6.S085 Machine Learning for Molecular Design

- We will start at 10:05 am
- Please fill the ML/science prior knowledge survey 👉🏻

Welcome to the class

Website: https://moldesign.github.io/

Wenhao Gao

Wenhao is a PhD candidate in the Coley Research Group at MIT. His research focuses on accelerating and scaling up the process of molecular discovery by leveraging the capabilities of AI for decision-making. He is the recipient of a Google Ph.D. fellowship and an MIT-Takeda fellowship. Wenhao also serves as one of the organizers of multiple AI for Science workshops at NeurIPS, ICML, and Machine Learning and AI for Organic Chemistry Symposium at ACS.

Ron Shprints

Ron is a third year undergraduate student at MIT, majoring in mathematics and computer science. His research interest lies in deep learning and its applications to the natural sciences. He has been doing research in the intersection of machine learning and molecular discovery since his freshman year. Ron joined the Coley Research Group as a UROP student in summer 2022 and has collaborated with Wenhao on several projects. Before that he worked at the Jensen Research Group where he collaborated with Andrew Zahrt on the machine learning discovery of electrochemical reactions.

Intended learning goals

By the end of this course, we hope you will be able to:

- comprehend and analyze applications of machine learning in molecular science from the primary literature.
- be familiar with common molecular design strategies and identify suitable strategies as well as machine learning algorithms for common real-world scenarios.
- implement machine learning algorithms for molecular design, using packages like scikit-learn, and common deep learning frameworks like PyTorch.

What we expect you do to

The final grade (P/F) for this course will be determined by:

- literature presentation (30%)
- in-class practice
- **a course project (70%)**

Literature presentation

- Understanding and analyzing research papers is crucial for learning about the latest advancements in research. We have compiled a selection of literature focusing on molecular design and the application of machine learning techniques in this area.
- Students will be assigned to groups and bid papers that they want to read and present.
- Presentation day: Tue (1/16/24) and Wed (1/24/24), 8 min with additional 2 min for Q&A.
- You may present a paper that is not included in the provided list, upon the approval of the teaching staff. Please submit your chosen paper along with the names of your team members for approval to the class email.

In-Class Practice

- Use Google Colab to help practice the concepts covered during each lecture.
- Exercises are intended to be completed within the class duration.
- These practices don’t contribute to the final grade. However, we will provide a structured exercise to guide you through a typical solution of the course project. So, we highly recommend to fully complete them.

Course Project

- A pharmaceutical-related small organic molecule design campaign (semi-real-world scenario).
- We will provide a dataset of ~1400 activity labeled molecules. Your task is to design novel molecules that have higher activity.
- Team up to groups of 3-5, sign up by this Friday.
- Each team has a budget of submitting/labeling 100 molecules each week,
- We will hold a live leaderboard showing top-10 teams on our webpage. After evaluating the submission, we will update the leaderboard.
- Top-3 team will be invited to present their workflow in the last day, and prize.

About pre-reqs

- We don’t have hard prerequisites, except comfortable in coding in Python.
- We provide some useful resources on the course website.

Office hour

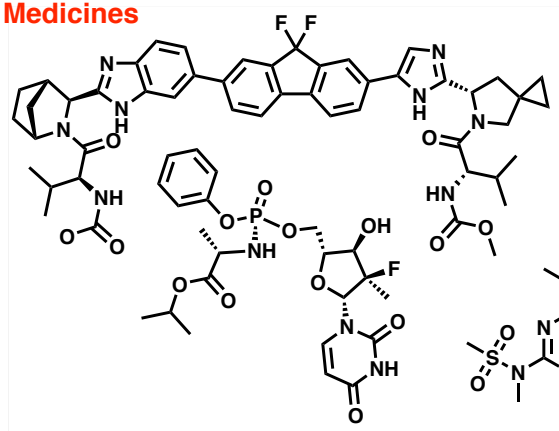
- Mon and Fri, 1-2 pm @ 26-168
- Questions: moleculedesigner@gmail.com

Schedule and content

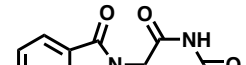
Date	Day of the Week	Topic	Link to Colab
1/8/2024	Monday	Course overview + Broad review of historical development and common workflows	[colab notebook] coming soon!
1/10/2024	Wednesday	Data process and analysis: focus on dimensionality reduction and clustering	[colab notebook] coming soon!
1/12/2024	Friday	Structure-property relationship modeling (Part 1): featurization of molecules	[colab notebook] coming soon!
1/16/2024	Tuesday	Literature presentation (session 1)	[slides] coming soon!
1/17/2024	Wednesday	Structure-property relationship modeling (Part 2): deep learning architectures	[colab notebook] coming soon!
1/19/2024	Friday	Molecular generation and design (Part 1): screening and generative AI	[colab notebook] coming soon!
1/22/2024	Monday	Molecular generation and design (Part 2): optimization approach	[colab notebook] coming soon!
1/24/2024	Wednesday	Literature presentation (session 2)	[colab notebook] coming soon!
1/26/2024	Friday	Guest Lecture 1	[slides] coming soon!
1/29/2024	Monday	Guest Lecture 2	[slides] coming soon!
1/31/2024	Wednesday	Guest Lecture 3	[slides] coming soon!
2/2/2024	Friday	Final project presentations	[slides] coming soon!

Molecules

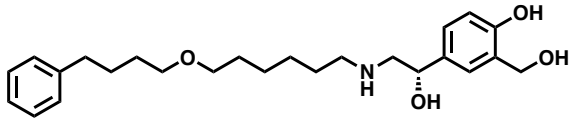
Medicines



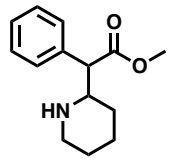
Harvoni
Ledipasvir/sofosbuvir (Hep C)



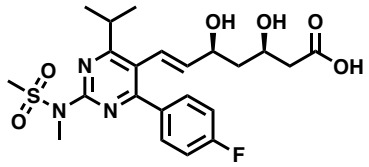
Revlimid
Lenalidomide (multiple myeloma)



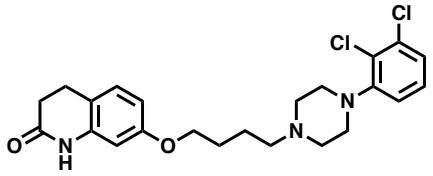
Serevent
Salmeterol (COPD)



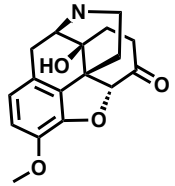
Ritaline
Methylphenidate (ADHD)



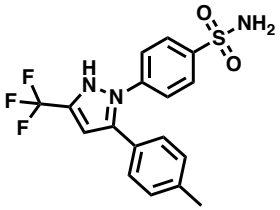
Cretor
Rosuvastatin (cholesterol)



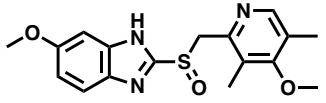
Abilify
Aripiprazole (antipsychotic)



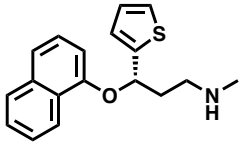
Percocet
Oxycodone (pain)



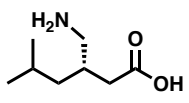
Celebrex
Celecoxib (NSAID)



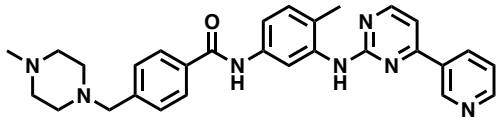
Nexium
Esomeprazole (dyspepsia)



Cymbalta
Duloxetine (SNRI)

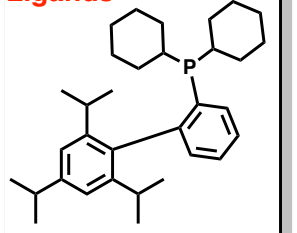


Lyrica
Pregabalin (epilepsy)



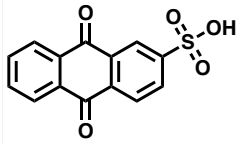
Gleevec
Imatinib (leukemia)

