RetroLens: A Human-AI Collaborative System for Multi-step Retrosynthetic Route Planning

Chuhan Shi
cshiag@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Yicheng Hu
yhubf@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Shenan Wang
s.wang@se22.qmul.ac.uk
Queen Mary,University of London
London, United Kingdom

Shuai Ma
shuai.ma@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Chengbo Zheng
cb.zheng@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Xiaojuan Ma
mxj@cse.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Qiong Luo
luo@cse.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China
The Hong Kong University of Science and Technology (Guangzhou)
Guang Zhou, China

ABSTRACT

Multi-step retrosynthetic route planning (MRRP) is the core task in synthetic chemistry, in which chemists recursively deconstruct a target molecule to find a set of reactants that make up the target. MRRP is challenging in that the search space is vast, and chemists are often lost in the process. Existing AI models can achieve automatic MRRP fast, but they only work on relatively simple targets, which leaves complex molecules under chemists' expertise. To facilitate MRRP of complex molecules, we proposed a human-AI collaborative system, RetroLens, through a participatory design process. AI can contribute by two approaches: joint action and algorithm-in-the-loop. Deconstruction steps are allocated to chemists or AI based on their capabilities and AI recommends candidate revision steps to fix problems along the way. A within-subjects study (N=18) showed that chemists who used RetroLens reported faster MRRP, broader design space exploration, higher confidence in their planning, and lower cognitive load.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581469

KEYWORDS

Human-AI collaboration, multi-step problem solving, multi-criteria decision making

ACM Reference Format:
Chuhan Shi, Yicheng Hu, Shenan Wang, Shuai Ma, Chengbo Zheng, Xiaojuan Ma, and Qiong Luo. 2023. RetroLens: A Human-AI Collaborative System for Multi-step Retrosynthetic Route Planning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3544548.3581469

1 INTRODUCTION

The synthesis of chemicals is central to chemistry [67] with applications in many fields (e.g., drug discovery). For example, the National Institute of General Medical Sciences puts about 44.3 million dollars on chemical synthesis per year[1]. A critical step towards the successful and efficient synthesis of chemical molecules is to identify feasible synthetic routes. Multi-step retrosynthetic route planning (MRRP) is one of the most widely used methods to design synthetic routes for the target molecules [7, 56]. It refers to recursively deconstructing a target molecule into its simpler precursors following a reversed chemical reaction until reaching a set of commercially or readily available molecules as starting materials [20]. Then the synthetic route for the target molecule can be obtained by reversing the derived retrosynthetic routes from MRRP.

Chemists typically rely on their prior knowledge, practical experience, and intuition to guide MRRP [56, 73]. Still, manually designing multi-step retrosynthetic routes is hard work for chemists (Figure 1). On the one hand, thousands of reactions can possibly

[1] https://www.nigms.nih.gov/about/budget/CJs/Documents/cj2015.pdf

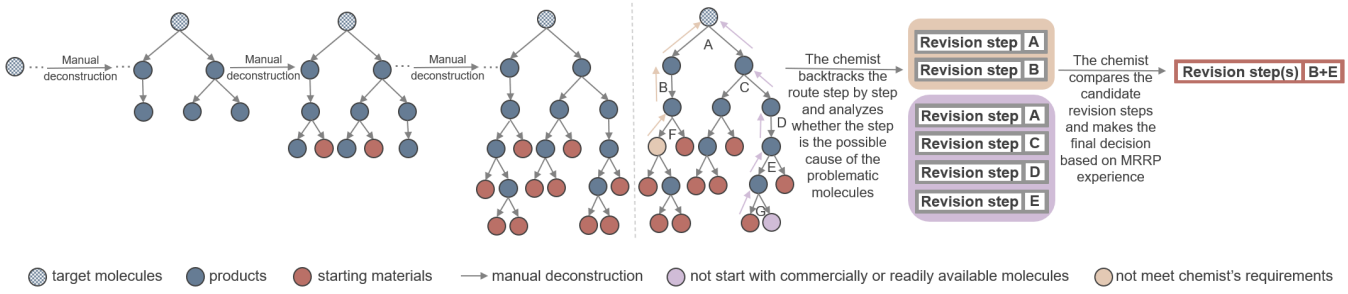
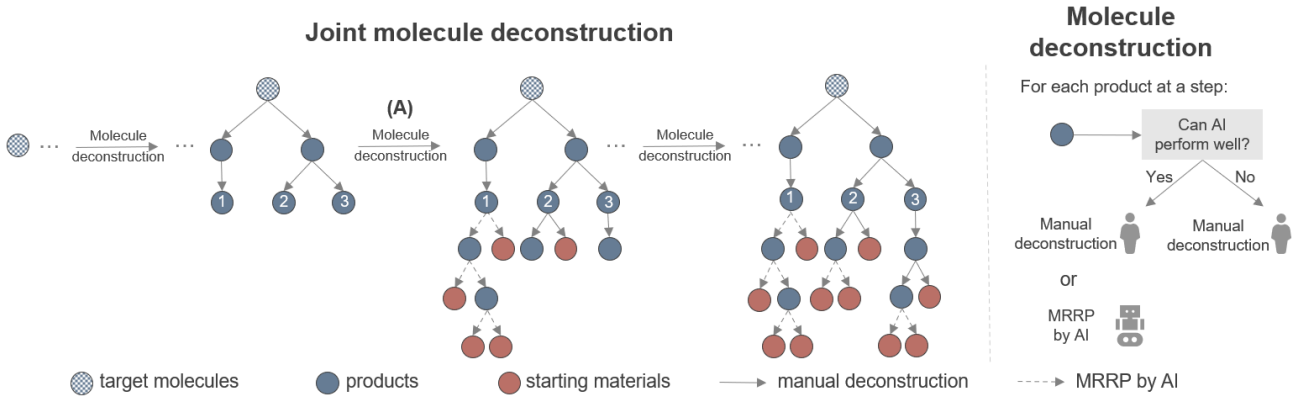
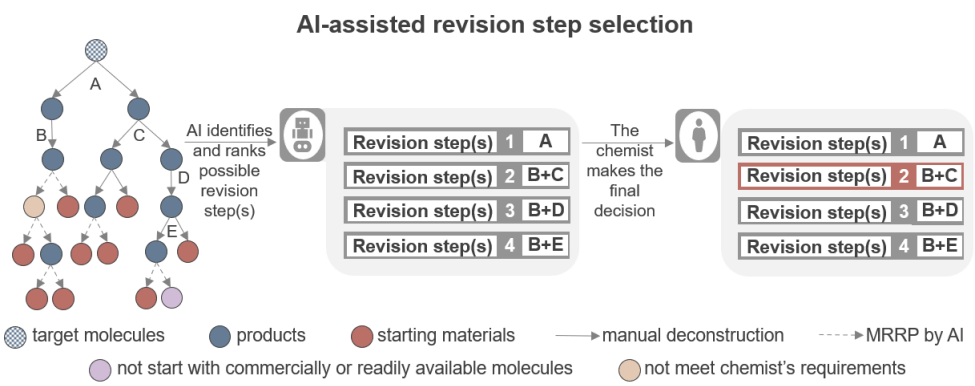


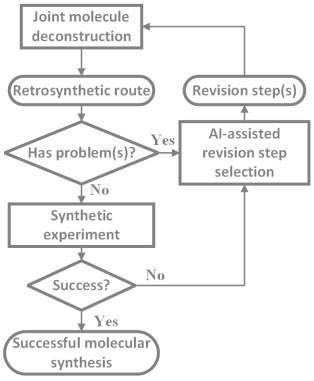
Figure 1: Human-only MRRP process



(a) The entire process of joint molecule deconstruction (left) and the task allocation method at each step of molecule deconstruction (right). Step (A) shows an example of: 1) product 1 is within the scope of AI's capability and the chemist chooses to adopt an AI-generated route; 2) product 2 is within the scope of AI's capability but the chemist chooses manual deconstruction at this step; 3) product 3 is beyond the scope of AI's capability.



(b) AI-assisted revision step selection



(c) The whole MRRP process with RetroLens

Figure 2: Illustration of the two forms of human-AI collaboration in RetroLens and the whole MRRP process with RetroLens.

produce the intended molecule at each step [75], which requires an enormous knowledge base to identify good candidates. Even experienced chemists need to search and integrate similar reaction precedents from existing works to find feasible candidate reactions [26]. Repeating such database and literature search at each step of MRRP makes the entire process time-consuming and tedious. It usually takes weeks to plan a retrosynthetic route for a single molecule [32]. On the other hand, since MRRP is a multi-step problem, the consequence of a particular deconstruction choice at a step may not be immediately obvious [72]. Chemists often fail to realize in time that they have made a wrong decision. When they find that the route under construction is unsuccessful after several steps, it is difficult, if not impossible, for chemists to locate the step(s) they made mistakes [29]. Furthermore, even if chemists can identify which reactions they choose may have caused the route failure, they still need to evaluate the trade-offs among a set of candidate revision steps with multiple objectives [82]. For instance, chemists not only hope to minimize the amount of revisions they have to

make to fix the problematic route but also tend to modify the steps they are relatively less confident about.

To assist chemists in MRRP, an increasing number of AI-powered methods have emerged [21, 34, 79, 100]. These AI models can automatically conduct MRRP based on large-scale knowledge bases and have been shown to offer more comprehensive plans faster than individual chemists [46]. However, these technologies are still limited in their capabilities to handle structurally complex molecules that are not well-studied [23] (e.g., natural products which usually have complex scaffolds [4]). This is because little related synthetic route data are available for AI model training to predict the retrosynthetic routes of complex molecules [8]. Also, the routes of such molecules are generally long, leading to explosive design space and planning complexity [15]. Furthermore, as AI algorithms only consider the theoretical synthesis possibility and cannot take into account practical factors (e.g., experimental conditions of individual researchers or labs), their recommended routes may not be feasible in subsequent chemical experimentation [84]. In these cases, the AI-generated routes, if any, would be unusable and the MRRP still needs to be carried out by human experts.

While AI cannot retrosynthetically analyze complex molecules, it may take the weight off the shoulders of chemists in the MRRP process when the target molecule becomes simple enough after a series of manual deconstructions. When the complexity of the product(s) (i.e., the molecule(s) to be deconstructed) at a step of MRRP falls within the scope of AI's capability, AI is likely to outperform chemists in handling these molecule(s) and can share the workload of the remaining steps of MRRP by automatically generating retrosynthetic routes of the products. Unfortunately, existing human-AI collaborative work often focuses on single-step problems (e.g., deceptive review detection) [51] and provides little guidance on when humans should hand tasks over to AI in such dynamic, multi-step problems. Also, there is still limited knowledge on how AI can help backtrack and locate possible causes of failure and suggest candidate revisions when humans make mistakes.

In this paper, we proposed RetroLens, a human-AI collaborative system (Figure 2) that facilitates chemists and AI to collaborate with each other on MRRP for complex molecules. Specifically, our proposed system integrates two forms of human-AI collaboration: joint action (i.e., a group works together like a single agent towards a shared goal; Figure 2(a)) [3, 106] and AI-assisted decision-making (i.e., AI informs human and human makes final decisions; Figure 2(b)) [31, 50]. First, when given a target molecule, chemists and AI have equal partnerships and would do the same type of tasks, namely analyzing and identifying the possible deconstructions of the input molecule – candidate reactions that can generate this molecule – at a given step of MRRP. The deconstruction tasks are allocated to humans or AI based on their capabilities and limitations. In particular, we communicated with six expert chemists to decide what kinds of molecules cannot be well analyzed retrosynthetically by existing AI solutions. If the product(s) of a step are such molecules, chemists should perform the deconstruction at this step manually. Otherwise, based on their preferences, chemists can choose manual deconstruction at the step or employ an AI model to automatically finish the MRRP of the product(s) at the current and remaining steps. Upon the completion of the initial retrosynthetic route, chemists may find that the output path has problem(s) (Figure 2(c)). For example, the route may not start with readily available molecules or chemists may think that some molecule(s) are not usable as they are too expensive or hazardous. To solve the problems in the route, AI plays an assistive role to chemists by suggesting candidate steps to fix. More specifically, we designed a pipeline for AI to 1) locate the possible manual step(s) that ultimately lead to each problematic molecule and 2) rank these potential revision steps based on multiple criteria that can be customized by chemists. Finally, human experts will make the final decision and perform the route revision according to their selected solution. We conducted a within-subjects user study with 18 chemistry researchers to evaluate the effectiveness of RetroLens. Results showed that compared with manual MRRP, joint molecule deconstruction helped chemists explore a broader design space. The two forms of human-AI collaboration together supported chemists to achieve faster MRRP, boost their confidence in the final planning, and lower their cognitive load.

