AI beyond the Hype: **Applications to ChemE**

Romain Lacombe
rlacombe@stanford.edu

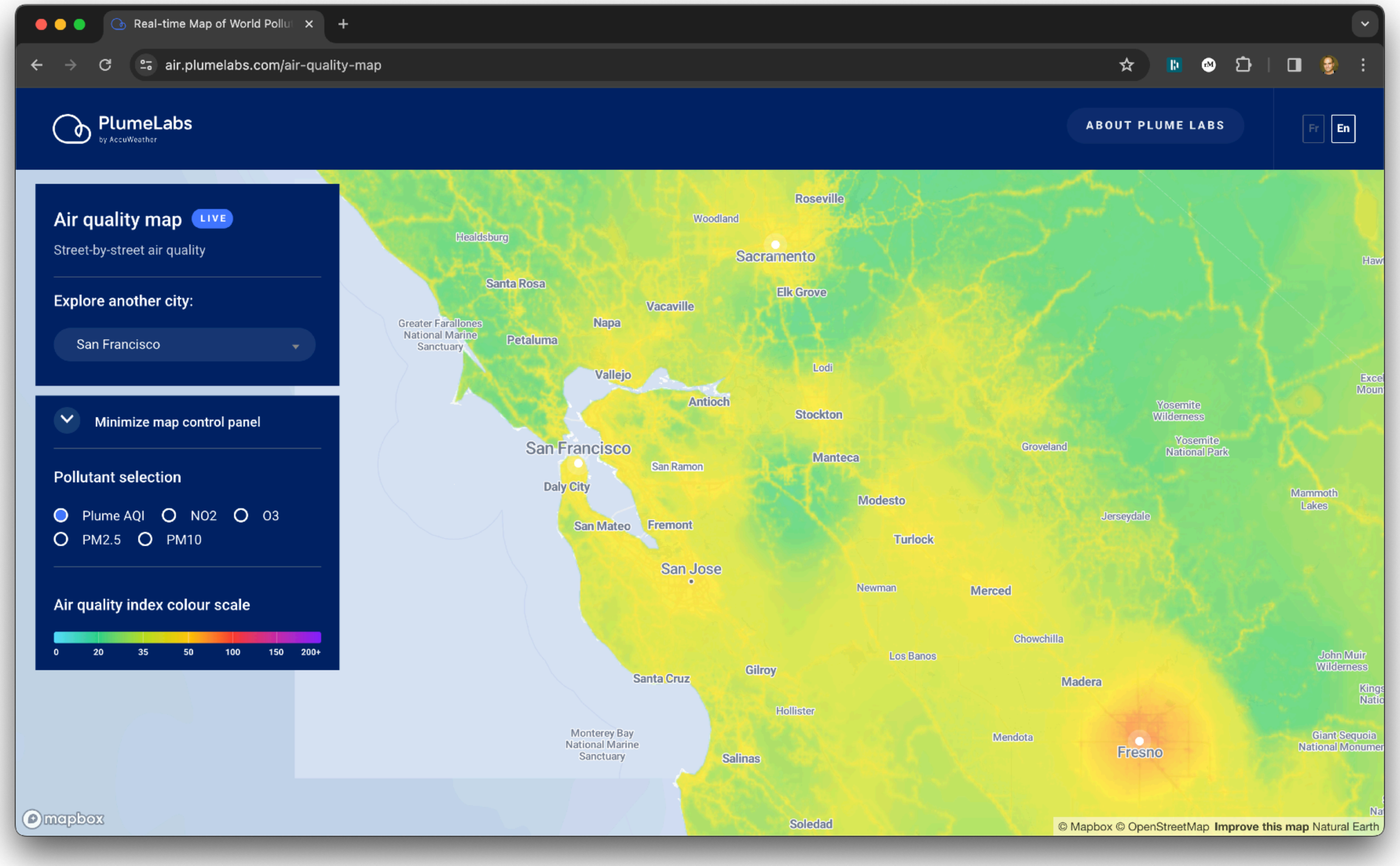
About me

- Undergraduate in Physics and Math (France) then MS Engineering Systems at MIT on climate economics (carbon markets and refineries)

- Climate technology entrepreneur (Plume Labs, acq. by AccuWeather)

- MS ChemEng HCP candidate with a focus on AI for chemistry.

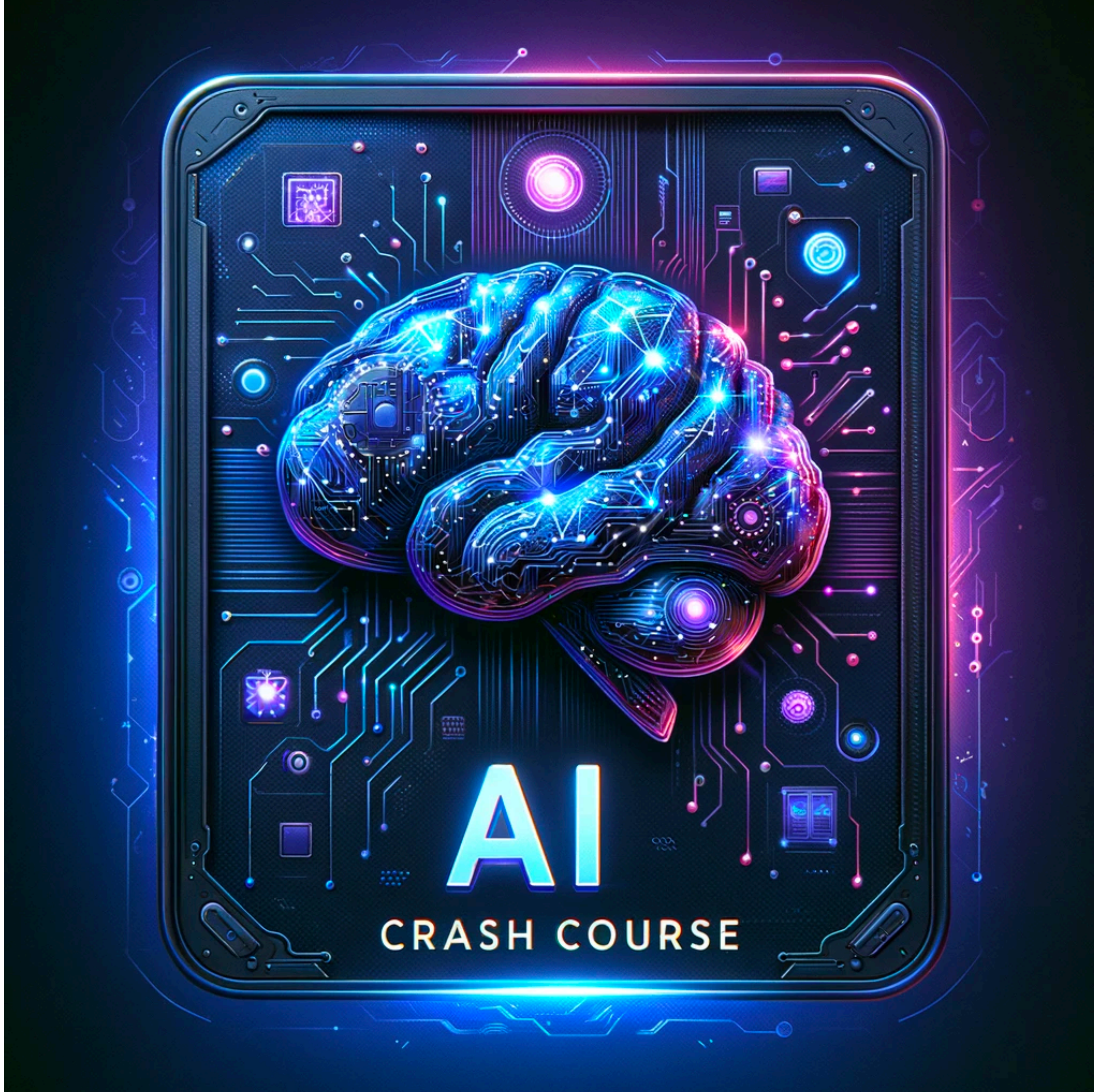
- Goal: build better materials/processes to help decarbonize at scale.



Outline

- AI crash course
- Extracting molecular properties from natural language
- Discovering catalysts with reinforcement learning
- Conclusion: new frontiers in AI for materials

An AI Crash Course

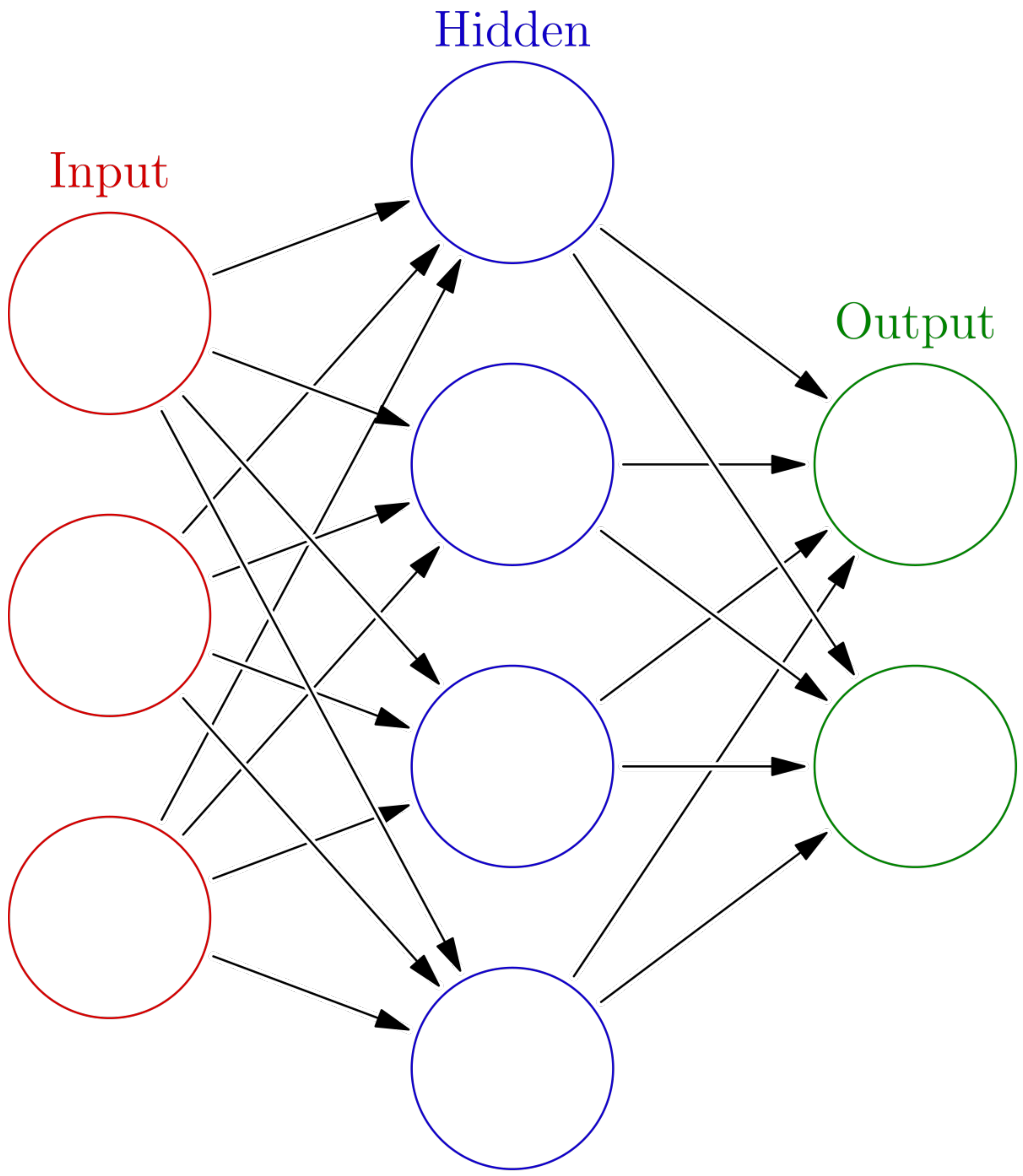


Neural networks

They’re just (very) complex functions.

- Inspired by biological neurons
- Activations: multiply inputs by matrix weights + apply non-linearities
- Universal Approximation Theorem (Cybenko, 1989)
- How to find the right weights? **Learn by gradient descent!**

Learn more: **CS 230**



WikiMedia

Why all the hype?

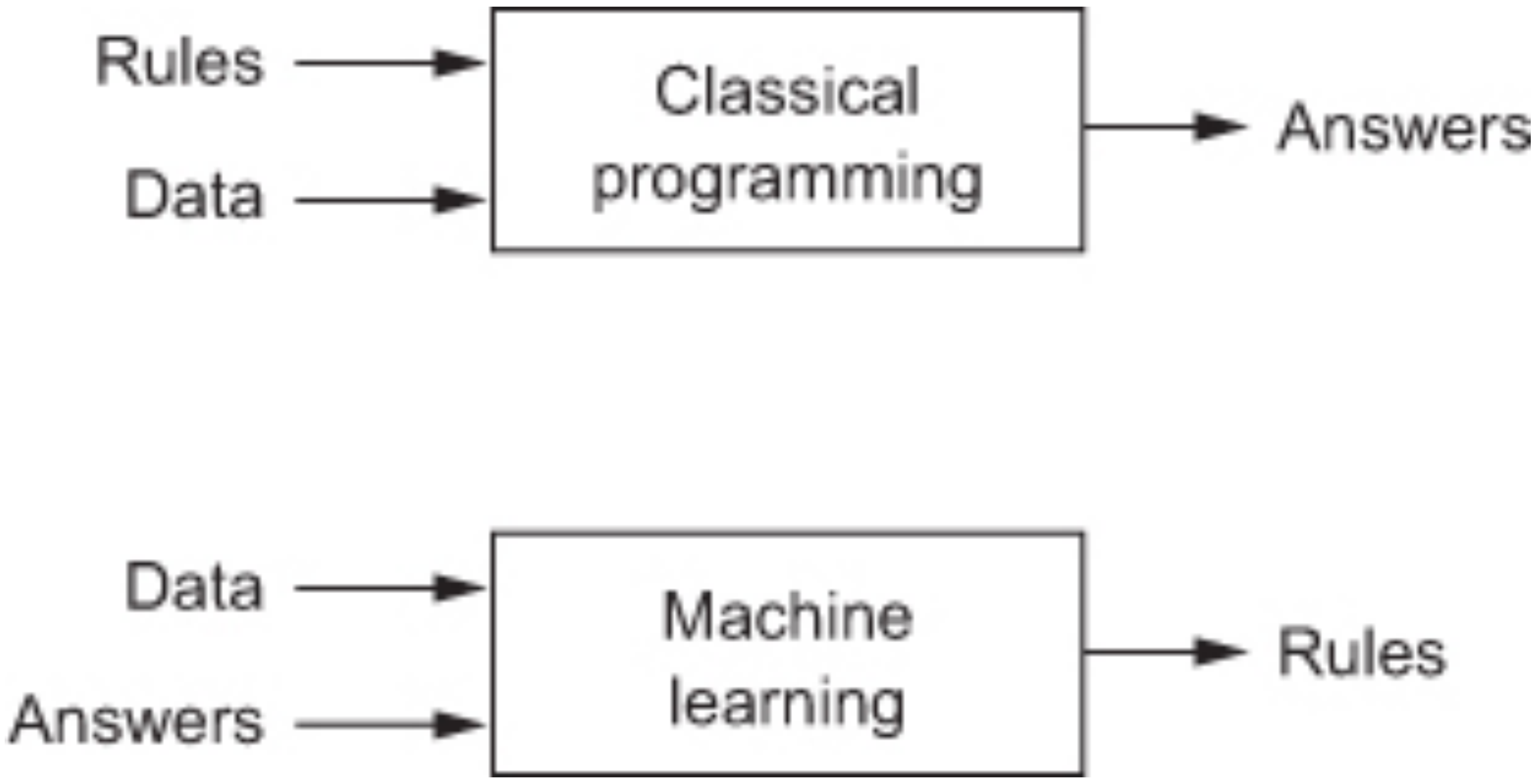
Because these ideas finally work!

