SynthLens: Visual Analytics for Facilitating Multi-step Synthetic Route Design

Qipeng Wang, Rui Sheng, Shaolun Ruan, Xiaofu Jin, Chuhan Shi, Min Zhu

Abstract—Designing synthetic routes for novel molecules is pivotal in various fields like medicine and chemistry. In this process, researchers need to explore a set of synthetic reactions to transform starting molecules into intermediates step by step until the target novel molecule is obtained. However, designing synthetic routes presents challenges for researchers. First, researchers need to make decisions among numerous possible synthetic reactions at each step, considering various criteria (e.g., yield, experimental duration, and the count of experimental steps) to construct the synthetic route. Second, they must consider the potential impact of one choice at each step on the overall synthetic route. To address these challenges, we proposed *SynthLens*, a visual analytics system to facilitate the iterative construction of synthetic routes by exploring multiple possibilities for synthetic reactions at each step of construction. Specifically, we have introduced a tree-form visualization in *SynthLens* to compare and evaluate all the explored routes at various exploration steps, considering both the exploration step and multiple criteria. Our system empowers researchers to consider their construction process comprehensively, guiding them toward promising exploration directions to complete the synthetic route. We validated the usability and effectiveness of *SynthLens* through a quantitative evaluation and expert interviews, highlighting its role in facilitating the design process of synthetic routes. Finally, we discussed the insights of *SynthLens* to inspire other multi-criteria decision-making scenarios with visual analytics.

Index Terms—Visual Analytics, Multi-criteria Decision Making, Synthetic Route Design.

I. INTRODUCTION

The synthesis of novel molecules is crucial to many fields, ranging from drug design to material science. For example, in drug design, a crucial step in successfully synthesizing safe and effective drugs is designing synthetic routes that can produce complex molecules and directly impact the efficiency and viability of the synthetic process. These routes enable chemists not only to discover but also to produce novel molecules of drugs. However, it is a challenging task to design a feasible synthetic route, often taking over a decade and costing upwards of US$100 million.

Currently, researchers typically use retrosynthesis [8], an automated method that breaks down the target molecule into starting molecules, to construct synthetic routes. However, they often find these routes impractical and prefer to rely on

Q. Wang, M. Zhu are with Sichuan University. Email: wangqipengscu@stu.scu.edu.cn, zhumin@scu.edu.cn

R. Sheng, X. jin are with Hong Kong University of Science and Technology. Email: rshengac, xjinao@connect.ust.hk

S. Ruan is with Singapore Management University. Email: slruan.2021@phdcs.smu.edu.sg

C. Shi is with Southeast University. Email: chuhanshi@seu.edu.cn

Corresponding author: M. Zhu and C. Shi.

reference papers. Therefore, they explore synthetic reactions from reference papers to manually design synthetic routes transforming starting molecules into intermediates step by step until obtaining the target molecule. At each step of the designing process, researchers must balance multiple criteria (e.g., yield, duration, and experimental difficulty), which may conflict when selecting the optimal synthetic reaction. Such multi-criteria decision-making task makes the synthetic route design process tedious and time-consuming. Additionally, each decision can affect not only current step's outcome but also subsequent choices and the overall synthetic route. For instance, choosing the former between two possible reactions at the early step: a high-yield but slower versus a low-yield but fast reaction may improve yield but extend the overall duration, leading to increased costs and delays. The decision may also restrict reagent options in later steps due to intermediate stability. Additionally, the reaction at this early step may limit the choice of reagents in subsequent steps due to the stability of the intermediates produced. The complexity is further heightened by the inclusion of qualitative factors (e.g., experimental procedures) in addition to quantifiable numerical values.

Several visualization tools can help users make multi-criteria decisions. For example, FSLens [1] can assist users in making decisions regarding the location of new fire stations by considering factors such as distance and time. WarehouseVis [2] aims to optimize and simplify the decision-making process for warehouse location selection according to heterogeneous data like renting costs, reachability, and traffic conditions. In the context of designing synthetic routes, where a decision sequence from the starting molecule to the target molecule involves a series of synthetic reactions, each with multiple experimental procedures, the complexity of the decision sequences requires a more comprehensive approach to ensure a successful design. However, current tools cannot aid users in decision-making among numerous possible choices at each step while considering the influences on subsequent choices. Additionally, they fail to provide users with a comprehensive view of decision sequences. To address the mentioned challenges, we propose *SynthLens*, a visual analytics system that assists researchers in designing synthetic routes by integrating and comparing potential synthetic reactions effectively. Firstly, *SynthLens* facilitates designing a synthetic route from the starting molecule to the target molecule by integrating synthetic reactions, whose information is extracted from papers. Secondly, *SynthLens* helps users make a series of decisions that is sequential decision-making by providing an overview of the decision sequences in a comprehensive visualization.

Thirdly, *SynthLens* provides users with multiple choices at each step of constructing sequence decisions and allows them to compare all decision sequences considering multiple factors. Consequently, *SynthLens* enables users to make informed decisions and identify the most promising decision sequences for exploration.

In conclusion, the contributions of our work are:

- The system design requirements are summarized by co-operating with six chemistry experts who are dedicated to designing the synthetic routes of novel molecules.
- A visual analytics system that can help chemists make sequential decisions considering subsequent impacts in designing synthetic routes. To be specific, we introduced a comprehensive visualization to compare and evaluate each decision or all constructed sequences at various steps of exploration, considering multiple criteria.
- Two case studies and expert interviews to validate the effectiveness of our system. Specifically, we found the workflow of *SynthLens* can be extended to multiple field tasks such as retrosynthesis.

II. PROBLEM FORMULATION

To clarify the complex design challenge of constructing synthetic routes, we have clearly defined several key terms that are extracted from our comprehensive literature review in the following content, which are illustrated in Fig.1.

- A **synthetic route** is a sequence of synthetic reactions that can transform a starting molecule into a target molecule [3]. We refer to the synthetic route that has not yet successfully achieved the target as an **intermediate synthetic route**.
- **Starting molecules** are the starting compounds in a synthetic route [4]. Experts usually determine the starting molecule based on their previous research experience. **Target molecules** represent novel molecules that experts endeavour to synthesize, given that these molecules are hitherto unavailable and have not been synthesized previously [4].
- A **synthetic reaction** is a chemical process involving bond cleavage and formation, by which two or more molecules react to form a more complex molecule [4].
- The **reactant** of a synthetic reaction refers to the molecule that acts as the starting material [5], which undergoes a chemical change during the reaction, breaking existing chemical bonds and forming new ones to generate the **products** [5].
- The **main chain** of a molecule is the basic skeleton and serves as the basis of organic molecules [6]. **Side chains** of a molecule refers to the components attached to the main chain [7], mainly determining the properties of a molecule.
- **Retrosynthesis** involves breaking down the structure of a target molecule into simpler, more easily synthesizable reactant, thereby completing the synthesis route [8].

Designing a **synthetic route** is the process of identifying various **synthetic reactions** and integrating them sequentially, with the goal of transforming the **starting molecule** into
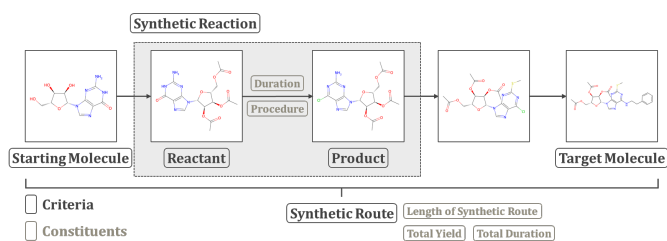

Fig. 1. A synthetic route from the starting molecule to the target molecule consists of several synthetic reactions, each involving specific reactants and products. The evaluation of a synthetic reaction contains criteria such as duration and experimental procedure.

intermediate products step by step until obtaining the desired **target molecule**. To construct a synthetic route, experts need to identify and assess various synthetic reactions from different papers. Then, for each step of the design process, they may filter multiple candidate reactions that have the potential to make up the whole synthetic route. To assess these candidate reactions, experts consider multiple criteria, including **reaction conditions**, **yield**, **reaction duration**, and **synthetic step count**. We will detail these criteria as follows.

- **Reaction conditions** are the various parameters that control the progress of a chemical reaction, such as temperature, pressure, solvent, and catalyst [9]. The difficulty of the reaction conditions is directly reflected in the complexity of carrying out the **experimental procedures**.
- **Yield** is a measure of the efficiency of one chemical reaction, representing the ratio of the actual amount of product obtained in an experiment to the theoretically calculated amount [10]. Improving yield is a significant objective as it directly impacts both the efficiency and cost-effectiveness of the reaction.
- **Reaction duration** refers to the time a synthetic reaction takes to complete [11]. Experts can optimize the reaction duration to speed up the overall synthetic route, enabling the fast production of the products and reducing the required material.
- **Synthetic step counts** is the number of reaction steps required to transform the starting molecule to the target molecule. Shortening the synthetic route may decrease the potential side reactions and impurities, leading to a more effective synthesis.

III. RELATED WORK

A. Multi-step Synthetic Route Design

Computer-Aided Synthesis Planning (CASP) aims to assist researchers in designing and optimizing multi-step synthetic routes by using computer technology. The goal of CASP is to automate the design of synthetic routes, reduce the trial-and-error process, and increase the efficiency and success rate of synthesis. Firstly, several studies on CASP can generate a variety of synthetic routes using predictive models. For example, WODCA [12] is a tool for forward synthetic and retrosynthetic route prediction using the fundamental nature of chemical bonding to guide suitable retrosynthetic breakpoints.

