A EXISTING PLANNING ALGORITHMS

	3N-MCTS	HgSearch	DFPN-E	RetroGNN	Metro	FusionRetro	Retro-fallback
Algorithm	Online	Offline	Offline	Offline	Online	Offline	Online
	EG-MCTS	**Retro***	**RetroGraph**	**GNN-Retro**	**SimulatedExp**	**GRASP**	**PDVN**
Algorithm	Online	Offline	Offline	Offline	Online	Online	Online

Table 4: Existing online and offline retrosynthetic planning methods.

Active reinforcement learning An active reinforcement learning(ARL) agent learns when to pay query costs and observe rewards(Daniel et al. (2014)) or other signals. A wide range of work has focused on ameliorating the problem of defining a complete reward function on trajectories in complicated real-world tasks, i.e. automated driving and robot grasping(Christiano et al. (2023), Saunders et al. (2018), Subramanian et al. (2016), Daniel et al. (2014)). To minimize reliance on human experts, Krueger et al. (2020), Bellinger et al. (2020), and Schulze & Evans (2018) study the active measure reinforcement learning(AMRL) framework under multi-armed bandit and tabular settings. Furthermore, Warnell et al. (2018) and Knox & Stone (2009) propose the TAMER framework which takes into account the time delays and noise when the human, a "teacher", provides rewards online to the agent, a "student".

B SINGLE STEP PROBABILITY

As the single-step model is trained to predict feasible reactant precursors, it is biased towards frequent reactions instead of those with high qualities. We verify the issue that frequently collected reactions in a single-step dataset are not necessarily high-yield, which we substantiate based on an analysis from Schwaller et al. (2021) that explores yields reported in the open-source USPTO dataset.

The USPTO dataset with reaction yields in sub-gram scale(Schwaller et al. (2021)) contains a large number of reactions and a broad range of superclasses, and a reaction distribution closely resembling that of the USPTO single-step dataset, such as USPTO-MIT. The actual reaction yield distribution of the above dataset, originally presented in Schwaller et al. (2021), is depicted in Fig 5c. Notably, a significant proportion of reactions within the dataset exhibits relatively low yields, affirming that the USPTO single-step dataset is not inherently biased to high-yield reactions. Fig 5a (originally presented in Schwaller et al. (2021)) shows various superclasses of reactions, where each color corresponds to a superclass and the coverage area of each color roughly represents the frequency of that superclass of reactions in the dataset. Combining Fig 5a and Fig 5b, we conclude that high-frequency superclasses do not show a significant correlation with high yields. For example, the superclasses annotated in purple and cyan demonstrate low yields, with only the green reaction superclass corresponding to high yields in Fig 5b.

In summary, frequently collected reactions in a single-step dataset are not inevitable to be high-yield ones and the single-step probabilities are not biased to high-quality but high-frequency reactions.

C BINING STRATEGY

A bining strategy $\mathcal{B}$ is performed to discretize the continuous reaction quality values in and obtain the associated bucket embedding. The preceding reaction cost is concatenated in the format its representation as $\mathcal{O}(a^r, a^q)$ in Eq 2. Concretely, when $a^q = 1$, we derive $\mathcal{O}(a^r, a^q)$ from (1) discretizing continuous reaction quality values into N^M discrete buckets, (2) learning N^M trainable embeddings in d^M dimensions for all