- First neural networks conceptualized in the 1950s: intractable to train.
- Started working well c. 2012: large networks learning from large datasets.
- 2010s-2020s: **Deep Learning Era**
- “Learning”: don’t program rules, learn them from the data
- Deep: represent these functions as large neural networks
  (“deep” = large number of hidden layers)

Learn more: **CS 230**

Paradigm shift: learning from data

Enabled by GPUs and data scale



Source: François Cholet, “Deep Learning with Python”

Source: ImageNet Large Scale Visual Recognition Challenge

Learn more: **CS 229**

Paradigm shift: deep neural nets

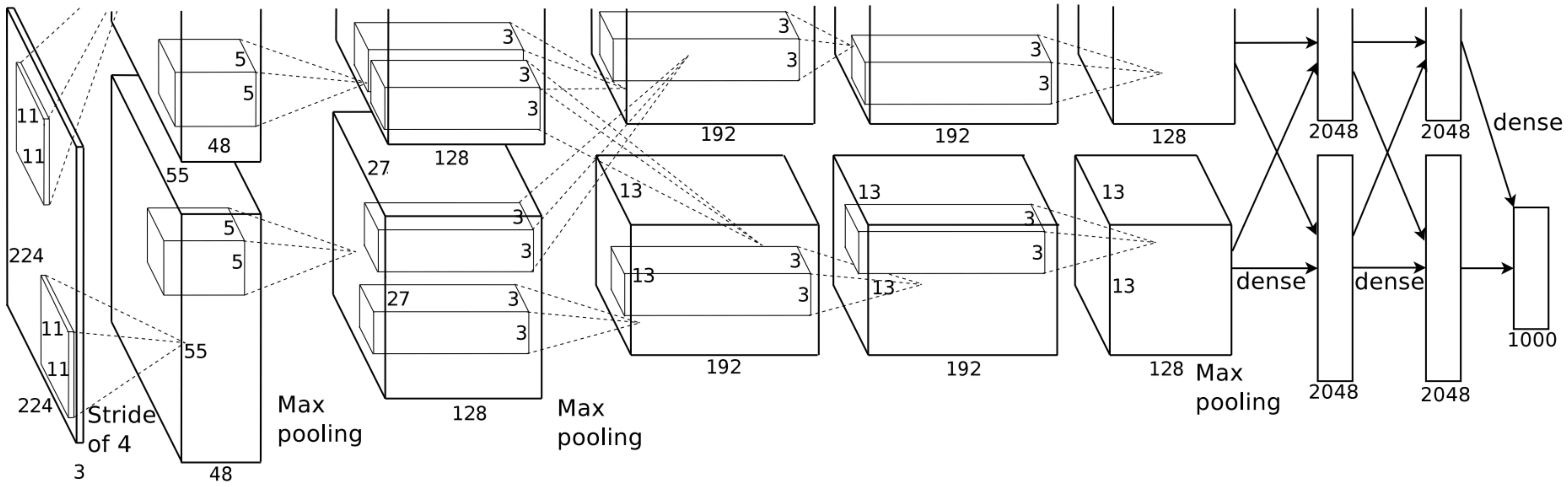
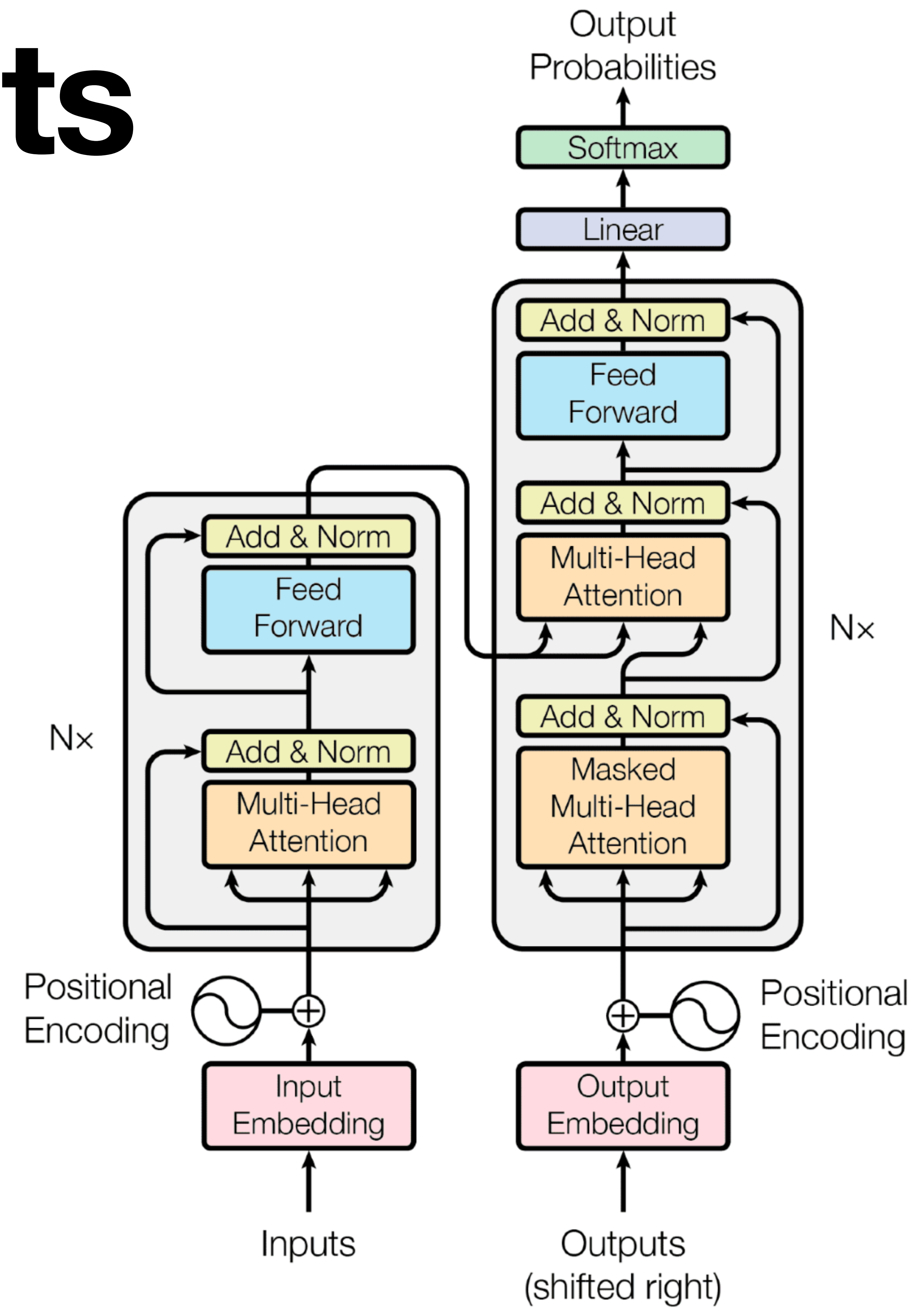


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

Krizhevsky et al., 2012, ImageNet Classification with Deep Convolutional Neural Networks

Learn more: **CS 231N**



Vaswani et al., 2017, Attention Is All You Need

Learn more: **CS 224N & CS 224U**

Flavors of learning

Different recipes, same neural nets

- **Supervised learning:** learn to match a value/label
  Ex: classification (image → dog or cat?), regression (molecule → solubility?)

Learn more: **CS 230**

- **Generative learning:** learn to generate an object
  Ex: AA sequence → 3d structure of a protein (AlphaFold)

Learn more: **CS 236 & CS 279**

- **Reinforcement learning:** learn through play
  Ex: play millions of games → beat humans at Go (AlphaGo)

Learn more: **CS 224R**

- **Contrastive learning**: learn to match/contrast samples
  Ex: several face photographs → are they the same person?

Learn more: **CS 224W**

Why is this relevant?

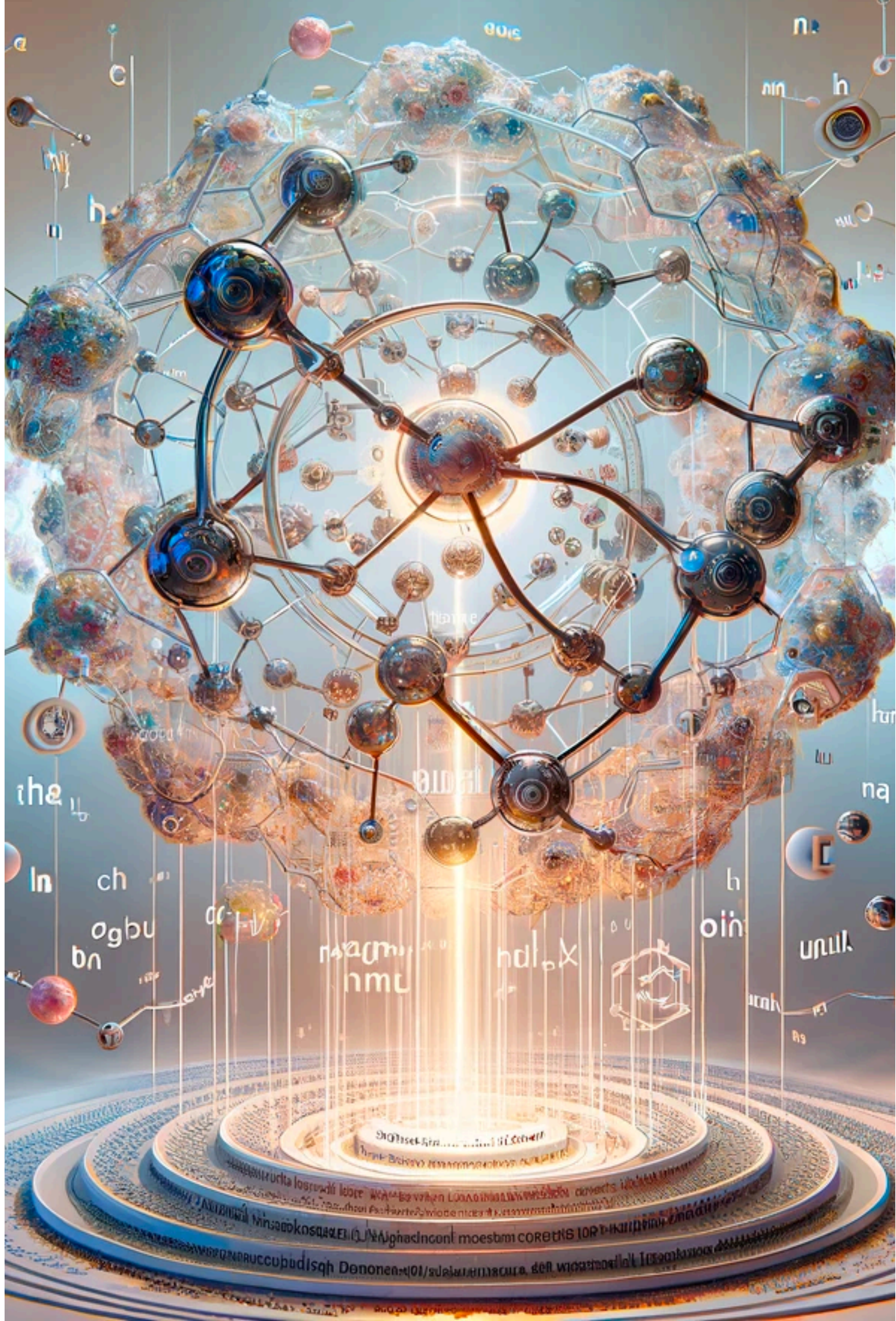
Two examples of ChemE applications

- **Molecular properties prediction:** learn to predict properties of an organic molecule given only its 2-d graph structure.
  ACS Fall 2023 AI for Organic Chemistry workshop

- **Catalyst discovery:** explore large spaces of possible metal catalysts fitting a target adsorption energies profile.
  NeurIPS 2023 Accelerated Materials Discovery workshop

Extracting molecular properties from natural language

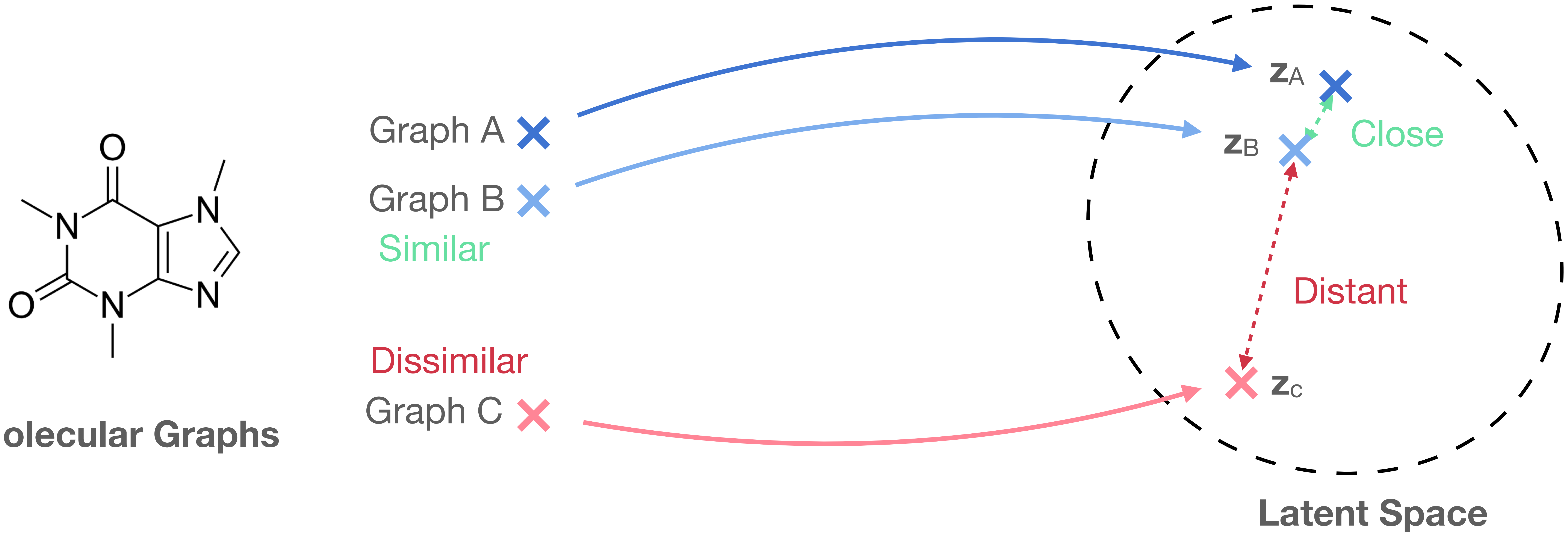
https://arxiv.org/abs/2307.12996



Contrastive learning

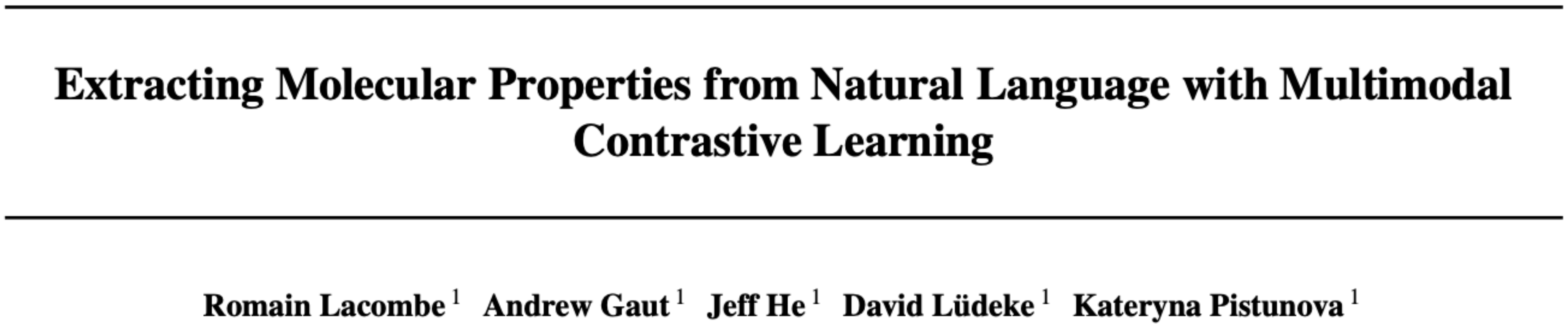
- Tasks in ML for chemistry require **deep molecular graph representations**

- GNNs can be trained to learn effective representations through **contrastive learning:**



Can we learn directly from scientific papers?

Treasure trove of collective knowledge now accessible.

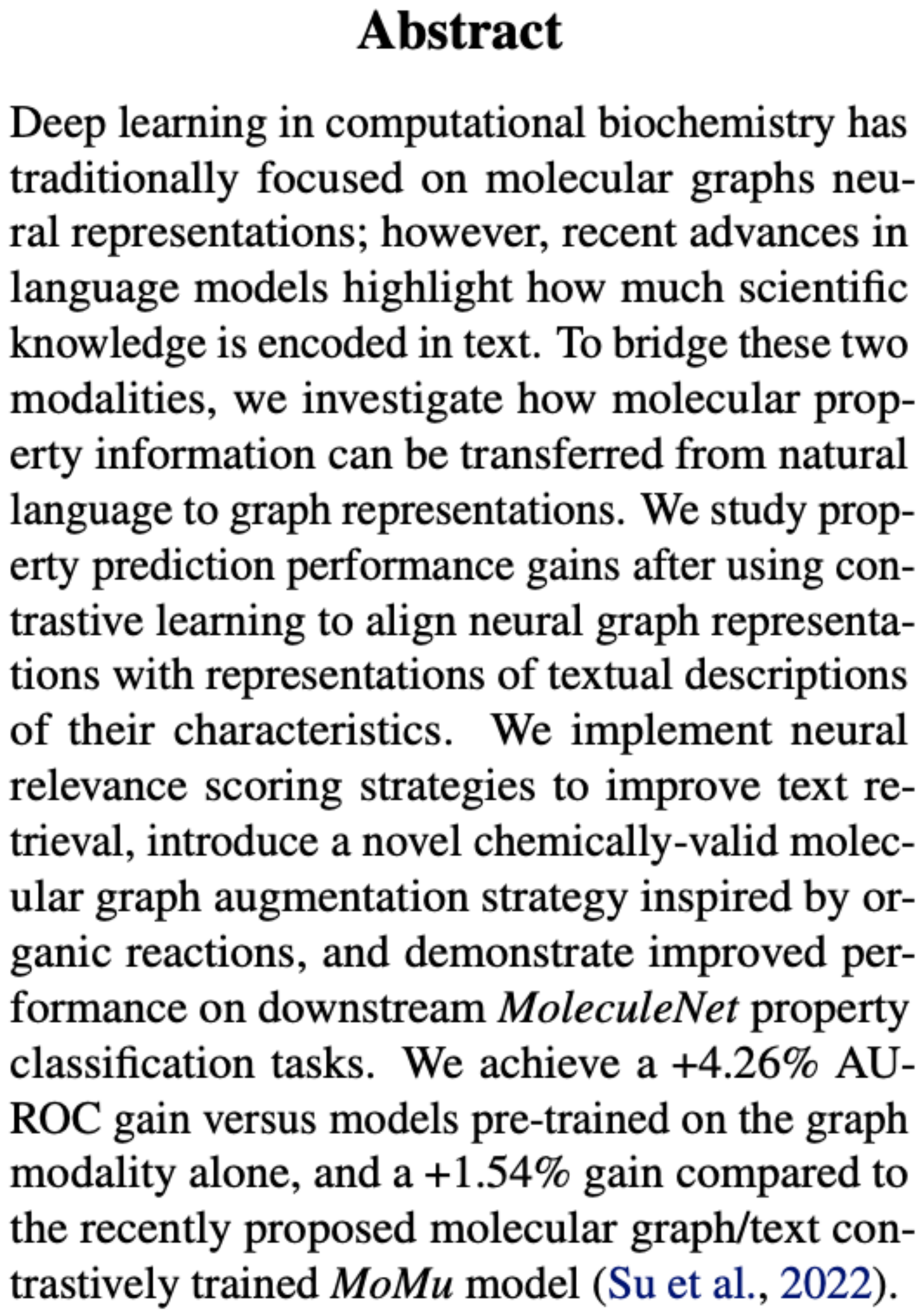


Extracting Molecular Properties from Natural Language with Multimodal Contrastive Learning