AiZynthFinder [13] is open-source software that uses an artificial neural network policy to quickly and efficiently perform retrosynthetic prediction, providing potential precursors for a given molecule. Nevertheless, the accuracy of predicted routes can sometimes fall short of expectations. Consequently, human expertise should be integrated to improve the reliability and precision of prediction. Some tools provide intuitive functions for the hands-on construction and iterative optimization of synthetic routes. ICSYNTH [14] utilizes reaction cores extracted from various databases to construct synthetic suggestion trees under user control. LinChemIn [15] is a Python toolkit that enables chemo-informatics operations on synthetic routes and reaction networks, allowing users to operate synthetic routes. However, existing studies lack the interaction for users to add synthetic reactions into their synthetic routes dynamically, which aligns with the decision process of chemists. Furthermore, current tools face challenges in facilitating researchers in choosing the most suitable synthetic routes from multiple candidates by considering various criteria, such as yield, duration, experimental procedures, and reaction conditions.

B. Visual Analytics for the Chemical Domain

Visual analysis has been widely employed in various fields in the chemical domain, such as drug discovery [16], chemistry education [17], and chemical reaction exploration [18]. Some of these studies can facilitate the discovery of novel molecules and identify relationships between different molecules. For instance, DataWarrior [19] can assist users in visualizing and analyzing chemical biology data and identifying correlations and associations among various molecules. Navejaet al. [20] presents the novel constellation plots to enable users to identify and interpret Structure-Activity Relationships (StARs). ChemoGraph [21] allows users to explore novel molecules in the form of hypergraphs. Furthermore, other studies can support users in leveraging visual analysis technology to study synthetic reactions or routes by enhancing the learning and understanding of complex chemical processes. For example, Chiu et al. [18] utilized enhanced visual units to help students develop a deeper understanding of chemical reactions by connecting and refining knowledge. Moreover, RetroLens [22] can provide the retrosynthetic routes of a target molecule and present candidate correction steps that AI recommends to promote collaboration between humans and AI. However, our work differs from all the prior studies in that the goal of our work is to assist researchers in designing synthetic routes based on numerous potential synthetic reactions, which is a complex multi-step decision-making problem and requires researchers to balance various criteria in the process.

C. Visual Analytics for Multi-criteria Decision-making

Common single-criteria decision-making tasks focus on only one objective, such as optimizing costs or maximizing benefits [23], while multi-criteria decision-making refers to identifying a satisfactory choice from a range of options evaluated against several criteria.

Furthermore, numerous multi-criteria decision-making systems have been proposed to cater to the unique requirements of specific domains such as location selection [1], [2], [28] and route planning [29]–[31]. Route planning is a typical multi-criteria decision-making problem that involves evaluating and prioritizing options based on multiple criteria [32]. Pathfinder [29] is designed to facilitate users with interactive capabilities for querying, ranking and comparing paths within large and multivariate network datasets. Weng et al. [31] designed BNVA that proposed a progressive route decision-making strategy to evaluate the performance of bus routes. SkiVis [30] enables users to receive customized route recommendations based on preferences such as steepness and crowdedness. In terms of location selection issues, WarehouseVis [2] combines visual analytics and interactive modelling to optimize and simplify the decision-making process for warehouse siting in retail logistics management. SmartAdP [28] facilitates the comparison of multiple billboard placement solutions by multiple attributes, such as traffic volume, speed, origins and destinations. FSLens [1] integrates fire records and collaborative decision-making technology to identify and improve the siting of fire stations. There are also various studies targeting specific multi-criteria decision-making approaches rather than particular applications or domains. WeightLifter [33] enhances multi-criteria decision-making by enabling the exploration of weight spaces for up to 10 criteria, improving the decision efficiency and credibility. Huang et al. [34] proposed a visual analytics framework to explore and compare the evolutionary processes in evolutionary multi-criteria optimization algorithms.

However, in designing chemical synthetic routes, researchers must consider a variety of quantitative factors (e.g., yield and reaction duration) and qualitative textual data in a long, multi-stage decision sequence, which poses significant challenges. Therefore, it is crucial to develop a new visual analytics system to streamline the extensive decision-making process involved in designing these routes.

IV. DESIGN STUDY

We collaborated with six experts (E_1-E_6) from the field of organic chemistry, each with varying years of experience and specialized in different research areas. We gained insights into their workflows and challenges in designing routes for novel molecules through semi-structured interviews, each lasting around 60 minutes. Additionally, we held biweekly meetings with experts to ensure our system aligned with their domain needs, iteratively refine the system based on their immediate feedback. This collaboration helped us identify the crucial factors influencing their decision-making process and the five key design requirements for our system.

R1 **Specify the starting molecule and required synthesis reaction.** The experts need to specify the starting molecule of the synthesis and search for potentially expected synthetic reactions from papers based on their expertise and knowledge. Therefore, our system should provide a flexible input interface for this purpose.

R2 **Examine the details of synthetic reaction.** Experts need to review numerous papers to find synthetic reactions that fulfil their requirements at the step, such as experimental procedures and specific products. For example, E_5 has

emphasized, *"Papers often include extensive details such as experimental procedures, flowcharts, molecule structure diagrams, and spectra diagrams, making it time-consuming to identify useful reactions."* Therefore, it is crucial to help experts effectively assess these details in each paper and decide whether to integrate a particular synthetic reaction into their synthetic route.

R3 **Construct synthetic routes based on synthetic reactions.** Experts will construct synthetic routes step by step through several identified synthetic reactions. Specifically, they might explore multiple potential synthetic reactions at each step. Therefore, our system should support users in manually constructing synthetic routes.

R4 **Support synthetic route comparison at various steps.** E_6 has highlighted that experts frequently assess the intermediate synthetic routes to choose the optimal one for further exploration in the middle step of the designing process. Furthermore, experts also need to compare all completed routes to make the final decision. To facilitate this crucial comparison process, our system should provide a flexible approach to support experts in comprehending multiple factors such as the synthetic step count, yield, duration, and experimental step details at any step. Besides, several experts have mentioned that experts may prioritize differently for various synthetic tasks. Therefore, our system should allow users to adjust the weight of different criteria based on their specific needs and preferences for ranking candidates.

R5 **Optimize synthetic routes iteratively.** Experts often need to optimize the constructed synthetic routes to achieve their desired criteria. Our system should support iterative optimization by allowing experts to modify and refine the synthetic routes based on their assessment and analysis iteratively.

V. DATA PROCESSING

In this section, we first present the overview of our system, followed by a detailed introduction to the paper access modules and the process of information extraction from papers. Then, we propose the workflow of constructing the synthetic route.

A. System Overview

We present *SynthLens*, a visual analytics system that assists organic chemists in designing synthetic routes by integrating synthetic reactions extracted from papers, enhancing the route design process. The data process of implementing *SynthLens* contains two modules: paper access and analysis, and information extraction (Fig.2). *SynthLens* can retrieve papers using the molecule that the users input as search terms. Then, *SynthLens* analyzes and processes the retrieved papers to demonstrate the distribution of these papers. In the information extraction module, we use an LLM-based information extraction technique to extract information on synthetic reactions from papers and transform them into structured data. As for front-end visualization, *SynthLens* allows the users to compose these reactions into synthetic routes and visualize these routes in

a tree-form visualization. *SynthLens* also provides views to assist the users in exploring the decision sequences and making decisions in the designing process.

B. Papers Access and Analysis

Users usually search for papers related to the reactants of each synthetic reaction. The amount of papers retrieved may be substantial, and each paper may propose multiple synthetic reactions. Therefore, it can be challenging for users to identify which paper is relevant to their requirements and what sections in a paper contain useful information. To address this, *SynthLens* needs to retrieve the papers using specific molecules as search terms and extract details about specific synthetic reaction from the chosen paper.

Requesting the user to enter the name of the starting molecule is not advisable because the molecule's name can be complex and may not accurately correspond to the correct molecular structure. As a result, *SynthLens* allows users to set the starting molecule by drawing the structure of a molecule. We have incorporated the open-source chemical structure editor Ketcher[1] into our system to assist users in drawing molecular structures. This editor can generate SMILES (Simplified Molecular Input Line Entry System) strings of a molecule once the user has drawn the molecular structure. SMILES can convey the accurate molecule structure in a compact and machine-readable format. We search papers using API provided by PubMed [35], which accepts SMILES strings as search terms. We can retrieve a list of relevant papers that include information such as title, abstract, keywords, digital object identifiers (DOI), and citation counts by this API.

As the users expect our system to categorize the retrieved papers according to the distribution of their research content, we transform the abstracts of the papers into vector embeddings to capture the semantic information of the text and map semantically similar abstracts to proximity locations in embedded space. To achieve this, we use the PubMedBERT model [36], which is a pre-trained language model designed for the biomedical domain, to convert the paper title and abstracts into high-dimensional embeddings. Then we use the t-SNE algorithm [37] to project high-dimensional embeddings into two-dimension points in the 2D plane, providing a spatial representation of papers. The positioning of points provides a visual cue for the semantic similarities among papers, helping users intuitively discover potential relationships that may not be apparent in tables.

C. Information Extraction from Papers

After users choose a paper to read, they may encounter difficulty in extracting a specific synthetic reaction. This is because one paper may propose multiple synthetic reactions and each synthetic reaction may contain various details. Therefore, our system aims to automate the extraction process. To achieve this, we first use the API provided by UnpayWall[2] to obtain a PDF file of the paper.

