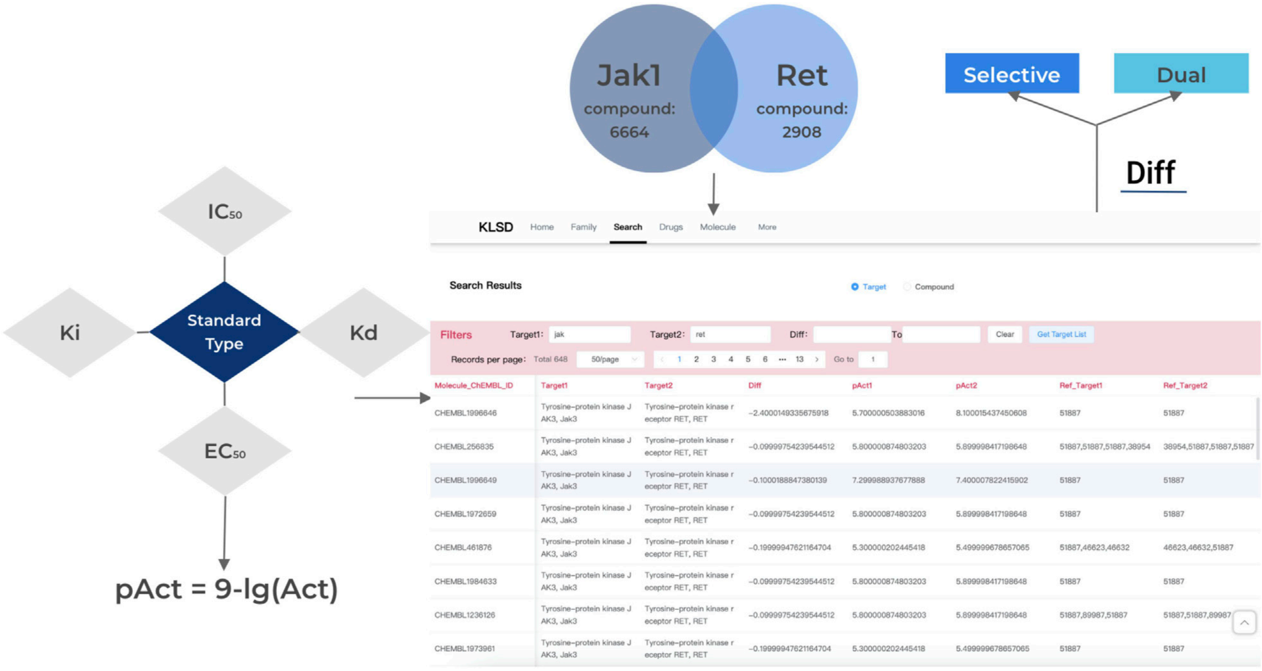
The Target difference analysis of kinase activity is the core function of KLSD. To compare the biological activity of different compounds or molecules in kinase target analysis, pAct values provide a convenient way to do so. pAct values standardize the representation of activity intensity, making comparisons between different activity values easier. This simplifies the interpretation of the data and enables researchers to have a clearer understanding of the relative activity of different molecules or compounds. The pAct value compares activity strengths quantitatively. Generally, higher pAct values indicate stronger biological activity, while lower pAct values may indicate weaker activity. This can be used in assessing the relative potency of kinase inhibitors. Overall, the pAct value is an important metric in kinase target analysis. It simplifies the interpretation and comparison of data and provides strong support for drug design and optimization. pAct values are given in the following general formulae:

$$pAct = 9 - lg(Act)$$

In the KLSD target difference analysis function, the user only needs to input two different targets and pAct value intervals. The platform will automatically retrieve all the relevant data for the two targets from the background and conduct the difference analysis. It will then provide the difference value between the two targets and all the relevant information for the corresponding compounds.

