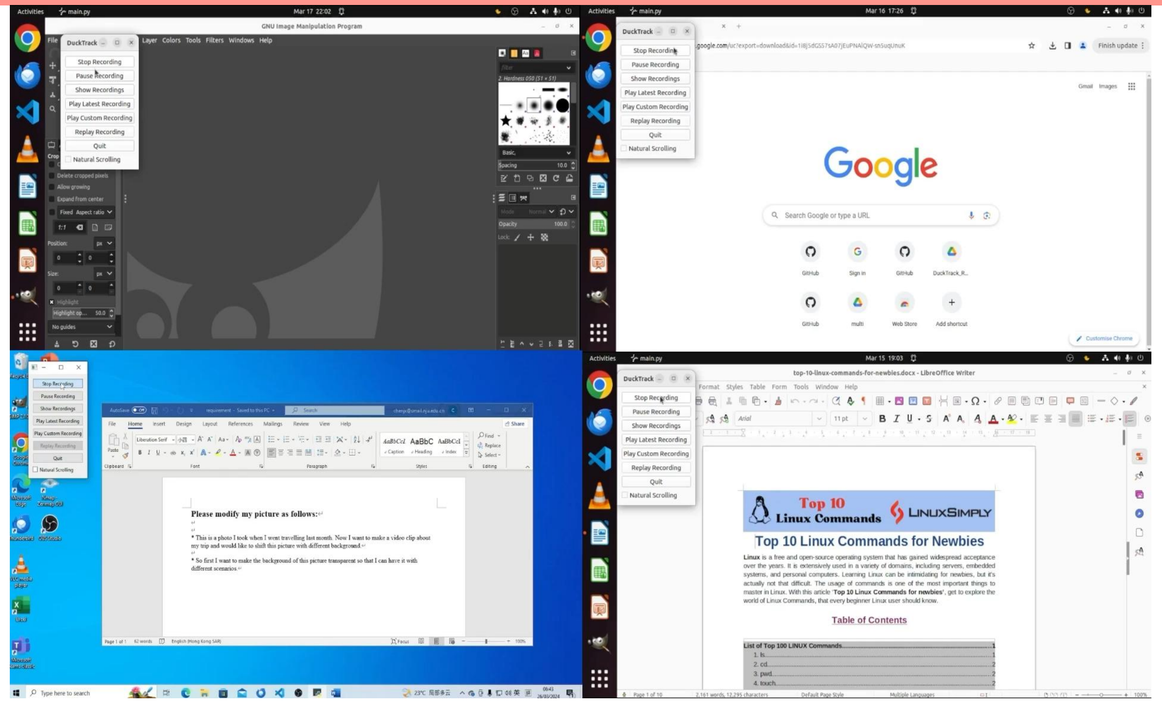
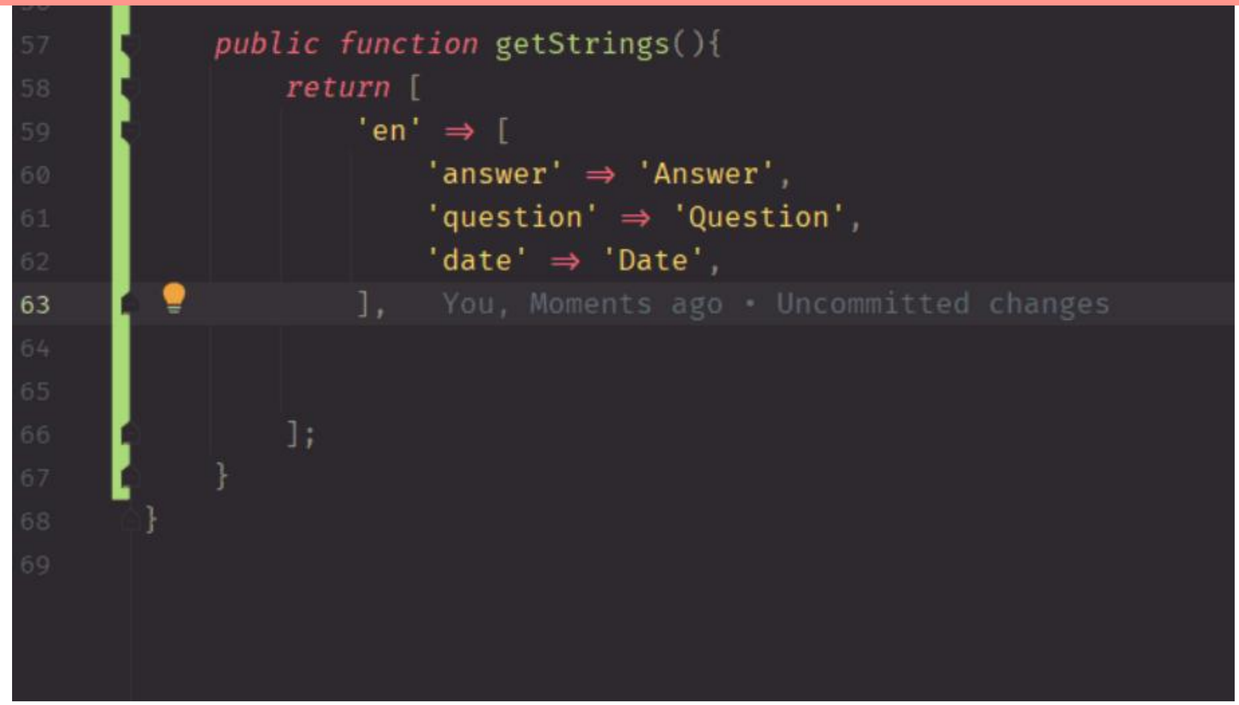
Large Language Model-augmented Optimization and Decision Programs

