Carnegie Mellon University

2025 Artificial Intelligence for Materials Science (AIMS) Workshop

Artificial Intelligence (AI) Solutions for Computational and Organic Chemistry

@olexandr

Olexandr Isayev

Department of Chemistry, Carnegie Mellon University

olexandr@olexandrisayev.com

http://olexandrisayev.com

Carnegie Mellon University

Postdocs:

Dylan Anstine
Peikun Zheng
Bhupalee Kalita
Tanya Zubatiuk
Hatice Gokcan

PhDs:

Phil Gusev
Polina Avdiunina
Ilkwon Cho
Nick Gao
Filipp Nikitin
Shuhao Zhang
Emma Bouchard

Isayev Lab circa 2024

Funding:

BASF We create chemistry

GOLD SCIENCE AMBASSADOR 2024 DASSAULT SYSTEMES

NSF

CHE-2154447

NSF CCI C-CAS

C-CAS Center for Computer Assisted Synthesis

Open Science Grid

ONR MURI N00014-21-1-2476

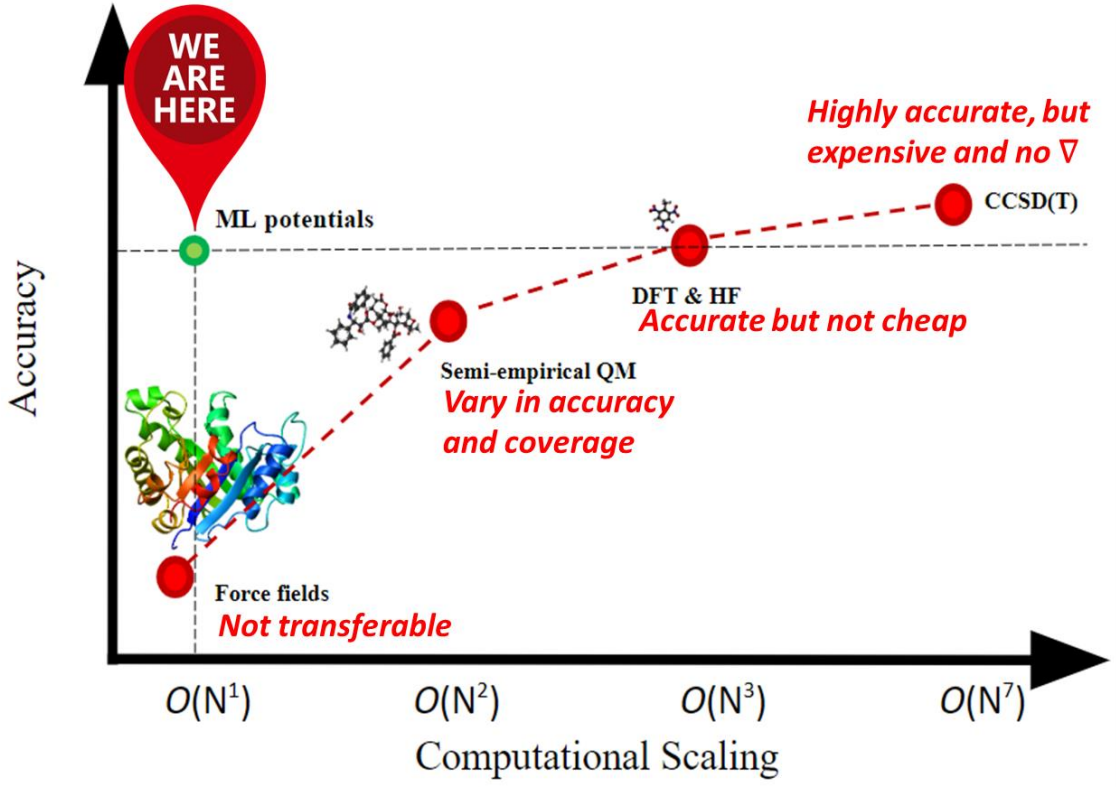
DEPARTMENT OF THE NAVY ONR Science & Technology

