Highlighting work from Wenhao Gao, Rocío Mercado (now Chalmers),
Sam Goldman (now MPM BioImpact), Matteo Aldeghi (now Bayer), Shitong Luo, and Jenna Fromer

Challenges in molecular optimization

Connor W. Coley

Assistant Professor

MIT Chemical Engineering

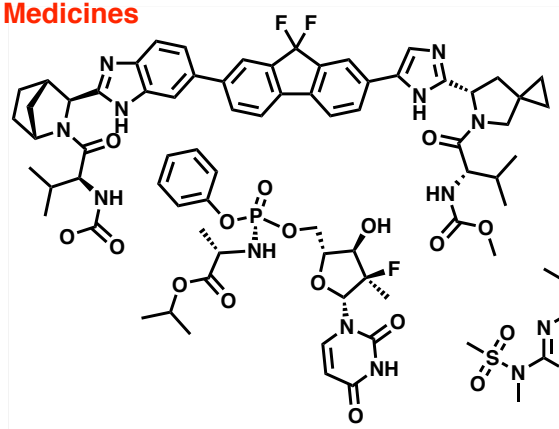
MIT Electrical Engineering and Computer Science

AI≡Science

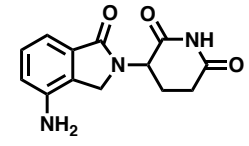
June 13, 2024

The kinds of molecules we are trying to find/optimize

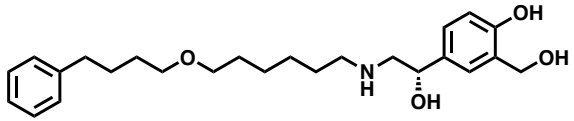
Medicines



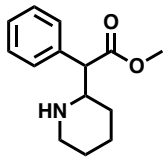
Harvoni
Ledipasvir/sofosbuvir (Hep C)



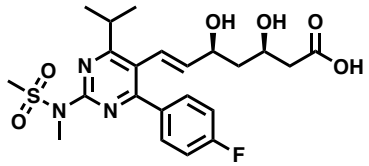
Revlimid
Lenalidomide (multiple myeloma)



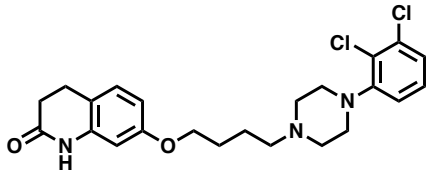
Serevent
Salmeterol (COPD)



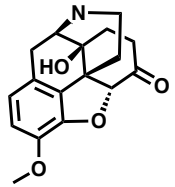
Ritaline
Methylphenidate (ADHD)



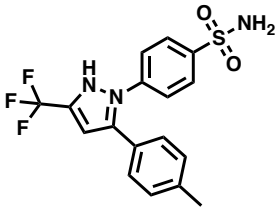
Cretor
Rosuvastatin (cholesterol)



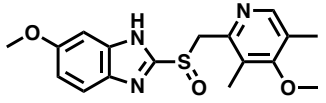
Abilify
Aripiprazole (antipsychotic)



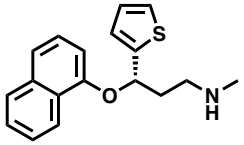
Percocet
Oxycodone (pain)



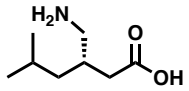
Celebrex
Celecoxib (NSAID)



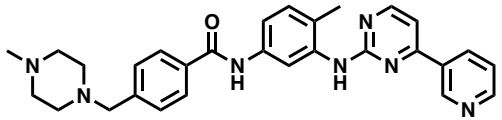
Nexium
Esomeprazole (dyspepsia)



Cymbalta
Duloxetine (SNRI)

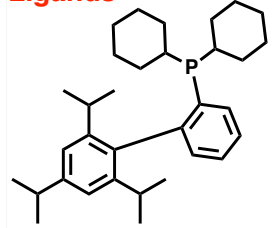


Lyrica
Pregabalin (epilepsy)



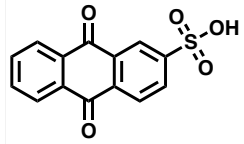
Gleevec
Imatinib (leukemia)

Ligands



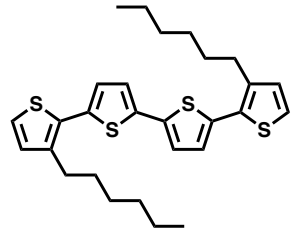
XPhos

Electrolytes



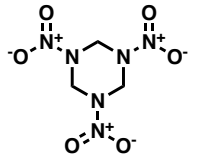
AMS

Organic electronics



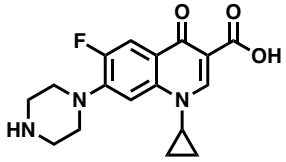
DH-4T

Energetics

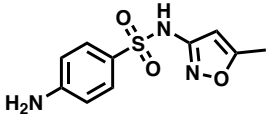


RDX

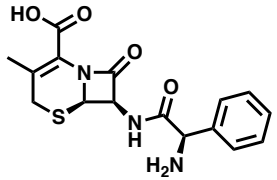
Antibiotics



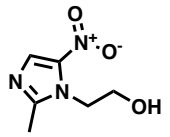
Ciprofloxacin



Sulfamethoxazole

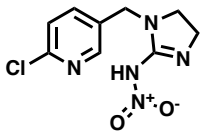


Cefalexin

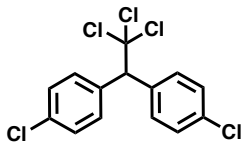


Metronidazole

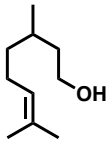
Pesticides



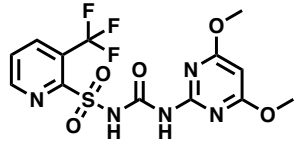
Imidacloprid



DDT

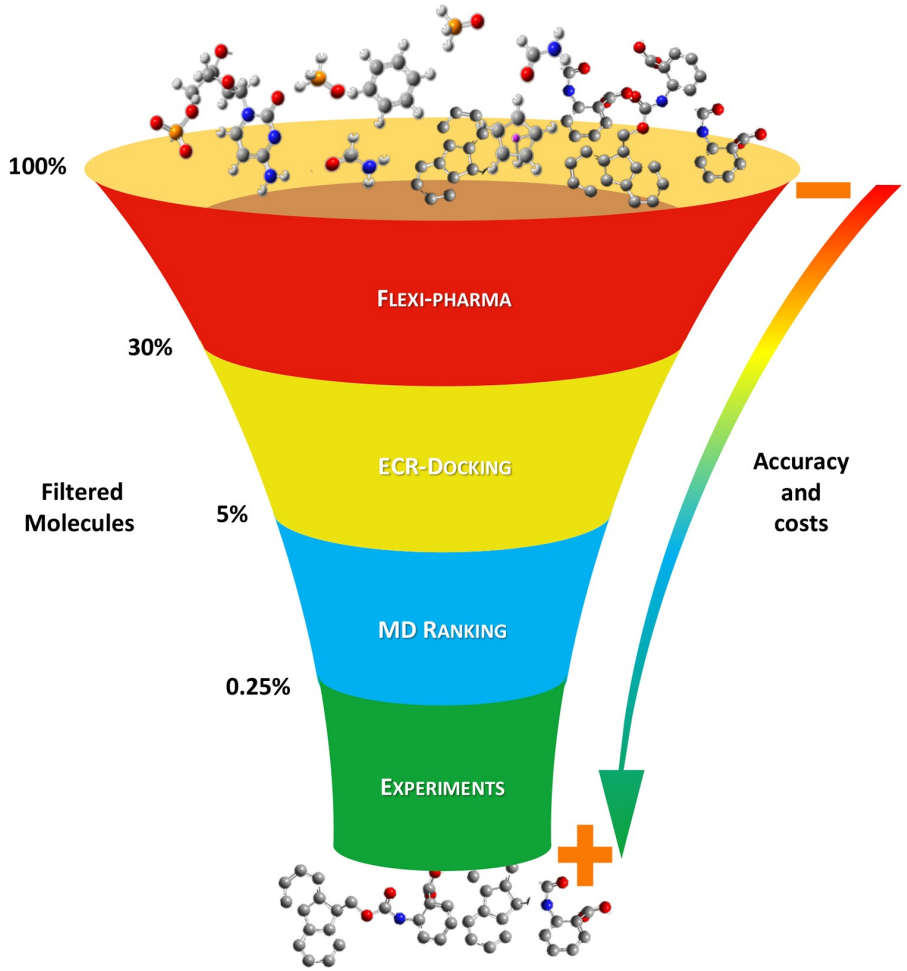


Citronellol



Flazasulfuron

Computer-aided molecular discovery pipelines still involve extensive manual intervention and are highly bespoke



1. "Considering a range of properties … as well as their commercial availability, 17 compounds were chosen as virtual screening hits"

2. "… the choice of these compounds was based on factors such as drug-likeness, availability for procurement, ligand efficiency and chemical diversity"

3. "The top-scoring molecules for the top-ranked 4,000 clusters were inspected for unfavourable features … From the remaining top-ranking clusters, we synthesized 17 richly functionalized THPs"

4. "all members were inspected … 40 molecules with ranks ranging from 16 to 246,721…were selected for de novo synthesis and testing."

[1] Lans, I. et al. *PLOS Computational Biology* **2020**, *16* (8), e1007898. https://doi.org/10.1371/journal.pcbi.1007898.
[2] Gorgulla, C et al. *Nature* **2020**, *580* (7805), 663–668. https://doi.org/10.1038/s41586-020-2117-z.
[3] Kaplan, A. L. et al. *Nature* **2022**, 1–10. https://doi.org/10.1038/s41586-022-05258-z.
[4] Stein, R. M. et al. *Nature* **2020**, *579* (7800), 609–614. https://doi.org/10.1038/s41586-020-2027-0.

The formulation of molecular optimization

best design

oracle

$$x^* = \arg\max_{x \in \mathcal{X}} f(x)$$

all possible molecules

surrogate oracle
(updated iteratively?)

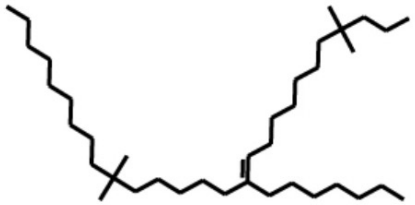
$$x^* \approx \arg\max_{x \in \mathcal{X}} \hat{f}(x)$$

1. Reliance on imperfect oracles
2. Constrained design spaces
3. Insufficient representations/surrogates
4. Non-sequential, batched design

Benchmarks for molecular optimization are toy problems

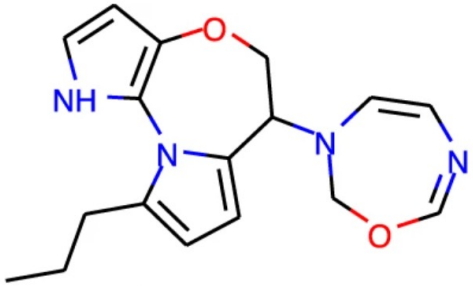
- The predominant benchmarks for molecular optimization are "penalized logP", a druglikeness heuristic (QED), and molecular "rediscovery" through similarity calculations Zhou et al. *Sci. Rep.* 2019

Penalized logP is trivially optimized with long alkyl chains



Penalized logP: 11.84

QED easily saturates at a score of 0.948



QED: 0.948

- But when these methods are evaluated with oracles that have any real level of complexity/utility (e.g., docking ≈ protein-ligand binding), performance is…uninspiring

Method	5HT1B	5HT2B	ACM2	CYP2D
CVAE for SMILES	-4.6	-4.2	-4.8	
GVAE for SMILES	-5.0	-4.6	-5.4	
LSTM for SMILES + REINFORCE	-9.8	-8.7	-9.8	-8.8
Training set (top 1%)	-11.5	-10.0	-10.0	-10.1
Virtual screening (top 1%)	-10.5	-9.8	-8.8	-9.3

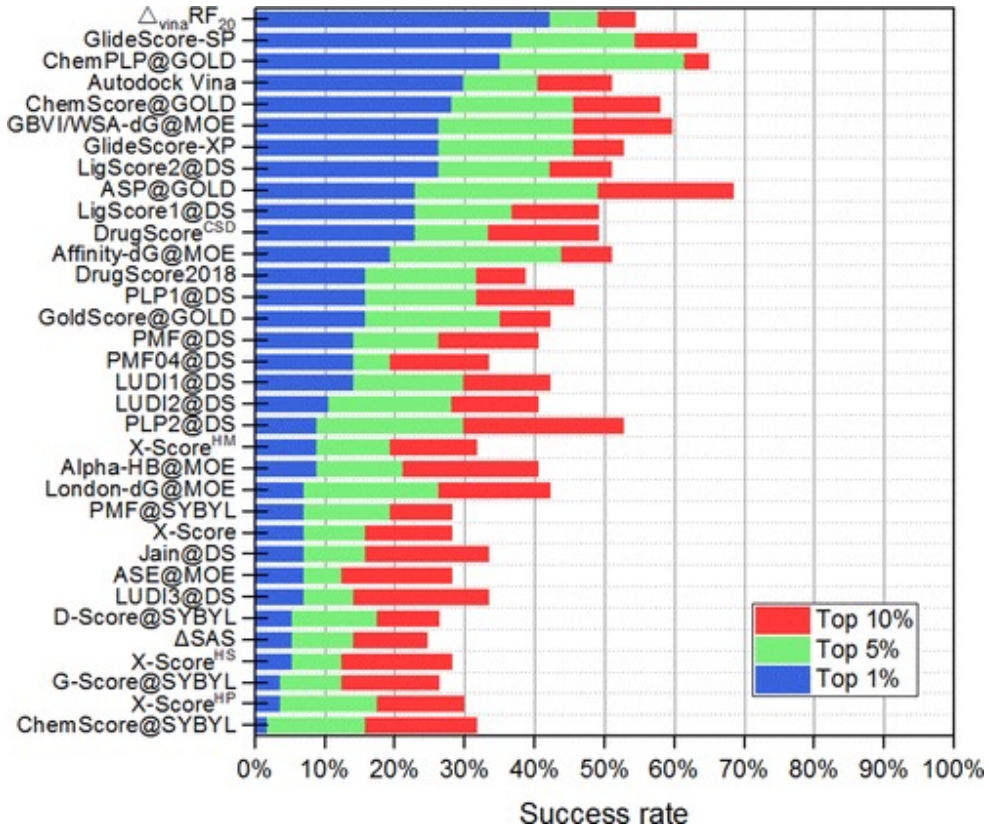
Average docking score of 250 compounds for different protein targets (lower = better)

Cieplinski et al. *J. Chem. Inf. Model.* 2023

We lack computational oracles for properties that matter

- Experimentally-relevant physical and biological properties cannot be predicted or simulated well
  - This includes binding affinity as a primary metric for therapeutic discovery

Docking scoring functions try to distinguish the highest-affinity ligands from decoys (CASF2016)



Su et al. *JCIM* 59, 895-913, 2018

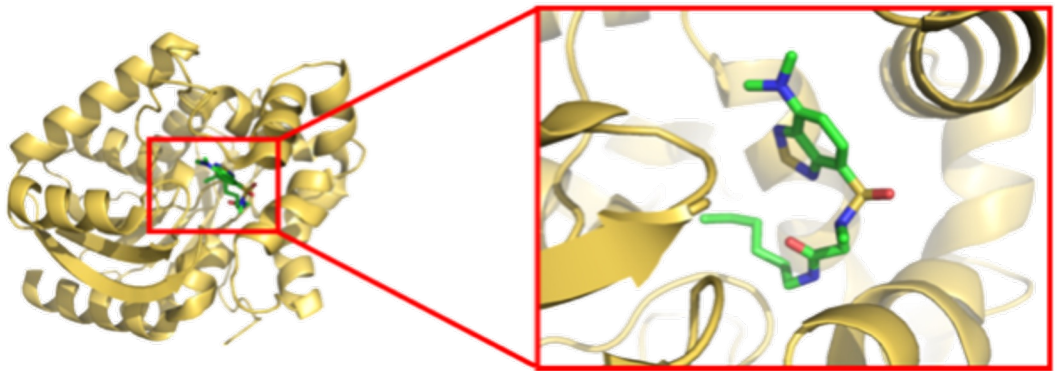


Image from profacgen.com

"$\Delta_{Vina}RF_{20}$ was calibrated on over 3300 protein–ligand complexes selected from the PDBbind v.2017 data set, which actually included 140 complexes (~50%) in the CASF-2016 test set."