Haorui Wang, Georgia Tech CSE

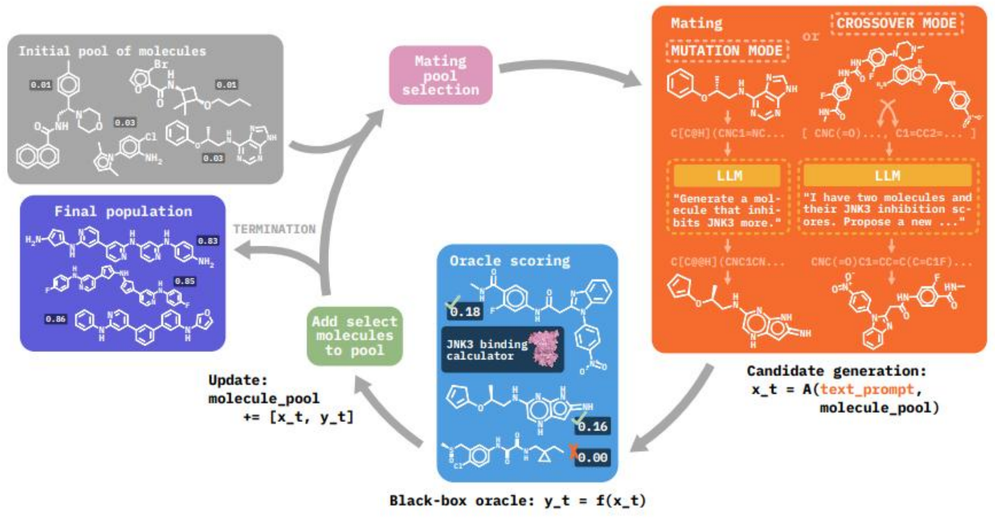
Large Language Models are Changing Everything!



Web agent



Code assistant



Molecule optimization

Do Large Language Models Understand Chemistry?

Table 2: The rank of fi
competitive, C: competit

Task	GPT-4	GPT-
Name Prediction	1	2
Property Prediction	1	2
Yield Prediction	1	3
Reaction Prediction	1	3
Reagents Selection	2	1
Retrosynthesis	2	3
Molecule Design	1	3
Molecule Captioning	1	2
Average rank	1.25	2.3

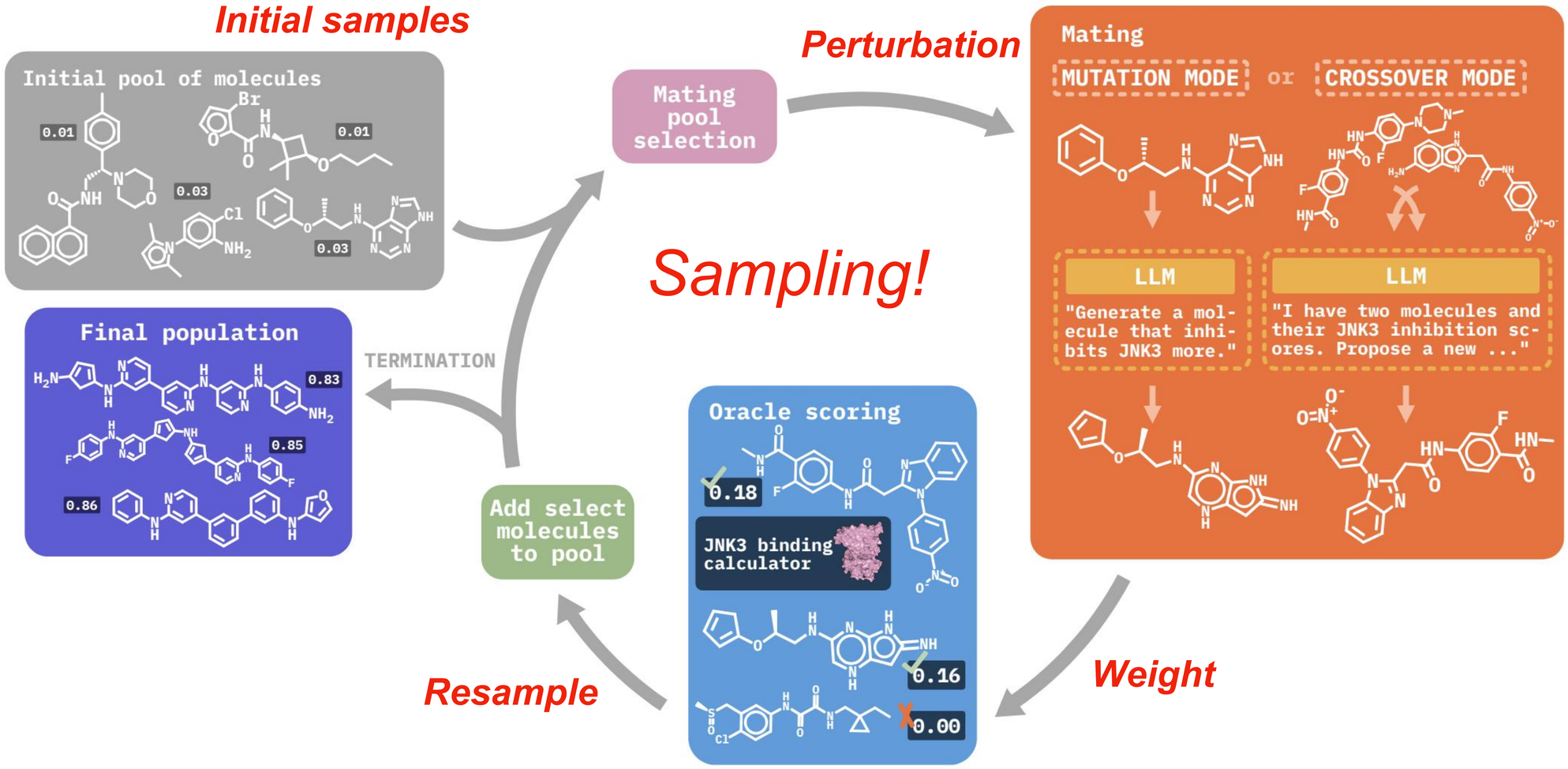
- 👍***Literature Review***: GPT-4 possesses extens such as density functional theory, Feynman dia theory, molecular dynamics simulations, and m
- 👍***Code Development***: GPT-4 is able to assist existing computational chemistry and physics
- 👍***Method Selection***: GPT-4 is able to recomm for specific research problems, taking into acc theory.
- 👍***Simulation Setup***: GPT-4 is able to aid in pre suggesting simulation parameters, including sp
- 👍***Experimental, Computational, and Theoretical Guidance***: GPT-4 is able to assist rese experimental, computational, and theoretical guidance.
- 😵***Hallucinations***: GPT-4 may occasionally generate incorrect information. It may struggle reasoning. Researchers need to independently verify and validate outputs and suggestion
- 😵***Raw Atomic Coordinates***: GPT-4 is not adept at generating or processing raw atomic c complex molecules or materials. However, with proper prompts that include molecular for supporting information, GPT-4 may still work for simple systems.
- 😵***Precise Computation***: GPT-4 is not proficient in precise calculations in our evaluated b usually ignores physical priors such as symmetry and equivariance/invariance. Currently, t numbers returned by GPT-4 may come from a literature search or few-shot examples. It is GPT-4 with specifically designed scientific computation packages or machine learning mo Graphormer and DiG.
- 😵***Hands-on Experience***: GPT-4 can only provide guidance and suggestions but cannot c experiments or run simulations. Researchers will need to set up and execute simulations c themselves or leverage other frameworks based on GPT-4, such as AutoGPT , HuggingGP on.