Romain Lacombe[1] Andrew Gaut[1] Jeff He[1] David Lüdeke[1] Kateryna Pistunova[1]

ACS Fall 2023 AI for Organic Chemistry workshop

https://arxiv.org/abs/2307.12996

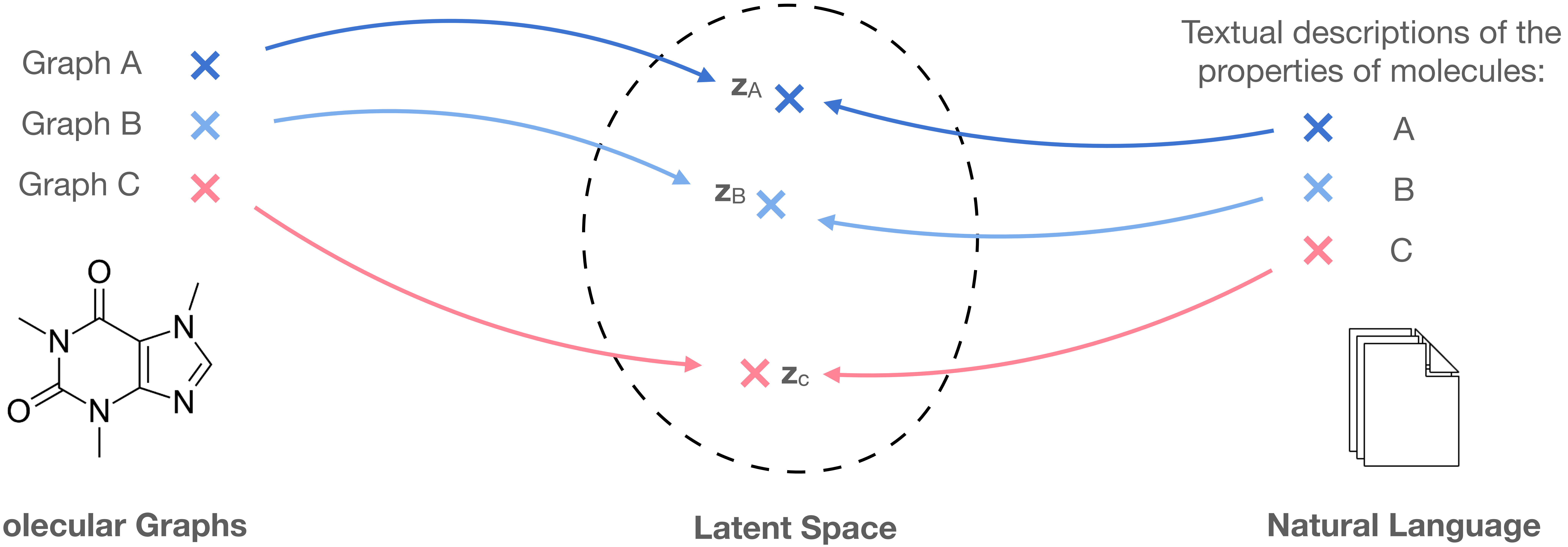


Abstract

Deep learning in computational biochemistry has traditionally focused on molecular graphs neural representations; however, recent advances in language models highlight how much scientific knowledge is encoded in text. To bridge these two modalities, we investigate how molecular property information can be transferred from natural language to graph representations. We study property prediction performance gains after using contrastive learning to align neural graph representations with representations of textual descriptions of their characteristics. We implement neural relevance scoring strategies to improve text retrieval, introduce a novel chemically-valid molecular graph augmentation strategy inspired by organic reactions, and demonstrate improved performance on downstream *MoleculeNet* property classification tasks. We achieve a +4.26% AUROC gain versus models pre-trained on the graph modality alone, and a +1.54% gain compared to the recently proposed molecular graph/text contrastively trained *MoMu* model (Su et al., 2022).

Align graph and text latent representations

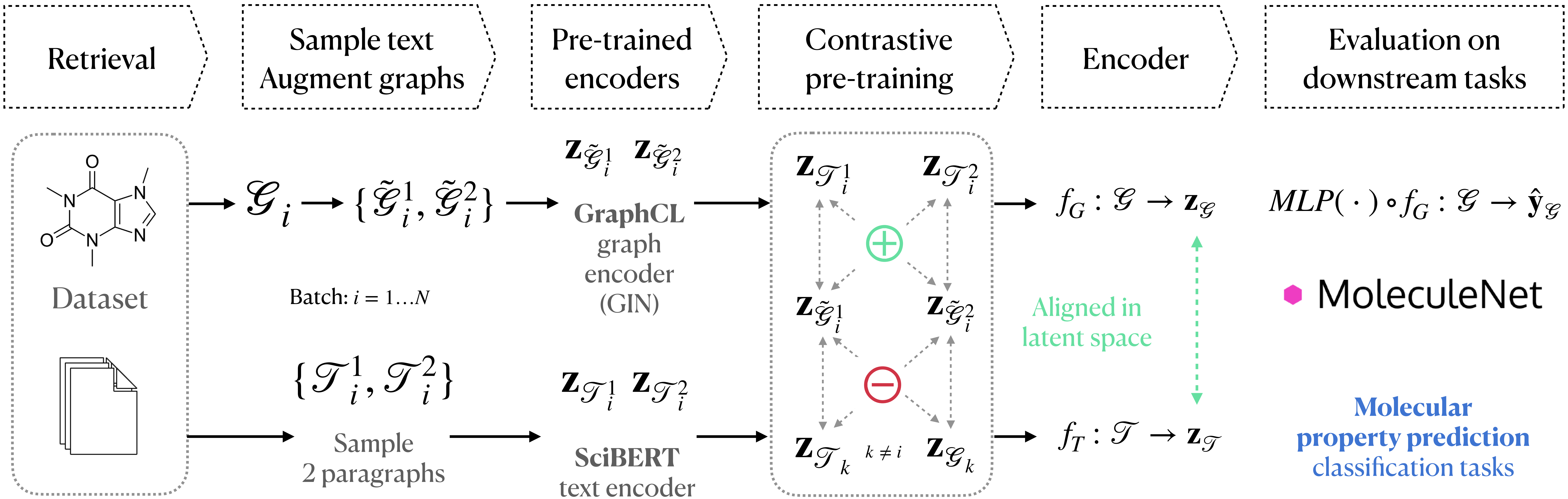
Using contrastive learning.



Do aligned representations improve property predictions?

Aligning graph and text representations

Using contrastive learning.



Retrieval

Sample text Augment graphs

Pre-trained encoders

Contrastive pre-training

Encoder

Evaluation on downstream tasks

Dataset

$\mathcal{G}_i \rightarrow \{\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2\}$

Batch: $i = 1 \ldots N$

$\{\mathcal{T}_i^1, \mathcal{T}_i^2\}$

Sample 2 paragraphs

$\mathbf{z}_{\tilde{\mathcal{G}}_i^1}$ $\mathbf{z}_{\tilde{\mathcal{G}}_i^2}$

GraphCL graph encoder (GIN)

$\mathbf{z}_{\mathcal{T}_i^1}$ $\mathbf{z}_{\mathcal{T}_i^2}$

SciBERT text encoder

$\mathbf{z}_{\mathcal{T}_i^1}$ $\mathbf{z}_{\mathcal{T}_i^2}$

⊕

$\mathbf{z}_{\tilde{\mathcal{G}}_i^1}$ $\mathbf{z}_{\tilde{\mathcal{G}}_i^2}$

⊖

$\mathbf{z}_{\mathcal{T}_k}$ $k \neq i$ $\mathbf{z}_{\mathcal{G}_k}$

$f_G : \mathcal{G} \rightarrow \mathbf{z}_{\mathcal{G}}$

Aligned in latent space

$f_T : \mathcal{T} \rightarrow \mathbf{z}_{\mathcal{T}}$

$MLP(\cdot) \circ f_G : \mathcal{G} \rightarrow \hat{\mathbf{y}}_{\mathcal{G}}$

MoleculeNet

Molecular property prediction classification tasks

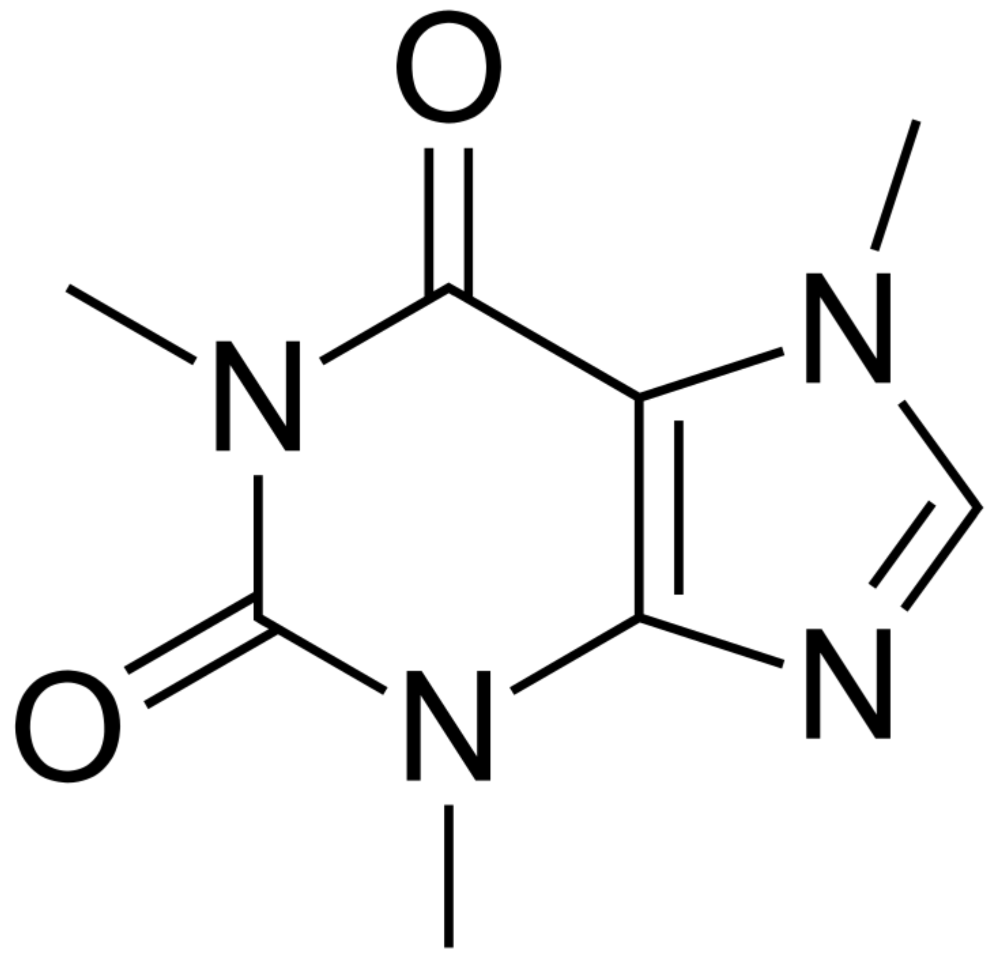
$$\ell(\mathcal{T}_i, \tilde{\mathcal{G}}_i) = -\log \frac{\exp\left(\cos(\mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_i^{\mathcal{G}})/\tau\right)}{\sum_{j\neq i} \exp\left(\cos(\mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_j^{\mathcal{G}})/\tau\right)}$$

/arxiv.org/abs/2209.05481

Could we generate molecules from text?

```
~$ sudo apt-get \
  install coffee
```

Text prompt
(`make me coffee’)



Molecular graph
(Caffeine ☕)

Answer: yes!

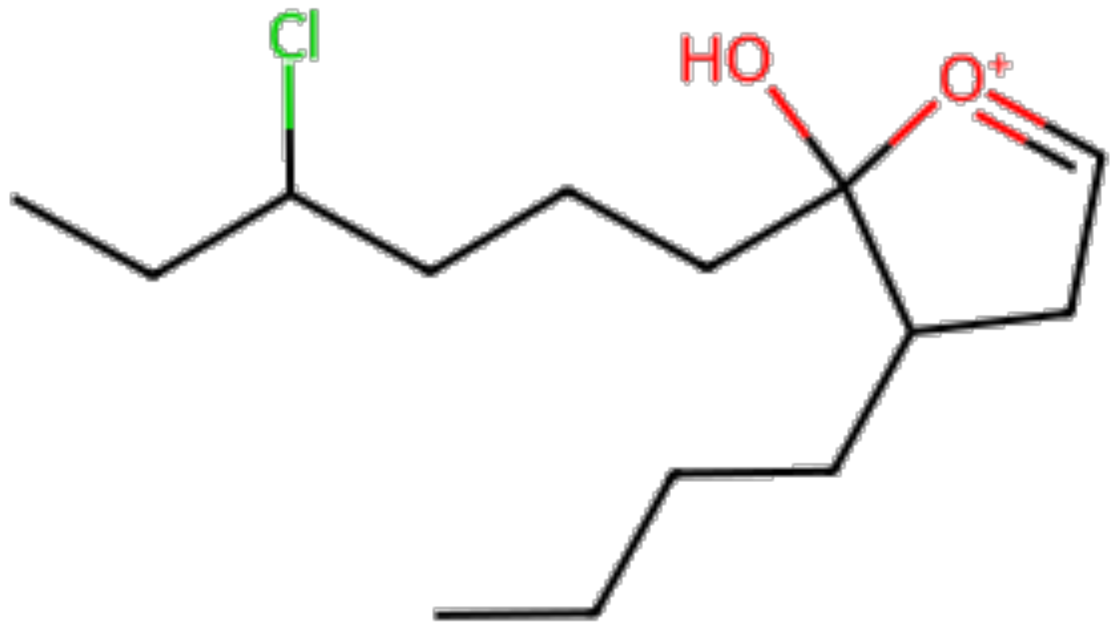
But not very well.

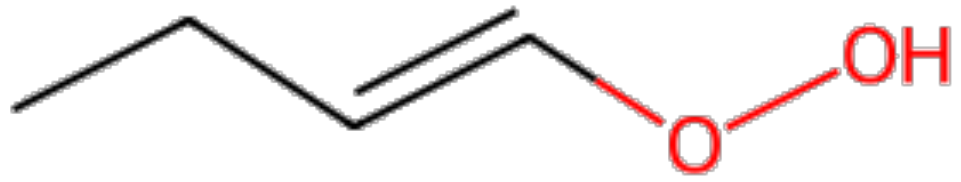
Prompt

“This molecule has a hydroxyl group and a carbonyl group”

“This molecule is hazardous for health”

Generation





Experiment: evaluation

MoleculeNet benchmark.

Evaluate graph representations on property prediction tasks (*MoleculeNet*)

- **BACE**: inhibitors of a human enzyme involved in Alzheimer.
- **BBBP:** blood-brain barrier penetration by small molecules.
- **Clintox:** classification of drugs approved/rejected by the FDA for toxicity.
- **MUV:** virtual molecule screening built on PubChem.
- **SIDER:** adverse side reactions of marketed drugs.
- **Tox21:** classification of toxicity measured by biological reactions and stress response.
- **ToxCast:** 600 tasks linked to in vitro toxicology data.