We rarely know the failure modes of oracles well

- Experimentally-relevant physical and biological properties cannot be predicted or simulated well
  - This includes binding affinity as a primary metric for therapeutic discovery

"On inspection, these are not molecules that fit the receptor uniquely well, but rather molecules that cheat the scoring function by exploiting its holes and approximations."

Modeling the expansion of virtual screening libraries

Jiankun Lyu[1], **John J. Irwin**[1,*], **Brian K. Shoichet**[1,*]

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94158, USA

- If we could quantify (epistemic) uncertainty perfectly or knew the systematic biases, then we could incorporate this into the optimization process more robustly or just fix the oracle

The formulation of molecular optimization

best design

oracle

$$x^* = \arg\max_{x \in \mathcal{X}} f(x)$$

all possible molecules

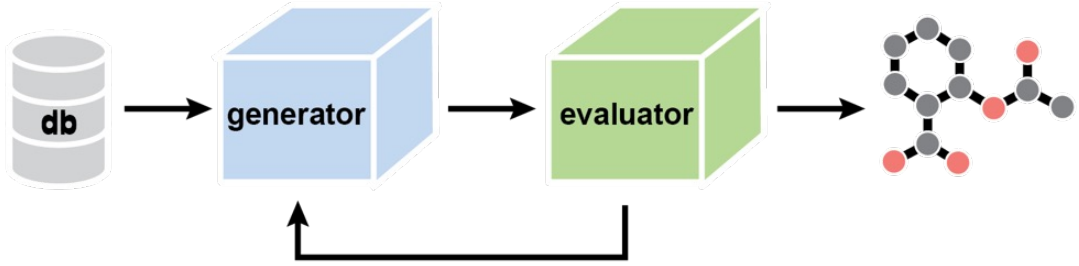
surrogate oracle
(updated iteratively?)

$$x^* \approx \arg\max_{x \in \mathcal{X}} \hat{f}(x)$$

1. Reliance on imperfect oracles
2. Constrained design spaces
3. Insufficient representations/surrogates
4. Non-sequential, batched design

Generative design is alluring due to its "creativity"

- De novo design of molecular structures can access chemical spaces beyond what is found in enumerated virtual libraries

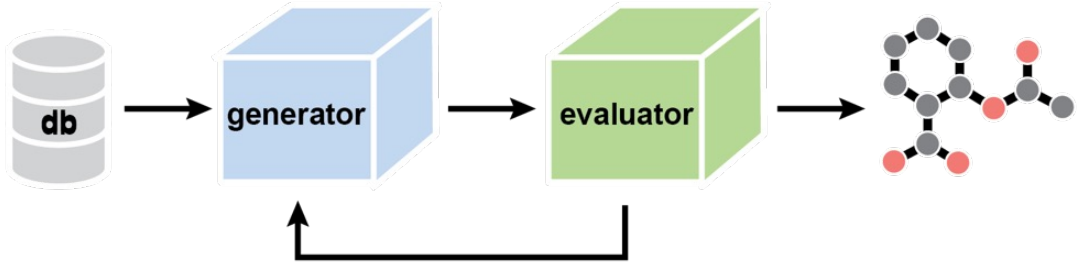


E.g., applying atom-by-atom generative modeling to PROTAC design

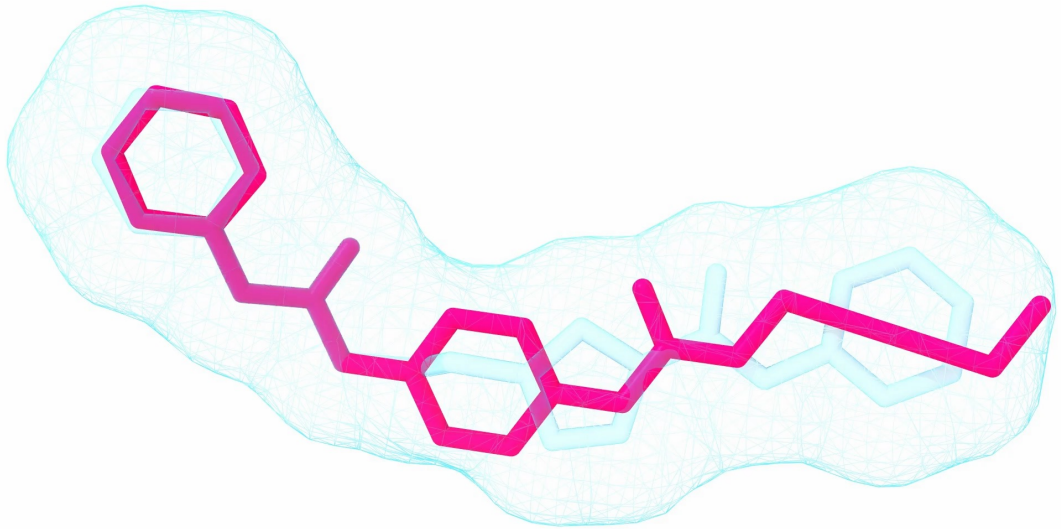
PROTAC = proteolysis targeting chimera

Generative design is alluring due to its "creativity"

- De novo design of molecular structures can access chemical spaces beyond what is found in enumerated virtual libraries

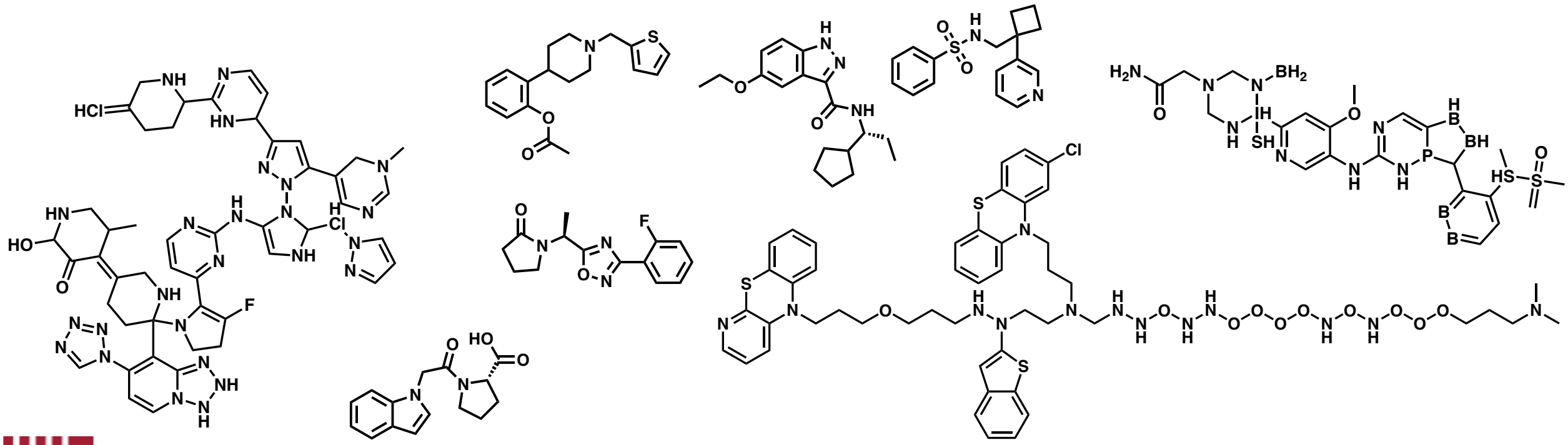


E.g., applying fragment-by-fragment generative modeling to 3D shape-conditioned design



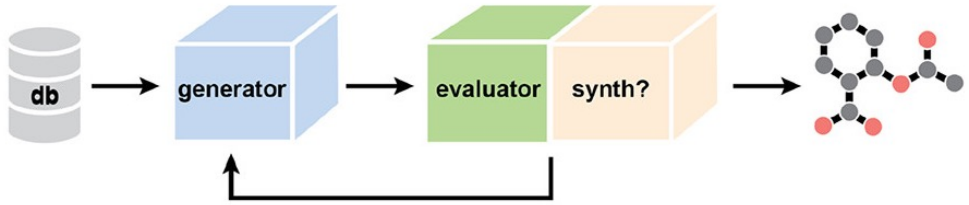
Generative design often results in bad solutions
(When applied to goal-conditioned molecular optimization, not distribution learning)

- There are many ways for molecules to be unreasonable despite being syntactically valid
  - E.g., lack of stability
  - E.g., lack of synthesizability
- Aside: the fact that we arrive at these structures as "optimal" molecules reflects the fact that our (surrogate) oracles are imperfect and have exploitable pathologies



Penalty functions can try to encourage "reasonableness"

1. Synthesizability heuristics (structure → scalar) can be incorporated into the objective function



SA Score
SCScore
RAScore
M1 Score
…

2. REINVENT (AstraZeneca), an LSTM that generates SMILES strings, is tied to its prior

$$\log \mathbf{P}_{\text{aug}}(T) = \log \mathbf{P}_{\text{prior}}(T) + \sigma \mathbf{S}(T)$$

"augmented likelihood" which should be higher for "better" molecules

"prior likelihood" which was trained on a large database of molecules

"score" which measures a property we are trying to maximize

Retrosynthetic planning can be applied as a filter

Output

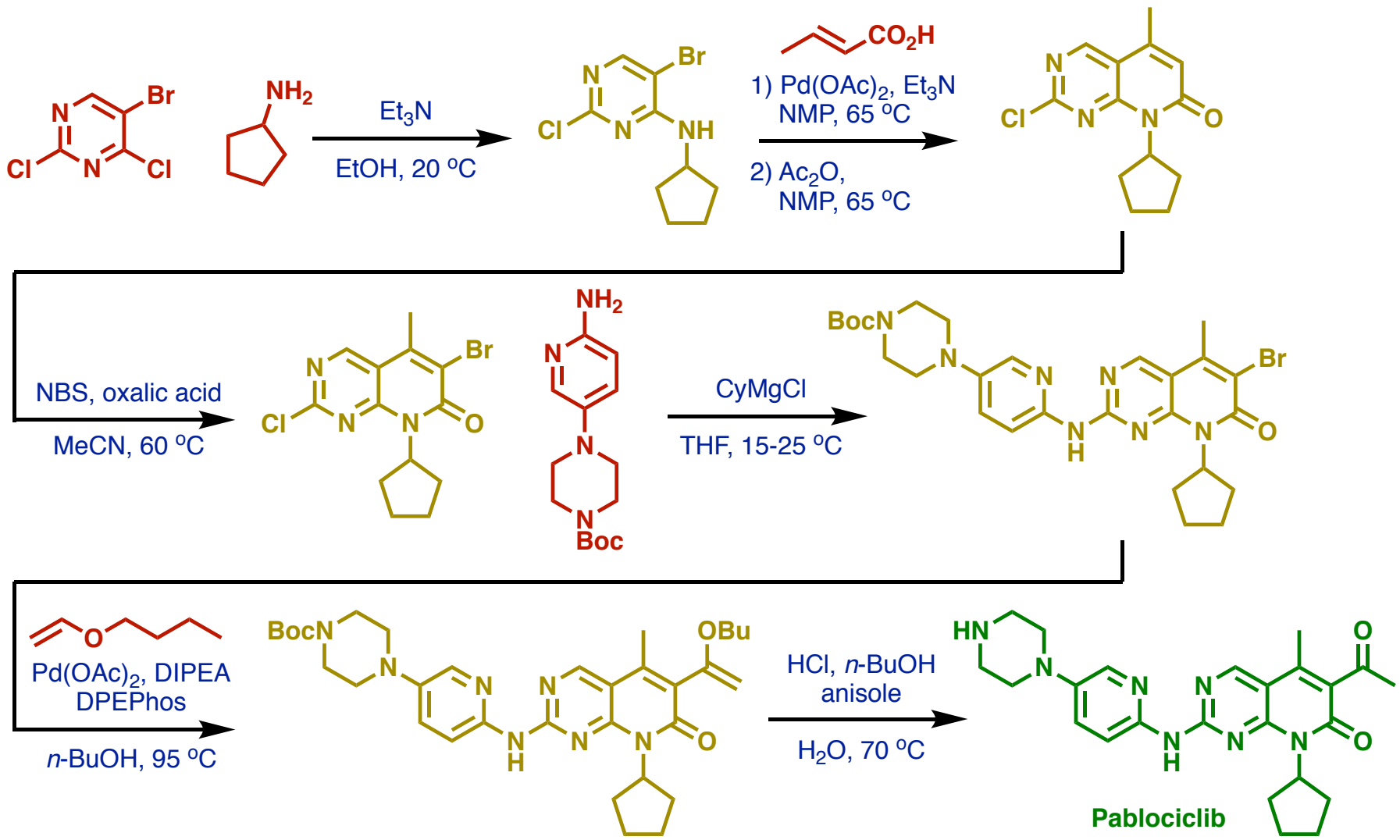
Reactants
(starting materials)