Table 3. Experimental results in terms of accuracy (%) on the textbook dataset. The best performing score is highlighted in **bold** and second-best is underlined. The average score is weighted by the number of problems in each textbook.

Model	Chemistry				Physics			Math			Avg.
	atkins	chemmc	quan	matter	fund	class	thermo	diff	stat	calc	
Zero-Shot Learning											
LLaMA-2-7B	0.00	0.00	0.00	0.00	1.37	0.00	0.00	2.00	5.33	0.00	1.03
LLaMA-2-70B	1.87	2.56	0.00	0.00	1.40	0.00	0.00	0.00	10.70	4.76	2.41
Mistral-7B	9.35	5.13	8.82	4.08	5.48	2.13	0.00	4.00	12.00	2.38	6.23
Claude2	15.00	12.83	14.71	10.20	12.33	6.40	9.00	4.00	38.70	16.70	14.94
GPT-3.5-Turbo	4.67	20.51	8.82	2.04	10.96	2.13	2.94	6.00	28.00	9.30	9.59
GPT-4	45.79	28.21	26.47	22.45	23.29	**25.53**	17.91	32.00	49.33	**54.76**	33.79
GPT-4-Turbo	**57.01**	**41.03**	**35.29**	**26.53**	**24.66**	21.28	**26.87**	**46.00**	**61.33**	52.38	**40.99**
Zero-Shot Learning + CoT Prompting											
LLaMA-2-7B	0.00	2.56	0.00	0.00	0.00	0.00	0.00	0.00	4.00	0.00	0.67
LLaMA-2-70B	0.93	2.56	0.00	0.00	0.00	0.00	1.49	0.00	10.70	0.00	1.89
Mistral-7B	6.54	5.13	2.94	0.00	0.00	2.12	1.49	6.00	10.67	9.52	4.63
Claude2	20.56	15.38	8.82	4.08	8.23	4.26	5.97	6.00	36.00	14.29	13.89
GPT-3.5-Turbo	6.54	23.08	2.94	10.20	12.33	2.12	5.97	12.00	33.33	9.30	12.17
GPT-4	28.04	**43.59**	14.71	20.41	21.92	19.15	17.91	22.00	50.67	42.86	28.52
GPT-4-Turbo	**60.75**	35.90	**29.41**	**28.57**	**30.14**	**31.91**	**25.37**	**38.00**	**64.00**	**54.76**	**42.37**
Few-Shot Learning + CoT Prompting											
LLaMA-2-7B	1.87	5.13	2.94	0.00	5.48	0.00	0.00	0.00	12.00	7.14	3.60
LLaMA-2-70B	13.10	12.83	14.71	4.08	12.33	0.00	0.00	0.00	13.30	9.52	8.40
Mistral-7B	6.54	10.26	2.94	2.04	2.74	2.13	4.48	4.00	14.67	9.52	6.17
Claude2	15.89	25.64	14.65	6.12	9.59	6.38	10.45	8.00	33.33	19.05	15.26
GPT-3.5-Turbo	8.41	20.51	8.82	6.12	10.96	2.12	1.49	10.00	38.67	6.98	11.99
GPT-4	41.12	33.33	17.65	16.33	17.81	17.02	20.90	30.00	49.33	45.24	30.36
GPT-4-Turbo	**59.81**	**35.90**	**26.47**	**18.37**	**23.29**	**19.15**	**32.84**	**32.00**	**65.33**	**50.00**	**39.45**
Few-Shot Learning + Python											
LLaMA-2-7B	0.93	2.56	0.00	0.00	0.00	0.00	0.00	0.00	6.67	0.00	1.20
LLaMA-2-70B	0.93	7.69	2.94	0.00	9.59	0.00	1.49	0.00	17.30	9.52	5.14
Mistral-7B	4.67	0.00	5.88	2.04	2.74	2.13	0.00	4.00	17.33	11.90	5.32
Claude2	6.54	12.82	14.71	4.08	17.81	8.51	5.97	20.00	40.00	16.67	14.92
GPT-3.5-Turbo	13.08	33.33	8.82	16.33	26.01	4.26	7.46	16.00	44.00	26.19	19.91
GPT-4	**57.01**	**38.46**	**44.12**	**34.69**	**28.77**	**23.40**	**34.33**	**44.00**	**68.00**	**38.10**	**43.22**
GPT-4-Turbo	32.71	33.33	17.65	26.53	27.40	12.76	16.42	34.00	42.67	30.95	28.47