To accomplish this function, the platform will first check the quality of the target-related information by eliminating abnormal values, missing data, or incorrect information. Only compounds with bioactivity_type corresponding to (p)Ki, (p) IC_{50}, (p)Kb, (p)Kd or (p) EC_{50} (represented here as Act) were kept while those with activity_ comments like “not active”, “inactive” were classified into an inactive group, and give an Act value of zero. In this study, our main focus was on analyzing SMKIs among all reported kinase targets. Therefore, we took the initiative to filter the Standard Type of compound activities. In our work, only (p)Ki, (p) IC_{50}, (p)Kb, (p)Kd or (p)EC_{50} were included. Hence, we treat other types as abnormal value. It will then standardize the activity data, specifically those with measurement units of (p)Ki, (p) IC_{50}, (p)Kb, (p)Kd or (p)EC_{50}. These data will be converted into uniform concentration units, such as nanomolar (nM), to calculate the difference value between the targets. The platform will also determine the threshold for kinase activity. For example, compounds with an (p) IC_{50} below 10 nM (López-López et al., 2022) are generally considered to have high activity. The activity data will be categorized into two groups: active and inactive, based on this threshold. The platform will then analyze the correlation between the structural or characteristic differences of the compounds and their activity. Figure 5 displays the homepage of the database platform, the list of kinase data, and the target difference analysis.

KLSD is specifically designed to provide a comprehensive analysis of SMKIs across all reported kinase targets, setting it apart from broader databases. KLSD used data from the ChEMBL database, it has been observed that ChEMBL regularly updates its database, with a frequency of 1-2 updates per year. To ensure synchronization and real-time information, regular monitoring is conducted. Through the development of scripts that can automatically check for updates and download new data from ChEMBL, documentation is updated promptly to reflect changes in the data source. Our web platform has been designed with user accessibility in mind, featuring a user-friendly interface and advanced data visualization tools. These features make it easier for researchers, even those without extensive backgrounds in bioinformatics, to navigate the platform and extract valuable insights effectively.



Molecule_ChEMBL_ID	Target1	Target2	Diff	pAct1	pAct2	Ref_Target1	Ref_Target2
CHEMBL1996646	Tyrosine-protein kinase JAK3, Jak3	Tyrosine-protein kinase receptor RET, RET	−2.4000149335675918	5.700000503883016	8.100015437450608	51887	51887
CHEMBL256835	Tyrosine-protein kinase JAK3, Jak3	Tyrosine-protein kinase receptor RET, RET	−0.09999754239544512	5.800000874803203	5.899998417198648	51887,51887,51887,38954	38954,51887,51887,51887
CHEMBL1996649	Tyrosine-protein kinase JAK3, Jak3	Tyrosine-protein kinase receptor RET, RET	−0.1000188847380139	7.299988937677888	7.400007822415902	51887	51887
CHEMBL1972659	Tyrosine-protein kinase JAK3, Jak3	Tyrosine-protein kinase receptor RET, RET	−0.09999754239544512	5.800000874803203	5.899998417198648	51887	51887
CHEMBL461876	Tyrosine-protein kinase JAK3, Jak3	Tyrosine-protein kinase receptor RET, RET	−0.19999947621164704	5.300000202445418	5.499999678657065	51887,46623,46632	46623,46632,51887
CHEMBL1984633	Tyrosine-protein kinase JAK3, Jak3	Tyrosine-protein kinase receptor RET, RET	−0.09999754239544512	5.800000874803203	5.899998417198648	51887	51887
CHEMBL1236126	Tyrosine-protein kinase JAK3, Jak3	Tyrosine-protein kinase receptor RET, RET	−0.09999754239544512	5.800000874803203	5.899998417198648	51887,89987,51887	51887,51887,89987
CHEMBL1973961	Tyrosine-protein kinase JAK3, Jak3	Tyrosine-protein kinase receptor RET, RET	−0.19999947621164704	5.300000202445418	5.499999678657065	51887	51887

FIGURE 5
Comparative Differential Analysis of Kinase Targets. It describes the analytical methods used to compare kinase targets, highlighting aspects of variation and the conditions under which the analyses are performed.