Intermediates

Conditions
+ concentrations
and reaction times

Input

Product
(target compound)



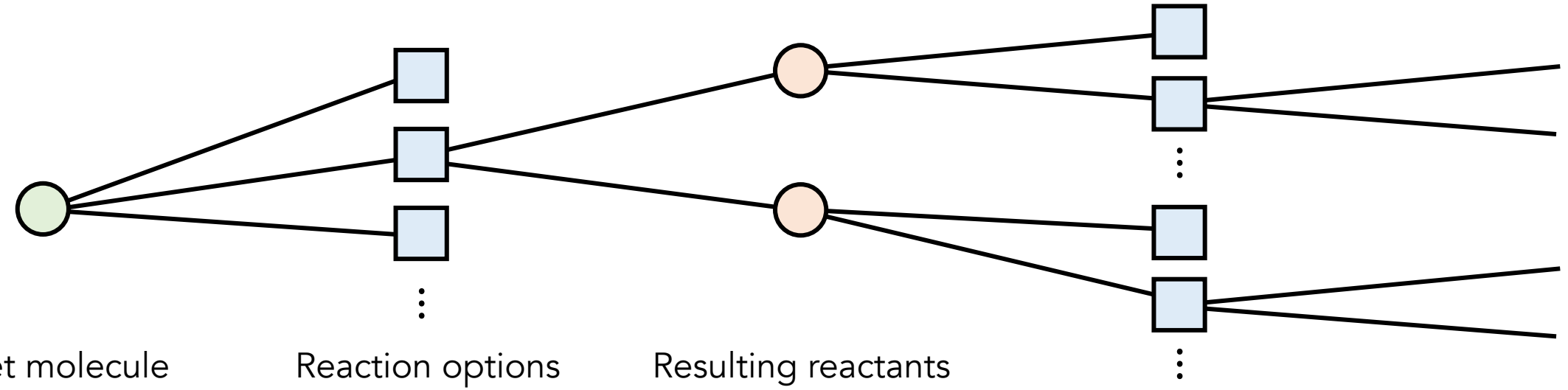
Retrosynthetic analysis requires a few key components

(1) A "one-step" method to generate plausible reactants given a product
Unlike game play, we do not have a world model for chemistry

(2) A method to apply this "one-step" method recursively and navigate the resulting combinatorial space of options
Because the graph of possible options must be generated on-the-fly from one-step predictions, exploration can be quite expensive

These are typically approached and evaluated separately

(3) Some termination criterion (e.g., commercial availability)



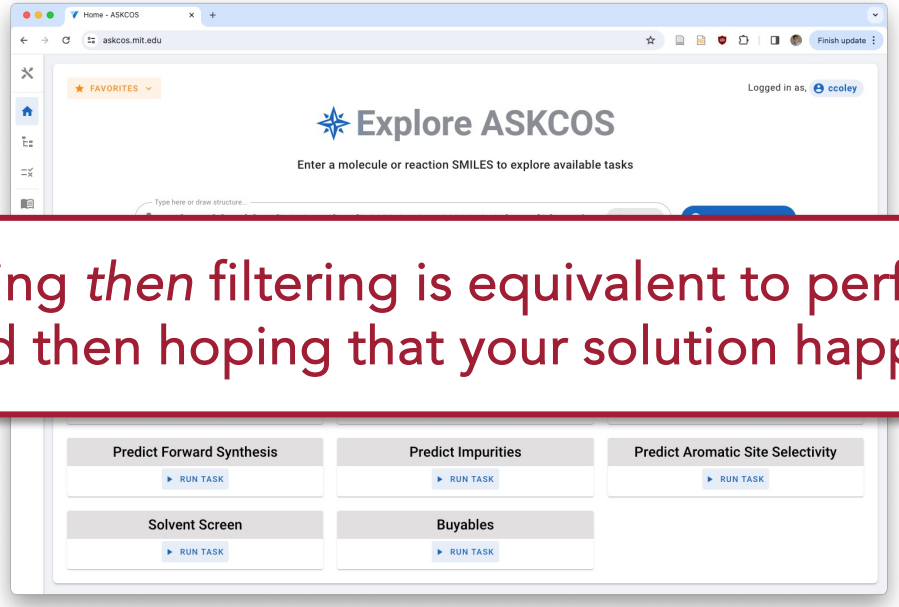
Retrosynthetic analysis requires a few key components

(1) A "one-step" method to generate plausible reactants given a product
Unlike game play, we do not have a world model for chemistry

(2) A method to apply this "one-step" method recursively and navigate the resulting combinatorial space of options
Because the graph of possible options must be generated on-the-fly from one-step predictions, exploration can be quite expensive

These are typically evaluated separately

(3) Some termination criterion (e.g., commercial availability)



Optimizing *then* filtering is equivalent to performing an unconstrained optimization and then hoping that your solution happens to be in your feasible region

Demo @ askcos.mit.edu
(running on small server)

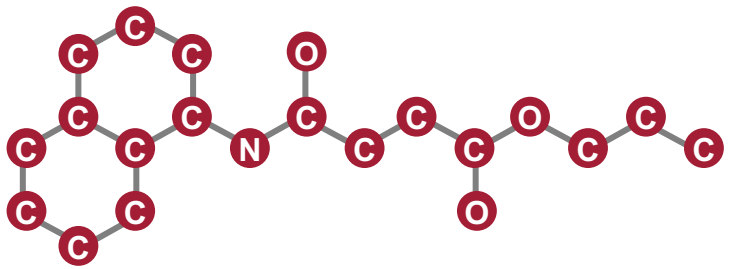
Enforcing strict design space constraints: synthesizability

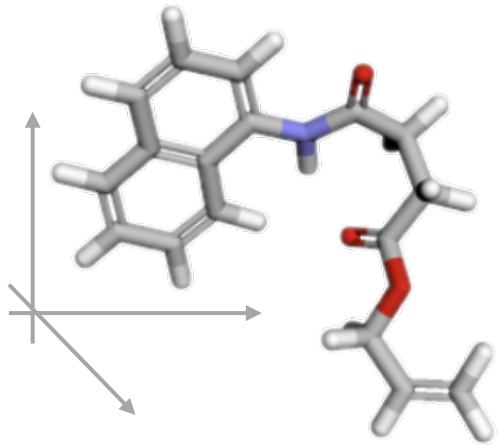
Instead of using generative AI to propose molecules, one can propose experimental procedures

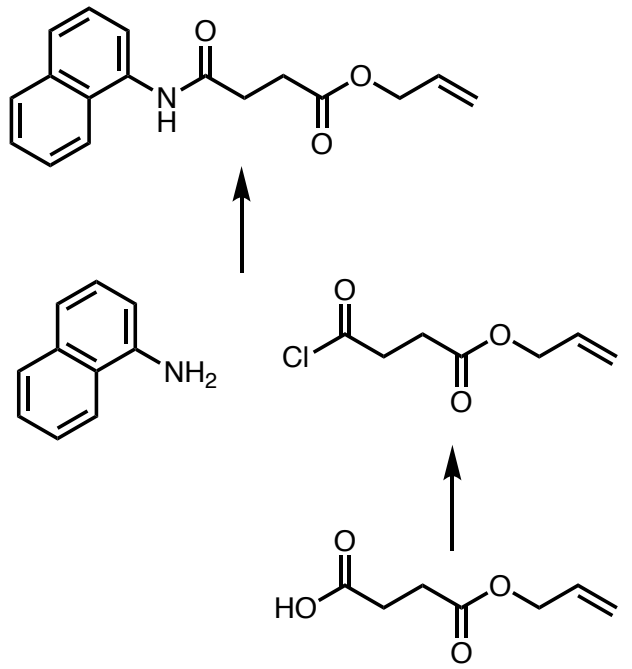
character-by-character
atom-by-atom
fragment-by-fragment

reaction-by-reaction

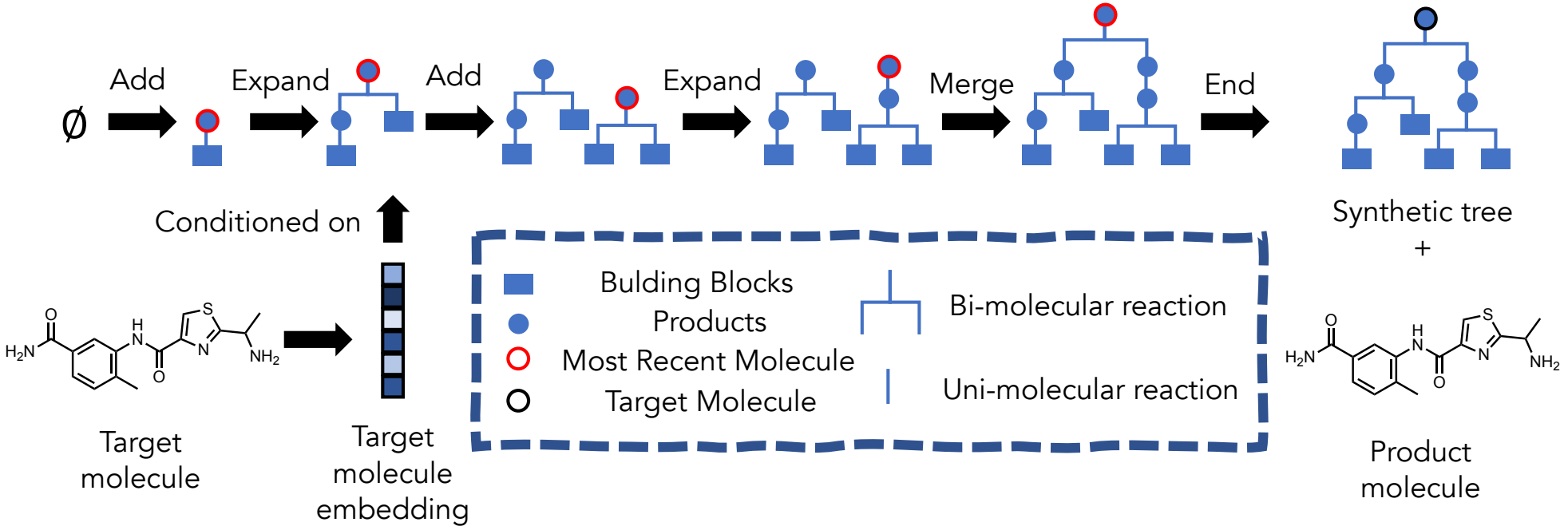

```
C=COC(=O)CCC(=O)
Nc1cccc2ccccc12
```







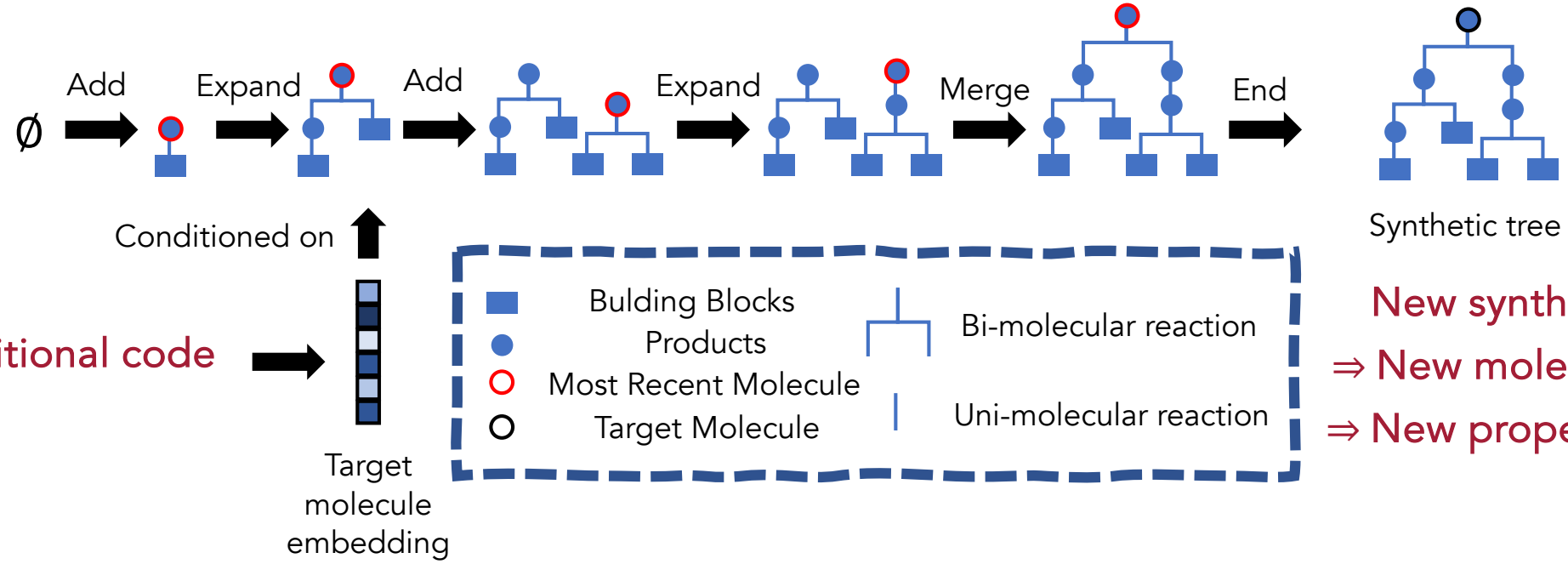
Generative design of experimental procedures



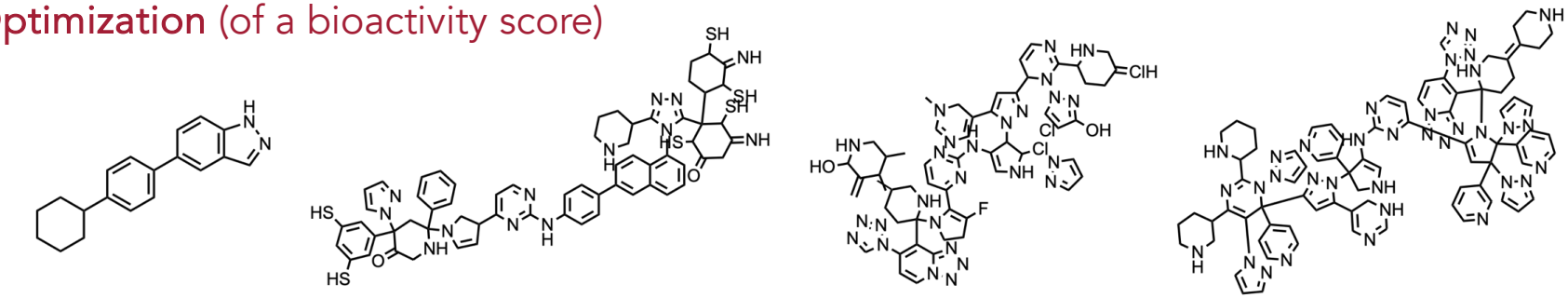
O(100,000) commercial building blocks

O(100) expert-defined reaction templates

Generative design of experimental procedures



Optimization (of a bioactivity score)



