MedChemLens: An Interactive Visual Tool to Support Direction Selection in Interdisciplinary Experimental Research of Medicinal Chemistry

Chuhan Shi, Fei Nie, Yicheng Hu*, Yige Xu*, Lei Chen, Xiaojuan Ma and Qiong Luo

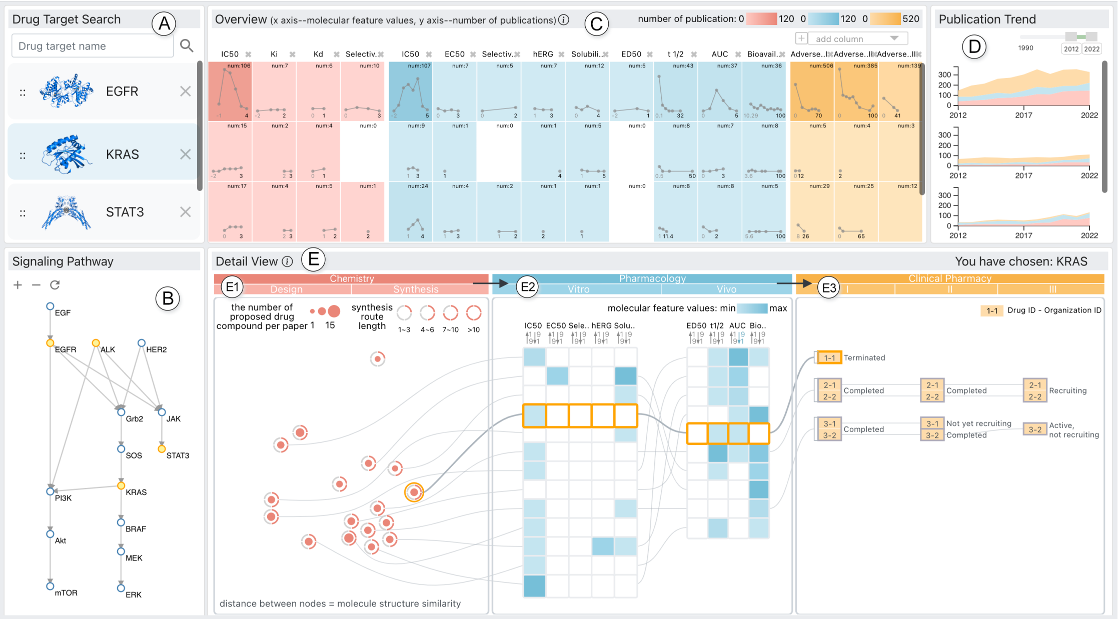


Fig. 1. MedChemLens: (A) The Drug Target Search view allows users to search drug targets by name. (B) The Signaling Pathway view presents the signaling pathways of the targets under search. (C) The Overview shows the overall distributions of the existing drug compound research. (D) The Publication Trend view displays the number of publications over time related to the targets under search. (E) The Detail View consists of (E1) the Chemistry panel, which summarizes the drug compounds proposed in chemical publications, (E2) the Pharmacology panel, which displays the molecular feature values of the drug compounds tested in *in vitro* and *in vivo* pharmacological assays, and (E3) the Clinical Pharmacy panel, which visualizes the clinical trial progress of the drug compounds.

Abstract— Interdisciplinary experimental science (e.g., medicinal chemistry) refers to the disciplines that integrate knowledge from different scientific backgrounds and involve experiments in the research process. Deciding "in what direction to proceed" is critical for the success of the research in such disciplines, since the time, money, and resource costs of the subsequent research steps depend largely on this decision. However, such a direction identification task is challenging in that researchers need to integrate information from large-scale, heterogeneous materials from all associated disciplines and summarize the related publications of which the core contributions are often showcased in diverse formats. The task also requires researchers to estimate the feasibility and potential in future experiments in the selected directions. In this work, we selected medicinal chemistry as a case and presented an interactive visual tool, MedChemLens, to assist medicinal chemists in choosing their intended directions of research. This task is also known as drug target (i.e., disease-linked proteins) selection. Given a candidate target name, MedChemLens automatically extracts the molecular features of drug compounds from chemical papers and clinical trial records, organizes them based on the drug structures, and interactively visualizes factors concerning subsequent experiments. We evaluated MedChemLens through a within-subjects study (N=16). Compared with the control condition (i.e., unrestricted online search without using our tool), participants who only used MedChemLens reported faster search, better-informed selections, higher confidence in their selections, and lower cognitive load.

Index Terms—Interdisciplinary experimental science, interactive visual analysis, scientific literature data

1 Introduction

Interdisciplinary experimental science (e.g., medicinal chemistry, biophysics, and biomedicine) refers to branches of knowledge that integrate the data, methods, and theories from different scientific backgrounds and employ experiments as the key research approach [29, 52]. Deciding "what research direction will be feasible and promising" is usually the first step of the entire research process in such disciplines [1]. Given that the research process of interdisciplinary experimental science is often time-consuming and resource intensive, identification

- *C. Shi, F. Nie, Y. Hu, L. Chen, X. Ma and Q. Luo are with the Hong Kong University of Science and Technology. E-mail: {cshiag, fnie, yhubf}@connect.ust.hk and {leichen, mxj, luo}@cse.ust.hk*
- *Y. Xu is with Nanyang Technological University. E-mail: yige002@e.ntu.edu.sg*
- **Both authors contributed equally to this research.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

of the right starting point in the initial stage of scientific inquiry can increase the chance of ultimate success. For instance, in medicinal chemistry, it may take more than 10 years from the identification of related chemical entities to marketed drugs [57], and cost at least a billion dollars [46]. If medicinal chemists choose to design drugs that interact with a human protein incapable of being modulated by any biological therapy, their design is likely to fail in the experimentation stage [14], leading to a substantial waste of time and money.

Despite its importance, research direction selection in interdisciplinary experimental science is never an easy task. First, researchers often need to comprehensively integrate and make sense of large-scale, heterogeneous data from all related disciplines to evaluate candidate directions from both theoretical and practical perspectives [34]. They want to establish solid knowledge grounds for later hypothesis-driven experimentation, learn from lessons of prior research to minimize risks, and assess their competitiveness down the chosen path. However, existing scientific literature analytic tools often focus on document organization [11, 19, 25, 61], retrieval [5, 10, 18, 28], and discovery [6, 7, 23, 40]. Few works support the decision-making in research process and help balance the considerations from both science and strategy aspects. Second, existing methods for extracting and organizing data from scholarly documentations (e.g., publications, lab reports, etc.) may not adequately meet the data integration needs of researchers in interdisciplinary experimental science fields. For example, medicinal chemists commonly organize literature by the chemical structures of drug compounds proposed in the papers [3], while documents are conventionally grouped and indexed by keywords [60, 63], author network [36, 62], and citation links [5, 27]. Third, it is challenging for researchers in interdisciplinary experimental science to estimate the feasibility and difficulties of future experimental testing, a critical research component, stemming from their decisions. When inspecting a candidate research direction, individual researchers or research groups may have different concerns, such as personal skills and laboratory resources available for experiments.

In this work, we selected medicinal chemistry as a case to demonstrate how an interactive scientific literature analytic system can help address the aforementioned challenges in the research direction selection in interdisciplinary experimental science. Medicinal chemistry is a scientific discipline at the intersection of chemistry, pharmacology, and clinical pharmacy [22]. In medicinal chemistry, for a specific disease, researchers need to select a particular drug target (i.e., disease-linked proteins in the human body that are agents being modulated by drugs to produce therapeutic effects) from a pool of candidates, and then design and test out drug compounds against the chosen target [30, 32]. We proposed an interactive visual tool called MedChemLens to assist medicinal chemists in the identification of a drug target that is most likely to lead to a promising research path of subsequent drug design. MedChemLens integrates and visualizes relevant literature and data from three related disciplines: chemistry, pharmacology, and clinical pharmacy. It retrieves drug compounds associated with the given drug target candidates that have been reported in scholarly publications and extracts the key molecular features of these compounds from the text, images, and tables of the returned documents. With these data, it enables the organization of the related papers by similarities in the chemical structures of the drug compounds in connection to each candidate target. Moreover, MedChemLens facilitates the exploration of potential research paths following different drug targets to help users evaluate the practicality and potential risks of the chemical experiments in future research processes. A within-subjects user study with 16 medicinal chemistry researchers of various levels of expertise provided support for the usefulness and effectiveness of our system.

In summary, our major contributions are:

- MedChemLens, an interactive visual tool to support medicinal chemists to evaluate possible research directions by analyzing and comparing relevant literature and experimental data.

- A within-subjects user study that demonstrates the effectiveness of our approach in helping users select research directions in the interdisciplinary experimental research of medicinal chemistry.