Encoder
Evaluation on downstream tasks

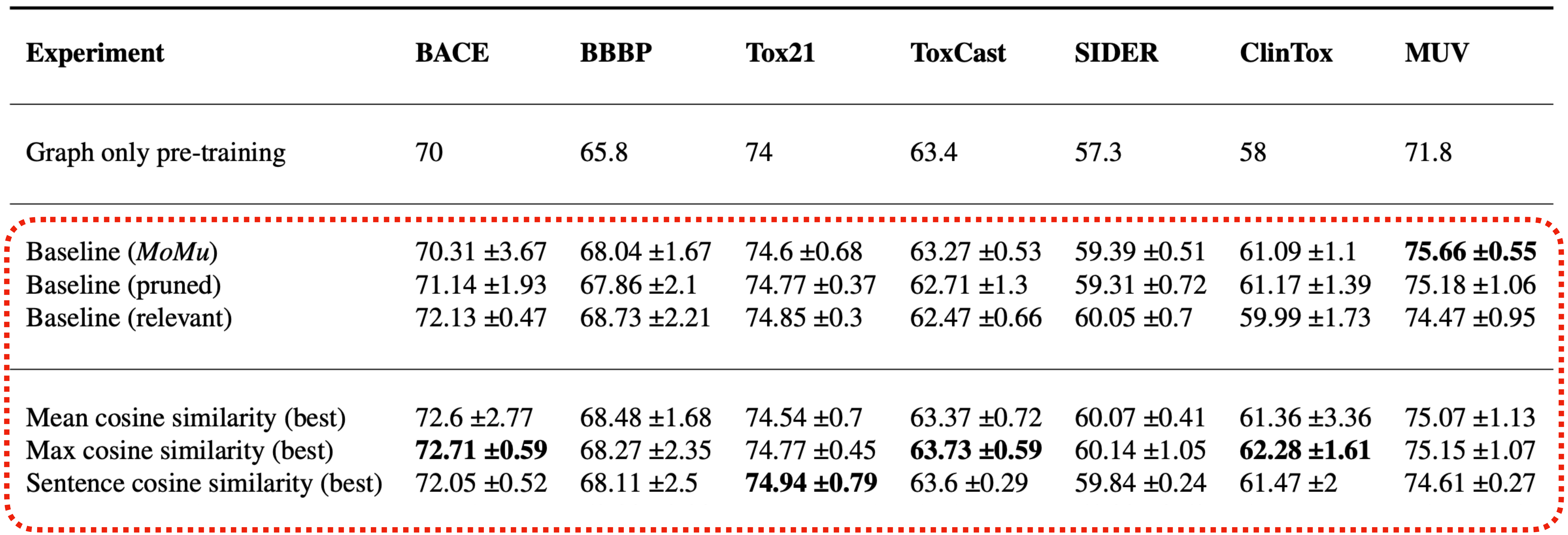
$f_G : \mathscr{G} \rightarrow \mathbf{z}_{\mathscr{G}}$ $\quad$ $MLP(\,\cdot\,) \circ f_G : \mathscr{G} \rightarrow \hat{\mathbf{y}}_{\mathscr{G}}$

MoleculeNet

Results

Graph only

Graph + natural language text



Experiment	BACE	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV
Graph only pre-training	70	65.8	74	63.4	57.3	58	71.8
Baseline (*MoMu*)	70.31 ±3.67	68.04 ±1.67	74.6 ±0.68	63.27 ±0.53	59.39 ±0.51	61.09 ±1.1	**75.66 ±0.55**
Baseline (pruned)	71.14 ±1.93	67.86 ±2.1	74.77 ±0.37	62.71 ±1.3	59.31 ±0.72	61.17 ±1.39	75.18 ±1.06
Baseline (relevant)	72.13 ±0.47	68.73 ±2.21	74.85 ±0.3	62.47 ±0.66	60.05 ±0.7	59.99 ±1.73	74.47 ±0.95
Mean cosine similarity (best)	72.6 ±2.77	68.48 ±1.68	74.54 ±0.7	63.37 ±0.72	60.07 ±0.41	61.36 ±3.36	75.07 ±1.13
Max cosine similarity (best)	**72.71 ±0.59**	68.27 ±2.35	74.77 ±0.45	**63.73 ±0.59**	60.14 ±1.05	**62.28 ±1.61**	75.15 ±1.07
Sentence cosine similarity (best)	72.05 ±0.52	68.11 ±2.5	**74.94 ±0.79**	63.6 ±0.29	59.84 ±0.24	61.47 ±2	74.61 ±0.27

Table 1. Results of our experiments: AUROC classifier task performance for multiple random seeds for each *MoleculeNet* dataset, reported for each pre-training experiment and baseline model/dataset.

Experiment: Can AI learn from chemistry?

- **GraphCL** (You et al. 2020) contrastive pre-training uses random node dropping and random subgraphs:

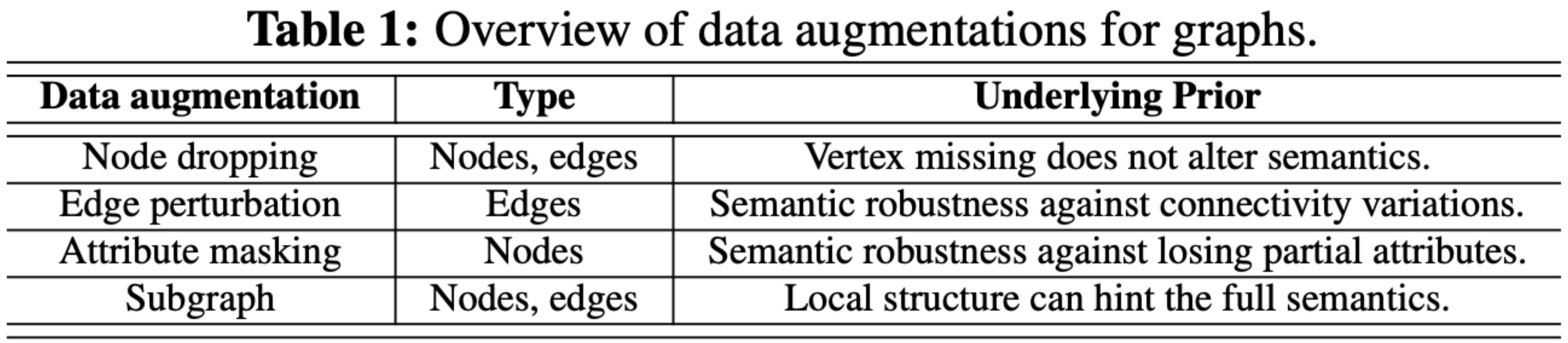


Table 1: Overview of data augmentations for graphs.

Data augmentation	Type	Underlying Prior
Node dropping	Nodes, edges	Vertex missing does not alter semantics.
Edge perturbation	Edges	Semantic robustness against connectivity variations.
Attribute masking	Nodes	Semantic robustness against losing partial attributes.
Subgraph	Nodes, edges	Local structure can hint the full semantics.

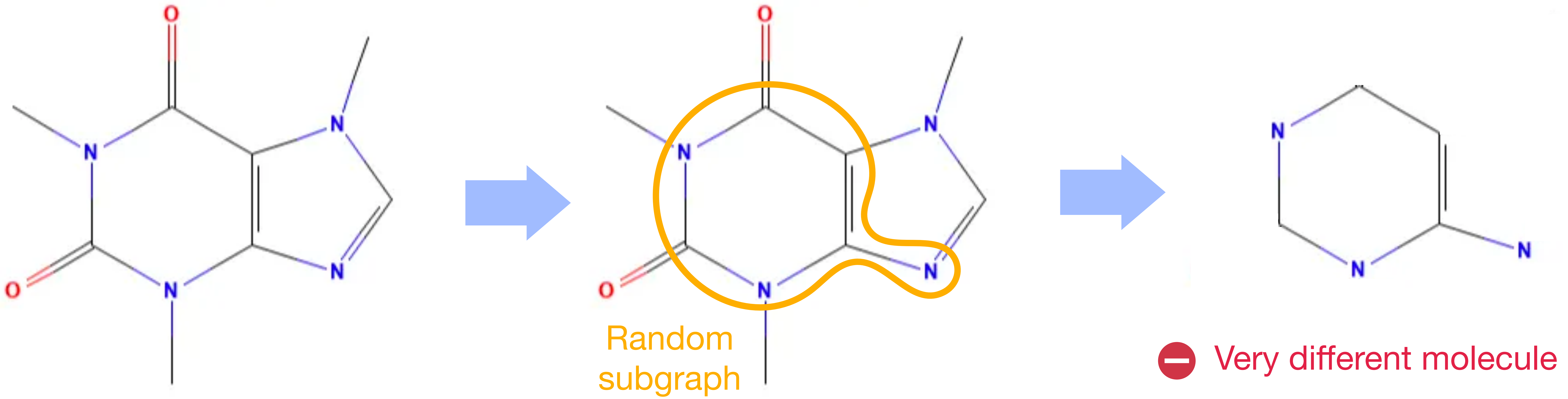
⊕ GraphCL GIN reached SOTA for unsupervised learning

⊖ No guarantee that augmented graphs are valid molecules!

You et al. 2020: https://arxiv.org/abs/2010.13902

Random graph augmentations can lead to strong contrasts in chemical space

- *Ex:* random subgraph.

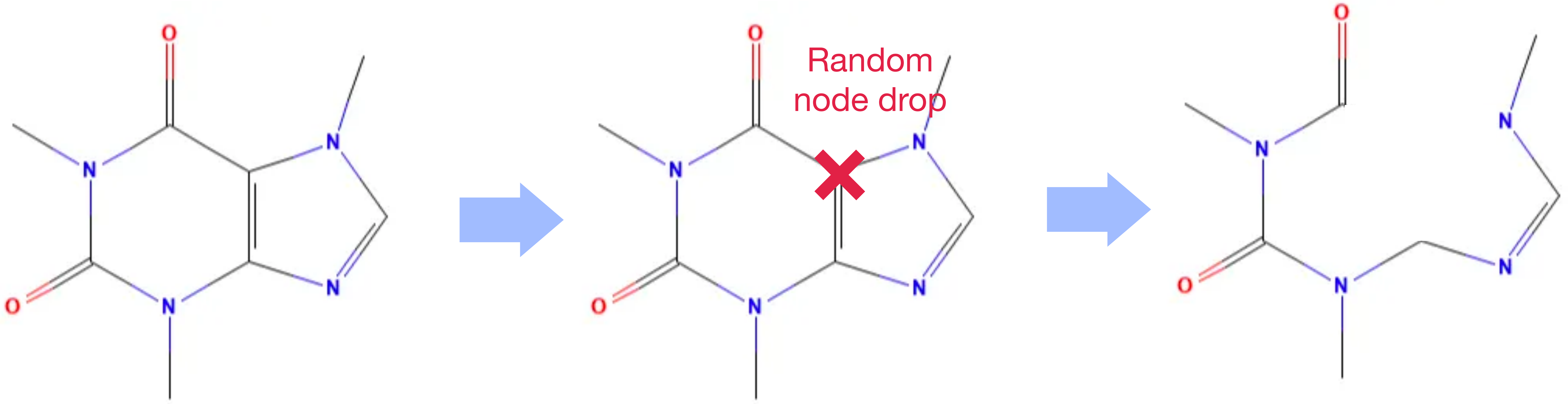


Random subgraph

⛔ Very different molecule

Random graph augmentations can lead to strong contrasts in chemical space

- *Ex:* drop random atom.

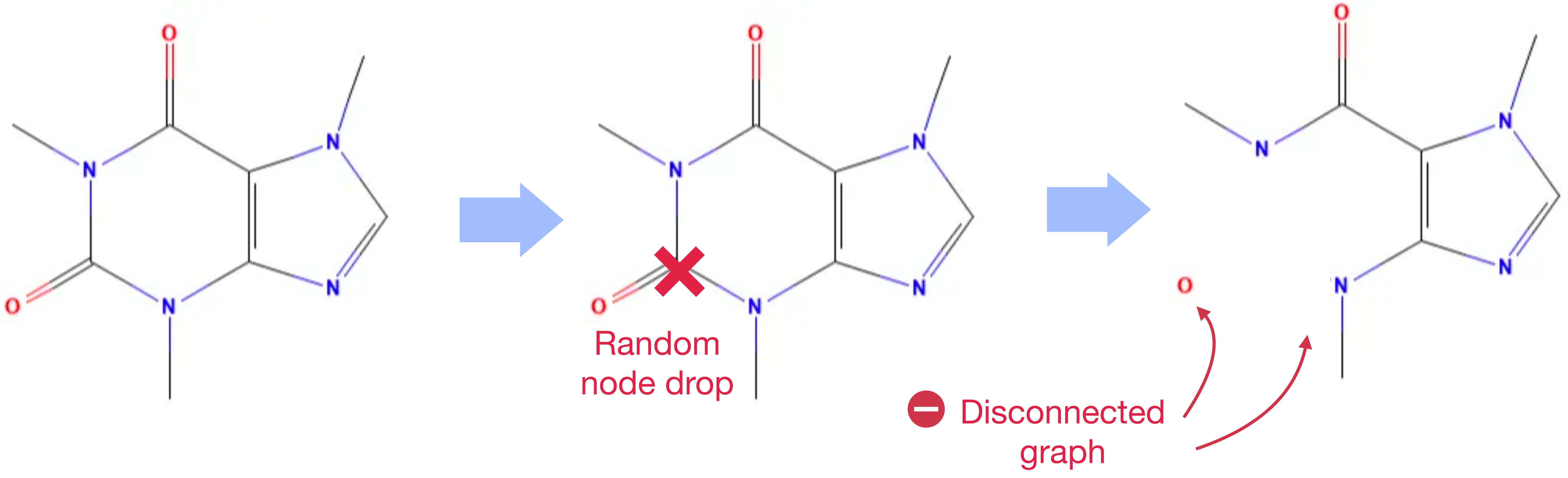


Random node drop

⊖ Very different molecule

Random graph augmentations can lead to invalid molecules

- *Ex:* drop random atom.



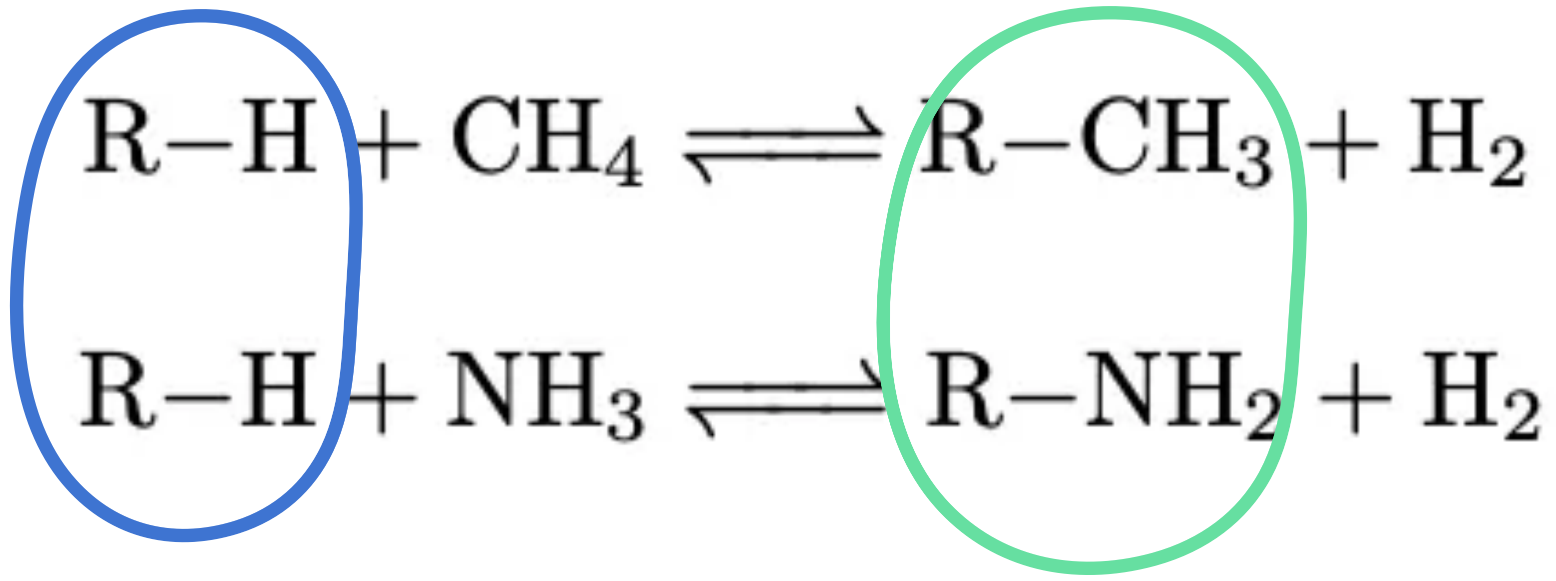
Random node drop

Disconnected graph

What if we used organic reactions as graph augmentations?

Idea: use addition/elimination organic reactions!

Transform initial graph into better behaved augmentations



$$\mathrm{R{-}H} + \mathrm{CH_4} \rightleftharpoons \mathrm{R{-}CH_3} + \mathrm{H_2}$$

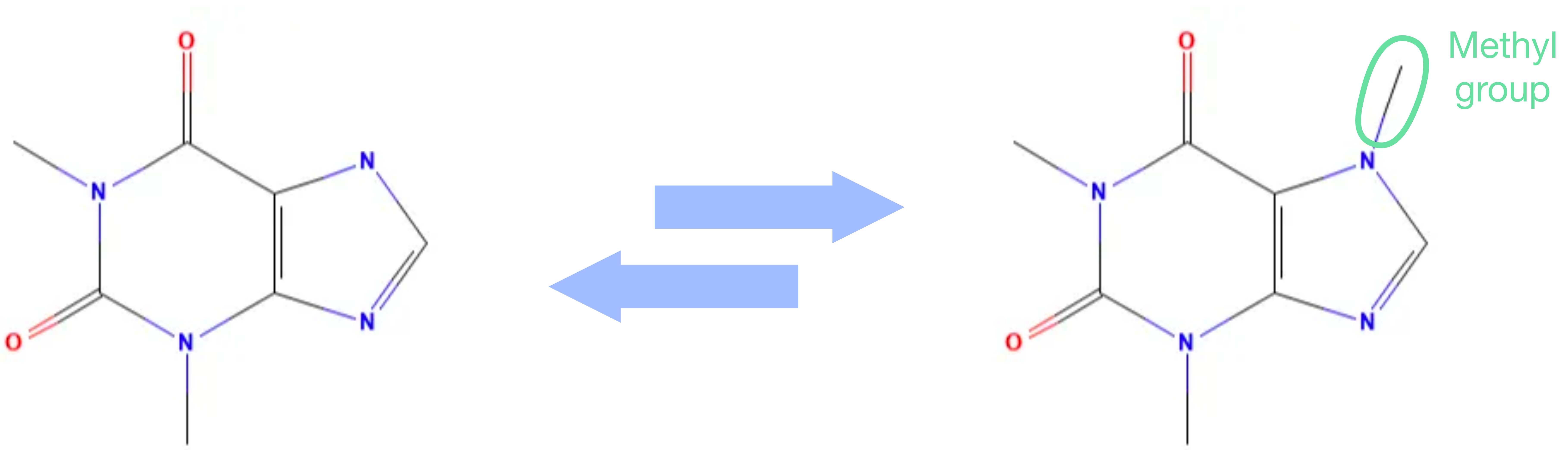
$$\mathrm{R{-}H} + \mathrm{NH_3} \rightleftharpoons \mathrm{R{-}NH_2} + \mathrm{H_2}$$

Initial molecule

➕ Valid augmented molecules

What if we used organic reactions as graph augmentations?