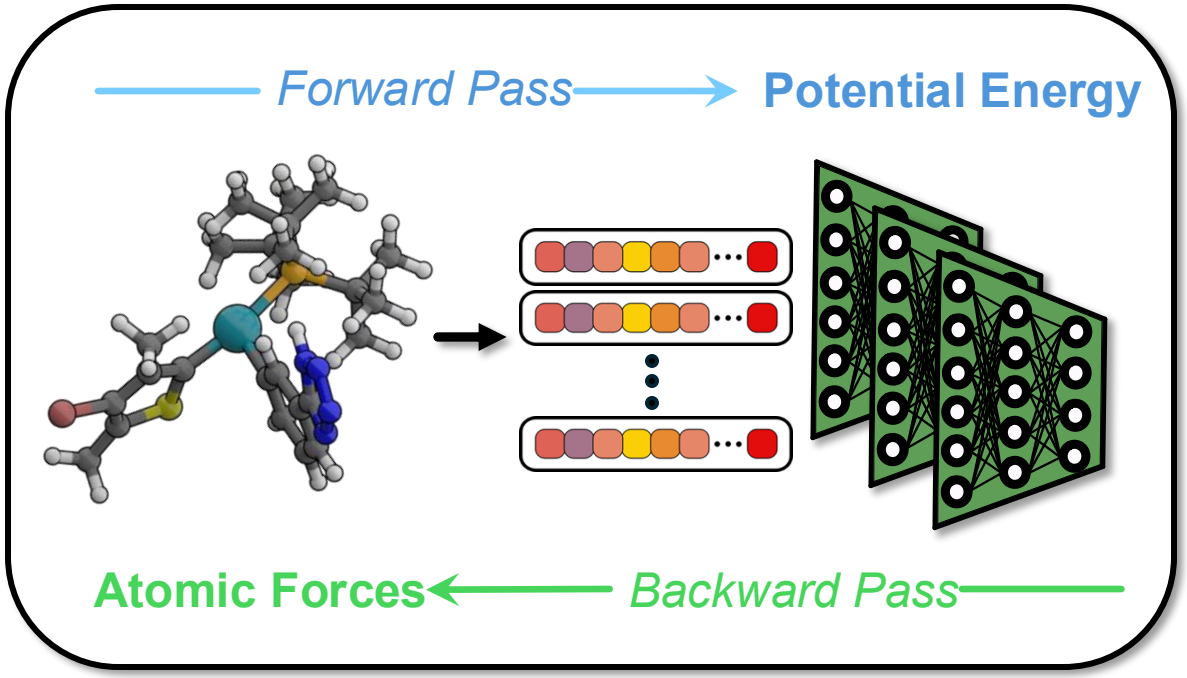
NVIDIA

Los Alamos NATIONAL LABORATORY EST. 1943

Motivation for Machine Learned Interatomic Potential (MLIP)

- Traditional physics-based methods for mechanistic modeling suffer from a **pervasive accuracy-efficiency trade-off**.
- MLIPs overcome this trade-off by **relating molecular structure directly to potential energy**, which is learned from a dataset of accurate quantum chemical calculations.