Large Language Models as Samplers

Goal: "extract knowledge" from LLMs -> ill-defined!
Reframed as: finding molecules with optimal properties

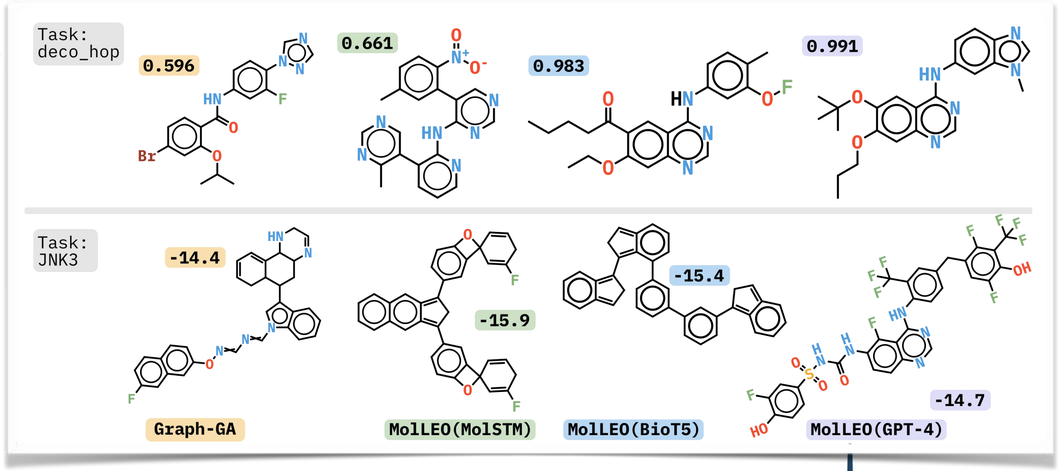


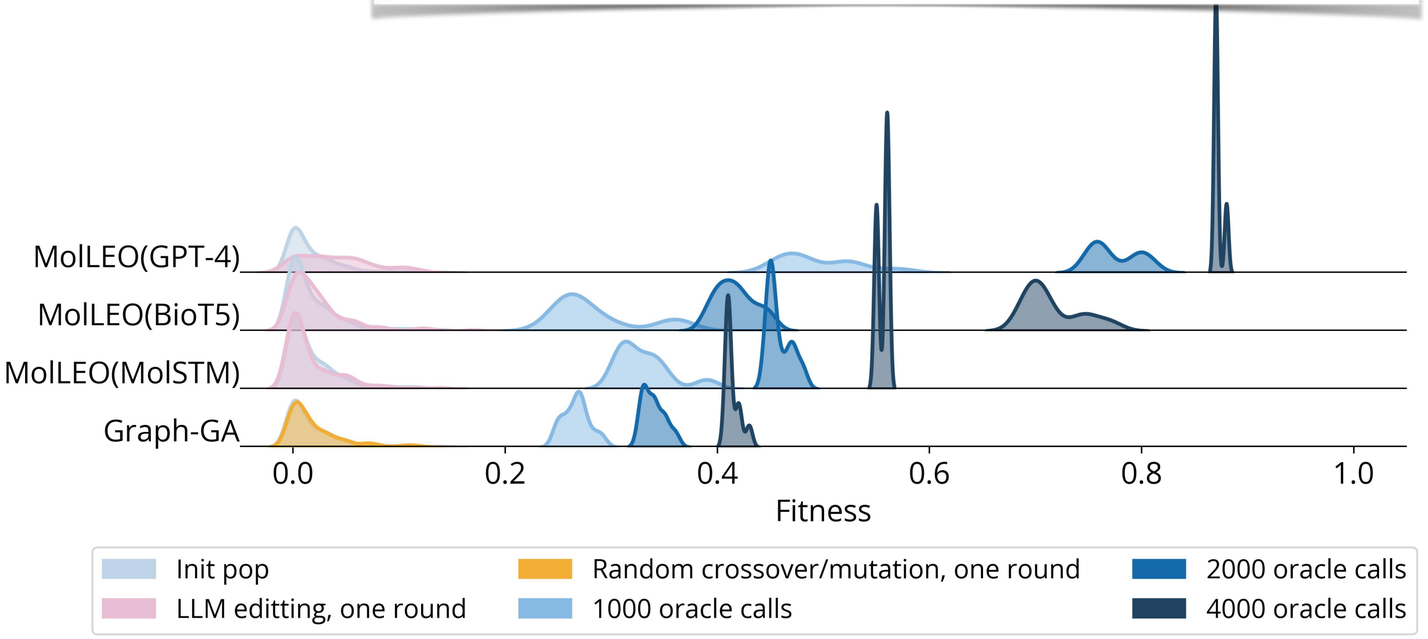
Large Language Models for Molecular Optimization

SOTA on 23 tasks!
Beat 25 strong baselines!