---

[1] https://github.com/epam/ketcher
[2] https://unpaywall.org/products/api

To extract synthetic reactions details from the paper, we use Eunomia [38], a chemist AI agent. This tool adopts the commonly used workflow for information extraction using LLMs [39]. Firstly, it converts both papers and queries into text embedding, which are then stored within an embedding database. It then conducts semantic similarity searches to find the most relevant paragraphs. Furthermore, Eunomia employs Chain-of-Verification [40] to reduce hallucinations in LLMs. Base on this automatic extraction pipeline, we created the prompt template in which **A** refers to the reactant of the reaction, i.e., the search term and **B** refers to the specific expected synthetic reaction:

> *1) You are an expert chemist. This document describes the synthetic route or synthetic reaction of the **A**.*
> *2) Find the information of the specific reaction **B** and the reactant of the reaction must be **A**.*

Then, we asked Eunomia to provide the answer, including the following items: reactants, products, solvent, reagent, catalysts, duration, instruments, operation, and yield, and we specified the output format:

> *Your final answer should be a structured JSON format including these items: reactants, products, solvent, reagent, catalysts, duration, instruments, operation, and yield. The answer should be "null" if you cannot find the expected reaction.*

We specified all the expected reactions as parameter **B** and performed several extractions. One example of the extracted information of a synthetic reaction is presented in Fig.4.

Although Eunomia claims that it is able to sace the agent's output in JSON format, we used the JSON module in Python to parse the model's output string in order to ensure that the output format fully meets our needs. If parsing fails, we retry by inserting a directive that asks the model to strictly adhere to the prompt following the original prompt until the maximum time of retrying is reached.

Additionally, we provide users with a key metric: *Context Relevancy* [42] to help them evaluate the extracted information. *Context Relevancy* is an indicator used to evaluate the performance of LLMs in retrieving information, and it measures the relevancy between the content of the answer and its context. Also, to avoid over-reliance on our system, we added disclaimers in SynthLens to remind users that there may be inaccuracy in the extracted information because of technological limitations. Meanwhile, we provide a link to each paper, allowing them to access papers online directly to verify extracted synthetic details.

D. Synthetic Route Construction

While there are automatic methods for generating synthetic routes like retrosynthesis [8], experts caution that the routes automatic generated may not be feasible in practice due to complexity and experimental challenges of these routes. Instead, they prefer designing synthetic routes based on information from reference papers. Therefore, *SynthLens* is designed to provide convenient access to valuable papers and directional suggestions, which helps reduce errors that could arise from full automation and construct synthetic routes. To facilitate the
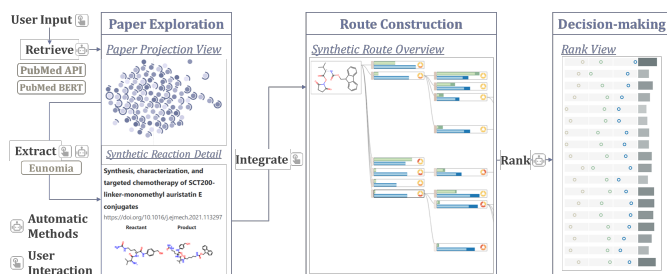

Fig. 2. The analyzing workflow of *SynthLens*: **User Input**: a user can define a starting molecule and expected synthetic reaction, then manually select one from *retrieved papers* for exploration. **Information Extraction**: the synthetic reaction details are then *extracted automatically*. **Synthetic Route Construction**: the user can choose to *integrate* the synthetic reaction into the synthetic route construction to form various decision sequences. **Decision-making**: finally, the user can select the optimal synthetic route from computed *rankings* supplemented by his own preferences. The whole process is the combination of *automatic methods* and *user interaction*.

construction process, we defined a tree-form data structure to organize decision sequences consisting of synthetic reactions from the starting molecule to the target molecule, comprising the following elements.

- The **root** of a tree represents the starting molecule of the synthetic route that users specified. After setting the starting molecule, *SynthLens* will automatically retrieve a number of papers using the starting molecule as a search term.

- A **node** in the tree represents a synthetic reaction extracted from retrieved papers, whose reactant is the product of the **parent node**. Each node also contains essential information, including yield, duration, and SMILES strings of the reactants and products. Upon choosing a node, *SynthLens* will automatically retrieve papers using the product of this node as the search term. The synthetic reaction extracted from specific retrieved papers can be integrated as subsequent nodes, i.e., **child nodes**. We also recorded the nodes' total yield and duration because these values will accumulate as the reaction proceeds sequentially. The output of one step serves as the input for the next. To be specific, if the yield of the first step is 80% and the yield of the second step is 90%, then the second step yields 72% totally (i.e., $80\% \times 90\% = 72\%$). Similarly, we obtain the total duration by calculating the sum of the duration of this node and the parent node.

- A **leaf node**, typically representing the product molecule, has no child nodes. Nodes that are neither root nor leaf are called **intermediate nodes**, which represent intermediate molecules. The node's **layer count** refers to the number of synthetic reactions from the root node(i.e., the starting molecule) to the current node(i.e., the intermediate molecule).

- **Deep and bread exploration strategies.** Deep exploration strategy involves going as far down a branch of the tree as possible before backtracking to explore new branches, while bread exploration strategy involves exploring all neighbour nodes at the present step before moving on to nodes at the next level.
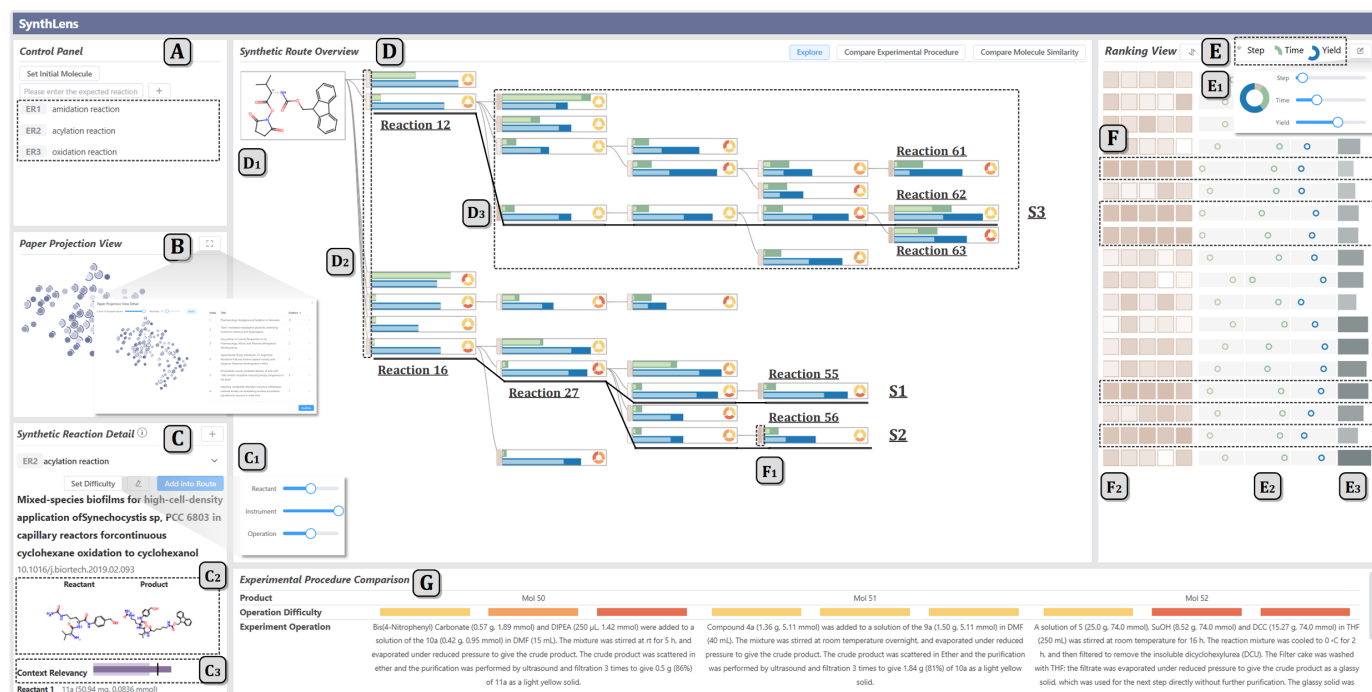
Fig. 3. *SynthLens*: (A) The Control Panel allows users to specify the starting molecule and potentially expected synthetic reactions before designing the synthetic routes. (B) The Paper Projection View presents the distribution of retrieved papers. (C) The Synthetic Reaction Detail shows the detail of extracted synthetic reactions. (D) The Synthetic Route Overview presents a tree-form visualization of the decision sequences of the synthetic routes. (E) The Similarity View shows the similarity in structure among specific molecules. (F) The Rank View visualizes the rank of decision sequences considering three factors with flexible weights. (G) The Experimental Procedure Comparison assists users in comparing the experimental procedures of multiple synthetic reactions.

VI. VISUAL DESIGN

In this section, we detail the visual design of *SynthLens*, which leverages a combination of manual inputs and automated processes to meet the design requirements specified in Section IV. As Fig.3 shows, users can set a starting molecule and input the expected synthetic reaction in the Control Panel (**R1**). Then, they can manually select a paper in the Paper Projection View, where the system has already performed searches and computations. The experimental procedures for the synthetic reactions are then automatically extracted and presented in the Paper Detailed View (**R2**). After annotating the difficulty of a synthetic reaction's experimental procedure, the user can integrate the reaction into the Synthetic Route Overview and end up producing multiple decision sequences (**R3**, **R5**). After route construction, the user can manually choose the most appropriate decision sequence from the automatically computed rankings in the Rank View (**R4**).

A. Paper Overview

To help users explore papers related to a synthetic reaction and decide whether to integrate the reaction into the decision sequence, we proposed the Paper Overview. This view consists of two sub-views: the Paper Projection View and the Paper Detail View. Specifically, the Paper Projection View provides a comprehensive overview of retrieved papers with diverse distribution of research content, while the Synthetic Reaction Detail presents the details of a particular synthetic reaction extracted from the corresponding paper.