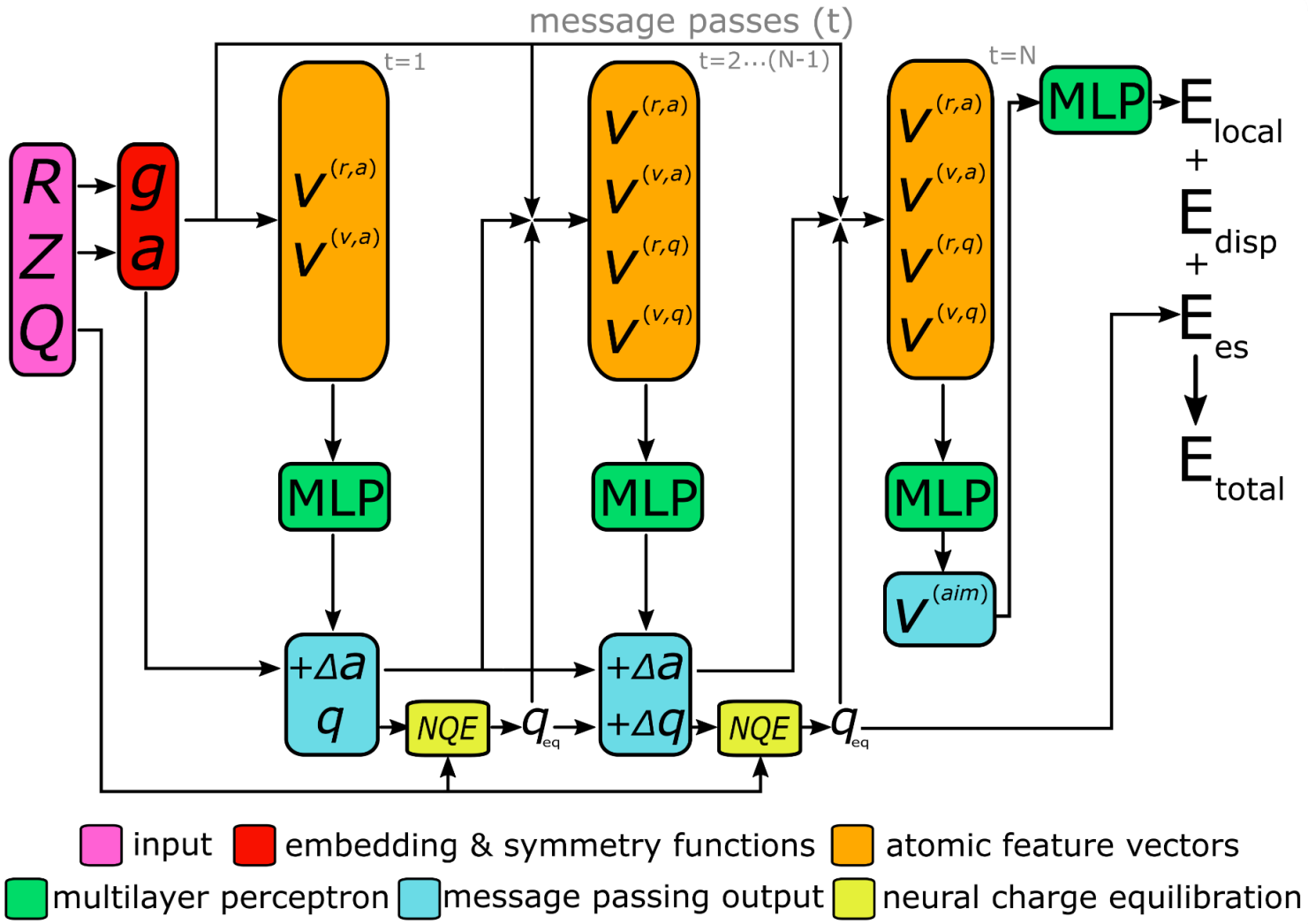




Neural Network Interatomic Potential

AIMNet2

The 2nd Generation Atoms-in-Molecules Model



Code and Data: https://github.com/isayevlab/aimnetcentral

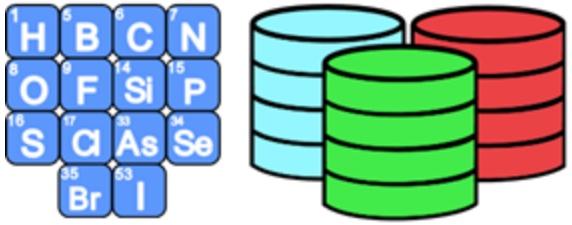
Anstine D, Zubatyuk R, Isayev O. AIMNet2: A Neural Network Potential to Meet your Neutral, Charged, Organic, and Elemental-Organic Needs. *Chem. Sci., 2025*; DOI: https://doi.org/10.1039/D4SC08572H