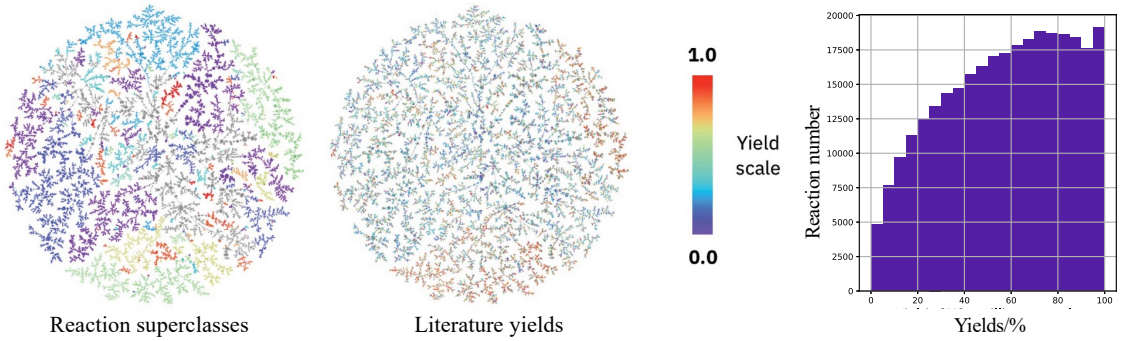

Figure 5: The figure is directly borrowed from Schwaller et al. (2021). USPTO yield analysis: (a) shows the superclasses which roughly reflect the reaction frequency in the dataset. (b) depicts the yield scales of reactions labeled by the superclasses in (a). and (c) displays the distribution of the reaction yields in the dataset.

buckets within our critic, and (3) determining the bucket index that the queried quality value u belongs to and thereby the associated bucket embedding. When $a^q = 0$, $\mathcal{O}(a^r, a^q) = \mathbb{M}$ as a d^M-dimensional trainable embedding. In our implementation, we consider $d^M = 512$ and $N^M = 18$ buckets which are defined in Fig 6 in the revision. These buckets are obtained via (1) collecting about 28M reactions during planning by GRASP and Retro*, (2) computing their reaction qualities by our surrogate model, and (3) defining the bucket boundaries to ensure that each bin covers a similar number of reactions.

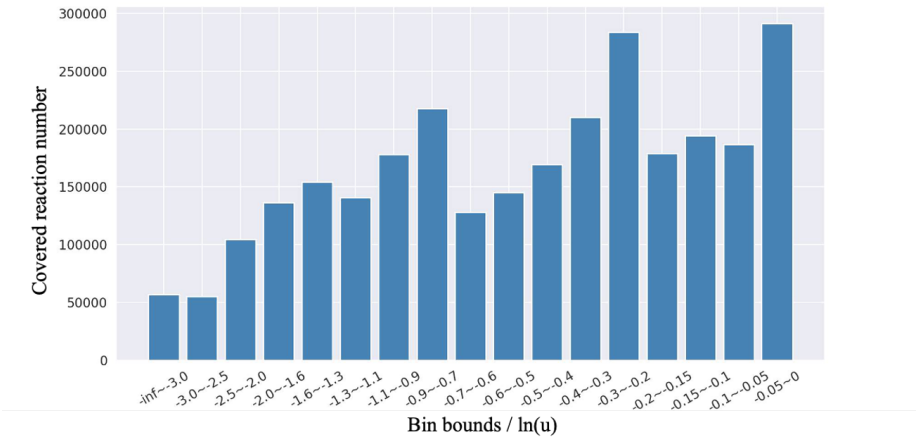

Figure 6: Bin bucket boundaries. Each bin covers a similar amount of collected reactions individually.

D SURROGATE TRAINING DETAILS

We utilize a 8-layer Transformer as the architecture of our surrogate model. The hyper-parameters are listed in Tab 5. The training of our surrogate model involves two steps: (1) pre-training on the USPTO-MIT dataset, and (2) finetuning on an in-house expert dataset of routes featuring high-yield reactions. It is important to note that we introduce step (2) precisely to ensure that high predictive probabilities from our surrogate model align with high yields.

Hyperparameters	Values
Encoder layers	4
Decoder layers	4
Encoder embedding dimension	2048
Encoder FFN embedding dimension	2048
Encoder attention heads	8
Decoder embedding dimension	2048
Decoder FFN embedding dimension	2048
Decoder attention heads	8
Optimizer	Adam
Learning rate	1e-4
Weight decay	0.0001
N epochs	12
Clip norm	0.25
Dropout rate	0.1

Table 5: The output of the cross-validation used for the hyperparameters optimization