1) Control Panel: The Control Panel (Fig.3(A)) allows users to draw the structure of the starting molecule of the whole synthetic route in the editor. Users can plan the synthetic route by identifying the expected reactions to transform the starting molecule into the target molecule based on their differences and chemical domain knowledge. So, the Control Panel allows users to input these expected reactions. This assists the following process of extracting synthetic reactions.

2) Paper Projection View: The retrieved papers present a diverse distribution in research content and include detailed information like titles, abstracts, citation counts, keywords, etc., making the exploration of a large number of papers a challenging task. After interviewing experts, we learned they usually rely on heuristic exploration methods to identify target papers for further exploration. This approach allows users to precisely determine the most relevant papers based on a deep understanding of their contents. Therefore, we employ a scatter plot in the Paper Projection View (Fig.3(B)), which helps users find semantically similar papers after identifying one paper of interest based on their expertise. For example, users are more likely to explore papers spatially close to the selected paper in the scatter plot, rather than just focusing on those ranking highest in the paper list.

Description. In the scatter plot of the Paper Projection View, each point represents a retrieved paper, with the color intensity indicating the paper's relevance to the search term. Rather than using point size to represent relevance, color intensity ensures points remain visible even when the paper's relevance to the search term is low. In contrast, smaller points might become

indistinguishable at a smaller scale. After interviewing experts, we found that they prefer manually exploring papers. This preference arises because automated filters can not capture the semantic difference between papers and carry the risk of omitting significant papers.

After transforming the abstract of these papers into embeddings and dimensionality reduction introduced in Section V, we obtained the coordinates of each paper projected in the 2D plane. However, if these points are mapped directly by coordinates, there will be an overlap among these points. To reduce the overlapping of the points in the scatter plot, we employed a force-directed algorithm provided by d3.js [41] that applies a repulsive force to maintain at least a minimum distance between points, without changing their relative positions. In addition, we added an arc surrounding each point encoding the citation count of this paper, which can indicate its credibility. Proximity in distance among points indicates a similarity in research content. Users are willing to prioritize papers that not only demonstrate a high degree of relevance to the key molecule but also have high citation numbers. Upon exploring a paper, subsequent exploring choices should focus on papers that are near to ensure the consistency of exploration. We display the 100 papers retrieved as related to the specified molecule in the Paper Projection View because experts usually retrieve less than 100 relevant papers when searching synthetic reactions. Showing too many can overwhelm the analysis with unrelated papers. If the user finds the points too dense, they can zoom in and drag to see a clearer picture of the scatter's positional relationships.
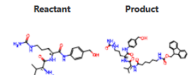
Interaction. Hovering over a point triggers a tooltip to appear, presenting the details of the corresponding paper, including title, abstract, DOI, citation number, and keywords. These details can guide the users' decision on whether to explore a specific paper. When a point is clicked, *SynthLens* will automatically extract the experimental procedures of the expected synthetic reactions from the paper. In addition, we provide a zoom-in button that, when clicked, opens a dialog containing an enlarged scatter plot. In this dialog, we provide two sliders: one to adjust the count of displayed points and another to modify the "perplexity" parameter in the t-SNE algorithm, which controls whether the algorithm emphasizes global patterns or local structures. We also include a tooltip to inform users about this parameter.

3) Synthetic Reaction View: The Synthetic Reaction View (Fig.3(C)) is designed to present the synthetic details extracted from the chosen paper. We allow the users to assess the experimental procedure in three aspects: the availability of necessary raw materials and reagents, the complexity of the experimental operations, and the difficulty in acquiring the necessary laboratory equipment (Fig.3(C$_1$)).

Description. We use RDKit.js[3] to display the molecular structures of the reactants and products (Fig.3(C$_2$)) and present the experimental procedure of the expected reactions the users have specified extracted from the chosen paper in the form of a table, including *reactants*, *products*, *solvents*, *catalysts*, *reagents*, *reaction time*, *instruments*, *operation*, *purification*,

[3]https://www.rdkitjs.com/

Fig. 4. The extracted synthetic details are presented in a table format in our system, consisting of several parts: *raw materials*, *experimental operations*, *duration*, and *yield*, etc. Moreover, our system allow users to modify the extracted details within a form.

and *yield*. We provide links to these papers, allowing users to access them online with a single click. Meanwhile, we provide a essential metric *Context Relevancy* [42] to aid users in assessing the trustworthiness of the extracted information (Fig.3(C$_3$)). We provide a view similar to a bullet chart to assist users in evaluating the effectiveness of information extraction. In this view, the narrower, darker bar represents the average Context Relevancy value of all extracted information with the wider, lighter bar representing the quartile range. The black vertical line indicates the value for this particular extraction result. If users find the extracted results unsatisfactory, they have the flexibility to modify the results based on the article's content and expertise (Fig.4).

Interaction. By clicking the *"Set Difficulty"* button, we allow the users to assess the experimental procedure in three aspects: *material*, *operation*, and *equipment*. We categorize these into three discrete tiers of difficulty: challenging, moderate, and simple, with values 1 to 3 (Fig.3(C$_1$)). By recording the users' annotation of difficulty levels in three aspects and corresponding descriptive text, we utilize PubMedBERT to compute similarity scores, thereby providing recommendations for difficulties of experimental procedures that align with the users' predefined difficulty preference in subsequent annotations. Then, the users can add the synthetic reaction to the end of the currently explored sequence after clicking the *"Add into Route"* button. Additionally, we allow users to manually modify or add the extracted synthetic details after reviewing the article content or practical experimental results.

B. Synthetic Route Construction View

To help users construct and intuitively view multiple synthetic routes while making decisions based on various factors, we propose the Synthetic Route Overview in combination with the Rank View. Additionally, the Molecule Similarity View assists users in selecting the most promising intermediate molecules, guiding them away from less viable options during exploration. The Experimental Procedure Comparison View further support users in scrutinizing the synthetic information to make more informed choices.

1) Synthetic Route Overview: The Synthetic Route View (Fig.3(D)) proposes tree-form visualization to present decision sequences constituting synthetic routes, enabling users to explore various options of synthetic reactions.

Description. In the synthetic route overview, the root represents the starting molecule of the whole synthetic route, and each subsequent node glyph (Fig.5) represents a synthetic reaction whose reactant is the product of the parent node. For instance, the product of *Reaction 16* serves as the reactant of *Reaction 27*. The dual-toned bar on the upper side of each node glyph (Fig.5(A) (B)) represents reaction duration, with the lighter color representing the experimental duration of the current reaction and the darker color representing the cumulative total duration. The bar on the lower side (Fig.5(C) (D)) indicates yield, with the darker bar representing the yield of the reaction in the node and the lighter representing the cumulative total yield traced from the root to the current node. The donut glyph on the right side (Fig.5(E)) reflects the degree of difficulty in the three aspects: material acquisition (Fig.5(E_1)), obtaining and utilizing experimental instruments (Fig.5(E_2)), and the complexity of the experimental operations (Fig.5(E_3)). Nodes within the same decision sequence are connected by lines, while nodes at the same layer are vertically aligned.

Interaction. When a node is hovered on, a pop-up tooltip (Fig.5(F)) shows the *reaction ID*, which consists of the number of layers and the vertical ordering of the nodes, *experiment duration*, *yield*, *total yield*, and the *reactant* and *product* of the corresponding synthetic reaction.

Design alternatives. Fig.6 illustrates alternative designs of the Synthetic Route Overview. The original design alternative consists of two layers of arcs, the outer ones representing the duration and yield, respectively, and the inner circle representing the difficulty of the experimental procedure. However, the dual-layer arcs could be visually complex and potentially overwhelming, making it difficult for users to discern the key information quickly. Moreover, comparing multiple nodes for yield or duration might not be intuitive, especially since these nodes are of the same sizes and there are many nodes in the view. Therefore, we proposed the current design (Fig.5), which juxtaposes two bar charts while positioning a donut glyph on the right of the node, thereby facilitating user comparison across various nodes. Since the "difficulty of a synthetic reaction" are driven from text describing experimental procedures, it is challenging to quantify difficulties. So, we use a donut glyph, with three segments and color intensity representing three aspects of difficulties, which allows users to view the overall difficulty of the synthetic route while primarily making decisions based on duration and yield.

Exploration. Users usually adopt two kinds of exploration strategies using the Synthetic Route Overview, i.e., deep or bread exploration strategies. The deep exploration strategy involves constructing a single decision sequence from the starting molecule to the target molecule while the bread exploration strategy considers various possibilities to integrate multiple reactions. Moreover, users often encounter multiple intermediates during the designing process, which may share similar structures, and must identify the most promising one
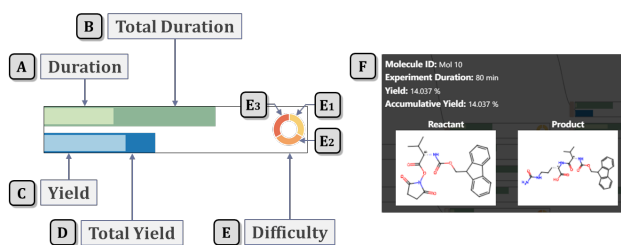

Fig. 5. The design of the node glyph that represents a synthetic reaction: (A) and (B) respectively represent the duration and total duration accumulated from the root to this node. (C) and (D) respectively represent the yield and total yield. The donut glyph (E) presents the annotated experimental procedures' difficulty of this synthetic reaction. Specifically, (E_1), (E_2), and (E_3) correspond to the difficulty in acquiring reactants, obtaining and utilizing experimental instrument, and the complexity of the experimental operations, respectively. The tooltip (F) displays some details about this synthetic reaction.
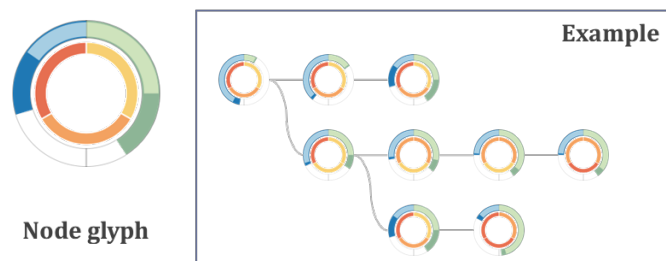

Fig. 6. Design alternatives. The node glyph is designed to represent a synthetic reaction, which consists of two layers of arcs.

for further exploration. To broaden the selection, users need to combine the deep and the broad exploration strategies to construct additional decision sequences that reach the step of producing similar intermediates.