Ligands



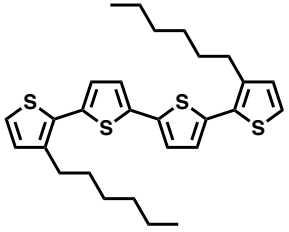
XPhos

Electrolytes



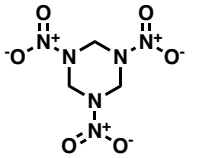
AMS

Organic electronics



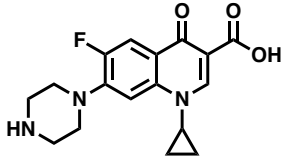
DH-4T

Energetics

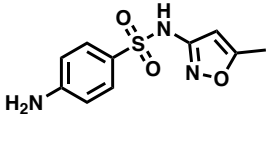


RDX

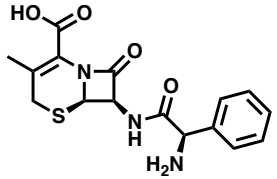
Antibiotics



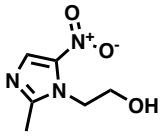
Ciprofloxacin



Sulfamethoxazole

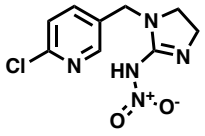


Cefalexin

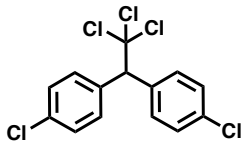


Metronidazole

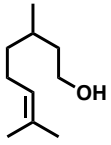
Pesticides



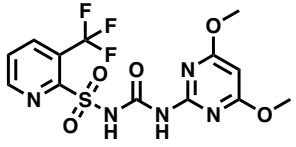
Imidacloprid



DDT



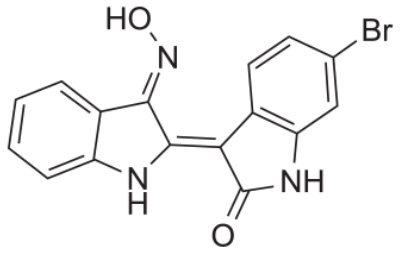
Citronellol



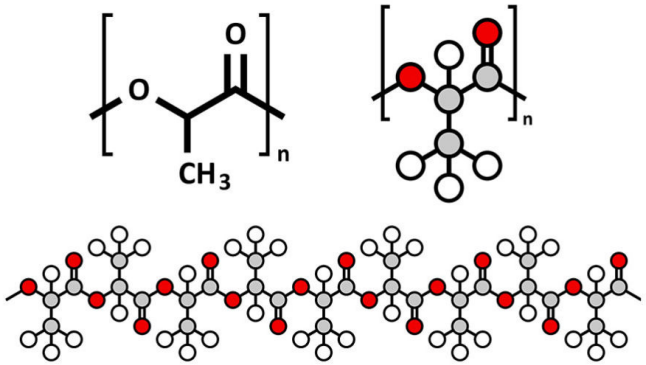
Flazasulfuron

Figure from Coley

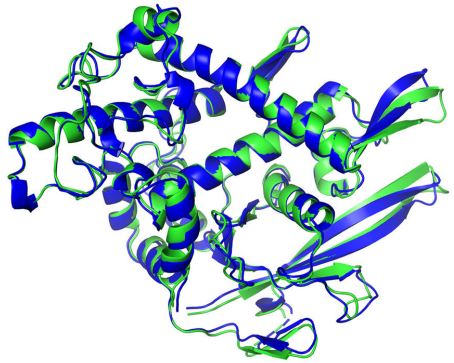
Larger and periodic



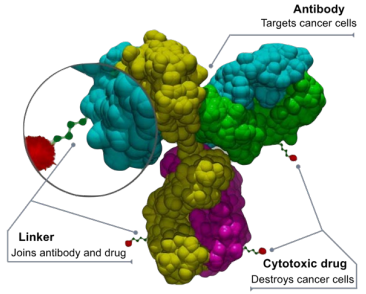
Small molecules



Polymers

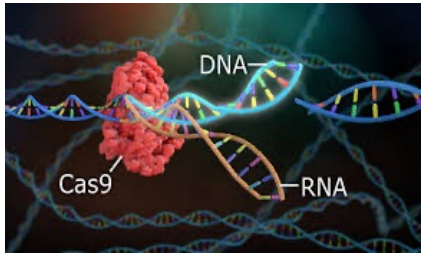


Proteins

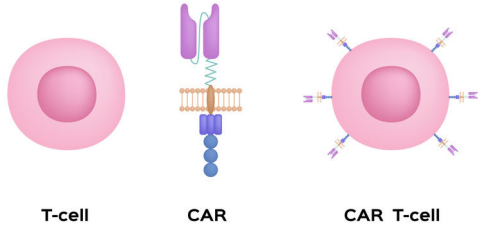


Anti- Nano-bodies

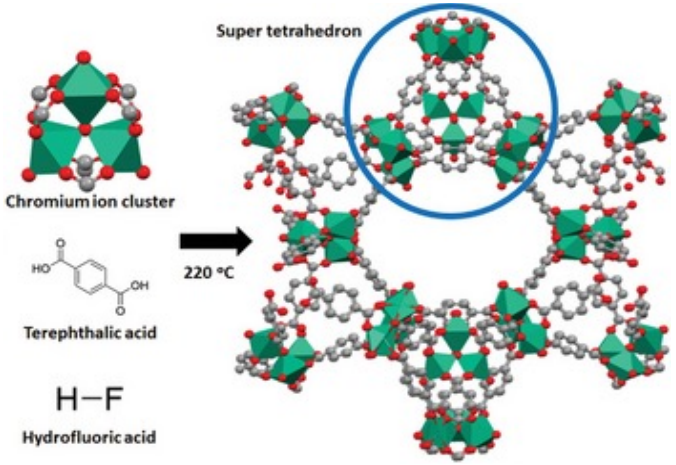
Vaccines



Gene-editing

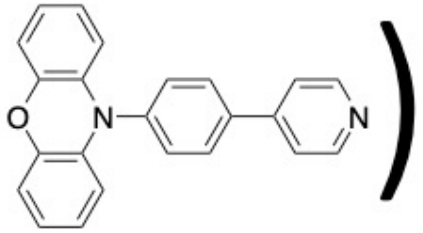


CAR-T

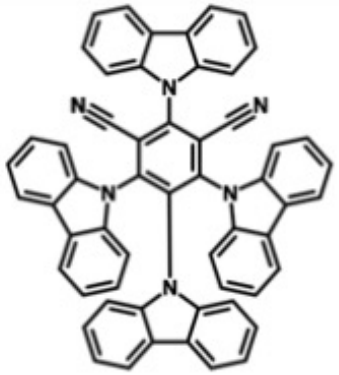


Crystal structure

Design problem: how can we find the molecular structure

$f($  $) =$ 

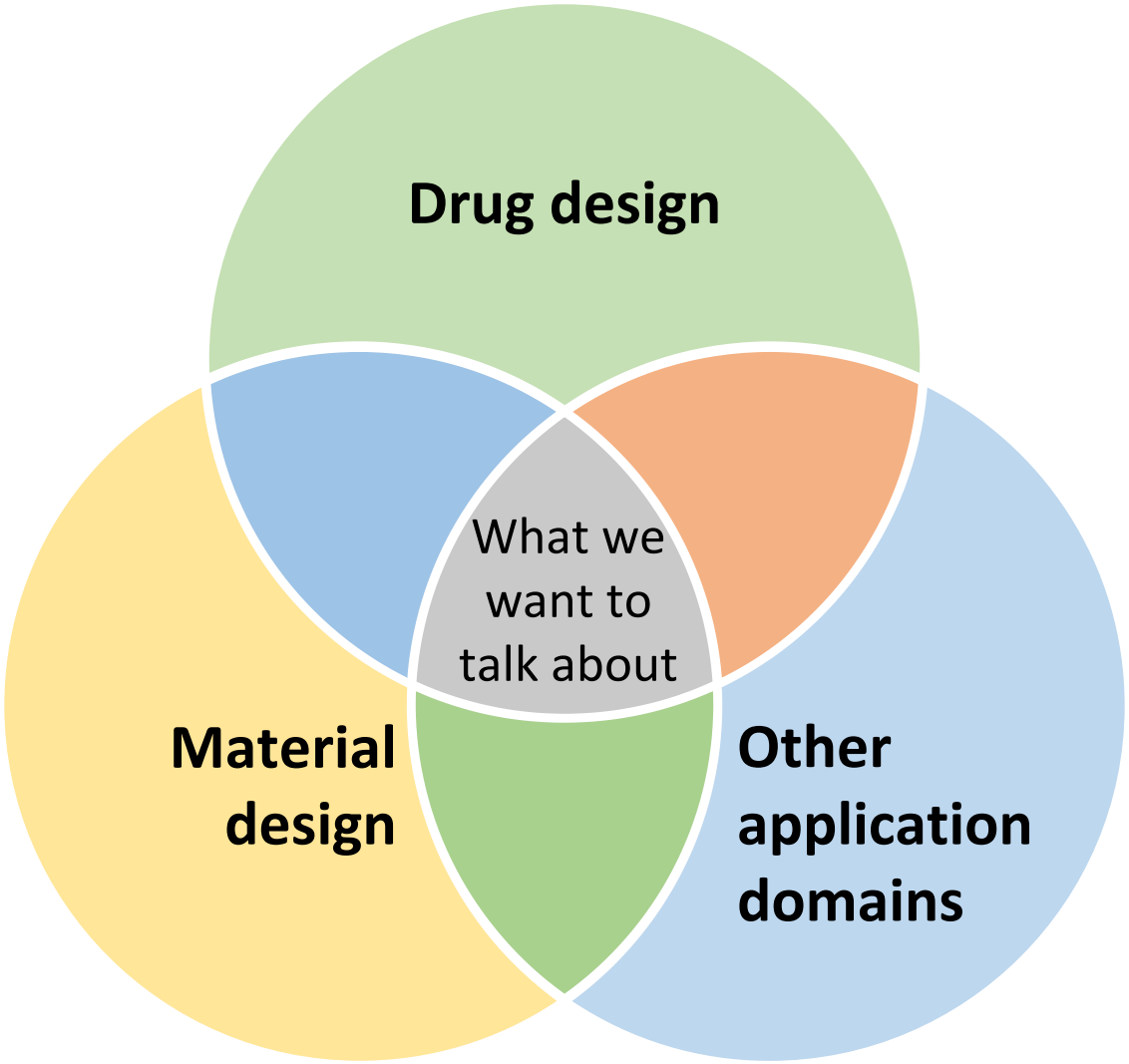
Macroscopic properties

$f^{-1}($  $) =$ 

Molecular structure

Figure from Rafa

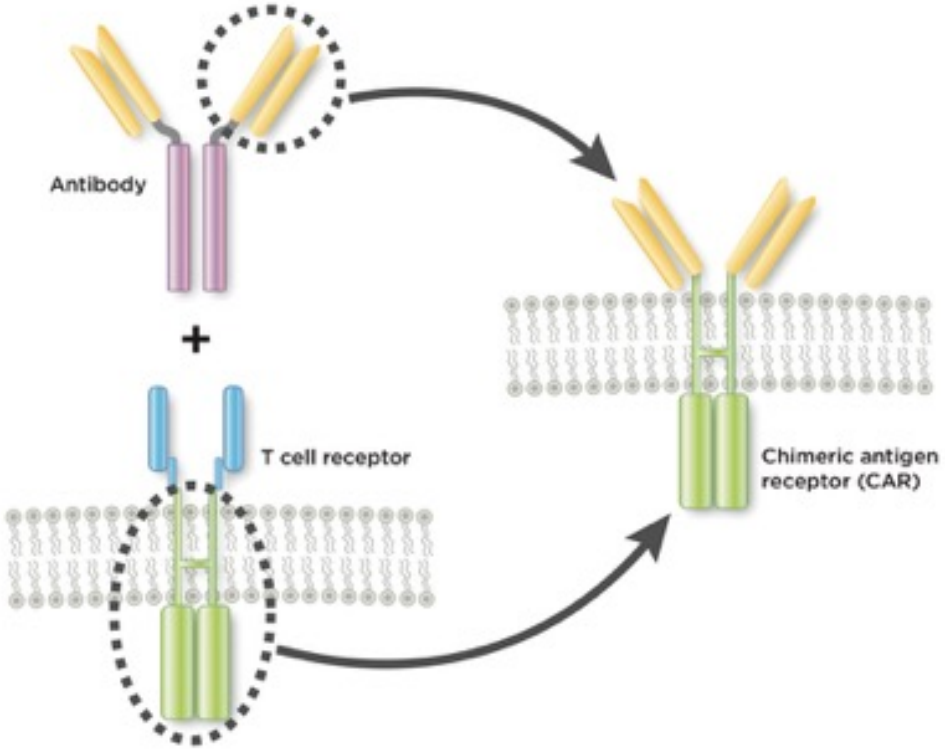
Common ground



$$m^* = \arg\max_{m \in \mathcal{M}} \mathcal{O}(m),$$

How can we design molecules?

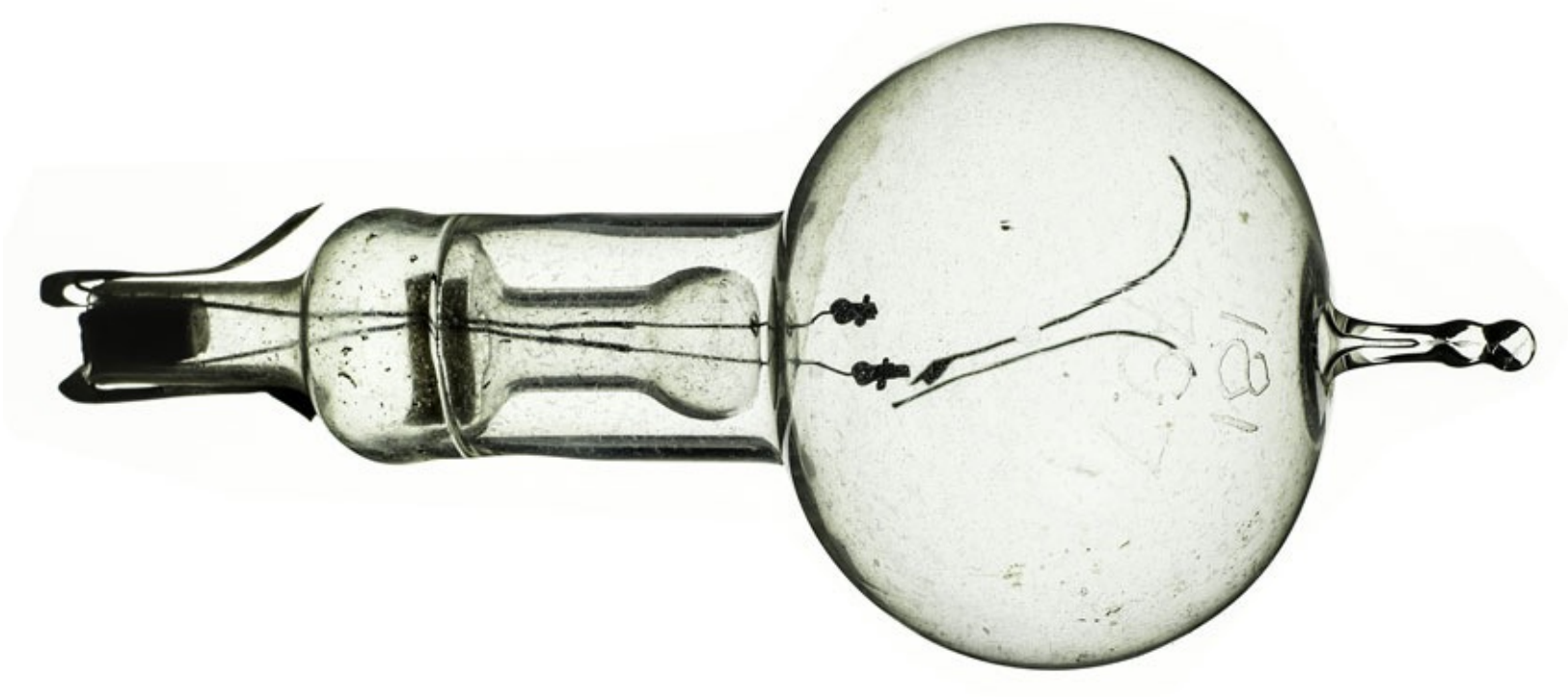
- Rational design



https://www.mskcc.org/timeline/car-t-timeline-progress

How can we design molecules?

- Screening/Trial and error (Edisonian approach)



https://www.bulbs.com/learning/history.aspx#:~:text=A%20Brief%20History%20of%20the,first%20commercially%20practical%20incandescent%20light.

Chemical space is too large

- The number of potential pharmacologically active molecules:

  ~10^60

- The number of possible sequence of a protein of 200 amino acids:

  ~20^200

- Largest chemical library you can potentially purchase:

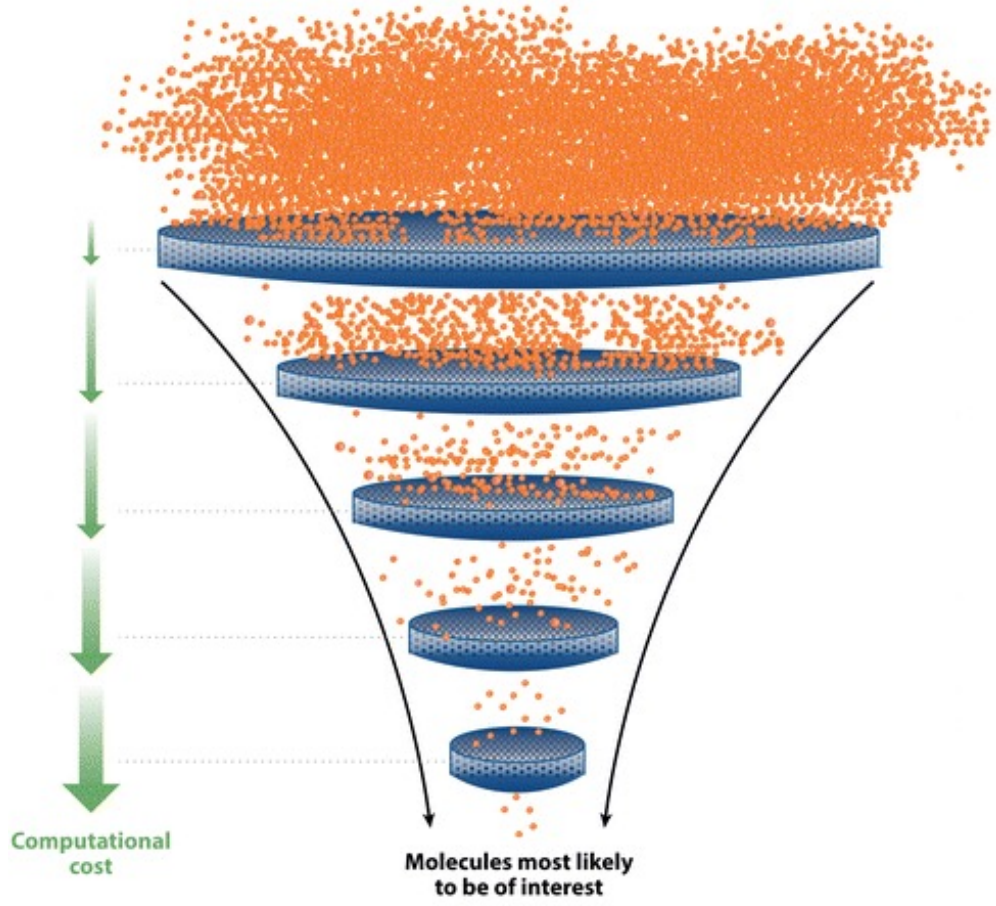
  38B

- Cost of running one plate of bioassay, evaluating ~50 make-on-demand molecules:

  In average $12k

High-Throughput Virtual screening

- When you can calculate the property, you can first computationally screen all candidates.
- Multi-fidelity:
  - fast and coarse methods
  - accurate but slow methods

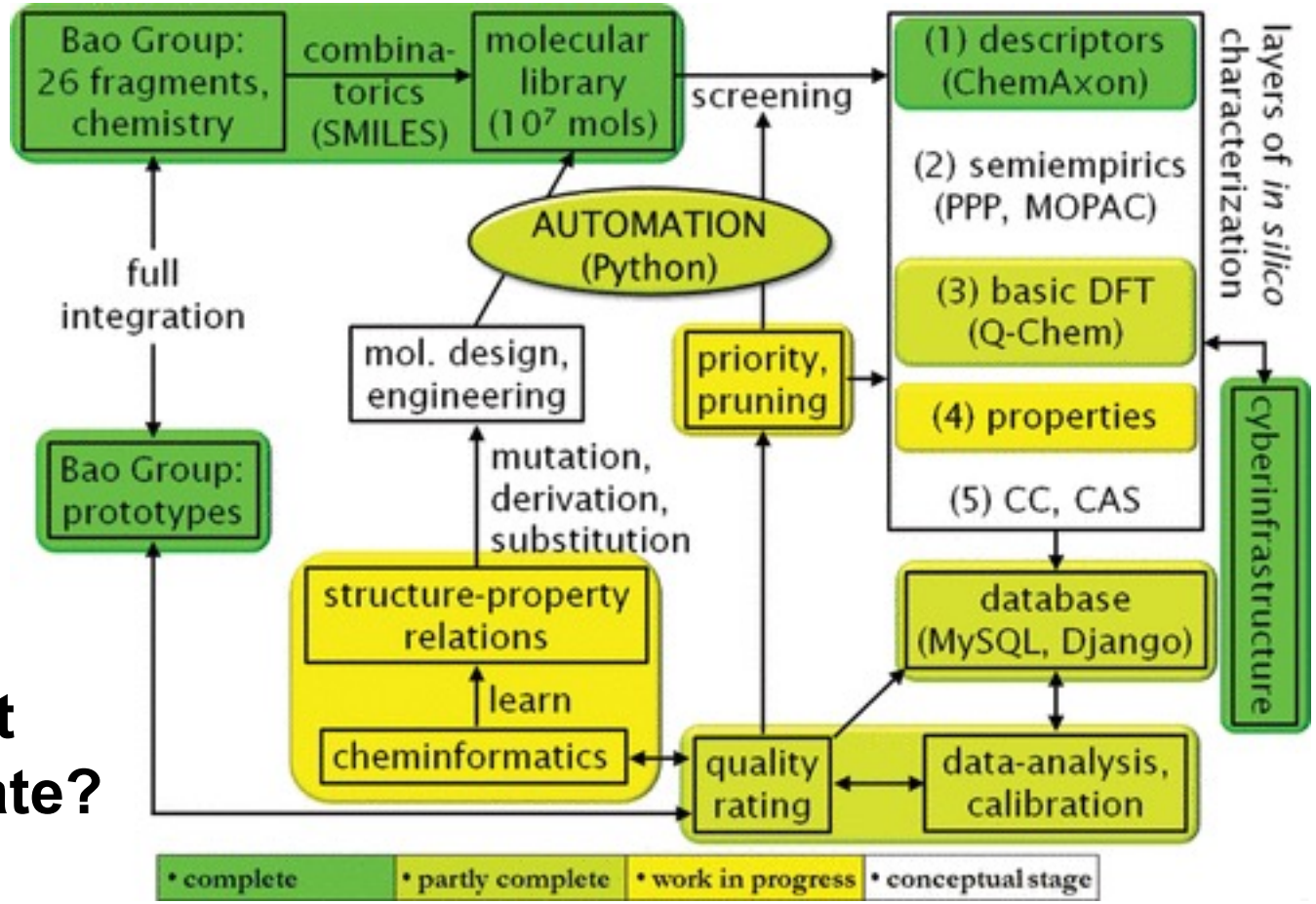


Pyzer-Knapp EO, et al. 2015.
Annu. Rev. Mater. Res. 45:195–216

Pyzer-Knapp, Edward O., et al. "What is high-throughput virtual screening? A perspective from organic materials discovery." *Annual Review of Materials Research* 45 (2015): 195-216.