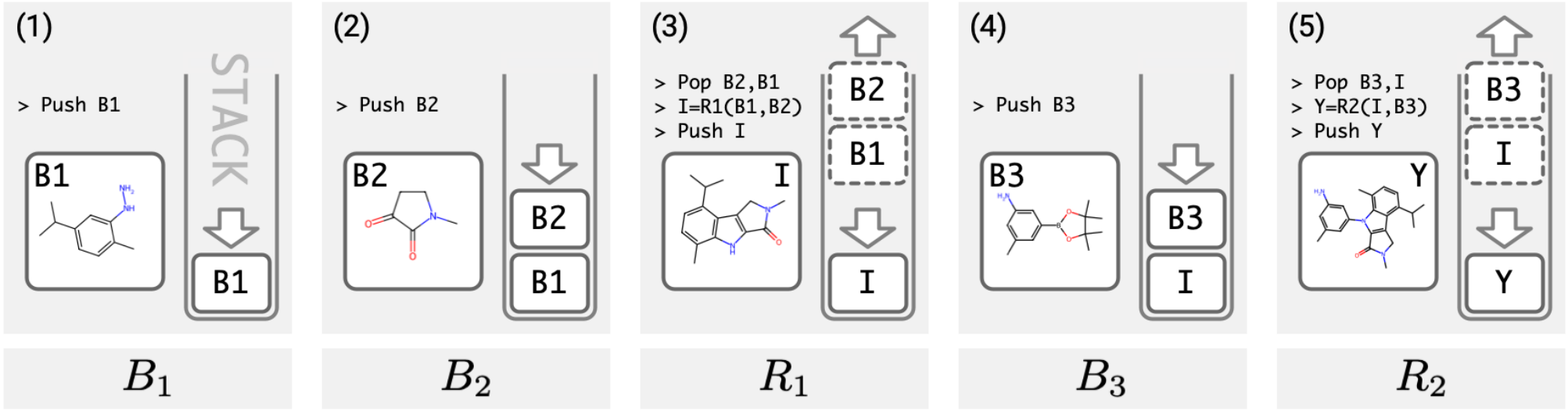
Top-1 from our model
GSK3β = 0.94

Top-1 from DST
GSK3β = 0.97

Top-1 from MARS
GSK3β = 0.95

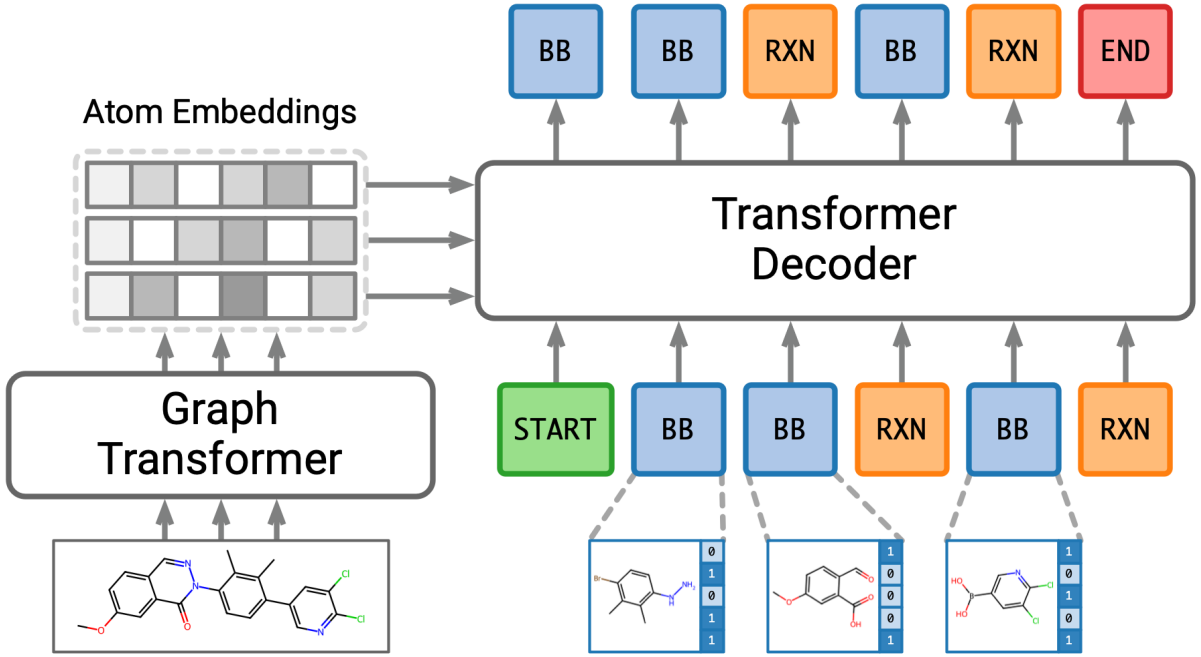
Top-1 from GA+D
GSK3β = 0.79

Generative design of experimental procedures (with transformers)



Encoder-decoder architecture

Postfix notation for sequence decoding
e.g., R2(R1(B1, B2), B3)



Generative design of experimental procedures (with transformers)

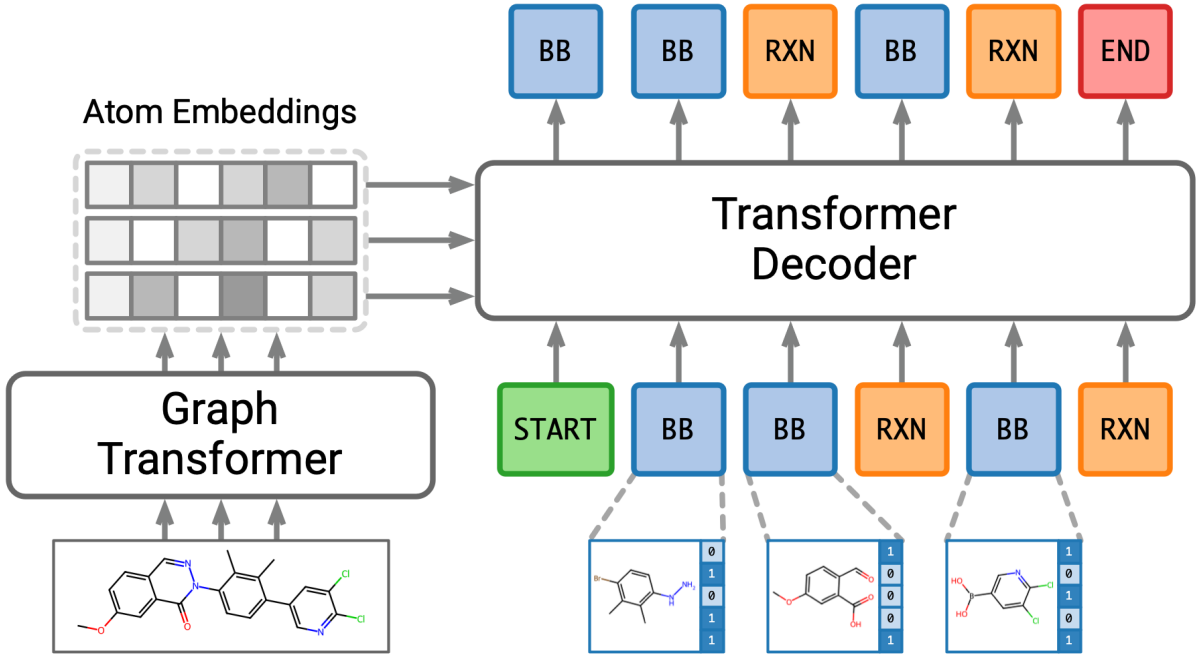
Even if we cannot precisely recover the molecule that is encoded, we get a synthesizable analog

Dataset	Method	Success%	Recons.%	Sim.(Morgan)	Sim.(Scaffold)	Sim.(Gobbi)
Test Set	SynNet	84.1%	10.7%	0.4575	0.5109	0.3465
	Proposed	**97.5%**	**28.4%**	**0.7167**	**0.7791**	**0.7273**

← Direct upgrade over prior work

Encoder-decoder architecture

Postfix notation for sequence decoding
e.g., R2(R1(B1, B2), B3)



The formulation of molecular optimization

best design

oracle

$$x^* = \arg\max_{x \in \mathcal{X}} f(x)$$

all possible molecules

surrogate oracle
(updated iteratively?)

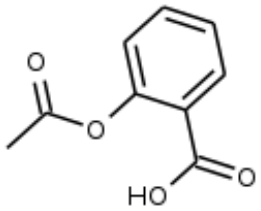
$$x^* \approx \arg\max_{x \in \mathcal{X}} \hat{f}(x)$$

1. Reliance on imperfect oracles
2. Constrained design spaces
3. Insufficient representations/surrogates
4. Non-sequential, batched design

What is a molecule?

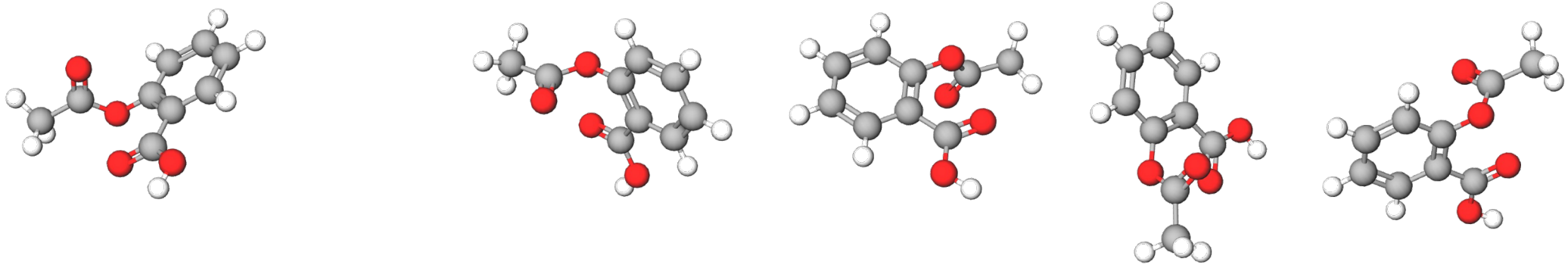
- Most models $\hat{f}(x)$ represent a molecular structure-relationship, requiring the choice of molecular representation and embedding strategy

Graph? SMILES string? 3D conformer? 4D ensemble of conformers?



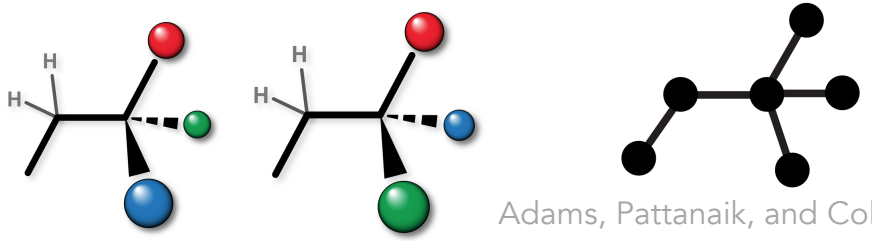
≡ aspirin ≡ `O=C(C)Oc1ccccc1C(=O)O`

```
c1ccc(C(=O)O)c(c1)OC(=O)C
O=C(C)Oc1c(cccc1)C(O)=O
O=C(C)Oc1c(C(O)=O)cccc1
c1cc(C(=O)O)c(OC(C)=O)cc1
c1c(C(O)=O)c(OC(=O)C)ccc1
...
```



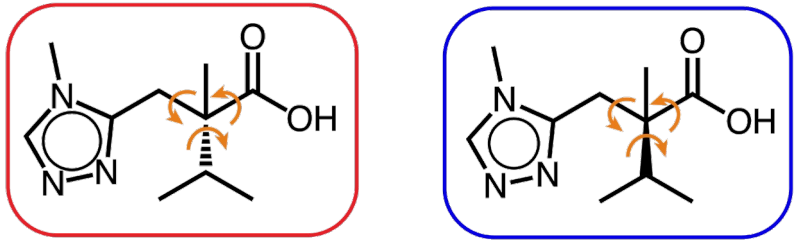
Molecular representations & stereochemistry

- Molecular ML pipelines have been overbuilt for SMILES strings parsed into covalent bond graphs
- Enantiomers have identical graph connectivity, so 'vanilla' GNNs cannot distinguish them
- Can a 3D model distinguish stereoisomers without getting confused by conformational flexibility?



Adams, Pattanaik, and Coley *ICLR* 2022

- Chiral InterRoto-Invariant Neural Network (ChIRo) uses continuous symmetries to make it invariant to single dihedral rotations



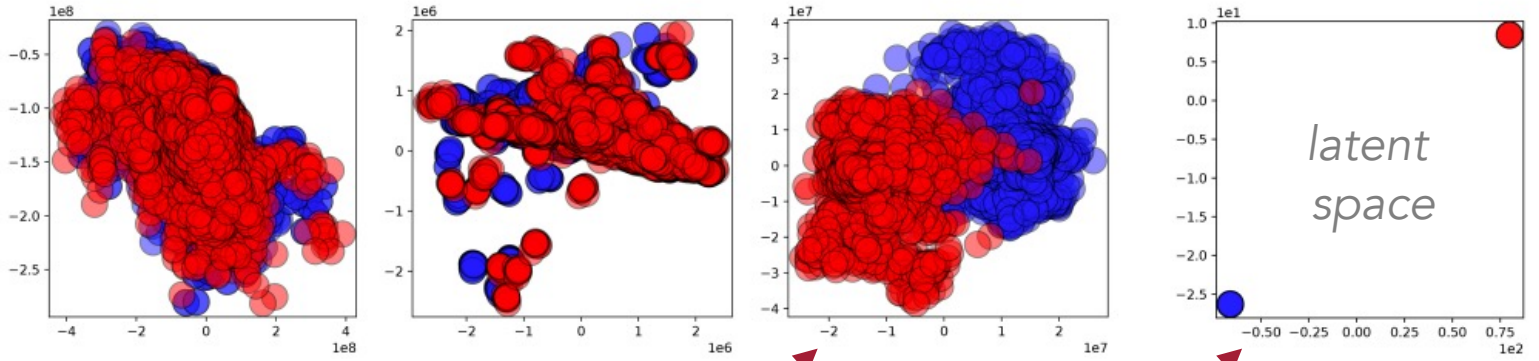
Consider two enantiomers: red and blue; we can enumerate conformers of each

SchNet
E(3)-invariant

DimeNet++
E(3)-invariant

SphereNet
SE(3)-invariant

ChIRo
{SE(3)+rotation}-invariant



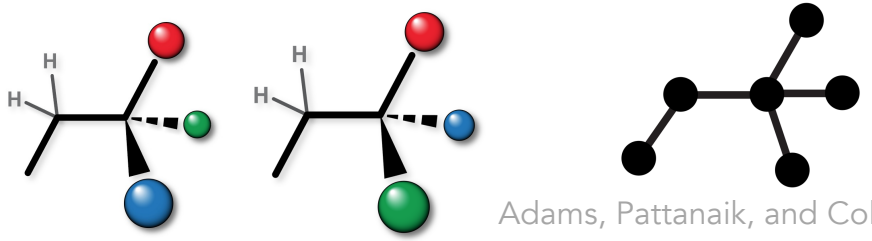
can't distinguish

can, but messy

can by design!