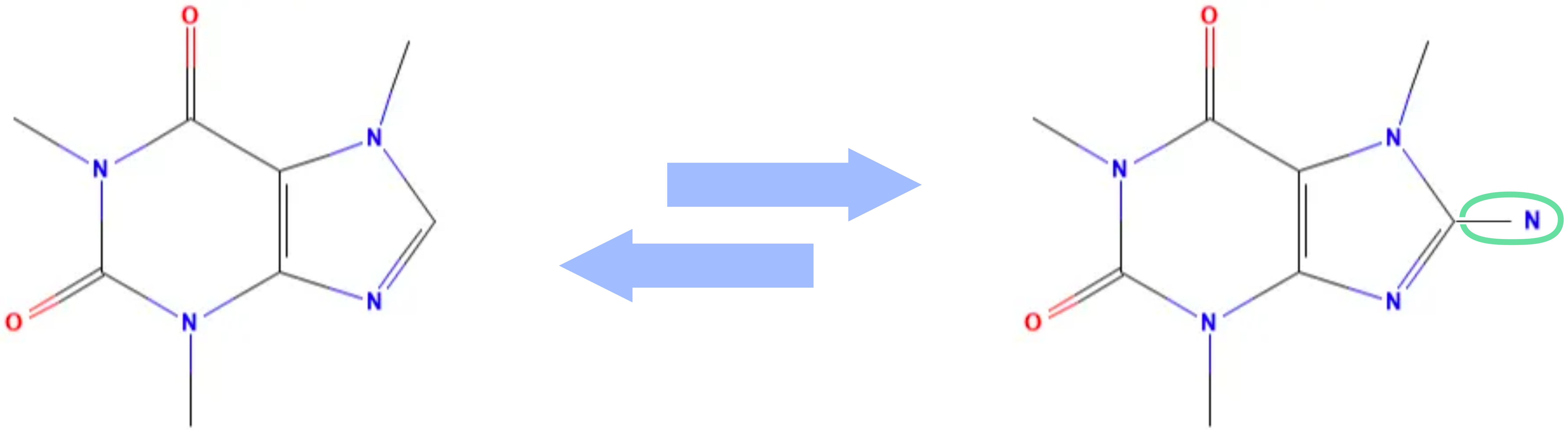
- *Ex:* methylation/de-methylation. $\mathrm{R{-}H} + \mathrm{CH_4} \rightleftharpoons \mathrm{R{-}CH_3} + \mathrm{H_2}$



➕ Valid + close to original molecule

What if we used organic reactions as graph augmentations?

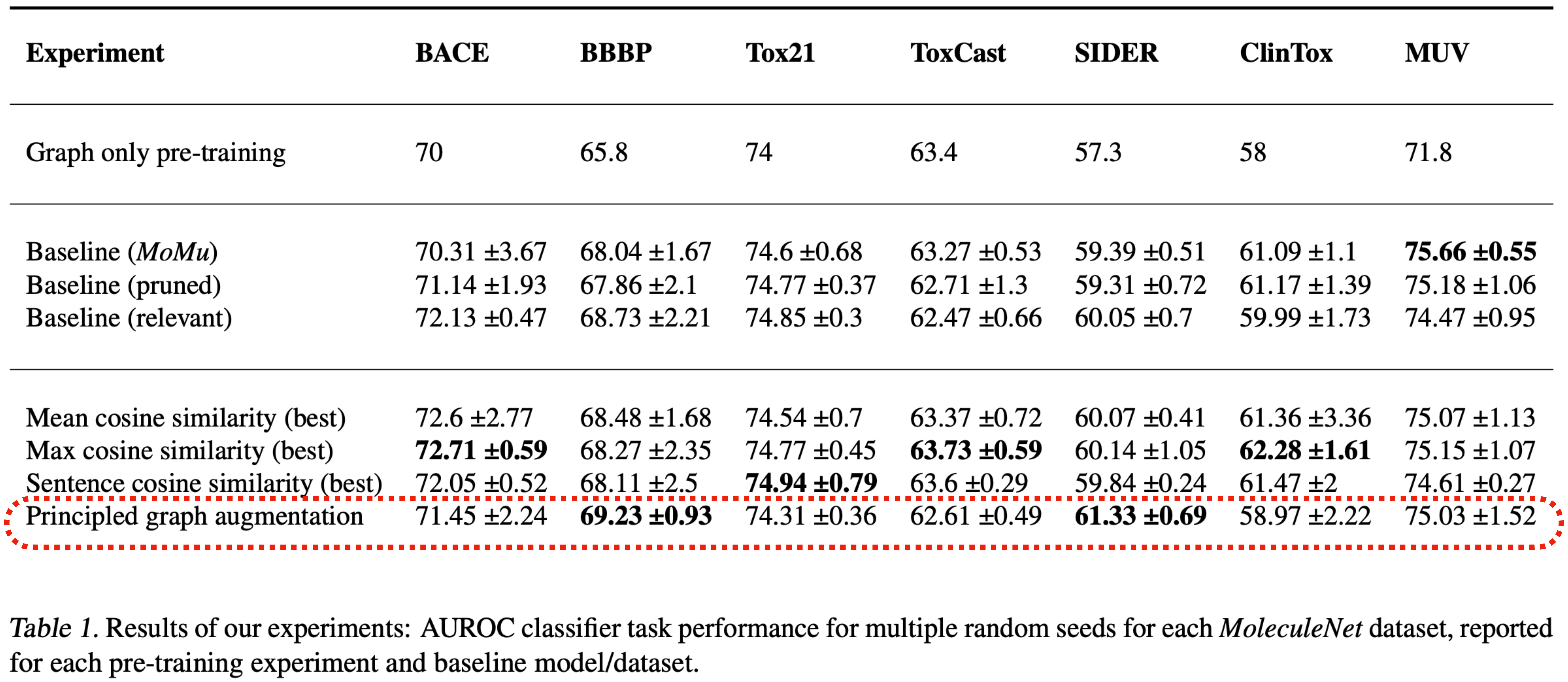
- *Ex:* amination/de-amination. $\mathrm{R{-}H + NH_3 \rightleftharpoons R{-}NH_2 + H_2}$



Amine group

➕ Valid + close to original molecule

Results



	Experiment	BACE	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV
Graph only	Graph only pre-training	70	65.8	74	63.4	57.3	58	71.8
Graph + natural language text	Baseline (*MoMu*)	70.31 ±3.67	68.04 ±1.67	74.6 ±0.68	63.27 ±0.53	59.39 ±0.51	61.09 ±1.1	**75.66 ±0.55**
	Baseline (pruned)	71.14 ±1.93	67.86 ±2.1	74.77 ±0.37	62.71 ±1.3	59.31 ±0.72	61.17 ±1.39	75.18 ±1.06
	Baseline (relevant)	72.13 ±0.47	68.73 ±2.21	74.85 ±0.3	62.47 ±0.66	60.05 ±0.7	59.99 ±1.73	74.47 ±0.95
	Mean cosine similarity (best)	72.6 ±2.77	68.48 ±1.68	74.54 ±0.7	63.37 ±0.72	60.07 ±0.41	61.36 ±3.36	75.07 ±1.13
	Max cosine similarity (best)	**72.71 ±0.59**	68.27 ±2.35	74.77 ±0.45	**63.73 ±0.59**	60.14 ±1.05	**62.28 ±1.61**	75.15 ±1.07
	Sentence cosine similarity (best)	72.05 ±0.52	68.11 ±2.5	**74.94 ±0.79**	63.6 ±0.29	59.84 ±0.24	61.47 ±2	74.61 ±0.27
	Principled graph augmentation	71.45 ±2.24	**69.23 ±0.93**	74.31 ±0.36	62.61 ±0.49	**61.33 ±0.69**	58.97 ±2.22	75.03 ±1.52

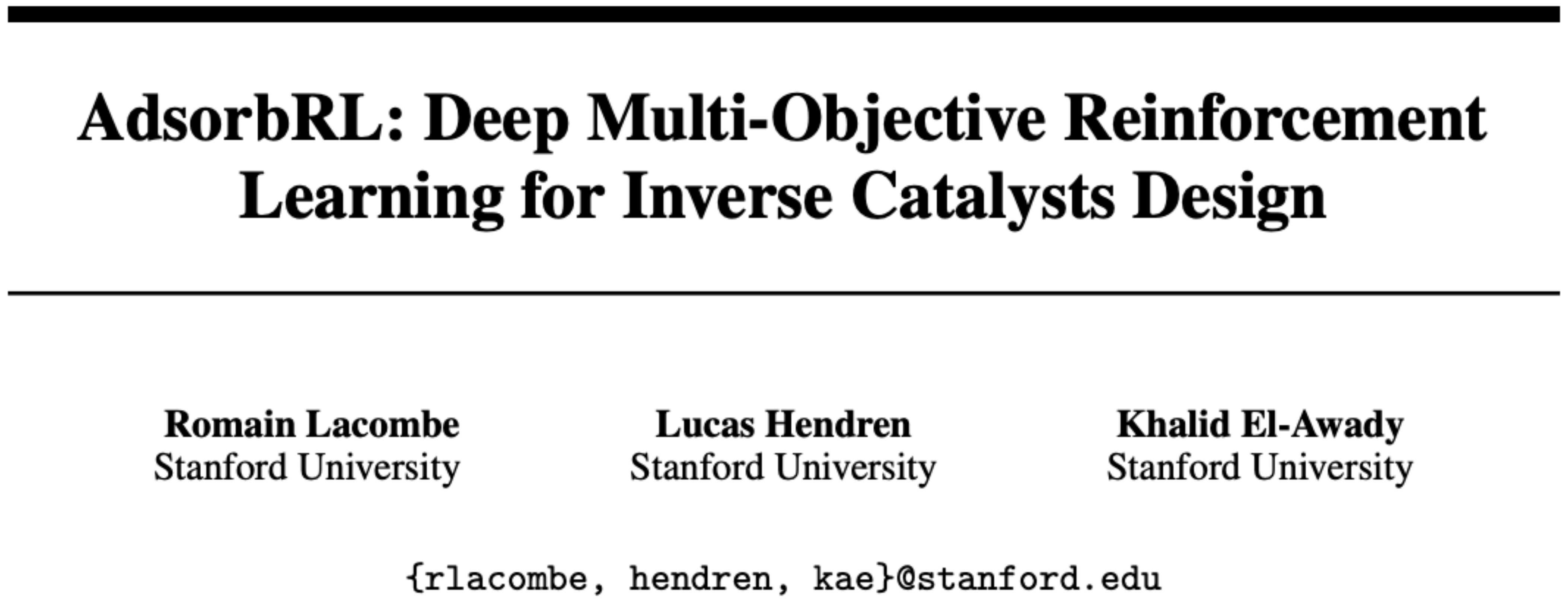
Table 1. Results of our experiments: AUROC classifier task performance for multiple random seeds for each *MoleculeNet* dataset, reported for each pre-training experiment and baseline model/dataset.

Catalysts discovery with reinforcement learning

https://arxiv.org/abs/2312.02308

RL for catalysts discovery

AI can master Go. What about materials?



AdsorbRL: Deep Multi-Objective Reinforcement Learning for Inverse Catalysts Design

Romain Lacombe
Stanford University

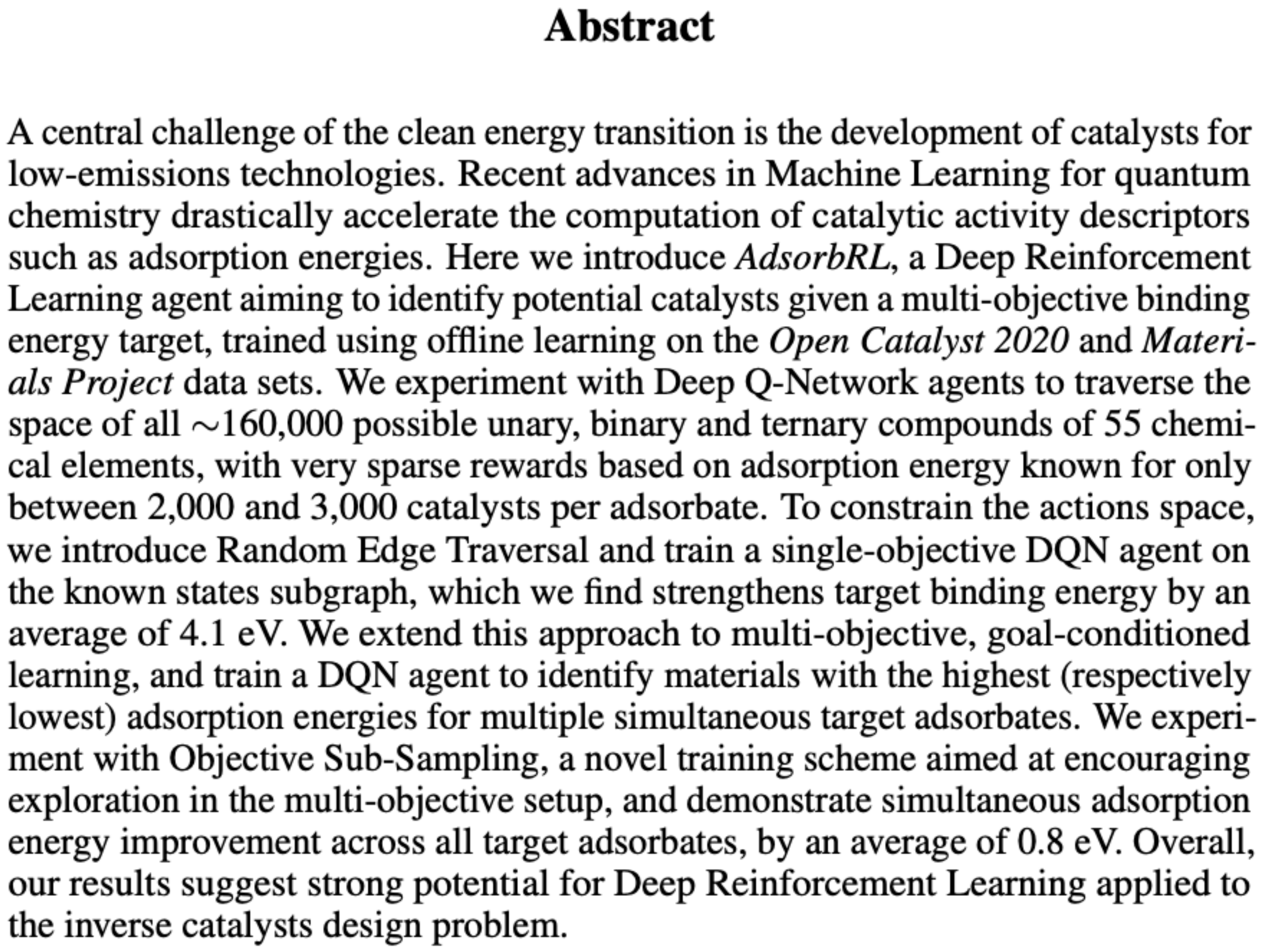
Lucas Hendren
Stanford University

Khalid El-Awady
Stanford University

{rlacombe, hendren, kae}@stanford.edu

NeurIPS 2023 AI for Accelerated Materials Design workshop

https://arxiv.org/abs/2312.02308



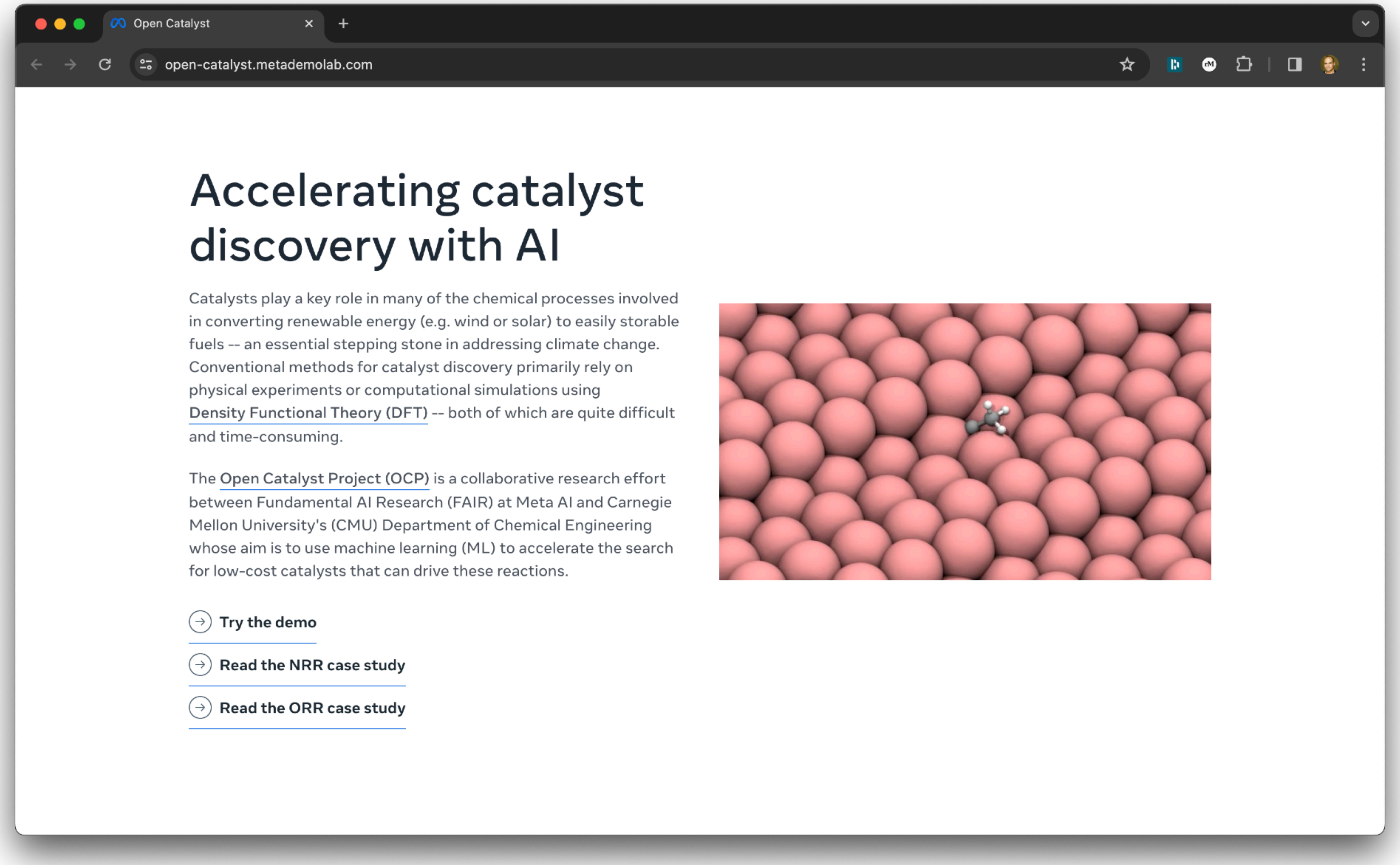
Abstract

A central challenge of the clean energy transition is the development of catalysts for low-emissions technologies. Recent advances in Machine Learning for quantum chemistry drastically accelerate the computation of catalytic activity descriptors such as adsorption energies. Here we introduce *AdsorbRL*, a Deep Reinforcement Learning agent aiming to identify potential catalysts given a multi-objective binding energy target, trained using offline learning on the *Open Catalyst 2020* and *Materials Project* data sets. We experiment with Deep Q-Network agents to traverse the space of all ∼160,000 possible unary, binary and ternary compounds of 55 chemical elements, with very sparse rewards based on adsorption energy known for only between 2,000 and 3,000 catalysts per adsorbate. To constrain the actions space, we introduce Random Edge Traversal and train a single-objective DQN agent on the known states subgraph, which we find strengthens target binding energy by an average of 4.1 eV. We extend this approach to multi-objective, goal-conditioned learning, and train a DQN agent to identify materials with the highest (respectively lowest) adsorption energies for multiple simultaneous target adsorbates. We experiment with Objective Sub-Sampling, a novel training scheme aimed at encouraging exploration in the multi-objective setup, and demonstrate simultaneous adsorption energy improvement across all target adsorbates, by an average of 0.8 eV. Overall, our results suggest strong potential for Deep Reinforcement Learning applied to the inverse catalysts design problem.

