Molecular Synthesizability and Synthetic Tree Generation for Synthesizable Molecular Design

Wenhao Gao

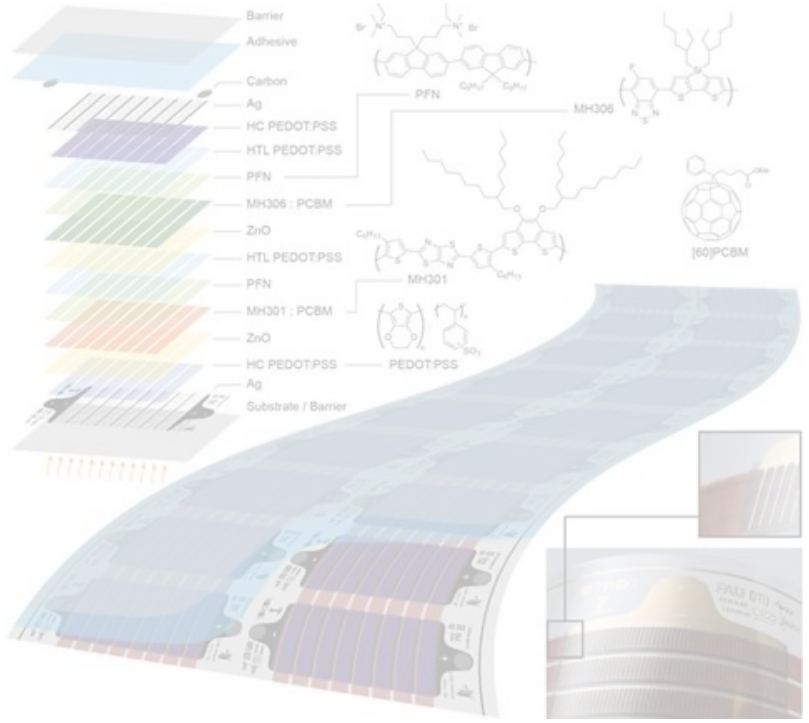
Chemical Engineering, MIT

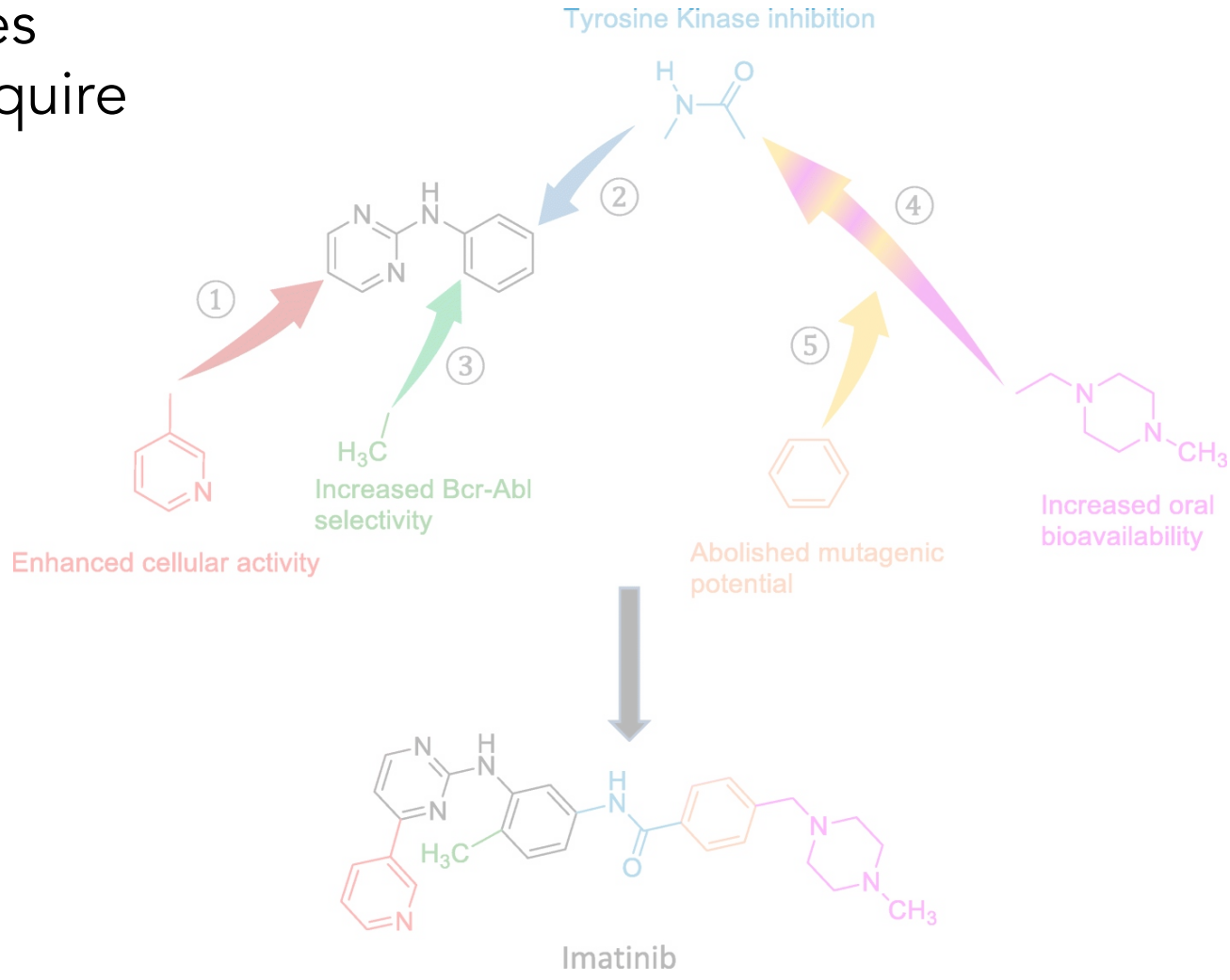
01/26/2022

Gao, W., & Coley, C. W. (2020). The synthesizability of molecules proposed by generative models. *JCIM*, *60*(12), 5714-5723.
Gao, W., Mercado, R., & Coley, C. W. (2021). Amortized Tree Generation for Bottom-up Synthesis Planning and Synthesizable Molecular Design. *ICLR 2022*.

Molecular design: key to many societal challenges

- Properties are fully determined by structure.
- Solutions to many of grand challenges (health, energy, sustainability, etc) require novel functional molecules.

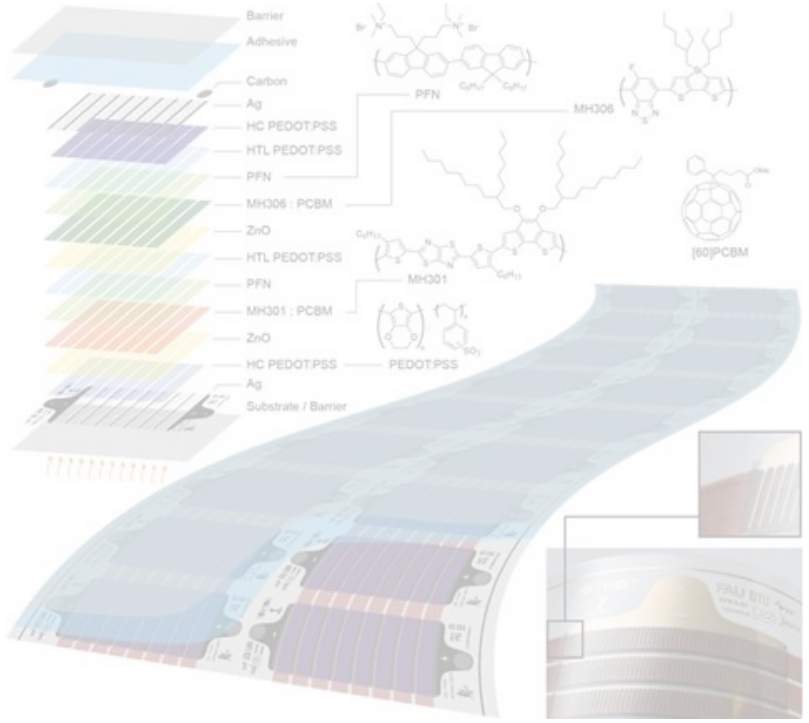


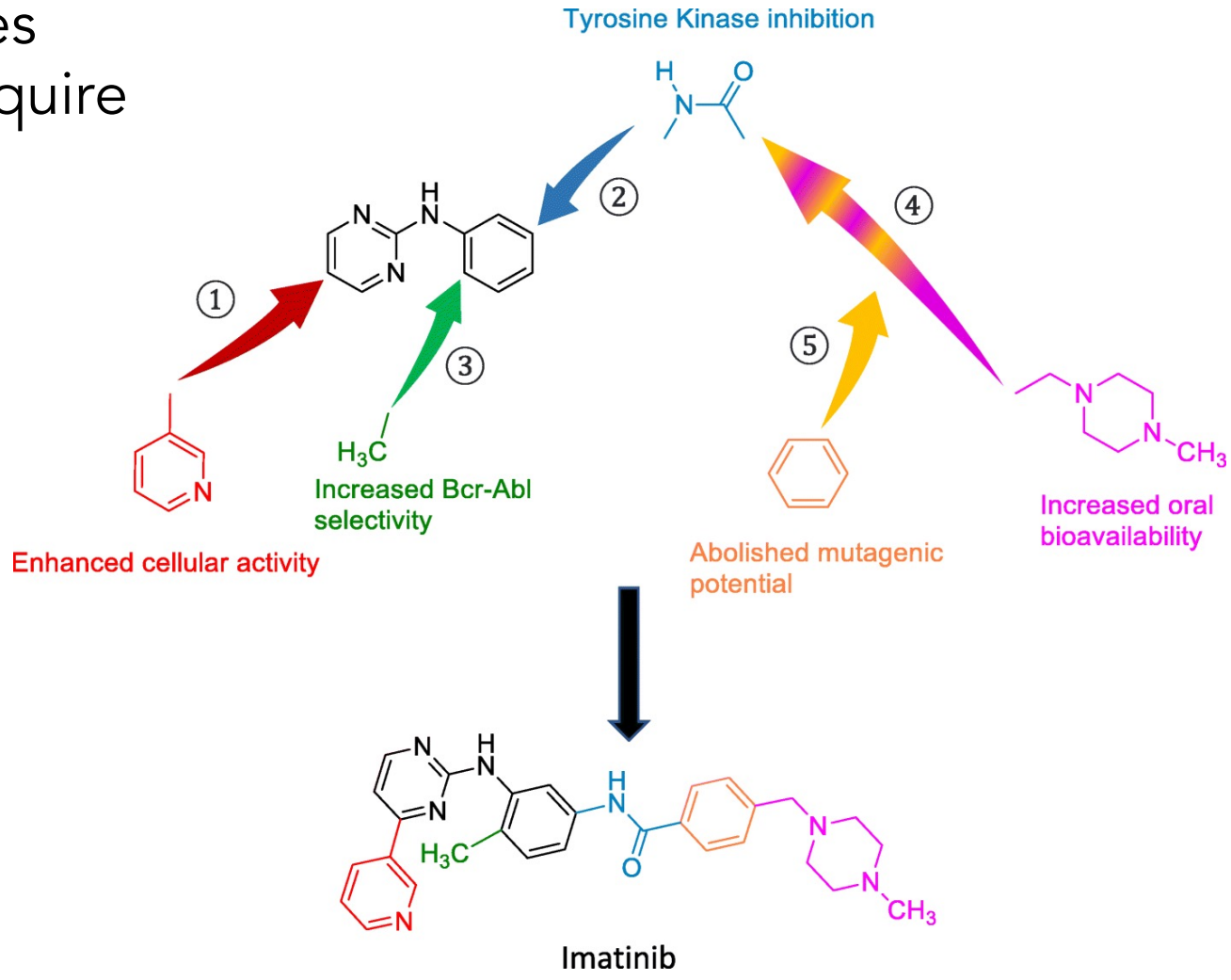


Rossari, F., Minutolo, F., & Orciuolo, E. (2018). *Journal of hematology & oncology*, *11*(1), 1-14.
https://www.chemistryworld.com/news/organic-solar-cells-reach-manufacturing-milestone/7439.article

Molecular design: key to many societal challenges

- Properties are fully determined by structure.
- Solutions to many of grand challenges (health, energy, sustainability, etc) require novel functional molecules.

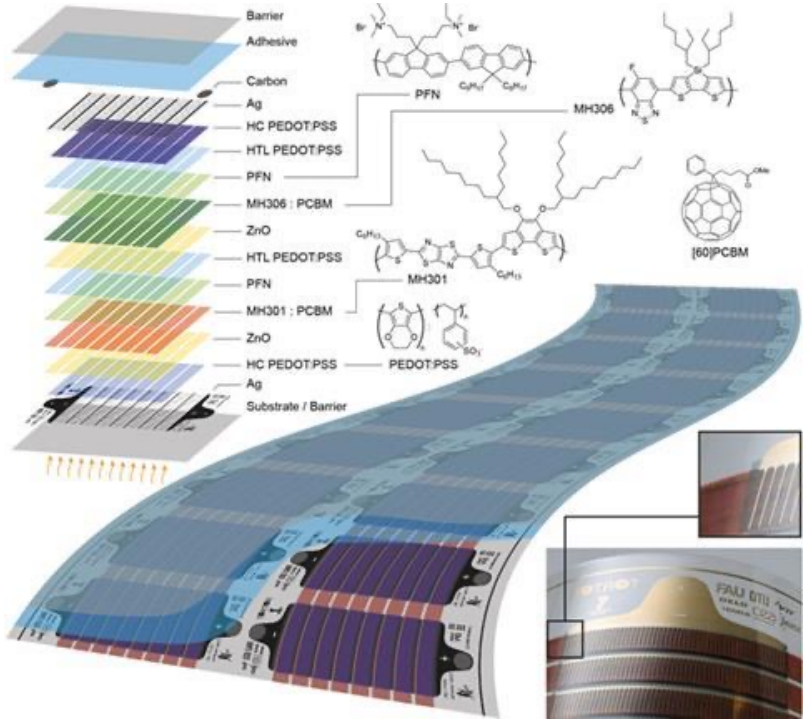


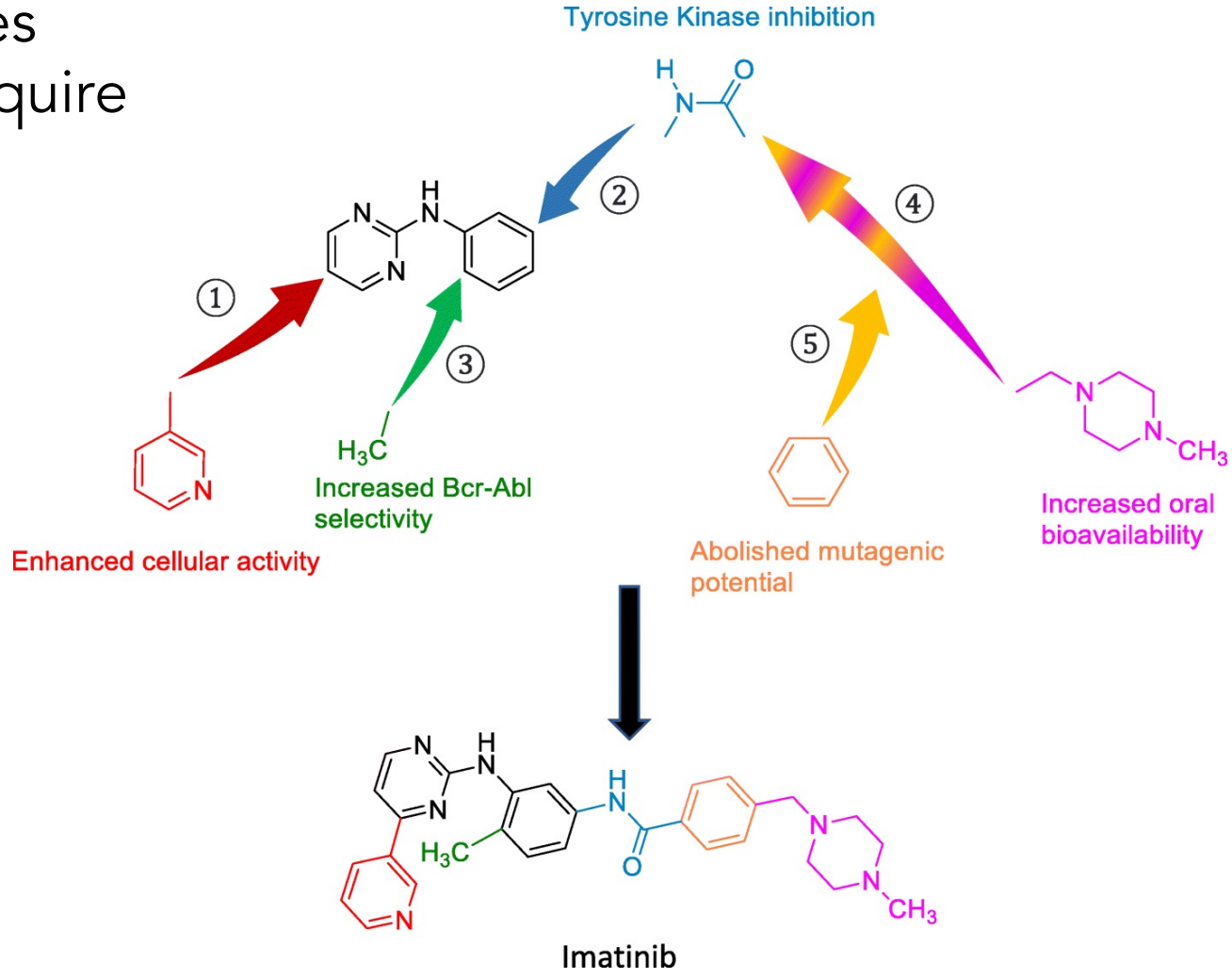


Rossari, F., Minutolo, F., & Orciuolo, E. (2018). *Journal of hematology & oncology*, *11*(1), 1-14.
https://www.chemistryworld.com/news/organic-solar-cells-reach-manufacturing-milestone/7439.article

Molecular design: key to many societal challenges

- Properties are fully determined by structure.
- Solutions to many of grand challenges (health, energy, sustainability, etc) require novel functional molecules.