In summary, the main contributions of this paper are:

(1) A Human-AI collaborative system, RetroLens, which integrates two forms of human-AI collaboration to support MRRP for complex molecules.
(2) A user study to evaluate how RetroLens impacts chemists' MRRP process, user experience, and perceptions of such human-AI collaborative system.
(3) Design considerations and opportunities for human-AI collaborative systems to better support MRRP and other multi-step problems.

2 RELATED-WORK

2.1 Multi-step Problem Solving

Multi-step problems have different definitions depending on the meaning of a "step". At a macro level, a step refers to an iteration of the whole process, and multi-step problems are those that require multiple iterations of solution construction to tackle. Previous work mainly helps solve such problems by prompting users on what needs work in the next iterations or helping humans to elicit guidance and feedback in the iterative process. For example, Zhang et al. designed a collaborative itinerary planning system, Mobi, which could check whether users' requirements are violated so that crowd participants could contribute to the iterative planning process appropriately based on current needs [103]. Similarly, Chilton et al. proposed a web application which provided actionable feedback on progress towards the global constraints on session creation to guide the iterative conference organization process [17]. Moreover, Louie et al. developed AI-steering tools enabling novices to direct the music generation model to iteratively create music [58].

At a micro level, a step refers to the smallest action or decision unit in the solution construction process which can be called correct or incorrect [90], and multi-step problems under this definition are those that require multiple units of decision-making within a single pass of solution construction (e.g., combinatorial games, programming). In this paper, we focus on this micro-level definition of multi-step problems which are underexplored in the human-AI collaboration area. Traditionally, solutions to such problems can be explored using heuristic search algorithms, such as Monte Carlo tree search (MCTS) [47, 98]. For example, Guillaume et al. applied

MCTS to enable computers to automatically play Go as human players. With the development of AI, some works began to apply deep learning (DL) models to solve this kind of multi-step problems. For instance, Silver et al. applied several machine learning (ML) algorithms (e.g., reinforcement learning) to MCTS and developed AlphaGo, which outperformed humans in the Go game [80]. However, in these scenarios, the number of solutions at each step is limited and certain, whereas there may be thousands of deconstruction choices at every step in MRRP. Thus, existing AI algorithms have not worked well on MRRP due to the complexity and large design space and require heavy involvement from chemists. An increasing amount of research has been done to provide problem solvers with post-hoc analysis on their problem-solving behavior and facilitate their operations (e.g., backtracking, reverting) during the multi-step problem solving process. For example, Wang et al. displayed different problem-solving patterns by visualizing users' sequences of intermediate steps of programming exercises in flow diagrams [94]. Kang et al. presented a prototype, AnalyticalInk, which modeled students' problem-solving directions as directed acyclic graphs and contrasted their solutions with the correct solution to tutor students to solve math problems [41]. Nevertheless, little work has investigated how to foresee the consequences after several steps stemming from the decisions that problem solvers have made.

2.2 Multi-criteria Decision Making (MCDM)

MCDM refers to making decisions based on multiple criteria [99]. Prior studies have proposed many algorithms to rank the candidates base on multiple criteria to facilitate MCDM. One of the most commonly used algorithms is simple additive weighting (SAW) [102]. It represents the importance of given criteria with weights, computes a SAW score based on the summation of the weighted criteria values, and ranks candidates accordingly. To support user-centric decision-making, interaction and visualization techniques have been widely applied to help users explore candidates based on multiple criteria. For example, Gratzl et al. designed LineUp, which allows users to interactively customize the criteria weights for ranking and uses stacked bar charts to illustrate the effects of different sets of criteria on the final ranking [30]. WeightLifter visualized the sensitivity of the criteria weights, enabling users to efficiently explore different weight spaces [68]. To facilitate MCDM in the real world, Weng et al. proposed a visual analytics system to assist people in searching for an ideal home based on multiple reachability-centric criteria [99]. To solve the problem that users may not have a clear understanding of the importance of some criteria to a decision, Wall et al. applied RankingSVM, which is a widely used ML algorithm for ranking tasks, to infer users' preferences based on their previous ranking behavior [92]. Additionally, some empirical research has been conducted to understand the difference between various MCDM supporting strategies. For instance, Kuhlman et al. investigated the impact of different user preference collection methods, including sub-list ranking, categorical binning, and pairwise comparisons, on users' decision-making process and found that the categorical binning technique enabled users to explore larger data space than the other two approaches [49]. Chan et al. studied the benefits and drawbacks of designer-led and optimization-driven designs in supporting the trade-offs between multiple design objectives. They found that with an optimizer, designers could obtain better solutions but felt a lower sense of engagement and expressiveness [14]. However, compared to scenarios in these studies (e.g., choosing ideal homes), MCDM for candidate revision steps in our work is more complicated. In particular, since MRRP is a multi-step problem, when choosing revision steps, chemists must consider not only the attributes of each individual candidate step but also their effects on the entire retrosynthetic route.

2.3 Human-AI Collaboration

Recent works have emphasized the importance of studying human-AI collaboration which describes the interactions between humans and AI. Three common forms of human-AI collaboration have been explored by researchers. First is AI-assisted decision-making, in which AI works as an assistant to provide decision recommendations while the final decision is made by humans [6, 104]. Several studies examined how people perceive and interact with the AI assistant. For example, Cai et al. investigated the onboarding needs of medical practitioners in their collaboration with diagnostic AI assistance [13]. Zhang et al. examined the effect of showing confidence value and explanation of AI prediction on human's perceived accuracy and trust in AI [104]. These studies pointed out that people often used AI assistance in unexpected ways and highlighted the importance of helping users understand AI's recommendations and correctly judge the reliability of AI.

Human-in-the-loop is an interactive training paradigm where the performance of AI can be improved based on human input. For example, Lee et al. designed a human-AI collaborative approach that allows therapists to identify weaknesses of the clinical decision-support AI by reviewing its output and provide feedback to improve AI [54]. Interactive ML also falls in this category, in which people are involved in the prediction process of AI models to improve the performance of AI [24, 25, 60, 61, 65, 86].

The third form of human-AI collaboration, joint action, refers to humans and AI working together towards a shared goal and acting as a single agent [3, 106]. Related work mainly studied the appropriate task allocation between humans and AI. For example, Lai et al. used deceptive review detection as a testbed, proposed a spectrum of allocation schemes ranging from full human agency to full automation, and divided work between human and AI accordingly [51, 52]. Mackeprang et al. designed a method based on a levels-of-automation framework to facilitate humans to find an optimal task allocation between humans and AI [63]. Lai et al. used content moderation as a case and proposed a novel paradigm of human-AI collaboration, conditional delegation. In this scheme, humans and AI together specify a trustworthy region within which AI's output is reliable, and the cases within this region are delegated to AI [50].

A common limitation of existing work is the focus on single-step problems, e.g., content moderation, where humans only need to make a final decision, not on multi-step problems, which require people to make a sequence of decisions to reach the final outcome. Moreover, although prior studies proposed various task allocation

mechanisms between humans and AI, they were mostly limited to providing guidance for the division of pre-defined tasks. Little work explored when humans and AI should hand over to one other in the dynamic environments of multi-step problems where pre-planned task delegation is impossible. To our knowledge, our work is the first study to extend human-AI collaboration to multi-step problem solving and integrate multiple forms of human-AI collaboration to support a single decision-making problem.

2.4 Retrosynthetic Route Planning (MRRP)

Existing AI-driven methods for MRRP can be categorized into template-based and template-free. Template-based methods plan retrosynthetic routes based on known reaction rules extracted from reactions reported in prior publications [7, 71, 76]. These approaches could effectively predict retrosynthetic routes for the molecules similar to those in the training data, but could not perform well on the prediction of unfamiliar reactions [79]. Template-free methods conduct MRRP utilizing DL models. Prior work typically treated deriving chemical reactions as a natural language processing problem, in which reactants and products at a step of a route were encoded as textual strings and the reaction can be regarded as a set of strings (reactants) being transformed into another (products) [42, 57, 87]. These AI-driven MRRP models have been limited to relatively simple target molecules so far, and no algorithm has been designed for complex targets, which have longer retrosynthetic routes and fewer reaction precedents that can be relied on [64].

3 PARTICIPATORY DESIGN PROCESS

To understand chemists' common practices and needs in MRRP for complex molecules and to form the design requirements and choices of our proposed human-AI collaborative system, we adopted an iterative participatory design approach.

3.1 Participants and Procedure

We brought six experts (E1-E6) in the field of chemistry to each phase of our iterative design process (Table 1). They all often perform MRRP in their research and have abundant MRRP experience. Our design process started with a two-hour semi-structured formative interview with each expert. We asked about their experiences of MRRP for complex molecules, including but not limited to their planning process and strategies, what kind of AI services, if any, they have used to analyze the retrosynthesis of molecules of interest, how they interacted with and perceived such services, and what challenges they faced during the planning process. Moreover, we asked the experts about their methods to solve the problems that cause MRRP to fail. The interview questions mainly covered how they locate the steps of deconstruction possibly resulting in route failure and how they choose effective revision methods. Through a thematic analysis [10] on experts' feedback by two authors of this paper, we derived a set of design requirements for potential human-AI solutions. Based on these requirements, we designed an initial version of a human-AI collaborative system supporting chemists and AI to design retrosynthetic routes for complex molecules together. We also proposed a pipeline to recommend revision steps for chemists when they find problems in the current routes. Then we carried out bi-weekly meetings with these experts for four months. We invited them to reflect on and co-design our system iteratively, ensuring that our updated implementation meets the design requirements and addresses new questions that emerged in the design process. Specifically, the experts contributed to the design of task allocation between chemists and AI, the pipeline to locate candidate revision steps, the criteria that chemists usually focused on to select ideal revision steps, and the interface of the system. We present the technical and design details of the final system in Section 4.

3.2 Insights from Formative Interviews

In this subsection, we summarize the key insights from the formative interviews.

All participants acknowledged that MRRP of complex molecules was challenging and time-consuming. E2, E5 and E6 reported that one main obstacle in this process was the difficulty in **searching and integrating the information associated with molecule deconstruction**. For each product at every step of MRRP, chemists need to comprehensively review the reaction precedents of the product from existing work to determine its deconstruction. Although current chemical data search engines (e.g., SciFinder[2]) can help collect related information from various online resources, it is still strenuous for chemists to repeatedly integrate and make sense of the large-scale search results. AI-powered MRRP supporting platforms (e.g., AiZynthFinder [28]) can alleviate these challenges to some extent since AI models can automatically predict molecule deconstruction based on large-scale reaction databases. But E3 and E5 reflected that they still needed to manually draw and input all products in the route, typically more than 10 or even 20, into the platforms one by one. E2 added that it was challenging for them to **distinguish what kinds of molecules should be delegated to AI to improve the efficiency of MRRP**. *"Sometimes I drew and input molecules to AI-powered platforms but finally found the molecules were too complex for AI to process, which was a waste of time"* (E2).

In addition, all experts reported that they often failed to **realize in time that they had made wrong deconstruction choices**. They pointed out that they usually could realize their mistakes only after several steps when they found that some resulting product(s) could not be further deconstructed or did not satisfy their requirements (e.g., the products could not be processed with their laboratory resources). For example, E4 shared a personal experience that he once chose an inappropriate deconstruction of a product at the second step of a retrosynthetic route, but he was not aware of it until he failed to decompose a product at the tenth step. This issue not only leads to a substantial waste of time and efforts but also increases the difficulty in backtracking and locating the problematic steps. Thus, the experts hoped AI to help them realize their wrong deconstruction decisions earlier.

Furthermore, all experts mentioned that the first rounds of MRRP were often unsuccessful due to the planning complexity. As a result, they need to iteratively revise the routes several times. They all complained about the difficulty of **locating the possible causes of the problems in retrosynthetic routes and deciding the**

[2]https://scifinder.cas.org

Table 1: Demographics of all participants in the participatory design process, including each participant's ID, gender, age, research area, research experience (number of years), and title.