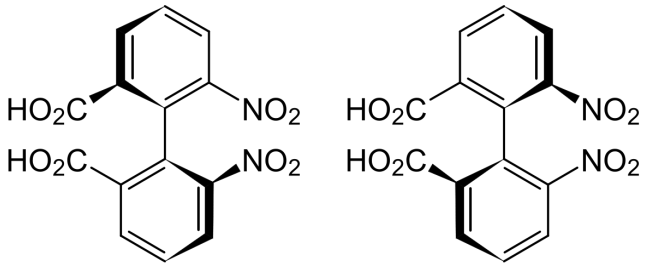
Molecular representations & stereochemistry

- Molecular ML pipelines have been overbuilt for SMILES strings parsed into covalent bond graphs
- Enantiomers have identical graph connectivity, so 'vanilla' GNNs cannot distinguish them
- Can a 3D model distinguish stereoisomers without getting confused by conformational flexibility?

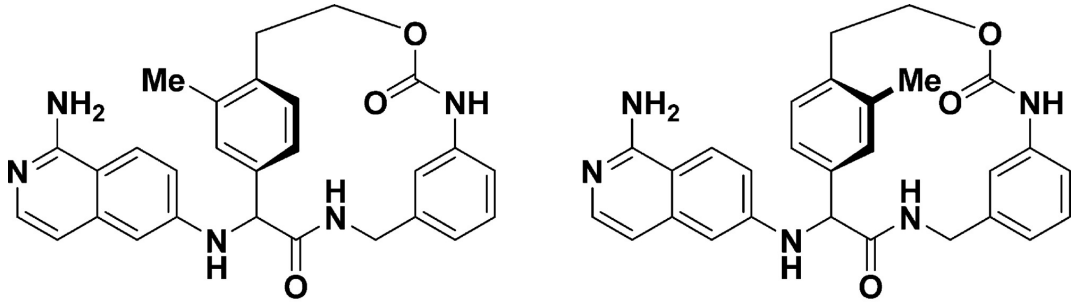


Adams, Pattanaik, and Coley *ICLR* **2022**

- Chiral InterRoto-Invariant Neural Network (ChIRo) uses continuous symmetries to make it invariant to single dihedral rotations
- However, we don't actually want to be invariant to single dihedral rotations…



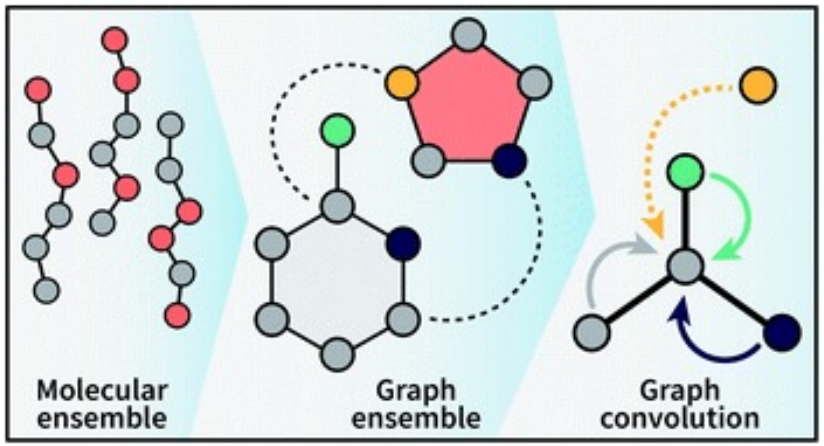
Wikipedia



J. Med. Chem. 59(8) 4007–4018, **2016**

Some materials of interest lack well-defined structures

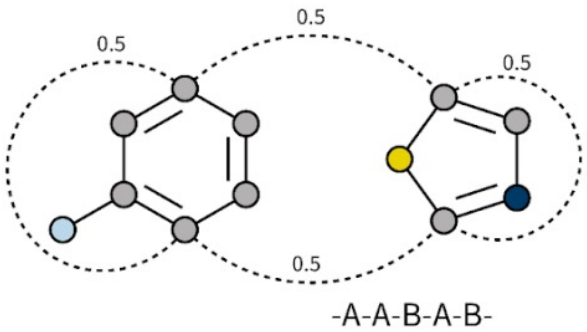
- Synthetic polymers are rarely *sequence-defined* like proteins and nucleic acid sequences are
- They are best described by distributions of chain lengths and monomer connections
- We represent an ensemble by parameters of the connections that generate them from constituent monomers



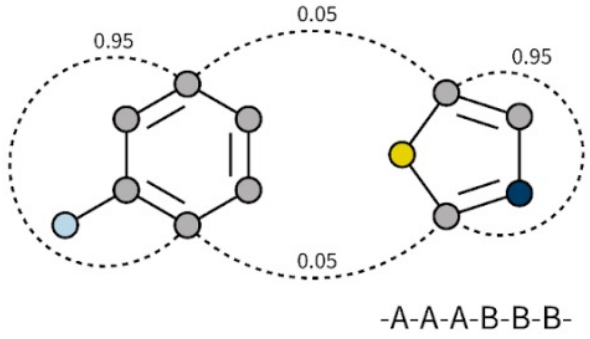
Some materials of interest lack well-defined structures

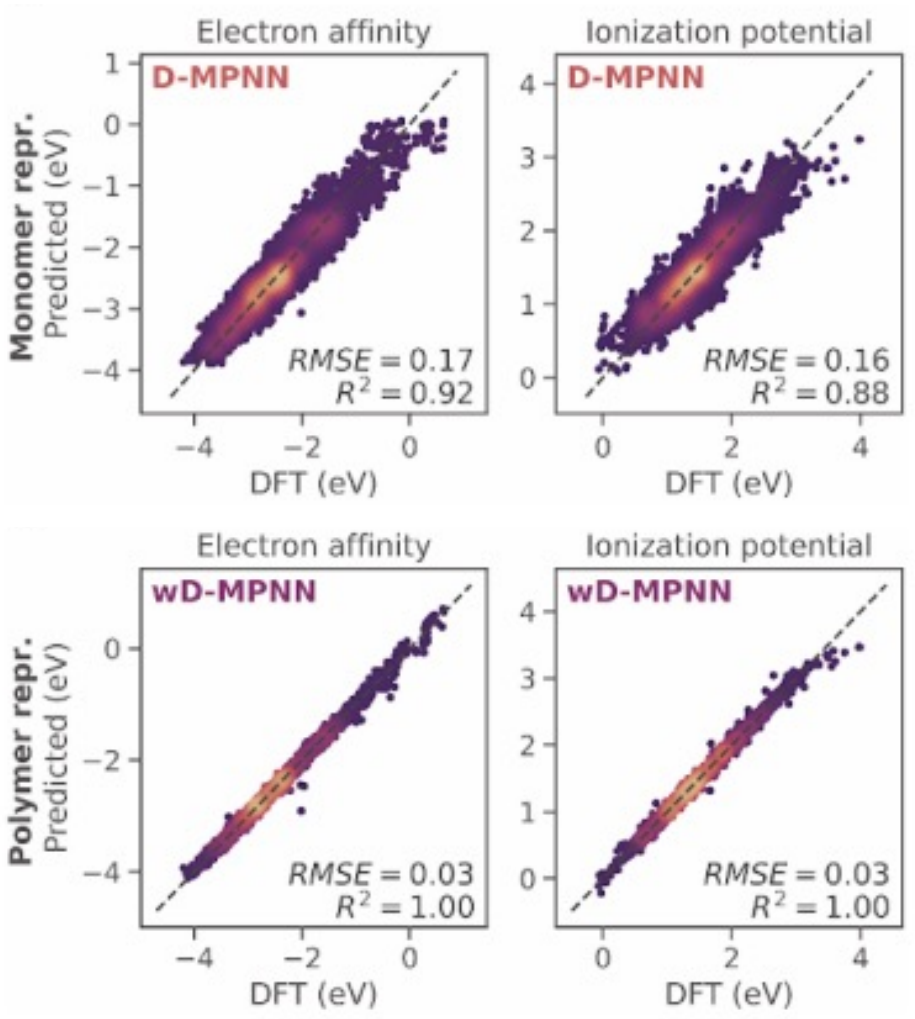
- Synthetic polymers are rarely *sequence-defined* like proteins and nucleic acid sequences are
- They are best described by distributions of chain lengths and monomer connections
- We represent an ensemble by parameters of the connections that generate them from constituent monomers
  - *Limitation:* this captures the first moment (average) of the distribution but not higher-order moments

Random copolymer



Block copolymer

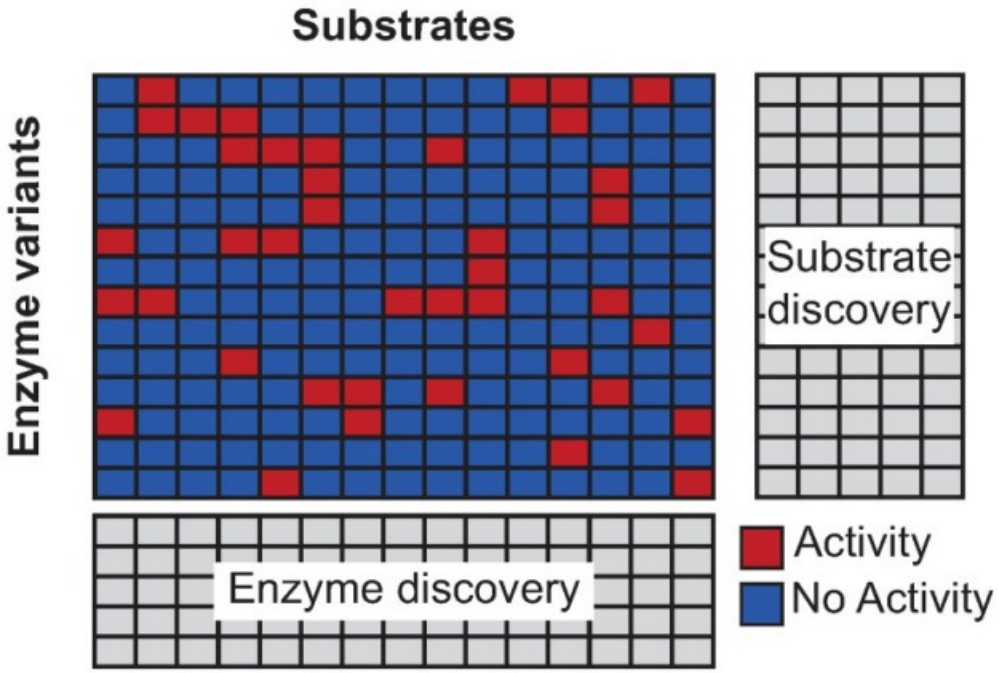




Weighted edges improve the fit to DFT data

We lack good representations of mixtures & interactions

- Interactions – exemplified by compound-protein interactions – are poorly captured by discriminative models (only apparent if the proper baselines are included)