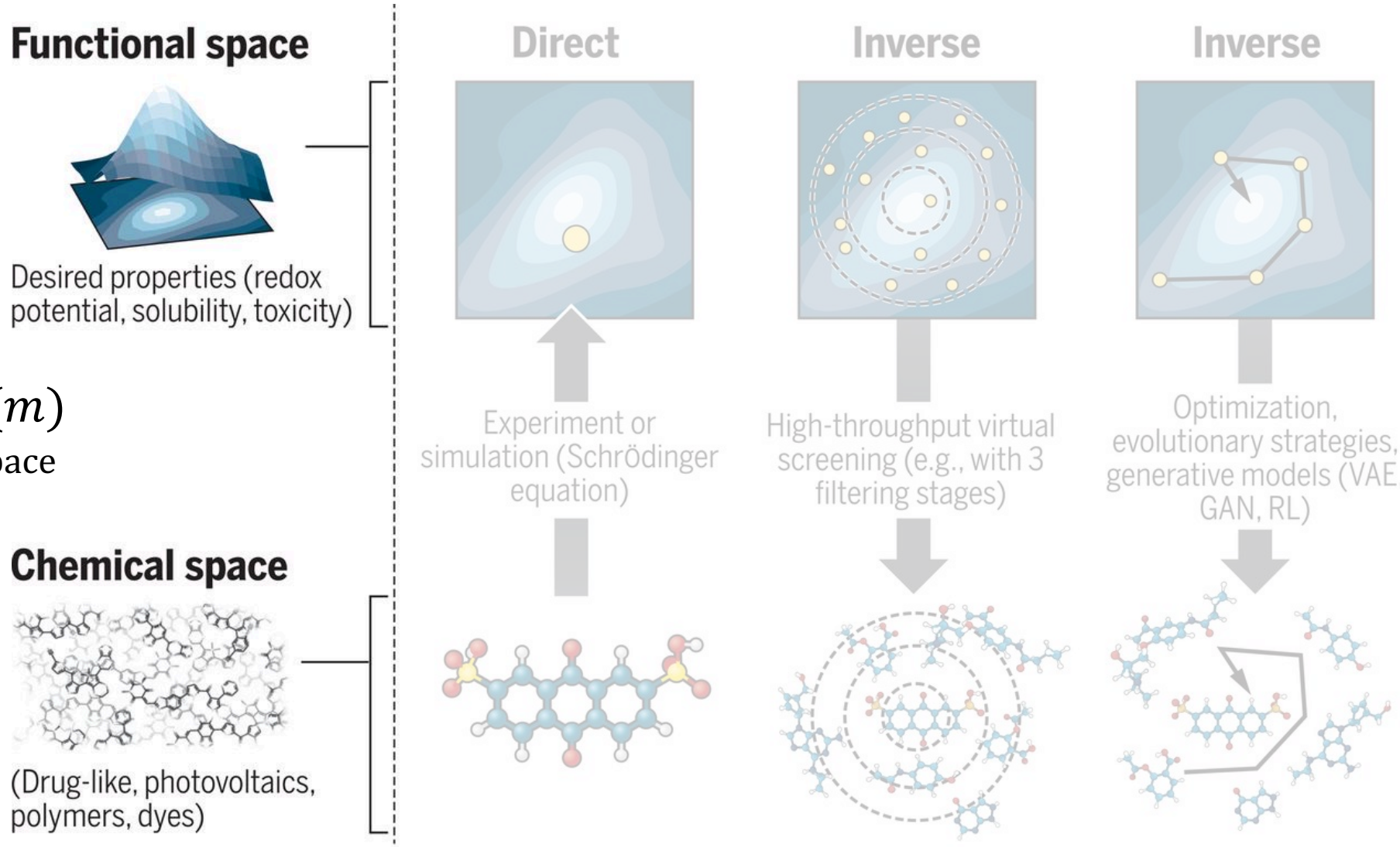




Rossari, F., Minutolo, F., & Orciuolo, E. (2018). *Journal of hematology & oncology*, *11*(1), 1-14.
https://www.chemistryworld.com/news/organic-solar-cells-reach-manufacturing-milestone/7439.article

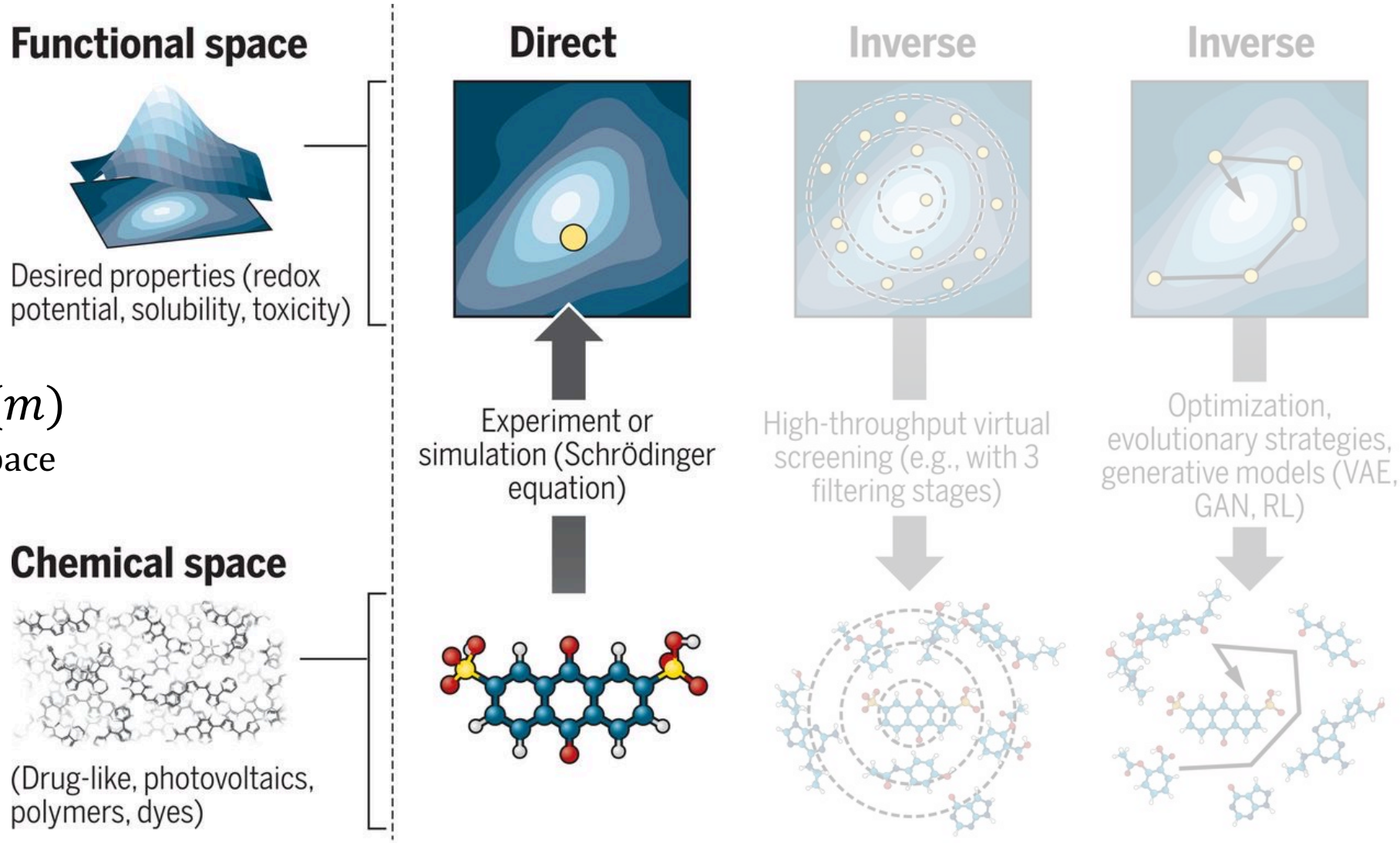
Molecular design: screening and *de novo* design

$$m = \text{argmax}\, f(m)$$
$$m \in \text{Chemical Space}$$



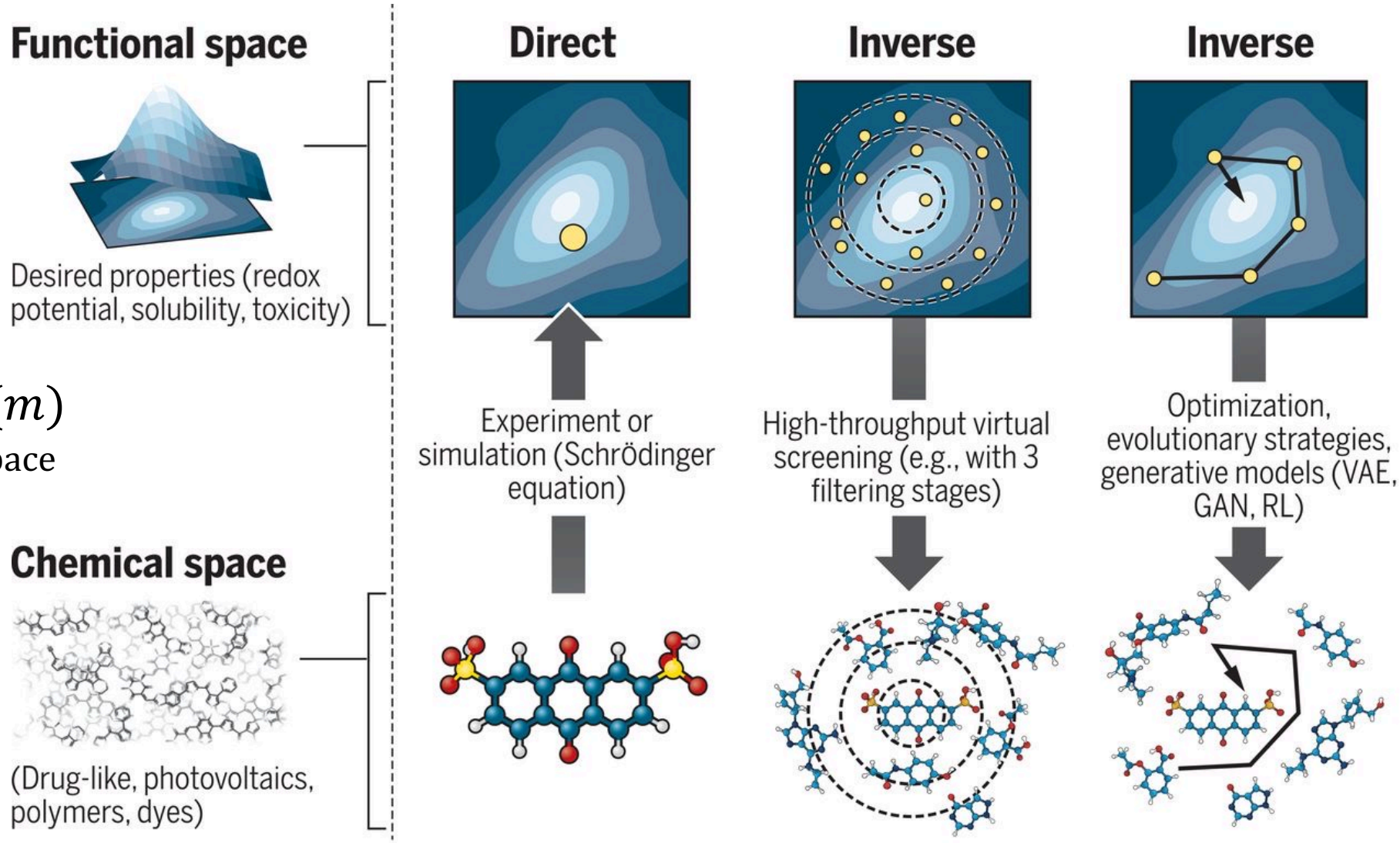
Molecular design: screening and *de novo* design

$$m = \text{argmax}\, f(m)$$
$$m \in \text{Chemical Space}$$



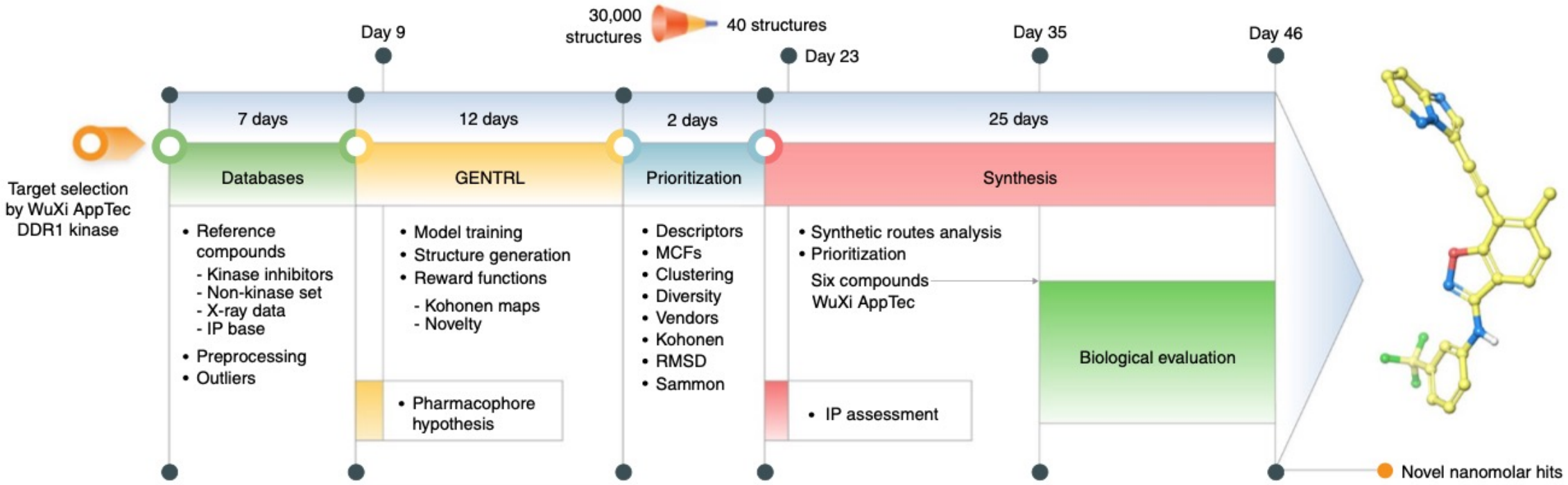
Molecular design: screening and *de novo* design