ID	Gender	Age	Research Area	Exp.	Title
E1	Male	33	Medicinal Chemistry	9	University professor
E2	Male	31	Medicinal Chemistry	8	Postdoc
E3	Male	28	Organic Synthesis	8	Postdoc
E4	Male	27	Computer-aided Drug Design	5	PhD
E5	Male	28	Computer-aided Organic Synthesis	5	PhD
E6	Female	28	Computer-aided Organic Synthesis	5	PhD

ideal revision step(s) to solve the problems. For each step with a problematic product, chemists would backtrack the previous steps on the path to identify the reaction(s) that potentially caused the occurrence of the product and regard them as the candidate steps to revise. E1, E3 and E6 commented that if several problems existed in the route, identifying the revision steps that could solve all issues was particularly difficult, especially when the route was complex – *"long"* (E6) or *"containing many branches"* (E1). For example, E3 said that in addition to trying to solve the problems one by one, he would also consider whether he could tackle multiple problems by modifying a single revision step (e.g., revising the step at the intersection of all the branches containing undesired molecules) to save his time and efforts in route modification. Hence, they suggested that it would be helpful if our system could assist in identifying possible revision steps that can solve all problems in the route. Moreover, E3 and E5 brought up the need for AI to recommend revision steps based on different criteria concerning revision efficiency and route quality. They expressed that even if they could identify all candidate revision steps, they still found it challenging to select the ideal ones because they had to comprehensively consider multiple criteria.

3.3 Final Design Requirements Emerged from Iterative Design and Feedback Process

We present the final set of design requirements for a human-AI collaborative system to support MRRP for complex molecules which contains the initial design requirements derived from the formative interview and the new ones identified in the subsequent design iterations. In the design process, we kept using these design requirements in updating to guide our design decisions.

- **R1:** The system should reduce repetitive information search and integration for molecule deconstruction.
- **R2:** The system should identify the scope of AI's capability and allocate molecule deconstruction tasks to chemists or AI accordingly.
- **R3:** If chemists make wrong decisions during the manual molecule deconstruction process, the system should facilitate them to become aware of the problems early.
- **R4:** When facing failures in retrosynthetic routes, the system should automatically identify possible revision steps to fix the problems and rank the candidate solutions based on multiple criteria set by chemists.
- **R5:** The system should allow chemists to customize different components of the collaborative MRRP process based on their preferences and concerns, and AI should keep consistent customization with chemists' manual analysis.

4 RETROLENS: HUMAN-AI COLLABORATIVE SYSTEM FOR MRRP

By iteratively co-designing with the experts, we developed a human-AI collaborative system, RetroLens, to support chemists and AI to conduct MRRP for complex molecules together. In this section, we describe the system architecture of RetroLens (Figure 3), including the mechanisms of joint molecule deconstruction and the pipeline for AI-assisted revision step selection.

4.1 Joint Molecule Deconstruction

In each step of MRRP, AI first examines whether an input molecule is already commercially or readily available by checking its existence in a purchasable molecule database called ZINC [81]. If the molecule is included in ZINC, it does not need to be further retrosynthetically analyzed. Otherwise, it is regarded as a product to be deconstructed. The task of MRRP for this product is then allocated to the human or AI based on their capabilities and limitations.

By discussing with the experts (Section 3), we identified an empirical rule-of-thumb of product deconstruction task allocation between chemists and AI **(R2)**. That is, if the number of atoms of the largest ring of a molecule is greater than 12, it should be deconstructed manually by chemists, as AI models usually cannot perform well given such molecule complexity. Therefore, AI would first calculate the number of atoms of the largest ring of each input product at the current step using RDKit [53]. The products are represented in textual strings composed based on the simplified molecular-input line-entry system (SMILES) [96]. For the products whose largest ring contains no more than 12 atoms, AI applies the API of IBM RXN[3] – one of the most widely used AI-powered MRRP tools – to predict their retrosynthetic routes based on their SMILES. IBM RXN allows users to specify a set of constraints (Supplementary material) on the output retrosynthetic routes so that the generated routes will be consistent with chemists' manual deconstructions **(R5)**. Still, the routes generated by AI may fail if they do not start with readily available molecules or do not meet chemists' constraints **(R3)**. After reviewing the AI-generated routes of an input product, chemists can either directly adopt one of them and incorporate it into the final plan **(R1)**, or reject all AI results and perform the deconstruction by themselves. For other input products that are deemed complex, AI would not carry out the MRRP, handing them to chemists for manual processing.

4.1.1 The Revision Step Identification Module. Following the feedback from the experts (Section 3), we first used RDKit to check

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

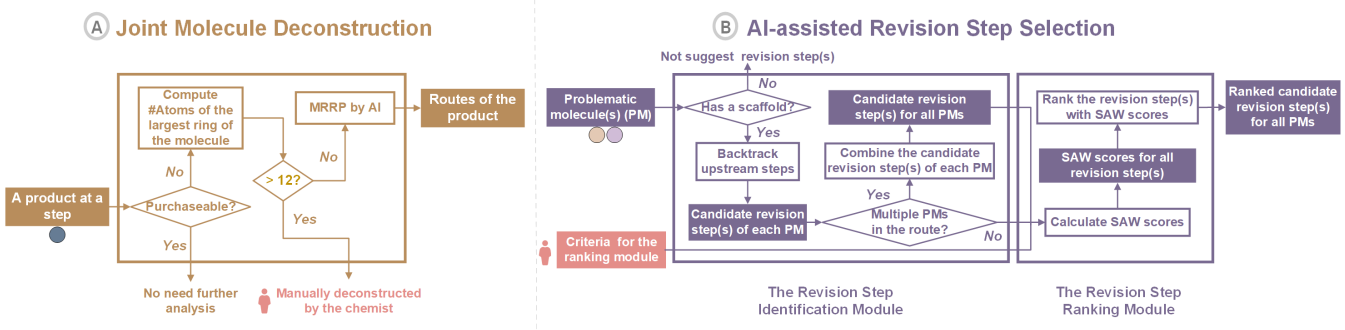


Figure 3: The system architecture of RetroLens: (A) the mechanisms of joint molecule deconstruction; (B) the pipeline for AI-assisted revision step selection module.

Algorithm 1 Algorithm to find candidate revision steps for a problematic molecule

Input: *molecule*: A problematic molecule that contains a scaffold
Output: *moleculeCandidateList*: A list of candidate revision steps of the input problematic molecule

```
procedure GetMoleculeCandidates(molecule)
    moleculeCandidateList ← []
    ancestor ← molecule.parent
    while ancestor ≠ null do
        scaffoldHasChanged ← true
        for all child ∈ ancestor.children do
            if child.scaffold == ancestor.scaffold then
                scaffoldHasChanged ← false
                break
        if ancestor.isHandledManually AND
           scaffoldHasChanged then
            if ancestor.isProblematic then
                moleculeCandidateList ← []
            else
                moleculeCandidateList.append(ancestor)
        ancestor ← ancestor.parent
    return moleculeCandidateList
```

whether a problematic molecule identified in the molecule deconstruction stage has a scaffold (i.e., the core structure of a molecule [74]). If not, RetroLens would not suggest any revision steps as such a problem can usually be solved by directly modifying the reaction that yielded the problematic molecule.

In other words, our revision step identification module mainly focuses on the problematic molecules with scaffolds (Algorithm 1) **(R4)**. In these cases, RetroLens would backtrack the retrosynthetic route from a failure position to locate the possible upstream steps that lead to it. Since AI can explore and return all possible retrosynthetic routes of an input molecule, if none of the AI results is acceptable, it is likely that the root issue occurs in the steps manually processed by humans. Therefore, RetroLens only inspects the manual steps for revision candidacy. The experts suggested that a candidate step for revision must satisfy two conditions (Section 3): 1) the step is a manual step located before where the problematic molecule appears; 2) the deconstruction reaction at that step involves modification of the structural scaffold. Thus, the identification module first narrows down the manual steps based on 1), and then extracts the scaffolds of the product and the reactant(s) (i.e., the molecules derived by deconstructing the product) of each remaining step using RDKit. If all scaffolds of the reactant(s) are different from the scaffold of the product, the corresponding step would be regarded as a possible site leading to the problematic molecule and be added to the candidate revision step pool of this problematic molecule.

If there are multiple problematic molecules in a retrosynthetic route, the identification module further pinpoints the potential revision step(s) to solve all problems together by combining the candidate revision steps of each molecule of concern (Algorithm 2). Different problematic molecules may share common candidate steps, because these steps are at the intersection of the branches containing those problematic molecules. Modifying such a step can influence the production of these molecules. They thus are directly regarded by AI as candidate revision steps. For other revision steps that are not shared by individual problematic molecules, the identification module extracts all possible combinations of them so that each combination may collectively solve the problems in the route.

4.1.2 The Revision Step Ranking Module. Once the identification module obtains a pool of candidate revision steps for each problematic molecule, the ranking module would rank these candidates based on a list of criteria derived from the literature review [1, 5, 19, 37, 84] and the feedback from the experts (Section 3) **(R4)**.

RetroLens allows chemists to select the criteria they care about for the ranking module from the following list **(R5)**.

- **Amount of revision:** the number of revisions chemists have to make to fix the problems in the retrosynthetic route. Modifying a step that possibly causes the problems may also lead to the revisions of the subsequent steps. Chemists are usually concerned about how much change each revision step will cause on the original retrosynthetic route and hope to minimize the amount of revisions they make. For each revision step, we defined its amount of revisions as the number of molecules in the subsequent steps of the revision step.

Algorithm 2 Recursive function to generate all candidate revision step combinations

Input: *molecule*: The target molecule of a retrosynthetic route
candidateList: a list containing all candidate revision steps of all problematic molecules that contain scaffolds
Output: *candidateCombList*: A list containing all candidate revision step combinations of the retrosynthetic route rooted at *molecule*

```
procedure CandidateComb(molecule, candidateList)
    candidateCombList ← []
    if molecule ∈ candidateList then
        for all child ∈ molecule.children do
            if child ∈ candidateList then
                childCandidateCombList
                    ← CandidateComb(child, candidateList)
                candidateCombList
                    .append(childCandidateCombList)
            else if child.isProblematic then
                candidateCombList ← []
                break
        candidateCombList
            ← candidateCombList × candidateCombList
        candidateCombList.append(molecule)
    return candidateCombList
```

- **Reaction confidence:** the chemist's confidence in the deconstruction in the revision step which is chosen in the molecule deconstruction stage. Among all the candidate revision steps, chemists tend to modify the steps they are relatively less confident about. By discussing with the experts, we found that the confidence in a manual step is estimated based on similar reaction precedents in existing work. Specifically, for each reactant R_i ($i = 1, 2, \cdots, n$) at the candidate revision step S, we first searched the reactants $\left\{A_i^{(1)}, A_i^{(2)}, \cdots, A_i^{(j)}\right\}$ with a similarity greater than 0.8 to R_i and obtain the number of them $\left\{N\left(A_i^{(1)}\right), N\left(A_i^{(2)}\right), \cdots, N\left(A_i^{(j)}\right)\right\}$ in the United States Patent Office (USPTO), which is one of the most widely adopted publicly available reaction dataset containing approximately one million reactions [15, 77]. We defined $p(R_i)$ as
$$p(R_i) = \frac{\sum_{x=1}^{j} \text{Similarity}\left(R_i, A_i^{(x)}\right) \times N\left(A_i^{(x)}\right)}{\sum_{x=1}^{j} N\left(A_i^{(x)}\right)}$$
to represent the possibility of the existence of R_i obtained by the deconstruction at step S. Then the confidence in the deconstruction choice in step S is calculated as the weighted average of $\{p(R_1), p(R_2), \cdots, p(R_n)\}$ with the relative size of R_i over that of $\{R_1, R_2, \cdots, R_n\}$ as the weight. That is
$$\text{Confidence}(S) = \sum_{i=1}^{n} p(R_i) \times \frac{a(R_i)}{\sum_{m=1}^{n} a(R_m)}$$
where $a(R_i)$ represents the number of atoms in R_i.
- **Complexity reduction:** the reduction of the complexity of the step's reactants compared to its product. In MRRP, the steps with more reduction of complexity are better and are less expected to be revised [84]. Therefore, we first calculated SCScore, an effective measure of molecular complexity in MRRP, which ranges from 1 to 5 with 5 as the highest complexity [19], of the product P and reactants $\{R_1, R_2, \cdots, R_n\}$ at the candidate revision step S. To keep consistent with other criteria that the smaller the value of the criteria of a step is, the more the step is preferred to be revised, we defined a simplicity score similar to existing work [73]:
$$\text{Simplicity}(R_i) = 1 - \frac{\text{SCScore}(R_i) - 1}{4}$$
and defined the complexity reduction of the step S as
$$\text{Complexity\_Reduction}(S) = \frac{\prod_{i=1}^{n} \text{Simplicity}(R_i)}{\text{Simplicity}(P)}.$$
- **Convergence:** the number of reactants in the step and their relative sizes. The retrosynthetic route should be as branched as possible rather than linear [1, 5]. Thus, the more reactants there are and the closer their relative sizes are, the more convergent the step will be and the less the step should be modified. According to [37], we computed the convergence of a candidate revision step S as follows:
$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \frac{a(P)}{n} - a(R_i)$$
$$\text{Convergence}(S) = \frac{1}{1 + \text{MAE}}$$
where $a(P)$ and $a(R_i)$ represent the number of atoms in the product and reactants of step S, respectively, and MAE is the mean absolute error.
- **Associated branch confidence:** the confidence in the associated branches (i.e., the sub-routes) under a step other than the branches containing the problematic molecule(s). When chemists choose to revise a step to fix a problematic molecule in the retrosynthetic route, in addition to the branch containing the problematic molecule, other associated branches will also be revised. Hence, chemists hope that the revision steps they choose affect the associated branches they are confident about as little as possible. The confidence of an associated branch is calculated by multiplying the confidence of all the steps contained in the branch. Specifically, the confidence of an AI-generated step is returned by IBM RXN, and the confidence of a manual step is computed as discussed in **Reaction confidence** criteria.