Open Catalyst Project

Very large DFT dataset

- OC20 and OC22 datasets by Meta AI and Carnegie Mellon
- 1.3 million molecular relaxations from over 260 million DFT calculations.
- Challenge and leaderboard
- Current lead: **~0.3 eV** MAE.



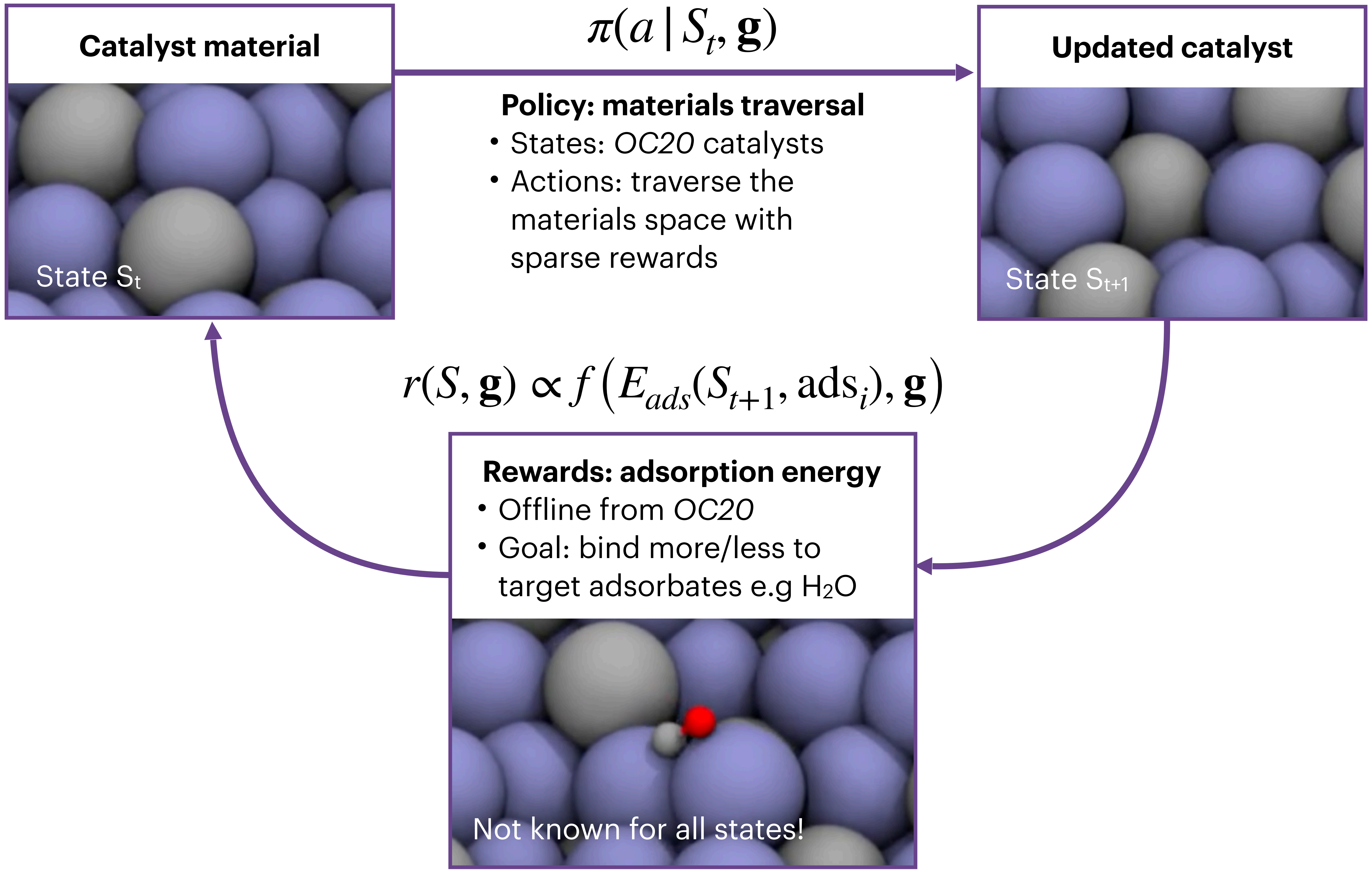
Accelerating catalyst discovery with AI

Catalysts play a key role in many of the chemical processes involved in converting renewable energy (e.g. wind or solar) to easily storable fuels -- an essential stepping stone in addressing climate change. Conventional methods for catalyst discovery primarily rely on physical experiments or computational simulations using Density Functional Theory (DFT) -- both of which are quite difficult and time-consuming.

The Open Catalyst Project (OCP) is a collaborative research effort between Fundamental AI Research (FAIR) at Meta AI and Carnegie Mellon University's (CMU) Department of Chemical Engineering whose aim is to use machine learning (ML) to accelerate the search for low-cost catalysts that can drive these reactions.

- Try the demo
- Read the NRR case study
- Read the ORR case study

RL for catalysts exploration



Catalyst material

State S_t

$\pi(a \mid S_t, \mathbf{g})$

Policy: materials traversal

- States: *OC20* catalysts
- Actions: traverse the materials space with sparse rewards

Updated catalyst

State S_{t+1}

$r(S, \mathbf{g}) \propto f\left(E_{ads}(S_{t+1}, \text{ads}_i), \mathbf{g}\right)$

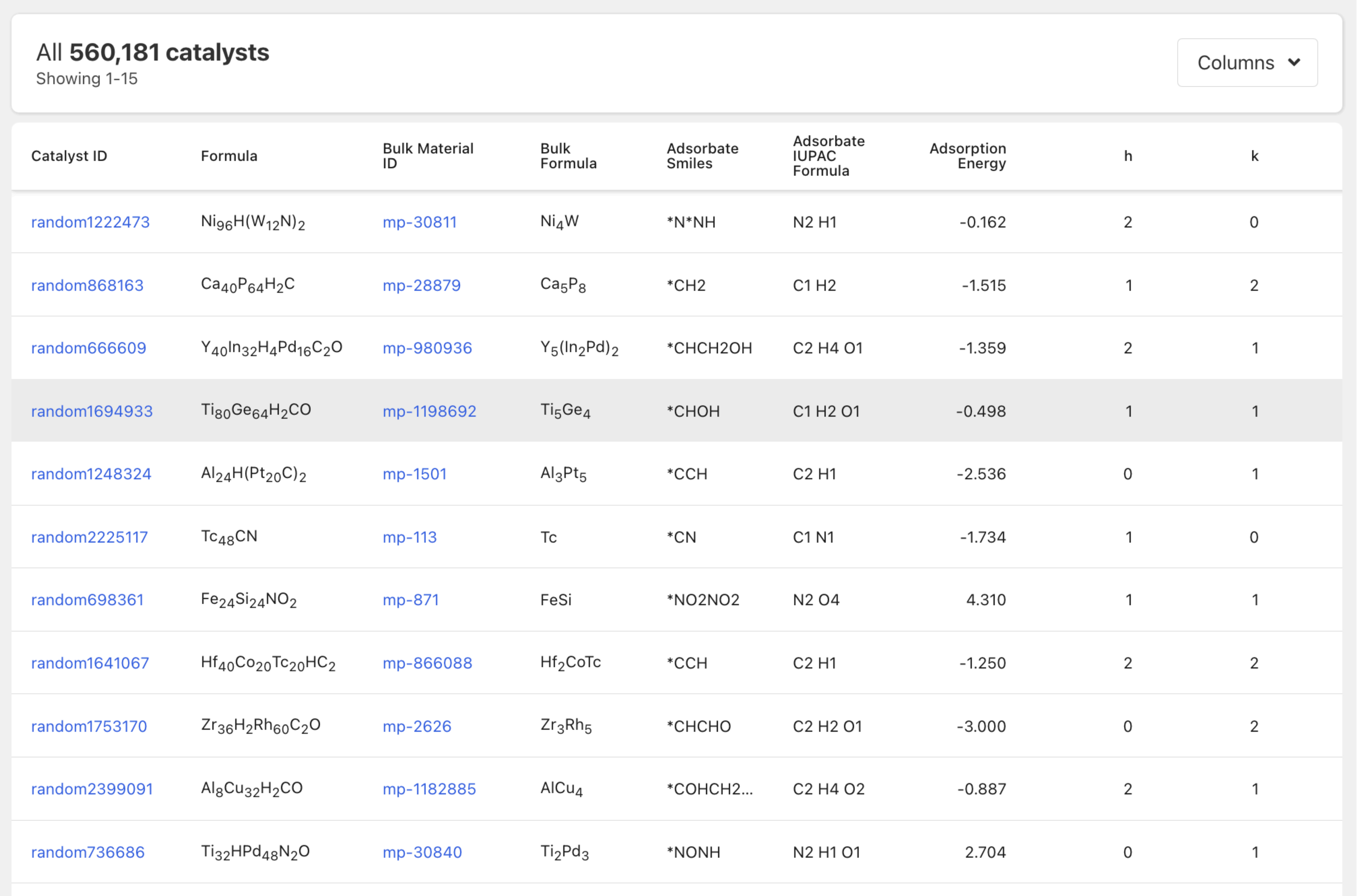
Rewards: adsorption energy

- Offline from *OC20*
- Goal: bind more/less to target adsorbates e.g H_2O

Not known for all states!

Simplifying the problem

- **Data set:** OC20 adsorption energies for (catalyst, adsorbate) pairs
- **States:** ~160,000 unary, binary and ternary compounds of 55 elements (ignoring stoichiometry).
- **Actions:** steps to traverse the dataset of materials.
- **Goals:** targets for each adsorbate (strong binding/low energy vs. weak binding/high energy)
- **Rewards** are functions of adsorption energy of catalysts for target adsorbates.



All **560,181 catalysts**

Showing 1-15

Catalyst ID	Formula	Bulk Material ID	Bulk Formula	Adsorbate Smiles	Adsorbate IUPAC Formula	Adsorption Energy	h	k
random1222473	$Ni_{96}H(W_{12}N)_2$	mp-30811	Ni_4W	*N*NH	N2 H1	-0.162	2	0
random868163	$Ca_{40}P_{64}H_2C$	mp-28879	Ca_5P_8	*CH2	C1 H2	-1.515	1	2
random666609	$Y_{40}In_{32}H_4Pd_{16}C_2O$	mp-980936	$Y_5(In_2Pd)_2$	*CHCH2OH	C2 H4 O1	-1.359	2	1
random1694933	$Ti_{80}Ge_{64}H_2CO$	mp-1198692	Ti_5Ge_4	*CHOH	C1 H2 O1	-0.498	1	1
random1248324	$Al_{24}H(Pt_{20}C)_2$	mp-1501	Al_3Pt_5	*CCH	C2 H1	-2.536	0	1
random2225117	$Tc_{48}CN$	mp-113	Tc	*CN	C1 N1	-1.734	1	0
random698361	$Fe_{24}Si_{24}NO_2$	mp-871	FeSi	*NO2NO2	N2 O4	4.310	1	1
random1641067	$Hf_{40}Co_{20}Tc_{20}HC_2$	mp-866088	Hf_2CoTc	*CCH	C2 H1	-1.250	2	2
random1753170	$Zr_{36}H_2Rh_{60}C_2O$	mp-2626	Zr_3Rh_5	*CHCHO	C2 H2 O1	-3.000	0	2
random2399091	$Al_8Cu_{32}H_2CO$	mp-1182885	$AlCu_4$	*COHCH2...	C2 H4 O2	-0.887	2	1
random736686	$Ti_{32}HPd_{48}N_2O$	mp-30840	Ti_2Pd_3	*NONH	N2 H1 O1	2.704	0	1

Reinforcement learning setting

The math behind training.

Multi-objective Goal-conditioned Deep Q-Network

S: compounds

a: actions

g: goal vector (+1, or -1 for adsorbate *i*)

r: reward (f(E_{ads}))

Bellman Equation for Q-learning:

$$\mathrm{Q}^*(a\,|\,S,\mathbf{g}) = r(a\,|\,S,\mathbf{g}) + \gamma \max_{a} \left(\mathrm{Q}^*(a\,|\,S',\mathbf{g})\right)$$

Evaluation metric Δ:

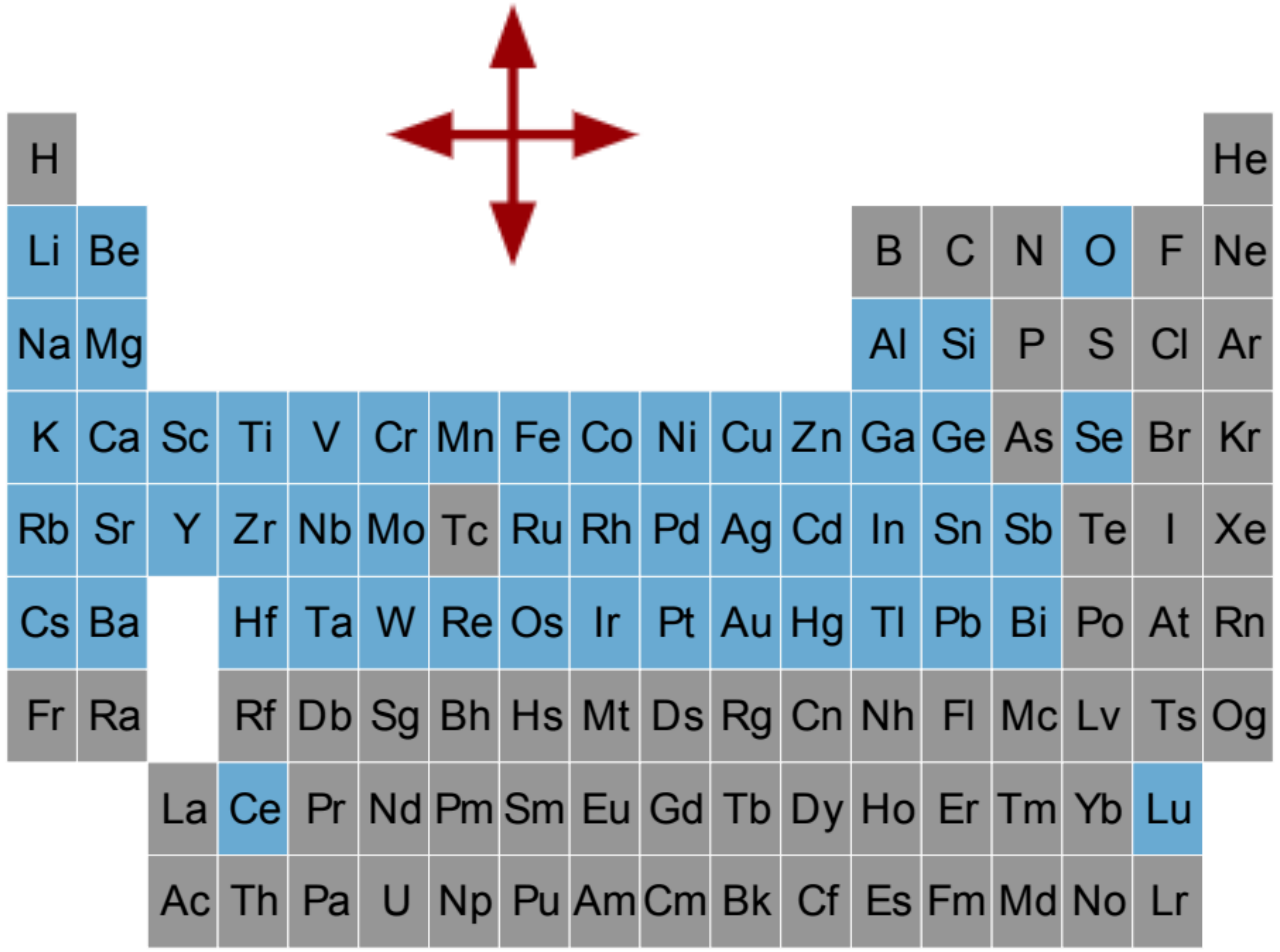
$$\Delta = \frac{1}{N}\sum_{i\in\text{final}} -\left(E_{ads}(S_i)\right) - \frac{1}{N}\sum_{j\in\text{initial}} -\left(E_{ads}(S_j)\right)$$

Learn more: **CS 224R**

Unary compounds

Navigating the Periodic Table.