$$m = \text{argmax}\, f(m)$$
$$m \in \text{Chemical Space}$$



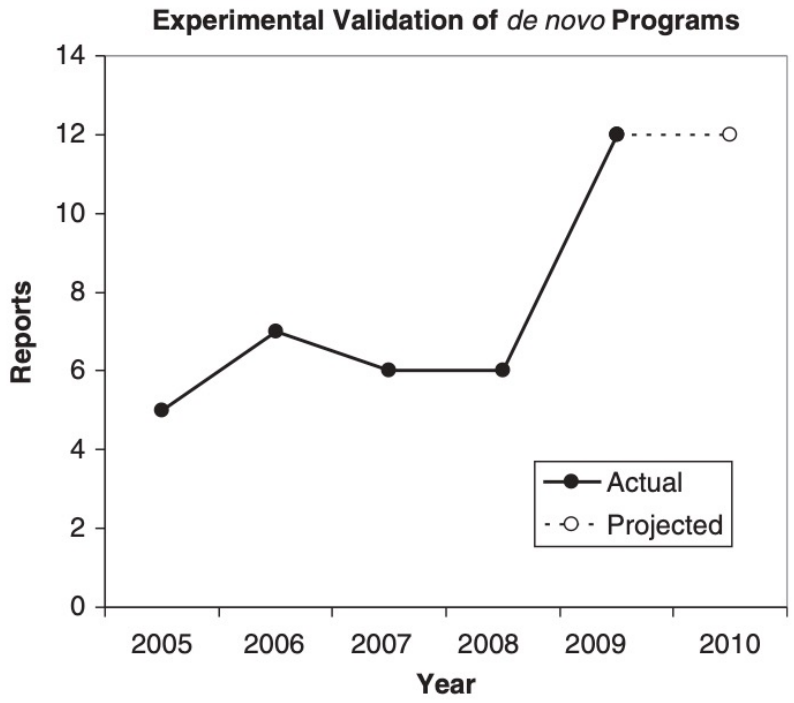
A successful example in drug discovery

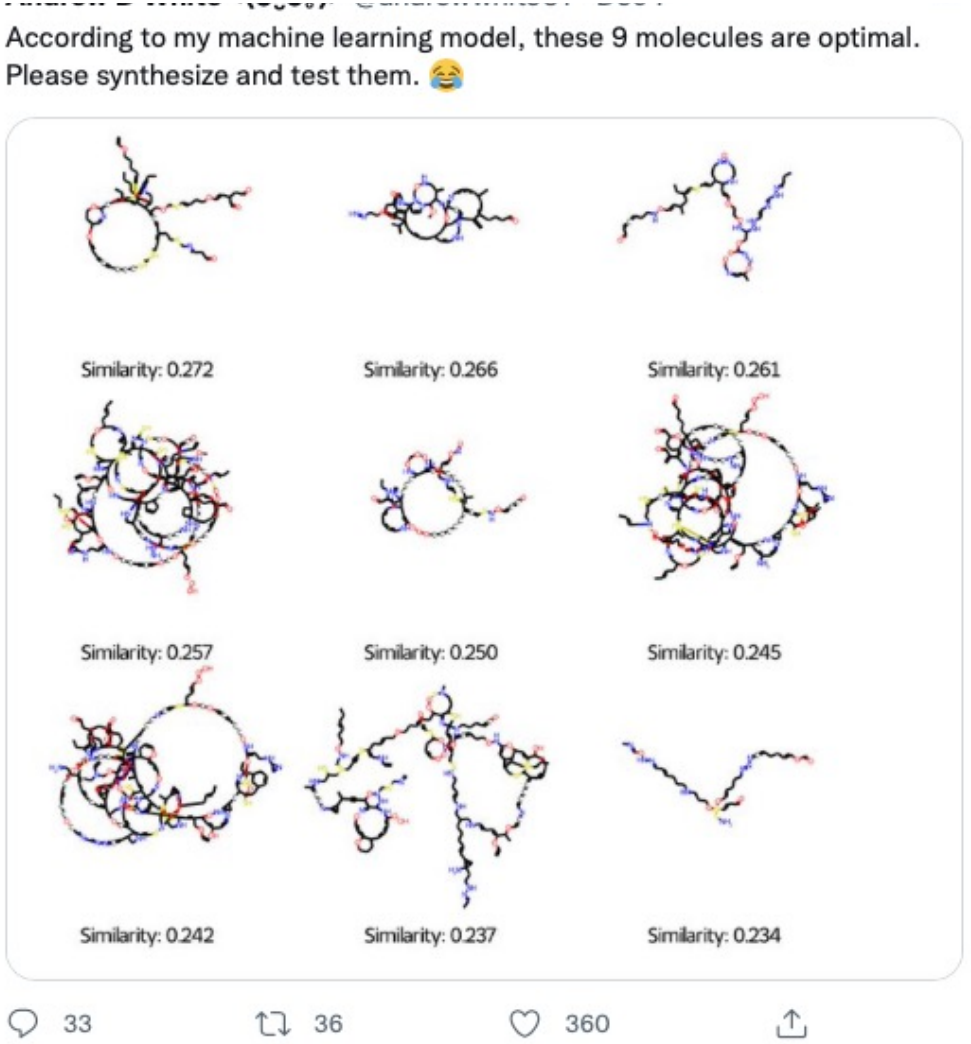
- Zhavoronkov et al. applied GENTRL successfully identified an inhibitor for DDR1 kinase.
- Years → 1.5 months



Synthesizability: a major bottleneck

- Zhavoronkov et al. manually selected only 6 molecules from 40 based on synthetic accessibility (initial generated pool: 30,000)



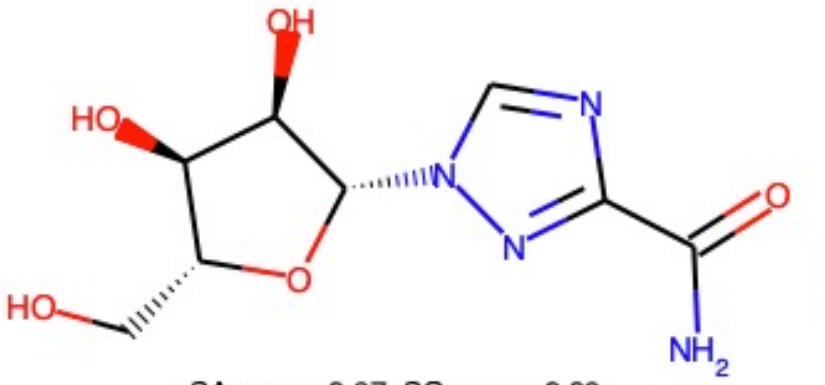


Kutchukian, Peter S., and Eugene I. Shakhnovich. *Expert opinion on drug discovery* 5.8 (2010): 789-812.
https://twitter.com/andrewwhite01/status/1468302573950148609

Synthesizability is not just another metric

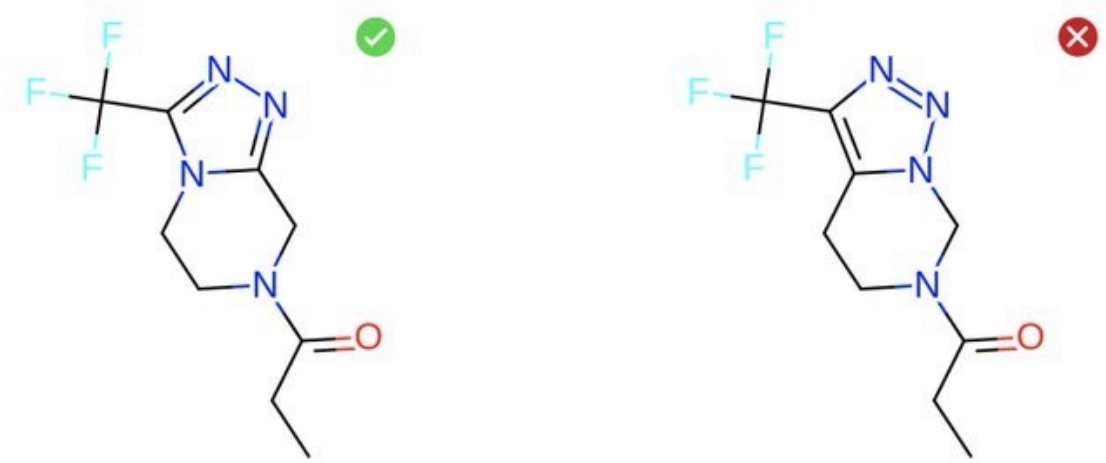
Challenges:

- Intuitive and subjective concept
- Highly “nonlinear” w.r.t. structure
- Sensitive to chemical availability

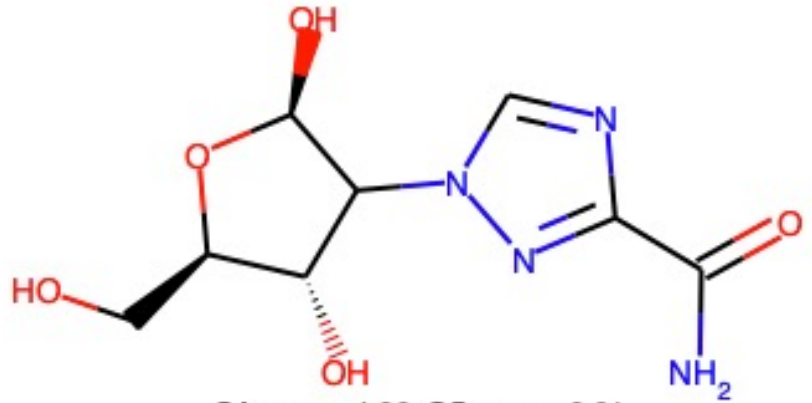


SA-score: 3.87, SC-score: 2.98

✓



The target on the left has a pretty straightforward synthesis, whereas the one on the right poses a significant challenge due to an unusual moiety. Our technology allows distinguishing between them extremely quickly.



SA-score: 4.22, SC-score: 2.91

X

- Sheridan, Robert P., et al. *Journal of chemical information and modeling* 54.6 (2014): 1604-1616.
- Ertl, Peter, and Ansgar Schuffenhauer. Journal of cheminformatics 1.1 (2009): 8.
- Coley, Connor W., et al. Journal of chemical information and modeling 58.2 (2018): 252-261.

Synthesizability is not just another metric

Former attempts:

- Crowd source scoring (e.g. meanComplexity)
- Based on structural complexity (e.g. SA_Score)
- Based on synthetic pathway (e.g. SCScore)

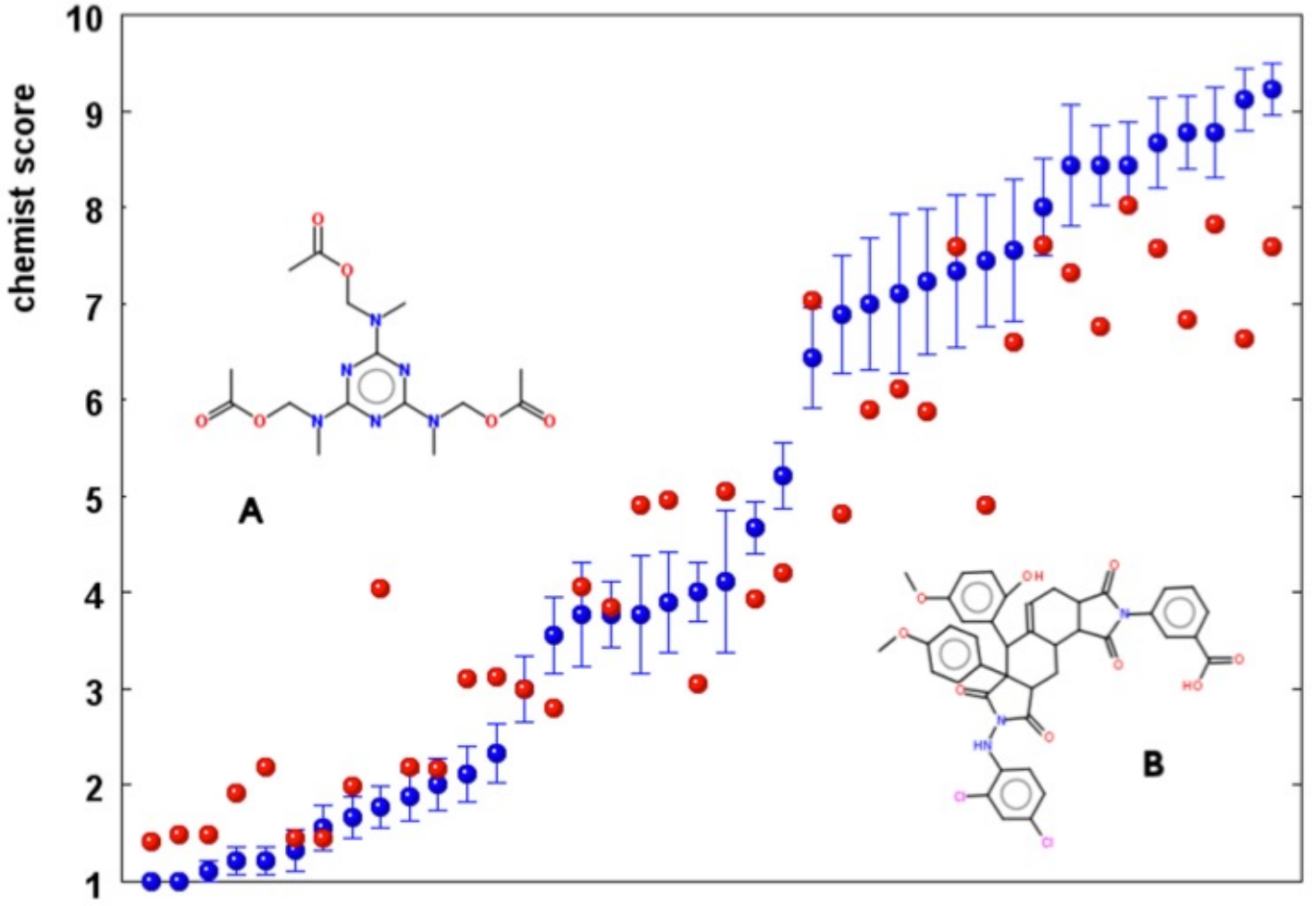
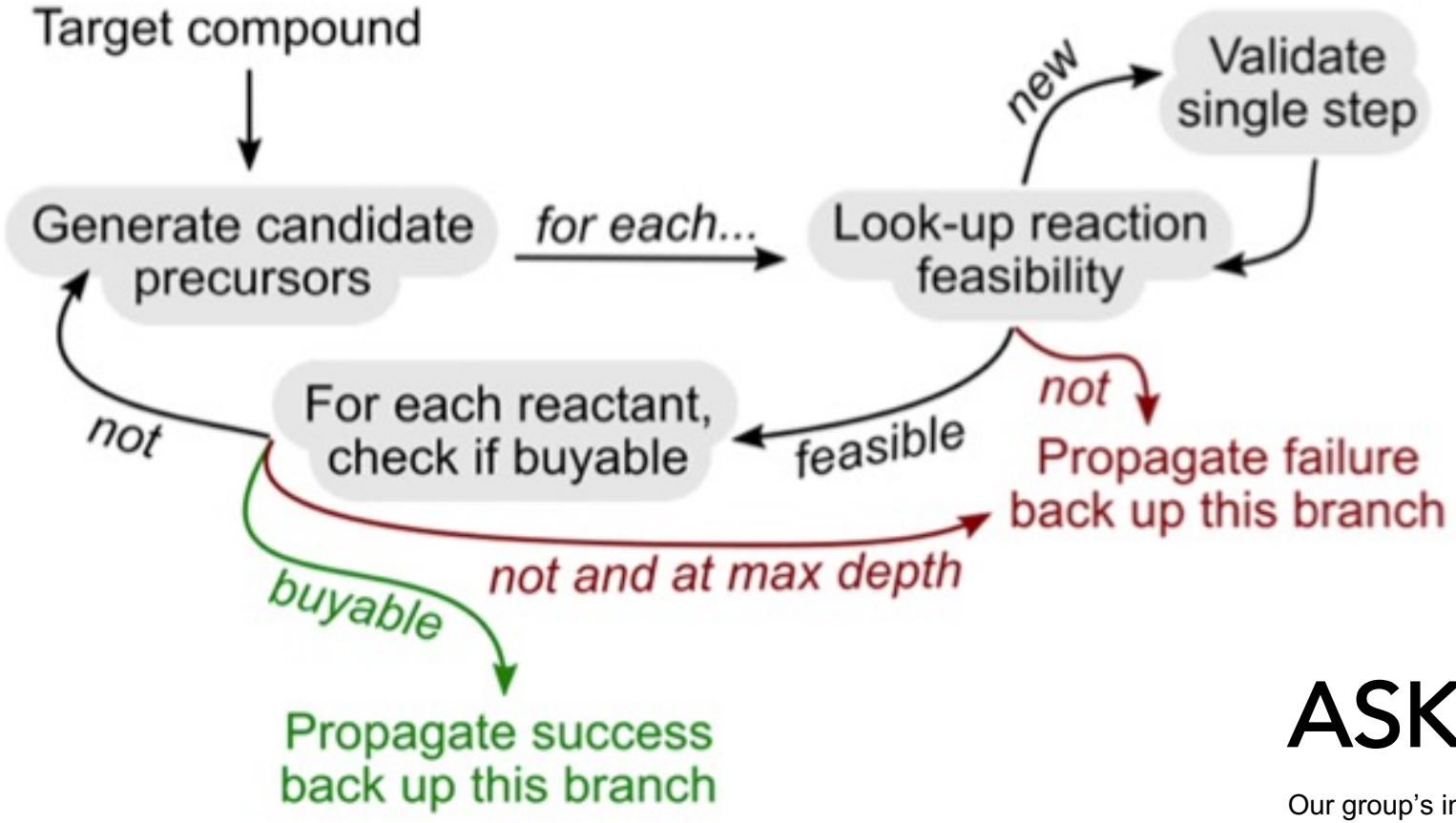


Figure 5
Average of chemist ranks for 40 test molecules (blue) compared with the computed SAscore (red). Error bars on blue points indicate standard error of mean of estimations by 9 chemists.