If there is one problematic molecule in the retrosynthetic route, we ranked its corresponding revision steps using the SAW algorithm, which is extensively applied in various MCDM scenarios [68, 83, 97, 105]. Suppose k criteria are selected by the chemist, we denoted the y-th criterion value of the candidate revision step S by $c_S^{(y)}$. Then, the SAW score of step S is computed with
$$\text{SAW}(S) = \sum_{y=1}^{k} w_y \times c_S^{(y)}$$
where w_y is the weight assigned to the y-th criteria by the chemist based on his/her preference, the sum of weights $\sum_y w_y = 1$, and

the value of SAW(S) ranges from 0 to 1. We therefore sort the SAW scores of candidate revision steps in ascending order.

If there are multiple problematic molecules in the route, AI computes the SAW scores of the candidate fixes based on the SAW scores of individual revision steps involved. Specifically, as a step S that can simultaneously affect several molecules (Section 4.1.1) may have different SAW scores as the candidate revision step of different problematic molecules, the ranking module defines its combination SAW score as the weighted average of these SAW scores. The weights are based on the number of molecules in the branches of concern under step S. For each of the other revision step combinations, the resulting SAW score is the sum of the SAW scores of every revision step in the combination. Note that the SAW scores of the combinations may be greater than 1 and we did not normalize them to guarantee the explainability of AI. In this way, users can easily understand that they need to make efforts to revise multiple steps if they choose this revision plan. AI displays these suggested plans for combined problems in the ascending order of their SAW scores.

5 EXPERIMENTAL DESIGN

In this section, we present our within-subjects user study design to evaluate whether and how RetroLens would influence chemists' MRRP process.

5.1 Experimental Conditions

As discussed in Section 4, our system provides both joint molecule deconstruction and AI-assisted revision step selection. To investigate their effects on MRRP, we compared the following three conditions:

- **Manual condition:** According to the interviews with experts (Section 3), it is a common practice for chemists to manually plan retrosynthetic routes based on their own knowledge and online search. Hence, we used the manual condition as the baseline. In this condition, participants were allowed to work with any search engines (e.g., SciFinder[4], Google Scholar) or AI-powered MRRP supporting systems they usually use in their routine practices.
- **Joint molecule deconstruction:** In this condition, only the joint molecule deconstruction component is provided to participants. Participants were also allowed to use any online search tools as in the manual condition.
- **Joint molecule deconstruction + AI-assisted revision step selection:** In addition to the online search in the manual condition, participants in this condition have access to both the joint molecule deconstruction and AI-assisted revision step selection features of RetroLens.

In the rest of this paper, we refer to these conditions as the "manual condition", "AI condition", and "AI^2 condition", respectively.

5.2 Experimental Website Design

We developed two separate experimental websites for the AI and AI^2 conditions. Both websites started with a canvas for users to draw or upload their target molecules. This canvas was developed based on an open-source web-based chemical structure editor, Ketcher[5], which contains common features of the canvas in the supporting tools (e.g., SciFinder) that chemists usually use in their research. When a user submits a target molecule through the canvas for the first time, a pop-up window would be shown asking the user to specify the constraints for retrosynthetic routes. In the default setting, no constraints are selected. The reactant(s) derived at each step of users' manual deconstruction are also inputted into our system using this canvas. Once users submit molecule(s) through the canvas, AI would automatically analyze their complexity. If AI could predict possible retrosynthetic routes of the molecule(s), the route with the highest confidence value would be displayed under the corresponding molecule(s) on the main page. Otherwise, only the molecule(s) inputted by the users would be shown.

The main page of both websites shows the retrosynthetic route jointly composed by a human and AI in a tree-like format (Figure 4). Such a design models after IBM RXN and is generally familiar to our users. The parts of the route processed by the chemist (thick) and AI (thin) are distinguished by the thickness of lines connecting the nodes in the tree. The molecules that are commercially or readily available are enclosed by a green frame, and those that cannot be further deconstructed by AI are in red. There are two buttons under each of the input molecules in the route: an edit reaction button (Figure 4(a)) and an AI-generated routes button (Figure 4(b)). Clicking on the edit reaction button would delete all the subsequent retrosynthetic steps, if any, under the molecule, and take users to the canvas to manually compose reactants from deconstructing this molecule. Hovering over the AI-generated routes button pops up a tooltip showing the number of alternative routes that AI has compiled for the molecule. If users click this button, a sidebar would appear on the right side of the main page (Figure 4(c)), listing all alternative routes and their corresponding confidence values. If the user selects a particular alternative route, it would replace the one currently displayed on the main page under the associated molecule. Note that as RetroLens does not allow modification of AI-predicted routes for simplicity, these two buttons are not available under the molecules in the parts of routes generated by AI. This is because our goal in this work is not to develop a full-fledged system but a research prototype to explore whether and how humans and AI can collaboratively conduct MRRP. We acknowledge that this may limit users' flexibility in the MRRP process and users may hope to edit the AI-returned results. Future work could enable users to revise AI's output for more customized MRRP. A reconfigure constraints button (Figure 4(d)) is in the upper right corner of the main page, allowing users to reset their constraints for retrosynthetic routes at any time.

The only difference between the two websites is that the one for AI^2 condition allows users to mark their undesired molecule(s) (enclosed in a blue frame) by clicking the molecule(s) in the route. The main page of AI^2 condition also provides an additional button – revise retrosynthetic routes (Figure 4(e)) – in the upper right corner for AI-assisted revision step selection. Pressing this button shows a popup box asking users to specify the criteria they care about for choosing revision steps and input their weights. If no option is selected, all criteria will be used and have the same weight by

[4]https://scifinder.cas.org

[5]https://github.com/epam/ketcher

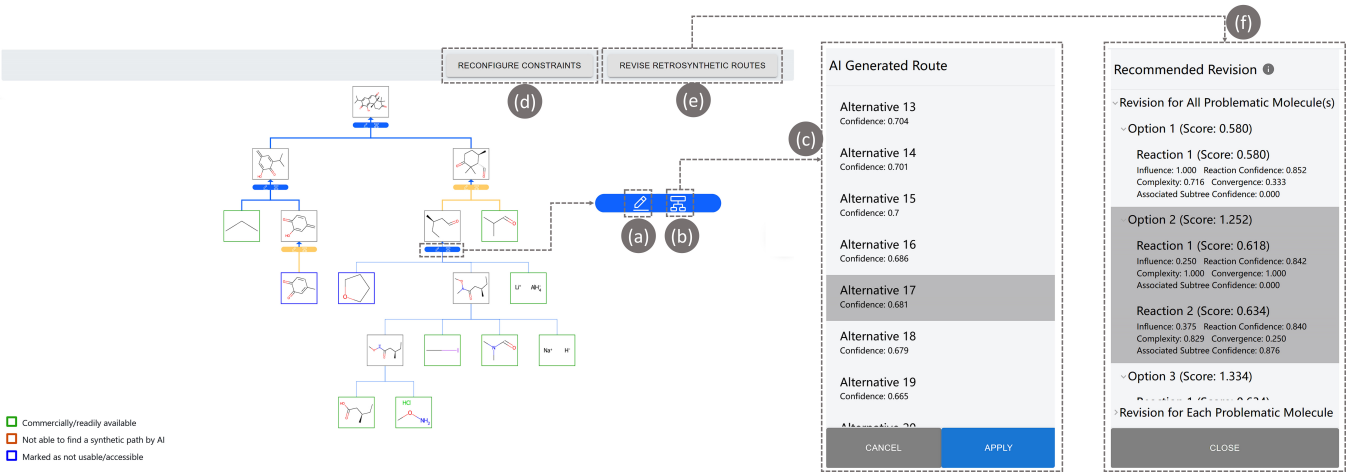


Figure 4: Screenshots of the main page of the experimental website for AI^2 condition. (a) edit reaction button. (b) AI generated routes button. (c) AI generated alternative routes sidebar. (d) reconfigure constraints button. (e) revise retrosynthetic routes button. (f) revision steps sidebar.

Table 2: Demographics of all participants in the within-subjects study, including each participant's ID, gender, age, research area, research experience (number of years), and MRRP skills.

ID	Gender	Age	Research Area	Exp.	MRRP Sills
1	M	28	Organic Chemistry	7	Expertise
2	M	21	Organic Chemistry	1	Novice
3	F	25	Medicinal Chemistry	3	Knowledgeable
4	M	32	Medicinal Chemistry	9	Expertise
5	M	27	Polymer Chemistry	4	Knowledgeable
6	Prefer not to say	26	Organic Chemistry	3	Knowledgeable
7	F	25	Organic Chemistry	4	Knowledgeable
8	F	25	Medicinal Chemistry	4	Knowledgeable
9	F	24	Chemical Engineering	3	Knowledgeable
10	F	24	Medicinal Chemistry	1	Novice
11	M	22	Medicinal Chemistry	1	Novice
12	M	23	Medicinal Chemistry	1	Novice
13	F	24	Medicinal Chemistry	1	Novice
14	F	23	Medicinal Chemistry	2	Novice
15	F	28	Medicinal Chemistry	5	Knowledgeable
16	F	25	Medicinal Chemistry	3	Knowledgeable
17	F	24	Medicinal Chemistry	2	Novice
18	M	24	Medicinal Chemistry	1	Novice

default. After the confirm button in the pop-up is clicked, a sidebar appears on the right side of the main page (Figure 4(f)) and presents both revision step combinations for all problematic molecules and separate revision steps for individual problematic molecules. The specific values of the criteria of each candidate revision step are also shown to users. When clicking on a revision plan (such as the "Option 2" in Figure 4(f)), the corresponding revision step(s) in the route would be highlighted in yellow.

5.3 Participants and Procedure

We recruited 18 participants with diverse expertise in chemistry through online advertising and word-of-mouth. To be eligible to participate in the study, participants were asked to self-report their chemical research experience and familiarity with MRRP for complex molecules (see details in Table 2).

We discussed with the experts (Section 3) and designed three MRRP tasks for our within-subjects user study using three complex molecules that are not well-studied as target molecules:

- *T1:* Design a multi-step retrosynthetic route for the molecule *3-oxoisotaxodione*.
- *T2:* Design a multi-step retrosynthetic route for the molecule *Fimbricalyxoid A*.
- *T3:* Design a multi-step retrosynthetic route for the molecule *Impatien A*.

These three molecules are all natural product molecules with promising anticancer effects. They also have similar synthetic complexity and have a similar amount of relevant literature available online.