TABLE 3 Detailed statistics of kinase family compounds.

Number	Family	Compound	Active	Inactive	Active (%)	Inactive (%)
1	NKF2	89762	4779	84983	5	95
2	MAPK	38070	21149	16921	56	44
3	Trbl	31060	7220	23840	23	77
4	PLK	27267	3214	24053	12	88
5	PKC	22872	14743	8129	64	36
6	CAMKL	19042	7103	11939	37	63
7	TBCK	18900	14757	4143	78	22
8	CDK	18620	15080	3540	81	19
9	PSK	16480	13061	3419	79	21
10	VEGFR	15303	13012	2291	85	15
11	EGFR	14112	10783	3329	76	24
12	Abl	14010	8666	5344	62	38
13	RSKR	12334	11458	876	93	7
14	PDGFR	10283	8648	1635	84	16
15	Jak2	9846	8637	1209	88	12

3 Results and discussion

3.1 Statistics and analysis of data

Based on the actual research and the drug development process of kinases, we obtained a total of 138 families of SMKIs, Active and Inactive by analysing the most researched kinase families. Table 3 shows the top 15 kinase families in terms of the number of compounds.

After analyzing the data on the compounds of kinase families, we employed Vue 3 and D3. js (Bostock et al., 2011) to create a radial tree diagram for visualizing kinase families data. Vue 3, a modern JavaScript framework, offers a reactive and composable architecture, making it ideal