2 RELATED WORK

2.1 Visual Analysis of Scientific Literature

Plenty of research has been conducted to provide interactive visualization to support the exploration and analysis of scientific literature. They mainly focus on assisting users in searching, organizing, and retrieving their desired research papers from broad sources. For example, Benito-Santos *et al.* [6] presented GlassViz that helps researchers explore a large document corpus by visualizing the entry points. Costagliola *et al.* [13] proposed a 3D analytical interface, *CyBiS*, that shows document items as spheres embedded in a 3D cylinder and supports operations such as rotate to refine search. To organize and reveal the relationships between documents, Zhao *et al.* [62] proposed PivotSlice which applies both node-link view and customized dynamic tabular view to represent the relationships across literature data items. Wang *et al.* [54] developed TopicPanorama which combines a radial icicle plot and a density-based graph to show a full picture of relevant topics from multiple sources. Moreover, some existing systems were proposed to support retrieving users' desired information from documents. For example, Beck *et al.* [5] presented SurVis which contains a word-sized sparkline enabling users to conduct textual search on details such as keywords, meta-information, and relationships. However, the visual analysis of scientific literature in our scenario is more complicated since we need to not only help researchers browse and explore related literature to establish solid knowledge backgrounds but also support the trade-off of risks and output of research paths to help researchers in decisions-making process. In addition, future experimental testing might affect researchers' research direction selection, yet few existing works have managed to help estimate its feasibility and difficulties.

In the existing visual scientific literature analytic tools, the publications are mainly organized and indexed by citation links, keywords, and author network. For instance, Burger *et al.* [7] applied citation contexts to develop a word-document 2D projection in their proposed visualization scheme cite2vec. Dattolo *et al.* [15] presented VisualBib which groups papers based on the corresponding bibliographies. Elmqvist *et al.* [20] comprehensively applied keywords, co-authorship, and citation to organize publications in their system CiteWiz. However, the information that researchers in interdisciplinary experimental fields are interested in goes beyond these data types. For example, it is a common practice in chemistry to organize publications by the images of molecular structures. While there is existing work emphasizing the importance of the images in publications (e.g., Chen *et al.* [9] proposed a dataset called VIS30K that represents visualization papers with figures), they still built the relationships between publications based on conventional data types, which cannot satisfy the needs in chemistry.

2.2 Visualization for Drug Target Selection

Previous works have explored different visualization methods to assist medicinal chemists in drug target selection. These methods were mainly used to present data within a specific discipline or across disciplines in the drug discovery process. For the visual representation of data within one area, node-edge network is widely applied in existing tools. For example, Promiscuous [50], Dinies [58], and TargetNet [59] were all web-based services that use node-edge networks to represent drug-target interactions. Furthermore, multimodal visualization interfaces are utilized for multifaceted data. For example, Open Targets Platform [33] integrated pathway overview maps and hierarchical networks to present drug-disease associations. Pharos [39] assigned different diagrams based on the data type, such as radial pie chart for categorical data and word cloud for textual data, to help aggregate information about drug targets from diverse resources. However, these tools only focus on a single area and fail to connect the disciplines involved in the drug discovery process. Some visual tools have been proposed to display data across disciplines. For example, ChEMBL [38] integrated medicinal chemistry data with pharmaceutical knowledge by creating a sunburst view to show the drug target classification and a heatmap view to show molecular bioactivity. However, such systems only provided a broad view of drug targets but lacked summaries of relevant research which would benefit medicinal chemists' future drug design.

3 Design Process

Our goal is to support an in-depth and systematic exploration of literature and experimental data to help medicinal chemists decide potential directions (i.e., drug target selection) in their interdisciplinary experimental research. Our design process started with a two-hour semi-structured interview with each of six researchers (E1-E6) in the field of medicinal chemistry to understand the current practices and challenges of drug target selection. E1 is a university professor with 9 years of research experience; E2 and E3 are postdocs in key drug discovery laboratory with 9 years of and 7 years of experience, respectively; E4 and E5 are senior PhD students (4 years of experience); and E6 is a junior (1-year) PhD student. Based on their feedback, we derived a set of design requirements which guided our initial system design. In the later stages, we carried out bi-weekly meetings with these six researchers for three months to iteratively update the system design to ensure that our implementation addresses the requirements.

3.1 Factors Related to Drug Target Selection

Based on the interview results, we summarized the factors considered by medicinal chemists in the drug target selection process.

1) Drug discovery process is a multifaceted process including the research of chemistry, pharmacology, and clinical pharmacy. E2-E5 reflected that medicinal chemists often integrate the knowledge of these related disciplines and evaluate the current research progress of potential drug targets as reported in the literature and experimental reports along this process. Usually, after deciding to proceed with a particular drug target, chemists are mainly concerned with structure activity designs and chemical synthesis of potential drug compounds that could interact with the selected target [30]. Then the drug compounds are progressed to the next step for pharmacological testing, which includes *in vitro* tests (i.e., experiments conducted on microorganisms or cells outside of a living organism [43]) and *in vivo* tests (i.e., experiments conducted in a living organism, such as animal models [43]). Subsequently, drugs go through clinical trials with human subjects [43]. The clinical pharmacy research consists of three phases – *phase I*, *phase II*, and *phase III*, during which the adverse effects of drugs are tested.

2) Drug target properties consist of the drug target structure, druggability, and signaling pathway. E3, E4 and E6 suggested that medicinal chemists are interested in these properties because these properties can facilitate them to initially filter targets. *Drug target structure* is the protein structure of the drug target. *Druggability* is the likelihood of the drug target being able to be modulated by a drug [41]. To evaluate the druggability, chemists need to know the research progress of existing drug compounds for each candidate target. A *Signaling pathway* shows a chain of proteins activated by drug compounds binding with drug targets [42]. Upstream proteins undergo biochemical reactions and transmit signals to downstream proteins until therapeutic effects are produced. A signaling pathway often contains several drug targets.

3) Molecular features of drug compounds often serve as the basis for evaluating the proposed drug compounds in terms of their viability as potential new drugs and are of major interest to medicinal chemists. All researchers we interviewed stated that they always spent the majority of their time searching and reading related research publications to know the existing drug compounds against each candidate drug target. The molecular features they care about are tested in different drug discovery stages and thus may appear in different sections of a paper in a wide variety of forms. We describe the details of them in Table 1.

3.2 Design Requirement

Based on the analysis of the medicinal chemists' feedback, we summarized six design requirements for our system design.

R1. Enable intuitive comparison of different targets on different scales. The system should support the comparison of candidate drug targets in different aspects, including the target properties, the research trend and popularity of targets over time, and the individual molecular feature of interest. For example, E3 said that medicinal chemists can directly filter out the candidates that do not satisfy their requirements for research directions by having an overview about the volume and stage of the related research. For the remaining candidate targets, researchers need to check detailed drug compound research information and progress to make their final decisions.

R2. Provide a comprehensive picture of the research about each candidate drug target in three relevant disciplines. The system should provide an overarching summary of the drug compound research about each drug target. As mentioned by E1 and E3, each publication has a research goal of designing new drug compounds to enhance certain molecular feature(s). Researchers want to know the number and the overall distribution of the drug compound research focusing on each molecular feature. In addition, E2, E4 and E5 mentioned that as the same drug compound should be studied by different disciplines in different drug discovery stages, the related scholarly documentations are scattered in large-scale online resources from various fields. Since integrating research about the same drug compound manually is difficult, research data on the same drug should be connected across disciplines and following the drug discovery process to facilitate medicinal chemists to streamline the literature survey process and track the development of each drug compound.

R3. Organize scholarly documentations following the practice of each individual discipline. Researchers require an organization and presentation of the scholarly documentations to show the corresponding research landscape and status in each individual discipline. All researchers pointed out that the chemical structures of the drug compounds are the core findings of medicinal chemical publications, and it is a common practice for medicinal chemists to organize literature based on chemical structures. In pharmacology, researcher focus on the values of molecular features tested in pharmacological assays. For clinical pharmacy, they want to know the status information for clinical trials, such as *"how many organizations are conducting clinical trials"* (E1) and *"why some clinical trials were terminated"* (E4). Visual designs should be adapted for different data types to help users process and digest the heterogeneous research data in different disciplines.

R4. Inspire drug target selection process and future drug compound design. Researchers demand inspiration for the drug target selection process. E4 and E5 said that the system should help inspect signaling pathways to find the connection between the candidate drug targets and remind users of previously overlooked targets that could be candidates. E2 and E6 added the importance of knowing the shortcomings of existing drug compounds and possible improvements so that users can get ideas for future research paths and drug design.

R5. Support estimation of the feasibility and difficulty of future practical experiments. The system should facilitate medicinal chemists to assess the feasibility and challenges of practical implementation when engaging in the medicinal chemical research related to the candidate drug targets. E1 and E4 hope the system can show the synthesis route lengths of the previous drug compounds against the candidate drug targets as indicators of synthesis difficulty and display what kinds of chemical structures can be advanced better in the course of drug discovery. Based on the molecular similarity principle [3] (i.e., two structurally similar molecules often have similar properties and can be analyzed using similar testing models), empirical information of existing studies can provide implications for the experimental feasibility of the new drug compounds designed by medicinal chemists.

R6. Facilitate an interactive and customized data exploration. We observed that individual medicinal chemical researchers or research groups may have different focuses, information needs, and exploration patterns. E2 and E6 added that their research interests, abilities, and available laboratory facilities may also vary. Hence, the system should enable users to customize their preferences on different evaluation metrics and decision-making patterns interactively.