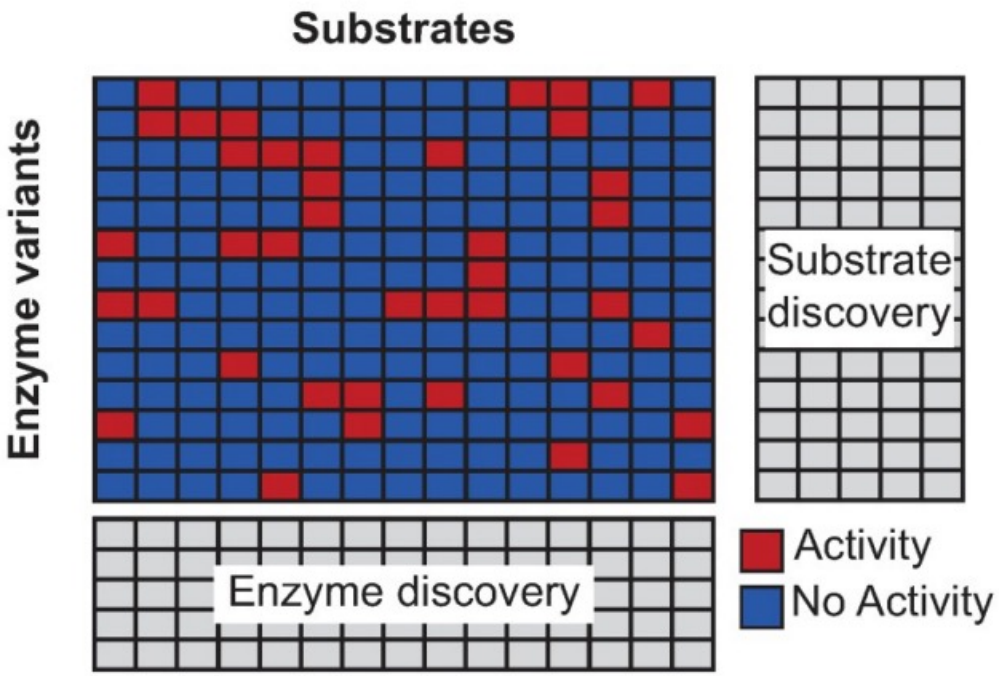


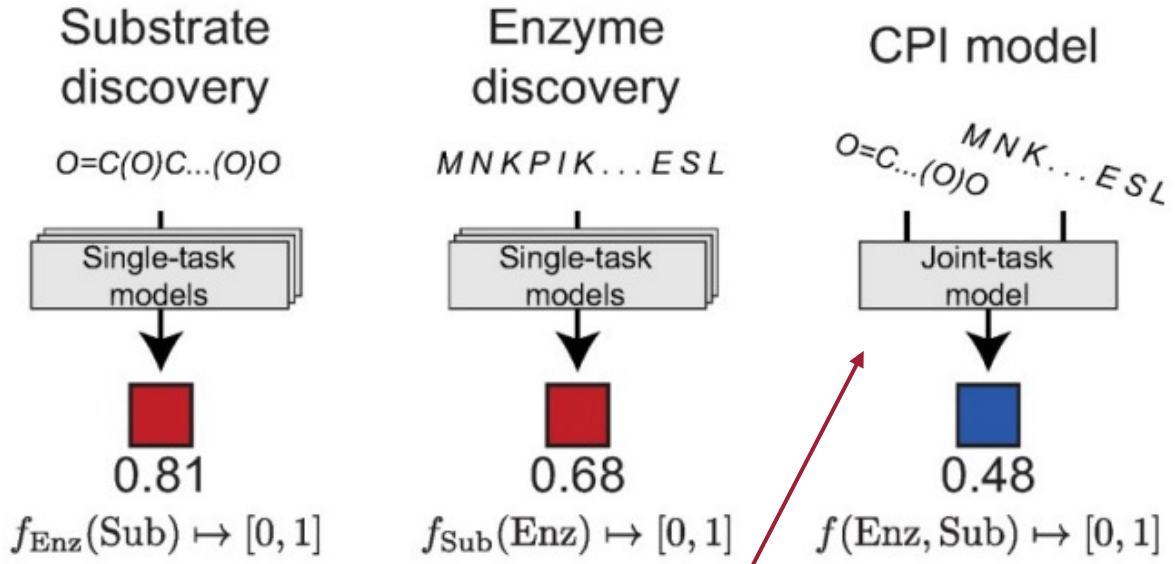
Dataset	# Enzymes	# Substrates	# Pairs
Halogenase	42	62	2604
Glycosyltransferase	54	90	4298
Thiolase	73	15	1095
BKACE	161	17	2737
Phosphatase	218	165	35970
Esterase	146	96	14016
Kinase (inhibitors)	318	72	22896

How well can we generalize from this 'dense' family-wide enzyme profiling data?

We lack good representations of mixtures & interactions

- Interactions – exemplified by compound-protein interactions – are poorly captured by discriminative models (only apparent if the proper baselines are included)



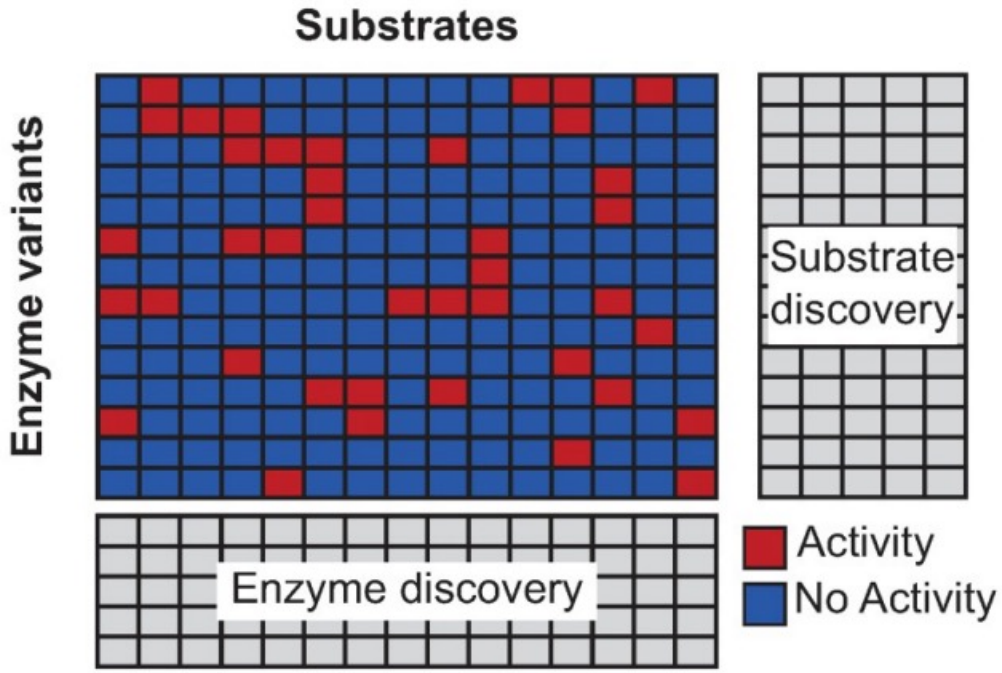


How well can we generalize from this 'dense' family-wide enzyme profiling data?

1. *two embedding MLP 'trunks'*
2. *interaction layer (concat., sum, outer prod., dot prod., …)*
3. *additional MLP*

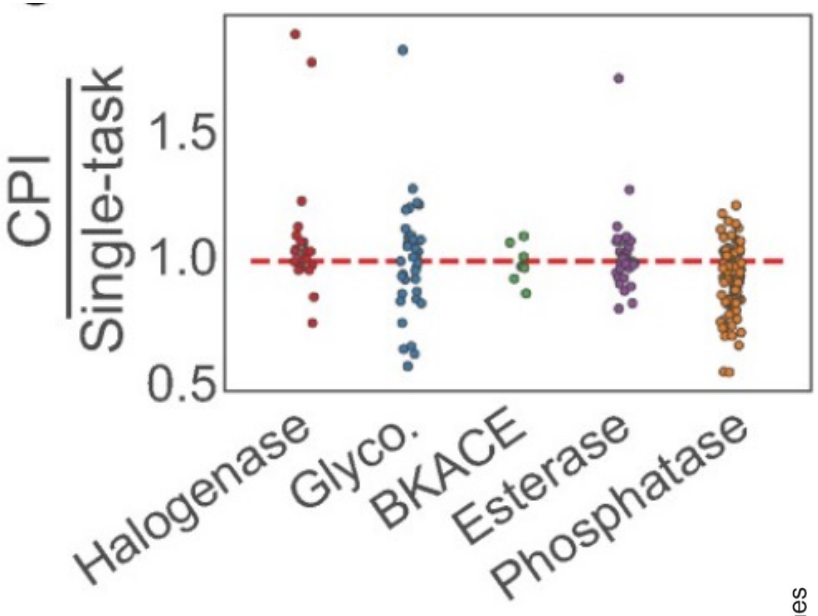
We lack good representations of mixtures & interactions

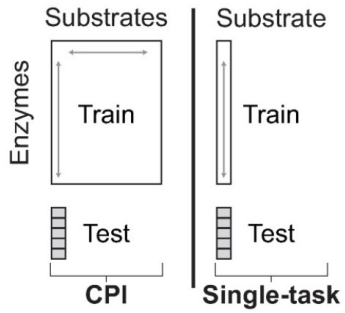
- Interactions – exemplified by compound-protein interactions – are poorly captured by discriminative models (only apparent if the proper baselines are included)



How well can we generalize from this 'dense' family-wide enzyme profiling data?

Enzyme-substrate compatibility (binary classification)

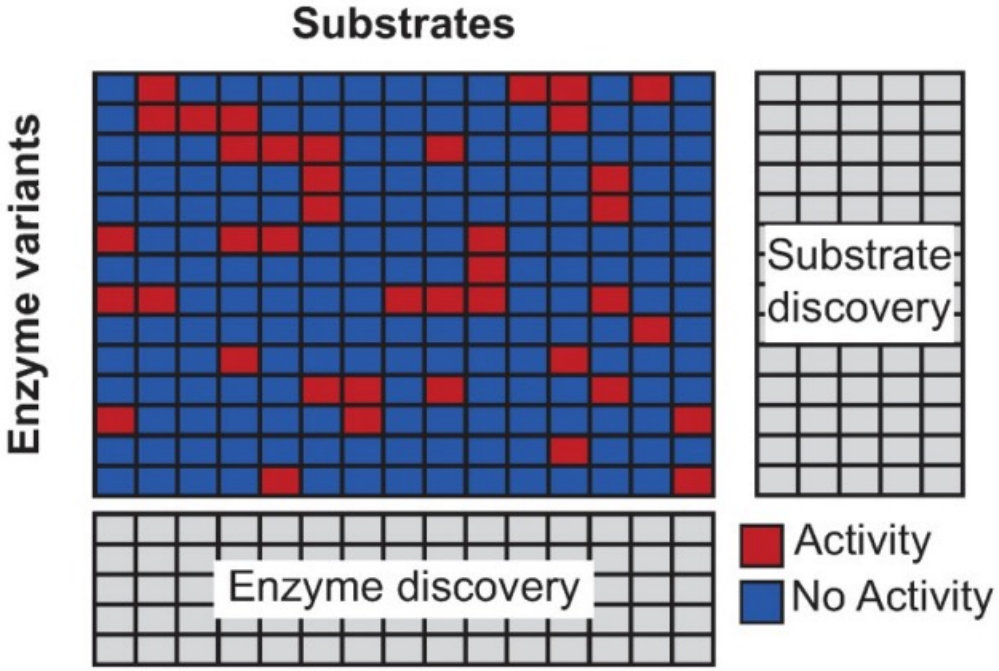




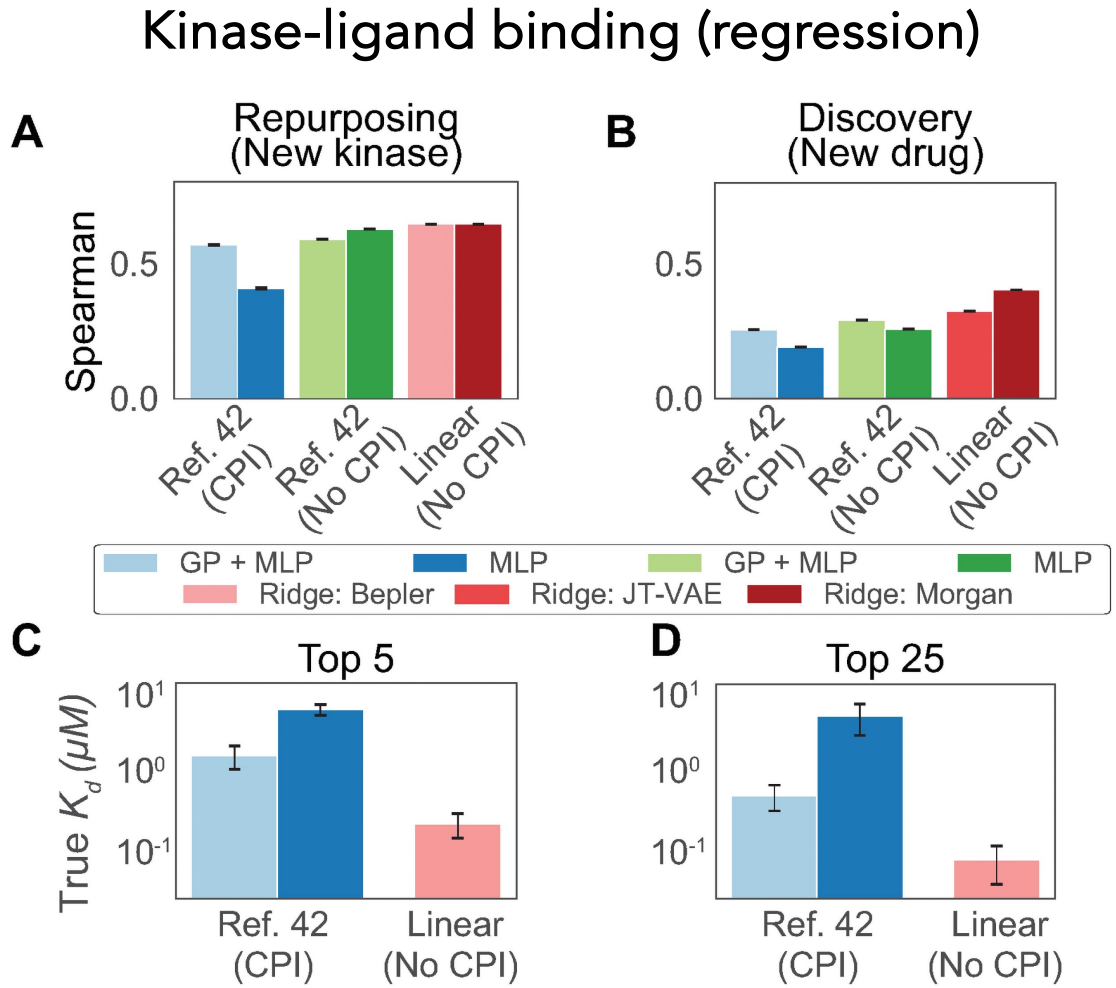
Goldman et al., *PLOS Comp. Bio* 2022

We lack good representations of mixtures & interactions

- Interactions – exemplified by compound-protein interactions – are poorly captured by discriminative models (only apparent if the proper baselines are included)



How well can we generalize from this 'dense' family-wide enzyme profiling data?