- Sheridan, Robert P., et al. *Journal of chemical information and modeling* 54.6 (2014): 1604-1616.
- Ertl, Peter, and Ansgar Schuffenhauer. Journal of cheminformatics 1.1 (2009): 8.
- Coley, Connor W., et al. Journal of chemical information and modeling 58.2 (2018): 252-261.

Computer-aided synthesis analysis (CASP)



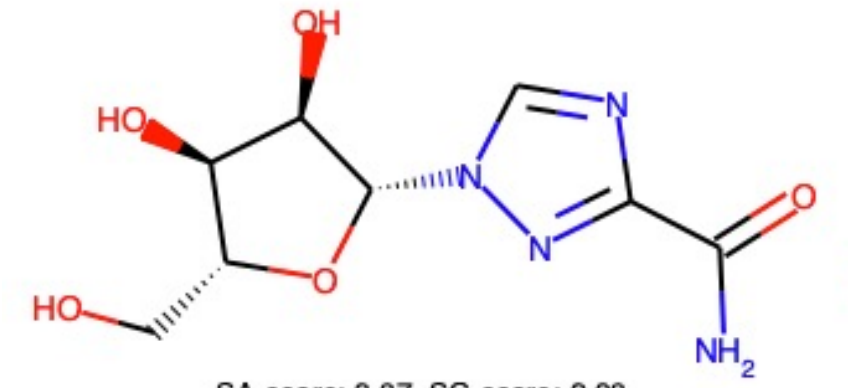
ASKCOS

Our group's implementation

CASP serves as a synthesizability measurement

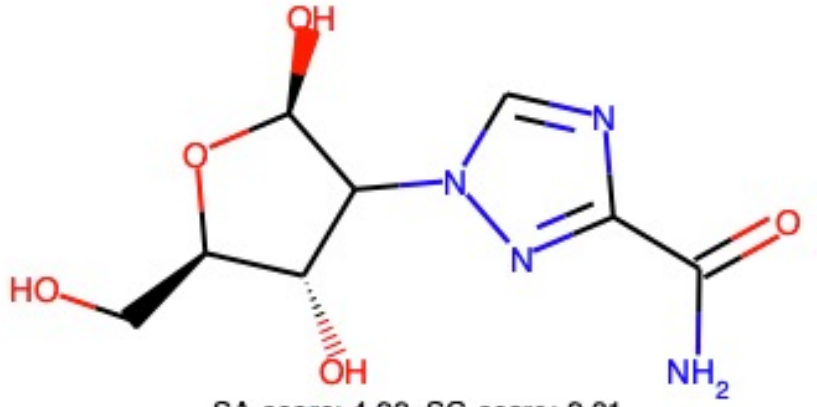
- Computer-Aided Synthesis Planning (CASP) is an alternative to expert scoring:
  - Capture the high ``non-linearity'' of synthesizability.
  - Recommend actionable synthetic pathways.
  - Can be accessed unlimitedly.

  But:
  - Time-consuming (~1min/molecule)



SA-score: 3.87, SC-score: 2.98

✓

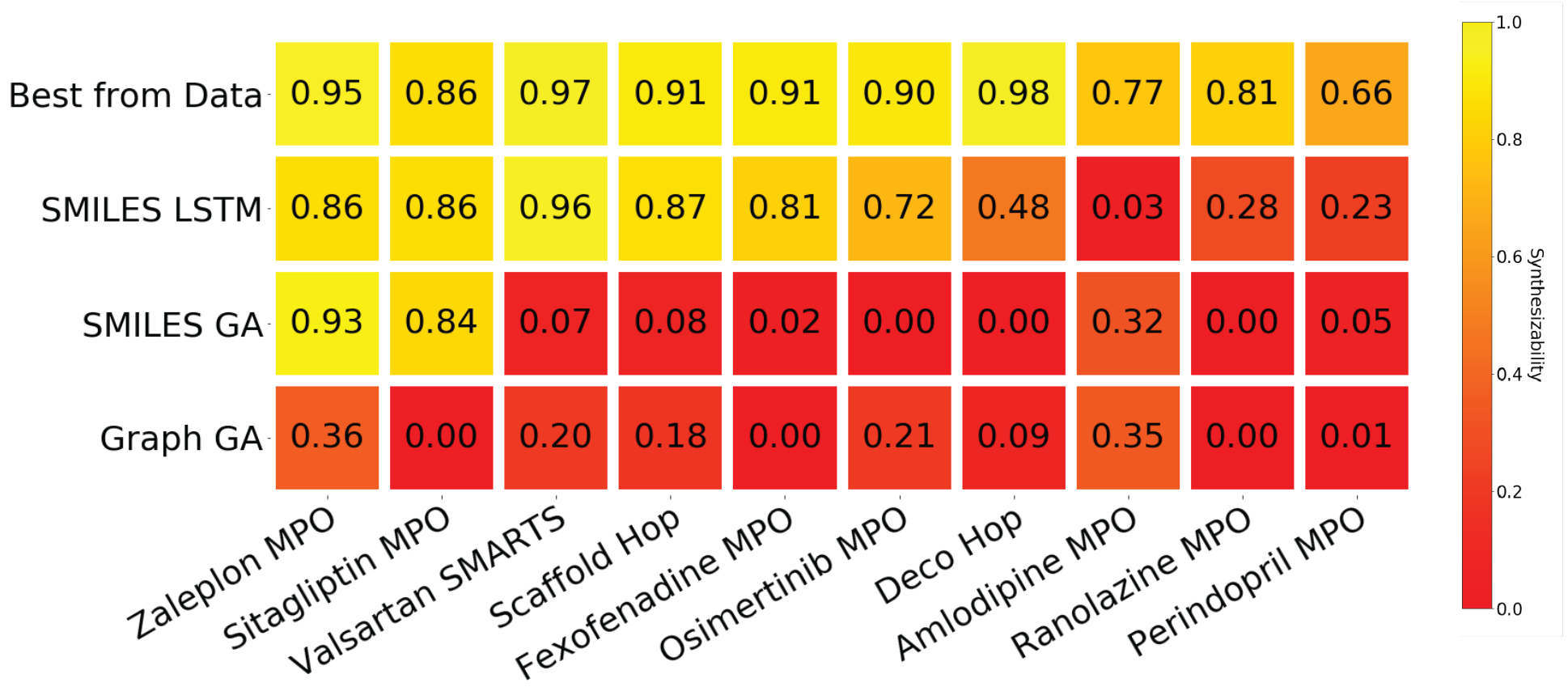


SA-score: 4.22, SC-score: 2.91

X

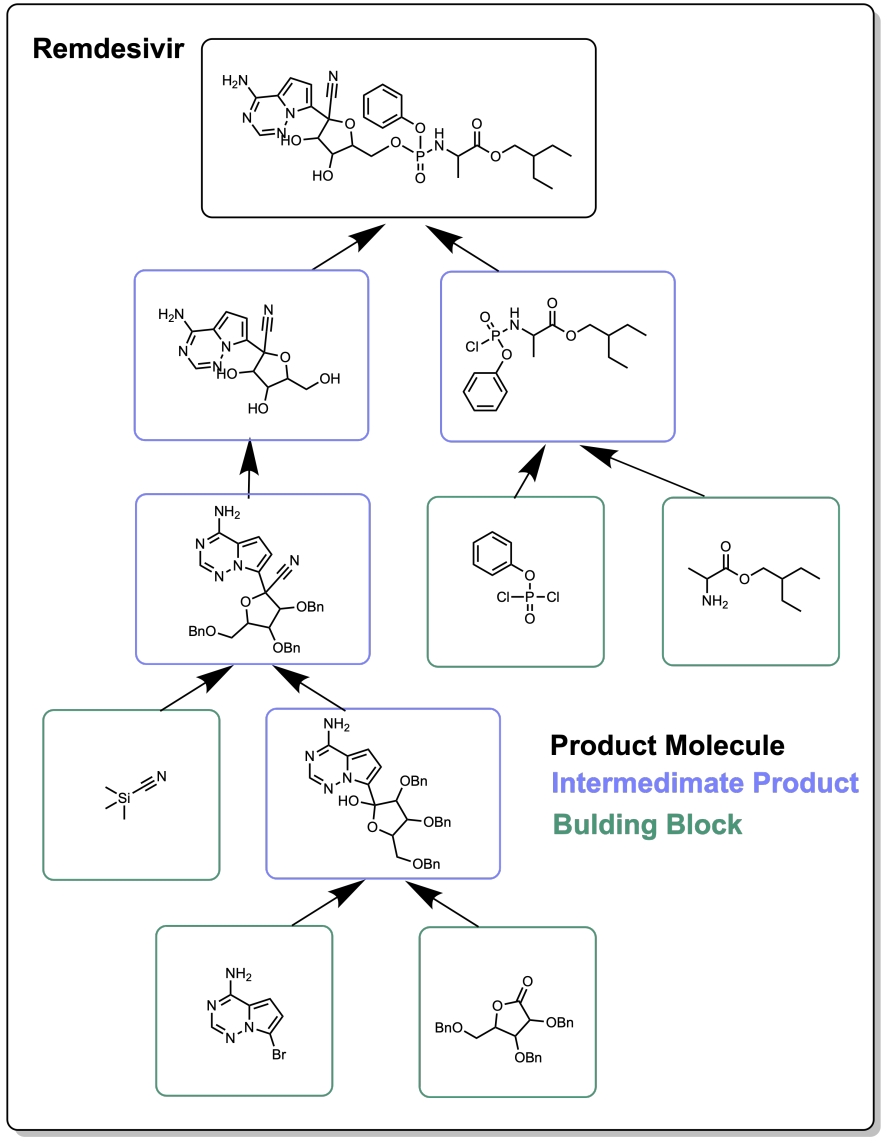
ASKCOS to benchmark synthesizability

- We evaluated the methods in Guacamol and found most molecular optimization methods are worse than screening.



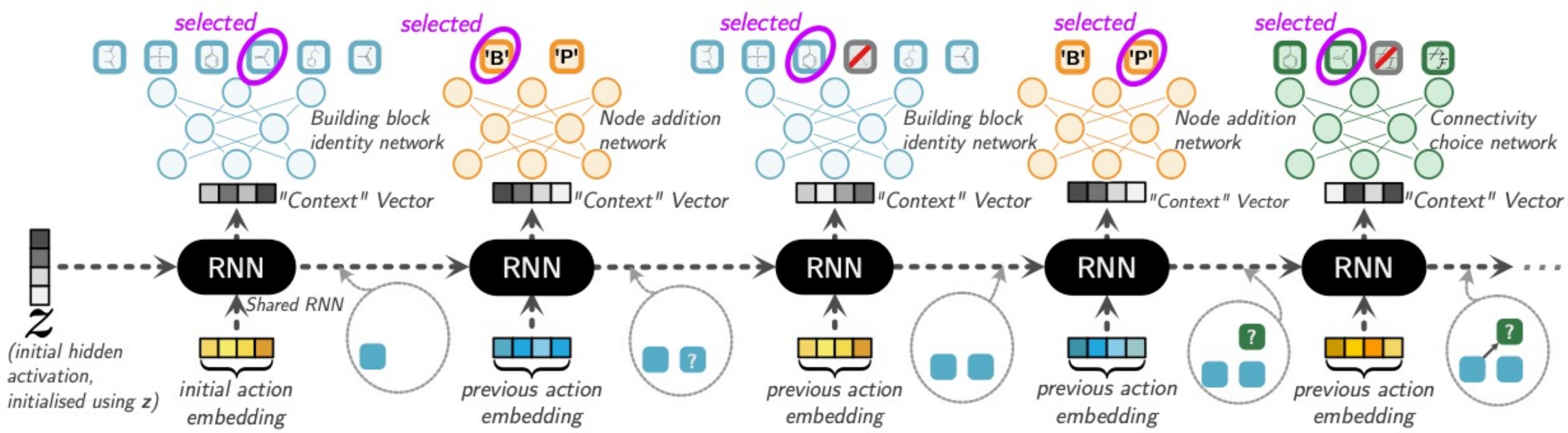
Synthetic tree generation: coupling design and CASP

- Synthetic pathway can be abstracted into a tree structure
- Both synthesizable molecular design and CASP involve looking for a synthetic tree:
  - **Synthesis Planning** is to generate synthetic trees whose product molecules matches the target molecule.
  - **Synthesizable Molecular Design** is to optimize the properties of interest of the product molecule w.r.t. the structure of a synthetic tree.



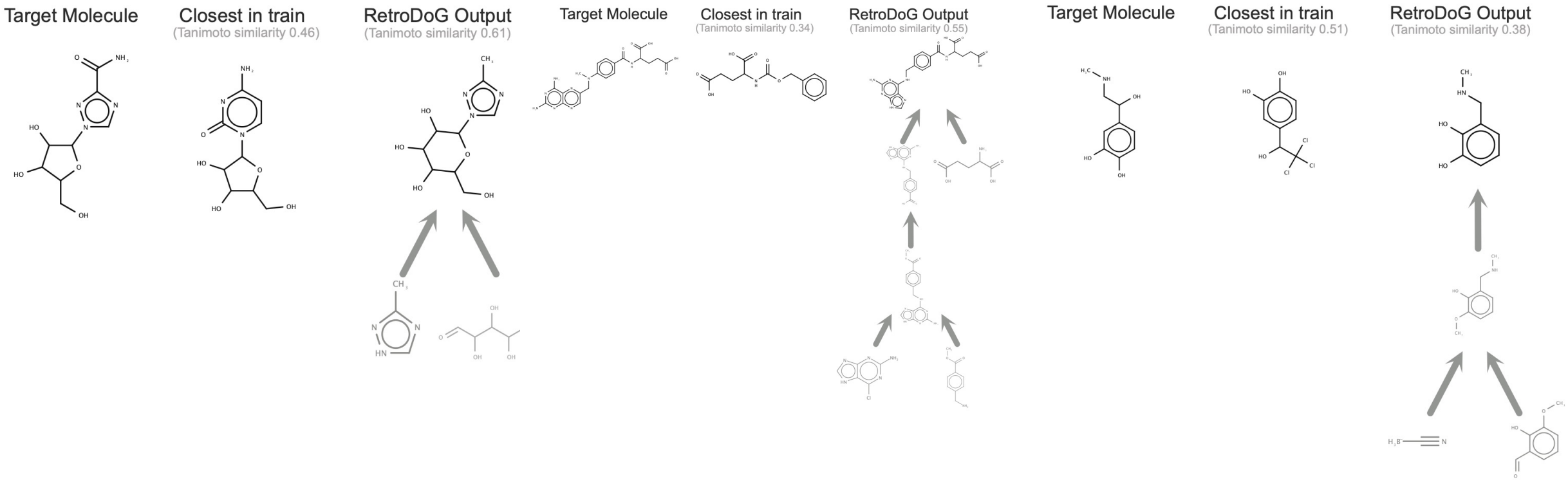
From DoG-AE/Gen and Retro-DoG

- The DoG models relies on a forward reaction predictor that:
  - All forward reaction predictors suffer from the bias of positive reaction data.
  - Don't explicitly use the information of intermediate molecular structures, thus the model need to learn to approximate the reaction prediction model.



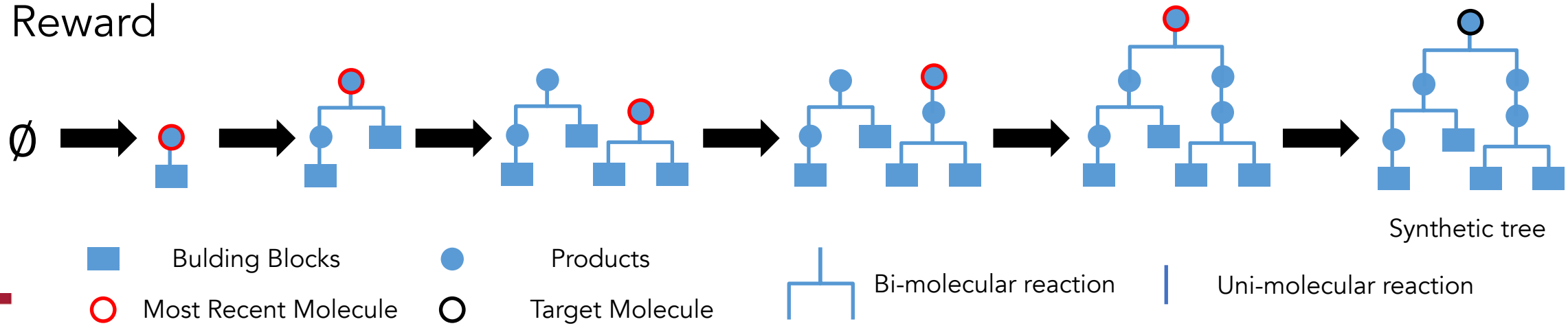
From DoG-AE/Gen and Retro-DoG

- Retro-DoG cannot recover target molecules, e.g. cannot perform synthesis planning.