4 MedChemLens

We presented a visual analytic system MedChemLens to aid medicinal chemists in exploring literature and experimental data to select drug targets. MedChemLens incorporates the following five views. The Drug Target Search view (Fig.1 A) allows users to search drug targets of interest and inspect their 2D structures (**R1**). The Signaling Pathway view (Fig.1 B) visualizes the interactions between the drug targets under search and prompts users for other possibly overlooked targets (**R4**).

Table 1. Molecular features about drug compounds that participants focus on during drug target selection

Discipline		Molecular Features	Description
Chemistry		IC_{50}	The concentration of an inhibitor needed to block a given predefined stimulus by 50% [44].
		Inhibition constant (K_i)	The concentration needed to perform half maximal inhibition [44].
		K_d	The equilibrium dissociation constant of a ligand receptor complex measured in a binding assay [30].
		Selectivity	The drug's ability to preferentially produce a desired versus a non-desired effect [8].
Pharmacology	*in vitro*	IC_{50}	Similar to IC_{50} in Chemistry but the inhibition is on cell receptors instead of enzyme [44].
		EC_{50}	The effective concentration of an agonist producing half maximum response to the particular drug [30].
		Selectivity	Similar to selectivity in Chemistry but with fewer reaction sites.
		hERG	The inhibition of hERG channel. Lower value indicates lower cardiotoxicity and therefore lower risks.
		Solubility	The maximum saturation concentration of a substance in a solvent [47]. High solubility is desired in drug design.
	in vivo	ED_{50}	A dose or concentration of an agonist producing half maximal pharmacological effect *in vivo* [17, 30].
		Half-life	The time required for the concentration of a drug to decline to half of its initial value [44].
		AUC	The area under the plasma drug concentration-time curve.
		Bioavailability	The extent a drug becomes completely available to its biological destinations [12].
Clinical Pharmacy		Adverse effects	Any undesired harmful effects in a patient or clinical investigation subject administered a pharmaceutical treatment and which is not required to have a causal relationship with this treatment [26].

The Overview (Fig.1 C) presents the overall performance distributions of existing drug compound research on the candidate targets to help understand the current research progress and difficulty as measured by assorted molecular features and inspire possible areas of improvement (**R1, R2, R4, R5**). The Publication Trend view (Fig.1 D) shows the research trend of each drug target searched by the user (**R1**). The Detail view (Fig.1 E) facilitates a detailed exploration of the research landscape of candidate drug targets in each discipline (i.e., chemistry, pharmacology, and clinical pharmacy) and helps integrate the research data across disciplines (**R1, R2, R3**). Furthermore, a collection of interactions, such as sorting, highlighting, and tooltips, is also provided for users to examine and compare the drug targets freely (**R6**).

4.1 Drug Target Search View

The Drug Target Search view (Fig.1 A) allows users to type the name of a drug target (e.g., "EGFR", "KRAS") of interest into the search box and returns a card containing its 2D structure (**R1**). Information about associated publications also appears in the same row in the Overview and Publication Trend view. Users can hover over a structure to enlarge it. They can also drag the target card up and down to place similar ones next to each other for easier comparison (**R6**). Corresponding information in the Overview and the Publication Trend view will also change position accordingly. Upon clicking on a card, the detailed research information of the selected target will be shown in the Detail View. Users can remove a target and all its related information by hitting the delete button on its card.

4.2 Signaling Pathway

The Signaling Pathway view (Fig.1 B) aims to help users understand the interactions between the candidate targets in the search view and remind users of other drug targets that may be overlooked by them (**R4**). Every time a new target is added to the Drug Target Search view, the Signaling Pathway view displays its corresponding signaling pathway with respect to other input targets in a tree format. This view allows users to get a sense of the interconnections (or the lack of connections) among various targets. In particular, each target is denoted by a unique node that may be shared by several paths. There could be several connected components appearing as separate trees. For example, as JAK (Janus kinase) is the downstream drug target of EGFR (Epidermal growth factor recepto), ALK (Anaplastic lymphoma kinase), and HER2 (Human epidermal growth factor receptor 2), the signaling pathways of EGFR, ALK, and HER2 share the node representing JAK in Fig.1 B. All nodes representing the targets searched by the user are highlighted for easily locating. In addition, the Signaling Pathway view supports zooming and panning to obtain a clearer view, especially when the tree visualization becomes overly complex.

Design alternatives. Initially, we have considered displaying the signaling pathway of each drug target under search separately in a tree format (Fig.2 A). However, it will be difficult for medicinal chemists to integrate the information across these signaling pathways to identify the relationships between the targets. We then tried to merge the signaling pathways in a subway map metaphor design (Fig.2 B). Users can click on a drug target of interest, and its signaling pathway will be highlighted for easier inspection. This makes the relationships between the drug targets clearer but causes visual clutters. Specifically, when the number of targets increases, additional overhead is required to distinguish the intertwined links. Also, the researchers we interviewed (Section 3) considered keeping redundant links in the combined pathway graph unnecessary for drug target selection. Thus, we merged the overlapping links between drug targets in our current design (Fig.2 C).

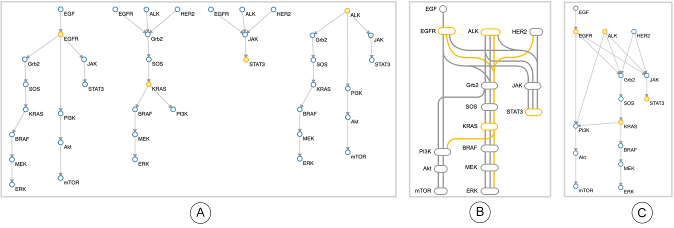


Fig. 2. Design alternatives for the Signaling Pathway view: A) separate trees; B) subway metaphor map; C) our current design.

4.3 Overview

For each drug target, the Overview (Fig.1 C) aims to provide an overarching picture of the drug compound designs related to the molecular features using a tabular design **(R2)**. Each row associates with a drug target and aligns with the target's position in the list of all input candidates in the Drug Target Search view; each column corresponds to a feature introduced in Table 1. The chemistry- (colored in a red theme), pharmacology- (blue), and clinical-pharmacy-related (orange) columns are arranged from left to right following the drug discovery process to show the research progress of the drug target **(R2)**. The background color shading of each cell in the chemistry- and pharmacology-related columns denotes the number of publications whose proposed compounds improved the corresponding molecular feature, while in the clinical-pharmacy-related columns it encodes the number of clinical studies in each phase of clinical trials. Darker color implies more publications fall in the cell; white means no related work exists. The number of publications is shown in the upper right corner of the cell.

To summarize the performance of related work on a molecular feature, we displayed the distributions of reported feature values in these works as a line chart in the corresponding cell **(R2)**. This distribution can also imply how difficult it is to improve the feature **(R5)**. The x-dimension represents the published feature values and the y-dimension indicates the number of publications/studies achieving a value. The minimum and maximum feature values are displayed on the x-axis indicating the progress of the research on a particular molecular feature (**R1**) and hinting researchers about what can be further improved **(R4)**. Hovering over each dot on the plot displays a tooltip of feature value and the number of publications/studies accordingly. Because ongoing or completed but confidential [48] clinical studies can not report their study results, there may be no distribution plot summarizing the clinical trial results even though the cell shows that there are clinical studies on the drug target. This discrepancy might confuse users about the research progress. Thus, hovering over a cell without a distribution plot pops up a tooltip clarifying the reason (*"no results reported"* or *"no studies completed"*). Upon searching a target in the Drug Target Search view, by default the table will add a new row accordingly containing all feature columns. Users can remove columns of features by clicking the delete button in the column headers and add features back from a drop-down menu in the upper right corner of the Overview **(R6)**.

4.4 Publication Trend View

The Publication Trend view (Fig.1 D) displays the temporal changes in the number of publications related to each candidate drug target in three disciplines (i.e., chemistry, pharmacology, clinical pharmacy), respectively, in area charts. It helps users explore the research trend and the evolution of each candidate's popularity over time **(R1)**. Upon searching a target in the Drug Target Search view, an area chart aligning with the target item in the search view appears, showing the publication trend from 1990 to the present by default. Users can adjust the date via a time slider and the area charts will be adjusted accordingly **(R6)**.

4.5 Detail View

Upon selecting a drug target in the Drug Target Search view, the Detail View (Fig.1 E) presents the research landscape of the existing works about it in each discipline (i.e., chemistry, pharmacology, and clinical pharmacy) **(R3)**. The Detail View aims to help users integrate research data of the drug compounds against selected target across disciplines **(R2)**, and compare the detailed drug compound research progress of the targets **(R1)**. The Detail View contains three panels (i.e., Chemistry panel (Fig.1 E1), Pharmacology panel (Fig.1 E2), and Clinical Pharmacy panel (Fig.1 E3)) arranged from left to right following the drug discovery process. The discipline-specific data points on the same drug compound in each panel are linked through lines. Hovering over a data point in any panel highlights the entire chain **(R2)**.