Task type	Method objective (↑)	REINVENT	Graph GA	GP BO	MolLEO (MolSTM)	MolLEO (BioT5)	MolLEO (GPT-4)
Property optimization	QED	0.941 ± 0.000	0.940 ± 0.000	0.937 ± 0.000	0.937 ± 0.002	0.937 ± 0.002	**0.948 ± 0.004**
	JNK3	0.783 ± 0.023	0.553 ± 0.136	0.564 ± 0.155	0.643 ± 0.226	0.728 ± 0.079	**0.790 ± 0.027**
	DRD2	0.945 ± 0.007	0.964 ± 0.012	0.923 ± 0.017	0.975 ± 0.003	**0.981 ± 0.002**	0.968 ± 0.012
	GSK3β	0.865 ± 0.043	0.788 ± 0.070	0.851 ± 0.041	**0.898 ± 0.041**	0.889 ± 0.015	0.863 ± 0.047
Name-based optimization	mestranol_similarity	0.618 ± 0.048	0.579 ± 0.022	0.627 ± 0.089	0.596 ± 0.018	0.717 ± 0.104	**0.972 ± 0.009**
	albuterol_similarity	0.896 ± 0.008	0.874 ± 0.020	0.902 ± 0.019	0.929 ± 0.005	0.968 ± 0.003	**0.985 ± 0.024**
	thiothixene_rediscovery	0.534 ± 0.013	0.479 ± 0.025	0.559 ± 0.027	0.508 ± 0.035	0.696 ± 0.081	**0.727 ± 0.052**
	celecoxib_rediscovery	0.716 ± 0.084	0.582 ± 0.057	0.728 ± 0.048	0.594 ± 0.105	0.508 ± 0.017	**0.864 ± 0.034**
	troglitazone_rediscovery	0.452 ± 0.048	0.377 ± 0.010	0.405 ± 0.007	0.381 ± 0.025	0.390 ± 0.044	**0.562 ± 0.019**
	perindopril_mpo	0.537 ± 0.016	0.538 ± 0.009	0.493 ± 0.011	0.554 ± 0.03		
	ranolazine_mpo	0.760 ± 0.009	0.728 ± 0.012	0.735 ± 0.013	0.725 ± 0.04		
	sitagliptin_mpo	0.021 ± 0.003	0.433 ± 0.075	0.186 ± 0.055	0.548 ± 0.06		
	amlodipine_mpo	0.642 ± 0.044	0.625 ± 0.040	0.552 ± 0.025	0.674 ± 0.01		
	fexofenadine_mpo	0.769 ± 0.009	0.779 ± 0.025	0.745 ± 0.009	0.789 ± 0.01		
	osimertinib_mpo	0.834 ± 0.046	0.808 ± 0.012	0.762 ± 0.029	0.823 ± 0.0(		
	zaleplon_mpo	0.347 ± 0.049	0.456 ± 0.007	0.272 ± 0.026	0.475 ± 0.01		
	median1	**0.372 ± 0.015**	0.287 ± 0.008	0.325 ± 0.012	0.298 ± 0.01		
	median2	0.294 ± 0.006	0.229 ± 0.017	**0.308 ± 0.034**	0.251 ± 0.03		
Structure-based optimization	isomers_c7h8n2o2	0.842 ± 0.029	0.949 ± 0.036	0.662 ± 0.071	0.948 ± 0.03		
	isomers_c9h10n2o2pf2cl	0.642 ± 0.054	0.719 ± 0.047	0.469 ± 0.180	0.871 ± 0.03		
	deco_hop	0.666 ± 0.044	0.619 ± 0.004	0.629 ± 0.018	0.613 ± 0.01		
	scaffold_hop	0.560 ± 0.019	0.517 ± 0.007	0.548 ± 0.019	0.527 ± 0.01		
	valsartan_smarts	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.0(		
	Total (↑)	14.036	13.823	13.182	14.557		
	Rank (↓)	4	5	6	3		

Table 1: Top-10 AUC of single-objective tasks. The best model for each t
three are underlined. We also report the sum of all tasks (total) and the ran

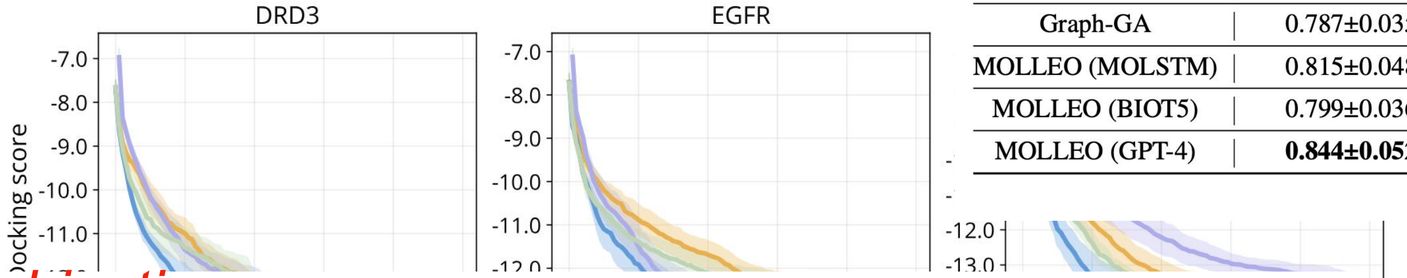




Large Language Models for Molecular Optimization

Optimize initial pool

Optimize docking scores



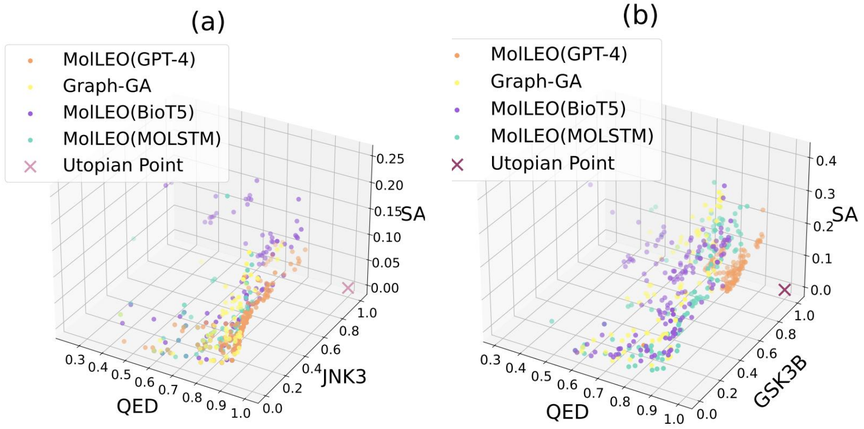
Model	JNK3 Top-10 AUC
Initial fitness	0.373±0.079
Graph-GA	0.787±0.035
MOLLEO (MOLSTM)	0.815±0.048
MOLLEO (BIOT5)	0.799±0.036
MOLLEO (GPT-4)	**0.844±0.052**

Table 3: Initializing MOLLEO with the best molecules from ZINC 250K [67]. The results of three different LLMs in MOLLEO and Graph-GA are compared. For all molecules in ZINC 250K, we run the JNK3 oracle and select the top 120 molecule pool. We run MOLLEO initializing from this pool of molecules and optimizing JNK3. We report the top-10 AUC on the output of MOLLEO. See the description of the models in the text.