The 2nd Generation Atoms-in-Molecules ML Potential

AIMNet2 is generalized potential with chemical space coverage of

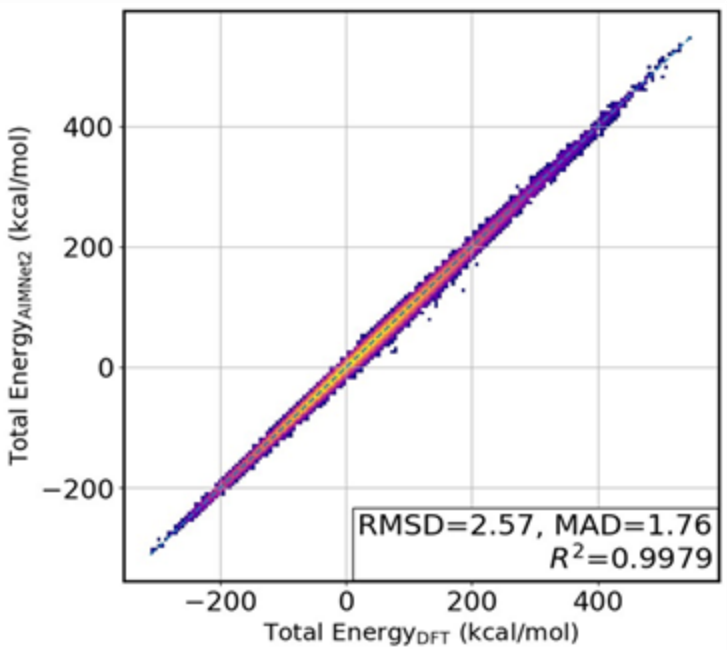
common non-metal and halogen elements with ~hybrid DFT accuracy

20M ωB97m/def2-TZVPP Calculations

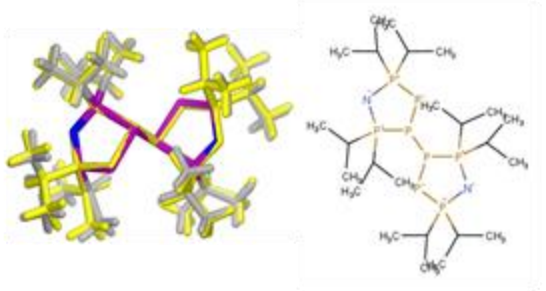


- Existing datasets
- Extraction from CSD
- Constrained Molecular Dyn.
- Normal Mode Sampling
- Active Learning
- Data Distillation
- ...