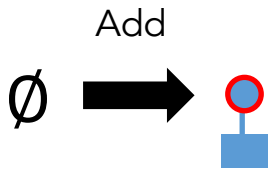
Synthetic tree generation as a Markov decision process

- State Space
  - States are defined as root molecule(s) of an intermediate synthetic tree.
  - We enforce a depth-first order thus at most two sub-trees can occur, which leads to (1) at most two root molecules and (2) expansions always take place from the most recent one.
- Action Space
  - One reaction step is an action step.
  - Define 4 types of actions: Add, Expand, Merge, End
- State Transition Dynamics
- Reward



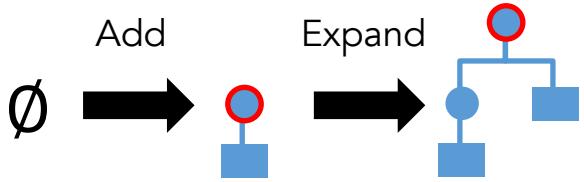
Synthetic tree generation as a Markov decision process

- State Space
  - States are defined as root molecule(s) of an intermediate synthetic tree.
  - We enforce a depth-first order thus at most two sub-trees can occur, which leads to (1) at most two root molecules and (2) expansions always take place from the most recent one.
- Action Space
  - One reaction step is an action step.
  - Define 4 types of actions: Add, Expand, Merge, End
- State Transition Dynamics
- Reward



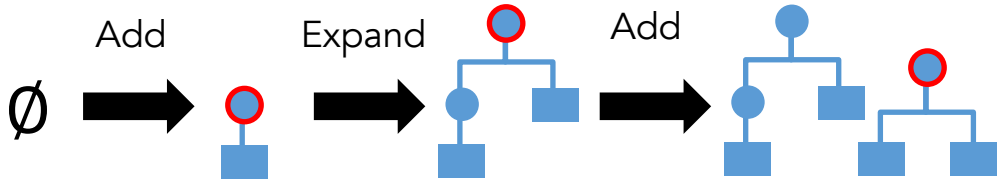
Synthetic tree generation as a Markov decision process

- State Space
  - States are defined as root molecule(s) of an intermediate synthetic tree.
  - We enforce a depth-first order thus at most two sub-trees can occur, which leads to (1) at most two root molecules and (2) expansions always take place from the most recent one.
- Action Space
  - One reaction step is an action step.
  - Define 4 types of actions: Add, Expand, Merge, End
- State Transition Dynamics
- Reward



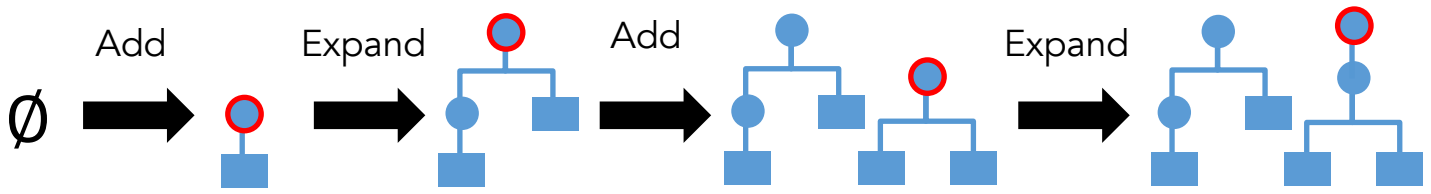
Synthetic tree generation as a Markov decision process

- State Space
  - States are defined as root molecule(s) of an intermediate synthetic tree.
  - We enforce a depth-first order thus at most two sub-trees can occur, which leads to (1) at most two root molecules and (2) expansions always take place from the most recent one.
- Action Space
  - One reaction step is an action step.
  - Define 4 types of actions: Add, Expand, Merge, End
- State Transition Dynamics
- Reward



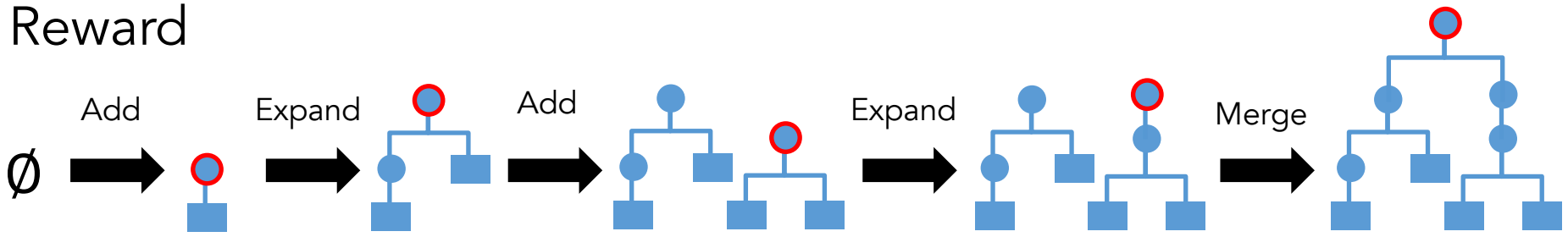
Synthetic tree generation as a Markov decision process

- State Space
  - States are defined as root molecule(s) of an intermediate synthetic tree.
  - We enforce a depth-first order thus at most two sub-trees can occur, which leads to (1) at most two root molecules and (2) expansions always take place from the most recent one.
- Action Space
  - One reaction step is an action step.
  - Define 4 types of actions: Add, Expand, Merge, End
- State Transition Dynamics
- Reward



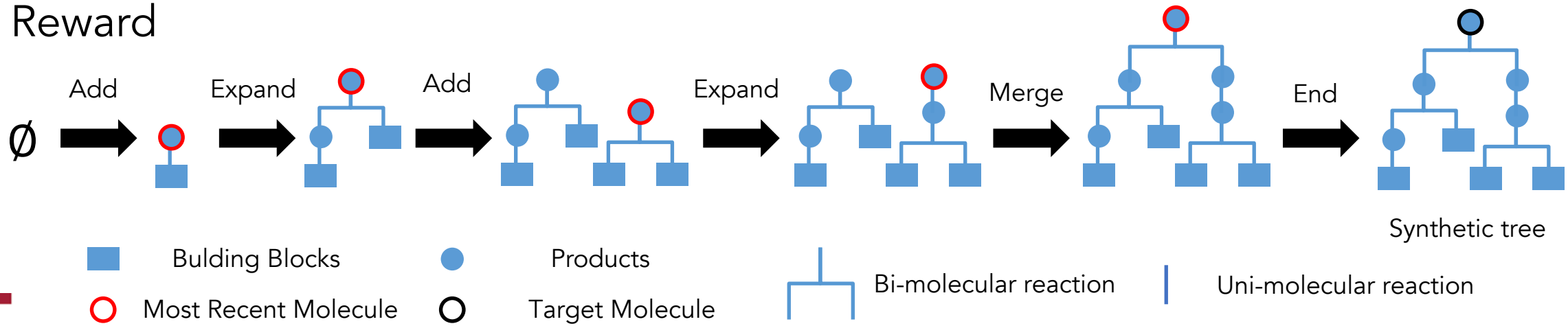
Synthetic tree generation as a Markov decision process

- State Space
  - States are defined as root molecule(s) of an intermediate synthetic tree.
  - We enforce a depth-first order thus at most two sub-trees can occur, which leads to (1) at most two root molecules and (2) expansions always take place from the most recent one.
- Action Space
  - One reaction step is an action step.
  - Define 4 types of actions: Add, Expand, Merge, End
- State Transition Dynamics
- Reward



Synthetic tree generation as a Markov decision process

- State Space
  - States are defined as root molecule(s) of an intermediate synthetic tree.
  - We enforce a depth-first order thus at most two sub-trees can occur, which leads to (1) at most two root molecules and (2) expansions always take place from the most recent one.
- Action Space
  - One reaction step is an action step.
  - Define 4 types of actions: Add, Expand, Merge, End
- State Transition Dynamics
- Reward



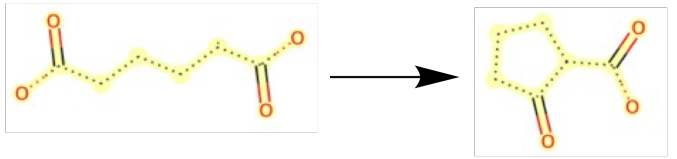
Synthetic tree generation as a Markov decision process

- State Space
- Action Space
- State Transition Dynamics
  - To ensure each reaction step is chemically plausible:
    - Machine learning reaction prediction model.
    - Domain-specific reaction rules encoded as reaction templates.
  - We reject all reactions don't follow a known template.
- Reward
  - Matching between product molecule and target molecule
  - The properties of interest of the product molecule.

Uni-molecular reaction template

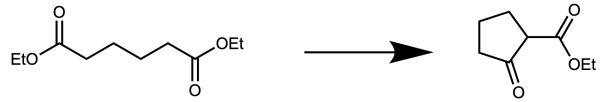
Reaction SMARTS

[O$(O(C)([CX4])):8][C:7](=[O:9])[CH:6][C:5][C:4][C:3][C:2]([O$(O(C)([CX4])):10]) =[O:1]>>[O:8][C:7](=[O:9])[C:6]1[C:5][C:4][C:3][C:2]1=[O:1]



Reaction SMILES (one macthed example)

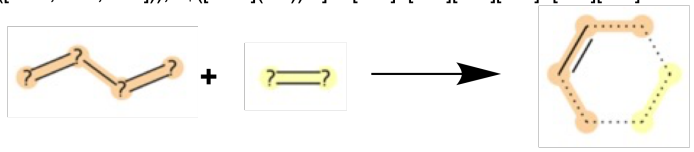
O=C(OCC)CCCCC(OCC)=O>>O=C1CCCC1C(OCC)=O



Bi-molecular reaction template

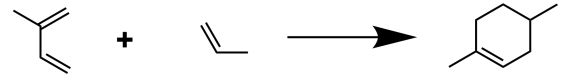
Reaction SMARTS

[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C):1]=[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([C H2](=C)):2].[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])) ,C$([CH2](=C)):3]=[C$([C](=C)(C)([CX4,OX2,NX3])),C$([CH](=C)(C)):4]-[C$([C](=C)(C)([CX4,OX2,NX3])),C$([CH](=C)(C)):5]=[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C $([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):6]>>[C:1]1[C:2][C:3][C:4]=[C:5][C:6]1



Reaction SMILES (one macthed example)

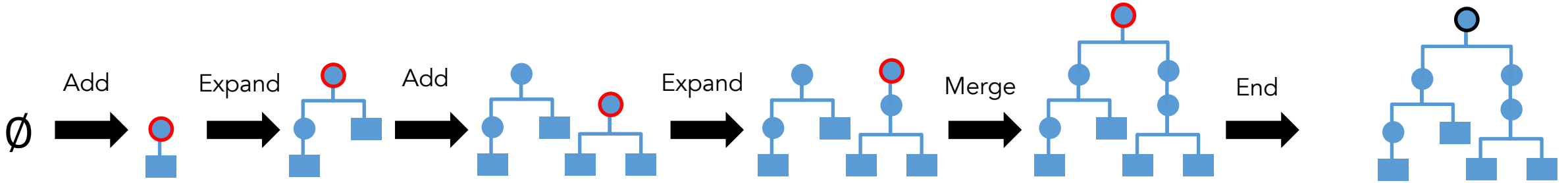
C=C(C)C=C.C=CC>>CC1=CCCCC1



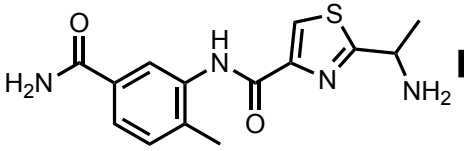
Conditional generation for synthesis planning

- Synthesis planning is a probabilistic modeling of synthetic trees conditioned on a target molecule.

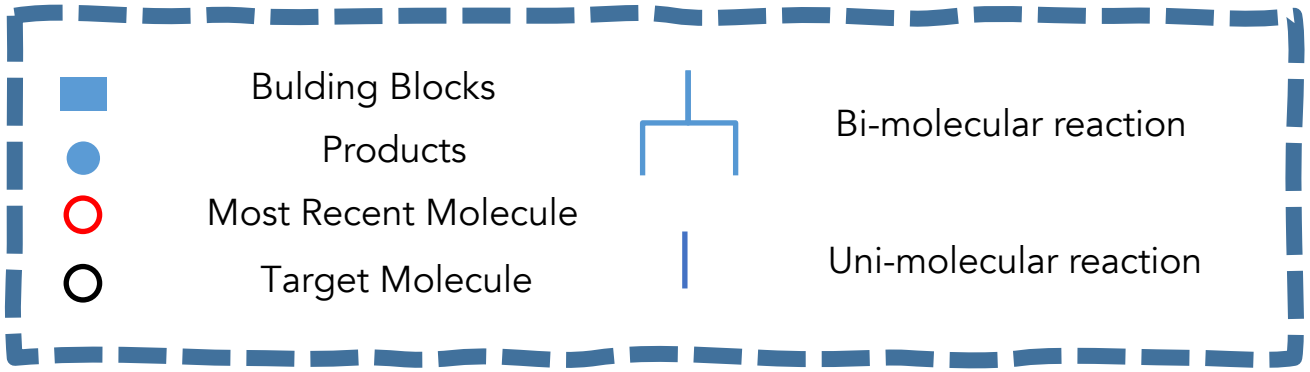
$$a^{(t)} = \left(a_{act}^{(t)}, a_{rt1}^{(t)}, a_{rxn}^{(t)}, a_{rt2}^{(t)}\right) \sim p(a^{(t)} | S^{(t)}, M_{target})$$