Chemistry Panel The Chemistry panel (Fig.1 E1) summarizes information about the design and synthesis of drug compounds. Each chemistry publication that designed new drug compounds against the user-selected drug target is projected into a 2D canvas as a circular glyph. The length of the colored segment in the outer ring of the glyph denotes the length of the synthesis route of the core drug compound proposed by the publication. The size of the inner core encodes the total number of drug compounds proposed by the publication. The distance between each pair of glyphs indicates the structural similarity between their core drug compounds. Glyphs close to each other indicates they have similar corresponding chemical structures. When a glyph is hovered on, the molecular features of the core drug compounds emphasized in chemistry research are displayed in a pop-up tooltip (Fig.5 (b)). The information of the corresponding publication, including graphic abstracts, title, author(s), publication year, DOI (Digital Object Identifier), venue, citation number, and affiliation, are also shown in the tooltip. Users can click on the DOI to check the publication online.

Glyph alternatives. The circular glyph in the Chemistry panel was designed and refined several times based on the feedback from the six researchers (Section 3). The first alternative (Fig.3 A) is similar to our final design except that the total number of drug compounds proposed by a publication is encoded using a monotonic, sequential color scale in red hue in the background of the inner core. A darker or lighter color indicates more or fewer compounds, respectively. However, we rejected this design as the researchers implied that they hoped to intuitively understand how big design space is that the compounds proposed by each paper occupy. Thus, in the second alternative (Fig.3 B), we used the size of the ring to represent the number of drug compounds in a paper. However, this design was rejected as it is difficult to identify the distance between two glyphs to figure out the structural similarity of their drug compounds. This leads us to the current design (Fig.3 C).

Pharmacology Panel The Pharmacology panel (Fig.1 E2) displays the common molecular features of each drug compound concerned in pharmacological testings. It contains two heatmaps corresponding to the *in vitro* (left) and *in vivo* (right) assays, respectively. In each heatmap, each column represents a molecular feature and each row corresponds to a drug compound. If a drug compound proposed in some chemistry publications never advances to pharmacological testing, it does not have a corresponding row in the *in vitro* heatmap. Similarly, the *in vivo* heatmap does not contain rows associated with drug compounds that have not proceeded to *in vivo* assays. We applied a monotonic, sequential color scale in blue hue to the background of a tile to encode its feature value. Darker (lighter) color indicates higher (lower) value. White is for the case when certain molecular features have not been tested and/or reported in the related publications though the drug compounds have been studied in pharmacological research. Detailed feature value is displayed in a tooltip when hovering over a tile. If all the rows cannot be fit into the panel, scrolling will be enabled. The Pharmacology panel also allows users to sort rows in a heatmap in ascending or descending order of the values in a specific column **(R6)**.

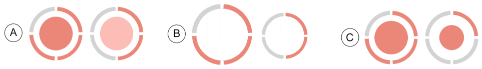


Fig. 3. Glyph alternatives. The length of the colored segment in the outer ring all represents the length of the synthesis route while the number of drug compounds in a paper is encoded differently: A) by the sequential color scale of the inner core, e.g., a darker color indicates more drug compounds studied; B) by the size of the ring, e.g., a bigger size indicates more drug compounds included; and C) our current design.

Clinical Pharmacy Panel The Clinical Pharmacy panel incorporates a Sankey diagram to show the clinical study information from *phase I* to *phase III* of the drug compounds that have been advanced to clinical trials (Fig.1 E2). Each section in the Sankey diagram represents a drug compound and consists of nine subsections arranged in vertical order, corresponding to nine statuses of clinical studies: 1) not yet recruiting; 2) recruiting; 3) enrolling by invitation; 4) active, not recruiting; 5) suspended; 6) terminated; 7) completed; 8) withdrawn; and 9) unknown status [24]. If there are no studies in certain status, the subsection will be left empty. A subsection contains one or more rectangles, each representing an organization that is conducting or has conducted clinical studies about the drug. The trace connecting rectangles in different phases shows the progress of each clinical trial conducted by an organization. Hovering over a rectangle highlights the trace and triggers a tooltip showing the organization and drug name. If the organization's clinical trial is terminated or withdrawn, reasons will be shown too. To help users easily grasp how many drugs against the selected target are tested in clinical trials and how many organizations are involved, we mark each rectangle in the form of 'drug ID - organization ID'.

5 Usage Scenario

We describe how Hannah, a PhD student in medicinal chemistry, uses MedChemLens to complete drug target selection. Hannah has picked four candidate drug targets for cancer, including EGFR, ALK, KRAS (Kirsten rat sarcoma virus), and STAT3 (Signal transducer and activator of transcription 3). Now she wants to use MedChemLens to investigate and compare these targets and choose one as her research direction.

She starts from the Drug Target Search view (Fig.1 A) by searching four drug targets and inspecting their structures. Then she examines the Signaling Pathway view (Fig.1 B) and notices that HER2 is an upstream target similar to EGFR and ALK which she missed in the target candidate collection. Thus, she searches "HER2" in the Drug Target Search view. Also, she finds that KRAS and STAT3 are downstream targets of EGFR, ALK, and HER2. Thus, if she chooses to study EGFR, ALK, and HER2, she also need to learn KRAS and STAT3, since upstream targets taking effect needs to go through downstream targets [2].

Then she turns to the Publication Trend view (Fig.1 D) to understand each drug target's research popularity and trend. She adjusts the publication date to 2012 - 2022 as she usually does. The view shows that EGFR has the most related publications, which indicates its high popularity and made her tentatively decide to rule out EGFR. Nevertheless, before making the final decision, she further clicks "EGFR" in the Drug Target Search view and examines existing drug compounds against EGFR in the Detail View. In the Chemistry panel, she easily notices that most glyphs are large and lie together, indicating that the chemical structures proposed were studied in-depth and that the latest proposed compounds were similar to the previous ones. In the Clinical

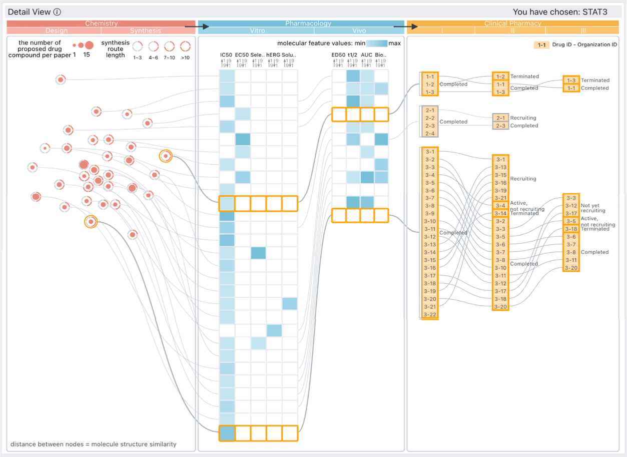
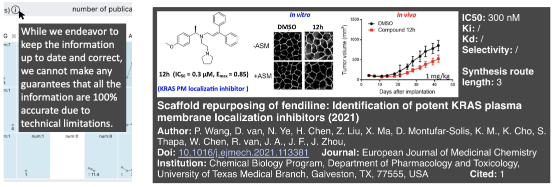

Detail View ⓘ
You have chosen: STAT3
Chemistry
Design
Synthesis
Pharmacology
Vitro
Vivo
Clinical Pharmacy
I
II
III
the number of proposed drug compound per paper 1 15
synthesis route length 1~3 4~6 7~10 >10
molecular feature values: min max
IC50 EC50 Sele.. hERG Solu..
ED50 t1/2 AUC Bio..
1-1 Drug ID - Organization ID
Completed
Terminated
Recruiting
Active, not recruiting
Not yet recruiting
distance between nodes ∝ molecule structure similarity

Fig. 4. Detail View for STAT3

Pharmacy Panel, she observes that several drugs have passed all three phases of clinical trials. These information all suggest that the research about the drug compounds against EGFR is relatively thorough. Hence, Hannah decides to delete EGFR from the Drug Target Search view. Following a similar process, she rules out HER2 and ALK.

She then turns to the Overview (Fig.1 C) to compare STAT3 and KRAS regarding research potential, value, and difficulty. Hannah estimates that the research progress of STAT3 is better than that of KRAS. First, there are more clinical trials and reported study results of STAT3 than those of KRAS. Second, when hovering on the cells of *phase III*, the tooltip of KRAS's cell shows "no studies completed" whereas the tooltip of STAT3's cell shows "no results reported". Hence, although there is no distribution plot in either cell, the tooltips show that some drugs against STAT3 have passed clinical trials while the drugs against KRAS not. Therefore, Hannah judges that STAT3 is more promising than KRAS. From the white cells in the Overview indicating no works focusing on the corresponding molecular features, she finds that some features (e.g., selectivity) in both KRAS and STAT3 are ignored by previous work, which could have high research potential. As Hannah's research lab focuses on the potency of drug compounds, she views related columns and removes the other columns in the Overview. She notices that the drug compounds against KRAS achieved better (lower value is better) IC_{50} in chemical research stage than those against STAT3 though there are more related works against STAT3, which suggests improving the potency of STAT3 may be challenging.