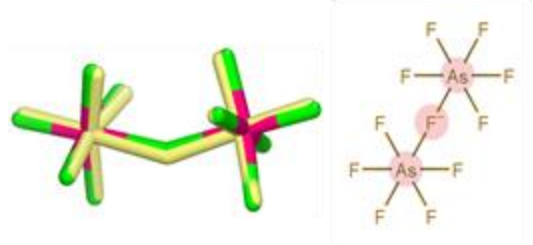
Neutral and Charged Species



zwitterionic bridged 5-membered phosphorous-nitrogen rings



Arsenic pentafluorides bridged by a negative fluorine

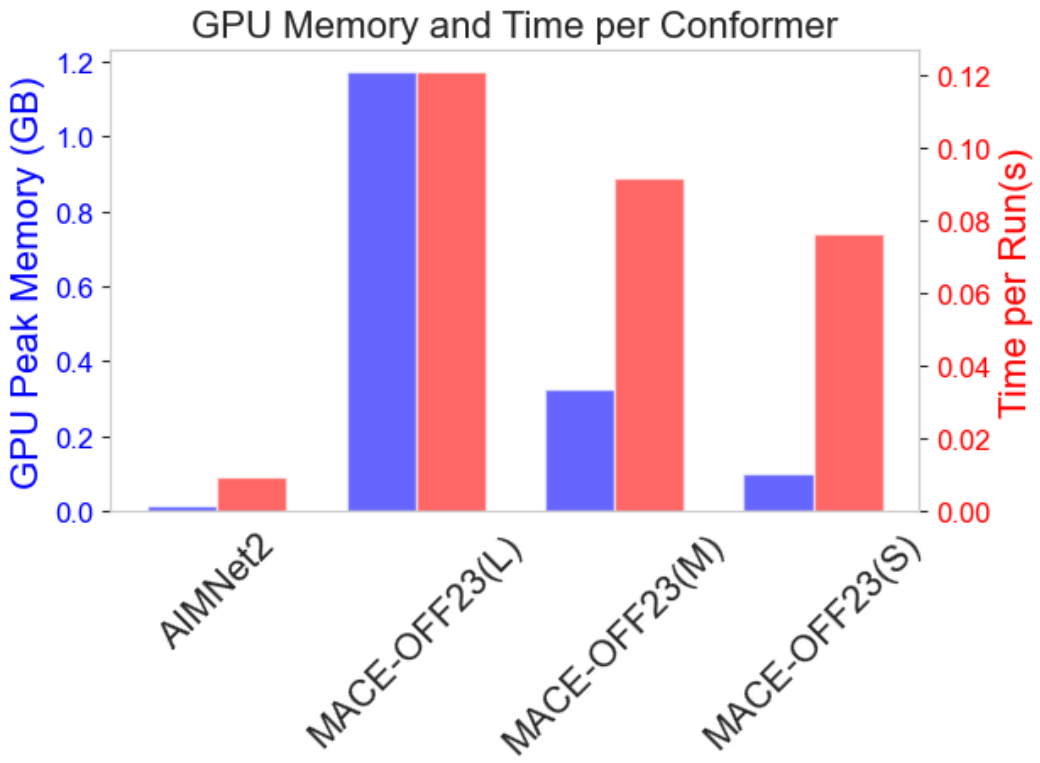


Pre-trained Models and Calculators Available: https://github.com/isayevlab/aimnetcentral

Anstine D, Zubatyuk R, Isayev O. AIMNet2: A Neural Network Potential to Meet your Neutral, Charged, Organic, and Elemental-Organic Needs. *Chem. Sci., 2025*; DOI: https://doi.org/10.1039/D4SC08572H

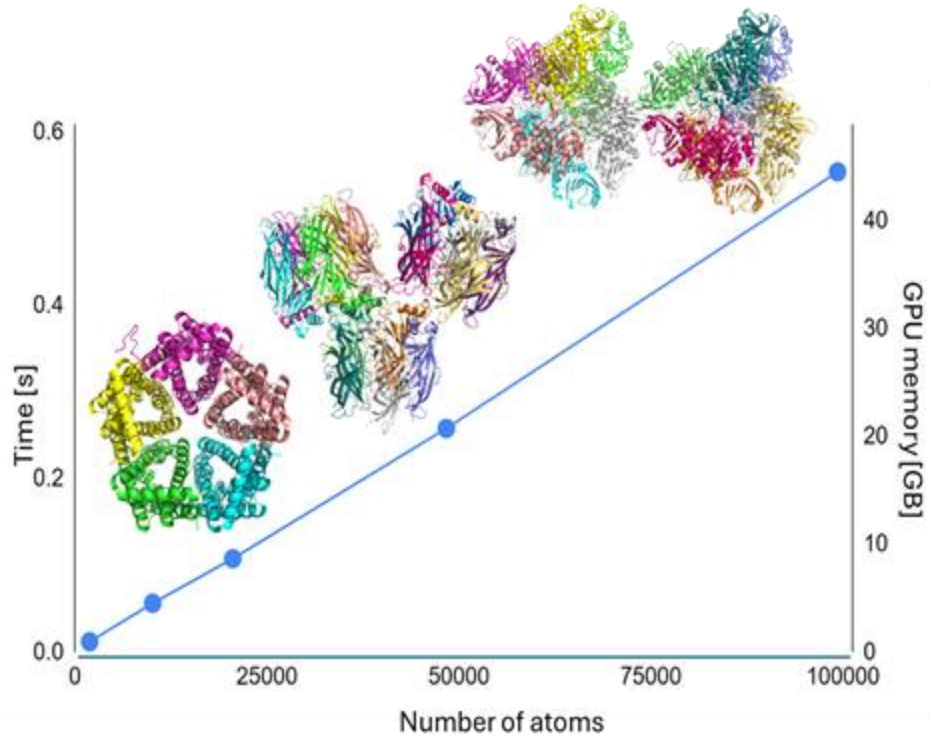
AIMNet2 computational efficiency

Single-point energy of macrocyclic peptide



Benchmarks performed on Nvidia NVIDIA V100 GPUs (32 GB SMX2).