2) Molecule Similarity View: We design the Molecule Similarity View (Fig.3(F)) to help the users find potential intermediate molecules to perform subsequent explorations to obtain molecules that can be compared. The first part of this view is the element left to each node glyph in the Synthetic Route Overview, which refers to the Molecule Similarity Mark (Fig.3(F_1)). When the users click a node glyph, the color of all the Molecule Similarity Mark encodes the similarity between the product molecules of all other nodes in the tree and the chosen node's product molecule. The other part of this view is the Molecule similarity Matrix (Fig.3(F_2)) right to the Synthetic Route Overview. Then, after the users choose a node in the Synthetic Route Overview and click the *"Add into Comparison"* button, an additional column is added to the Molecule Similarity Matrix with a maximum number of five. The color intensity of each element in this column represents the molecular similarity between the product of this node and the products of all decision sequences.

3) Rank View: The users are required to rank various decision sequences represented by the routes from the root to the leaf in a tree structure considering the flexible weighting of multiple factors. Additionally, they need to observe each factor and the weighted sum to make informed decisions.

Description. The horizontal coordinate of dot plots in the Rank View (Fig.3(E_2)) indicates the values of each factor, including *the number of steps*, *yield*, and *duration*, which are

the three most important factors in the decision-making of designing synthetic routes. The color and length of the bar on the right side of the view (Fig.3(E_3)) refer to the weighted total score for the three factors.

Initially, we presented the scores for three attributes using bar charts. However, expert feedback indicated that presenting multiple bar charts together could lead to visual overload. Therefore, considering the need for ranking, visual simplicity, and quick task completion, we opt for dot plots. Moreover, the human eye is not as sensitive in visual perception to assessing length, i.e., the length of bars, as it is to dots. We use bars with color intensity in the last column to represent the weighted total score. Because the bar chart uses multiple encodings, including the length of the bars, color intensity, and position on the common scale, it can represent the weighted total score while also allowing users to assess the rankings of each row intuitively.

Interaction. We provide a sorting feature that allows rows in the Rank View to be arranged based on their scores, facilitating users in selecting the optimal synthetic routes. When a user clicks on a row, the entire corresponding route in the Synthetic Route Overview will be highlighted, facilitating users' review.

Legend. The legend of the Rank View presents the three factors, where the ratio of the length of each arc to the circumference of the corresponding circle represents the weight of each factor. The users can edit the weights by clicking the button beside the legend (Fig.3(E_1)).

4) Experimental Procedure Comparison View: Before finalizing their decision on the specific synthetic route, the users can examine and assess the experimental procedures of several synthetic reactions(Fig.3(G)). By choosing a node within the tree and clicking the *"Add into Comparison"* button, the users can compare the experimental procedure with other nodes in the form of a table. The examination facilitates the intuitive assessment of the difficulties in the experimental procedure of each synthetic reaction.

VII. CASE STUDY

We invited the interviewed experts to explore *SynthLens* remotely online and recorded their exploration process. Through these case studies, we summarized their insights to demonstrate the practical applications of our system.

A. Case Study 1: Synthesis of Linker

E_4, who specializes in anti-tumor targeting drugs, has been concentrating on Adcetris, a cancer treatment drug. He sought to use *SynthLens* to design the synthetic route for Mc-Vak-Cit-PABC-PNP, a linker of Adcetris. He stated that the primary purpose of designing this synthetic route was to serve as a reference for experimental procedures intended for trials, not necessarily to actually produce the target molecules.

User Input (R1). According to the existing research results and the actual laboratory situation, he chose the Fmoc-Val-OSu as the starting molecule (Fig.3(D_1)). He assigned initial weights to three factors: step count, duration, and yield. The

weights for these factors were set as 0.1, 0.3, and 0.6, respectively, in (Fig.3(E_1)). *"I am willing to accept long duration if it means achieving higher yields as often the significance in designing new synthetic route is the maximization of yield."* After observing the difference between the starting and target molecules, he found that the first step should be an amidation reaction based on his prior knowledge. Then, he inputted the name of the expected reaction in the Control Panel (Fig.3(A_1)).

Combining deep and bread exploration strategies (R3). *SynthLens* automatically retrieved papers related to the starting molecule and presented them in the Paper Projection View (Fig.3(B)). E_4 noted a point surrounded by a relatively long arc, indicating a high citation. He checked the topic of this paper and found that this paper was about generating novel linkers for Adcetris. Therefore, he determined this paper for further exploration. Once clicking the corresponding point in the view, *SynthLens* automatically extracted several synthetic reaction information according to the inputted expected reaction whose reactant is the starting molecule, as (Fig.3(C_1)) shows. Adopting the bread exploration strategy, he integrated the identified candidate synthetic reactions into the Synthetic Route Overview (Fig.3(D)) following the root node. Subsequently, he wanted to explore if there were more candidates for the first step. Therefore, he began to explore two other papers. Then, E_4 chose the most satisfied reaction *Reaction 16* by checking the weighted score of these candidates in the Rank View to complete the decision sequence until the synthetic route reached the target molecule, i.e., *S1*. E_4 indicated, *"We need to construct a decision sequence from the starting molecule to the target molecule to ensure that this synthetic route is chemically feasible before subsequent full-scale exploration."* Once the synthetic route exists in principle, he can construct multiple decision sequences for comparison. It is worth noting that the product of different decision sequences may not be the same but may be structurally similar, and he can still compare these decision sequences.

Choosing intermediate reactions for another exploration (R5). The first route *S1* he completed had great difficulty in the experimental procedure, all of the experimental procedures in this route are difficult in experimental operations, so he decided to choose an intermediate synthetic reaction to explore other decision sequences. First, he chose an intermediate reaction *Reaction 27* and constructed a decision sequence *S2*. Although *S2* is not as difficult as *S1* in experimental procedures, it has a significantly lower yield than *S1*. After checking the Molecule Similarity Mark in the (Fig.3(D_2)) and examining the Rank View, E_4 noted that *Reaction 12* not only boasted a high weighted score but also had a product structure that is structurally similar to the product of *Reaction 16*. This observation led him to construct a new decision sequence, ultimately choosing *Reaction 12* for further exploration. He started with *Reaction 12* and constructed decision sequences such as those shown in (Fig.3(D_3)).

Decision making (R4). After completing the construction of decision sequences, E_4 compared the structural similarity of *Reaction 61*, *62*, *63*, *55*, and *56* in the Molecule Similarity View (Fig.3(F_2)) and found that the product of these reactions had the same main chain and only differed in side chains.
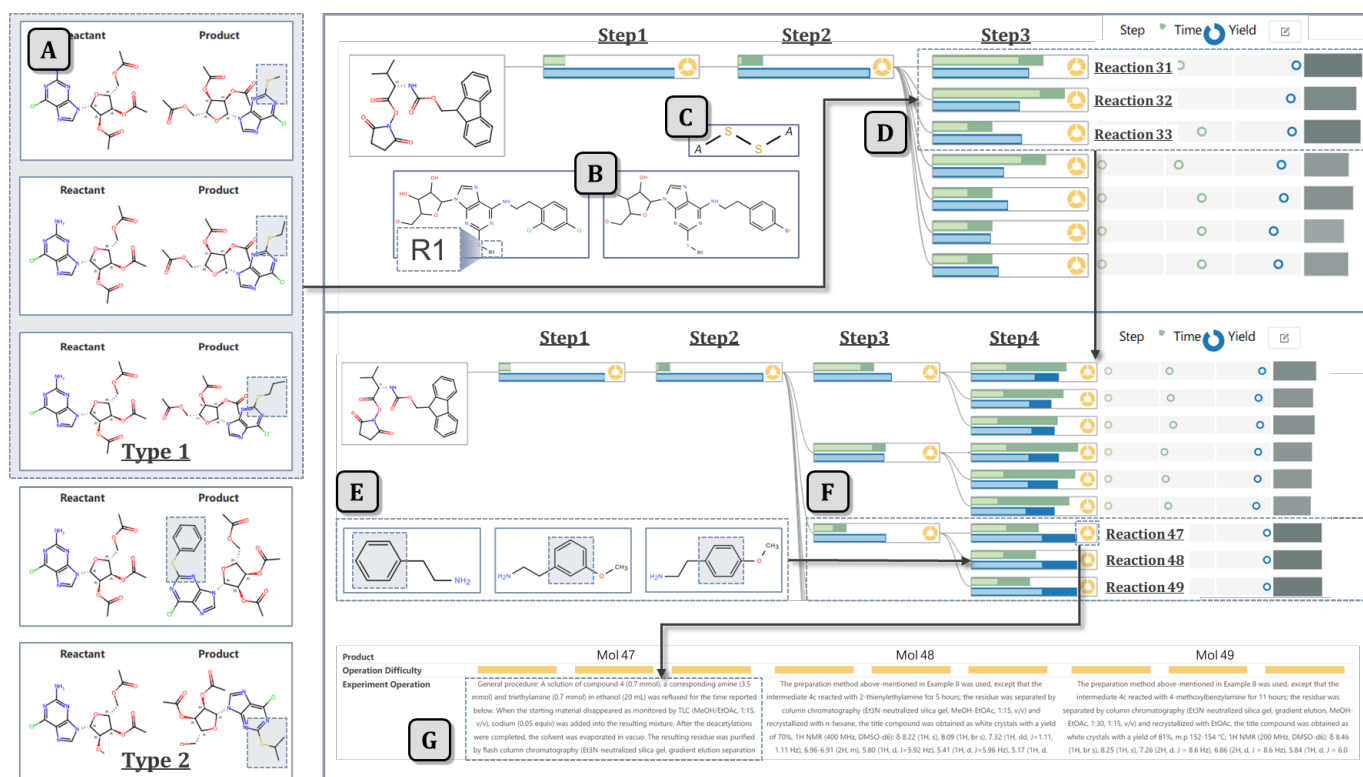
Fig. 7. Design a synthetic route of a class of molecules in Case Study 2. (A) presents several synthetic reactions integrated in *Step3*. (B) shows two of the target molecules. (C) refers to a class of molecules containing double sulfur bonds (-S-S-). (D) and (F) presents decision routes with relatively high weighted scores. (E) presents molecules with similar structures as the reactants of the synthetic reactions in *Step 4*. (G) details several experimental procedures of certain synthetic reactions, which have difficulty in experimental operations.