Harvard Clean Energy Project (2011)

- Organic Photovoltaics (OPVs) discovery.

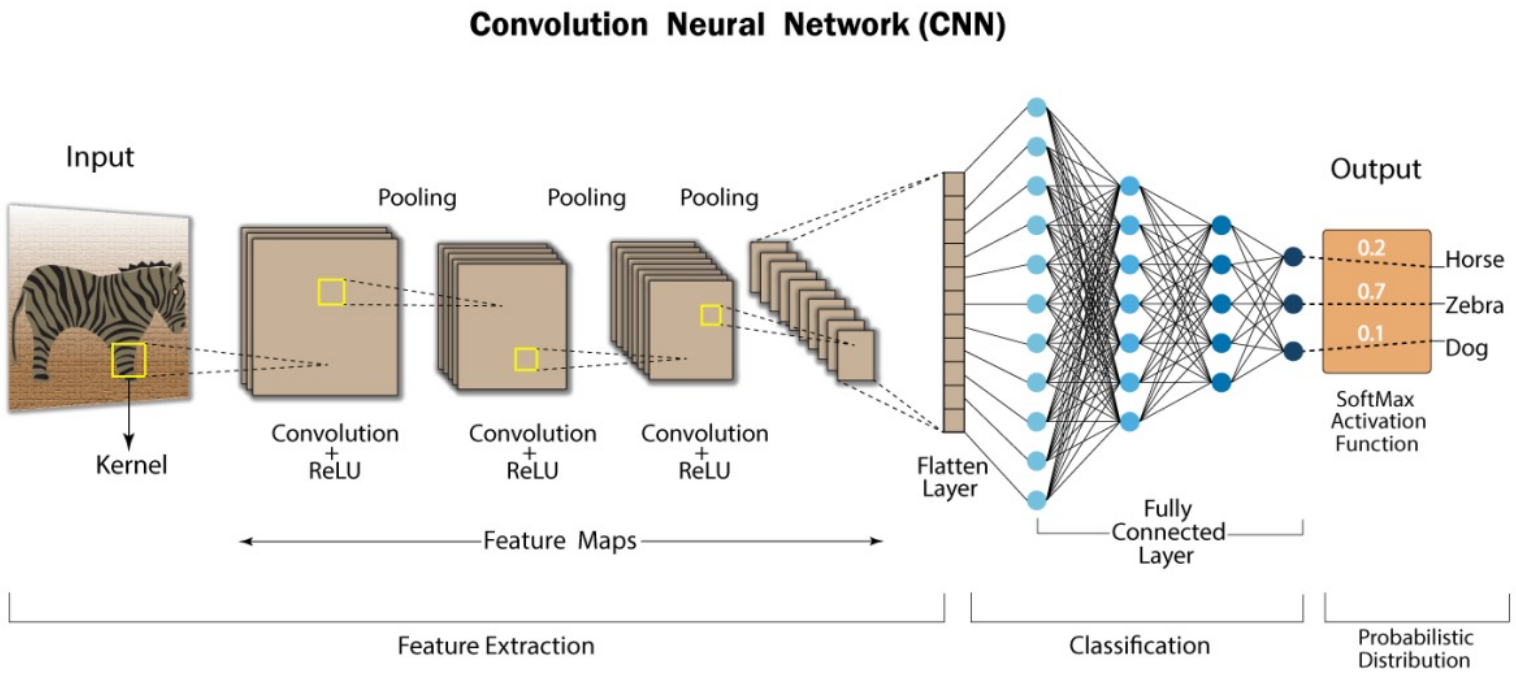


What if we cannot accurately simulate?

Hachmann, Johannes, et al. "The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid." *The Journal of Physical Chemistry Letters* 2.17 (2011): 2241-2251.

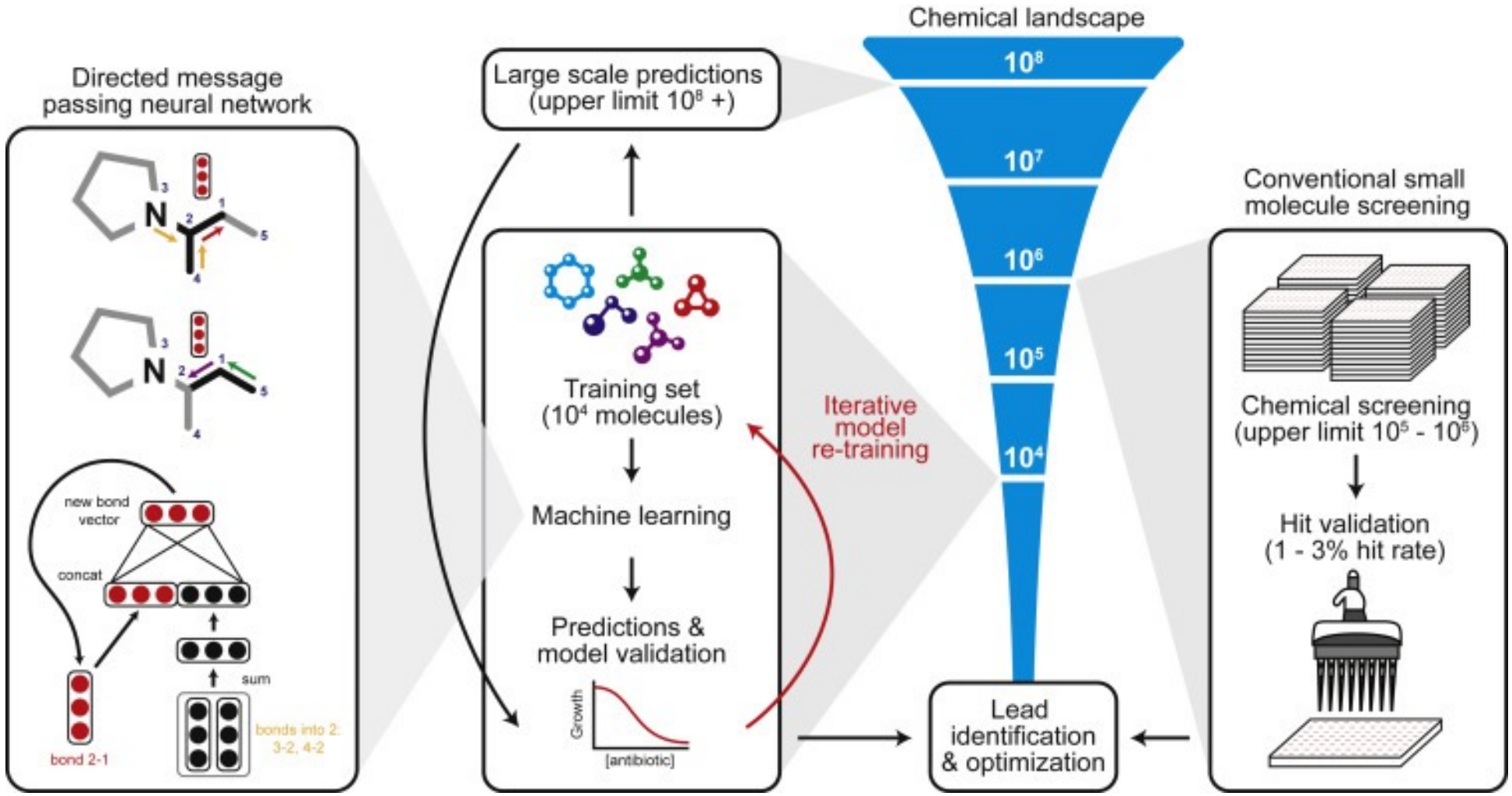
Machine Learning

- Wiki: Machine learning (ML) is an umbrella term for solving problems for which development of algorithms by human programmers would be cost-prohibitive, and instead the problems are solved by helping machines 'discover' their 'own' algorithms.



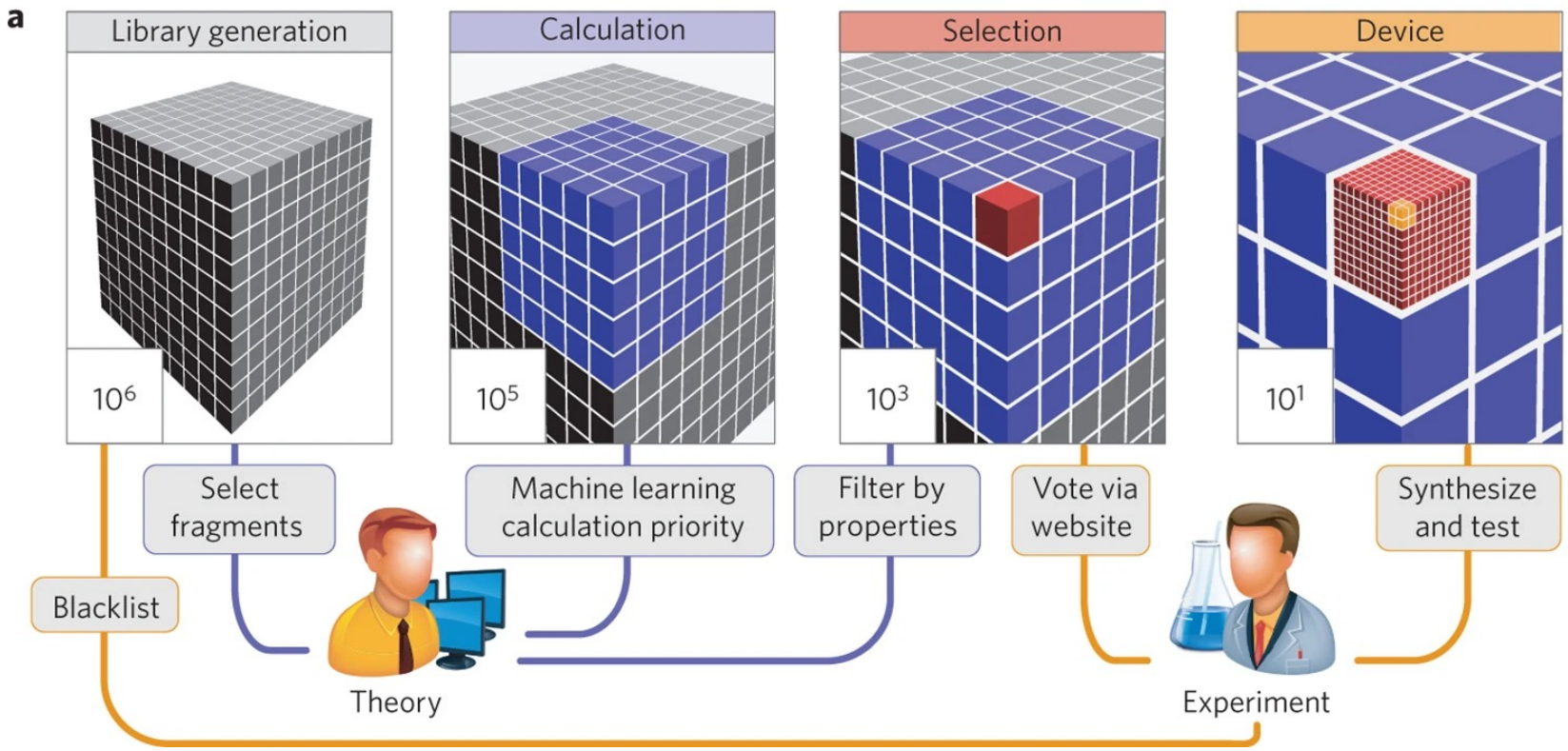
https://developersbreach.com/convolution-neural-network-deep-learning/

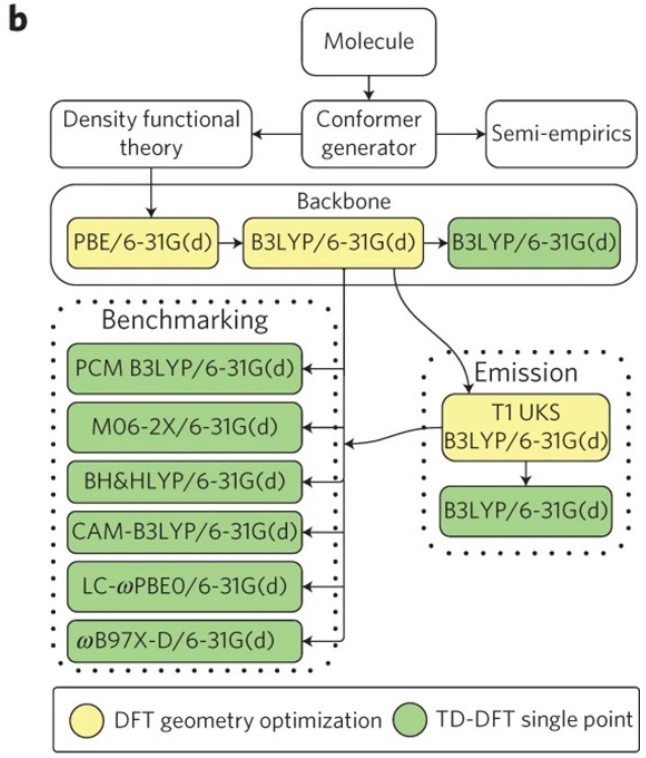
Machine learning enabled virtual screening



Stokes, Jonathan M., et al. "A deep learning approach to antibiotic discovery." *Cell* 180.4 (2020): 688-702.