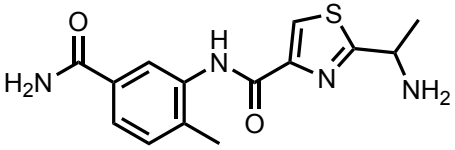
Conditioned on



Target molecule



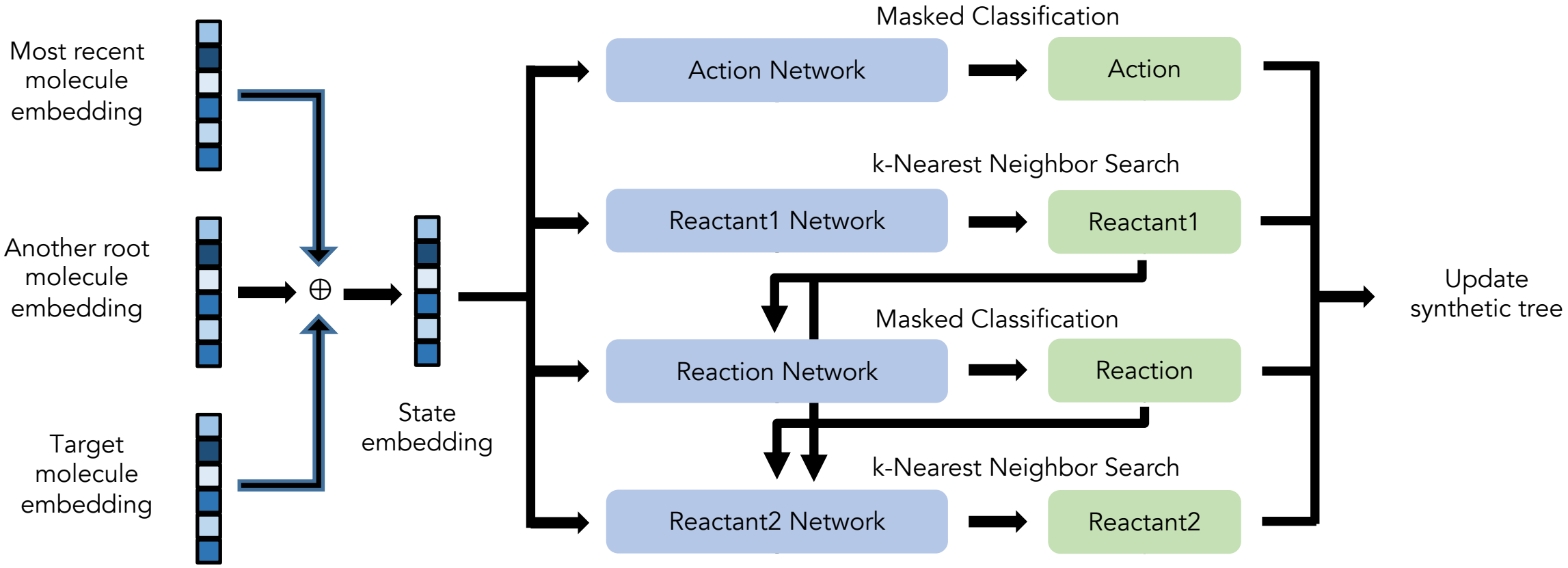
Synthetic tree
+



Product molecule

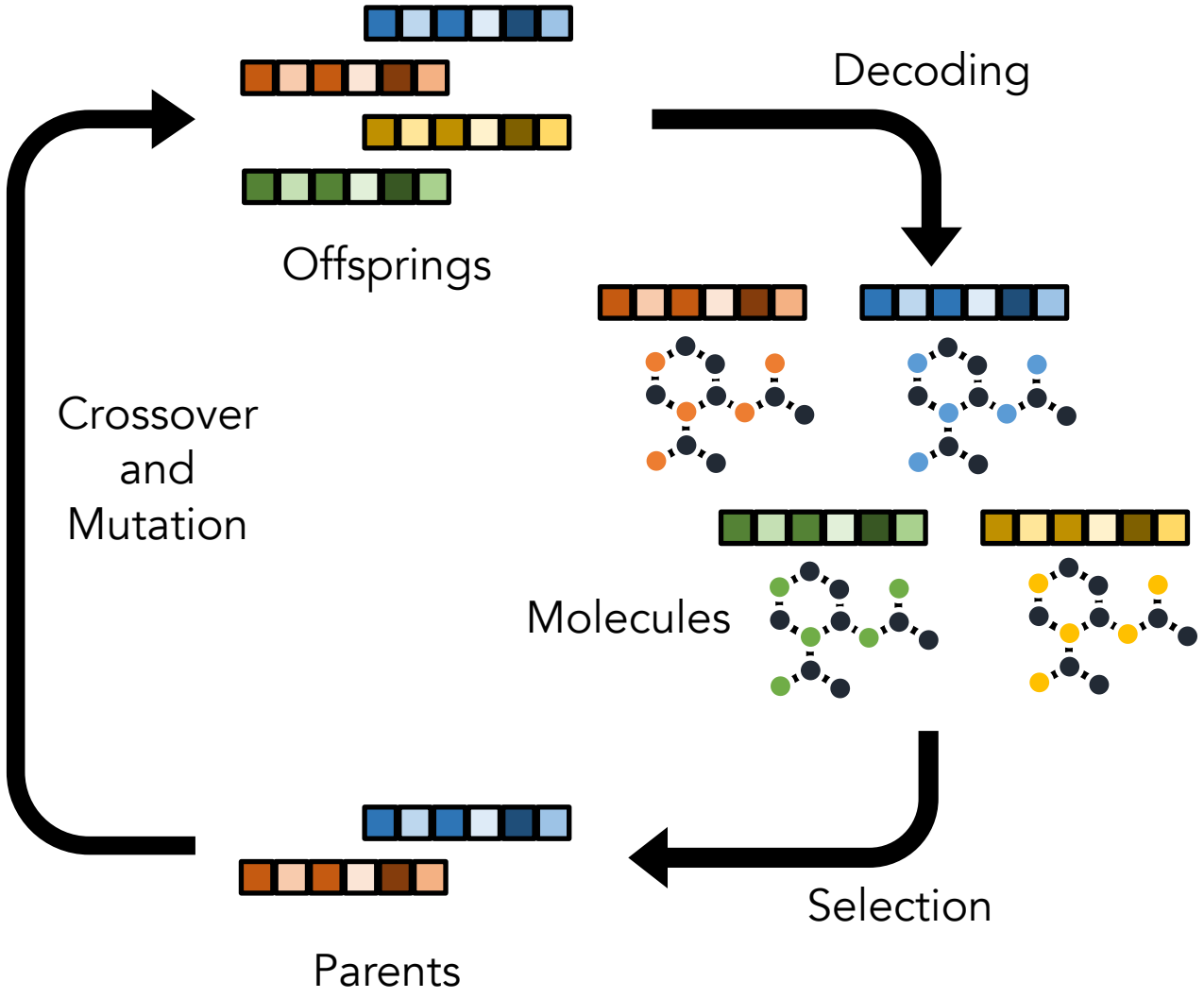
Model architecture

- Morgan fingerprints with radius 2 and 4096 bits are used to represent molecules.
- Action and Reaction networks are classifiers, reactants networks are regressors.
- We mask out illegal actions and conduct k-NN search to select reactants.



Genetic algorithm for synthesizable molecular design

- GA on fingerprints:
  - From a random sampled 128 from ZINC, offspring size is 512, tried up to 200 generations
  - Crossover: inherit about half from one and remaining from another, higher probability to sample high scored element.
  - Mutation: with a probability (0.5) flip a number of bits (24)



Data preparation and training

- **Reaction templates:** Combined 91 reaction templates (42 from Button's, 49 from Hartenfeller's), 13 uni-mol, 78 bi-mol.
- **Purchasable compounds:** Enamine building blocks, US stock (147,505).
- **Data**: Synthetic trees are generated by randomly applying applicable templates to randomly selected purchasable compounds, filtered by QED (drug-likeliness) of root molecules: 208,644 synthetic paths for training, 69,548 for validation and testing each.

- Each network is trained as a separate supervised learning problem using a subset of information from the known synthetic routes.

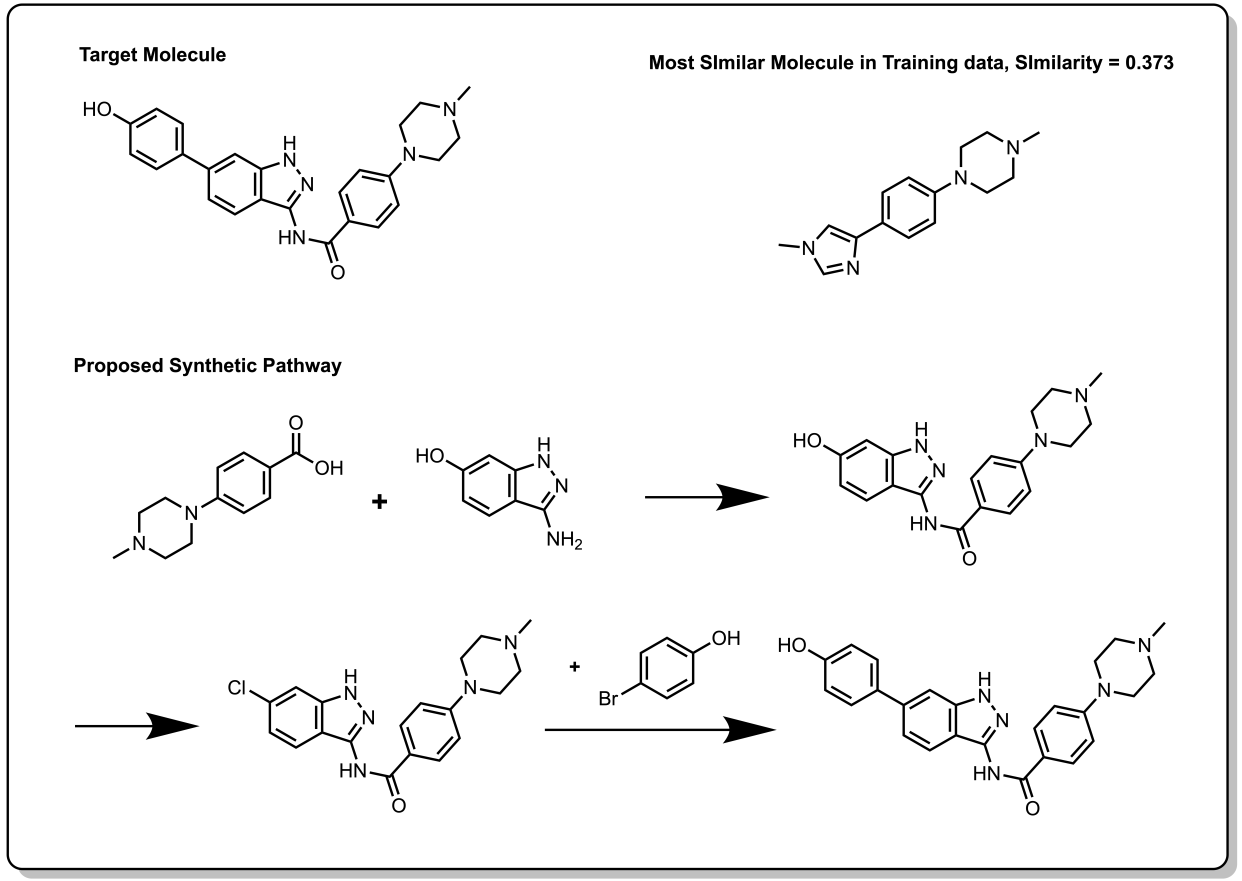
Hartenfeller, M.,... & Renner, S. (2011). *Journal of chemical information and modeling*, *51*(12), 3093-3098.
Button, A., Merk, D., Hiss, J. A., & Schneider, G. (2019). *Nature machine intelligence*, *1*(7), 307-315.

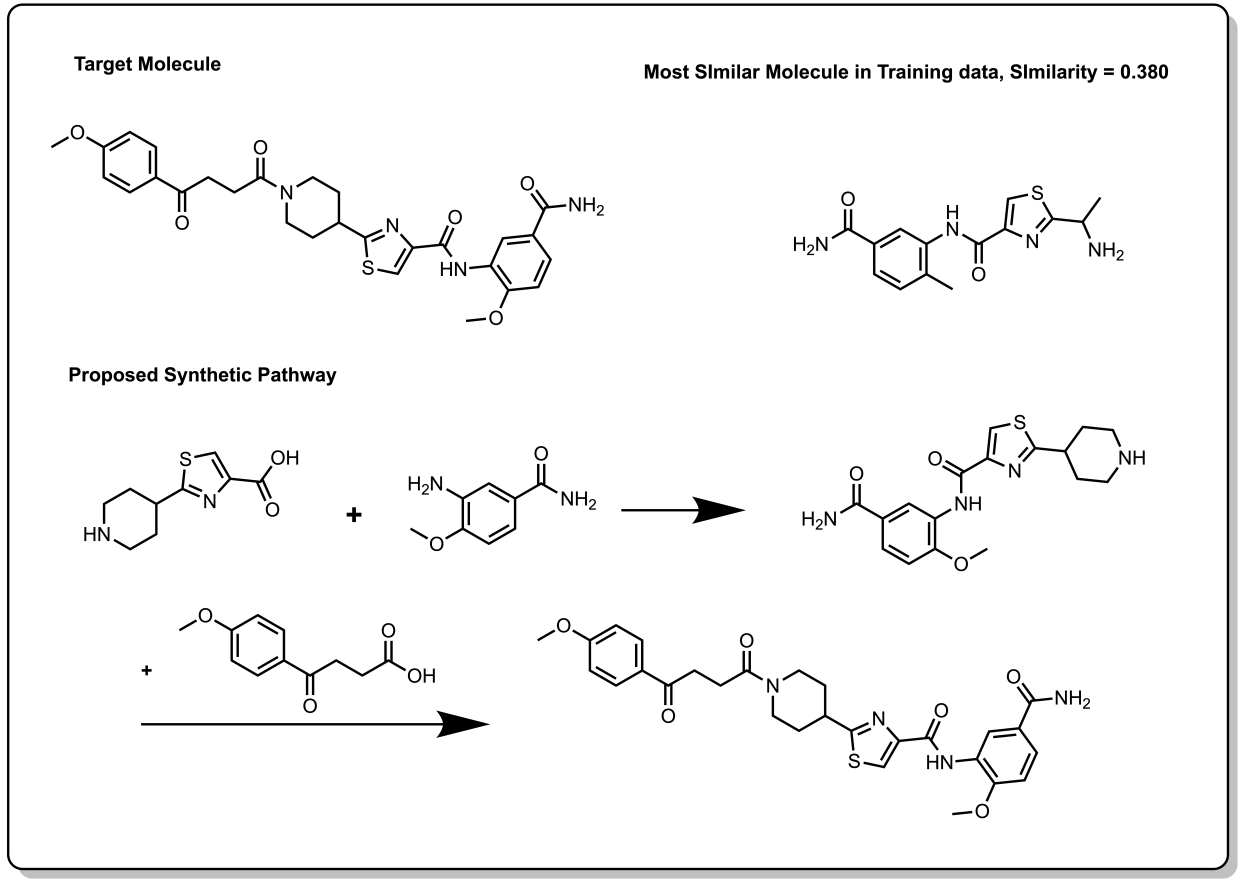
Synthesis Planning

- We construct synthetic trees for testing data as "reachable" data and a random sample from ChEMBL as "unreachable" data.
- We use k=3 in the nearest neighbor search of first reactant, and k=1 for the remaining. (~1s/mol, ~1min/mol for MCTS)
- In the unrecovered cases, the output molecules could also serve as a synthesizable structural analog.

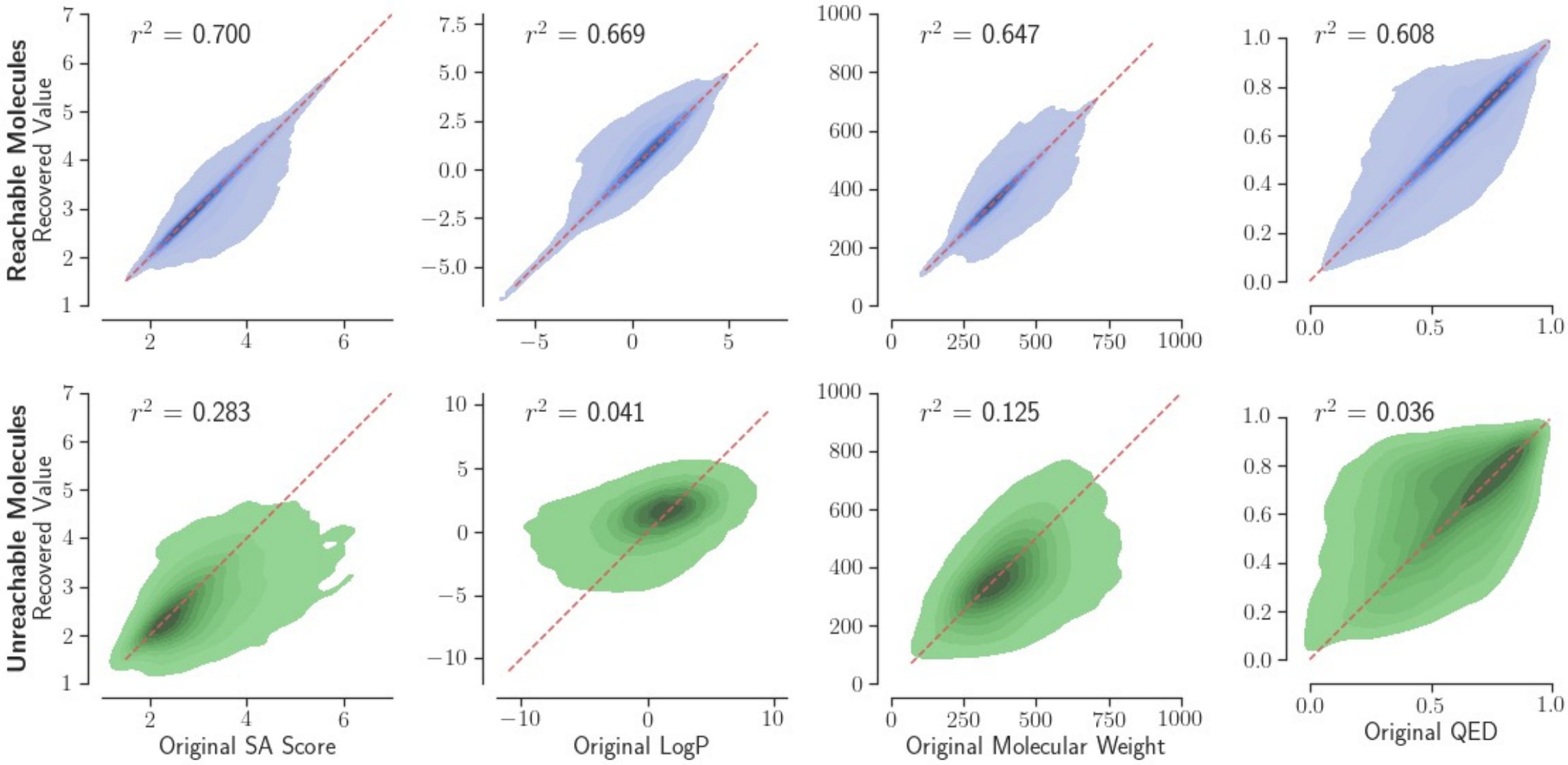
	N	Recovery rate	Average Similarity	KL Divergence	FC DIstance
Reachable (test set)	69,548	51.0%	0.508	0.995	0.067
Unreachable (ChEMBL)	20,000	4.5%	0.396	0.966	1.994