E_4 indicated that when the product molecules of different sequences share structural similarities, they desire to compare these sequences and choose the optimal one. In such cases, he only needed to make minor adjustments to the chosen constructed route to adapt it for synthesizing the target molecule. Consequently, he decided to rank the synthetic route by checking the Rank View (Fig.3(F_2)) and taking the difficulty of the experimental procedure into account. He thought that although *S1* had a higher weighted score in the Rank View, the experimental procedure of this route was more difficult. So he chose *S3*. *"Although S3 did not have the highest total score, its experimental procedures are relatively moderate in difficulty. This indicates that we can try hands-on in the real experiment."*

In this case, E_4 explored *SynthLens* using flexible strategies, balancing various factors to guide decision-making, and constructing decision sequences in about 50 minutes, which results in the choice of a practical and promising synthetic route for a novel molecule.

B. Case Study 2: Synthesis of A Class of Molecules

Purine is a type of organic molecule that is found in nucleic acids and plays important roles in the human body, especially in fighting against viruses and bacteria. E_2 had designed several new molecules based on purine, as (Fig.7(B)) shows, in which *R1* refers to any chains only with carbon and hydrogen atoms. Since these molecules are newly designed, he needs to

refer to the synthetic route of similar molecules to design the synthetic routes of these novel molecules. Based on the results of his earlier experiments, E_2 set guanosine as the starting molecule. Considering the structural gap between the starting molecule and the target molecule, the expert specified that there would be four steps in this synthetic route.

Compare intermediate synthetic routes (R2, R4). After completing the first two steps, E_2 started to examine papers to complete the third step of his construction process. He found that the product of *Reaction 21* could react with various molecules containing a disulfide bond (-S-S-)(Fig.7(C)) according to the extracted information of synthetic reactions from papers. Based on his domain knowledge in chemistry, E_2 realized there are two possibilities for the main chain of a molecule containing a disulfide bond: either the main chain is straight (*type1*) or the main chain branches out into branched chains (*type2*). As a result, he explored six different papers in the Paper Projection View and the Synthetic Reaction Detail to integrate the synthetic reactions with seven different disulfide bond-containing reactants into the Synthetic Route Overview (Fig.7(D)). Each of these reactions can produce molecules that share a common main chain. He assessed these intermediate synthetic routes considering yield and duration in the Rank View and found that when the main chain of the molecule containing the disulfide bond used in the third step has a straight chain, the yield and reaction time are superior to those of the molecule containing a branched chain (Fig.7(A)). So, he chose *Reaction 41*, *42*, and *43* to complete

the decision sequences. In the fourth step, he found that these reactants could react with a variety of molecules that shared similar structures (Fig.7(E)). As a result, the products of these reactions and the target molecules are similar in structure. After having constructed the decision sequences, he checked the Rank View and found the reactions in (Fig.7(F)) had the highest scores.

Compare the Experimental Procedure (R2). Then, he chose the last three nodes in the fifth step of the tree in the Synthetic Route Overview to compare their experimental procedures. He found that the experimental method of *Reaction 57* required the construction of a vacuum environment during evaporation (Fig.7(G)), which was relatively difficult to achieve. Therefore, he chose to refer to the experimental procedures of *Reaction 58* and *59*.

The case demonstrates the importance of domain knowledge in assessing the choices during the designing process. Additionally, the case indicates that the users need to compare textual data on several experimental procedures to ensure the chosen reactions are practically feasible, particularly avoiding complex operations. E_2 successfully completed this case within 35 minutes.

VIII. EVALUATION

A. Automatic Extraction Evaluation

We conducted an experiment to validate the accuracy of the automatic extraction method of our system, introduced in Section IV. We randomly collected 100 papers proposing synthetic routes, published from 2023 to the present, from four major chemical journals: *The Journal of Organic Chemistry*, *Organic Chemistry Frontiers*, *Organic Process Research & Development*, and *European Journal of Organic Chemistry*. We invited four domain experts, each with over four years of experience in organic chemistry, to annotate the papers. We assigned 50 papers to each expert, ensuring each paper was annotated by two experts. Experts were asked to annotate the reactants, products, and yields of the synthetic routes. It is worth mentioning that experts may give multiple annotations to a paper if it proposes more than one synthetic route. Then, we manually checked each annotation for conflicts. When finding conflicts, we consulted the opinions of the most experienced expert among those we invited to determine the correct annotation. Next, we used the method in our system to extract the synthesis routes from each paper. Using the current standard evaluation metrics for NLP information extraction tasks [38], we employed Precision, Recall, and F1 Score to evaluate the accuracy of our method in extracting synthetic details from chemical papers. An extraction result was considered a *True Positive* only if the reactants, products, and yields were all correctly identified. Any extraction result with at least one error was considered a *False Positive*, and those missed in the extraction but present in the annotations were considered *False Negatives*. The Precision was defined as the *True Positives* out of all extracted details, indicating the accuracy of the method. The Recall was defined as the ratio of *True Positives* to the total number of actual annotations, measuring the ability of the method to find information from

papers. The *F1 Score* was the harmonic mean of Precision and Recall, which is a comprehensive assessment of method performance. Additionally, we tested these metrics using common information extraction tools for chemical papers: ChemDataExtractor [43], ReactionDataExtractor [44], and ChemRxnExtractor [45]. Table I shows the performance comparison among our system and common information extraction tools in chemistry. The precision, recall and F1 scores of our method are higher than those of the other methods.

TABLE I
COMPARISON OF OUR METHOD AND COMMON INFORMATION EXTRACTION TOOLS IN EXTRACTING INFORMATION FROM CHEMISTRY PAPERS

	Precision	Recall	F1 Score
ChemicalDataExtractor [43]	0.871	0.527	0.653
ReactionDataExtractor [44]	0.722	0.436	0.544
ChemRxnExtractor [45]	0.942	0.627	0.753
Our System	**0.944**	**0.798**	**0.865**

B. User Study

To validate the efficacy of *SynthLens*, we conducted a semi-structured interview with the previous experts (E_1-E_6) mentioned in Section IV. Besides, we also involved four new experts (P_1-P_4) in our evaluation, with more than four years of experience in organic chemistry. Each interview lasted one hour and a half. We first introduced the background of our study (10 minutes). Then, we introduced our system in detail (15 minutes). Next, experts could familiarize themselves with the system through free exploration (15 minutes). Then, they were asked to design a synthetic route using our system according to their individual requirements (30 minutes). Finally, we had a semi-structured interview with them to collect their feedback (20 minutes). After the interview, we asked the participants to complete a questionnaire multiple choice using a 5-point Likert scale to evaluate the effectiveness of *SynthLens* in view representation, workflow enhancement, and overall usability. We documented their findings and comments on our system during the process.

Workflow and System Performance. All the experts involved in the evaluation of *SynthLens* expressed their appreciation for its workflow, as it aligned closely with their prior analysis process in organic chemistry. P_1 mentioned, *"The system can accelerate my design process significantly. Specifically, it makes it possible for me to compare different routes even if I have not completed them, resulting in substantial time savings."* Specifically, we found that experts usually prioritize exploring potential synthetic reactions as much as possible in the early steps of their design process. Then, they will identify the most promising one to conduct the subsequent exploration to complete a synthetic route. E_2 said that our ranking view provided him with effective insights to determine which one deserves further exploring. Furthermore, P_3 appreciated our system's capability to support retrospective exploration, recognizing that there might be several potential synthetic reaction candidates in the earlier exploration steps that could lead to improved synthetic routes. *"After completing

a particular route, I often review previous steps to identify any new possibilities. The system is particularly helpful in quickly identifying these potential alternatives through molecular similarity analysis." Particularly, E_5 praised the tree-form visualization for exploration, and he mentioned that this component is also suited for retrosynthetic tasks. "*In retrosynthesis, chemists employ a recursive approach, systematically working backwards from a target molecule to identify feasible synthetic routes. I can also construct a retrosynthetic route by inputting the target molecule through the tree-form visualization of SynthLens.*"

Visualization and Interaction. E_1, E_2, and P_2 all mentioned that the Paper Projection View provides an intuitive visualization of the distribution of retrieved papers, making it easy to identify which papers need to be examined first. Moreover, E_3 and P_1 also highlighted the usefulness of allowing them to specify expected reactions and extract information from the papers quickly. Additionally, P_4 specifically pointed out that the Synthetic Route View combined with tooltips can effectively present the necessary information due to its straightforward design. Finally, P_3 and P_4 pointed out the benefits of combining point visualizations with bar charts in the Rank View. They noted that this combination effectively highlighted the most crucial information (i.e., the weighted score displayed through the rightmost bar charts). In summary, all the experts provided positive feedback on various aspects of the system's visualizations and interactions.