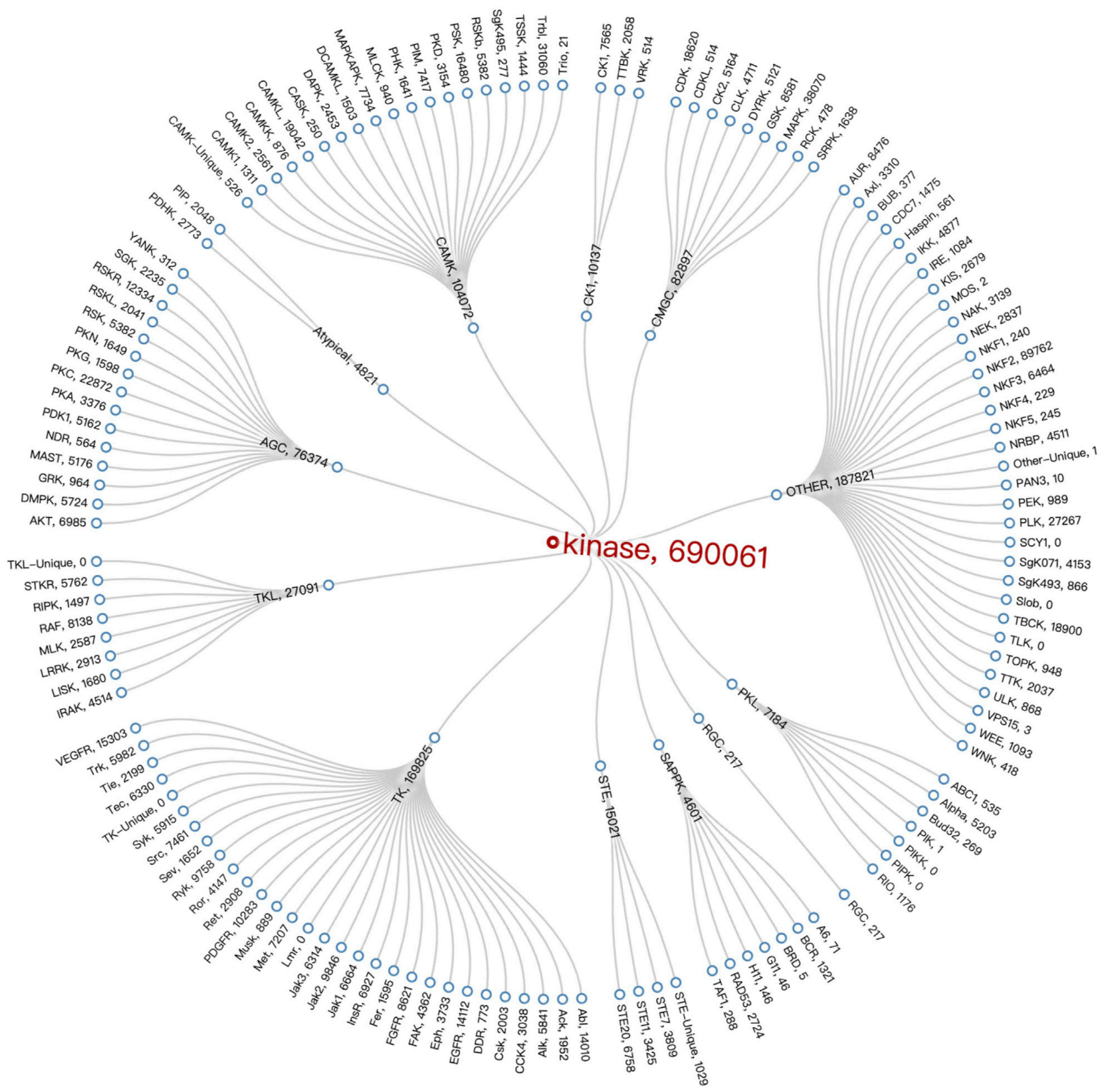


FIGURE 6
Kinase Family Radial Tree diagram. It employs specific technologies (Vue 3 and D3.js) that embody the interactive and data-driven nature of visualisation to represent the structure and number of compounds within the kinase family.

for developing interactive web applications. D3.js complements this by providing robust data-driven techniques for creating and manipulating graphical elements, which we used to calculate and render the complex structures of our radial tree diagram. It was observed that the TK Group had the highest number of compounds at 169,825. When looking at specific Families, the NKF2 Family, belonging to the OTHER Group, stood out with a number of compounds at 89,762, followed by the MAPK Family in the OTHER Group, which had a number of compounds at 38,070. The Radial Tree diagram illustrating the relationships between kinase families is displayed in Figure 6.

Upon a thorough analysis of the activity profiles of compounds across different kinase families, we observed that the MAPK family within the CMGC group demonstrated the most significant number of active compounds. Specifically, while the MAPK family encompasses a total of 38,070 compounds, the number of active compounds identified was 21,149. This was followed by the CDK Family in the CMGC Group with an activity of 15,080. Furthermore, it is worth noting that the NKF2 Family had the highest inactivity of 84,983, and this was followed by the PLK Family in the OTHER Group with an inactivity value of 24,053. The heat map illustrating the activity of kinase families, known as Compound_Activity, is displayed in Figure 7. Figure 7 serves as an overall heat map that showcases the top six kinase families based on the number of compounds linked to each family.