Hannah further uses the Detail View to understand the existing drug compounds against KRAS and STAT3. Since medicinal chemists design new drug compounds based on existing ones, Hannah wants to examine the research potential by comparing the existing drugs on them. Thus, she studies the Chemistry panel. She notices the glyphs representing drug compounds against STAT3 mostly lie together (Fig.4), whereas there are outliers among the glyphs against KRAS (Fig.1 E1), indicating that the chemical structures in the corresponding publications have not been studied thoroughly. Hannah therefore comprehensively examines the information about these publications and pharmacological properties (in the Pharmacology panel) of the proposed drug compounds. She also read some papers in detail through the DOI in the tooltip (Fig.5 (b)). In addition, she figures the synthesis routes of these chemical structures are not long, and some structures have been evaluated by pharmacological testings. Thus, she estimates that designing drug compounds against KRAS based on these scaffolds would be feasible. Hannah finally compares STAT3 and KRAS in the Clinical Pharmacy panel. It shows that three drugs against STAT3 have been advanced to clinical trials with two of them have passed *phase III*. Following the lines across panels, she observes that the chemical structures of the three drugs are not similar. In contrast, the two glyphs in the Chemistry panel corresponding to the two drugs against KRAS that have been advanced to clinical trials are close, implying that only one scaffold was explored thoroughly. Comprehensively considering these factors, Hannah finally selects KRAS as her future research direction.



While we endeavor to keep the information up to date and correct, we cannot make any guarantees that all the information are 100% accurate due to technical limitations.

IC50: 300 nM
Ki: /
Kd: /
Selectivity: /

Synthesis route length: 3

Scaffold repurposing of fendiline: Identification of potent KRAS plasma membrane localization inhibitors (2021)

Author: P. Wang, D. van, N. Ye, H. Chen, Z. Liu, X. Ma, D. Montufar-Solis, K. M., K. Cho, S. Thapa, W. Chen, R. van, J. A., J. F., J. Zhou,
Doi: 10.1016/j.ejmech.2021.113381 **Journal:** European Journal of Medicinal Chemistry
Institution: Chemical Biology Program, Department of Pharmacology and Toxicology, University of Texas Medical Branch, Galveston, TX, 77555, USA **Cited:** 1

(a) (b)

Fig. 5. (a) The disclaimer; (b) An example of the tooltips shown in the Chemistry panel. The paper shown in the tooltip is [53].

6 Implementation

MedChemLens has an interactive web interface built with React framework. It is published on a web server so that users can easily retrieve the website with a link and run it on their own laptops. After users input a drug target, the tool will automatically extract and pre-process relevant data and store it in a pre-cached memory for further visualization use. In this section, we describe the system architecture (Fig.6) of MedChemLens for extracting the information needed by medicinal chemists in their drug target selection process and constructing visualizations.

6.1 Data Collection

First, to provide users with drug target properties, we collected the image of the drug target structure from PDBe[1] and signaling pathway information from OmniPath database [49]. Then we automatically collected the publications and experimental reports about the drug target. Specifically, as suggested by the researchers (Section 3), we chose three top journals of each discipline, that is: *European Journal of Medicinal Chemistry, Journal of Medicinal Chemistry, Drug Discovery Today* for chemistry; *Nature Reviews Drug Discovery, Journal of Pharmacology and Experimental Therapeutics, Advanced Drug Delivery Reviews* for pharmacology; *the New England Journal of Medicine, the Lancet, the Journal of the American Medical Association* for clinical pharmacy. To get the number of publications in each journal related to the inputted target name, our program accessed the publisher site of the journal and obtained the search results using the target name as the query string and the journal name as the restriction. For example, the publisher of *European Journal of Medicinal Chemistry* is ScienceDirect. Then we used its official Search API[2] to get the search results of the user input target name. The number of publications per year about the drug target in each discipline is counted by summing up the numbers of publications per year in the three journals of that discipline. According to the interviews, medicinal chemists mainly focus on chemistry articles. Therefore, we collected the full texts of the publications in the three chemistry journals using DOIs of publications in the search results. These full texts contain the metadata, structural information, and main text of each publication. We wrote a script to automatically discard the publications that did not propose new drug compounds (e.g., surveys) by checking whether the main text contains the names of the molecular features and whether the graphical abstracts of the articles contain chemical structures using ChemSchematicResolver [4].

Next, we extracted the number of all proposed drug compounds from publications. We randomly sampled 50 medicinal chemistry articles and checked with the researchers about some general writing patterns in chemical publications. We found that the authors of the chemical articles usually assigned IDs (e.g., "6", "5b") to all their proposed drug compounds, and we identified the common patterns of the IDs. In this way, we got the number of all new drug compounds the publication proposed by counting the number of unique IDs that following the naming pattern identified. If the core drug compound in a paper had been advanced to clinical trials, it would be given a specific drug name (e.g., "mZIENT", "AZD9150"). In the same way we identified the drug compound IDs, we extracted the drug name of the core drug compound from chemical publications. Based on the extracted drug names, we collected the information in the clinical pharmacy discipline that medicinal chemists need about the clinical trials of the drugs using

[1] https://www.ebi.ac.
[2] https://dev.elsevier.com/text_mining.html

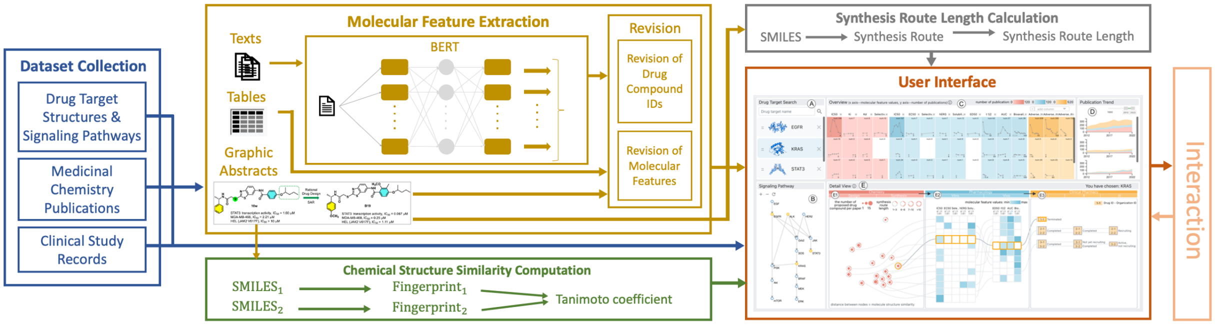

Dataset Collection
Drug Target Structures & Signaling Pathways
Medicinal Chemistry Publications
Clinical Study Records
Molecular Feature Extraction
Texts
Tables
Graphic Abstracts
BERT
Revision
Revision of Drug Compound IDs
Revision of Molecular Features
Chemical Structure Similarity Computation
SMILES$_1$
SMILES$_2$
Fingerprint$_1$
Fingerprint$_2$
Tanimoto coefficient
Synthesis Route Length Calculation
SMILES
Synthesis Route
Synthesis Route Length
User Interface
Interaction

Fig. 6. The system architecture and pipeline of MedChemLens (The graphic abstract image is from [21]).

the official API of ClinicalTrials.gov[3].

6.2 Molecular Feature Extraction

We developed a pipeline to extract the molecular features. The pipeline consists of two modules: an NLP (Natural Language Processing) module and a revision module.

6.2.1 The NLP Module

Although molecular features are numerical values, their textual patterns in publications may vary, and the features of core drug compounds and derivative compounds are mixed, such as "K_i = 176 nM", "...the IC_{50} values for compounds 5a and 5b on EGFRT790M were 5.52 and 25.8 nM, respectively". Thus, we used a NLP model, BERT [16], to automatically extract molecular features of the core drug compound from the textual contents of each publication. Two authors of this paper annotated 528 papers and the data was randomly split into 90% training set and 10% testing set [51, 55]. Before we passed the articles into the model, we first pre-processed the documents to construct the vocabulary and perform word-to-index mapping. The BERT model achieved an accuracy of 93.9% on core drug compound ID identification. However, it had a limited performance on the molecular features with an accuracy of 66.6% on average. One reason for the relatively low performance is that many molecular features are reported in tables or figures resulting in the failure of data extraction from the textual contents. Thus, we further proposed a revision module to revise and complement the extraction results of the BERT module by extracting the information from tables and figures in publications.

6.2.2 The Revision Module

We first validated the results of the NLP module by format checking. Molecular features are numerical values and we have identified the general patterns of the drug compound IDs (Section 6.1). Therefore, if the core drug compound IDs or certain molecular features extracted by the NLP module were empty or did not conform to those patterns, we marked those extracted fields to be revised or filled.

Revision of drug compound ID We used EasyOCR[4] to extract the textual words and their positions in a graphic abstract and identified drug compound IDs from the extracted text based on the patterns we summarized. Nevertheless, many graphic abstracts may contain two or more drug compound IDs. Based on our sampling and checking with medicinal chemical researchers, we found that the core compound is commonly at the rightmost position. Hence, if there were multiple drug compound IDs in a graphic abstract, we utilized the absolute position of each compound ID to retrieve the rightmost one.