Synthesis Planning



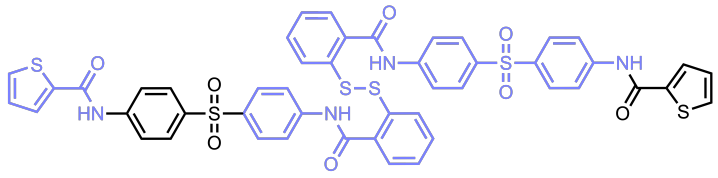


Synthesizable Analog Recommendation

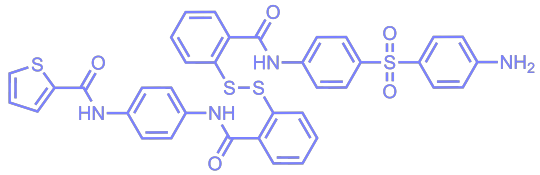


Synthesizable Analog Recommendation

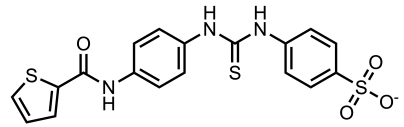
Target Molecule



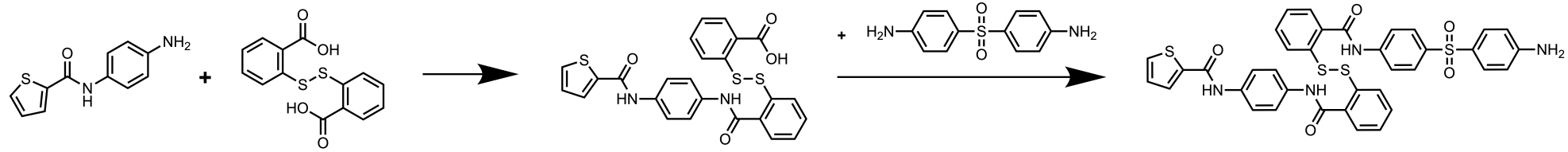
Product Molecule, similarity = 0.894



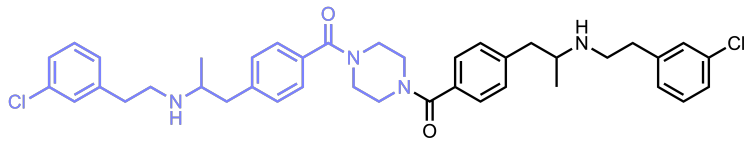
Most SImilar Molecule in Training data, SImilarity = 0.549



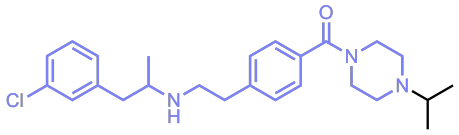
Proposed Synthetic Pathway



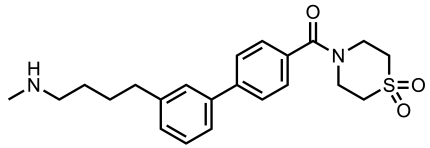
Target Molecule



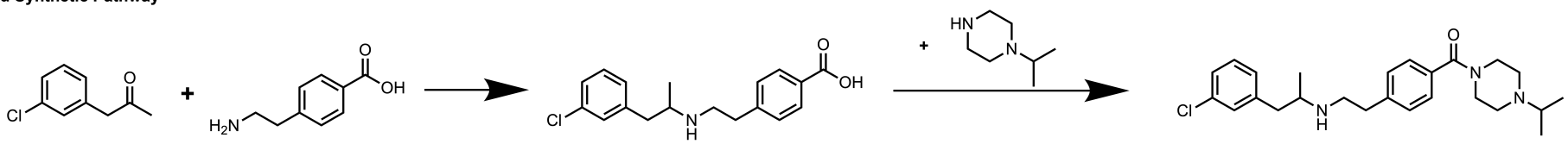
Product Molecule, similarity = 0.830



Most SImilar Molecule in Training data, SImilarity = 0.453



Proposed Synthetic Pathway



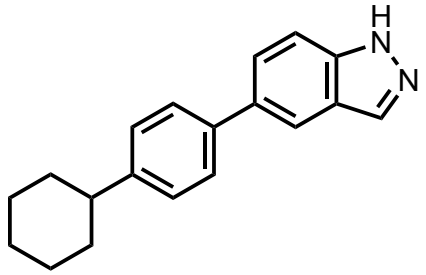
Synthesizable molecular optimization

- To validate our model, we first consider common heuristic oracle functions relevant to drug discovery
- Our model consistently outperforms GCPN and MolDQN, and is comparable to GA+D and MARS across different tasks.

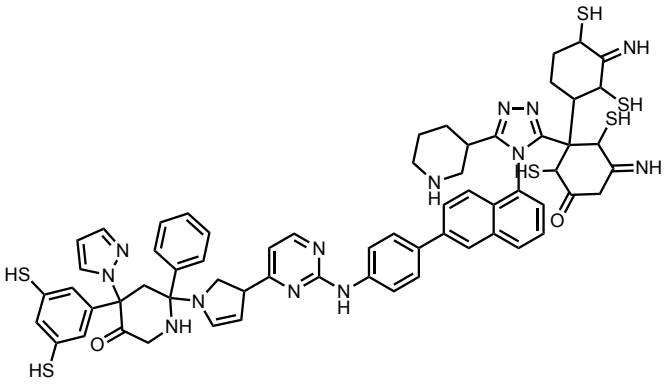
	JNK3			GSK3β			QED		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
GCPN	0.57	0.56	0.56	0.57	0.56	0.56	**0.948**	0.947	0.946
MolDQN	0.64	0.63	0.63	0.54	0.53	0.53	**0.948**	**0.948**	**0.948**
GA+D	0.81	0.80	0.80	0.79	0.79	0.78	-	-	-
MARS	0.92	0.91	0.90	**0.95**	0.93	0.92	**0.948**	**0.948**	**0.948**
DST	**0.97**	**0.97**	**0.97**	**0.95**	**0.95**	**0.95**	0.947	0.946	0.946
Our method	0.80	0.78	0.77	0.94	0.93	0.92	**0.948**	**0.948**	**0.948**

GCPN: You, J., Liu, B., Ying, R., Pande, V., & Leskovec, J. (2018). *arXiv preprint arXiv:1806.02473*.
MolDQN: Zhou, Z., Kearnes, S., Li, L., Zare, R. N., & Riley, P. (2019). *Scientific reports*, *9*(1), 1-10.
GA+D: Nigam, A., Friederich, P., Krenn, M., & Aspuru-Guzik, A. (2019). *arXiv preprint arXiv:1909.11655*.
MARS: Xie, Y., Shi, C., Zhou, H., Yang, Y., Zhang, W., Yu, Y., & Li, L. (2021). *arXiv preprint arXiv:2103.10432*.
DST: Fu, T., Gao, W., Xiao, C., Yasonik, J., Coley, C. W., & Sun, J. (2021). *arXiv preprint arXiv:2109.10469*.

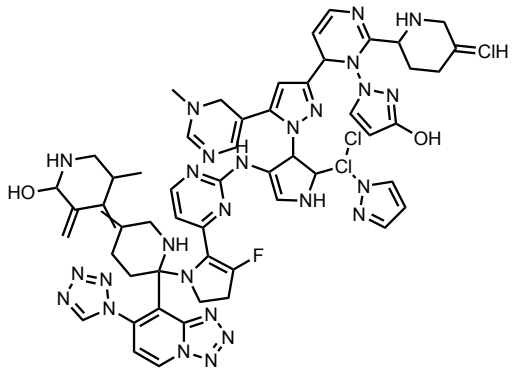
Synthesizable Molecular Optimization



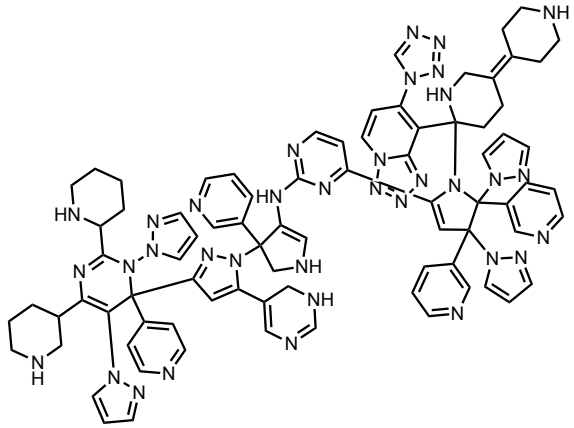
GSK3β = 0.94
Top-1 from our model



GSK3β = 0.97
Top-1 from DST

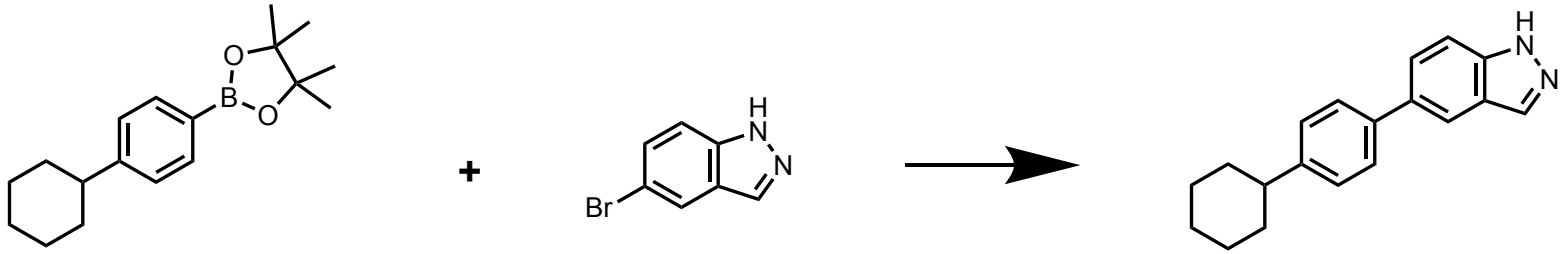


GSK3β = 0.95
Top-1 from MARS

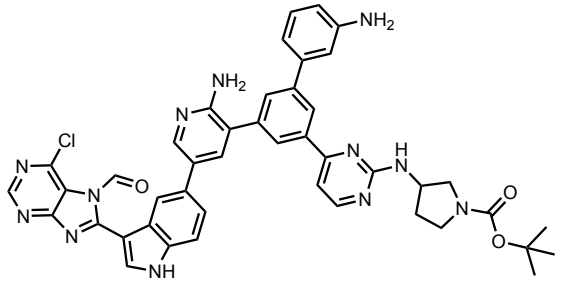


GSK3β = 0.79
Top-1 from GA+D

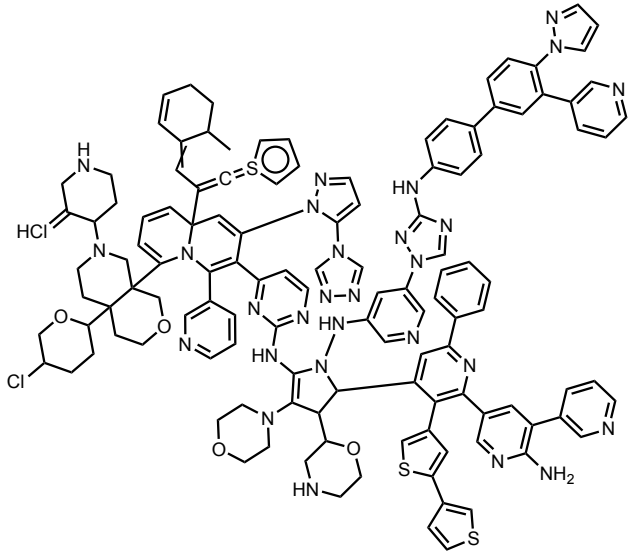
Synthetic pathway:



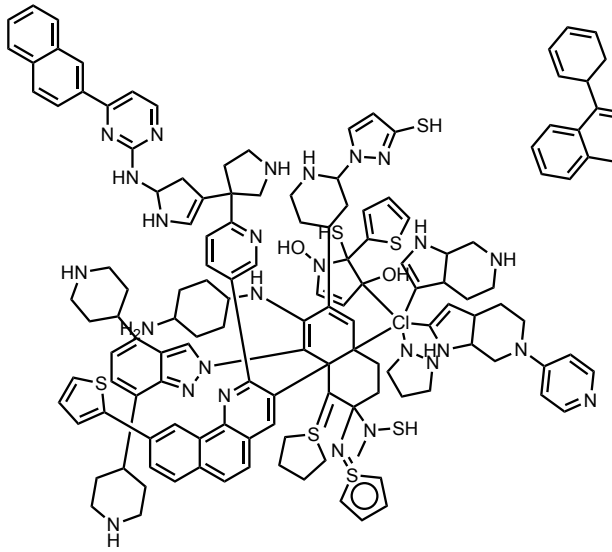
Synthesizable Molecular Optimization



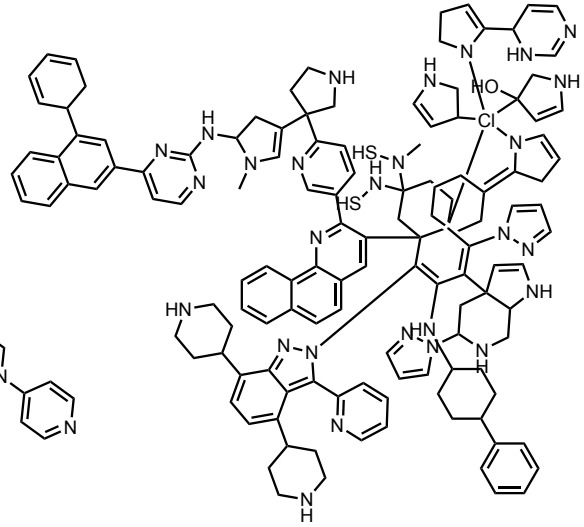
JNK3 = 0.80
Top-1 from our model



GSK3B = 0.97
Top-1 from DST

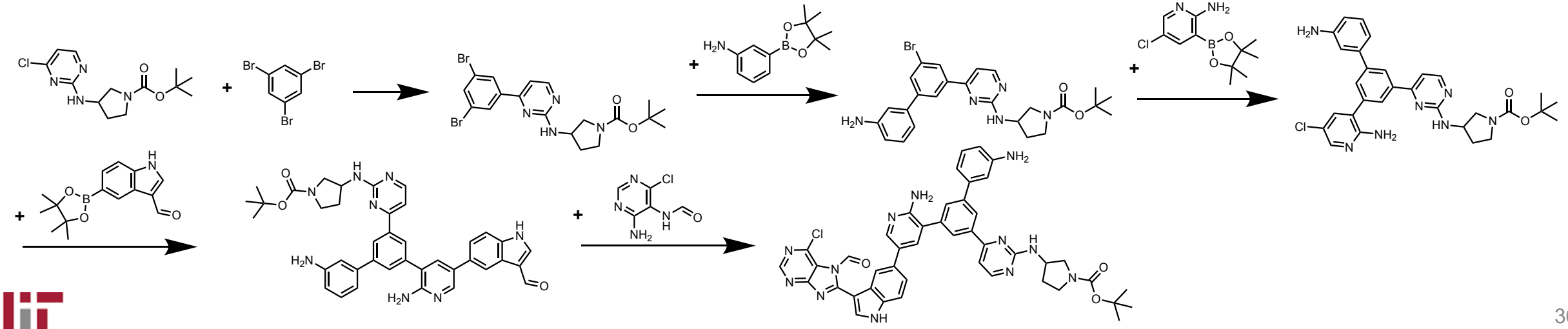


GSK3B = 0.92
Top-1 from MARS



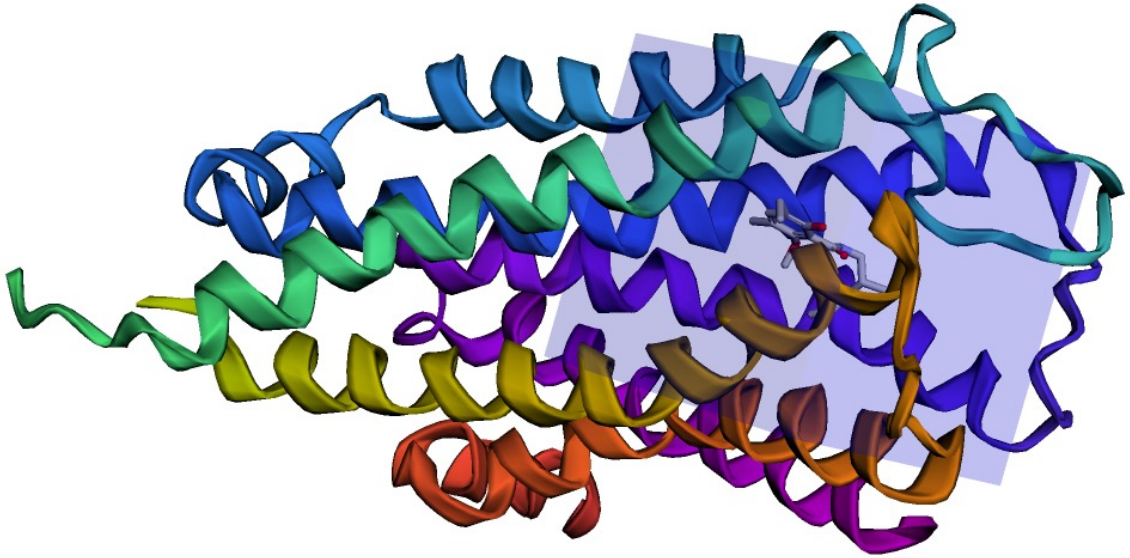
GSK3B = 0.81
Top-1 from GA+D

Synthetic pathway:

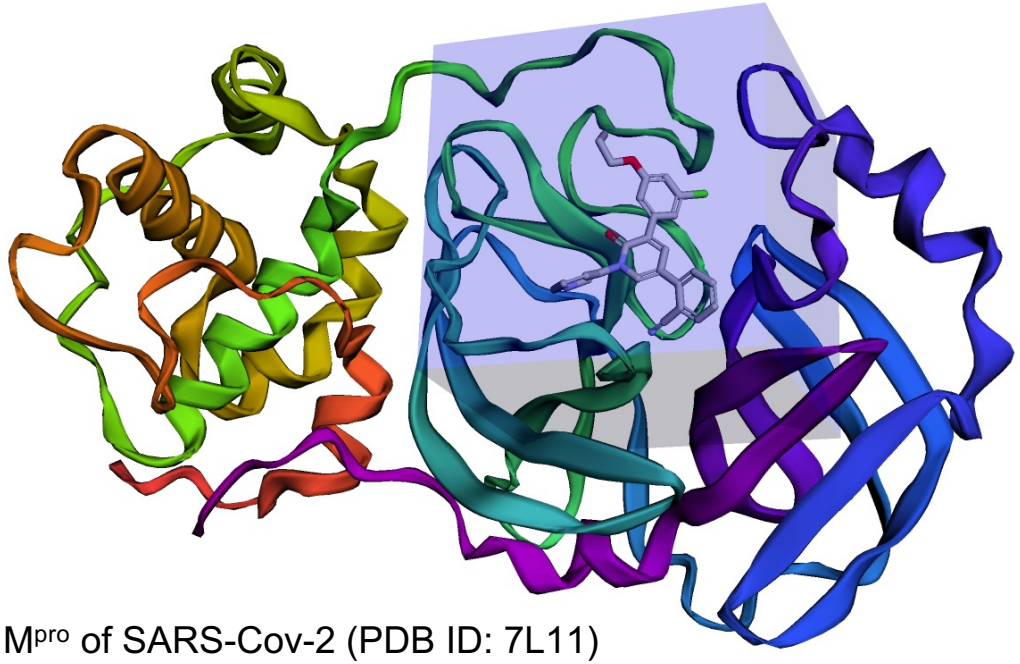


Optimizing docking score w/ TDC generative benchmark

- To simulate a more realistic case, we optimized docking score against two important disease targets
- We limit the number of oracle calls less than 5000.



dopamine D_3 receptor (PDB ID: 3PBL)



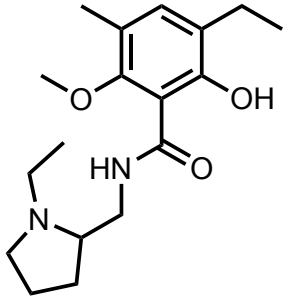
M^{pro} of SARS-Cov-2 (PDB ID: 7L11)

Optimizing docking score against dopamine D_3 receptor

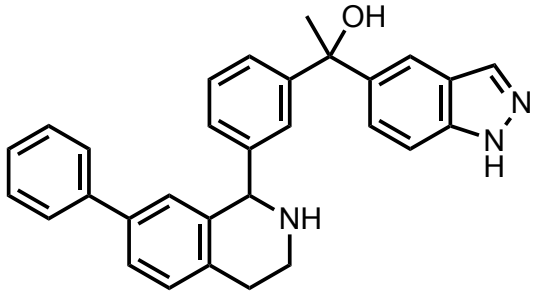
- **Good structure quality: Our** model achieved high passing rate of quality filter and low SA_Score.

Method Category			Domain-Specific Methods		State-of-the-Art Methods in ML				Ours
Metric	Best-in-data	# Calls	Screening	Graph-GA	LSTM	GCPN	MolDQN	MARS	SynNet
Top100 (↓)	-12.080	5000	-10.542±0.035	**-14.811±0.413**	-13.017±0.385	-10.045±0.226	-8.236±0.089	-9.509±0.035	-11.133
Top10 (↓)	-12.590		-11.483±0.056	**-15.930±0.336**	-14.030±0.421	-11.483±0.581	-9.348±0.188	-10.693±0.172	-12.020
Top1 (↓)	-12.800		-12.100±0.356	**-16.533±0.309**	-14.533±0.525	-12.300±0.993	-9.990±0.194	-11.433±0.450	-12.300
Diversity (↑)	0.864		0.872±0.003	0.626±0.092	0.740±0.056	**0.922±0.002**	0.893±0.005	0.873±0.002	0.821
Novelty (↑)	-		-	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000
%Pass (↑)	0.780		0.683±0.073	0.393±0.308	0.257±0.103	0.167±0.045	0.023±0.012	0.527±0.087	**0.800**
Top1 Pass (↓)	-11.700		-10.100±0.000	**-14.267±0.450**	-12.533±0.403	-9.367±0.170	-7.980±0.112	-9.000±0.082	-12.300
SA_Score (↓)	2.973		3.036±0.014	4.783±1.195	**2.611±0.238**	6.843±0.210	6.687±0.049	3.103±0.011	2.801

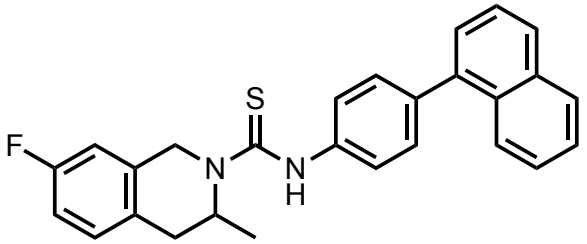
Optimizing docking score against dopamine D_3 receptor



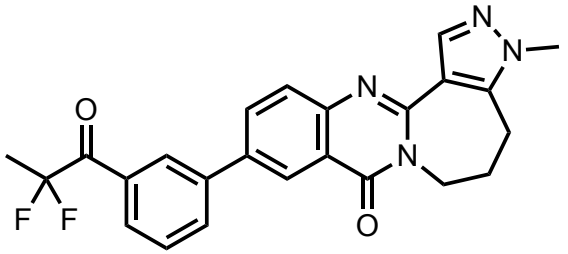
Vina score = -8.62 kJ/mol
Known Inhibitor



Vina score = -12.3 kJ/mol
Top-1 from our model

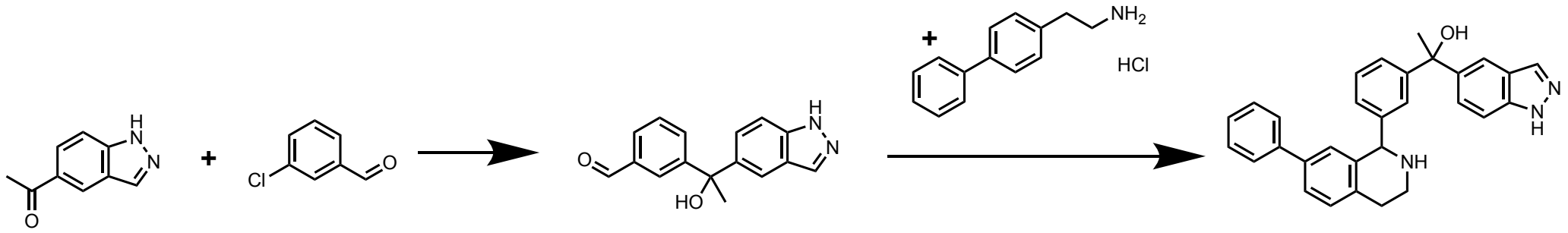


Vina score = -11.9 kJ/mol
2nd from our model

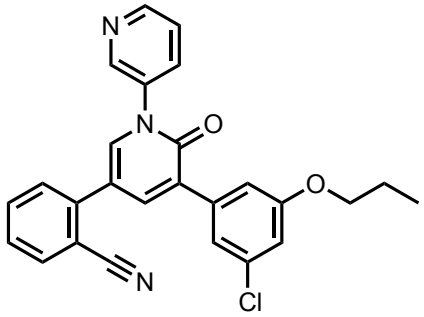


Vina score = -11.8 kJ/mol
3rd from our model

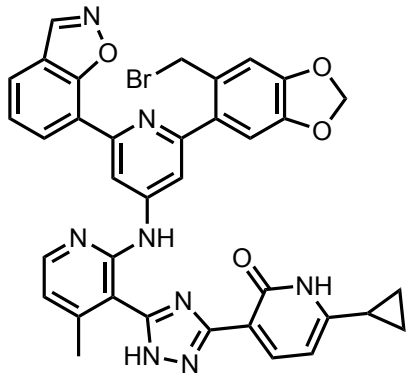
Synthetic pathway:



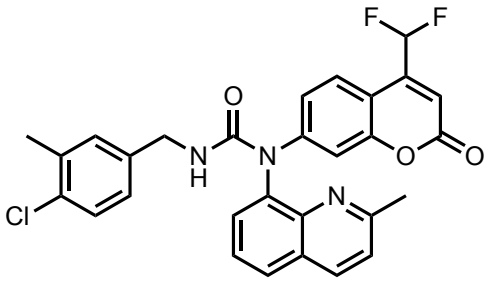
Optimizing docking score against M^{pro} of SARS-Cov-2



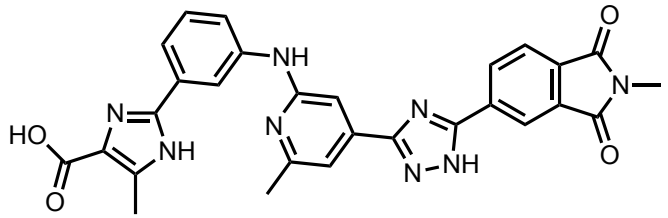
Vina score = -8.96 kJ/mol
Known Inhibitor



Vina score = -10.50 kJ/mol
Top-1 from our model

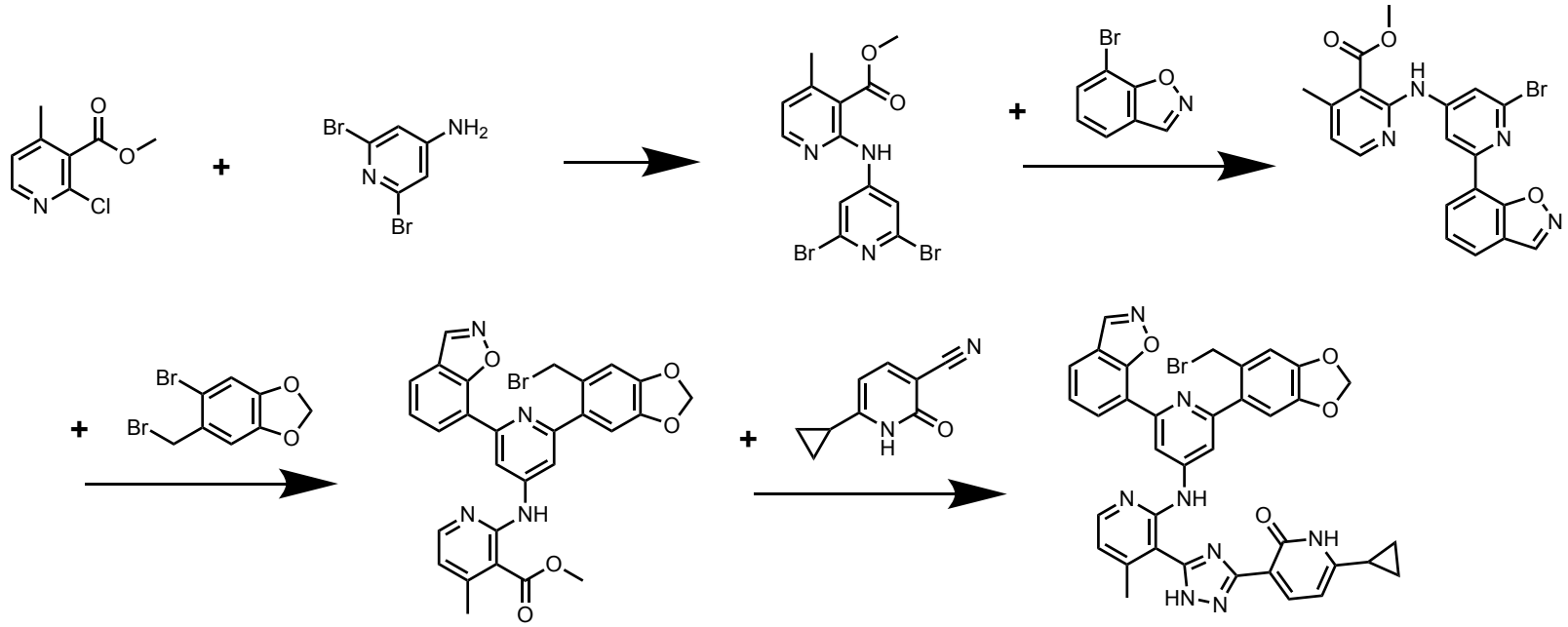


Vina score = -9.31 kJ/mol
2nd from our model



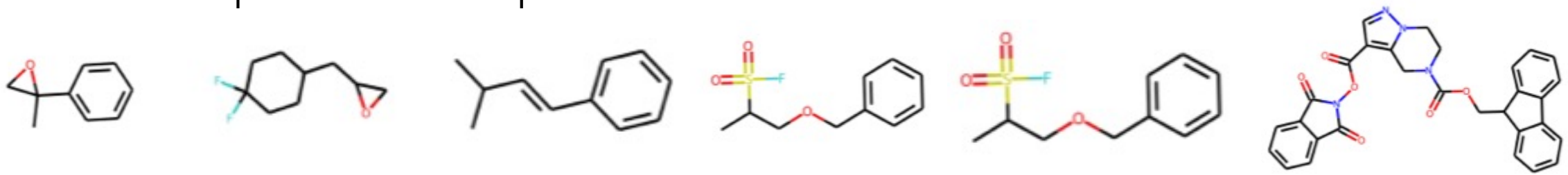
Vina score = -9.25 kJ/mol
3rd from our model

Synthetic pathway:



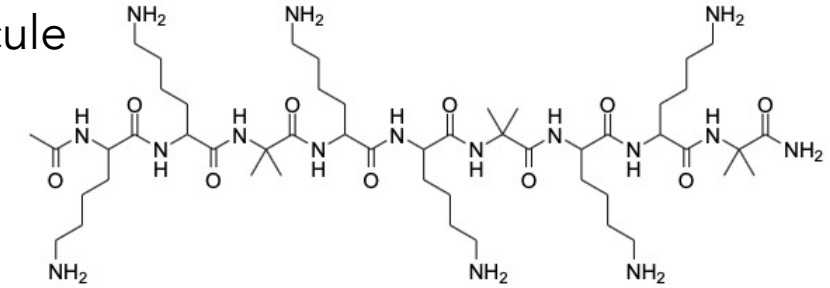
Limitation

- Reaction templates are not perfect.

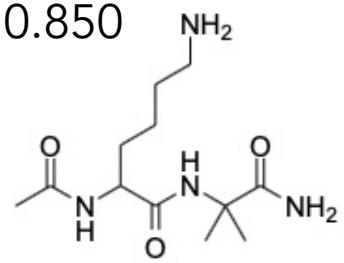


- A depth-first order leads to a canonical order of reactants, which is unphysical (DAG-type MDP and tree-type MDP).
- A binary presence-based fingerprint cannot distinguish repeating units.

Target Molecule



Output Molecule, similarity=0.850



- The first reactant selection is the bottleneck (~30%).

Conclusion

- We formulate the tasks of multi-step synthesis planning and synthesizable molecular design as **a single shared task of conditional synthetic tree generation**.
- We **formulate a Markov decision process** to model the generation of synthetic trees, allowing the generation of multi-step and convergent synthetic pathways.
- We propose a model that is capable of (1) rapid bottom-up **synthesis planning** and (2) constrained **molecular optimization** that can explore a chemical space defined by reaction templates and purchasable starting materials.
- We demonstrate encouraging results on the recovery of molecules via conditional generation and on de novo molecular optimization with multiple objective functions relevant to bioactive molecule design and drug discovery.

Take home message

whgao@mit.edu | coley.mit.edu

Synthetic tree design could solve molecule design and synthesis planning simultaneously.

Acknowledgement

- This work was supported by the ONR and MLPDS.
- Coley Research Group
  - Connor W. Coley
  - Rocío Mercado
  - Aparajita Dasgupta
  - Thijs Stuyver
  - Matteo Aldeghi
  - John Bradshaw
  - David Graff
  - Itai Levin
  - Samuel Goldman
  - Priyanka Raghavan
  - Zhengkai Tu
  - Keir Adams
  - Herry (Tianyi) Jin
  - Ria Das
  - Katherine Lim
  - Jacob Yasonik
  - Saul A. Vega Sauceda