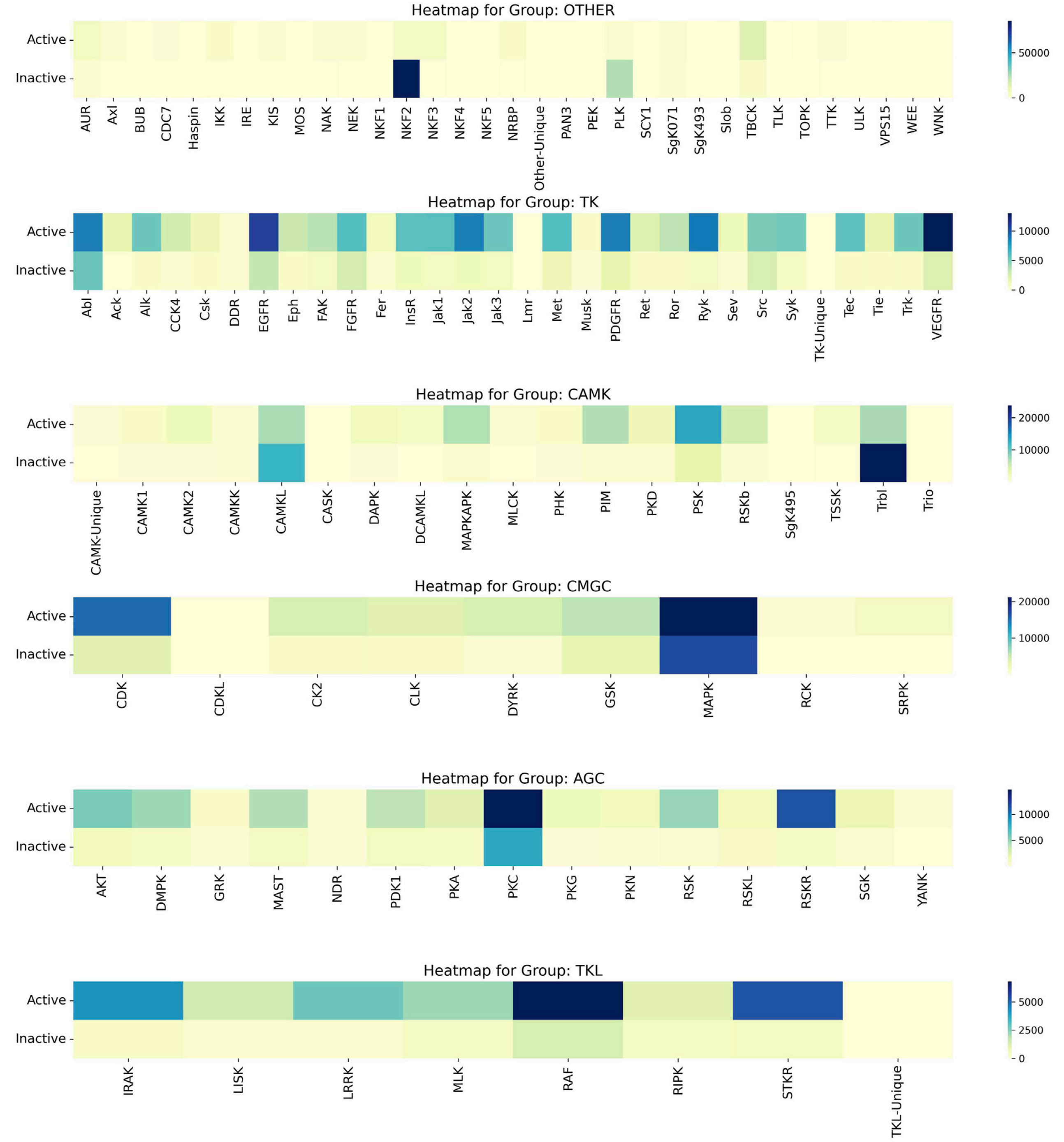


FIGURE 7
Heatmap of Kinase Families Activity. It uses heat maps to visually analyse the activity of the first 6 kinase families, reflecting the number of active and inactive. The varying color gradients in the heatmap signify different ranges in the number of compounds, with darker shades indicating higher numbers and lighter shades indicating lower numbers. Each color on the heatmap corresponds to specific range of compound numbers, as indicated on the right side.

The term "Standard Type" pertains to the specific metric employed to quantify the activity of compounds within the scope of pharmacological data. Common measurements under this category include (p)IC$_{50}$, (p)EC$_{50,}$ (p)Kd, and (p)Ki, among others. These parameters are pivotal as they denote the potency of a compound in its interaction with a biological target, be it through inhibition or activation. Understanding these measurements is fundamental to grasping the pharmacodynamics of the compounds under discussion. Furthermore, the 'BAO label', which stands for BioAssay Ontology label, provides a descriptor for the biological or pharmacological context within which an assay is conducted. This label encompasses classifications that can specify the nature of the target, such as a single protein, a protein complex, or a subcellular structure. The BAO label is indispensable for categorizing assays according to their biological specificity, thereby facilitating a clear interpretation of the experimental design and the resultant data.