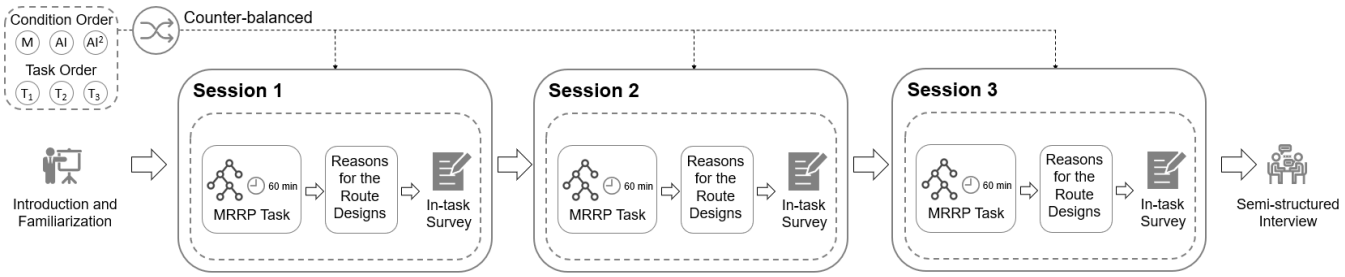


Figure 5: Procedure of the within-subjects user study. Each participant needs to complete the three MRRP tasks (*T1, T2, T3*) under different conditions (M: manual condition, AI: AI condition, AI^2: AI^2 condition) in a counter-balanced order.

Figure 5 shows the procedure of the experiment. After obtaining participants' consent, we first carefully introduced the two experimental websites to the participants and gave them 15 minutes to get familiar with the interfaces. Then, we invited each participant to complete the three tasks separately under the three conditions. We counterbalanced the task assignment and the order of the three conditions to alleviate the potential order effect. Each participant were given 60 minutes for each task. We recorded the video of each user study session during the study. At the end of each session, we asked the participants to write down their reasons for their final route designs and fill out an in-task questionnaire (See Appendix A). The questionnaire contains three parts: 1) user confidence in the final retrosynthetic routes [101]; 2) user cognitive load during the tasks, measured using the NASA Task Load Index (NASA-TLX) [35]; and 3) user perceptions of RetroLens [59, 78]. Upon the completion of the three sessions, we further conducted a semi-structured interview with the participants about their experience, attitudes, and concerns towards RetroLens and collaborating with AI in their research.

6 RESULTS

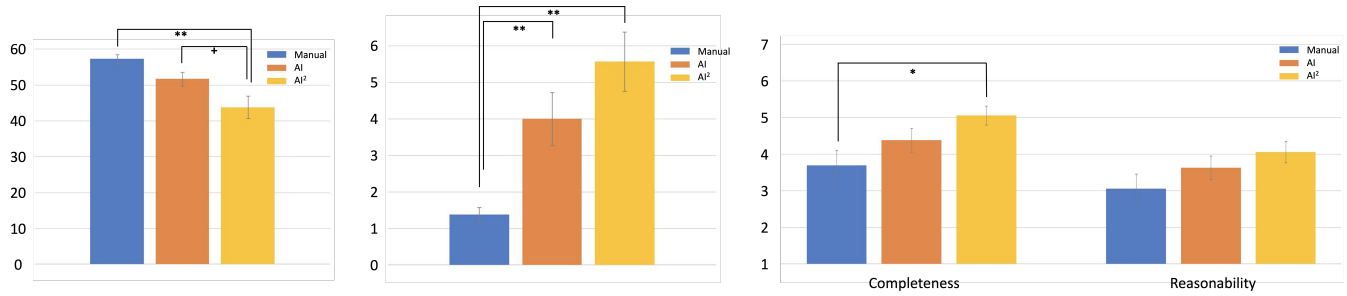
To probe the impact of RetroLens on chemists during the MRRP process, we conducted a series of statistical analyses on our user study data. We first run the Shapiro-Wilk test on all quantitative measures (i.e., user behavior data coded from video recordings and responses on questionnaires from participants) and the results show that they all do not conform to normality ($p < 0.05$). Therefore, we performed the Friedman test and post-hoc Wilcoxon signed-rank test with Bonferroni correction [55] to assess the difference in the participants' experience regarding each measure across the three conditions (i.e., manual, AI, and AI^2). Finally, two authors of this paper conducted a thematic analysis [10] on the transcripts of the semi-structured interview and identified key themes in participants' feedback. In this section, we present the quantitative results concerning user performance, user confidence and cognitive load, and qualitative findings from the interview.

6.1 User Performance

To examine how well RetroLens helps users with MRRP, we conducted a statistical analysis on participants' performance in the user study, including the task completion time, the number of retrosynthetic routes explored, and the quality of their final outcome.

6.1.1 Two forms of human-AI collaboration save MRRP Time. We compared users' task completion time in three conditions using the Friedman test and found a significant difference ($\chi^2(2) = 13.37$, $p < .01$). The pairwise comparisons showed that participants in the AI^2 condition spent significantly ($p < .01$) and marginally significantly ($p = .06$) less time completing the MRRP task than when they were in the manual condition and AI condition, respectively (Figure 6(a)). However, the difference between the manual condition and AI condition is not significant. Despite this, participants reflected in the post-study interview that both forms of human-AI collaboration in RetroLens contributed to saving their planning time. 12 (of 18) participants mentioned that with joint molecule deconstruction, they **do not need to process the products that can be handled by AI**. The time for manually inputting these products into other search platforms and mentally integrating and transforming the reactions found into deconstruction plans is thus reduced. P8 added that RetroLens is particularly helpful in this way because it **can examine whether a molecule could be handled by AI**. *"In my research routine, I always only find out that the existing AI-powered platforms cannot process the molecules I inquire after I have spent a lot of time drawing the molecules and waiting for their analysis results"*. Seven participants reported that this module also enables them to **detect mistakes in their deconstruction decisions early** when seeing that all AI-generated subsequent routes were infeasible. In addition, 12 participants think that the AI^2 condition system allows them to **locate the possible steps leading to the problems in a route faster**. P16 explained, *"To find out how to solve the problems [in the route], I usually need to check my previous deconstruction choices one by one, which is really time-consuming. It is always difficult for me to determine whether I should backtrack and revise some earlier steps or design a completely new route from scratch. I cannot tell which would require fewer efforts. In contrast, I can easily make decisions based on AI-recommended solutions"*.

6.1.2 Broader retrosynthetic route design space is covered with the help of AI. To assess the effectiveness of our system in supporting users to explore retrosynthetic route design space, we counted the number of retrosynthetic routes each participant explored in each condition. As shown in Figure 6(b), participants covered significantly bigger design space in both the AI condition ($p < .01$) and the AI^2 condition ($p < .01$) than in the manual condition ($\chi^2(2) = 22.81$, $p < .01$); no statistical difference is found between the two AI-related conditions. In the manual condition, chemists typically



(a) Completion time (min)

(b) The number of retrosynthetic routes explored

(c) Final retrosynthetic routes

Figure 6: The results of participants' performance during the MRRP process in terms of task completion time, the number of retrosynthetic routes each participant explored and final retrosynthetic routes (+ : .05 < p < .1, ∗ : p < .05, ∗∗ : p < .01).

take a greedy, depth-first approach, keeping deconstructing a target molecule with the optimal reaction they could think of in each step until they finally obtain a seemly feasible route. In this way, chemists often can only come up with one plan in the theoretical MRRP stage and then try to synthesize the target molecule accordingly in the laboratory (P14). They would turn to explore other alternative routes only if the synthesis experimentation fails. In comparison, in the other two conditions, participants are more aware that a retrosynthetic route that seems viable at the current step may run into problems later (Section 6.1.1), which **reminded them to explore other alternatives before the practical experimentation**. Moreover, three participants indicated that **the cost of failure became lower with the assistance of RetroLens, making them willing to explore more routes**. P16 said, *"I would first use RetroLens to explore as many routes as possible. Even if I finally learn that the AI-returned results fail because of technical limitations and I need to manually plan the whole route by myself, it would not cost me much extra time"*. Also, AI can recommend and rank different revision steps in case all the routes under investigation fail. With such information, participants tended to **feel more confident about the change of direction, if needed**. *"If AI tells me that creating a new route would cost the least effort to solve the problems, I would not hesitate to explore a new one"* (P12). On the contrary, without RetroLens, it is hard for them to abandon the current plan even though they have found problems in it (P6), due to both the Sunk Cost Fallacy [33] and the lack of clue regarding where to start over.

6.1.3 Two forms of human-AI collaboration together prompt more complete MRRP. To investigate the effectiveness of RetroLens in helping identify high-quality retrosynthetic routes, we invited two experts with over 10 years of experience in chemistry research to evaluate the final routes designed by the participants. We chose completeness (i.e., whether participants' final retrosynthetic routes were complete) and perceived reasonableness (i.e., whether participants' final retrosynthetic routes fulfil known reaction rules, the experts' experimental chemical synthesis experience, and common route evaluation criteria such as reaction selectivity and economics [95]) as the criteria for decision quality assessment after consulting the experts. Specifically, each participant's final outcomes of the planning tasks and the corresponding justifications in the three study conditions were given to the experts. We randomized the order of presentation, and both experts were blind to the conditions. Then we asked the experts to rate the completeness and reasonableness of each final decision on a 7-point Likert scale (1 - not complete/reasonable at all, 7 - extremely complete/reasonable). The results show that experts perceived route completeness in the three conditions to be significantly different ($\chi^2(2) = 22.81$, p < .01), but not so regarding reasonableness. The participants created significantly more complete routes in the AI^2 condition than in the manual condition (p < .05), while we did not find significant differences between other pairs of conditions (Figure 6(c)). The reason for the relatively low reasonableness across all conditions may be that the experts tend to have reservations when retrosynthetic routes have not yet been empirically verified, just as the situation in the scope of this paper. The experts could only provide rather conservative projections of the reasonableness of the routes. Despite this, in the post-study interview, some participants reported that RetroLens helped them design more reasonable retrosynthetic routes. For instance, P5 and P12 reflected that RetroLens provided more theoretically feasible routes for a product than they could have thought of by themselves.

6.2 User Confidence and Cognitive Load

To inspect users' experience in the MRRP process, we statistically analyzed their confidence in the final route constructed and their perceived cognitive load during the tasks based on their responses in the post-study questionnaires.

6.2.1 More confident in the final decision with AI's support. Since chemists' confidence in their designed retrosynthetic routes is an important factor that directly influences whether they would proceed to the experiment stage and synthesize molecules based on the routes, we compared participants' confidence in their final decisions in the three conditions. This measurement could help us to explore whether RetroLens can facilitate chemists to find retrosynthetic routes that they are inclined to test in the lab earlier, thereby accelerating the entire MRRP process. Compared to the manual condition (Figure 7), participants reported to be significantly more confident (p < .05) in their final decisions in the AI condition ($\chi^2(2) = 24.53$, p < .01). The pairwise comparisons also show that their confidence ratings in the AI^2 condition are significantly higher

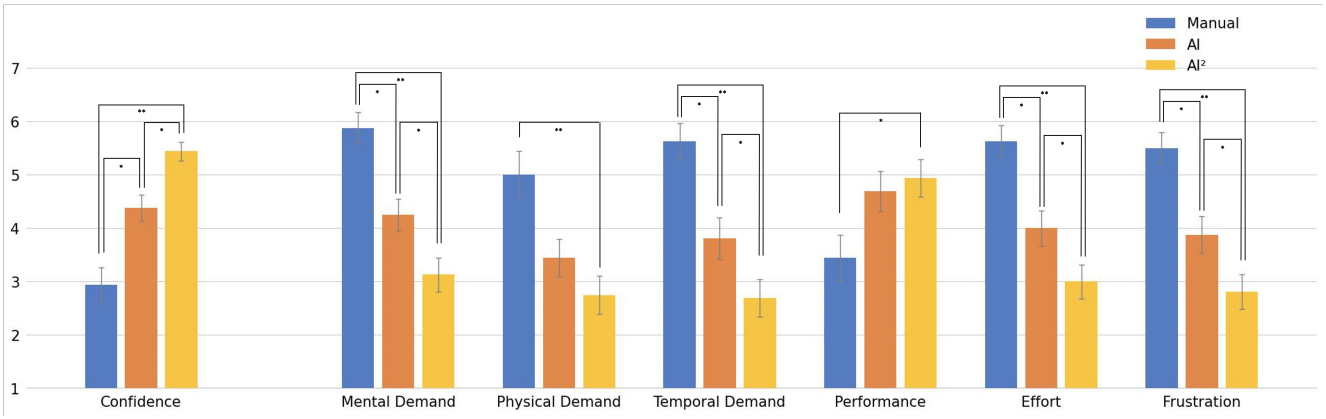


Figure 7: Means and standard errors of the participants' confidence in their final retrosynthetic routes (left) and cognitive load in MRRP process (right) on a 7-point Likert scale ($*$: p < .05, $**$: p < .01).

than those in the manual condition (p < .01) and the AI condition (p < .05). From the interview, we learned that the participants with different MRRP experiences may have different reasons for having higher confidence in their designed routes when using RetroLens. Novice participants mainly (6/8) suggested that RetroLens enabled them to **have a clearer and more objective understanding of the output routes**. In particular, the confidence of AI-generated routes in the deconstruction stage implies the feasibility and success rate of the routes. Interestingly, P10 reflected that the AI-assisted revision step selection function can also help quantify the quality of the retrosynthetic routes. In the manual condition, all participants complained that they usually cannot find related works that study the exact same molecules as their targets. Thus, although they have carefully studied the precedents of similar molecules to infer possible deconstruction for their targets, they were still not sure whether the routes are feasible and whether the reactants generated from their deconstruction indeed exist. In contrast, the specific values of the assessment criteria (Section 4.1.2) for ranking candidate revision steps shown in our system could provide quantitative evidence of the rationality of each manual step. For example, to compute the reaction confidence of a manual step, RetroLens would estimate the possibility of the existence of the reactants and the success rate of this step. The participants with more MRRP experience reported that the **AI-generated routes allowed them to reconfirm their preliminary ideas of retrosynthetic route design, if any**, and our system thus increased their confidence. *"If the routes recommended by AI generally align with my deconstruction ideas, I would feel more confident in the final route design"* (P4).