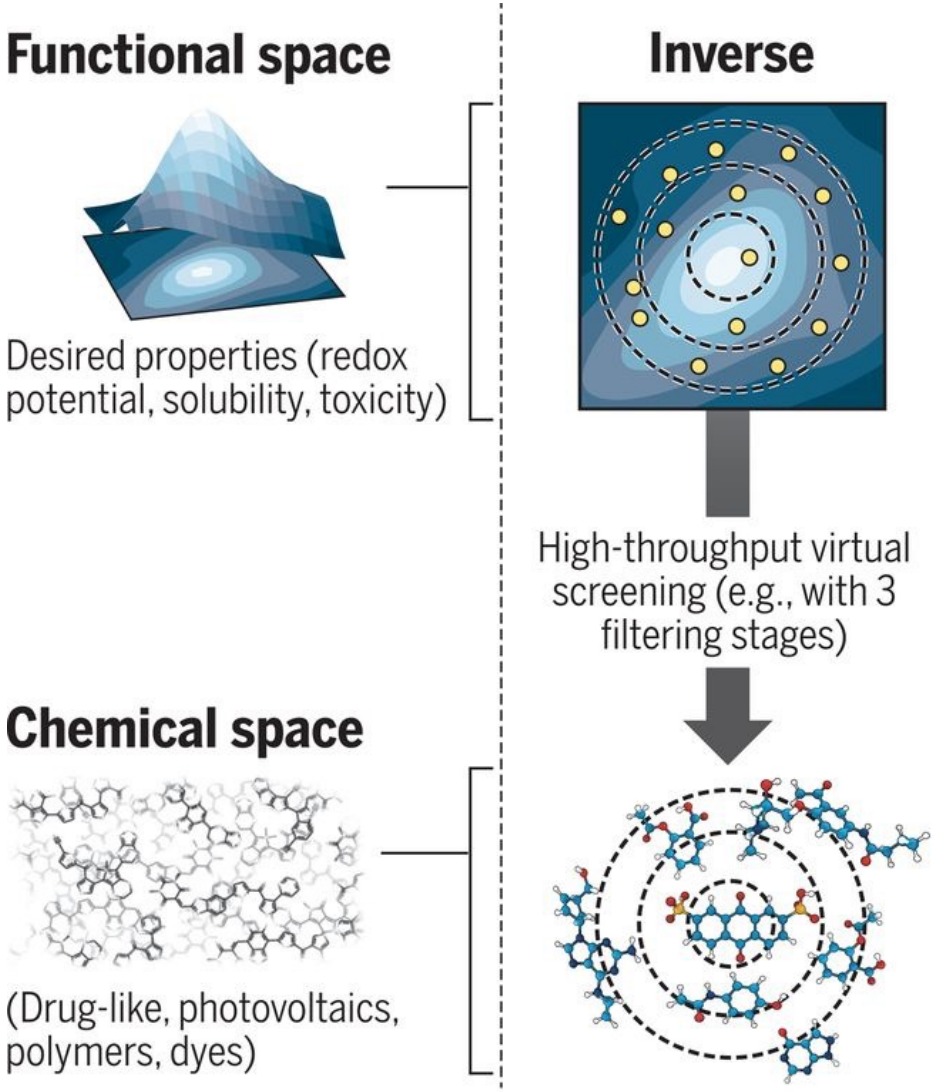
ML also serves as faster surrogates



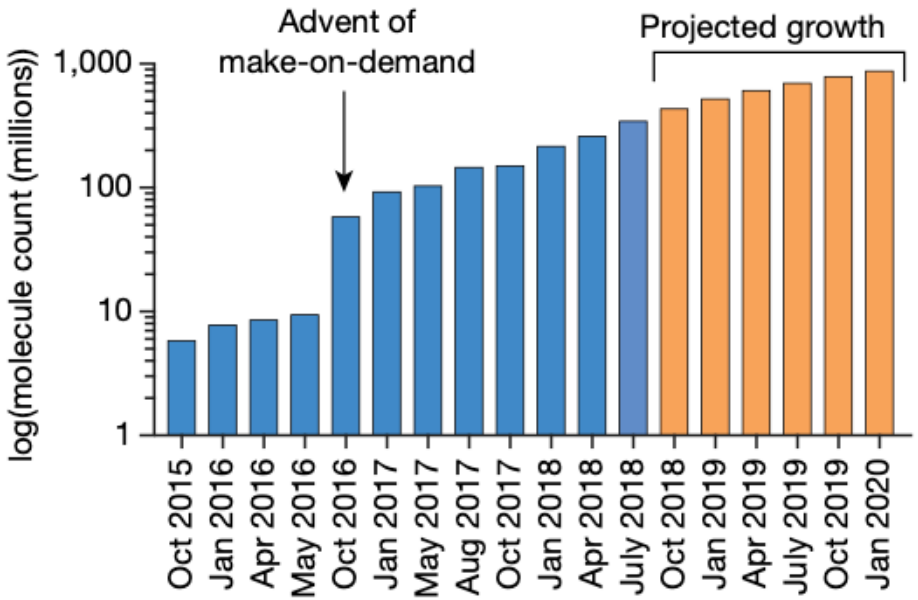


Gómez-Bombarelli, Rafael, et al. "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach." *Nature materials*15.10 (2016): 1120-1127.

Molecular design: screening and *de novo* design



The drug-like chemical: 10^{60}

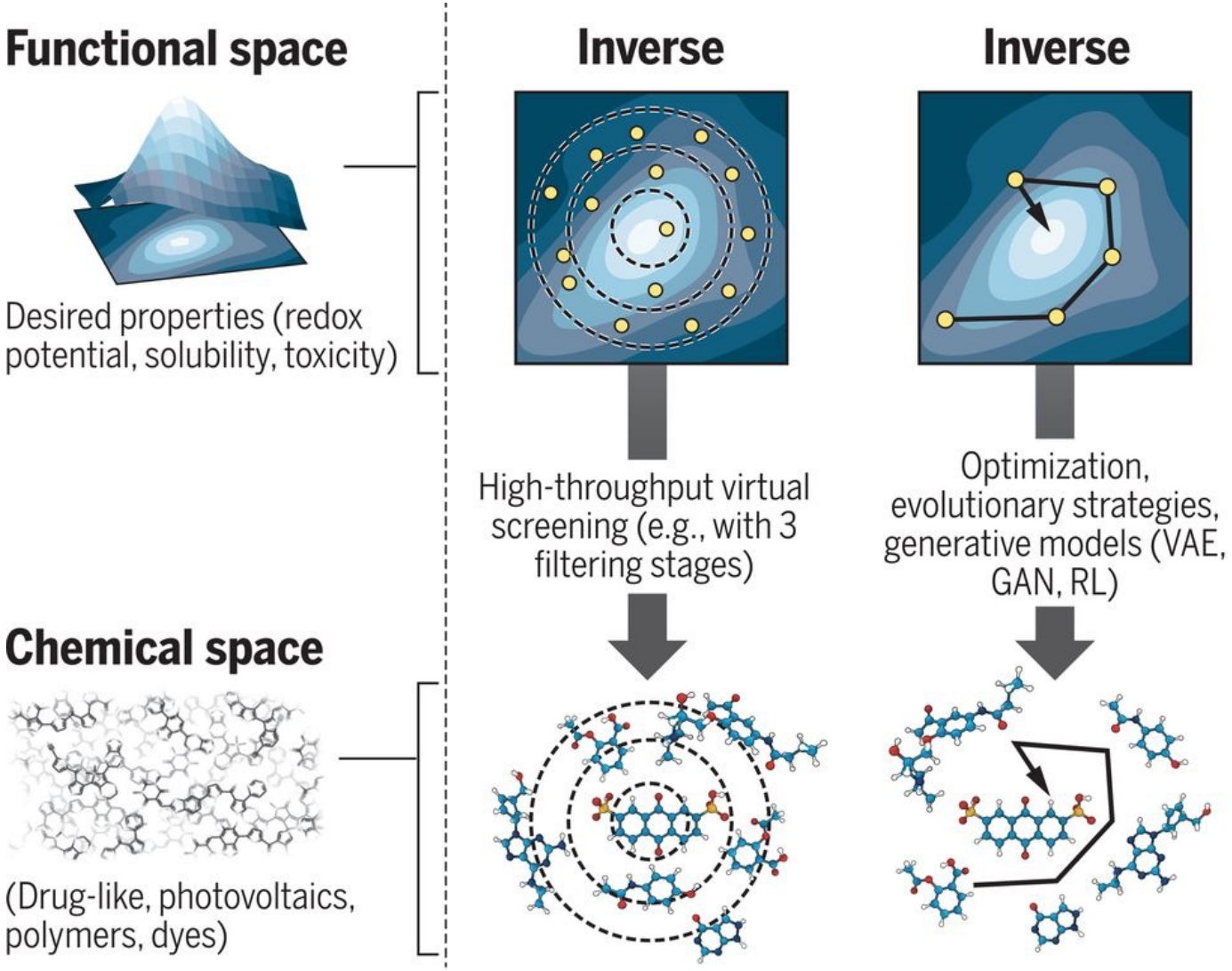


Challenges:

- Limited and finite chemical space
- Time and resources consuming

Modified from Sanchez-Lengeling, B., & Aspuru-Guzik, A. (2018). *Science*, *361*(6400), 360-365.
Lyu, Jiankun, et al. *Nature* 566.7743 (2019): 224-229.

Molecular design: screening and *de novo* design

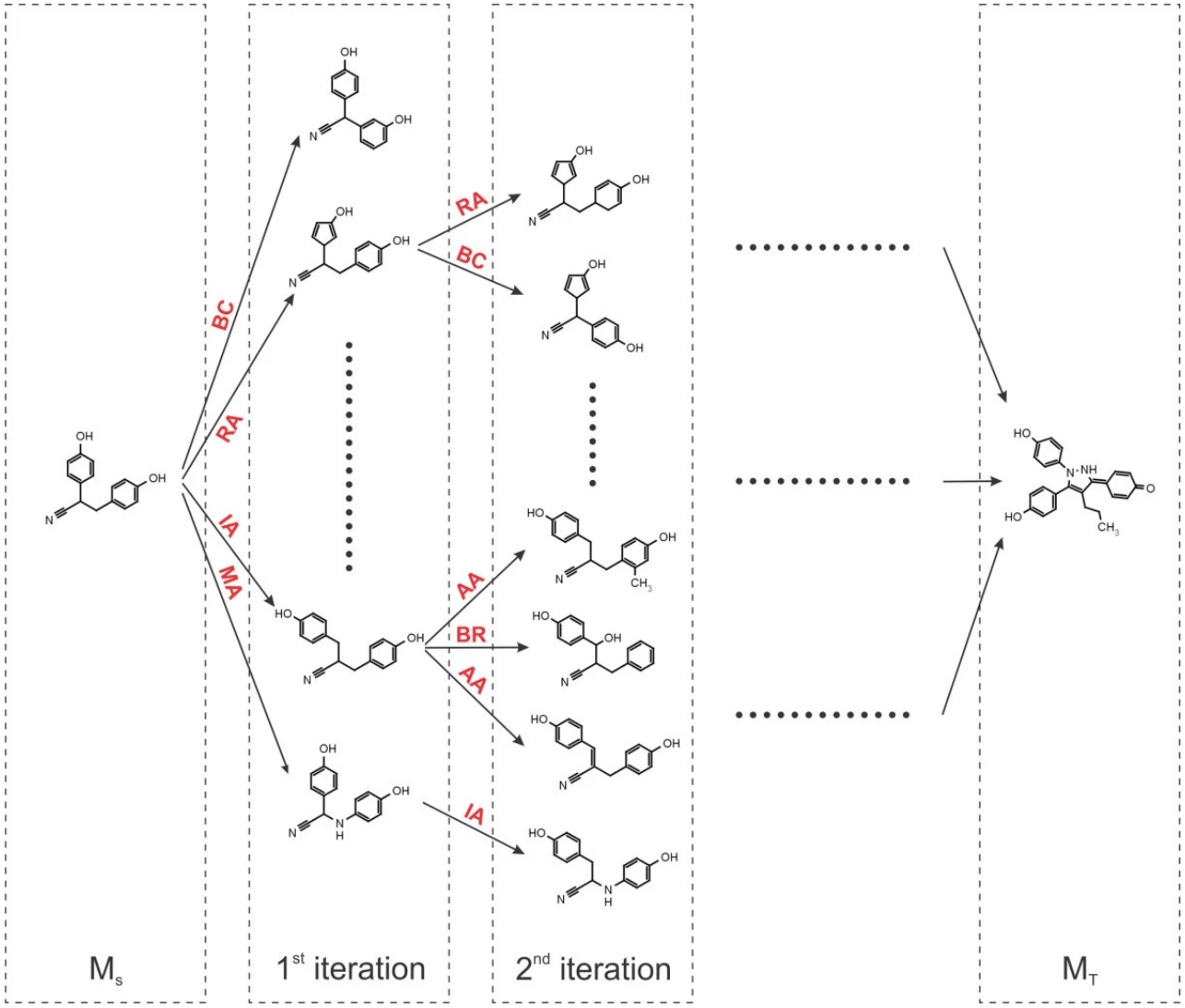


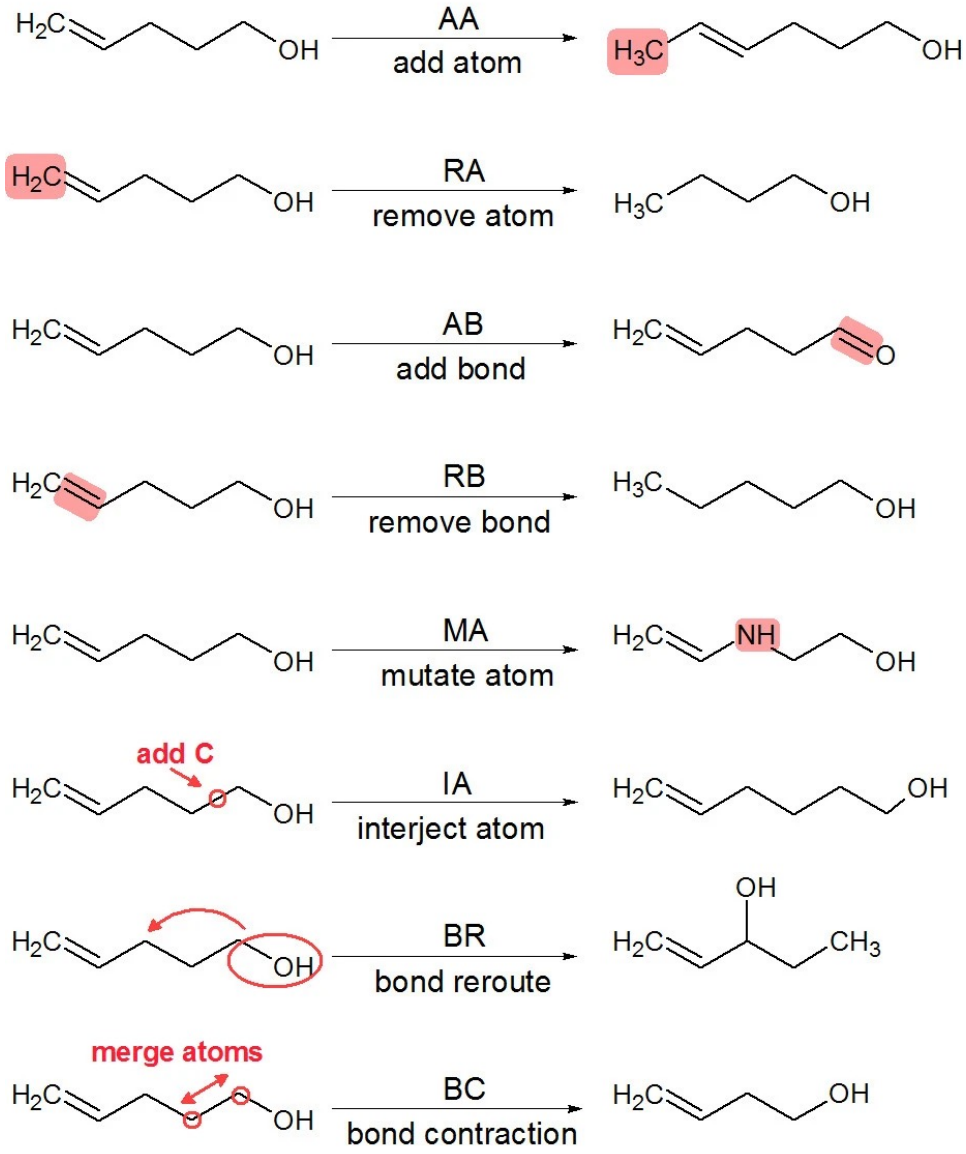
Advantages:

- Implicitly define a chemical space
- Potentially more efficient

Modified from Sanchez-Lengeling, B., & Aspuru-Guzik, A. (2018). *Science*, *361*(6400), 360-365.
Kutchukian, Peter S., and Eugene I. Shakhnovich. *Expert opinion on drug discovery* 5.8 (2010): 789-812.

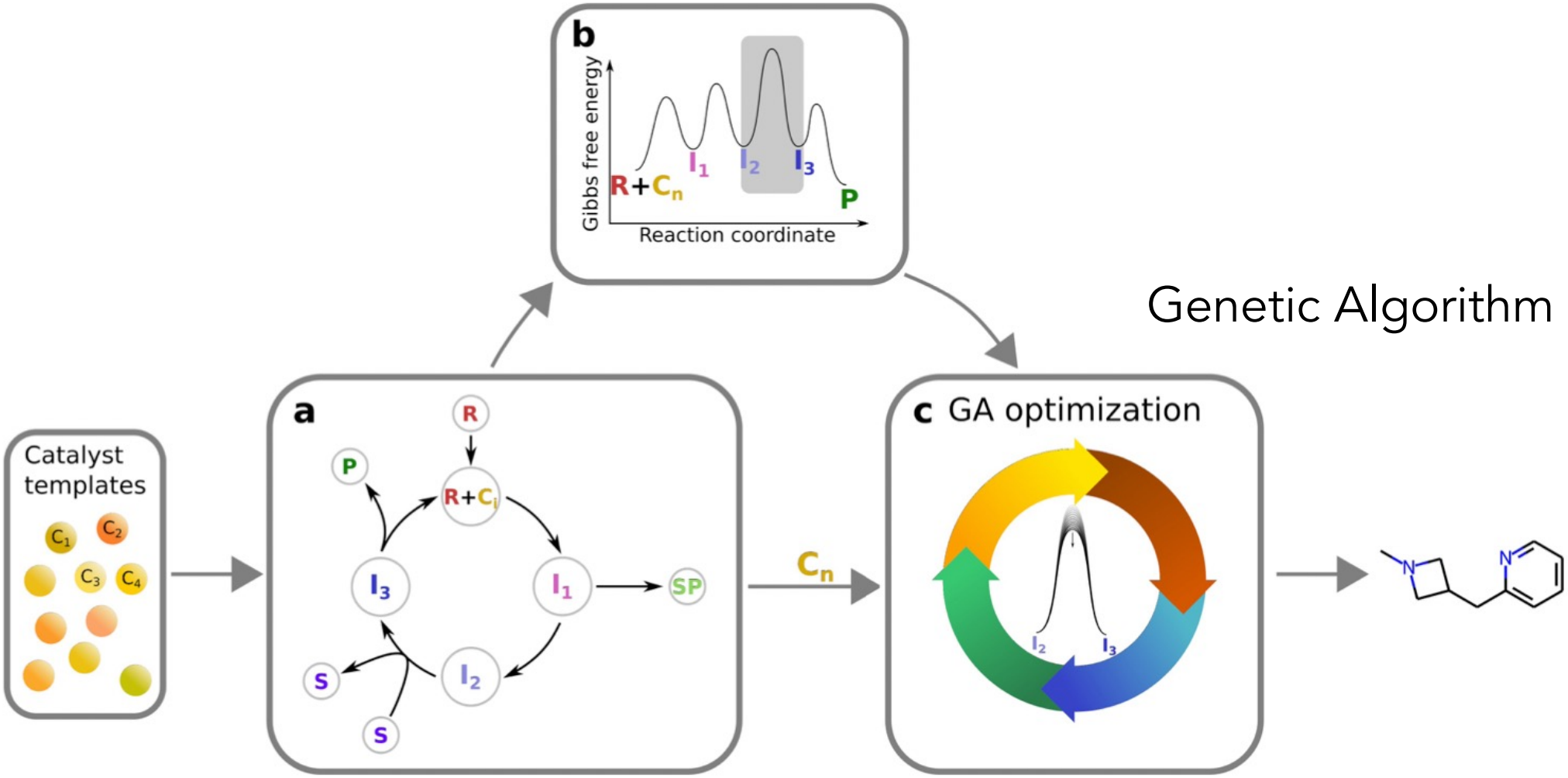
Combinatorial optimization





Hoksza, David, et al. "Molpher: a software framework for systematic chemical space exploration." *Journal of cheminformatics* 6.1 (2014): 1-13.

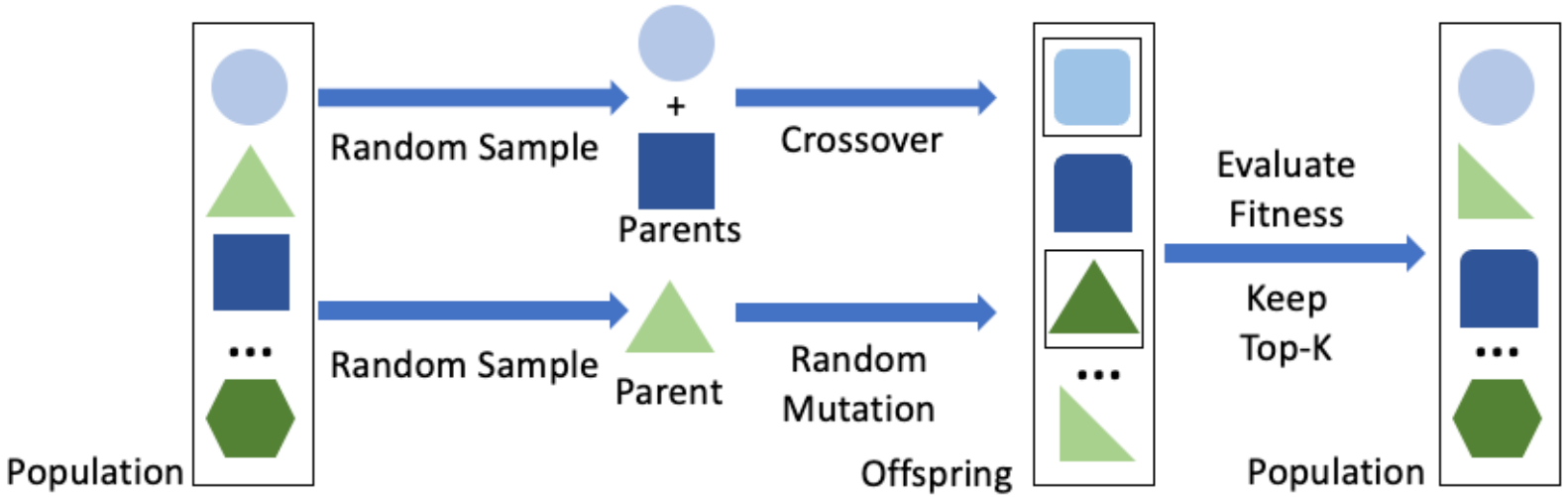
MBH reaction catalyst design



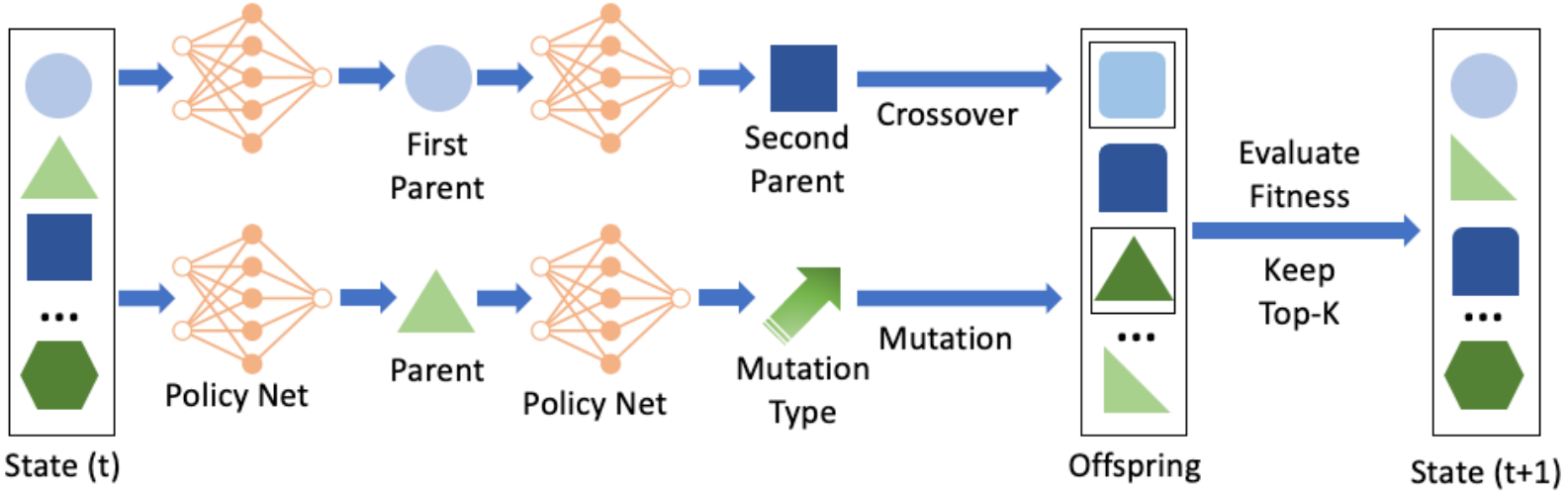
Rasmussen, Maria H., Julius Seumer, and Jan H. Jensen. "Toward De Novo Catalyst Discovery: Fast Identification of New Catalyst Candidates for Alcohol-Mediated Morita–Baylis–Hillman Reactions." *Angewandte Chemie* 135.49 (2023): e202310580.

Machine learning helps decision making

Genetic Algorithm

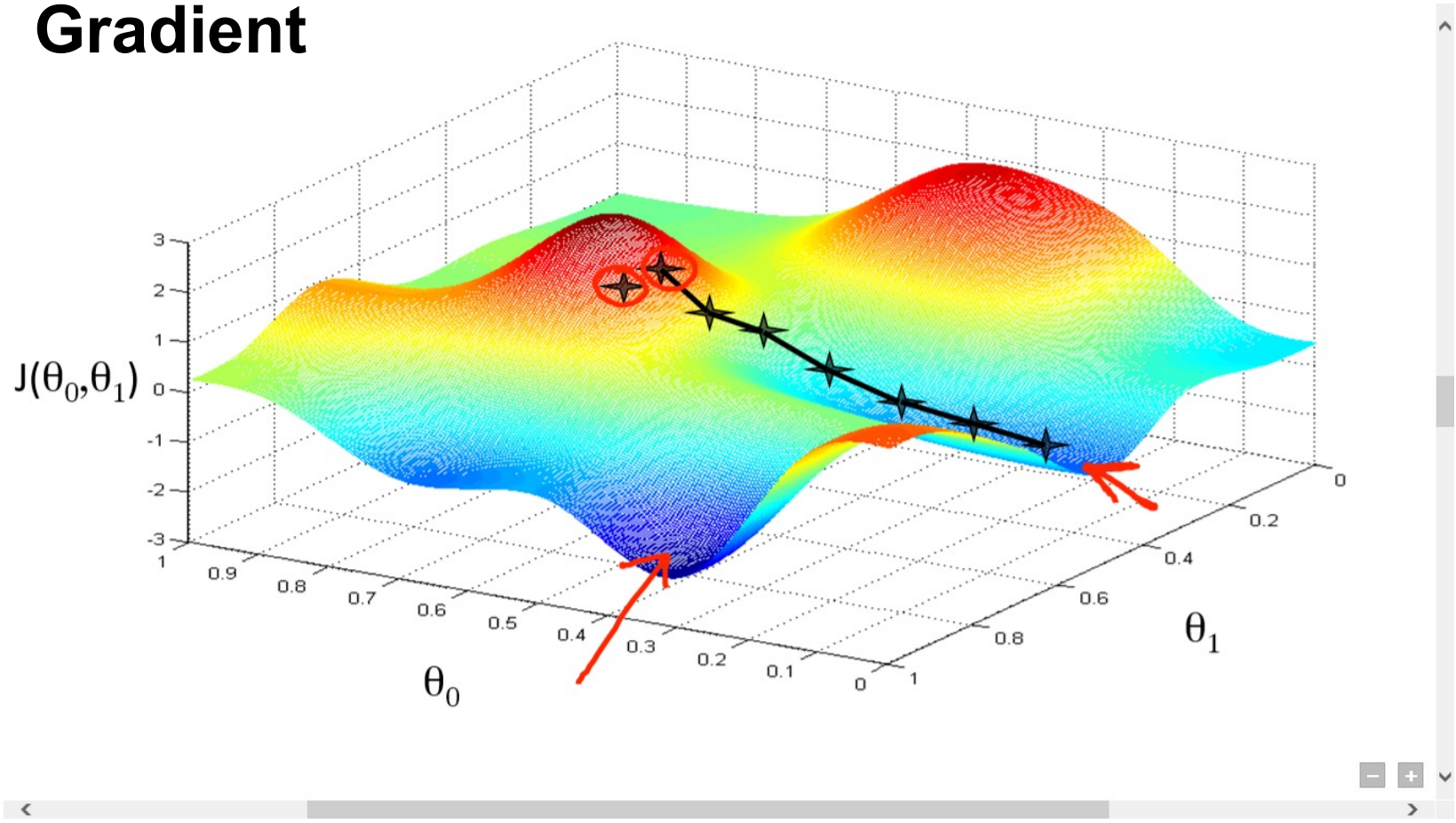


Machine learning enhanced Genetic Algorithm



Fu, Tianfan, et al. "Reinforced genetic algorithm for structure-based drug design." *Advances in Neural Information Processing Systems* 35 (2022): 12325-12338.