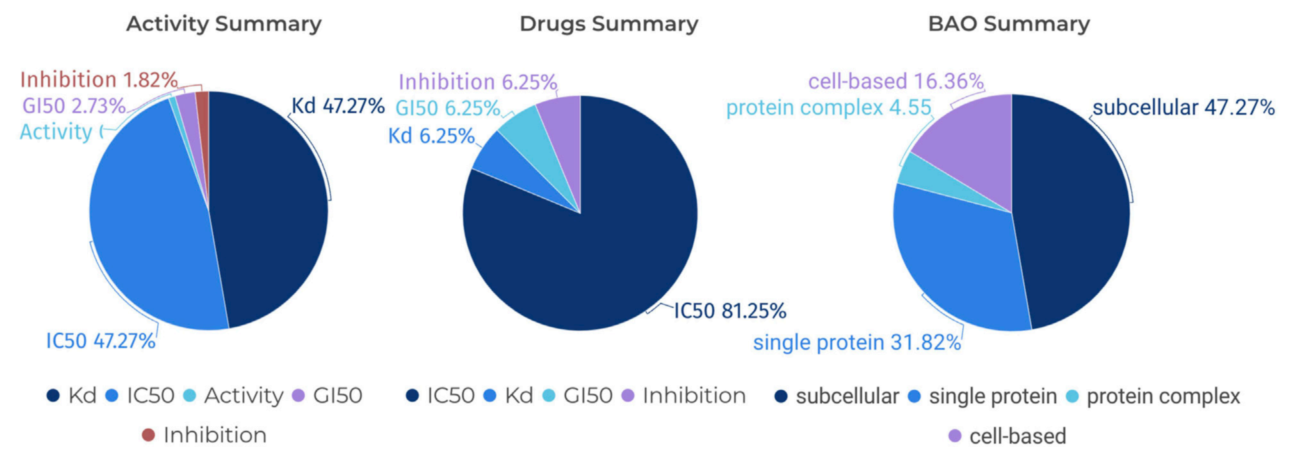


FIGURE 8
Pie Chart Analysis of Various Assays for a Specific Compound. It analyses a specific compound (CHEMBL3301607) in detail from three different perspectives: Standard Type for Activity, Standard Type for Drugs and BAO Label for BAO.

In the field of kinase research, the study of JAK1 (Janus Kinase 1) is crucial as JAK1 plays a significant role in signaling pathways that impact cellular processes, including immune response and inflammation (Harris and Cummings, 2021). FILGOTINIB, a selective inhibitor of JAK1, is proven to be highly effective in the treatment of rheumatoid arthritis (Feagan et al., 2021). Therefore, we have used CHEMBL3301607, which corresponds to Filgotinib, as an example in Figure 8. Specifically, we can search for "Filgotinib" in the "Drugs" section of the KLSD database, the platform automatically retrieves the corresponding compound name, Standard Type, and BAO Label among other information. These pieces of information are crucial for understanding the mechanisms of action of drugs, evaluating drug efficacy, and optimizing drug design. As shown in Figure 8, we analysed CHEMBL3301607 from three perspectives: the standard type, the standard type of the drug and the BAO label of the BAO. This approach not only enhances our understanding of Filgotinib's role as a JAK1 inhibitor but also demonstrates the KLSD database's utility in facilitating detailed drug analysis and research involving SMKIs.