6.2.2 Less cognitive load perceived when collaborating with AI. All pairwise comparisons between the three study conditions reveal significant differences in the Mental Demand, Temporal Demand, Effort, and Frustration dimensions of cognitive load (Figure 7). The participants experienced significantly less cognitive load in these four dimensions in the AI^2 condition than in both the manual condition (p < .01) and AI condition (p < .05). These dimensions of cognitive load also received significantly lower ratings in the AI condition than in the manual condition (p < .05). As for the other two dimensions of cognitive load, the AI^2 condition is perceived to be significantly less Physically demanding (p < .01) and with better Performance (p < .05) than the manual condition, while the differences between other pairs of conditions are not significant (Figure 7). 11 participants explained that they perceived less cognitive load when using RetroLens because **RetroLens saved a lot of their effort in designing retrosynthetic routes of the products which can be performed by AI**. For these products, chemists no longer need to retrieve relevant information from their existing knowledge base. Also, they do not have to search and make sense of similar reaction precedents in the literature and store such large-scale new information in their working memory to infer possible deconstruction of the products on their own. Additionally, six participants stated that in the manual condition, they overlooked some criteria which should be considered when deciding how to fix the problems in the retrosynthetic routes. P2 complained, *"this caused me to make more mistakes in the revision process"*. In contrast, they reported that **RetroLens helped them identify possible revision steps more accurately and comprehensively, enabling them to get theoretically feasible routes through fewer iterations** and thus significantly reduced their cognitive load.

6.3 Qualitative Feedback on User Experience and Concerns towards RetroLens

Through the post-study interview, we investigated how participants perceived and used RetroLens. We also discuss their concerns about collaborating with AI during the MRRP process (Table 3).

6.3.1 Perception towards RetroLens. In general, the subjective ratings in the post-study questionnaire show that participants were generally positive about RetroLens (Figure 8). We further derive the following three reasons behind the positive ratings via thematic analysis of the post-study interview.

- **RetroLens supports intuitive inspection and comparison of retrosynthetic routes.** Four participants reported that they can examine and compare alternative retrosynthetic route designs conveniently with RetroLens. Without this system, chemists have to switch between various platforms, from search engines for chemical reactions to existing AI-powered services, to explore possible retrosynthetic reactions at each step. They also need to manually piece the

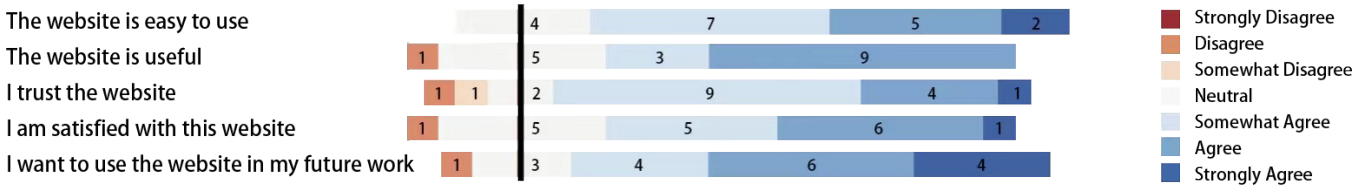
Table 3: Top: User perception towards RetroLens and possible underlying reasons. Middle: User adaptation of workflows and their specific usage of RetroLens. Bottom: User concerns about collaborating with AI and possible underlying reasons.

Perception towards RetroLens	Possible Reasons
Support intuitive inspection and comparison	Enable chemists to compare alternative retrosynthetic route designs by embedding them into the existing part of the route.
Inspire novel perspectives	Manual and AI parts of the retrosynthetic route can be reviewed as a whole.
Facilitate a comprehensive trade-off evaluation	Provide globally optimal revision plans for all problematic molecules. Help comprehensively consider multiple criteria to find ideal revision steps. Support comparisons of candidate revision steps on different dimensions.
Adaption of workflows	**Usages**
Use RetroLens in theoretical MRRP process	Collaborate with AI to iteratively design and revise for a final theoretically possible retrosynthetic route. Quickly verify ideas of MRRP with RetroLens.
Use RetroLens in practical synthetic experimentation	Revise problematic molecules encountered in practical synthesis with RetroLens.
Concerns about collaborating with AI	**Possible reasons**
Being misled by AI	Chemists cannot properly examine the quality of AI-generated routes. Over-trust the imperfect AI.
Being over-influenced by AI	Chemists hope to hand deconstruction tasks over to AI earlier and the thinking process is thus constrained.

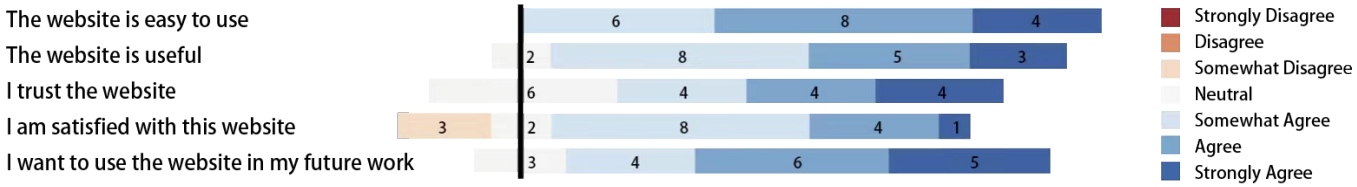
step-wise decisions together to construct the final plan, the workload of which increases dramatically when multiple options exist in each step. In contrast, RetroLens enables users to compare alternative designs by embedding them into the existing part of the route (P5).

- **RetroLens inspires new perspectives throughout MRRP.** Five participants claimed that the joint design of retrosynthetic routes enabled by RetroLens stimulates their creativity during the MRRP process. In the manual condition, chemists tend to take a depth-first exploration approach and only focus on the deconstruction of the product(s) at the current step. Although they also use AI-powered tools in their MRRP routine, the predicted results would only influence their local decisions. Nevertheless, RetroLens combines the outcomes of humans and AI together, enabling chemists to review these two parts as a whole. For example, as P13 said, *"the AI-generated routes reminded me of a better deconstruction of the target molecule, so I returned back to revise that manual step I made earlier"*.
- **RetroLens facilitates a comprehensive trade-off evaluation during MRRP process.** Without RetroLens, it is extremely difficult, if not impossible, for chemists to identify a globally optimal solution for all problems in the route. In comparison, six participants reported that RetroLens can address this challenge since it provides revision plans for all undesired molecules. Additionally, chemists always find it *"a struggling process to comprehensively consider multiple factors to find an ideal revision step, especially when the retrosynthetic routes are complex"* (P1). P2 encountered similar issues, *"in the past, I simply hoped to revise as few steps as possible, but I ignored the impact of my revisions on the associated branches that do not have problems. This often causes me to make more mistakes in the revision process"*. With RetroLens, seven participants feel it much easier to make revision decisions, as users only need to input the criteria they care about and the corresponding weightings based on their preferences, and AI can automatically rank possible revision steps accordingly. Furthermore, RetroLens displays the confidence of AI-generated retrosynthetic routes and the SAW scores of candidate revision steps. Five participants conveyed that such information makes assessing the trade-offs among different revision plans a lot simpler. For instance, P1 commented, *"I can not only know the difference between the cost of choosing different revision steps based on the overall scores but also compare the revision steps on different dimensions based on my needs"*.

6.3.2 Adaptation of Workflows. All participants reported that they want to use RetroLens in their future research (Figure 8(b)), but possibly in different stages according to their own needs and preferences. First, about 89% of participants (16/18) stated that they would use RetroLens **in the theoretical MRRP process**. They could collaborate with AI to design an initial retrosynthetic route and iteratively revise the route with the support of the AI-assisted revision step selection function until they obtain a theoretically possible route. Three participants (17%) suggested that RetroLens may support them **verify their ideas of MRRP**. For example, P4 said *"sometimes I would have an initial idea of the first few steps of deconstruction when I see a target molecule. Then RetroLens could help me quickly complete the routes and help me check whether my thinking is in the right direction"*. Second, since chemists usually find new problems with the routes in actual molecule synthesis, four participants (22%) indicated that RetroLens can also be used **during practical synthetic experimentation**. When encountering problematic molecules (e.g., products that cannot be synthesized in the real world following the retrosynthetic route), they would input these molecules back into our system and revise the routes based on RetroLens's recommendations. Interestingly, P1 proposed that such usage is also applicable to the synthesis of simple molecules. *"Although existing AI-powered MRRP tools can handle simple molecules,*



(a) AI system



(b) AI^2 system

Figure 8: User perception towards the (a) AI and (b) AI^2 systems.

I often have problems with the practical synthesis process based on their returned routes. RetroLens can help me to modify the routes".

6.3.3 Concerns about collaborating with AI. In addition to the aforementioned positive feedback about RetroLens, participants also expressed concerns about the process of collaborating with AI. We summarized their views into two key themes below.

- **Fear of being misled by AI.** Six participants worried that they would be misguided by AI due to the inaccuracy of RetroLens. For example, when finding that the AI-generated retrosynthetic routes failed, P17 went back and checked whether she had made mistakes in her previous manual steps. Nevertheless, she finally discovered that the failure occurred because the molecule was still too complex for AI to analyze, although it was originally deemed doable according to the empirical rule of thumb (Section 4.1). This concern comes particularly from the novice chemists (5/6). For one thing, they were not able to properly tell the complexity of a molecule from experience and examine the quality of AI-generated routes as expert participants did. For another, we found that these participants tended to over-trust the imperfect AI. For example, P11 said, *"since I do not have abundant knowledge [in MRRP], I am not confident in my deconstruction decisions. So I was more inclined to doubt my previous decisions rather than questioning system performance when I found the MRRP is a failure"*. P18 even expressed that *"I think AI would never perform worse than me"*.
- **Fear of being over-influenced by AI.** Three participants reported that their thinking process was influenced by the AI collaborating with them. For instance, P9 indicated that in addition to considering synthesis-related knowledge, she would think of how to deconstruct molecules into reactants that AI can handle so that she could hand over the deconstruction tasks to AI as soon as possible. However, she claimed that *"even if this did help me plan retrosynthetic routes faster and reduce my workload, my manual MRRP is thus more constrained. Sometimes, I found that the quality of my manual route design declined as a result"*.

7 DISCUSSION

In this section, we first discuss the generalizability of our work. Built upon the key findings in the user study, we then derive several design issues in human-AI collaboration for multi-step problems and propose corresponding design considerations (**DC**) adapted from and extended beyond existing human-AI interaction guidelines [2]. We also discuss the limitations and future work of our research.

7.1 Generalizability of Our Work

Although our system is designed for MRRP, our proposed mechanisms and pipeline for human-AI collaboration could be easily extended to other multi-step problems (e.g., programming and multi-step mathematical questions). Various AI models have been proposed to help people solve multi-step problems [38], but they are usually limited to solving relatively simple and repetitive cases [9]. Our human-AI collaborative approach could be adapted to help solve complex cases by 1) adjusting the rules for task allocation between humans and AI and 2) unifying the criteria used by humans and AI to assess the cost and gain of each problem step. For example, to modify our system for complex programming, we can discuss with domain experts to identify which kind of cases can be handled well by existing AI, how they debug their codes, and how they determine the severity of errors and warnings as well as places to improve. By translating these insights into computational rules, the system can automatically assign programmers and AI to work on different code segments of a complex project. Once the codes written by humans and AI are integrated together, the system could help suggest and rank lines of codes that could possibly be optimized based on user preferences. Some of the criteria employed in MRRP may be applicable in this process. For example, programmers may hope to reduce the time complexity of the program on the premise of modifying as few lines of code as possible and affecting as few code segments they are satisfied with as possible.