E CORRELATION BETWEEN THE SURROGATE MODEL AND REACTION YIELDS

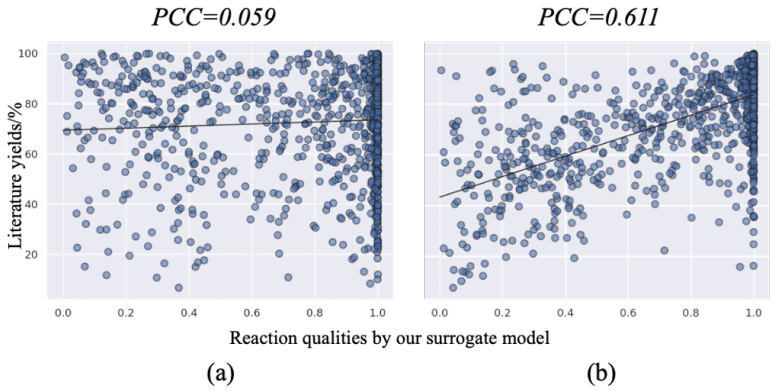

Figure 7: PCC against reaction yields. (a) shows the PCC of the pre-trained model and (b) shows the PCC of the finetuned model.

To evaluate our surrogate model, we resort to a route-with-yield test set. Following the method described in Chen et al. (2020), we extract synthesis routes with yields from the USPTO-milligram-scale reaction yield dataset Schwaller et al. (2021). For evaluation purposes, we randomly select 200 routes, encompassing approximately 1000 reactions. We thereby calculate the Pearson correlation coefficient(PCC) between the reaction quality predicted by our surrogate model and the literature yield. In Fig 7 of the revised manuscript, (a) illustrates the 0.059 PCC of the pre-trained model while (b) shows the 0.611 PCC of the finetuned model, providing strong evidence that the surrogate model accurately predicts yields.

F SUCCESS RATE AND ROUTE QUALITY INCONSISTENCY

There are two separate objectives to optimize in our framework: the success rate and the route quality. However, optimizing these two objectives together can lead to certain trade-offs, as demonstrated by the following case study.

As shown in Fig. 8, the root molecule M_0 has three candidate reactions, and the R_0 is identified as a high quality reaction. However, the child molecule M_1 of R_0 is a unexpandable dead node with no further reaction candidates. If the planner makes the selection with observable next state molecular structures of M_1, M_2 and M_3 and unobservable reaction quality values of R_0, R_1 and R_2, it might selects R_2 for its most synthesizable next state molecule M_3. However, with observable reaction quality values, the planner could be misled into selecting R_0 due to its highest route quality expectation,which demonstrate an inconsistency between the two optimization obejectives.

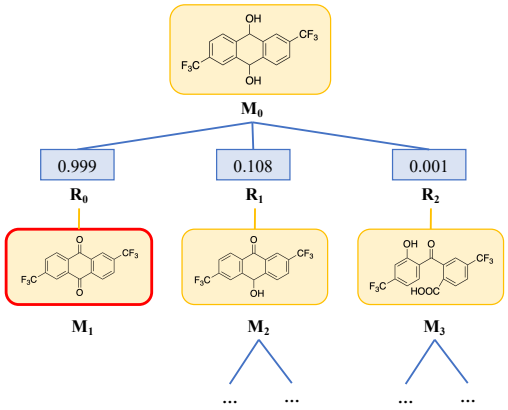

Figure 8: A case for illustrating two objective inconsistency. The root molecule M_0 has three candidate reactions, and the R_0 is identified as a high quality reaction. However, the child molecule M_1 of R_0 is a unexpandable dead node with no further reaction candidates.

G REAL-LIFE RETROSYNTHETIC PLANNING SCENARIOS

The quality metric required by our framework in a real-life scenario should be expensive but not prohibitively so. While a single-step model is not competent enough, a lab validation might be excessively expensive and time-consuming. This consideration constitutes the primary motivation behind our active planning framework, aiming to query a minimum number of reaction quality annotations while still planing high-quality routes.

While our current implementation involves querying the surrogate model, our inspiration is drawn directly from real-life retrosynthesis planning scenarios, such as in online softwares like SYNTHIA, where chemists are pivotal end users. In this context, integrating chemists as valuable resources into the AI planning process will be invaluable for planning routes that are not only feasible but also of practical high-quality. We envision the successful deployment of our framework in this scenario for several reasons.