Multi-objective optimization

Aggregate objective	Model	Task 1: QED (↑), JNK3 (↑), SAscore (↓) Sum	Task 1: QED (↑), JNK3 (↑), SAscore (↓) Hypervolume	Task 2: QED (↑), GSK3β (↑), SAscore (↓) Sum	Task 2: QED (↑), GSK3β (↑), SAscore (↓) Hypervolume	Task 3: QED (↑), JNK3 (↑), SAscore (↓),GSK3β (↓), DRD2 (↓) Sum	Task 3: QED (↑), JNK3 (↑), SAscore (↓),GSK3β (↓), DRD2 (↓) Hypervolume
Sum	Graph-GA	1.967 ± 0.088	0.713 ± 0.083	2.186 ± 0.069	0.719 ± 0.055	3.856 ± 0.075	0.162 ± 0.048
	MOLLEO (MOLSTM)	2.177 ± 0.178	0.625 ± 0.162	2.349 ± 0.132	0.303 ± 0.024	**4.040 ± 0.097**	0.474 ± 0.193
	MOLLEO (BIOT5)	1.946 ± 0.222	0.592 ± 0.199	2.306 ± 0.120	0.693 ± 0.093	3.904 ± 0.092	0.266 ± 0.201
	MOLLEO (GPT-4)	**2.367 ± 0.044**	**0.752 ± 0.085**	**2.543 ± 0.014**	**0.832 ± 0.024**	4.017 ± 0.048	**0.606 ± 0.086**
PO	Graph-GA	2.120 ± 0.159	0.603 ± 0.082	2.339 ± 0.139	0.640 ± 0.034	4.051 ± 0.155	0.606 ± 0.052
	MOLLEO (MOLSTM)	2.234 ± 0.246	0.472 ± 0.248	2.340 ± 0.254	0.202 ± 0.054	3.989 ± 0.145	0.381 ± 0.204
	MOLLEO (BIOT5)	2.325 ± 0.164	0.630 ± 0.120	2.299 ± 0.203	0.645 ± 0.127	3.946 ± 0.115	0.367 ± 0.177
	MOLLEO (GPT-4)	**2.482 ± 0.057**	**0.727 ± 0.038**	**2.631 ± 0.023**	**0.820 ± 0.024**	**4.212 ± 0.034**	**0.696 ± 0.029**

Table 2: Summation and hypervolume scores of multi-objective tasks. We report the results for two aggregation methods: Summation (Sum) and Pareto optimality (PO). The best model for each task is bolded.

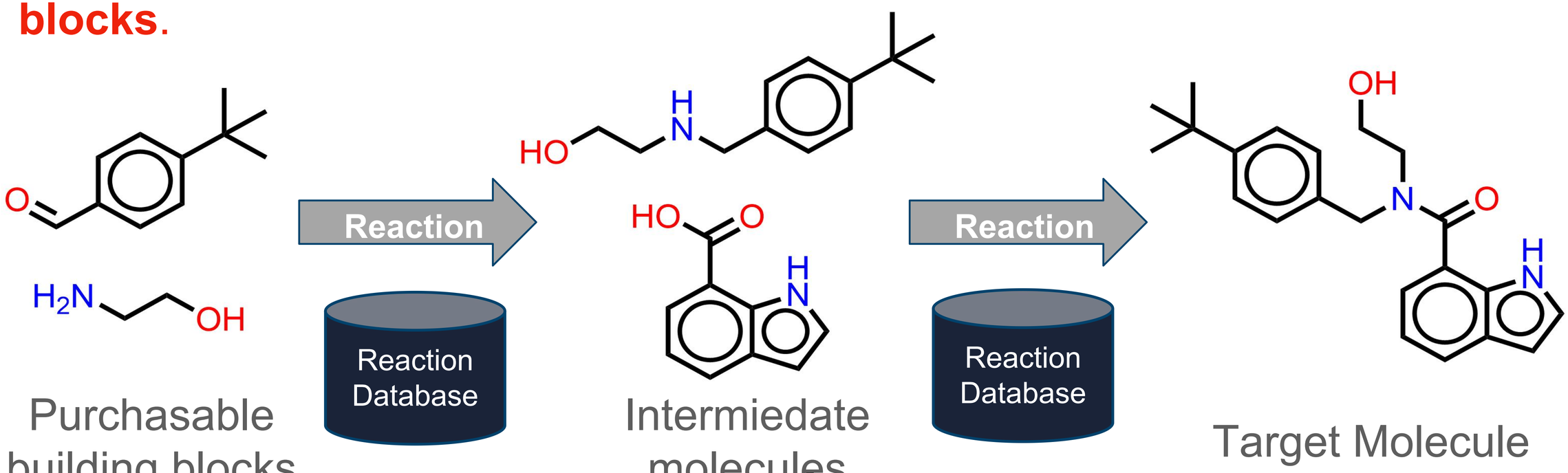


Retrosynthesis Planning

Complex reasoning: Can LLMs generate synthetic routes for a given molecule?