Suggestions. The experts expressed their willingness to use the system but suggested some improvements. Specifically, they noted that the system currently only retrieves academic papers on chemical synthesis. Incorporating the ability to search for patents of this domain would broaden the coverage of synthesis information sources, thereby enhancing the usability of *SynthLens*. P_1 indicated that experts in the field sometimes retrieve relevant papers by inputting multiple molecules with similar structures. Therefore, P_1 suggested that *SynthLens* can recommend structurally similar molecules and allow for the inclusion of more than one starting molecule. This approach would enhance the informativeness of the constructed synthetic routes. Moreover, P_4 suggested that we could take the price of raw material into account, which is also important in deciding whether to integrate a synthetic reaction.

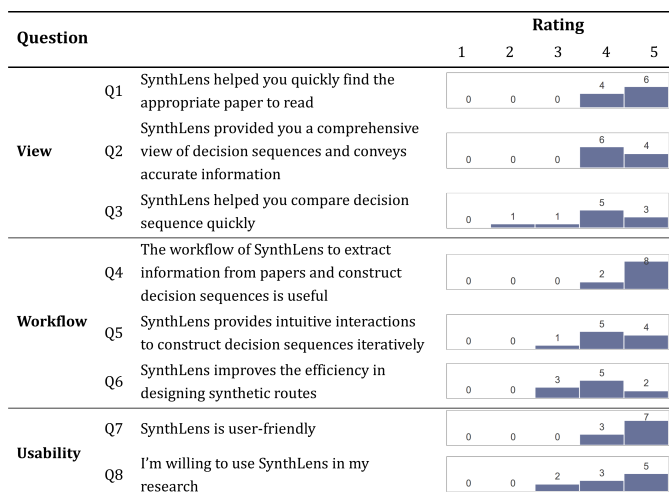| Question			Rating				
			1	2	3	4	5
View	Q1	SynthLens helped you quickly find the appropriate paper to read	0	0	0	4	6
	Q2	SynthLens provided you a comprehensive view of decision sequences and conveys accurate information	0	0	0	6	4
	Q3	SynthLens helped you compare decision sequence quickly	0	1	1	5	3
Workflow	Q4	The workflow of SynthLens to extract information from papers and construct decision sequences is useful	0	0	0	2	8
	Q5	SynthLens provides intuitive interactions to construct decision sequences iteratively	0	0	1	5	4
	Q6	SynthLens improves the efficiency in designing synthetic routes	0	0	3	5	2
Usability	Q7	SynthLens is user-friendly	0	0	0	3	7
	Q8	I'm willing to use SynthLens in my research	0	0	2	3	5

Fig. 8. Questionnaire results. We designed our questionnaire in three aspects: the system's view, workflow, and usability. Experts are asked to rate by a 5-point Likert scale (from strongly disagree to strongly agree).

IX. DISCUSSION

Runtime Performance. *SynthLens* uses the PubMed API for paper retrieval, followed by the conversion of the paper content into high-dimensional embedding representations and dimensionality reduction. This is followed by the information extraction through LLM, which costs almost 1 minute or more and contributes to the main time consumption of the analytic process. However, *SynthLens*' analysis process is designed in parallel mode, allowing users to continue exploring other papers while awaiting the results of the current analysis, thus ensuring the continuity and efficiency of the research process. Notably, *SynthLens* markedly reduces the time consumption and preserves the essential element of manual decision-making in comparison to traditional expert-driven approaches, where the expert manually searches, filters and extracts information from papers, subsequently constructing synthetic routes.

Design Implications. In our interviews, experts indicated that they can adopt the proposed workflow for designing synthetic routes in practice, treating the process as constructing a tree structure where a node represents a synthetic reaction, and the path from root to leaf defines the synthesis route, with attributes like cumulative yield and synthesis time to help identify the optimal route. Moreover, we observed that experts frequently review intermediate nodes to explore promising synthetic route completions. To facilitate this, *SynthLens* includes the Molecule Similarity View, allowing users to select any node and automatically receive alternative possibilities of structure-similar molecules. Additionally, users' feedback indicates that our workflow is suitable for retrosynthetic tasks, where chemists deconstruct a target molecule to find feasible synthetic routes [3]. Our method may also apply to various multi-criteria decision-making scenarios, involving multiple options at each stage and enables users to efficiently evaluate the impact of choices on the outcome, such as fisheries management [47], architecture construction [48], and clinical decision-making [49]. For instance, in highway route selection, involving multiple objectives like safety, cost etc., our tree-form visualization can help decision-makers compare and rank construction segment plans [50].

Scalability. *SynthLens* provides a tree-form visualization to assist experts in constructing and exploring synthetic routes. Although the constructed routes can usually be fully displayed within the current interface size, exploring the entire routes becomes challenging when the node number increases. To address this, our system supports zooming in and out to facilitate exploration, with corresponding adjustments in the size of the items in the Molecule Similarity View and the Rank View. Additionally, we provide a sorting feature that ranks various paths, making it easier for users to select the optimal synthetic route. In the future, we plan to enhance the

Synthetic Route Construction View with an overview feature that presents a summary of the tree-form visualization, which will facilitate navigation through complex options even when dealing with a large number of nodes.

Limitations and Future Work. There are several areas that can be addressed to enhance *SynthLens* further. Firstly, our system supports experts in specifying one currently starting molecule. However, expanding the capability to support multiple starting molecules would give experts more flexibility. Additionally, we can also recommend relevant papers about molecules similar to the current key molecule. This feature is valuable because molecules with similar structures, even if they differ slightly (such as in side chains), may share similar reaction conditions or types. By including these related molecules, the system can help users explore a wider range of relevant synthesis routes. Finally, incorporating additional information from papers, such as spectrograms and validation methods, would enrich the decision-making context for experts. Our main contribution is the proposed analytics analytic pipeline, which allows integration of advanced tools in place of our method, making it both adaptable and scalable to accommodate future advancements.

X. Conclusion

We introduce *SynthLens*, a visual analytics system designed to assist organic chemists in designing synthetic routes by choosing optimal options at each step. The novel tree-form visualization of *SynthLens* simplifies understanding of decision sequences and helps optimal routes with the Rank View. The collaboration with domain experts helps us to establish design requirements and obtain valuable user feedback to refine *SynthLens*. Two case studies demonstrate the practical benefits and effectiveness of *SynthLens* in practical applications. We also envision expanding the application of *SynthLens* to other decision-making scenarios in chemistry fields and scientific researches.

Acknowledgments

References

[1] L. Chen, H. Wang, Y. Ouyang, Y. Zhou, N. Wang, and Q. Li, "FSLens: A Visual Analytics Approach to Evaluating and Optimizing the Spatial Layout of Fire Stations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, pp. 847-857, 2024.

[2] Q. Li, Q. Q. Liu, C. F. Tang, Z. W. Li, S. C. Wei, X. R. Peng, M. H. Zheng, T. J. Chen, and Q. Yang, "Warehouse Vis: A Visual Analytics Approach to Facilitating Warehouse Location Selection for Business Districts," *Computer Graphics Forum*, vol. 39, pp. 483-495, 2020.

[3] S. Ishida, K. Terayama, R. Kojima, K. Takasu, and Y. Okuno, "AI-Driven Synthetic Route Design Incorporated with Retrosynthesis Knowledge," *Journal of Chemical Information and Modeling*, vol. 62, pp. 1357-1367, 2022.

[4] W. A. F. S. Finnigan, "Enzyme Cascade Design: Retrosynthesis Approach,", S. A. R. F. Kara, Ed. Cham: Springer International Publishing, 2021, pp. 7–30.

[5] N. Gerasimchuk, *Chemical Literacy and Writing Chemical Reactions*: Cambridge Scholars Publishing, 2022.

[6] W. Huang, "Combined Main-Chain/Side-Chain Liquid Crystalline Polymers: Synthesis and Supramolecular Ordering,", V. K. A. K. Thakur, Ed. Cham: Springer International Publishing, 2016, pp. 363–389.

[7] S. A. J. E. Li, "Using automated synthesis to understand the role of side chains on molecular charge transport," *Nature Communications*, vol. 13, p. 2102, 2022.

[8] E. J. Corey, "Robert Robinson Lecture. Retrosynthetic thinking—essentials and examples," Chem. Soc. Rev., vol. 17, no. 0, pp. 111–133, 1988.

[9] Y. Kwon, D. Lee, J. W. Kim, Y. Choi, and S. Kim, "Exploring Optimal Reaction Conditions Guided by Graph Neural Networks and Bayesian Optimization," *ACS Omega*, vol. 7, pp. 44939-44950, 2022.

[10] P. Schwaller, A. C. Vaucher, T. Laino, and J. Reymond, "Prediction of Chemical Reaction Yields using Deep Learning," *ChemRxiv*, 2020.

[11] M. Xu, X. Kong, T. He, X. Wu, L. Xie, and J. Li, "Reaction duration-dependent formation of two Cu(ii)-MOFs with selective adsorption properties of C3H4 over C3H6," *Dalton Transactions*, vol. 48, pp. 8617-8624, 2019.

[12] R. Höllering, J. Gasteiger, L. Steinhauer, K. Schulz, and A. Herwig, "Simulation of Organic Reactions: From the Degradation of Chemicals to Combinatorial Synthesis," *Journal of Chemical Information and Computer Sciences*, vol. 40, pp. 482-494, 2000.

[13] S. A. T. A. Genheden, "AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning," *Journal of Cheminformatics*, vol. 12, p. 70, 2020.