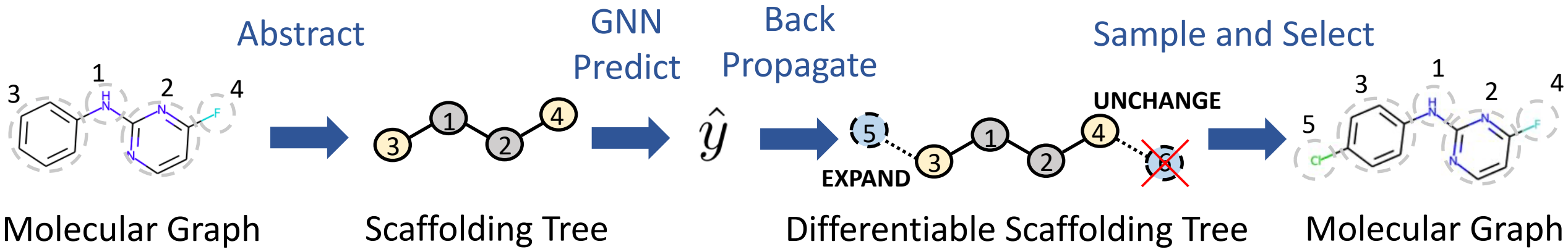
How do we optimize?



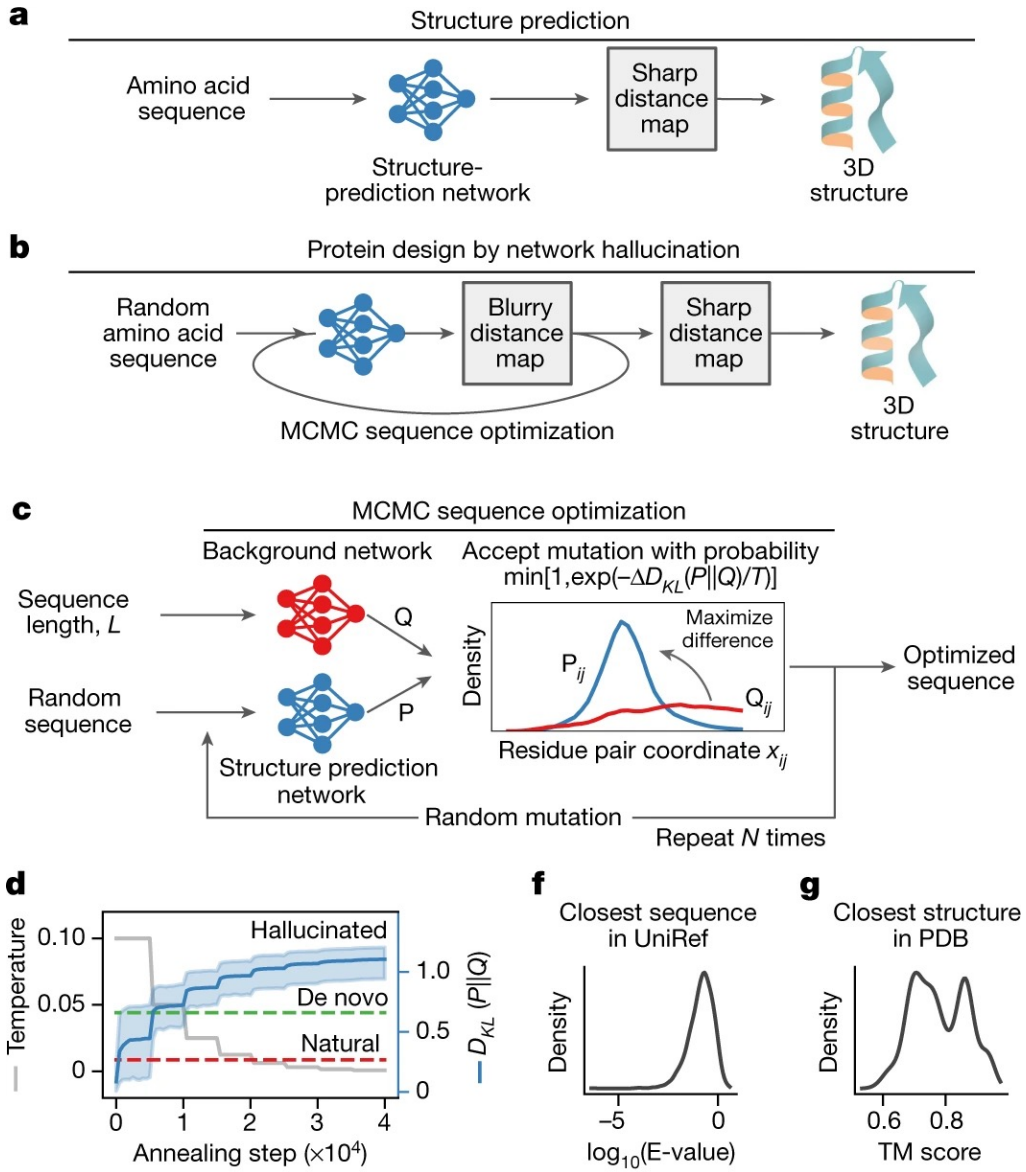
- https://medium.com/hackernoon/gradient-descent-aynk-7cbe95a778da

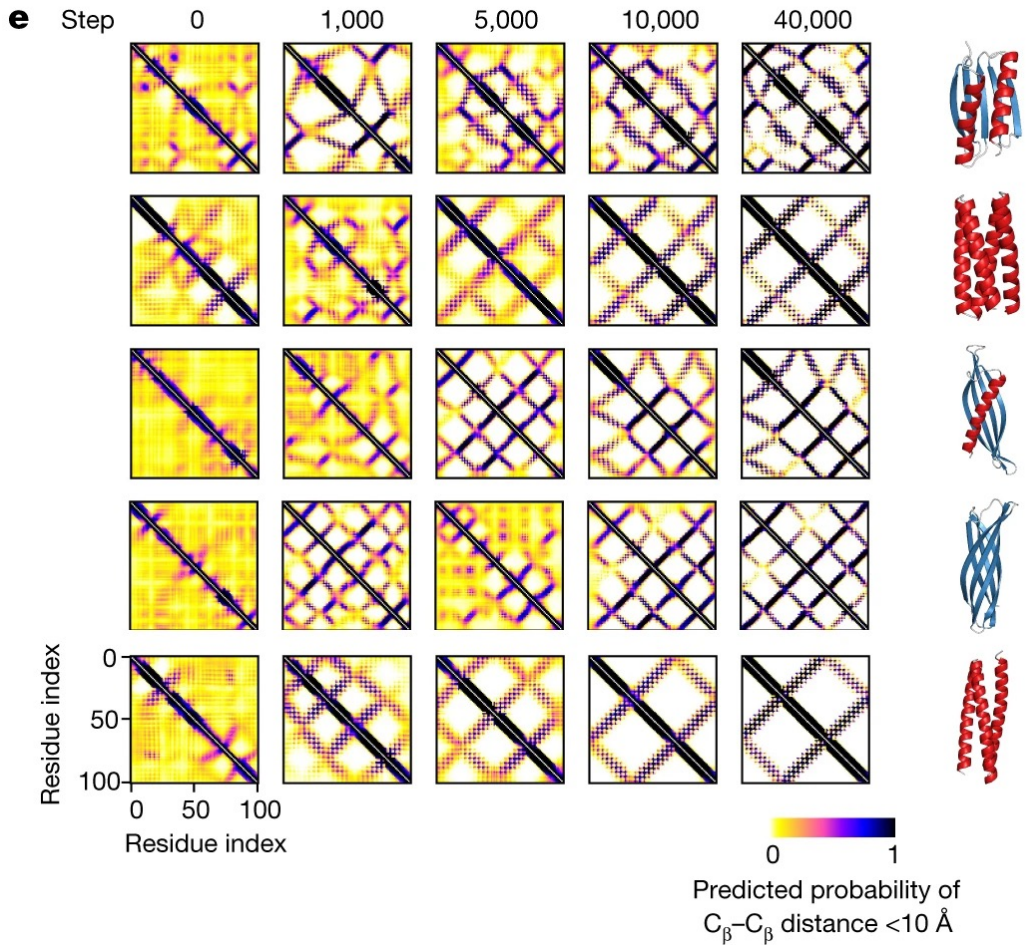
DST: Differentiable Scaffolding Tree

- Differentiable learning on graph representation

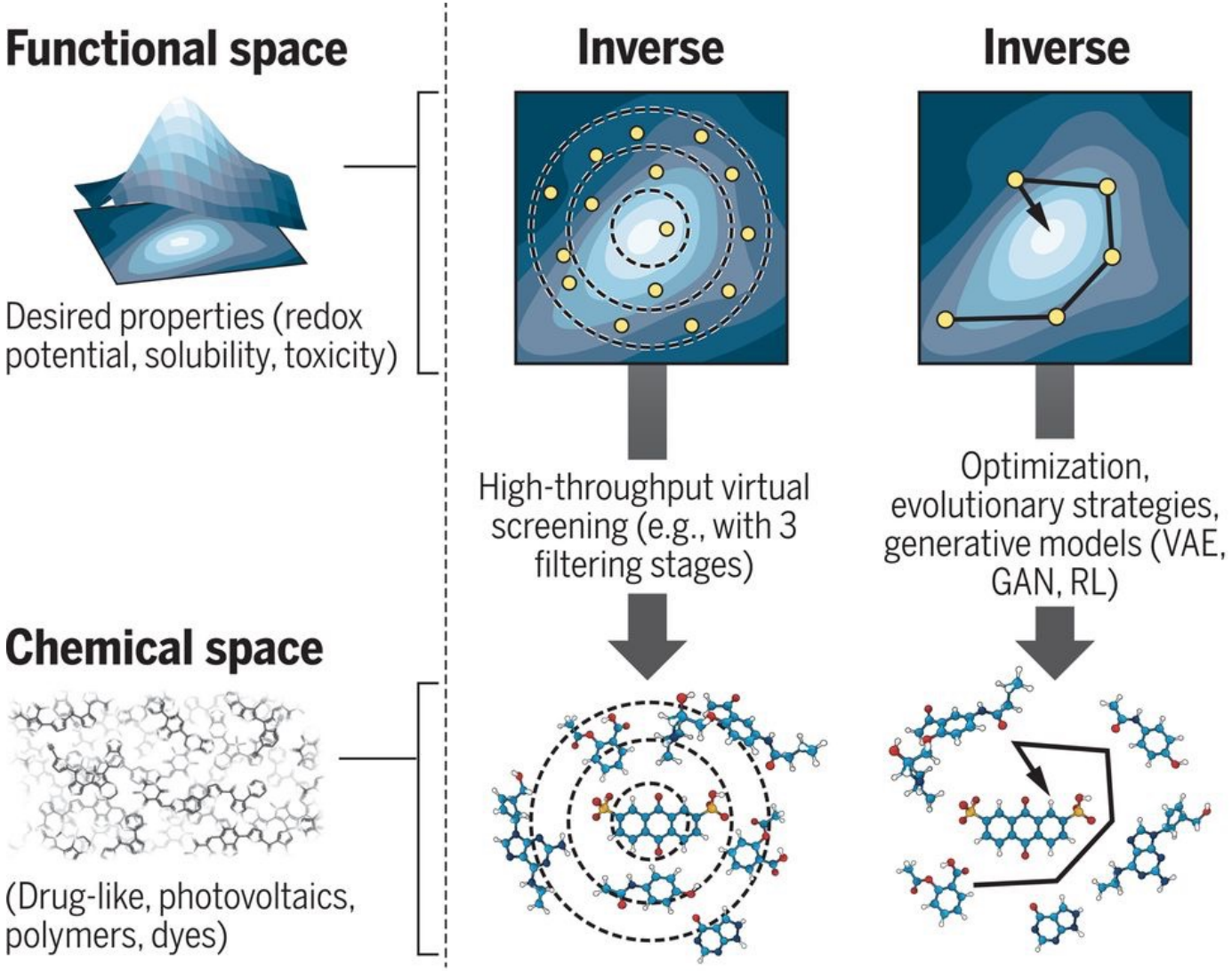


Protein hallucination





Molecular design: screening and *de novo* design



Advantages:

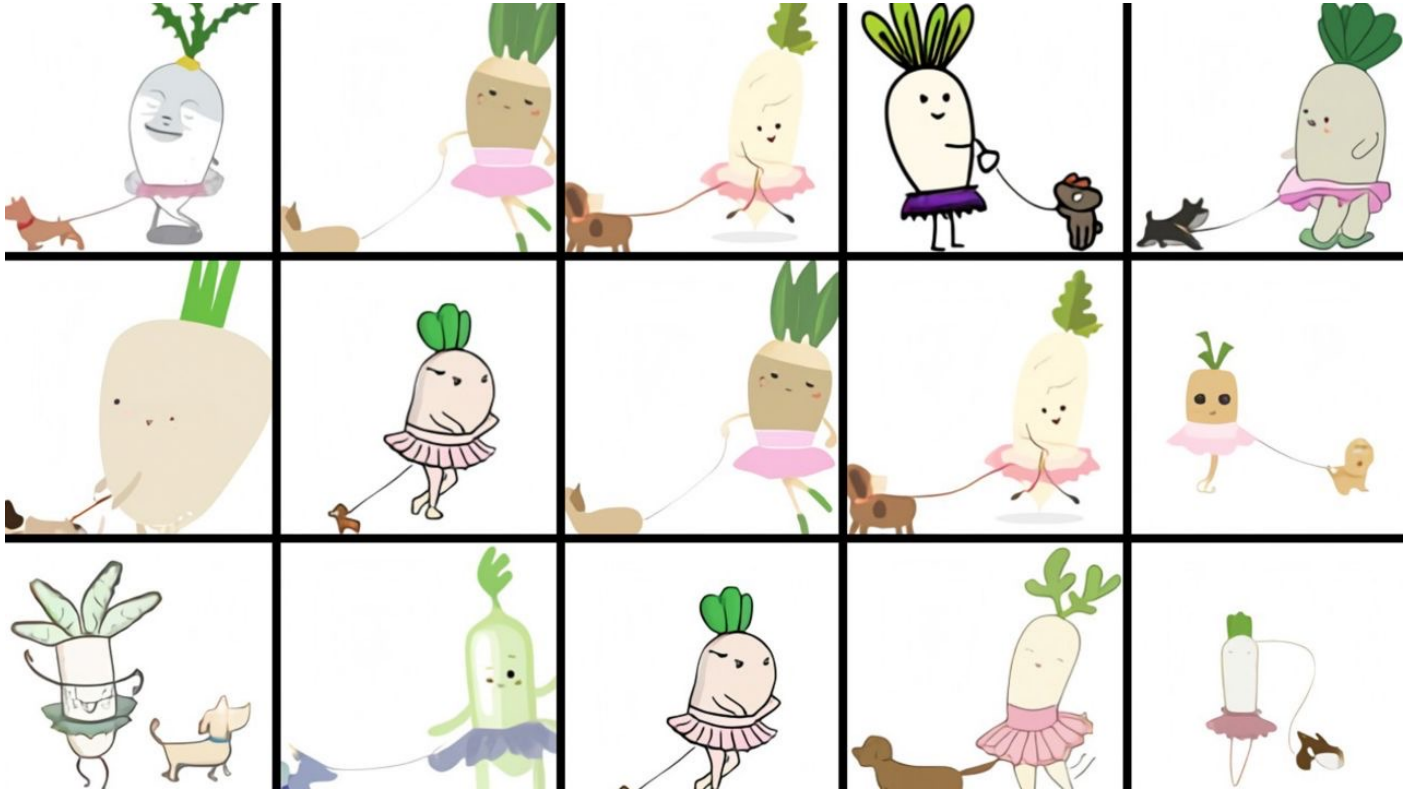
- Implicitly define a chemical space
- Potentially more efficient

Modified from Sanchez-Lengeling, B., & Aspuru-Guzik, A. (2018). *Science*, *361*(6400), 360-365.
Kutchukian, Peter S., and Eugene I. Shakhnovich. *Expert opinion on drug discovery* 5.8 (2010): 789-812.