Revision of molecular features Firstly, we extracted the values of the molecular features reported in graphical abstracts using a similar method to that of revising drug compound IDs. Then we extracted the molecular features of the core drug compounds from tables in the articles. Specifically, for each table in a publication, we first verified whether it contained our identified core drug compound ID. If so, we would locate the cell containing the value of the molecular feature by identifying the row (or column) whose number corresponding to the drug compound ID and the column (or row) whose header is the name of the expected molecular feature.

After executing the revision module, our pipeline finally reached an accuracy score of 97.0% on drug compound ID extraction and 80.6% on average on molecular feature extraction. We acknowledge that we did not further evaluate the pipeline outside of our training data due to the lack of publicly available large-scale labeled dataset. Our main goal is to propose a basic method for automatically extracting molecular features from chemical publications. Future work could fine-tune our model based on their research data (e.g., publications and lab reports). Also, to avoid the over-reliance on our system, we added disclaimers in MedChemLens (Fig.5 (a)) to remind users that there may be inaccuracy in the returned results because of technological limitations.

6.3 Chemical Structural Similarity Computation

We calculated the structural similarity between the core drug compounds of each pair of publications based on the simplified molecular-input line-entry system (SMILES) of the drug compounds, which is a line notation for describing chemical structures in textual strings [56]. The chemical structures of the core drug compounds are generally shown in the graphic abstracts of papers. Therefore, we first used ChemSchematicResolver [4] to resolve the chemical structures in the graphic abstracts to SMILES. Then using RDKit [35], we obtained the extended connectivity fingerprint with bond diameter 4 (ECFP4), which encodes the topological information of a chemical structure as a fixed-length binary bit vector [45], of each core drug compound based on its SMILES. Finally, we calculated the Tanimoto coefficient [37], a similarity coefficient of two fingerprints, to represent the similarity between two drug compounds.

6.4 Synthesis Route Length Calculation

To calculate the synthesis route length of the core drug compound in each publication, we used the API of IBM RXN[5] to predict the synthesis routes based on the SMILES of the core drug compound. Since each drug compound would have several different synthesis routes, we defined its synthesis route length as the length of the shortest synthesis route with higher than 90% confidence that starts from commercially available chemical entities.

7 USER STUDY

We conducted a within-subjects study with 16 participants to evaluate the effectiveness of our proposed system, MedChemLens. According to the interviews with the medicinal chemical researchers (Section 3), online search, which is a common practice in drug target selection, is used as the baseline in the control condition.

7.1 Participants

We recruited 16 participants (8 males, 7 females, and 1 prefer not to say; age range 22-31) through online advertising and word-of-mouth. Two of them are postdocs who had more than eight years of medicinal chemistry studying experience and had multiple top research publications. They also had experience working in company labs. Two participants

[3] https://clinicaltrials.gov

[4] https://github.com/JaidedAI/EasyOCR

[5] https://rxn.res.ibm.com

had PhD degrees and had more than five years of research experience. Seven participants had between two to five years of experience, and the remaining five had only one year of experience. Participants self-reported their familiarity with drug targets about cancer and central nervous system (CNS) disease. 10 participants reported themselves as novices (N), four as knowledgeable (K), and two as experts (E).

7.2 Procedure

We designed two drug target selection tasks, both of which are representative in current medicinal chemistry research and have a similar size of related publications in our experiments:

1. *T1:* Rank five drug targets for cancer – EGFR, HER2, ALK, KRAS, STAT3, based on how much the participant would like to choose the target as their research direction.

2. *T2:* Rank five drug targets for CNS diseases – Amyloid-beta precursor protein (APP), Catechol-O-methyltransferase (COMT), Dopamine transporter (DAT), Monoamine oxidase B (MAO-B), and Serotonin transporter (SERT), based on how much the participant would like to choose the target as their research direction.

Each participant was invited to complete the two tasks separately in the control and experiment conditions. In the control condition, participants were allowed to use any search engines they usually use in their routine practices, e.g., pubChem [31], Google Scholar, to find any online resources. In the experiment condition, participants were allowed to use MedChemLens only. Before the task with MedChemLens, participants were asked to spend 5 minutes familiarizing themselves with the tool. We counterbalanced the task assignment and the order of the two conditions to minimize the potential order effect. Each participant was given 60 minutes for each task. They were encouraged to think aloud during these two sessions, and we recorded each session with participants' permission. After each task, we asked the participants to write down their reasons for the final ranking and fill out a questionnaire (please see the supplementary material) to rate their experience on a 7-point Likert scale. To better understand participants' ratings and behavior, we further conducted a semi-structured interview with them upon the completion of the two sessions.

8 RESULTS

In this section, we summarize quantitative results on participants' performance, user confidence and cognitive load, and qualitative feedback from the user study.

8.1 User Performance

To investigate how well MedChemLens helps users select drug targets, we statistically analyzed the participants' performance of the target selection tasks in the user study. We measured the user performance using task completion time, the number of publications each participant inspects, and the quality of their final selections.

Completion time We conducted a paired samples t-test to compare users' task completion time as we found the completion time followed the normal distribution, in both control and experimental conditions. Compared with using online search (47.37, [41.64, 53.10] 95% CI), participants using MedChemLens (34.35, [27.71, 40.99] 95% CI) spent significantly less time ($t = 5.52$, $df = 15$, $p < .01$) completing the target selection task.

Number of publications each participant inspects To assess the effectiveness of our system in helping users filter desired information, we counted the number of publications each participant inspected in each task. As a Shapiro-Wilk test showed a significant departure from the normal distribution for MedChemLens ($W(16) = .83$, $p = .006$), we conducted a Wilcoxon signed-rank test to compare the number of publications each participant inspected in both control and experiment conditions. The results show that participants using MedChemLens (6.69, [3.11, 10.27] 95% CI) clicked on significantly fewer articles for a detailed read ($Z = -2.14$, $p < .05$) than when they used online search (11.00, [7.66, 14.34] 95% CI). To figure out whether MedChemLens indeed saves users' efforts in screening relevant publications, we further interviewed the participants to understand their intent when opening certain papers. Five participants said that they opened some papers with online search, but found the papers not related to what they wanted after reading the paper for a while. In contrast, when using MedChemLens, participants could easily narrow down to their desired papers via various filtering mechanisms. P4 (M, 30, E) explained, *"I can easily find articles I need, such as the ones on the core of clusters [in Chemistry Panel] that proposed representative drug compounds, and the ones whose proposed compounds have been advanced to clinical trials"*.

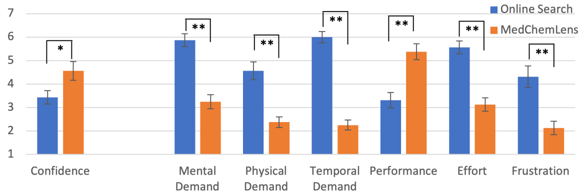


Fig. 7. Means and standard errors of the participants' confidence in their selections (left) and cognitive load in drug target selection process (right) on a 7-point Likert scale (*: $p < .05$, **: $p < .01$)

Final selections To investigate the effectiveness of MedChemLens on supporting drug target selection, we invited two experts who have more than 10 years of medicinal chemistry research experience to evaluate participants' final decisions (i.e., the rankings of the given drug targets). After discussing with the experts, we selected rationality and comprehensiveness as measures to evaluate whether the participants' final decisions were rational and whether they examined the targets comprehensively. Specifically, we provided each participant's final rankings and the corresponding justifications to experts and asked them to rate participants' final decisions in terms of rationality and comprehensiveness on a 7-point (1 – not rational/comprehensive at all, 7 - extremely rational/comprehensive) Likert scale. Both experts were blind to the study condition, and the order of the participant results was randomized. We analyzed the experts' ratings using Wilcoxon signed-rank tests. The results show that participants' final decisions were perceived by experts to be significantly more rational ($Z = -2.47$, $p < .05$) and comprehensive ($Z = -2.13$, $p < .05$) in the MedChemLens condition (rationality: 4.28, [3.80, 4.76] 95% CI; comprehensiveness: 4.59, [4.21, 4.97] 95% CI) than in the online search condition (rationality: 3.34, [2.83, 3.86] 95% CI; comprehensiveness: 3.91, [3.40, 4.41] 95% CI), which indicates that MedChemLens helped users make better-informed choices. We acknowledge that experts' ratings may be subjective and biased. To minimize such effects, we asked the experts to rate participants' final decisions following the common criteria in the field of medicinal chemistry and based on whether the participants considered aspects comprehensively and made correct decisions accordingly instead of comparing the participants' rankings with theirs.

8.2 User Confidence and Cognitive Load

To examine users' experience of the drug target selection process, we conducted statistical analysis on participants' ratings in the post-study survey about their confidence in the final selections and their cognitive load of completing the target selection tasks.