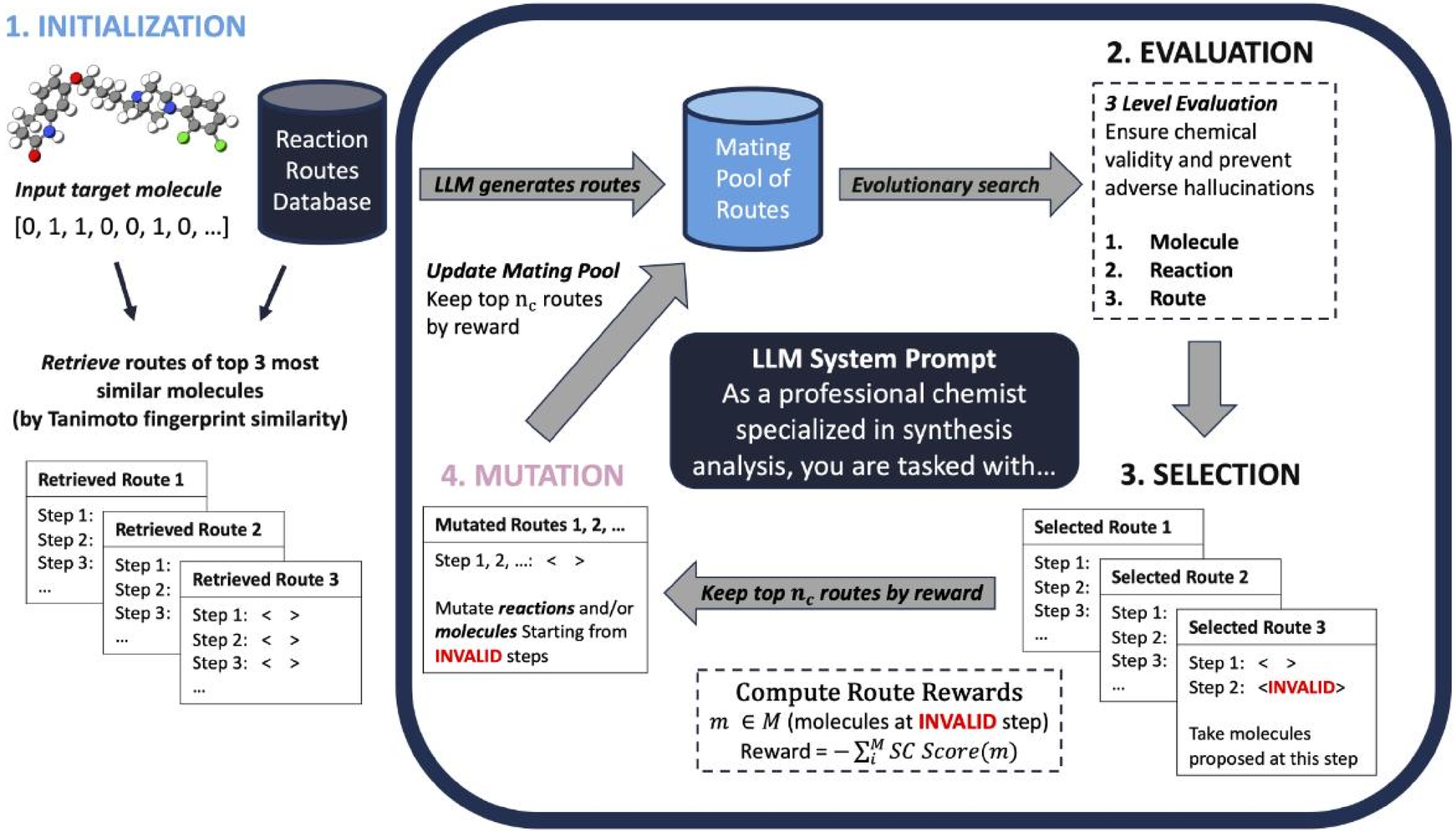
With reaction template set R and purchasable building blocks set C

Retrosynthesis planning: a sequential decision-making process, **starting from the target molecule** and **ending with a set of purchasable building blocks**.



Retrosynthesis Planning

LLM-Syn-Planner



Retrosynthesis Planning

LLM-Syn-Planner achieve comparable performance compared with specialized models

Algorithm	USPTO Easy			USPTO-190			Pistachio Reachable			Pistachio Hard		
	Solve Rate (%)			Solve Rate (%)			Solve Rate (%)			Solve Rate (%)		
	N=100	300	500	N=100	300	500	N=100	300	500	N=100	300	500
Graph2Edits(MCTS)	90.0	93.5	95.5	42.7	54.7	60.5	77.3	88.4	94.2	26.0	41.0	59.0
RootAligned(MCTS)	98.0	98.0	98.0	79.4	79.4	79.4	**99.3**	**99.3**	**99.3**	**83.0**	**83.0**	**83.0**
LocalRetro(MCTS)	92.5	94.5	95.5	44.3	50.9	58.3	86.7	90.0	95.3	52.0	55.0	62.0
Graph2Edits(Retro*)	92.0	95.5	97.0	51.1	59.4	78.5	94.0	95.0	95.5	71.0	74.0	79.0
RootAligned(Retro*)†	**99.0**	**99.0**	**99.0**	**86.8**	**86.8**	**86.8**	98.7	98.7	98.7	78.0	78.0	78.0
LocalRetro(Retro*)	95.5	97.5	98.0	51.0	65.8	73.7	97.3	**99.3**	**99.3**	63.0	69.0	72.0
LLM(MCTS)	54.5	68.5	75.5	25.8	27.2	31.3	12.7	17.3	20.7	0.0	4.0	5.0
LLM(Retro*)	56.0	69.0	75.5	23.2	26.8	30.6	14.7	19.3	13.3	0.0	2.0	5.0
LLM-Syn-Planner	91.0	98.0	98.5	64.7	70.0	80.5	93.3	94.7	96.7	72.0	73.0	80.0