The formulation of molecular optimization

best design

oracle

$$x^* = \arg\max_{x \in \mathcal{X}} f(x)$$

all possible molecules

surrogate oracle
(updated iteratively?)

$$x^* \approx \arg\max_{x \in \mathcal{X}} \hat{f}(x)$$

1. Reliance on imperfect oracles
2. Constrained design spaces
3. Insufficient representations/surrogates
4. Non-sequential, batched design

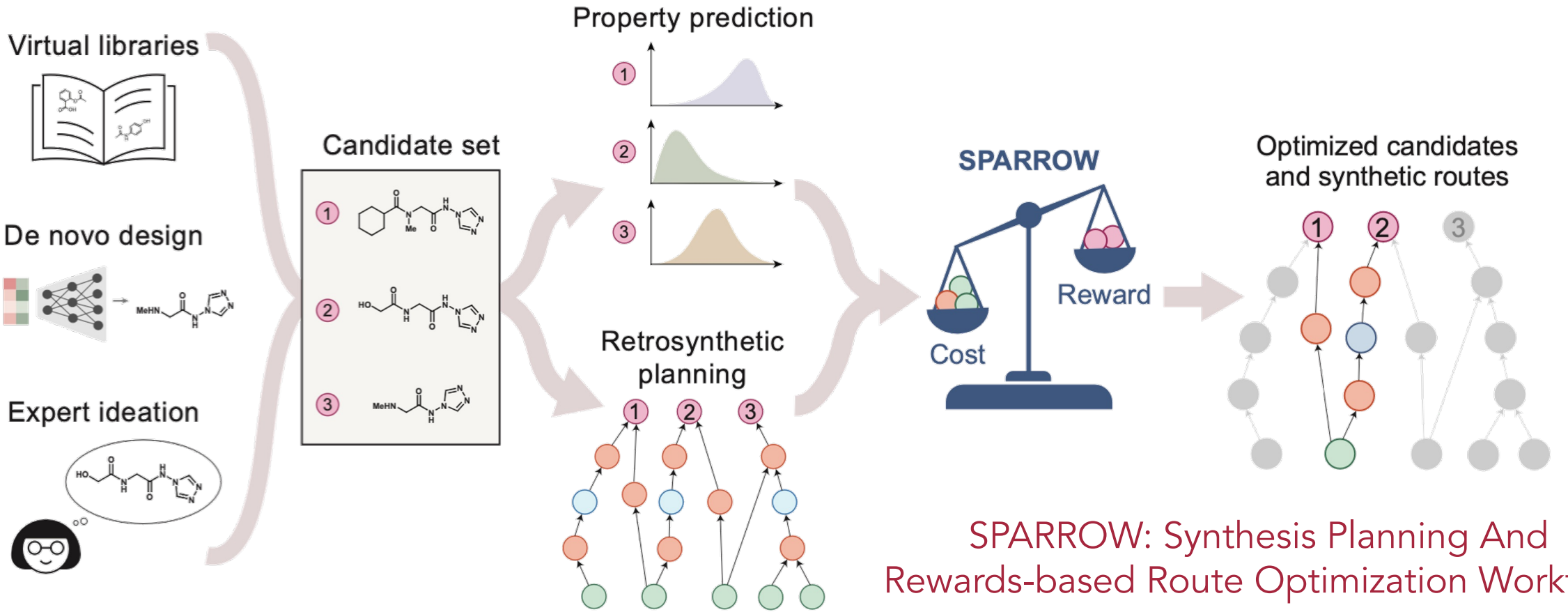
Molecular discovery workflows are not truly sequential

- We do not just make (or buy) and test a single molecule at a time – we do so in *batches*
- We should really consider the utility <u>of the batch</u> against the cost <u>of the batch</u> (not traditional BO)

$$x^* = \arg\max_{x \in \mathcal{X}} \; f(x) \quad \xrightarrow{\textit{one option}} \quad \mathcal{X}_b^* = \arg\max_{\mathcal{X}_b \subset \mathcal{X}, |\mathcal{X}_b| = b} \; P(x^* \in \mathcal{X}_b)$$

Molecular discovery workflows are not truly sequential

- We do not just make (or buy) and test a single molecule at a time – we do so in *batches*
- We should really consider the utility <u>of the batch</u> against the cost <u>of the batch</u> (not traditional BO)



SPARROW: Synthesis Planning And Rewards-based Route Optimization Workflow

Synthesis-aware design of *batches* of molecules

- We do not just make (or buy) and test a single molecule at a time – we do so in *batches*
- We should really consider the utility <u>of the batch</u> against the cost <u>of the batch</u> (not traditional BO)

Decision variable defining if candidate j is selected

Reward for candidate j

Likelihood that reaction i is successful

Expected reward

Set of reactions selected to produce candidate j

$$\underset{\boldsymbol{c},\boldsymbol{r}}{\arg\max} \frac{\sum_{j\in\mathcal{T}}\left(c_j U_j \prod_{i\in\mathcal{R}_j} L_i\right)}{\text{cost}\left(\{\mathcal{R}_j \,\forall j \in \mathcal{T} : c_j = 1\}\right)}$$

Set of all candidate molecules

Total cost of synthesizing all selected routes

Constraints

(1) $c_j \geq r_j \quad \forall j \in \mathcal{P}_j, i \in \mathcal{R}$

If a compound node is selected, at least one of its parent reactions must be selected.

(2) $\sum_{i\in\mathcal{P}_j} r_i \geq c_j \quad \forall j \in \mathcal{C}$

If a reaction is selected, all of its parent compound nodes (its reactants) must also be selected.

(3) $\sum_{i\in\mathcal{Y}} r_i \leq \text{length}(\mathcal{Y}) - 1 \quad \forall \mathcal{Y}$

For each cycle in the graph, every reaction node in the cycle cannot be simultaneously selected.

Additional notation

- j: An index referring to a reaction node
- i: An index referring to a compound node
- $\mathcal{R}$: Set of reaction node indices
- $\mathcal{C}$: Set of compound node indices
- c_j: Decision variable defining whether compound node j is selected
- r_i: Decision variable defining whether reaction node i is selected
- $\mathcal{S}$: Set of dummy reaction node indices
- $\mathcal{P}_{i \text{ or } j}$: Set of parent nodes for the node corresponding to index i or j
- $\mathcal{Y}$: A cycle in a retrosynthetic graph

Synthesis-aware design of *batches* of molecules

- We do not just make (or buy) and test a single molecule at a time – we do so in *batches*
- We should really consider the utility <u>of the batch</u> against the cost <u>of the batch</u> (not traditional BO)

Cost of starting material produced by dummy reaction i

Penalty for selecting reaction i

$$\underset{\boldsymbol{c},\boldsymbol{r}}{\arg\min} \; \underbrace{-\lambda_1 \sum_{j \in \mathcal{T}} c_j U_j}_{\text{Select candidates with high rewards}} + \underbrace{\lambda_2 \sum_{i \in \mathcal{S}} D_i r_i}_{\text{Select cheap starting materials}} + \underbrace{\lambda_3 \sum_{i \in \mathcal{R}} \min\{L_i^{-1}, 20\} r_i}_{\text{Select few reactions and ones that are likely to be successful}}$$

Whether reaction i is selected

Constraints

(1) $c_j \geq r_j \quad \forall j \in \mathcal{P}_j, i \in \mathcal{R}$

If a compound node is selected, at least one of its parent reactions must be selected.

(2) $\sum_{i \in \mathcal{P}_j} r_i \geq c_j \quad \forall j \in \mathcal{C}$

If a reaction is selected, all of its parent compound nodes (its reactants) must also be selected.

(3) $\sum_{i \in \mathcal{Y}} r_i \leq \text{length}(\mathcal{Y}) - 1 \quad \forall \mathcal{Y}$

For each cycle in the graph, every reaction node in the cycle cannot be simultaneously selected.

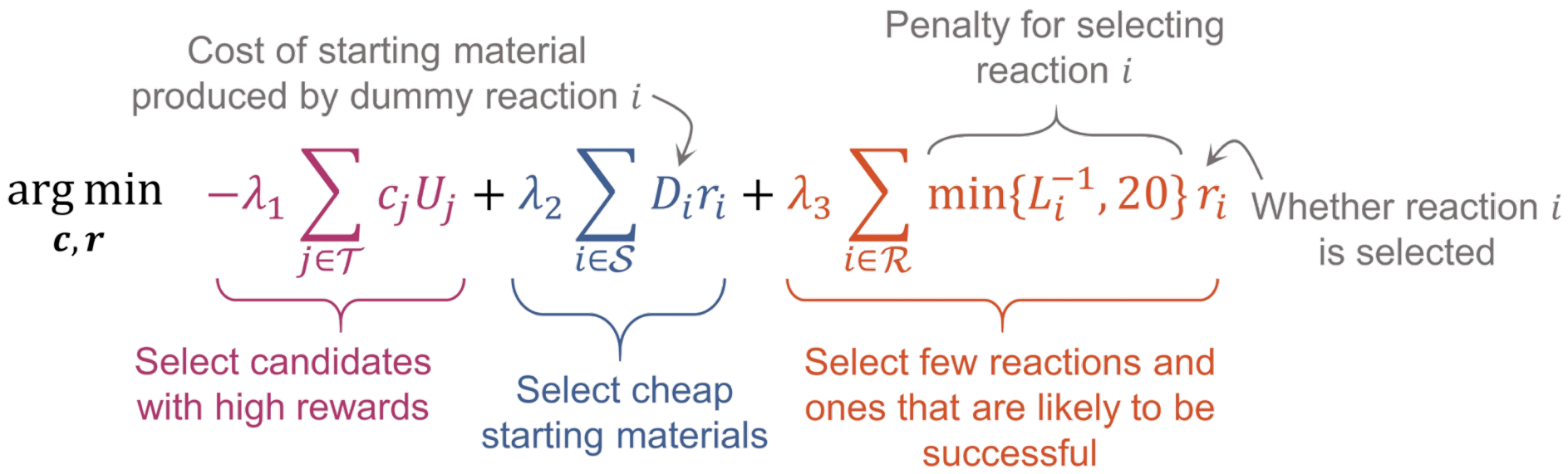
Additional notation

- j: An index referring to a reaction node
- i: An index referring to a compound node
- $\mathcal{R}$: Set of reaction node indices
- $\mathcal{C}$: Set of compound node indices
- c_j: Decision variable defining whether compound node j is selected
- r_i: Decision variable defining whether reaction node i is selected
- $\mathcal{S}$: Set of dummy reaction node indices
- $\mathcal{P}_{i \text{ or } j}$: Set of parent nodes for the node corresponding to index i or j
- $\mathcal{Y}$: A cycle in a retrosynthetic graph

Jenna Fromer; https://arxiv.org/abs/2311.02187

Synthesis-aware design of *batches* of molecules

- We do not just make (or buy) and test a single molecule at a time – we do so in *batches*
- We should really consider the utility <u>of the batch</u> against the cost <u>of the batch</u> (not traditional BO)

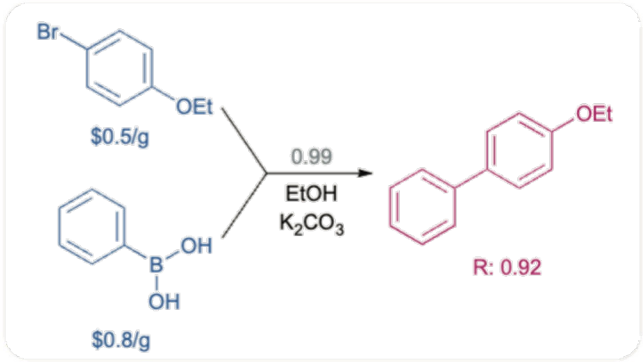


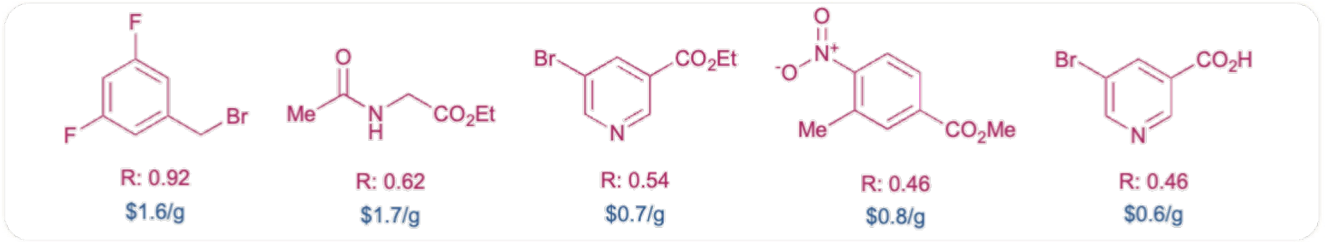
Rewards ≈ acquisition functions (a la Bayesian optimization), predicted properties, docking, etc.	Starting material (or screening compound) costs can come from vendor catalogs, e.g., Chemspace	Potential synthetic pathways and likelihoods of success for each reaction are from ML predictions

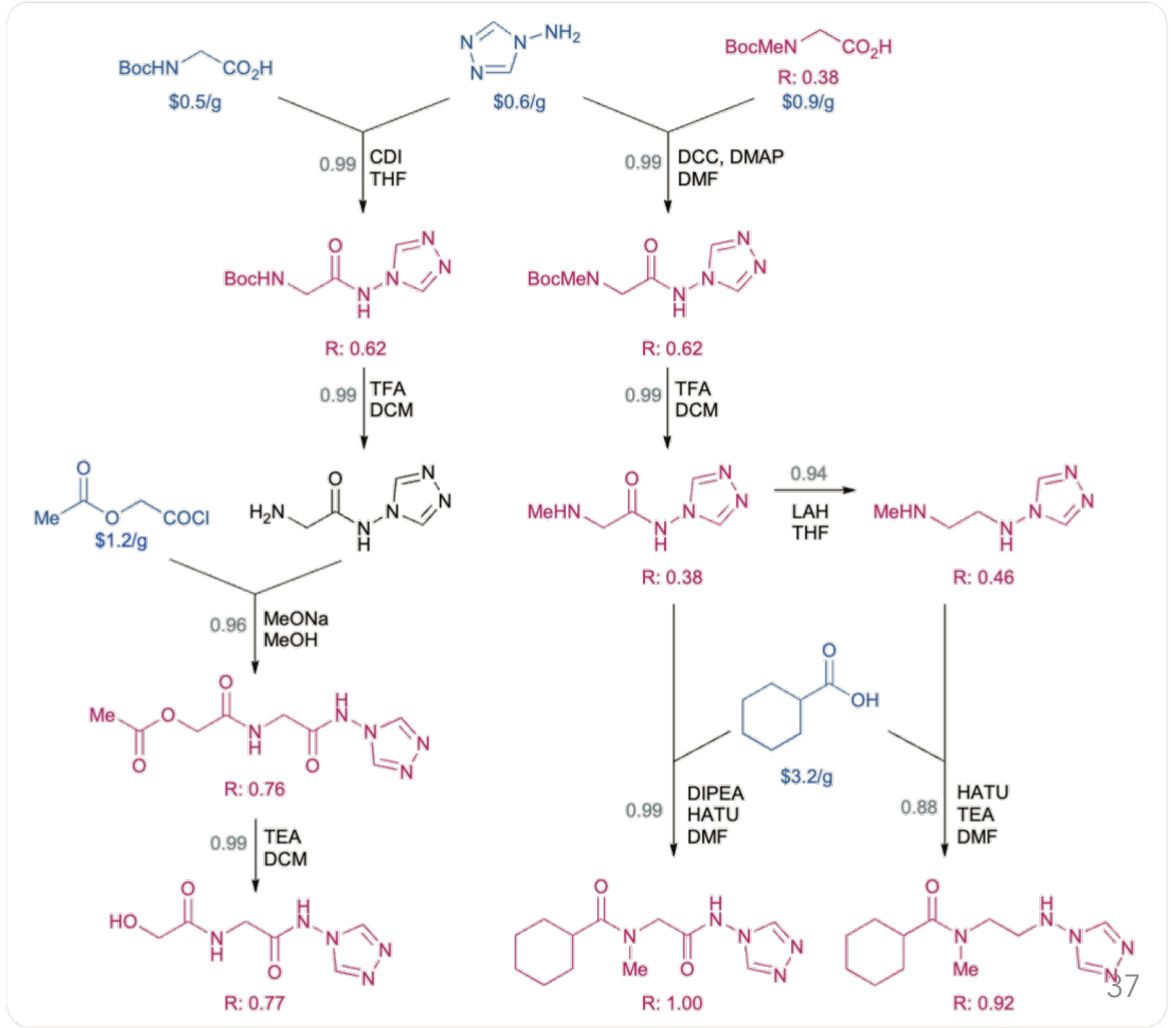
Synthesis-aware design of *batches* of molecules