Online annotation by chemists introduces minimal time delays and manageable labor costs, making it an ideal candidate for a route quality metric that is expensive but not prohibitively so. Our framework is intentionally designed to be compatible with various types of annotations, including a coarse-grained quality rating from 0 to 10. We believe such a rating is sufficient for the planner to make satisfactory decisions. Additionally, this rating can also be seamlessly integrated into our current framework by replacing the bucket index to which a quality value belongs (see details in Section 3.2 in the revision) with this discrete rating. Chemists contribute valuable insights beyond mere reaction yields, such as knowledge about preferred reactions in real-world synthesis contexts, which can include factors like toxicity, material costs and work-up difficulty (post-process, like purification or separation).

H FUTURE WORK

Although we focused on the high-quality routes, the retrosynthetic planning has other essential considerations like the green chemistry. In future work, we intend to investigate Active Retrosynthetic Planning with multi-objective optimization in order to find eco-friendly routes of high chemical feasibility.

I CASE STUDY

We conduct a double-blind test to check the route quality generated by **Retro**, ARP with **Retro***, **GRASP**, and ARP with **GRASP**. We collect top-1 successful routes from the experimental results of the benchmark dataset and the chemists tag the route with a rating from 0 to 10. 10 refers to a high-quality route while 0 refers to a low-quality one. We lie the average rating in Tab. 6. Compared with the original methods, ARP with **Retro*** outperforms **Retro*** by 1.7 and ARP with **GRASP** outperforms **GRASP** by 2.2.

	Retro*	ARP with **Retro***	**GRASP**	ARP with **GRASP**
Route rating(1-10)	7.8	8.5	6.9	9.1

Table 6: Double blind test on the top-1 route quality.

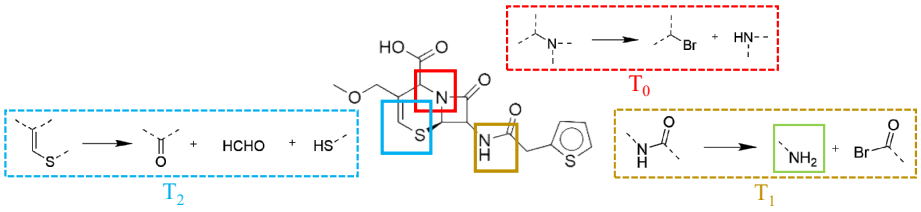

Figure 9: A target molecule.

Furthermore, we study a case to illustrate the active query capability. In Fig. 9, a target molecule has three basic molecular structures that need to be broken down by respective templates, T_0, T_1, T_2. Simplified,

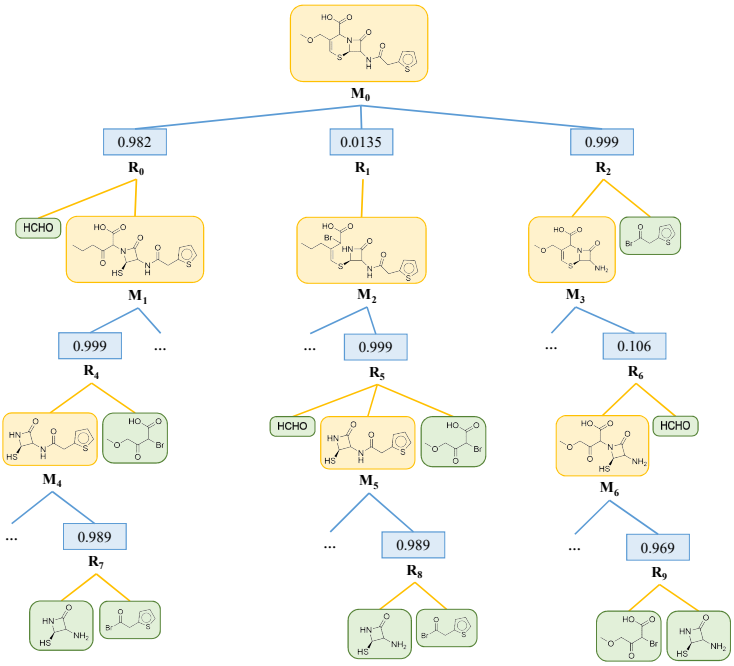

Figure 10: A search tree by ARP based on **GRASP**.