Generative modeling

Can we generate molecules?

- The output of a model is text, images, etc.
- ChatGPT, DALL-E, Midjourney

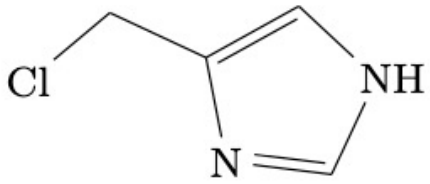


Images generated by DALL-E upon the prompt: "an illustration of a baby daikon radish in a tutu walking a dog"

https://en.wikipedia.org/wiki/DALL-E

Chemical language model

Graph:



SMILES: ClCc1c[nH]cn1

One-hot encoding:

	Cl	C	c	1	c	nH	c	n	1
C	0	1	0	0	0	0	0	0	0
c	0	0	1	0	1	0	1	0	0
n	0	0	0	0	0	0	0	1	0
1	0	0	0	1	0	0	0	0	1
nH	0	0	0	0	0	1	0	0	0
Cl	1	0	0	0	0	0	0	0	0

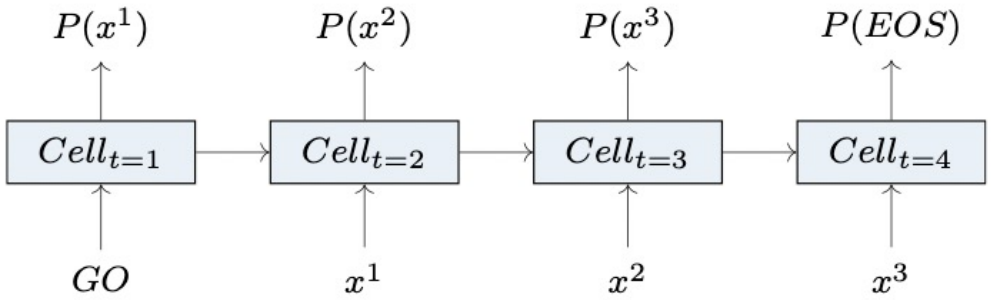


Figure 1 Learning the data. Depiction of maximum likelihood training of an RNN. x^t are the target sequence tokens we are trying to learn by maximizing $P(x^t)$ for each step.

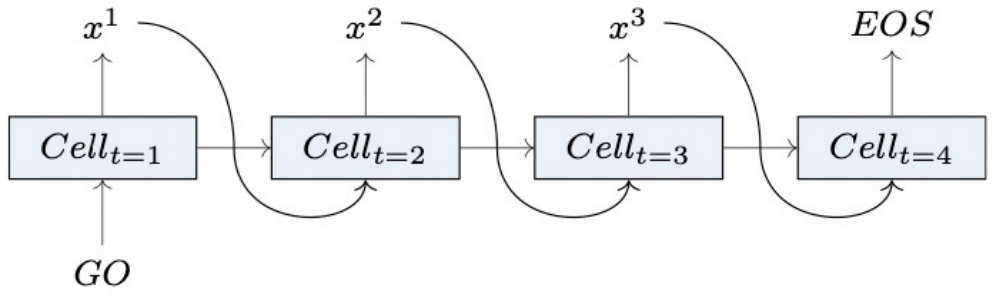


Figure 2 Generating sequences. Sequence generation by a trained RNN. Every timestep t we sample the next token of the sequence x^t from the probability distribution given by the RNN, which is then fed in as the next input.

Transformers, of course

nature machine intelligence

Article

https://doi.org/10.1038/s42256-022-00580-7

Large-scale chemical language representations captur and properties

Received: 18 April 2022

Accepted: 3 November 2022

Published online: 21 December 2022

Check for updates

Jerret Ross, Brian
Youssef Mroueh &

Models based on ma

nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00740-3

Neural scaling of deep chemical models

Received: 14 June 2022

Accepted: 15 September 2023

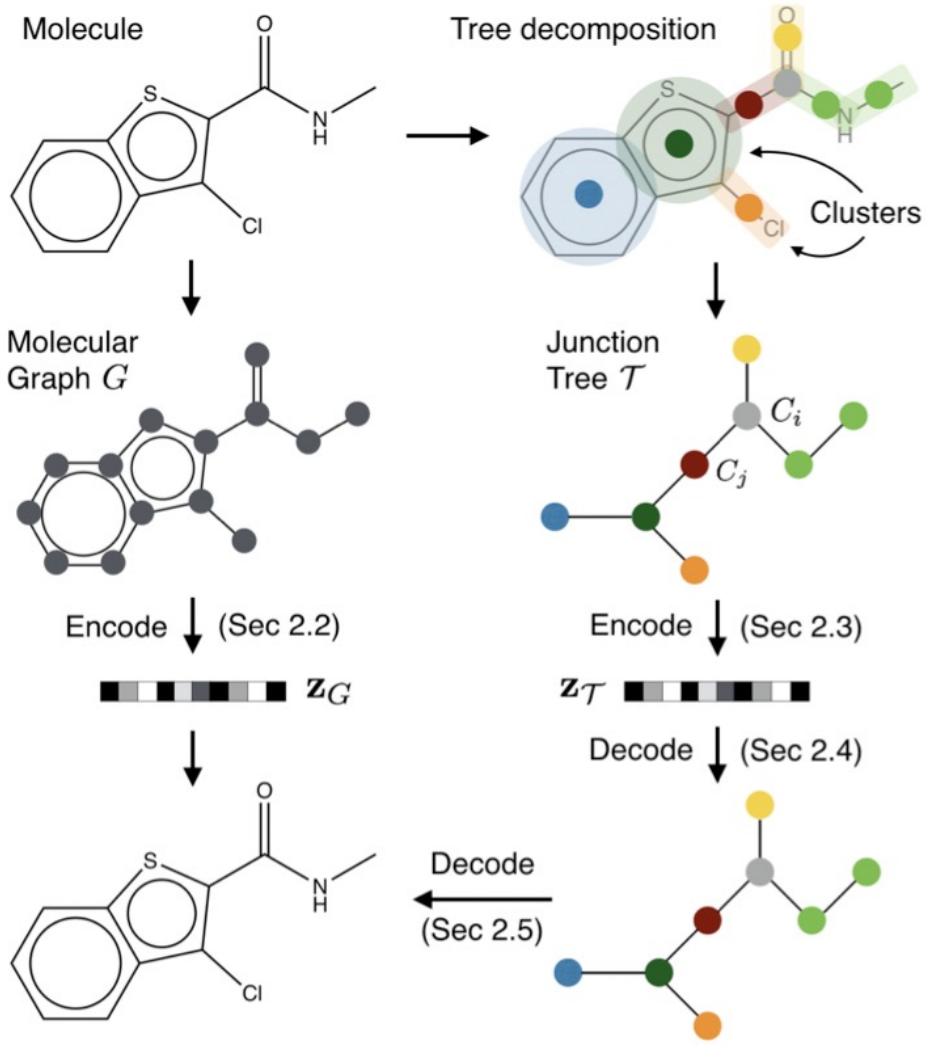
Published online: 23 October 2023

Check for updates

Nathan C. Frey [1,6], Ryan Soklaski[1,7], Simon Axelrod[2,3], Siddharth Samsi[1], Rafael Gómez-Bombarelli [2], Connor W. Coley [4,5] & Vijay Gadepally[1]

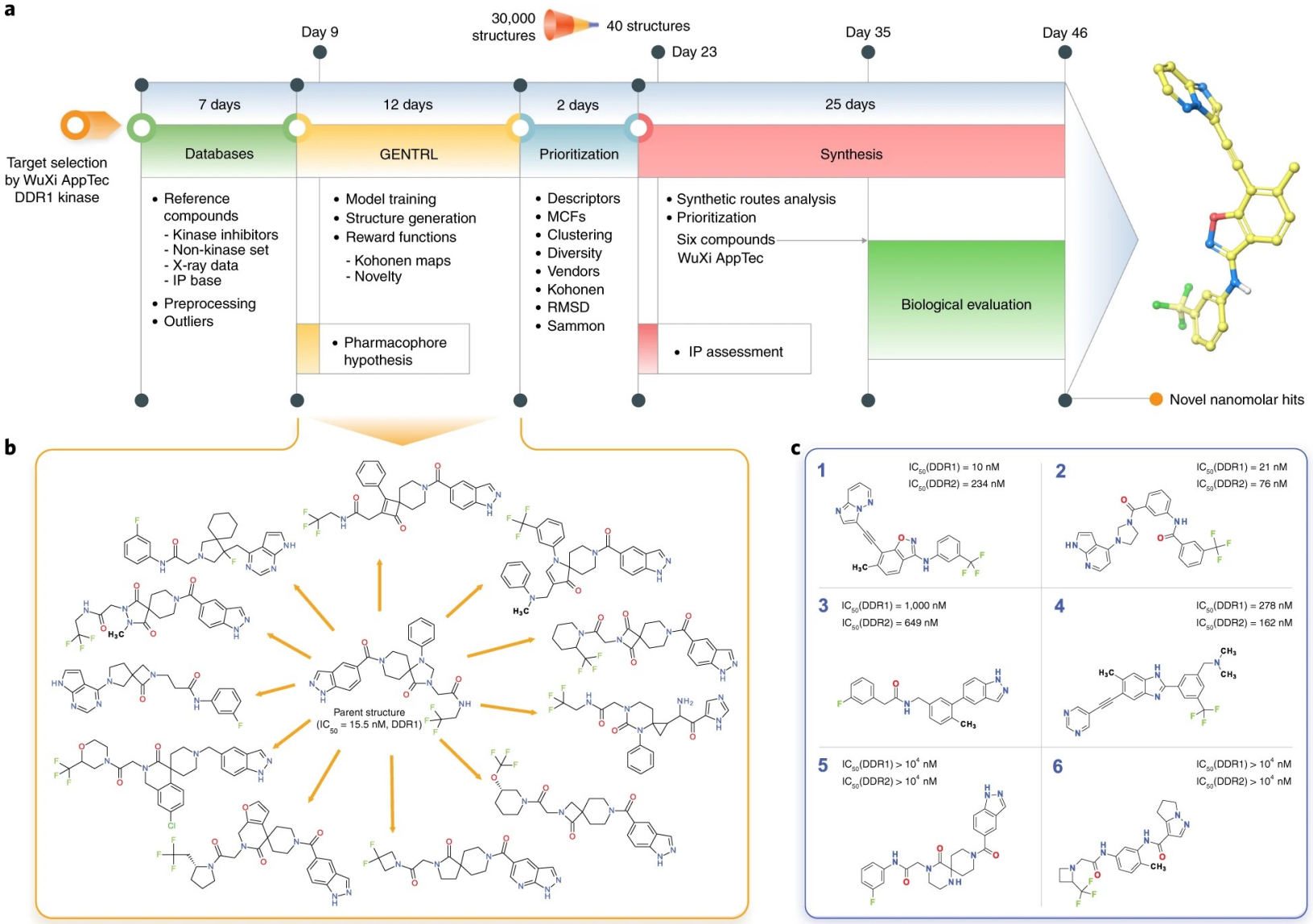
Massive scale, in terms of both data availability and computation, enables important breakthroughs in key application areas of deep learning such as natural language processing and computer vision. There is emerging evidence that scale may be a key ingredient in scientific deep learning, but the importance of physical priors in scientific domains makes the strategies

Graph generative model

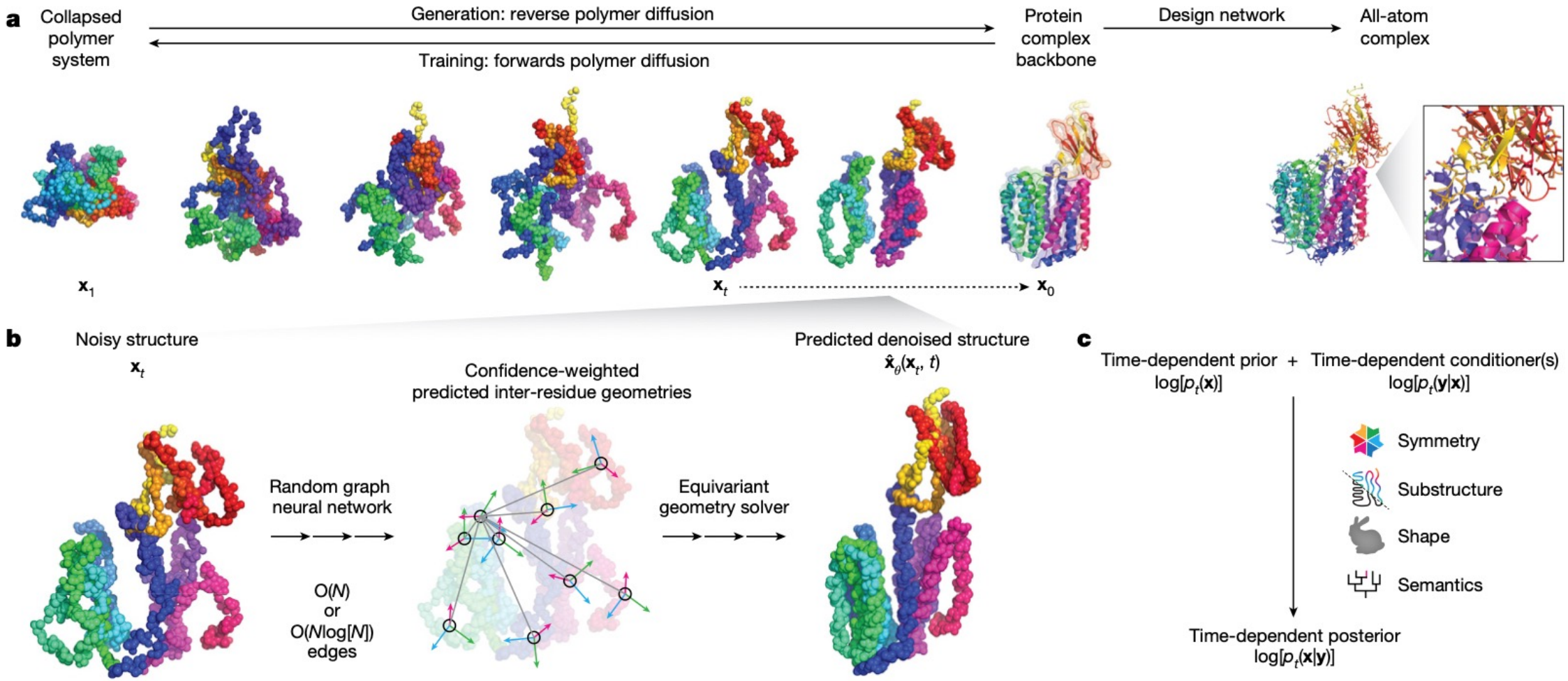


Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation." *International conference on machine learning*. PMLR, 2018.