[14] S. A. T. A. Genheden, "AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning," *Journal of Cheminformatics*, vol. 12, p. 70, 2020.

[15] M. A. S. M. Pasquini, "LinChemIn: SynGraph—a data model and a toolkit to analyze and compare synthetic routes," *Journal of Cheminformatics*, vol. 15, p. 41, 2023.

[16] L. J. Medina-Franco, K. Martinez-Mayorga, A. M. Giulianotti, A. R. Houghten, and C. Pinilla, "Visualization of the Chemical Space in Drug Discovery," *Current Computer-Aided Drug Design*, vol. 4, pp. 322-333, 2008.

[17] H. Wu, J. S. Krajcik and E. Soloway, "Promoting understanding of chemical representations: Students' use of a visualization tool in the classroom," *Journal of Research in Science Teaching*, vol. 38, pp. 821-842, 2001.

[18] J. L. A. L. Chiu, "Supporting Knowledge Integration in Chemistry with a Visualization-Enhanced Inquiry Unit," *Journal of Science Education and Technology*, vol. 23, pp. 37-58, 2014.

[19] T. Sander, J. Freyss, M. von Korff, and C. Rufener, "DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis," *Journal of Chemical Information and Modeling*, vol. 55, pp. 460-473, 2015.

[20] J. J. U. S. Naveja and J. E. L. Medina-Franco, "Finding constellations in chemical space through core analysis," *Frontiers in chemistry*, vol. 7, p. 510, 2019.

[21] B. Kale, A. Clyde, M. Sun, A. Ramanathan, R. Stevens, and M. E. Papka, "ChemoGraph: Interactive Visual Exploration of the Chemical Space," *Computer Graphics Forum*, vol. 42, pp. 13-24, 2023.

[22] C. Shi, Y. Hu, S. Wang, S. Ma, C. Zheng, X. Ma, and Q. Luo, "RetroLens: A Human-AI Collaborative System for Multi-step Retrosynthetic Route Planning," in *CHI '23* New York, NY, USA, 2023.

[23] C. D. Brumar, S. Molnar, G. Appleby, K. Potter, and R. Chang, "A Typology of Decision-Making Tasks for Visualization," *ArXiv*, vol. abs/2404.08812, 2024.

[24] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, "LineUp: Visual Analysis of Multi-Attribute Rankings," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, pp. 2277-2286, 2013.

[25] Q. Liu, Y. Ren, Z. Zhu, D. Li, X. Ma, and Q. Li, "RankAxis: Towards a Systematic Combination of Projection and Ranking in Multi-Attribute Data Exploration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, pp. 701-711, 2023.

[26] T. Joachims, "Optimizing search engines using clickthrough data," in *KDD '02* New York, NY, USA, 2002, pp. 133–142.

[27] L. Q., L. Q., Z. Z., Y. T., and M. X., "Inspecting the Process of Bank Credit Rating via Visual Analytics," in 2021 IEEE Visualization Conference (VIS), 2021, pp. 136-140.

[28] D. Liu, D. Weng, Y. Li, J. Bao, Y. Zheng, H. Qu, and Y. Wu, "SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, pp. 1-10, 2017.

[29] C. Partl, S. Gratzl, M. Streit, A. M. Wassermann, H. Pfister, D. Schmalstieg, and A. Lex, "Pathfinder: Visual Analysis of Paths in Graphs," *Computer Graphics Forum*, vol. 35, pp. 71-80, 2016.

[30] J. Rauscher, R. Buchmüller, D. A. Keim, and M. Miller, "SkiVis: Visual Exploration and Route Planning in Ski Resorts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, pp. 869-879, 2024.

[31] D. Weng, C. Zheng, Z. Deng, M. Ma, J. Bao, Y. Zheng, M. Xu, and Y. Wu, "Towards Better Bus Networks: A Visual Analytics Approach," *IEEE Transactions on Visualization and Computer* Graphics, vol. 27, pp. 817-827, 2021.

[32] D. Shin, J. Jo, B. Kim, H. Song, S. Cho, and J. Seo, "RCMVis: A Visual Analytics System for Route Choice Modeling," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, pp. 1799-1817, 2023.

[33] S. Pajer, M. Streit, T. Torsney-Weir, F. Spechtenhauser, T. Möller, and H. Piringer, "WeightLifter: Visual Weight Space Exploration for Multi-Criteria Decision Making," IEEE Transactions on Visualization and Computer Graphics, vol. 23, pp. 611-620, 2017.

[34] Y. Huang, Z. Zhang, A. Jiao, Y. Ma, and R. Cheng, "A Comparative Visual Analytics Framework for Evaluating Evolutionary Processes in Multi-Objective Optimization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, pp. 661-671, 2024.

[35] N. C. F. B. US, "Entrez Programming Utilities Help,", 2010.

[36] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," A*CM Trans. Comput. Healthcare*, vol. 3, 2021.

[37] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, no. 86, pp. 2579–2605, 2008.

[38] M. Ansari and S. M. Moosavi, "Agent-based Learning of Materials Datasets from Scientific Literature,", 2023.

[39] J. Lála, O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodriques, and A. D. White, "PaperQA: Retrieval-Augmented Generative Agent for Scientific Research," arXiv preprint arXiv:2312.07559, 2023.

[40] S. Dhuliawala et al., "Chain-of-Verification Reduces Hallucination in Large Language Models," in Findings of the Association for Computational Linguistics: ACL 2024, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 3563–3578. doi: 10.18653/v1/2024.findings-acl.212.

[41] M. Bostock, V. Ogievetsky and J. Heer, "D³ Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, pp. 2301-2309, 2011.

[42] E. S. Shahul, J. James, L. E. Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," *ArXiv*, vol. abs/2309.15217, 2023.

[43] J. Mavračić, C. J. Court, T. Isazawa, S. R. Elliott, and J. M. Cole, "ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science," *Journal of Chemical Information and Modeling*, vol. 61, pp. 4280-4289, 2021.

[44] D. M. Wilary and J. M. Cole, "ReactionDataExtractor 2.0: A Deep Learning Approach for Data Extraction from Chemical Reaction Schemes," *Journal of Chemical Information and Modeling*, vol. 63, pp. 6053-6067, 2023.

[45] J. Guo, A. S. Ibanez-Lopez, H. Gao, V. Quach, C. W. Coley, K. F. Jensen, and R. Barzilay, "Automated Chemical Reaction Extraction from Scientific Literature," *Journal of Chemical Information and Modeling*, vol. 62, pp. 2035-2045, 2022.

[46] M. C. Swain and J. M. Cole, "ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature," *Journal of Chemical Information and Modeling*, vol. 56, pp. 1894-1904, 2016.

[47] M. Booshehrian, T. Möller, R. M. Peterman, and T. Munzner, "Vismon: Facilitating Analysis of Trade-Offs, Uncertainty, and Sensitivity In Fisheries Management Decision Making," *Computer Graphics Forum*, vol. 31, pp. 1235-1244, 2012.

[48] D. Jato-Espino, E. Castillo-Lopez, J. Rodriguez-Hernandez, and J. C. Canteras-Jordana, "A review of application of multi-criteria decision making methods in construction," *Automation in Construction*, vol. 45, pp. 151-162, 2014.

[49] J. G. Dolan, "Multi-criteria clinical decision support: A primer on the use of multiple criteria decision making methods to promote evidence-based, patient-centered healthcare.," *Patient*, vol. 3, no. 4, pp. 229–248, 2010, doi: 10.2165/11539470-000000000-00000.

[50] G. S. Kalamaras, L. Brino, G. Carrieri, C. Pline, and P. Grasso, "Application of Multicriteria Analysis to select the best highway alignment," *Tunnelling and Underground Space Technology*, vol. 15, pp. 415-420, 2000.

Qipeng Wang received his BS in computer science from Sichuan University in 2023. He is currently pursuing a master's degree in the College of Computer Science at Sichuan University. His research interests include visualization and visual analytics.

Rui Sheng is currently a third-year Ph.D. candidate in the HKUST VisLab, within the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. Under the guidance of Professor Huamin Qu, his research focuses on data visualization, human-AI collaboration, and decision-making, particularly in critical domains such as healthcare. For more information, please visit https://dylansheng.github.io/.
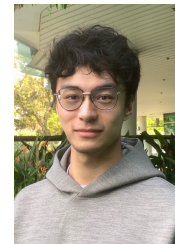
Shaolun Ruan is currently a Ph.D. candidate in School of Computing and Information Systems at Singapore Management University (SMU). His work focuses on developing novel graphical representations that enable a more effective and smoother analysis for humans using machines, leveraging the methods from Data Visualization and Human-computer Interaction. He received his bachelor's degree from the University of Electronic Science and Technology of China (UESTC) in 2019. For more information, kindly visit https://shaolun-ruan.com/.

Xiaofu Jin is currently a third-year Ph.D. candidate in the HKUST VisLab, within the Department of Academy of Interdisciplinary Study at the Hong Kong University of Science and Technology. Her research focuses on accessibility, cognitive modelling, VR/AR, and health applications. For more information, please visit https://xiaofu-jin.github.io/.

Chuhan Shi is an associate professor in the School of Computer Science and Engineering at Southeast University, China. She obtained her Ph.D. degree in Computer Science and Engineering from the Hong Kong University of Science and Technology in 2023. Her research interests include data visualization, visual analytics and human-computer interaction. For more details, please refer to https://shichuhan.github.io/.

Min Zhu received the PhD degree in applied mathematics from Sichuan University, in 2004. She is currently a professor at the College of Computer Science, Sichuan University, and has presided over a number of national and provincial research projects. Based on the research works, she has published more than 100 academic papers in journals and conferences. Her current research interests include visualization, visual analysis, and bioinformatics.