3.2 Conclusion

Currently, the kinase database platform has accumulated a substantial amount of kinase-related data, including information on kinase structure, function, regulation, and biological roles. KLSD can provide personalized data query, analysis, and display tailored to researchers' specific needs. By analyzing the kinase data, KLSD seeks to identify patterns and features, offering valuable insights for research and improving researchers' efficiency. Furthermore, KLSD ensures data quality and accuracy, making it highly beneficial in fields such as biomedical research and drug discovery. However, there is still ample room for growth in kinase deep learning and artificial intelligence. We must leverage artificial intelligence technology to further analyze large-scale kinase data, uncover potential associations and patterns, predict kinase functions and biological effects, and broaden the functionality and application areas of the kinase database platform. These advancements will enable users to conduct more comprehensive studies and analyses beyond the current capabilities of the platform, meeting the evolving needs of users and providing greater support and resources for kinase research.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

YY: Conceptualization, Data curation, Software, Visualization, Writing–original draft. XT: Validation, Writing–review and editing. HL: Writing–review and editing. XL: Methodology, Writing–review and editing. CL: Writing–review and editing. YS: Writing–review and editing. SS: Conceptualization, Formal Analysis, Validation, Writing–review and editing. YY: Conceptualization, Funding acquisition, Writing–review and editing. ZZ: Validation, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by National Natural Science Foundation of China (82103985) and 2023 Jiangsu Province Graduate Research Innovation Program (KYCX23_2074).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Attwood, M. M., Fabbro, D., Sokolov, A. V., Knapp, S., and Schiöth, H. B. (2021). Trends in kinase drug discovery: targets, indications and inhibitor design. *Nat. Rev. Drug Discov.* 20 (11), 839–861. doi:10.1038/s41573-021-00252-y

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3: data-driven documents. *IEEE Trans. Vis. Comput. Graph.* 17 (12), 2301–2309. doi:10.1109/TVCG.2011.185

Capecchi, A., Probst, D., and Reymond, J. L. (2020). One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. cheminformatics* 12, 43–15. doi:10.1186/s13321-020-00445-4

Charles, S., Abhinav, V., Nathan, H., et al. (2018). ChemNet: a transferable and generalizable deep neural network for small-molecule property prediction. *Abstr. Pap. Am. Chem. Soc.*, 255. doi:10.48550/arXiv.1712.02734

Cohen, P., Cross, D., and Jänne, P. A. (2021). Kinase drug discovery 20 years after imatinib: progress and future directions. *Nat. Rev. Drug Discov.* 20 (7), 551–569. Epub 2021 May 17. PMID: 34002056; PMCID: PMC8127496. doi:10.1038/s41573-021-00195-4

Cortés-Ciriano, I., and Bender, A. (2019). KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. *J. Cheminformatics* 11 (1), 41–16. doi:10.1186/s13321-019-0364-5

DuBois, P. (2014) *MySQL cookbook: solutions for database developers and administrators*. Sebastopol, California, United States: O'Reilly Media, Inc.

Duvenaud, K. D., Maclaurin, D., Aguilera-Iparraguirre, J., et al. (2015). Convolutional networks on graphs for learning molecular fingerprints. *CoRR*, 09292. abs/1509. doi:10.48550/arXiv.1509.09292

Feagan, B. G., Danese, S., Loftus, E. V., Jr, Vermeire, S., Schreiber, S., Ritter, T., et al. (2021). Filgotinib as induction and maintenance therapy for ulcerative colitis (SELECTION): a phase 2b/3 double-blind, randomised, placebo-controlled trial. *Lancet* 397 (10292), 2372–2384. Epub 2021 Jun 3. PMID: 34090625. doi:10.1016/S0140-6736(21)00666-8

Fischer, P. M. (2016). Approved and experimental small-molecule oncology kinase inhibitor drugs: a mid-2016 overview. *Med. Res. Rev.* 37 (2), 314–367. doi:10.1002/med.21409

Han, J., Pei, J., and Tong, H. (2022) *Data mining: concepts and techniques*. Burlington, Massachusetts, United States: Morgan kaufmann.