As shown in Fig.7, participants reported to be significantly more confident (Wilcoxon signed-rank test: $Z = -2.43$, $p < .05$) in their final selections using MedChemLens than searching online themselves. This result is mainly because participants thought that they were able to investigate each target more sufficiently and therefore gained more comprehensive insight regarding each target with the assistance of MedChemLens (P12, F, 24, K). Thematic analysis on users' target selection process and the justifications for their selections also reveal that in the online search condition, 12 participants overlooked several aspects (e.g., research popularity), of which the importance was emphasized by them in the MedChemLens condition. In addition, 11 participants, especially self-reported novices, stated that they had more control of their target selection process with MedChemLens. For example, P2 (M, 27, N) complained that he did not know where to start when facing thousands of search results in the online search condition. P11 (M, 22, N) added, *"The information I got through online search is not systematic, and I do* *not know where this information lies in the big picture of the research about that drug target. In contrast, I know how much I have understood about the drug target when using MedChemLens"*.

Using Wilcoxon signed-rank tests, we analyzed participants' cognitive load of drug target selection in the online search condition and MedChemLens condition based on their ratings. The results (Fig.7) show that MedChemLens significantly reduced users' cognitive load in all related dimensions. This result indicates that MedChemLens did not overwhelm users while assisting users in processing more information.

8.3 Qualitative Feedback

In general, most participants showed positive responses to the usability ($M = 5.75$, $SD = 1.13$) and usefulness ($M = 5.75$, $SD = 1.13$) of MedChemLens. To further understand the reasons behind the scores, two authors of this paper conducted a thematic analysis on the transcripts of the post-study interview.

8.3.1 Intuitive, Systematic and Time-saving System

All participants regarded the system as "intuitive", "systematic" and "time-saving". Five participants thought the Signaling Pathway view was one of the most helpful views. For example, when using online search, P2 (M, 27, N) got confused as an article reported the properties of drug compounds against not only the candidate target but also some downstream targets that he did not know. In contrast, he said the Signaling Pathway view showed him why publications included other targets in addition to the candidate targets. Another benefit of MedChemLens is that it provides a holistic picture of the existing work about each drug target and helps users intuitively compare them. In general, participants believed that they could easily know which targets attract more research (13/16), against which targets some drugs passed clinical trials (12/16), and whether most publications designed drug compounds based on similar scaffolds or based on different scaffolds (7/16). Additionally, four participants commented on the convenience of interactions. For instance, P4 (M, 30, E) said, *"it is helpful to allow me to drag similar targets together and compare them"*.

8.3.2 Inspiration and Insightfulness

The system was considered "inspiring" and "insightful". In general, participants were excited about MedChemLens as it not only shows what has been done but also uncovers potential opportunities as to what can be done in future research. Six participants reported that they gained insights from Overview about what molecular features of drug compounds could be further improved. Interestingly, we found that participants who explored the Detail View starting from different panels might get different insights. More specifically, 12 participants began with the Chemistry panel and easily found the papers at the center of the research clusters that proposed classical drug compound structures. Two participants began with the Pharmacology panel and used the sorting function to locate the papers whose corresponding drug compounds performed best on pharmacological features. Other participants began with the Clinical Pharmacy panel and went back to the Chemistry panel following the lines across the panels to find the druggable chemical structures. Moreover, P14 (M, 22, N) pointed out that he got useful information that he never realized that he needed to know. He explained, *"In the past, I mainly focused on the papers that proposed representative molecules. Now the Chemistry panel reminds me of the papers that are the outliers of the clusters. The chemical structures in the outlier papers seem to be more creative and may have greater research potential"*. Eight participants commented that MedChemLens helped them evaluate the possible 'benefits' and 'risks' of choosing the targets. For example, P2 (M, 27, N) said *"some papers focus on designing drugs against this target [KRAS], but no drug compounds have passed the clinical trials. There seems to be an opportunity to make breakthroughs if I choose to study this target, but it might be too risky for me as a novice [researcher]"*.

8.3.3 Adaption of Workflows

From the user study, we observed that participants would adapt their own workflows to the capabilities of MedChemLens. Most participants (14/16) stated that they would prefer beginning with MedChemLens to make decisions since it allows them to quickly get a general picture of the drug target and navigate to specific areas of interest for a focused analysis. The other two participants would like to first search online for general information (e.g., latest news) about the drug targets to gain an initial understanding and then use MedChemLens to explore the scholarly documentations about the targets. Interestingly, four participants proposed that MedChemLens may support their other research tasks in addition to drug target selection. For example, P2 (M, 27, N) said *"sometimes my professor would directly tell me that a certain molecular feature of drug compounds against a certain target may need to be further improved. Then MedChemLens could help me check whether the feature indeed could be improved and help me filter related papers to analyze how to accomplish it"*.

9 Discussion

Generalizability Although our system is domain-specific, our visual design and pipeline could be easily extended to other interdisciplinary experimental science fields (e.g., biomedicine). In these areas, researchers always need to collect and integrate information from multiple areas. Our molecular feature extraction pipeline could be adapted to help extract other types of textual, numerical, and/or visual information from related publications and be adjusted based on the characteristics of the disciplines. The components of our pipeline could be made into individual modules for users to plug-and-play and customize easily. For instance, if key features are reported in tables in publications, the part of the pipeline that processes tables can be applied. In addition, the idea behind the Chemistry panel of organizing publications around figures could be applicable to other disciplines that rely on images to showcase their contributions, such as data visualization.

Lessons learned. We learned several practical lessons for visualization research during our system design and evaluation. 1) *Choose the data organization method that best fits the field.* We organized chemical articles based on their proposed chemical structures in our system. Researchers confirmed that such design matches their intuition well and helps them gain quick insights into the research landscapes of drug targets and the relations between the papers. 2) *Provide flexibility by customizing configurations.* We found that the decision-making strategies vary across researchers. Thus, it is important to allow users to adjust the organization methods of visual information as needed. For example, in our user study, participants thought that dragging drug targets with their related information was helpful for the target comparison.

Limitations First, we only focused on three top journals of each discipline as a proof of concept, and the set of molecular features presented in our system may be incomplete. Second, we did not evaluate our pipeline outside of the training data. The imperfection of the pipeline may affect the effectiveness of MedChemLens. Third, as medicinal chemistry research often takes many years [57], within the scope of our study, we could not examine users' satisfaction with their decisions after they researched into their selected drug targets for a long time.

10 Conclusion

In this paper, we presented MedChemLens, an interactive visual tool to support medicinal chemists in selecting drug targets. MedChemLens integrates information from three disciplines (i.e., chemistry, pharmacology, clinical pharmacy) and organizes scholarly documentations following the practice of each individual discipline. Also, MedChemLens captures and visualizes factors implying the possible difficulty of experiments. Through a within-subjects study, we demonstrated the effectiveness of MedChemLens in helping users analyze relevant literature and experimental data to select research directions.

Acknowledgments

Many thanks to Zifei Chen, Huoming Huang, Shicheng Hou, Qihao Xu, and Shenan Wang for their valuable inputs. This work is supported by the Research Grants Council of the Hong Kong Special Administrative Region, China under Grant No. C6030-18G.

References

[1] B. B. Abbott and K. S. Bordens. *Research design and methods: A process approach*. McGraw-Hill, 2018.

[2] V. Abhyankar, P. Bland, and G. Fernandes. The role of systems biologic approach in cell signaling and drug development responses—a mini review. *Medical Sciences*, 6(2):43, 2018.

[3] J. Bajorath. Molecular similarity concepts for informatics applications. In *Bioinformatics*, pp. 231–245. Springer, 2017.

[4] E. J. Beard and J. M. Cole. Chemschematicresolver: a toolkit to decode 2d chemical diagrams with labels and r-groups into annotated chemical named entities. *Journal of Chemical Information and Modeling*, 60(4):2059–2072, 2020.

[5] F. Beck, S. Koch, and D. Weiskopf. Visual analysis and dissemination of scientific literature collections with survis. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):180–189, 2015.

[6] A. Benito-Santos and R. Therón. Glassviz: Visualizing automatically-extracted entry points for exploring scientific corpora in problem-driven visualization research. In *2020 IEEE Visualization Conference (VIS)*, pp. 226–230. IEEE, 2020.

[7] M. Berger, K. McDonough, and L. M. Seversky. cite2vec: Citation-driven document exploration via word embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):691–700, 2016.

[8] H. Buschmann, R. Mannhold, and J. Holenz. *Drug Selectivity: An Evolving Concept in Medicinal Chemistry*, vol. 72. John Wiley & Sons, Germany, 2018.

[9] J. Chen, M. Ling, R. Li, P. Isenberg, T. Isenberg, M. Sedlmair, T. Möller, R. S. Laramee, H.-W. Shen, K. Wünsche, et al. Vis30k: A collection of figures and tables from ieee visualization conference publications. *IEEE Transactions on Visualization and Computer Graphics*, 27(9):3826–3833, 2021.

[10] J. Choo, C. Lee, H. Kim, H. Lee, Z. Liu, R. Kannan, C. D. Stolper, J. Stasko, B. L. Drake, and H. Park. Visirr: Visual analytics for information retrieval and recommendation with large-scale document data. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 243–244. IEEE, 2014.

[11] J. Choo, C. Lee, C. K. Reddy, and H. Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1992–2001, 2013.

[12] S.-C. Chow. Bioavailability and bioequivalence in drug development. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(4):304–312, 2014.