And it (to some extend) works!

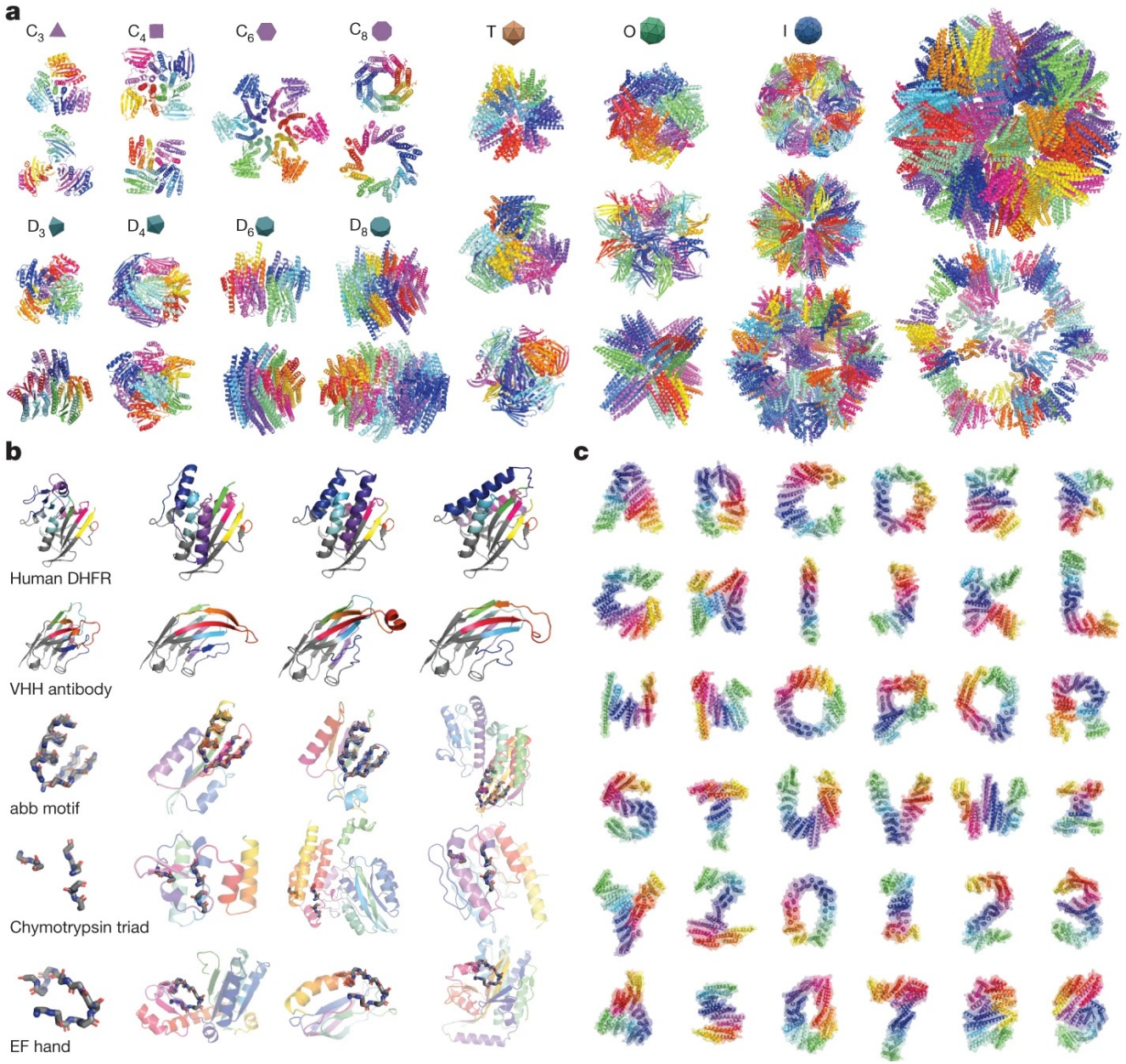


3D structure

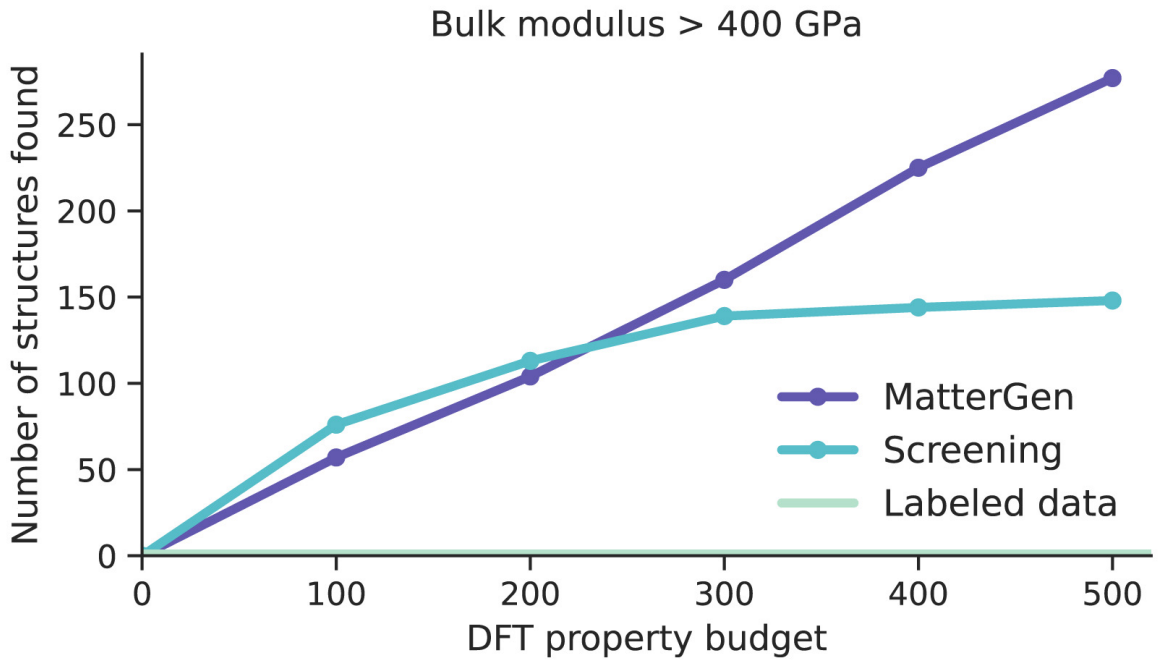


Ingraham, John B., et al. "Illuminating protein space with a programmable generative model." *Nature* (2023): 1-9.

3D structure

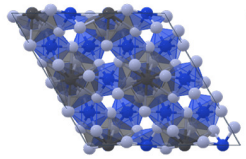


Crystal structure

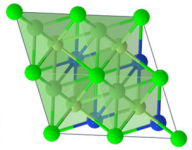


High Symmetry

Cu4PbF12 (Im3)

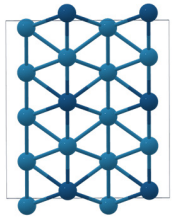


LiZrGaCu ($F\bar{4}3m$)

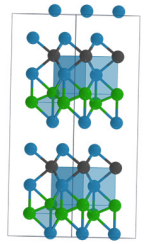


High Bulk Modulus

Re6Ir2 (Cmcm)

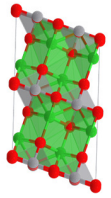


Re3B2C (P3m1)

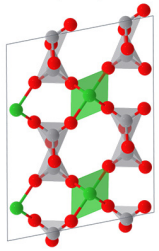


Target Chemistry

SrVO3 (I4/mmm)

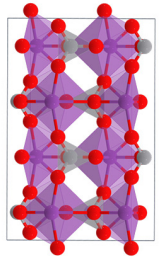


SrV2O6 (Cm)

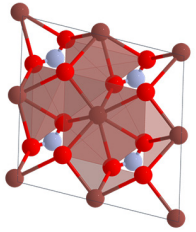


Target Band Gap

V2B2O6 (P21/m)

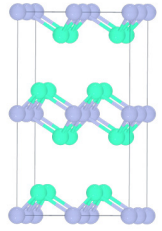


TINO3 (Cm)

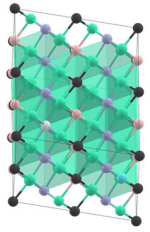


High Magnetic Density

Gd2N ($P\bar{3}m1$)

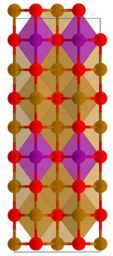


Gd6H2CN3 (C2/m)

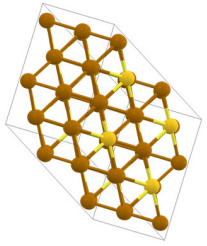


High Magnetic Density + Low HHI Index

MnFe3O4 (Pm)



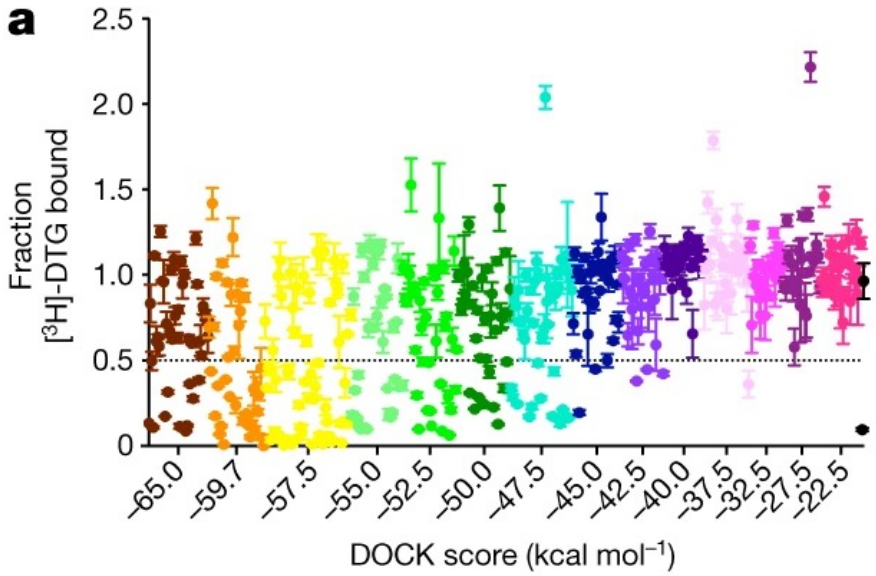
Fe8Au (C2/m)



Zeni, Claudio, et al. "MatterGen: a generative model for inorganic materials design." *arXiv preprint arXiv:2312.03687*(2023).

Course Project: a simulated drug design campaign

- We developed a simulation protocol based on molecular docking. We treat this oracle as ground truth to simulate experimental evaluation.
- We also add drug likeliness and synthesizability into consideration, resulting in a single scalar score, ranging from 0 to 1, to optimize.
- Your task is to design novel molecules that have higher activity under limited evaluation budget.



What makes a good design protocol?

- Efficient
- Novel
- Diverse
- Synthesizable
- …

Course Project: a simulated drug design campaign

- Performance metric: (average score of top-30) + 0.3 * (Internal diversity of top-30)
- We will hold a leaderboard showing top-10 teams.

Internal diversity

We define the *internal diversity* I of a set of molecules A of size $|A|$ to be the average of the Tanimoto-distance T_d of molecules of A with respect to each other. Formally, we have:

$$I(A) = \frac{1}{|A|^2} \sum_{(x,y) \in A \times A} T_d(x, y) \qquad (1)$$

For a sufficiently large set A, any sufficiently large subset $A' \subset A$, sampled with uniform probability, has the same internal diversity as A. This property follows from the law of large numbers. We can thus define the internal diversity of a generative model, by computing the internal diversity of a sufficiently large generated sample. This allows to formalize our challenge:

Benhenda, Mostapha. "ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity?." *arXiv preprint arXiv:1708.08227* (2017).

Course Project: a simulated drug design campaign

- The class is centered around solving this problem:

Date	Day of the Week	Topic	Link to Colab
1/8/2024	Monday	Course overview + Broad review of historical development and common workflows	[colab notebook] coming soon!
1/10/2024	Wednesday	Data process and analysis: focus on dimensionality reduction and clustering	[colab notebook] coming soon!
1/12/2024	Friday	Structure-property relationship modeling (Part 1): featurization of molecules	[colab notebook] coming soon!
1/16/2024	Tuesday	Literature presentation (session 1)	[slides] coming soon!
1/17/2024	Wednesday	Structure-property relationship modeling (Part 2): deep learning architectures	[colab notebook] coming soon!
1/19/2024	Friday	Molecular generation and design (Part 1): screening and generative AI	[colab notebook] coming soon!
1/22/2024	Monday	Molecular generation and design (Part 2): optimization approach	[colab notebook] coming soon!
1/24/2024	Wednesday	Literature presentation (session 2)	[colab notebook] coming soon!
1/26/2024	Friday	Guest Lecture 1	[slides] coming soon!
1/29/2024	Monday	Guest Lecture 2	[slides] coming soon!
1/31/2024	Wednesday	Guest Lecture 3	[slides] coming soon!
2/2/2024	Friday	Final project presentations	[slides] coming soon!

Finally

- Please fill the ML/science prior knowledge survey
- We will invite every registered students into our slack channel, where you can reach out and team up.
- For anyone who doesn't mind who to team with, please send an email with title "[Your name] + subject to team assignment" to moleculedesigner@gmail.com
- Try not to drop after this Wed.

Other questions?