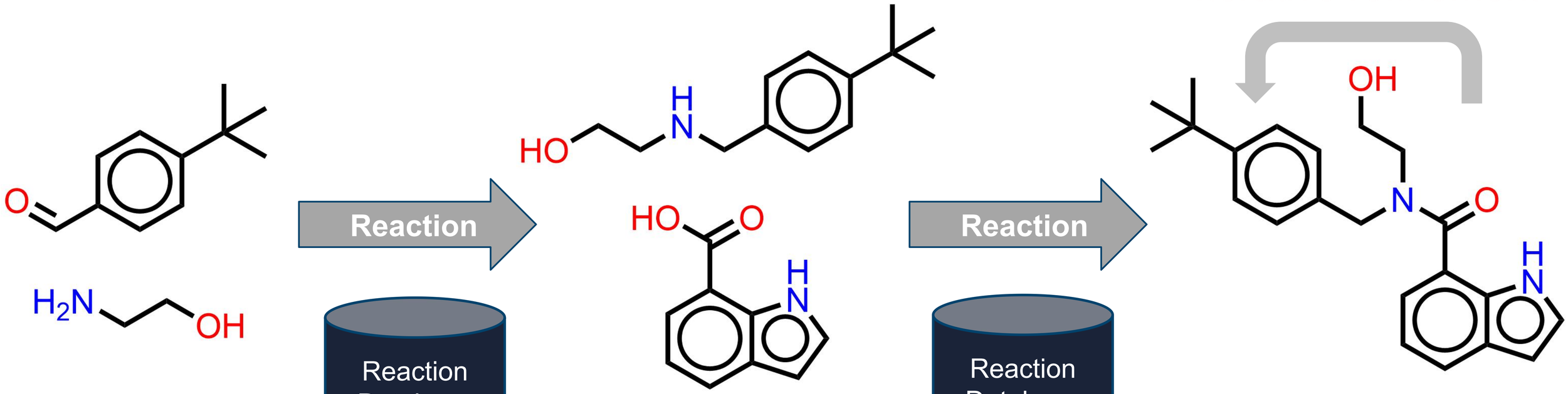
Baselines: Combine single-step specialized models with searching algorithms

LLM doesn’t work well as a single-step retrosynthesis-predictor

Retrosynthesis design

LLM-Syn-Designer: Optimize molecules towards a given property while keeping synthesizability

Iterative refinement



Purchasable building blocks

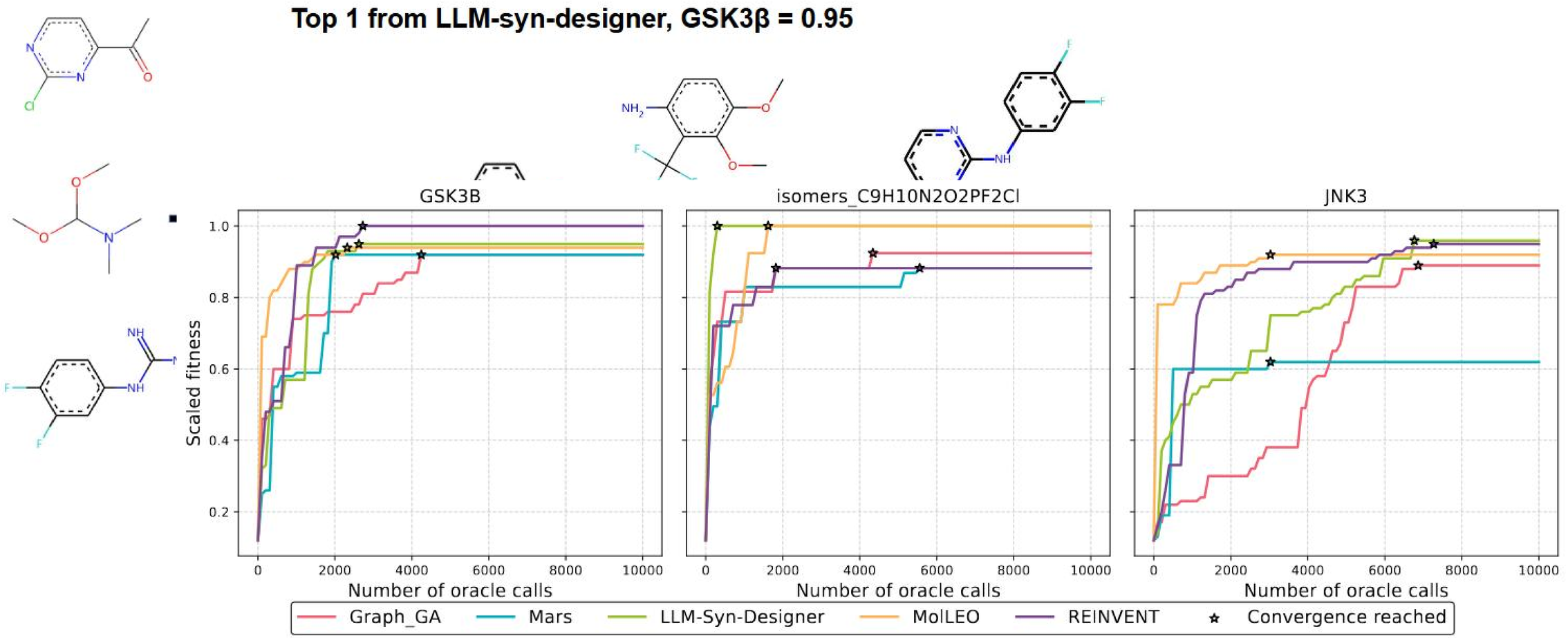
Interneidate molecules

Target Molecule

Combining MolLEO and LLM-Syn-Planner together:
filter out molecules with SC score > 3 in each iteration

Retrosynthesis design

LLM-Syn-Designer: Optimize molecules towards a given property while keeping synthesizability



Thank you for listening!

Codes are all public!

Questions and discussions!

hwang984@gatech.edu

Thanks to Yuanqi Du for providing parts of the content and valuable suggestions for this presentation.