- 86 single element states
- 5 actions: { _ | ← | → | ↓ | ↑ }
- Goal: strong binding (minimize E_{ads})
- Simple Q-learning reaches top-2 states for ~95% of roll-outs.
- High performance agent: **Δ = -5.9 eV.**



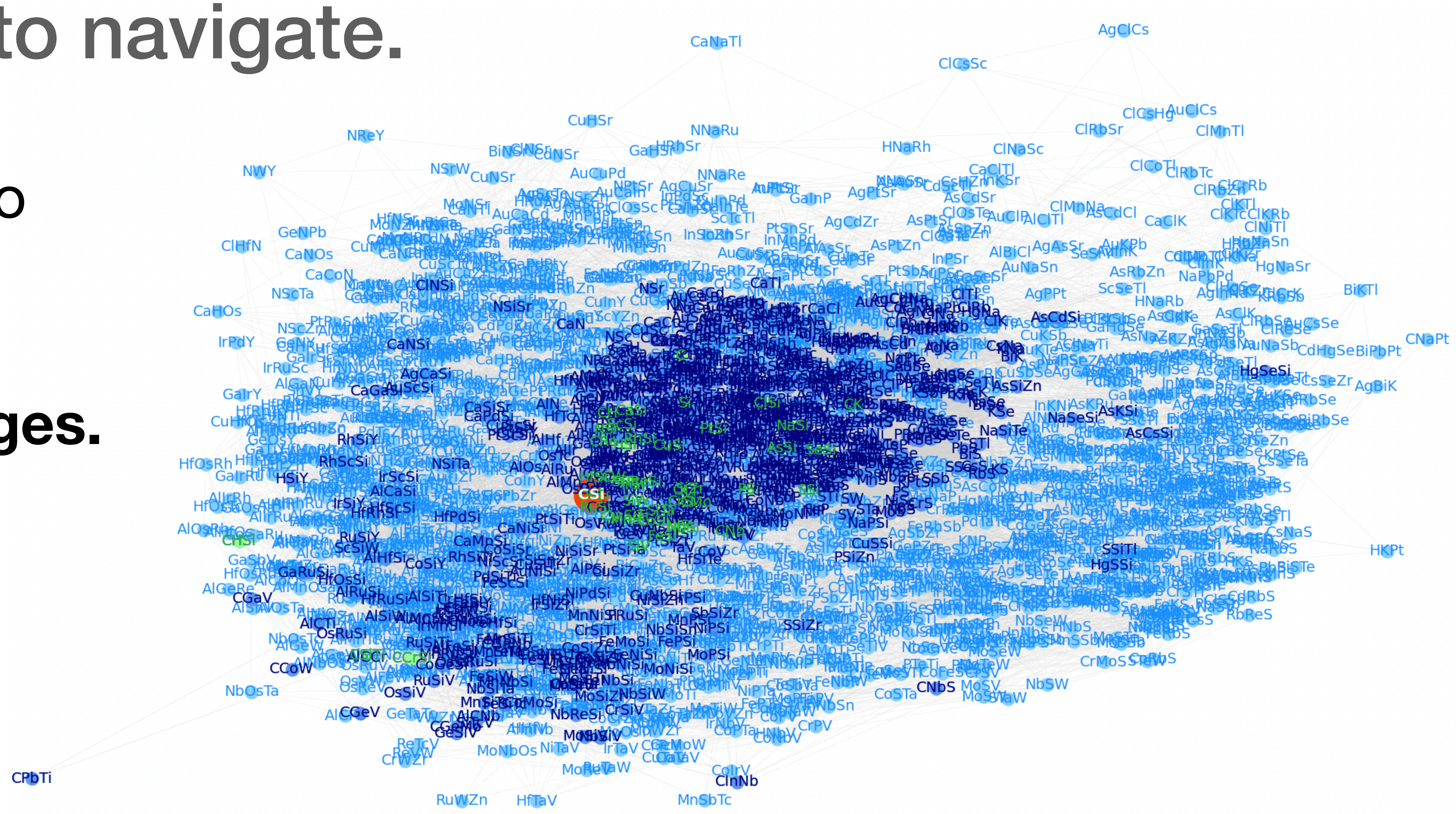
H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba		Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra		Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Cn	Nh	Fl	Mc	Lv	Ts	Og
		La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	
		Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr	

55 elements in *OC20*

Compounds: random edge traversal

Larger, sparser dataset, hard to navigate.

- Insight: constrain states and actions to make DQN learning more tractable
- **Only traverse known energy states.**
- **Traverse subgraph with random edges.** Learn only 5 actions:
  - <stop>
  - <add> a random element
  - <delete> element 1, 2, or 3.
- High performance agent: **Δ = 4.1 eV.**



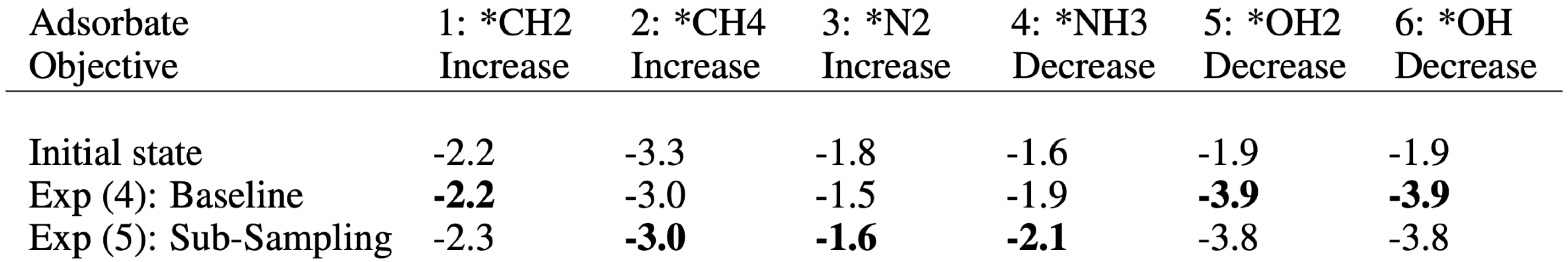
3-hop ego graph of lowest energy state for *OH_2 adsorbate (SiC)

Multi-objective setting

Different targets for different adsorbates.

- **Learn 6 objectives at once!**
  - Increase E_{ads} for some adsorbates
  - Decrease E_{ads} for others
- Multi-objective DQN
- Simultaneously improves adsorption energy in the desired direction by **Δ = 0.8 eV** on average across all 6 adsorbates.

Motivation: could we **break linear scaling relationships?**



Adsorbate Objective	1: *CH2 Increase	2: *CH4 Increase	3: *N2 Increase	4: *NH3 Decrease	5: *OH2 Decrease	6: *OH Decrease
Initial state	-2.2	-3.3	-1.8	-1.6	-1.9	-1.9
Exp (4): Baseline	**-2.2**	-3.0	-1.5	-1.9	**-3.9**	**-3.9**
Exp (5): Sub-Sampling	-2.3	**-3.0**	**-1.6**	**-2.1**	-3.8	-3.8

Conclusions & Future work

RL for generalized inverse catalyst design

Identify promising catalysts for any combination of target adsorbates:

- In practice: conduct a large number of roll-outs.
- **Most common terminal states are promising materials on which to focus computational and experimental resources.**

Future work

- Better handling of unknown energy states while traversing state space
- Scalar goal-conditioning to find compounds with any given target E_{ads}
- Actor-critic using *AdsorbML* [2], ML-based DFT for binding energies

Key References

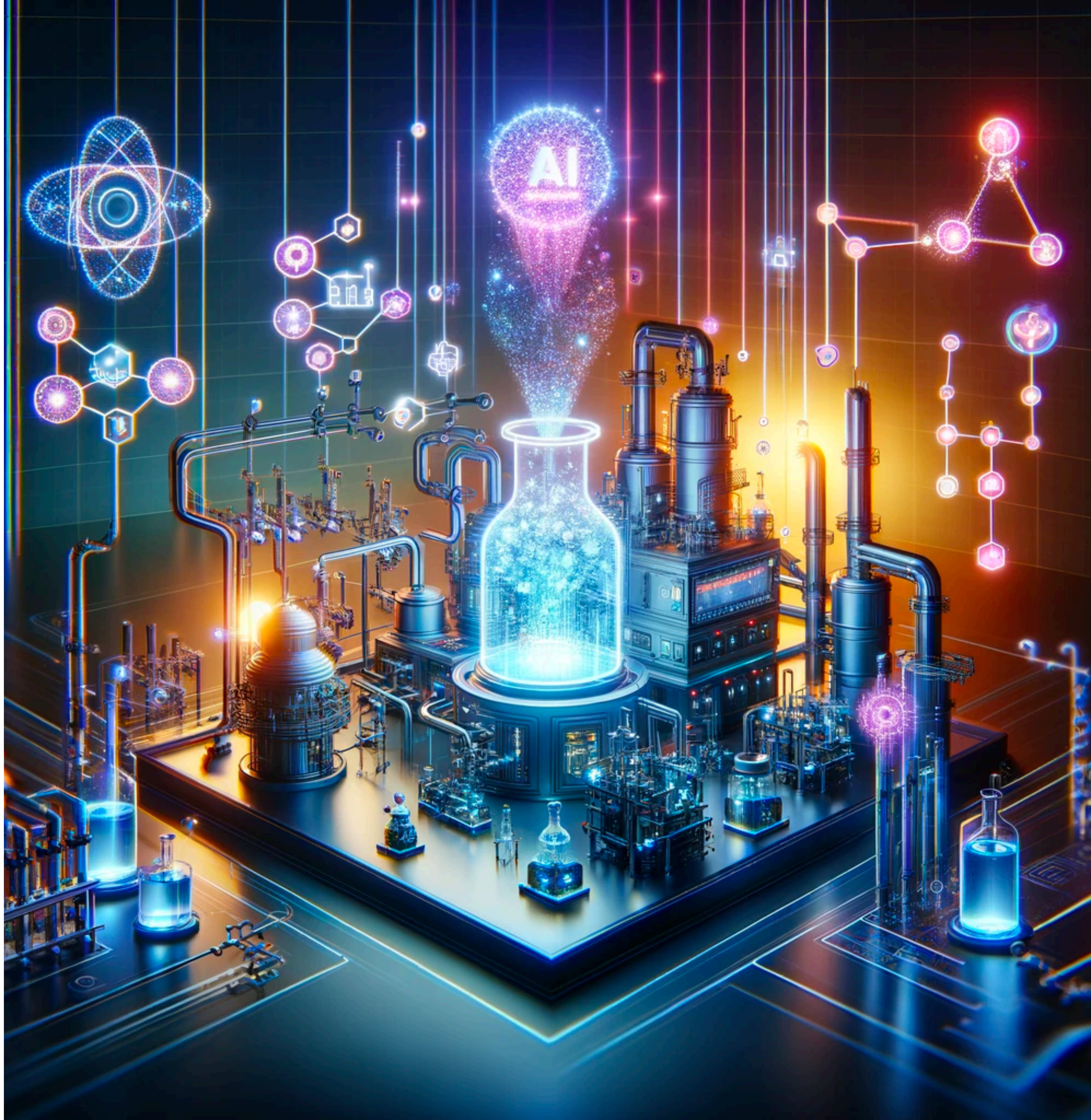
[1] Zitnick et al.: “**An Introduction to Electrocatalyst Design using Machine Learning for Renewable Energy Storage**”, 2020; arXiv:2010.09435.

[2] Lan et al. "**AdsorbML: Accelerating Adsorption Energy Calculations with Machine Learning.**" 2022; arXiv:2211.16486 (2022).

[3] **Materials Project**, https://materialsproject.org

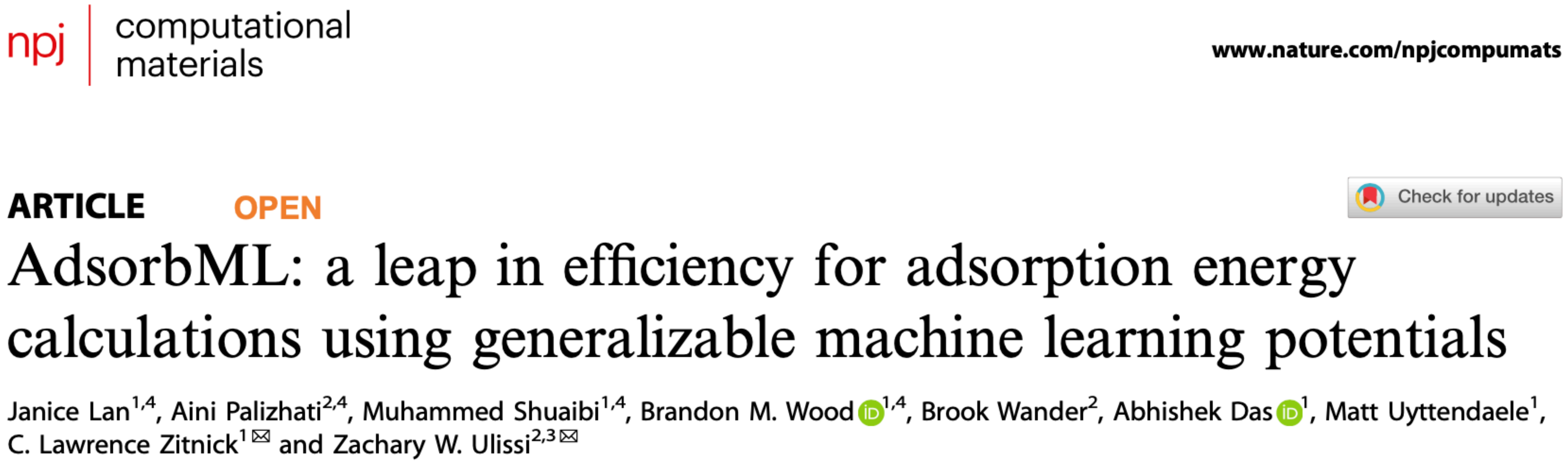
[4] **Open Catalyst Project**, https://opencatalystproject.org

Frontiers in AI for chemical engineering



ML for DFT computations

Accelerating simulations



npj | computational materials

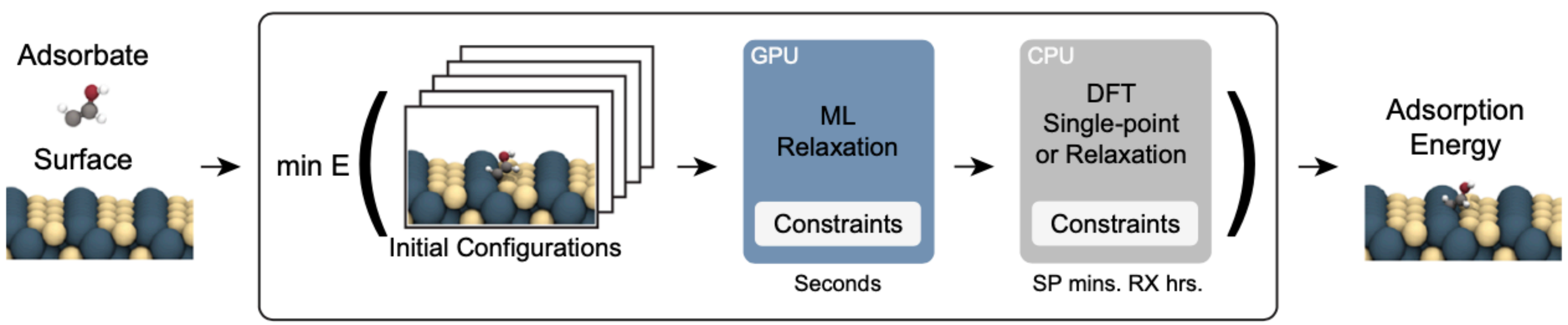
www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