[13] G. Costagliola and V. Fuccella. Cybis: A novel interface for searching scientific documents. In *2011 15th International Conference on Information Visualisation*, pp. 276–281. IEEE, 2011.

[14] C. V. Dang, E. P. Reddy, K. M. Shokat, and L. Soucek. Drugging the'undruggable'cancer targets. *Nature Reviews Cancer*, 17(8):502–508, 2017.

[15] A. Dattolo and M. Corbatto. Visualbib: narrative views for customized bibliographies. In *2018 22nd International Conference Information Visualisation (IV)*, pp. 133–138. IEEE, 2018.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. doi: 10.18653/v1/N19-1423

[17] S. Dimmitt, H. Stampfer, and J. H. Martin. When less is more–efficacy with less toxicity at the ed50. *British Journal of Clinical Pharmacology*, 83(7):1365, 2017.

[18] M. Dörk, N. H. Riche, G. Ramos, and S. Dumais. Pivotpaths: Strolling through faceted information spaces. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2709–2718, 2012.

[19] C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63(12):2351–2369, 2012.

[20] N. Elmqvist and P. Tsigas. Citewiz: a tool for the visualization of scientific citation networks. *Information Visualization*, 6(3):215–232, 2007.

[21] D. Gao, N. Jin, Y. Fu, Y. Zhu, Y. Wang, T. Wang, Y. Chen, M. Zhang, Q. Xiao, M. Huang, et al. Rational drug design of benzothiazole-based derivatives as potent signal transducer and activator of transcription 3 (stat3) signaling pathway inhibitors. *European Journal of Medicinal Chemistry*, 216:113333, 2021.

[22] L. Gomez. Decision making in medicinal chemistry: The power of our intuition. *ACS Medicinal Chemistry Letters*, 9(10):956–958, 2018.

[23] C. Görg, Z. Liu, J. Kihm, J. Choo, H. Park, and J. Stasko. Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *IEEE Transactions on Visualization and Computer Graphics*, 19(10):1646–1663, 2012.

[24] G. Gresham, J. L. Meinert, A. G. Gresham, and C. L. Meinert. Assessment of trends in the design, accrual, and completion of trials registered in clinicaltrials. gov by sponsor type, 2000-2019. *JAMA Network Open*, 3(8):e2014682–e2014682, 2020.

[25] B. Gretarsson, J. O'donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth. Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):1–26, 2012.

[26] I. H. T. Guideline. Clinical safety data management: definitions and standards for expedited reporting e2a. In *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use*, 1994.

[27] F. Heimerl, Q. Han, S. Koch, and T. Ertl. Citerivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):190–199, 2015.

[28] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, 2012.

[29] R. Hvidtfeldt. *The structure of interdisciplinary science*. Springer, 2018.

[30] P. Imming. Chapter 1 - medicinal chemistry: Definitions and objectives, drug activity phases, drug classification systems. In C. G. Wermuth, D. Aldous, P. Raboisson, and D. Rognan, eds., *The Practice of Medicinal Chemistry (Fourth Edition)*, pp. 3–13. Academic Press, San Diego, fourth edition ed., 2015. doi: 10.1016/B978-0-12-417205-0.00001-8

[31] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, et al. Pubchem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, 2016.

[32] J. Knowles and G. Gromo. Target selection in drug discovery. *Nature Reviews Drug Discovery*, 2(1):63–69, 2003.

[33] G. Koscielny, P. An, D. Carvalho-Silva, J. A. Cham, L. Fumis, R. Gasparyan, S. Hasan, N. Karamanis, M. Maguire, E. Papa, et al. Open targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research*, 45(D1):D985–D994, 2017.

[34] T. S. Kuhn. 9. the essential tension: Tradition and innovation in scientific research. In *The Essential Tension*, pp. 225–239. University of Chicago Press, 2011.

[35] G. Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling, 2013.

[36] B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson. Understanding research trends in conferences using paperlens. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, pp. 1969–1972, 2005.

[37] C. Mellor, R. M. Robinson, R. Benigni, D. Ebbrell, S. Enoch, J. Firman, J. Madden, G. Pawar, C. Yang, and M. Cronin. Molecular fingerprint-derived similarity measures for toxicological read-across: Recommendations for optimal use. *Regulatory Toxicology and Pharmacology*, 101:121–134, 2019.

[38] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 2019.

[39] D.-T. Nguyen, S. Mathias, C. Bologa, S. Brunak, N. Fernandez, A. Gaulton, A. Hersey, J. Holmes, L. J. Jensen, A. Karlsson, et al. Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Research*, 45(D1):D995–D1002, 2017.

[40] F. Osborne, E. Motta, and P. Mulholland. Exploring scholarly data with rexplore. In *International Semantic Web Conference*, pp. 460–477. Springer, 2013.

[41] J. Owens. Determining druggability. *Nature Reviews Drug Discovery*, 6(3):187–187, 2007.

[42] E. A. D. Ralph A. Bradshaw. *Handbook of Cell Signaling*. Cell Biology. Academic Press, 2004.

[43] V. S. Rao and K. Srinivas. Modern drug discovery process: An in silico approach. *Journal of Bioinformatics and Sequence Analysis*, 3(5):89–94, 2011.

[44] S. M. Roberts and A. J. Gibb. Introduction to enzymes, receptors and the action of small molecule drugs. In *Introduction to Biological and Small Molecule Drug Research and Development*, pp. 1–55. Elsevier, 2013.

[45] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

[46] S. D. Satyanarayanajois and R. A. Hill. Medicinal chemistry for 2020. *Future Medicinal Chemistry*, 3(14):1765–1786, 2011.

[47] K. T. Savjani, A. K. Gajjar, and J. K. Savjani. Drug solubility: importance and enhancement techniques. *International Scholarly Research Notices*, 2012, 2012.

[48] R. Stephens, R. Langley, P. Mulvenna, M. Nankivell, A. Vail, and M. Parmar. Interim results in clinical trials: Do we need to keep all interim randomised clinical trial results confidential? *Lung Cancer*, 85(2):116–118, 2014.

[49] D. Türei, A. Valdeolivas, L. Gul, N. Palacio-Escat, M. Klein, O. Ivanova, M. Ölbei, A. Gábor, F. Theis, D. Módos, et al. Integrated intra-and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, 17(3):e9923, 2021.

[50] J. Von Eichborn, M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne, and R. Preissner. Promiscuous: a database for network-based drug-repositioning. *Nucleic Acids Research*, 39(suppl_1):D1060–D1066, 2010.

[51] E. M. Voorhees and D. M. Tice. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. European Language Resources Association (ELRA), Athens, Greece, May 2000.

[52] C. S. Wagner, J. D. Roessner, K. Bobb, J. T. Klein, K. W. Boyack, J. Keyton, I. Rafols, and K. Börner. Approaches to understanding and measuring interdisciplinary scientific research (idr): A review of the literature. *Journal of Informetrics*, 5(1):14–26, 2011.

[53] P. Wang, D. van der Hoeven, N. Ye, H. Chen, Z. Liu, X. Ma, D. Montufar-Solis, K. M. Rehl, K.-J. Cho, S. Thapa, et al. Scaffold repurposing of fendiline: identification of potent kras plasma membrane localization inhibitors. *European Journal of Medicinal Chemistry*, 217:113381, 2021.

[54] X. Wang, S. Liu, J. Liu, J. Chen, J. Zhu, and B. Guo. Topicpanorama: A full picture of relevant topics. *IEEE Transactions on Visualization and Computer Graphics*, 22(12):2508–2521, 2016.

[55] A. Warstadt, A. Singh, and S. R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

[56] D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[57] G. Wu, T. Zhao, D. Kang, J. Zhang, Y. Song, V. Namasivayam, J. Kongsted, C. Pannecouque, E. De Clercq, V. Poongavanam, et al. Overview of recent strategic advances in medicinal chemistry. *Journal of Medicinal Chemistry*, 62(21):9375–9414, 2019.

[58] Y. Yamanishi, M. Kotera, Y. Moriya, R. Sawada, M. Kanehisa, and S. Goto. Dinies: drug–target interaction network inference engine based on supervised analysis. *Nucleic Acids Research*, 42(W1):W39–W45, 2014.

[59] Z.-J. Yao, J. Dong, Y.-J. Che, M.-F. Zhu, M. Wen, N.-N. Wang, S. Wang, A.-P. Lu, and D.-S. Cao. Targetnet: a web service for predicting potential drug–target interaction profiling via multi-target sar models. *Journal of Computer-Aided Molecular Design*, 30(5):413–424, 2016.

[60] Y. Yu, E. Xu, E. Xia, H. Huang, B. Hao, and S. Zhang. A method to accelerate and visualize iterative clinical paper searching. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pp. 1332–1336. IOS Press, 2019.

[61] J. Zhang, C. Chen, and J. Li. Visualizing the intellectual structure with paper-reference matrices. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1153–1160, 2009.

[62] J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan. Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2080–2089, 2013.

[63] Z. Zhou, X. Wen, Y. Wang, and D. Gotz. Modeling and leveraging analytic focus during exploratory visual analysis. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.