- We do not just make (or buy) and test a single molecule at a time – we do so in *batches*
- We should really consider the utility <u>of the batch</u> against the cost <u>of the batch</u> (not traditional BO)
- E.g., re-analyzing one design cycle of 121 candidates from Koscher et al. *Science* 2023

Out of 121 candidates…
- purchase 6
- synthesize 9 using 6 building blocks & 10 reaction steps







The correlation between molecules in a batch matters

$$x^* = \arg\max_{x \in \mathcal{X}} \; f(x) \quad \xrightarrow{\textit{one option}} \quad \mathcal{X}_b^* = \arg\max_{\mathcal{X}_b \subset \mathcal{X}, |\mathcal{X}_b| = b} \; P(x^* \in \mathcal{X}_b)$$

- Typically, medicinal chemists will cluster candidates to design "diverse" batches
- My assertion: *the use of diversity/clustering for compound selection is just a proxy for trying to decorrelate the risk of failure – we do not want every molecule in the batch to underdeliver*

Decorrelating risk is especially important when workflows are non-iterative

Summary

best design

oracle

$$x^* = \arg\max_{x \in \mathcal{X}} f(x)$$

all possible molecules

surrogate oracle
(updated iteratively?)

$$x^* \approx \arg\max_{x \in \mathcal{X}} \hat{f}(x)$$

1. We lack good computational oracles; scoring is the bottleneck for discovery
2. Experimental oracles require synthesis, which constrains our design space
3. Representation learning and property prediction (+interactions) is not "solved"
4. Sequential molecular design does not reflect the reality of batched design

Group & funding

ccoley@mit.edu | coley.mit.edu

Graduate students & postdocs

Group Members
Alex Stoneman
Anji Zhang
Babak (Bo) Mahjour
Itai Levin
Jenna Fromer
Jihye Roh
Joonyoung Joung
Joules Provenzano
Keir Adams
Kento Abeywardane
Kevin Yu
Mingrou Xie
Mrunali Manjrekar
Nick Casetti
Priyanka Raghavan
Qianxiang Ai
Runzhong Wang
Tianyi Jin
Wenhao Gao
Xiaoqi Sun
Zhengkai Tu

Software dev.

Huiqian Lin
Mun Hong Fong
Sourabh Choure

UROPs

Alexandra, Ghazal, Jonathan, Ron, Vlad, Ne, Ray, Tony, Alec, Giselle, Montgomery

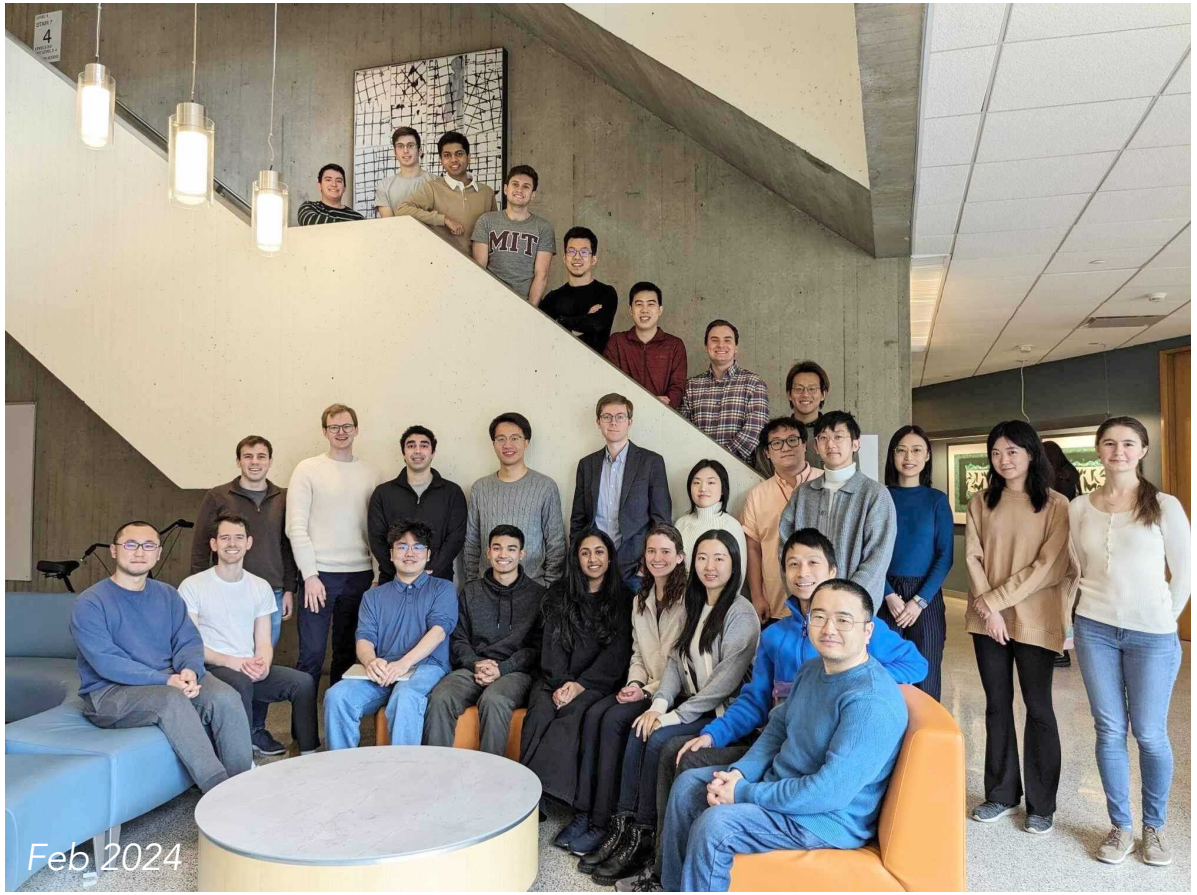
CAREER CHE-2144153
CCI CHE-2202693

1R21GM141616-01
1U18TR004149-01
1R21AI169342-01

N00014-21-1-2195

HR00111920025

(do not receive funding from these two)



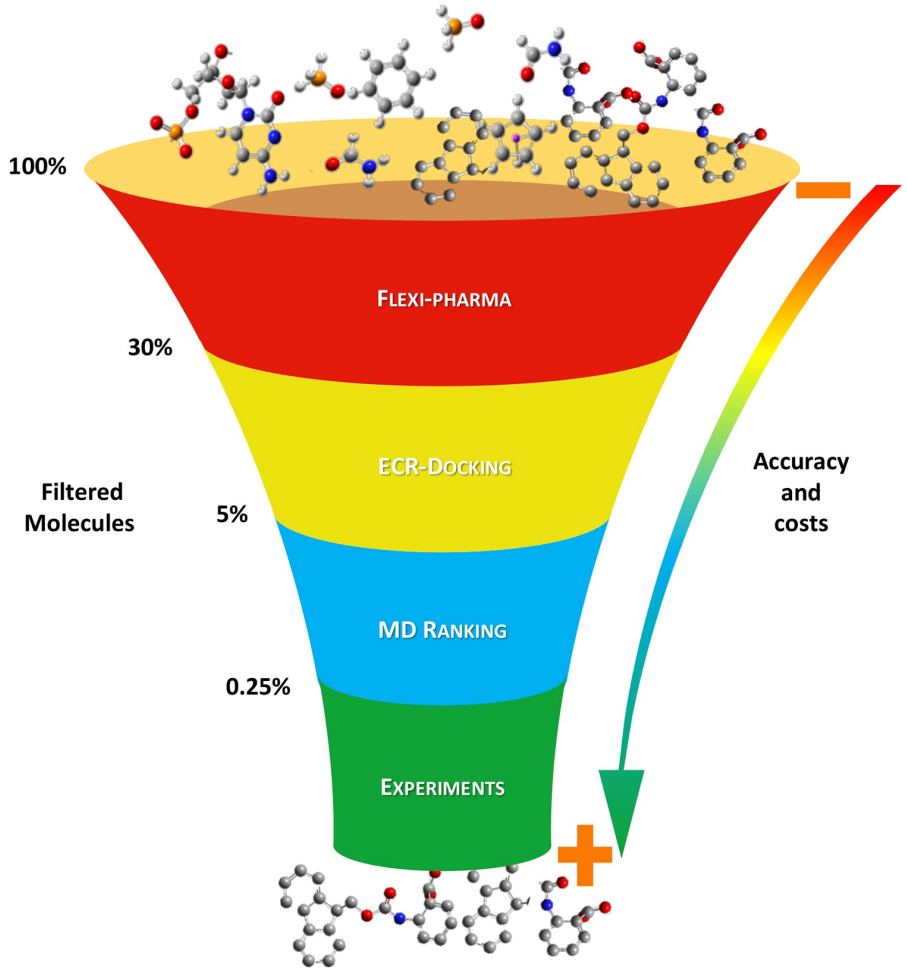
MLPDS
Machine Learning for Pharmaceutical Discovery and Synthesis

(members from over the years are shown; not all are current)

Data: CAS, Pistachio

Group Guide: see website

Computer-aided molecular discovery pipelines still involve extensive manual intervention and are highly bespoke



1. "Considering a range of properties … as well as their commercial availability, 17 compounds were chosen as virtual screening hits"

2. "… the choice of these compounds was based on factors such as drug-likeness, availability for procurement, ligand efficiency and chemical diversity"

3. "The top-scoring molecules for the top-ranked 4,000 clusters were inspected for unfavourable features … From the remaining top-ranking clusters, we synthesized 17 richly functionalized THPs"

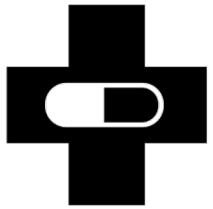
4. "all members were inspected … 40 molecules with ranks ranging from 16 to 246,721…were selected for de novo synthesis and testing."

[1] Lans, I. et al. *PLOS Computational Biology* **2020**, *16* (8), e1007898. https://doi.org/10.1371/journal.pcbi.1007898.
[2] Gorgulla, C et al. *Nature* **2020**, *580* (7805), 663–668. https://doi.org/10.1038/s41586-020-2117-z.
[3] Kaplan, A. L. et al. *Nature* **2022**, 1–10. https://doi.org/10.1038/s41586-022-05258-z.
[4] Stein, R. M. et al. *Nature* **2020**, *579* (7800), 609–614. https://doi.org/10.1038/s41586-020-2027-0.

Our primary research threads

AI for synthetic organic chemistry

Machine learning models that learn what chemical transformations are possible



AI for medicinal chemistry

Computer-aided design or selection of molecular structures, considering synthesis

AI for analytical chemistry

Spectral prediction and structure elucidation through mass spectrometry

Foundational capabilities

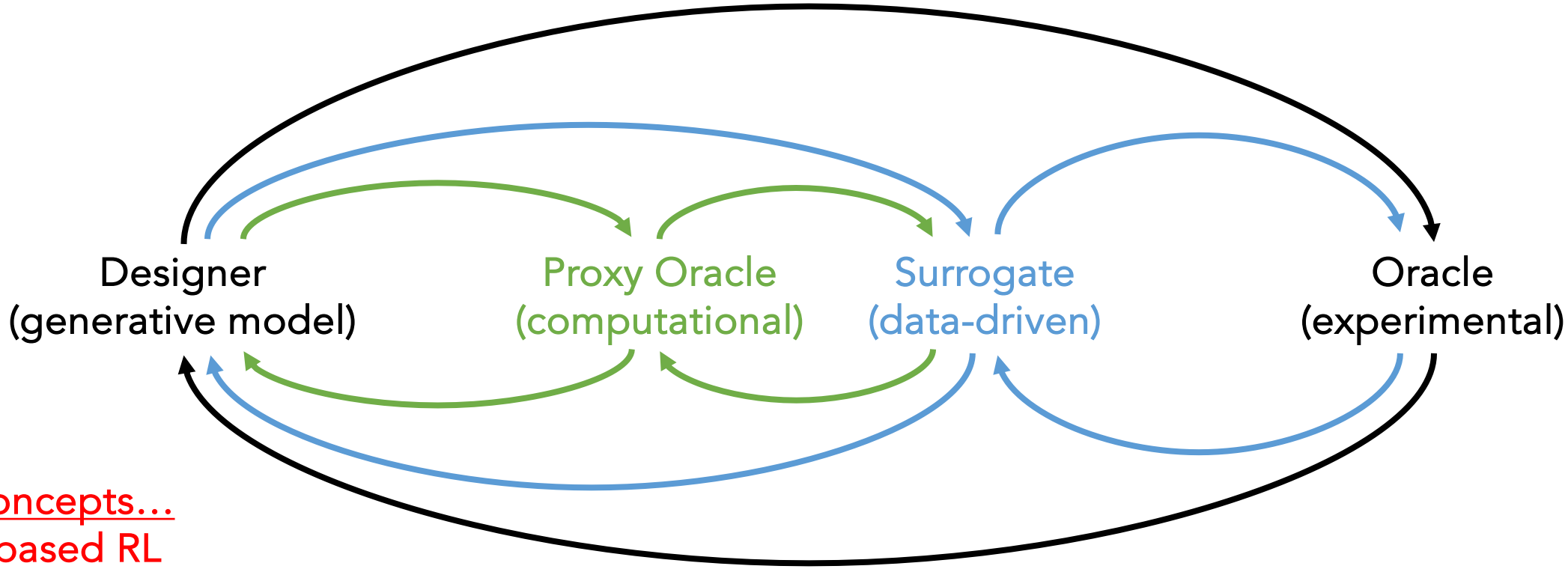
Chemistry-tailored models for molecular representation learning

Data sharing to facilitate modeling for chemistry and drug discovery

Autonomous chemistry laboratories for molecular and reaction discovery

Aligning AI for molecular design with the real world

- What should computer-aided molecular design workflows look like? What is the best role for generative modeling – hit finding or optimization? Does it even *need* to be sample efficient?



More concepts…
Model-based RL
Multi-fidelity learning
Low data surrogate models