7.2 Design Issues and Design Considerations

7.2.1 Misled by inappropriate task allocation. One pain point we found from the user study is that chemists may be misled by faulty results generated by AI during the joint action process due to inappropriate task allocation (Section 6.3.3). Although we have carefully discussed with the experts to define AI's capability boundary for task distribution, participants still found some special cases which AI was expected to handle well but failed to do so. The steps containing such molecules were wrongly assigned to AI, and its failure negatively affected users' decision-making. On the one hand, when seeing failed routes returned by the AI, the participants would doubt and reinspect their manual deconstruction, resulting in a waste of time. On the other hand, users felt confused about how to assess AI's capability and spent extra effort on determining task allocation later. This imposed a learning curve on the participants. The first few times they encountered special cases where AI returned failed retrosynthetic routes as the inputted molecules went beyond its ability, the participants had to diagnose whether the failures were caused by inappropriate task allocation or their previous decisions. As their understanding of AI's capability built up, they would ignore AI's output and directly did manual deconstruction when they encountered similar molecules. Although RetroLens statistically lowered chemists' overall cognitive load during the MRRP process (Section 6.2.2), six participants reported in the interview that it could be cognitively demanding when there was a need to analyze AI's faulty predictions to update their knowledge of AI.

DC1. Provide more fine-grained task allocation. We recommend designers to explore more accurate specifications of the AI capability boundary and more fine-grained task allocation in human-AI collaborative systems [50]. For example, rather than using heuristic rules, we could train a predictive model to estimate case by case whether a molecule inputted by the user is within AI's scope of capabilities so that the steps could be directed to humans or AI dynamically.

DC2. Instantly communicate how well AI can do what it can do as collaboration unfolds. Prior work proposed human-AI interaction design guidelines suggesting that AI systems should help human *"make clear how well the system can do what it can do"* when users *initially* interact with AI [2]. However, in the context of multi-step problems, providing a general knowledge heuristic of AI's capability boundary in the initial stage of human-AI collaboration is insufficient, as AI's ability at each detailed step could vary. Hence, we propose that the previous design guidelines could be further improved to *"update and communicate how well the system can do what it can do" during interaction* to mitigate users' cognitive load of learning AI's capability boundary on their own. Designers can enhance model transparency and interpretability at a instance level to help users detect the inaccuracy of AI early on. To achieve this, in addition to the model's confidence in AI-generated retrosynthetic routes, which has been shown in our system, RetroLens could provide the reaction precedents in training datasets similar to the routes [16, 18, 36, 48] and the data sources of these precedents [45, 89], making it easier for users to understand and evaluate these routes. Moreover, the systems could store special cases that cannot be handled by AI previously and inform chemists that AI may make mistakes when they meet similar molecules in the future [62].

7.2.2 Tendency of over-reliance on AI. We observed in our user study that, when finding AI-generated routes were successful, most participants (11/18) would directly check the returned routes and choose one rather than analyzing the products by themselves. In this way, users could collaborate with AI more smoothly and may hand over the deconstruction tasks to AI earlier to reduce their own workload. Some participants even changed their thinking process to adapt to AI's capability (Section 6.3.3) and consciously deconstructed molecules into reactants that AI can process (Section 6.3.3). They thus were less likely to be misled by AI's wrong route prediction. However, such adjustments could be a double-edged sword. For one thing, chemists' manual MRRP process would consequently be constrained by the output of AI. The diversity of the final design might be endangered [39]. For another, this may cause users' over-reliance on AI (i.e., the user could design a better retrosynthetic route for a product on their own than AI but accept AI's recommendation [12]) [43, 70]. Even though AI enables users to obtain complete routes faster (Section 6.1.1 and 6.1.3), it is possible that users may only obtain suboptimal decisions which are less reasonable and feasible than those designed by human alone [22].

DC3. Watch out for and mitigate human over-reliance on AI. We suggest extending design guidelines for human-AI interaction and collaboration [2] and recommend that systems should *"watch out for and mitigate human over-reliance on AI" during interaction.* Past research highlights the dual-process theory in the human decision-making process [11, 93] and demonstrates that cognitive forcing strategy could effectively reduce users' over-reliance on AI [12, 69]. Therefore, when recognizing users are trying to turn over too much work to AI, we suggest applying cognitive forcing functions to systems, such as having users make independent decisions before seeing AI's prediction. Moreover, an option to enable/disable AI could be provided to users so that the deconstruction tasks would be delegated to AI only upon users' request even if the molecules are within the scope of AI's capability [27]. In addition, showing counterfactual explanations for AI's output would also encourage users to explore more route design rather than simply accepting whatever returned by AI [66, 88].

7.2.3 Needs for personalized support. Previous studies on human-AI interaction stress the need to *"learn from user behavior" over time* [2]. Similarly, our experiment reveals that the participants really demand and appreciate personalized support during the MRRP process. For example, P9 said that *"it would be better if AI can rank the AI-generated retrosynthetic routes and candidate revision steps based on my previous choices".*

DC4. Maintain a shared mental model of humans. When conducting retrosynthetic analysis, chemists may have their own preferences and concerns (e.g., laboratory resources available for experiments). To achieve effective human-AI collaboration, we suggest that AI should maintain a shared mental model of users to make its service personalized and adaptive. Many existing methods for inferring human mental models (e.g.,[44]) can be applied to help AI derive users' preferences on retrosynthetic routes from their historical decisions, update the estimated user model based on current interactions, and adjust the ranking of AI-generated routes and the weightings of the criteria for assessing the candidate revision steps accordingly.

7.3 Limitations and Future Work

Our system and experiments have several limitations. First, as a proof-of-concept system, RetroLens does not allow chemists to edit the parts of retrosynthetic routes returned by AI. This may affect the efficiency and flexibility of MRRP. Second, the set of criteria considered for ranking the candidate revision steps in RetroLens came from the literature survey and the participatory design process, which may be incomplete. Third, as chemical synthesis takes a long time and needs lots of resources [91], we could not test participants' final retrosynthetic plans by practical synthesis experimentation within the scope of our study. We can only invite experts to assess the routes based on their experience to estimate the feasibility of the routes. Furthermore, human expertise plays an important role in human-AI collaboration (e.g., influencing human's preference for delegating to AI [40]). However, we only considered user expertise as a random variable in our user study and did not analyze its impact on how chemists would interact with and perceive towards RetroLens.

In the future, we will apply more advanced AI MRRP models in joint molecule deconstruction (e.g., models trained on a larger database) and improve the AI-assisted revision step selection mechanisms by considering a more comprehensive set of criteria and exploring other MCDM algorithms (e.g., fuzzy Choquet integral [85]). To help users better collaborate with AI, we could provide more information about AI's output, such as the evidence of AI's prediction (i.e., similar reaction precedents in existing work). Moreover, we will incorporate more customizable settings in our system. For example, RetroLens could enable users to modify the AI-predicted routes to improve users' flexibility in the MRRP process. To allow users to better screen the desired revision plans, RetroLens also could display only the revision steps whose SAW scores are lower than a value specified by a user. In addition, we will recruit more chemists to our user study and balance the number of participants with different levels of molecule synthesis experience to systematically explore whether they would have different user needs and collaboration patterns with our system.

8 CONCLUSION

In this paper, we presented RetroLens, a human-AI collaborative system that facilitates humans and AI to complement each other for MRRP for complex molecules. Through an iterative participatory design process, we co-designed the RetroLens system with six chemical experts. RetroLens adopts two forms of human-AI collaboration: joint action for molecule deconstruction and AI-assisted decision-making for revision step selection. RetroLens allocates deconstruction tasks to chemists and AI based on an empirical rule-of-thumb about the strengths of both parties. It also helps identify possible causes of the problems in the initial route, and recommends candidate revision steps based on chemists' preferences. A follow-up within-subjects user study demonstrated that RetroLens improved chemists' efficiency in MRRP, increased their confidence in the final decisions, and lowered their cognitive load. We believe our work takes the first step to explore human-AI collaboration in dynamic, multi-step problem solving.

ACKNOWLEDGMENTS

Many thanks to Zifei Chen and Jinwen Zhang for their valuable inputs. This work is supported by the Research Grants Council of the Hong Kong Special Administrative Region under General Research Fund (GRF) with Grant No. 16203421 and No. 16209821.

REFERENCES

[1] Laura KG Ackerman-Biegasiewicz, Daniela M Arias-Rotondo, Kyle F Biegasiewicz, Elizabeth Elacqua, Matthew R Golder, Laure V Kayser, Jessica R Lamb, Christine M Le, Nathan A Romero, Sidney M Wilkerson-Hill, et al. 2020. Organic Chemistry: A Retrosynthetic Approach to a Diverse Field. , 1845–1850 pages.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[3] Zahra Ashktorab, Q Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020. Human-ai collaboration in a cooperative game setting: Measuring social perception and outcomes. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–20.

[4] Atanas G Atanasov, Sergey B Zotchev, Verena M Dirsch, and Claudiu T Supuran. 2021. Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery* 20, 3 (2021), 200–216.

[5] Tomasz Badowski, Karol Molga, and Bartosz A Grzybowski. 2019. Selection of cost-effective yet chemically diverse pathways from the networks of computer-generated retrosynthetic plans. *Chemical Science* 10, 17 (2019), 4640–4651.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[7] Javier L Baylon, Nicholas A Cilfone, Jeffrey R Gulcher, and Thomas W Chittenden. 2019. Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. *Journal of Chemical Information and Modeling* 59, 2 (2019), 673–688.

[8] William Bort, Igor I Baskin, Timur Gimadiev, Artem Mukanov, Ramil Nugmanov, Pavel Sidorov, Gilles Marcou, Dragos Horvath, Olga Klimchuk, Timur Madzhidov, et al. 2021. Discovery of novel chemical reactions by deep generative recurrent neural network. *Scientific Reports* 11, 1 (2021), 1–15.

[9] Riccardo Bovo, Nicola Binetti, Duncan P Brumby, and Simon Julier. 2020. Detecting errors in pick and place procedures: detecting errors in multi-stage and sequence-constrained manual retrieve-assembly procedures. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 536–545.

[10] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597.

[11] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[12] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[13] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[14] Liwei Chan, Yi-Chi Liao, George B Mo, John J Dudley, Chun-Lien Cheng, Per Ola Kristensson, and Antti Oulasvirta. 2022. Investigating Positive and Negative Qualities of Human-in-the-Loop Optimization for Designing Interaction Techniques. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[15] Binghong Chen, Chengtao Li, Hanjun Dai, and Le Song. 2020. Retro*: learning retrosynthetic planning with neural guided A* search. In *International Conference on Machine Learning*. PMLR, 1608–1616.

[16] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems* 32 (2019).

[17] Lydia B Chilton, Juho Kim, Paul André, Felicia Cordeiro, James A Landay, Daniel S Weld, Steven P Dow, Robert C Miller, and Haoqi Zhang. 2014. Frenzy: collaborative data organization for creating conference sessions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1255–1264.

[18] Penny Chong, Ngai-Man Cheung, Yuval Elovici, and Alexander Binder. 2021. Toward Scalable and Unified Example-Based Explanation and Outlier Detection. *IEEE Transactions on Image Processing* 31 (2021), 525–540.