the planner needs to decide the order of executing three templates. However, if T_0 is executed after T_1, it will produce a low-quality reaction because T_1 reveals a high-activity amino group blocked green. From a chemical perspective, T_1 can be regarded as a deprotection reaction to suppress side reactions on the amino group for T_0. Thus T_0 must precede T_1. We visualize a search tree in Fig. 10 planned by ARP based on **GRASP** to solve the target molecule with the query cost equals 0, 0.01, and 0.05. For simplicity, we ignore some molecule nodes and reaction nodes. We tag the reaction qualities on the blue reaction nodes, the non-building block molecules on the yellow nodes, and the building block molecules on the green nodes. The empty blue nodes present reaction nodes of which the qualities are not annotated. Furthermore, we tag the Q value near the respective molecule nodes to explore the reaction quality annotation's impact. In the three search trees, the molecule node selection among M_0, M_1, and M_2 is a key decision that determines the next following expansion of the whole search tree. M_0 will results in a high-quality route while M_1 and M_2 will lead to low-quality routes. M_1 has a low preceding reaction quality and M_2 has a low future-quality expectation. M_1 is the best next molecule node to expand. We compare the situations with different query costs.

Ful observation: With a query cost of 0, the actor in ARP queries every reaction qualities in the search tree. The search tree is depicted in Fig. 11. Q values reflect the molecule's high-quality route expectation properly. The planner selects M_1 as the next molecule state node properly.

Partial observation: With a query cost of 0.01, the actor in ARP selects partial reaction qualities in the search tree to observe. The search tree is depicted in Fig. 12. It is observed that two reaction qualities are annotated. The molecule with the maximum Q value maintains M_1. Nevertheless, the unannotated reaction quality of R_2 misdirects the Q value estimate of M_3 to some extent. Though the ranking prior between

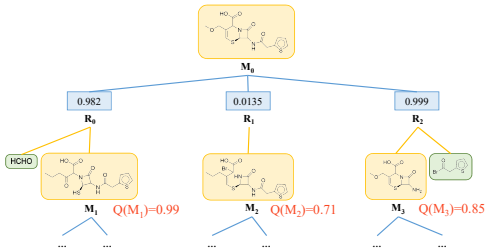

Figure 11: The search tree with query cost of 0.0

M_2 and M_3 changed compared with 11, the planner still selects M_1 to expand next. This phenomenon demonstrates the query ability of ARP to select the most impactful reactions to annotate qualities.

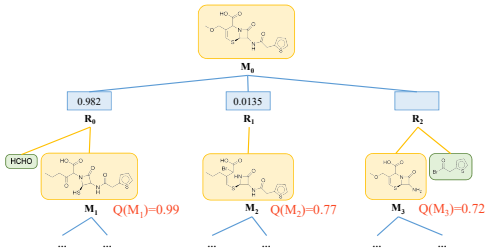

Figure 12: The search tree with query cost of 0.01

None observation: With a query cost of 0.05, the actor in ARP selects no reaction qualities in the search tree to observe. The search tree is depicted in Fig. 13. It is observed that the unannotated reaction qualities misdirect the Q value estimates of three molecules. In contrast to Fig. 11 and Fig. 12, the next selected molecule changed into M_2. This issue, on the one hand, illustrates how reaction qualities benefit retrosynthetic planning, on the other hand, proves the active capability of utilizing the reaction quality annotations to find high-quality routes.

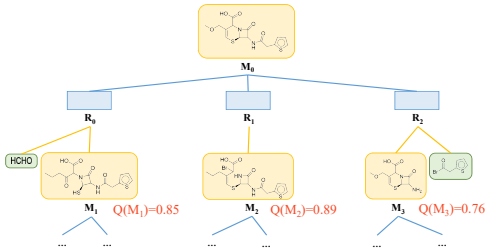

Figure 13: The search tree with query cost of 0.05