Harris, C., and Cummings, J. R. F. (2021). JAK1 inhibition and inflammatory bowel disease. *Rheumatol. Oxf.* 60 (Suppl. 2), ii45–ii51. doi:10.1093/rheumatology/keaa896

Jan, W., Hans, M., and Friedemann, S. (2019). Predictive multitask deep neural network models for ADME-tox properties: learning from large data sets. *J. Chem. Inf. Model.* 59 (3), 1253–1268. doi:10.1021/acs.jcim.8b00785

Kanev, G. K., de Graaf, C., Westerman, B. A., de Esch, I. J. P., and Kooistra, A. J. (2021). KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Res.* 49 (D1), D562–D569. doi:10.1093/nar/gkaa895

Karpov, P., Godin, G., and Tetko, V. I. (2020). Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J. Cheminformatics* 12 (1), 17–554. doi:10.1186/s13321-020-00423-w

Kevin, Y., Kyle, S., Wengong, J., Coley, C., Eiden, P., Gao, H., et al. (2019). Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* 59 (8), 3370–3388. doi:10.1021/acs.jcim.9b00237

Kumari, P., Abhigyan, N., and Radha, C. (2015). Identification of human drug targets using machine-learning algorithms. *Comput. Biol. Med.* 56, 175–181. doi:10.1016/j.compbiomed.2014.11.008

Landrum, G. (2013). Rdkit documentation. *Release* 1 (1-79), 4.

Li, X., and Fourches, D. (2021). SMILES pair encoding: a data-driven substructure tokenization algorithm for deep learning. *J. Chem. Inf. Model.* 61 (4), 1560–1569. doi:10.1021/acs.jcim.0c01127

López-López, E., Fernández-de Gortari, E., and Medina-Franco, J. L. (2022). Yes SIR! On the structure-inactivity relationships in drug discovery. *Drug Discov. Today* 27 (8), 2353–2362. Epub 2022 May 11. PMID: 35561964. doi:10.1016/j.drudis.2022.05.005

Lutz, M. (2013) *Learning python: powerful object-oriented programming*. Sebastopol, California, United States: O'Reilly Media, Inc.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science* 298 (5600), 1912–1934. PMID: 12471243. doi:10.1126/science.1075762

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., et al. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* 47 (D1), D930–D940. PMID: 30398643; PMCID: PMC6323927. doi:10.1093/nar/gky1075

Mengying, S., Sendong, Z., Coryandar, G., Elemento, O., Zhou, J., and Wang, F. (2020). Graph convolutional networks for computational drug development and discovery. *Briefings Bioinforma.* 21 (3), 919–935. doi:10.1093/bib/bbz042

Mukhtar, M., Zheng, H., Yafen, Li, et al. (2017). Study on the *in vitro* anti-fine-grained Echinococcus granulosus protocephalicus effect of the rapamycin-targeted protein inhibitor Deforlimus. *Chin. J. Pathogenic Biol.* 12 (12), 1157–1160+1164. doi:10.13350/j.cjpb.171206

Nathan, H., Charles, S., Abhinav, V., et al. (2018). SMILES2vec: an interpretable general-purpose deep neural network for predicting chemical properties. *Abstr. Pap. Am. Chem. Soc.*, 256. doi:10.48550/arXiv.1712.02034

Popova, M., Isayev, O., and Tropsha, A. (2018). Deep reinforcement learning for *de novo* drug design. *Sci. Adv.* 4 (7), eaap7885. doi:10.1126/sciadv.aap7885

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50 (5), 742–754. doi:10.1021/ci100050t

Taylor, K. (2024). MACCS key definition. Available at: https://list.indiana.edu/sympa/arc/chminf-l/2007-11/msg00058.html (Accessed April 25, 2024).

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28 (1), 31–36. doi:10.1021/ci00057a005

Wildman, S. A., and Crippen, G. M. (1999). Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* 39, 868–873. doi:10.1021/ci990307l

Wu, P., Nielsen, T. E., and Clausen, M. H. (2015). FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol. ences* 36 (7), 422–439. doi:10.1016/j.tips.2015.04.005

Xiang, W. S., Xian, Z., Feng, Z., Wang, Y. l., Qin, C., Tan, Y., et al. (2021). Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat. Mach. Intell.* 3 (4), 334–343. doi:10.1038/s42256-021-00301-6