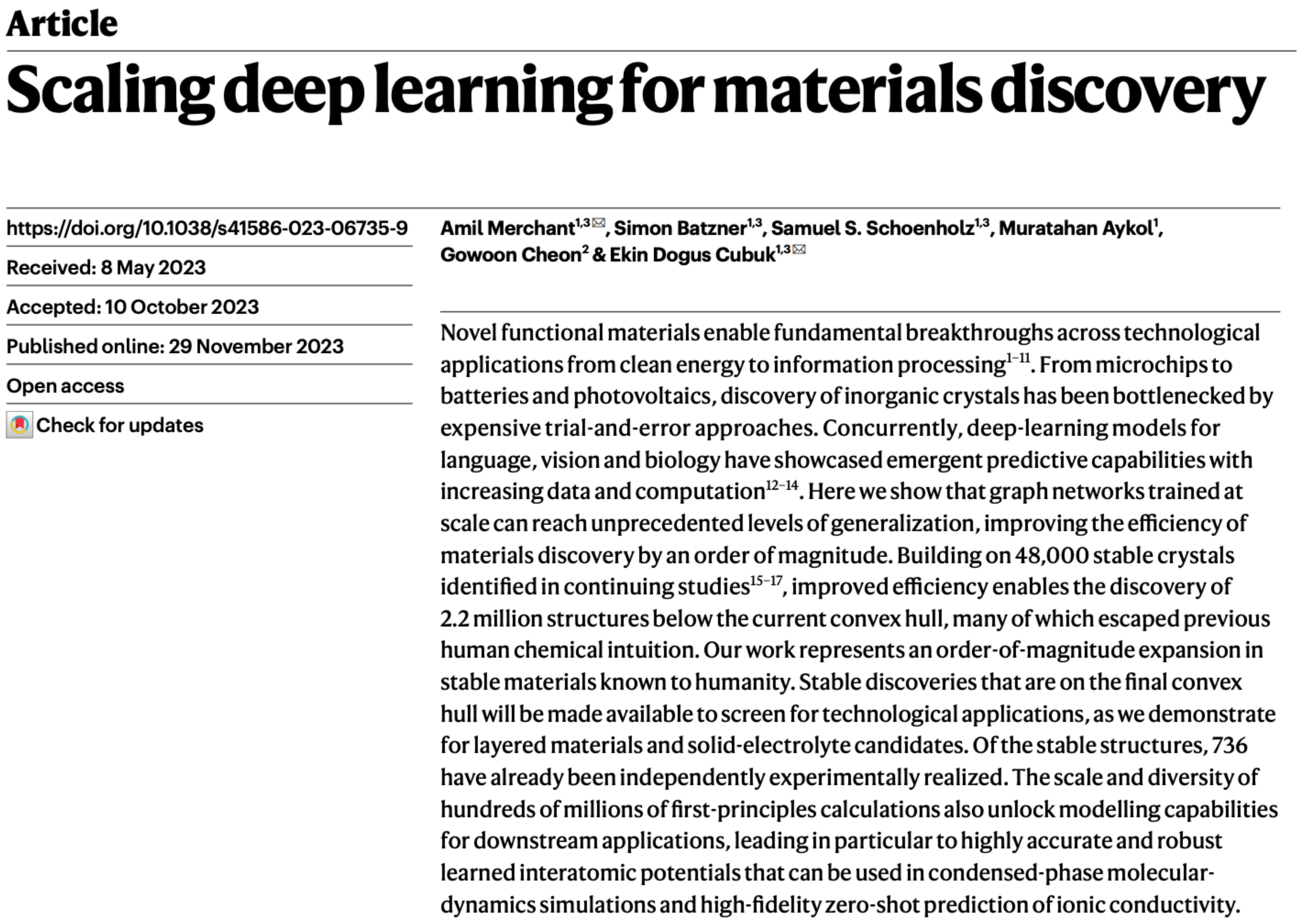
AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials

Janice Lan[1,4], Aini Palizhati[2,4], Muhammed Shuaibi[1,4], Brandon M. Wood[1,4], Brook Wander[2], Abhishek Das[1], Matt Uyttendaele[1], C. Lawrence Zitnick[1] ✉ and Zachary W. Ulissi[2,3] ✉



Materials generation

Generative AI for solid-state structures



Article

Scaling deep learning for materials discovery

https://doi.org/10.1038/s41586-023-06735-9

Received: 8 May 2023

Accepted: 10 October 2023

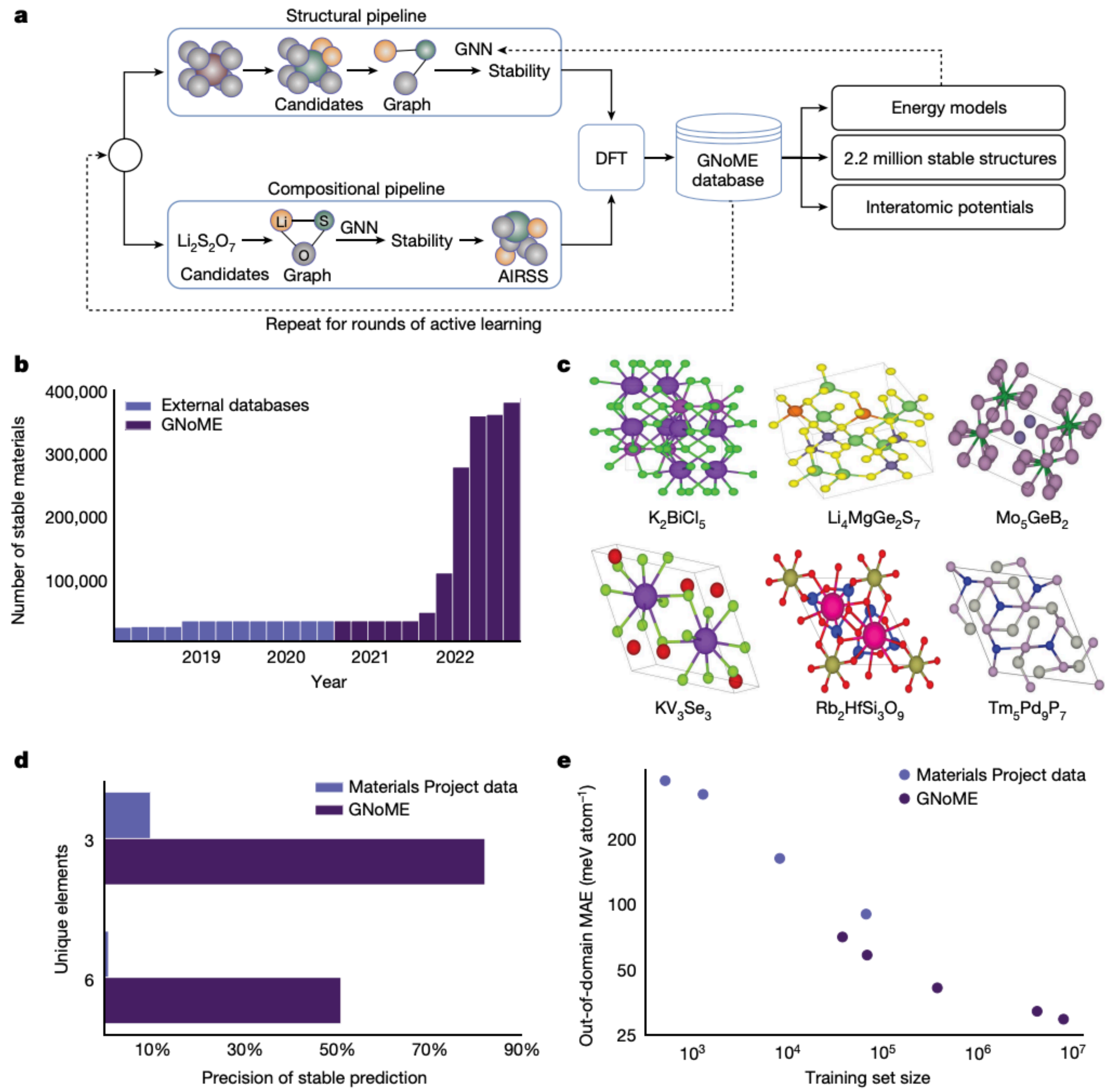
Published online: 29 November 2023

Open access

Check for updates

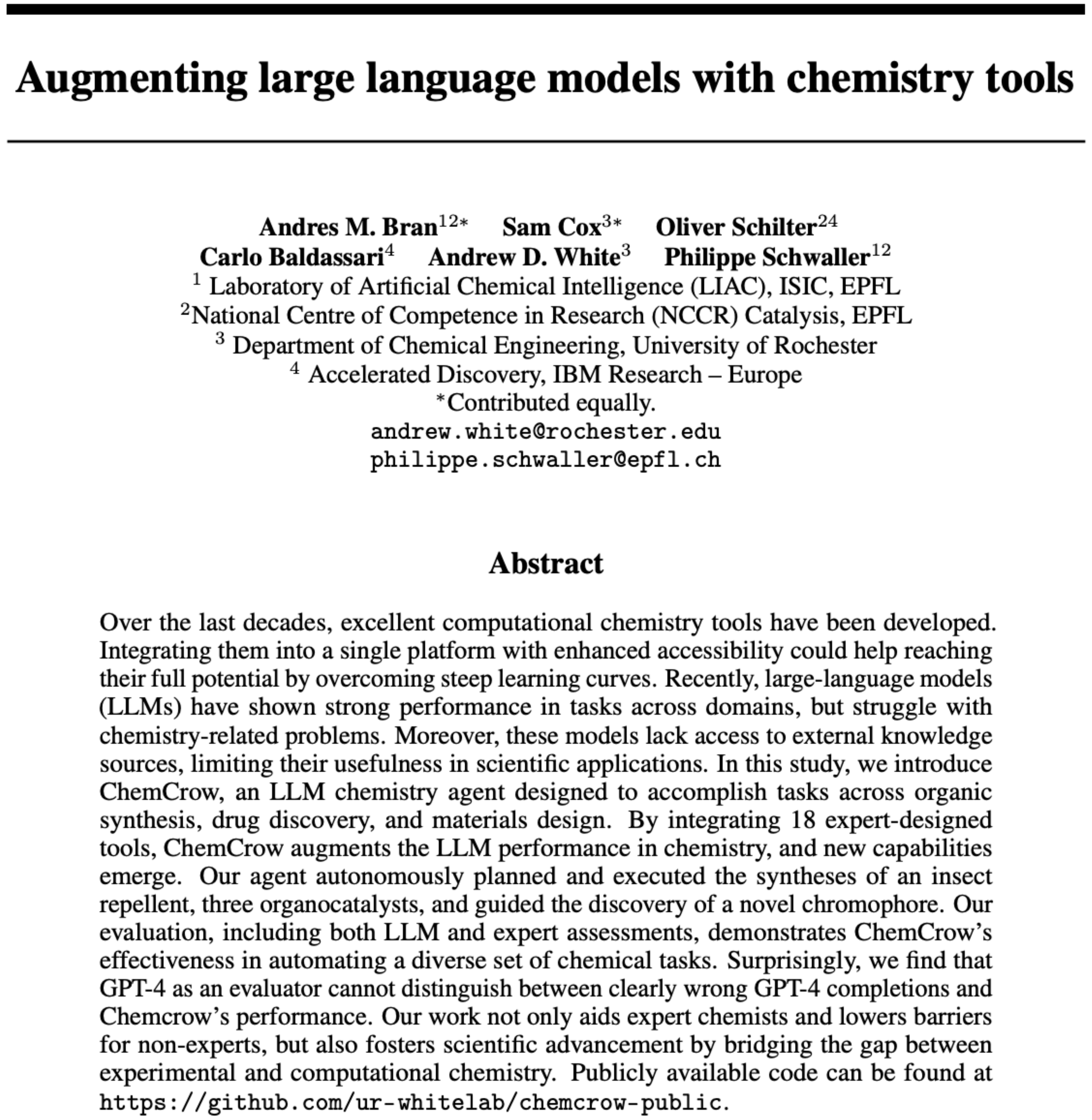
Amil Merchant[1,3]✉, Simon Batzner[1,3], Samuel S. Schoenholz[1,3], Muratahan Aykol[1], Gowoon Cheon[2] & Ekin Dogus Cubuk[1,3]✉

Novel functional materials enable fundamental breakthroughs across technological applications from clean energy to information processing[1–11]. From microchips to batteries and photovoltaics, discovery of inorganic crystals has been bottlenecked by expensive trial-and-error approaches. Concurrently, deep-learning models for language, vision and biology have showcased emergent predictive capabilities with increasing data and computation[12–14]. Here we show that graph networks trained at scale can reach unprecedented levels of generalization, improving the efficiency of materials discovery by an order of magnitude. Building on 48,000 stable crystals identified in continuing studies[15–17], improved efficiency enables the discovery of 2.2 million structures below the current convex hull, many of which escaped previous human chemical intuition. Our work represents an order-of-magnitude expansion in stable materials known to humanity. Stable discoveries that are on the final convex hull will be made available to screen for technological applications, as we demonstrate for layered materials and solid-electrolyte candidates. Of the stable structures, 736 have already been independently experimentally realized. The scale and diversity of hundreds of millions of first-principles calculations also unlock modelling capabilities for downstream applications, leading in particular to highly accurate and robust learned interatomic potentials that can be used in condensed-phase molecular-dynamics simulations and high-fidelity zero-shot prediction of ionic conductivity.



AI assistant

“CoPilot for Research”



Augmenting large language models with chemistry tools

Andres M. Bran[12*] **Sam Cox**[3*] **Oliver Schilter**[24]
Carlo Baldassari[4] **Andrew D. White**[3] **Philippe Schwaller**[12]
[1] Laboratory of Artificial Chemical Intelligence (LIAC), ISIC, EPFL
[2]National Centre of Competence in Research (NCCR) Catalysis, EPFL
[3] Department of Chemical Engineering, University of Rochester
[4] Accelerated Discovery, IBM Research – Europe
*Contributed equally.
andrew.white@rochester.edu
philippe.schwaller@epfl.ch

Abstract

Over the last decades, excellent computational chemistry tools have been developed. Integrating them into a single platform with enhanced accessibility could help reaching their full potential by overcoming steep learning curves. Recently, large-language models (LLMs) have shown strong performance in tasks across domains, but struggle with chemistry-related problems. Moreover, these models lack access to external knowledge sources, limiting their usefulness in scientific applications. In this study, we introduce ChemCrow, an LLM chemistry agent designed to accomplish tasks across organic synthesis, drug discovery, and materials design. By integrating 18 expert-designed tools, ChemCrow augments the LLM performance in chemistry, and new capabilities emerge. Our agent autonomously planned and executed the syntheses of an insect repellent, three organocatalysts, and guided the discovery of a novel chromophore. Our evaluation, including both LLM and expert assessments, demonstrates ChemCrow's effectiveness in automating a diverse set of chemical tasks. Surprisingly, we find that GPT-4 as an evaluator cannot distinguish between clearly wrong GPT-4 completions and Chemcrow's performance. Our work not only aids expert chemists and lowers barriers for non-experts, but also fosters scientific advancement by bridging the gap between experimental and computational chemistry. Publicly available code can be found at `https://github.com/ur-whitelab/chemcrow-public`.

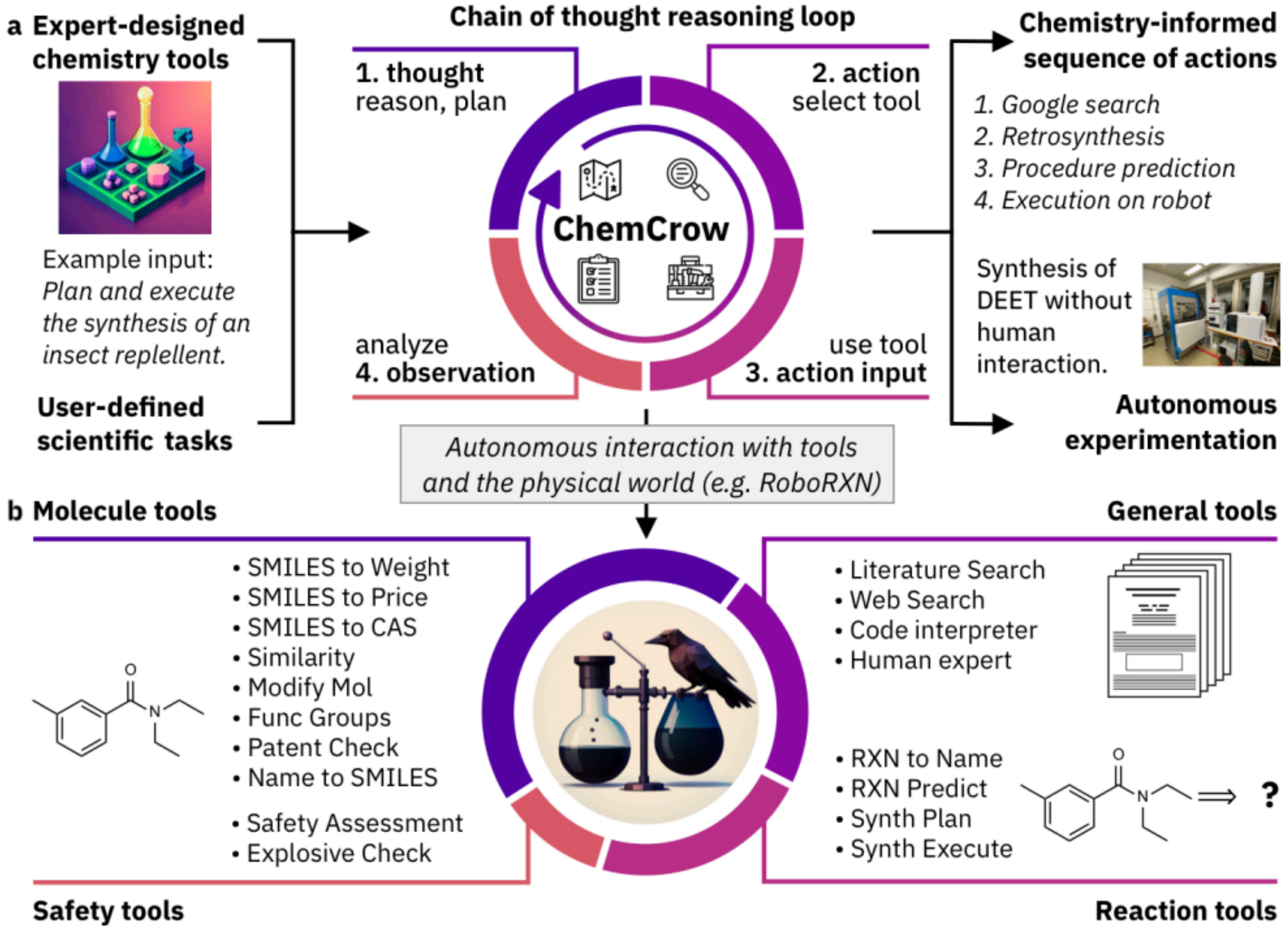


Figure 1: **Overview and toolset.** a) An overview of the task-solving process. Using a variety of chemistry-related packages and software, a set of tools is created. These tools and a user input are then given to an LLM. The LLM then proceeds through an automatic, iterative chain-of-thought process, deciding on its path, choice of tools, and inputs before coming to a final answer. The example shows the synthesis of DEET, a common insect repellent. b) Toolsets implemented in ChemCrow: reaction, molecule, safety, search, and standard tools.

Thank you!

Questions?

- Link to papers:
  https://arxiv.org/abs/2307.12996
  https://arxiv.org/abs/2312.02308

SCAN ME

- Follow up questions?
  rlacombe@stanford.edu
- **Please get in touch!**