[19] Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. 2018. SCScore: synthetic complexity learned from a reaction corpus. *Journal of*

Chemical Information and Modeling 58, 2 (2018), 252–261.
[20] Elias James Corey. 1967. General methods for the construction of complex molecules. *Pure and Applied chemistry* 14, 1 (1967), 19–38.
[21] Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. 2019. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems* 32 (2019).
[22] Kahneman Daniel. 2017. Thinking, fast and slow.
[23] Ian W Davies. 2019. The digitization of organic synthesis. *Nature* 570, 7760 (2019), 175–181.
[24] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37.
[25] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*. 39–45.
[26] William Finnigan, Sabine L Flitsch, Lorna J Hepworth, and Nicholas J Turner. 2021. Enzyme Cascade Design: Retrosynthesis Approach. In *Enzyme Cascade Design and Modelling*. Springer, 7–30.
[27] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digital Medicine* 4, 1 (2021), 1–8.
[28] Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. 2020. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics* 12, 1 (2020), 1–9.
[29] Subhash Ghosh and Tapan Kumar Pradhan. 2008. The first total synthesis of emericellamide A. *Tetrahedron Letters* 49, 22 (2008), 3697–3700.
[30] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. 2013. Lineup: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2277–2286.
[31] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[32] Bartosz A Grzybowski, Sara Szymkuć, Ewa P Gajewska, Karol Molga, Piotr Dittwald, Agnieszka Wołos, and Tomasz Klucznik. 2018. Chematica: a story of computer code that started to think like a chemist. *Chem* 4, 3 (2018), 390–398.
[33] Corina Haita-Falah. 2017. Sunk-cost fallacy and cognitive ability in individual decision-making. *Journal of Economic Psychology* 58 (2017), 44–59.
[34] Peng Han, Peilin Zhao, Chan Lu, Junzhou Huang, Jiaxiang Wu, Shuo Shang, Bin Yao, and Xiangliang Zhang. 2022. GNN-Retro: Retrosynthetic Planning with Graph Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 4014–4021.
[35] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*. Vol. 52. Elsevier, 139–183.
[36] Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. 2019. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 32–40.
[37] Shoichi Ishida, Kei Terayama, Ryosuke Kojima, Kiyosei Takasu, and Yasushi Okuno. 2022. AI-Driven Synthetic Route Design Incorporated with Retrosynthesis Knowledge. *Journal of Chemical Information and Modeling* 62, 6 (2022), 1357–1367.
[38] Ellen Jiang, Edwin Toh, Alejandra Molina, Kristen Olson, Claire Kayacik, Aaron Donsbach, Carrie J Cai, and Michael Terry. 2022. Discovering the Syntax and Strategies of Natural Language Programming with Generative Language Models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.
[39] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. 2018. To trust or not to trust a classifier. *Advances in Neural Information Processing Systems* 31 (2018).
[40] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2020. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. (2020).
[41] Bo Kang, Arun Kulshreshth, and Joseph J LaViola Jr. 2016. Analyticalink: An interactive learning environment for math word problem solving. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 419–430.
[42] Pavel Karpov, Guillaume Godin, and Igor V Tetko. 2019. A transformer model for retrosynthesis. In *International Conference on Artificial Neural Networks*. Springer, 817–830.
[43] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[44] Harmanpreet Kaur, Alex Williams, and WS Lasecki. 2019. Building shared mental models between humans and ai for effective collaboration. *CHI'19, May 2019, Glasgow, Scotland* (2019).
[45] Junhan Kim, Jana Muhic, Lionel Peter Robert, and Sun Young Park. 2022. Designing Chatbots with Black Americans with Chronic Conditions: Overcoming Challenges against COVID-19. In *CHI Conference on Human Factors in Computing Systems*. 1–17.
[46] Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P Gajewska, et al. 2018. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* 4, 3 (2018), 522–532.
[47] Levente Kocsis and Csaba Szepesvári. 2006. Bandit based monte-carlo planning. In *European Conference on Machine Learning*. Springer, 282–293.
[48] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*. PMLR, 1885–1894.
[49] Caitlin Kuhlman, Diana Doherty, Malika Nurbekova, Goutham Deva, Zarni Phyo, Paul-Henry Schoenhagen, MaryAnn VanValkenburg, Elke Rundensteiner, and Lane Harrison. 2019. Evaluating preference collection methods for interactive ranking analytics. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
[50] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. 1–18.
[51] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[52] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.
[53] Greg Landrum et al. 2013. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling.
[54] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[55] Yang Li, Sayan Sarcar, Yilin Zheng, and Xiangshi Ren. 2021. Exploring Text Revision with Backspace and Caret in Virtual Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.
[56] Kangjie Lin, Youjun Xu, Jianfeng Pei, and Luhua Lai. 2020. Automatic retrosynthetic route planning using template-free models. *Chemical Science* 11, 12 (2020), 3355–3364.
[57] Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. 2017. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Central Science* 3, 10 (2017), 1103–1113.
[58] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[59] Hao Lü and Yang Li. 2012. Gesture coder: a tool for programming multi-touch gestures by demonstration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2875–2884.
[60] Shuai Ma, Mingfei Sun, and Xiaojuan Ma. 2022. Modeling Adaptive Expression of Robot Learning Engagement and Exploring its Effects on Human Teachers. *ACM Transactions on Computer-Human Interaction* (2022).
[61] Shuai Ma, Zijun Wei, Feng Tian, Xiangmin Fan, Jianming Zhang, Xiaohui Shen, Zhe Lin, Jin Huang, Radomír Měch, Dimitris Samaras, et al. 2019. SmartEye: assisting instant photo taking via integrating user preference with deep view proposal network. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.
[62] Shuai Ma, Taichang Zhou, Fei Nie, and Xiaojuan Ma. 2022. Glancee: An adaptable system for instructors to grasp student learning status in synchronous online classes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–25.
[63] Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. 2019. Discovering the sweet spot of human-computer configurations: A case study in information extraction. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
[64] Barbara Mikulak-Klucznik, Patrycja Gołębiowska, Alison A Bayly, Oskar Popik, Tomasz Klucznik, Sara Szymkuć, Ewa P Gajewska, Piotr Dittwald, Olga Staszewska-Krajewska, Wiktor Beker, et al. 2020. Computational planning of the synthesis of complex natural products. *Nature* 588, 7836 (2020), 83–88.
[65] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Designing Interactive Transfer Learning Tools for ML Non-Experts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[66] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In

Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 607–617.

[67] Vishnu H Nair, Philippe Schwaller, and Teodoro Laino. 2019. Data-driven chemical reaction prediction and retrosynthesis. *CHIMIA International Journal for Chemistry* 73, 12 (2019), 997–1000.

[68] Stephan Pajer, Marc Streit, Thomas Torsney-Weir, Florian Spechtenhauser, Torsten Möller, and Harald Piringer. 2016. Weightlifter: Visual weight space exploration for multi-criteria decision making. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2016), 611–620.

[69] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.

[70] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[71] Koji Satoh and Kimito Funatsu. 1999. A novel approach to retrosynthetic analysis using knowledge bases derived from reaction databases. *Journal of Chemical Information and Computer Sciences* 39, 2 (1999), 316–325.

[72] John S Schreck, Connor W Coley, and Kyle JM Bishop. 2019. Learning retrosynthetic planning through simulated experience. *ACS Central Science* 5, 6 (2019), 970–981.

[73] Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. 2020. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science* 11, 12 (2020), 3316–3325.

[74] Oliver B Scott and AW Edith Chan. 2020. ScaffoldGraph: an open-source library for the generation and analysis of molecular scaffold networks and scaffold trees. *Bioinformatics* 36, 12 (2020), 3930–3931.

[75] Marwin HS Segler, Mike Preuss, and Mark P Waller. 2018. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 7698 (2018), 604–610.

[76] Marwin HS Segler and Mark P Waller. 2017. Modelling chemical reasoning to predict and invent reactions. *Chemistry–A European Journal* 23, 25 (2017), 6118–6128.

[77] Seung-Woo Seo, You Young Song, June Yong Yang, Seohui Bae, Hankook Lee, Jinwoo Shin, Sung Ju Hwang, and Eunho Yang. 2021. GTA: Graph truncated attention for retrosynthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 531–539.

[78] Chuhan Shi, Zhihan Jiang, Xiaojuan Ma, and Qiong Luo. 2022. A Personalized Visual Aid for Selections of Appearance Building Products with Long-term Effects. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[79] Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, and Jian Tang. 2020. A graph to graphs framework for retrosynthesis prediction. In *International Conference on Machine Learning*. PMLR, 8818–8827.

[80] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484–489.

[81] Teague Sterling and John J Irwin. 2015. ZINC 15–ligand discovery for everyone. *Journal of Chemical Information and Modeling* 55, 11 (2015), 2324–2337.

[82] Sarah Jean Stevens. 2011. *Progress toward the synthesis of providencin*. Ph. D. Dissertation. Colorado State University.

[83] Dong Sun, Renfei Huang, Yuanzhe Chen, Yong Wang, Jia Zeng, Mingxuan Yuan, Ting-Chuen Pong, and Huamin Qu. 2019. PlanningVis: A visual analytics approach to production planning in smart factories. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 579–589.

[84] Sara Szymkuć, Ewa P Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk, and Bartosz A Grzybowski. 2016. Computer-assisted synthetic planning: the end of the beginning. *Angewandte Chemie International Edition* 55, 20 (2016), 5904–5937.

[85] Chunqiao Tan and Xiaohong Chen. 2010. Intuitionistic fuzzy Choquet integral operator for multi-criteria decision making. *Expert Systems with Applications* 37, 1 (2010), 149–157.

[86] Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 239–245.

[87] Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. 2020. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature Communications* 11, 1 (2020), 1–11.

[88] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 2017. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 465–474.

[89] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[90] Kurt VanLehn. 2016. Regulative loops, step loops and task loops. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 107–112.

[91] Chandrabhan Verma, LO Olasunkanmi, Eno E Ebenso, and MA Quraishi. 2018. Substituents effect on corrosion inhibition performance of organic compounds in aggressive ionic solutions: a review. *Journal of Molecular Liquids* 251 (2018), 100–118.

[92] Emily Wall, Subhajit Das, Ravish Chawla, Bharath Kalidindi, Eli T Brown, and Alex Endert. 2017. Podium: Ranking data using mixed-initiative visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 288–297.

[93] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.

[94] Yiting Wang, Walker M White, and Erik Andersen. 2017. Pathviewer: Visualizing pathways through student data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 960–964.

[95] Zhuang Wang, Wenhan Zhang, and Bo Liu. 2021. Computational Analysis of Synthetic Planning: Past and Future. *Chinese Journal of Chemistry* 39, 11 (2021), 3127–3143.

[96] David Weininger. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* 28, 1 (1988), 31–36.

[97] Di Weng, Ran Chen, Zikun Deng, Feiran Wu, Jingmin Chen, and Yingcai Wu. 2019. SRVis: Towards Better Spatial Integration in Ranking Visualization. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 459–469. https://doi.org/10.1109/TVCG.2018.2865126

[98] Di Weng, Chengbo Zheng, Zikun Deng, Mingze Ma, Jie Bao, Yu Zheng, Mingliang Xu, and Yingcai Wu. 2020. Towards better bus networks: A visual analytics approach. *IEEE transactions on visualization and computer graphics* 27, 2 (2020), 817–827.

[99] Di Weng, Heming Zhu, Jie Bao, Yu Zheng, and Yingcai Wu. 2018. Homefinder revisited: Finding ideal homes with reachability-centric multi-criteria decision making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[100] Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, Jinyu Yang, Yang Yu, and Junzhou Huang. 2020. Retroxpert: Decompose retrosynthesis prediction like a chemist. *Advances in Neural Information Processing Systems* 33 (2020), 11248–11258.

[101] Litao Yan, Elena L Glassman, and Tianyi Zhang. 2021. Visualizing Examples of Deep Neural Networks at Scale. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[102] Stelios H Zanakis, Anthony Solomon, Nicole Wishart, and Sandipa Dublish. 1998. Multi-attribute decision making: A simulation comparison of select methods. *European Journal of Operational Research* 107, 3 (1998), 507–529.

[103] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. 2012. Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 217–226.

[104] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

[105] Yong Zheng. 2017. Criteria chains: a novel multi-criteria recommendation approach. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 29–33.

[106] Oren Zuckerman, Viva Sarah Press, Ehud Barda, Benny Megidish, and Hadas Erel. 2022. Tangible Collaboration: A Human-Centered Approach for Sharing Control With an Actuated-Interface. In *CHI Conference on Human Factors in Computing Systems*. 1–13.

A APPENDIX

A.1 Measurements of User Experience in MRRP

Table 4: In-task questionnaire used in the AI and AI^2 conditions in 7-point Likert scale.

Category	Question
User confidence in final planning	Q1. How confident are you in your decision?
User cognitive load during MRRP	Q2. How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Q3. How much physical activity was required (e.g.. pushing, pulling, turning, controlling, activating, etc.)? Q4. How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Q5. How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? Q6. How hard did you have to work (mentally and physically) to accomplish your level of performance? Q7. How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?
Website usability	Q8. What do you think of the usability of the website?
Website usefulness	Q9. What do you think of the usefulness of the website?
Trust of website	Q10. Do you trust the website?
User satisfaction with website	Q11. Are you satisfied with the website?
Adoption and use intention	Q12. Do you intend to use the website in future research?