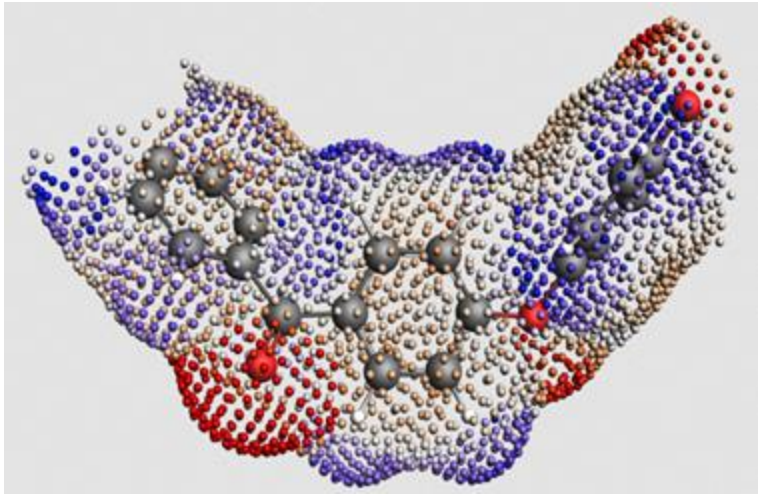
Protein energy & gradient evaluation



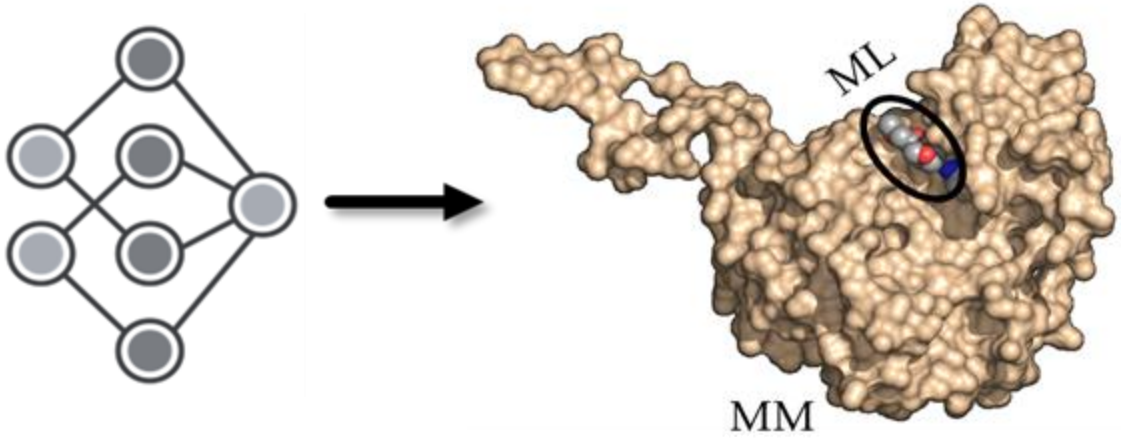
Benchmarks performed on a Nvidia H100 (80GB VRAM)
LBFGS optimizer in ASE has overhead of 4-5% beyond forces evaluation with AIMNet2.

AIMNet2 Foundation Model: Enabling Diverse Application

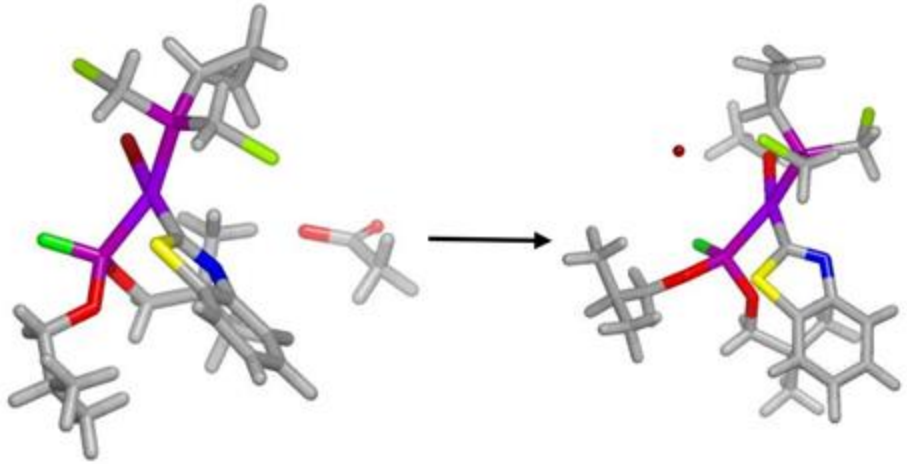
AIMNet2-COSMO-SAC/ RS



Machine Learning / Molecular Mechanics Simulations



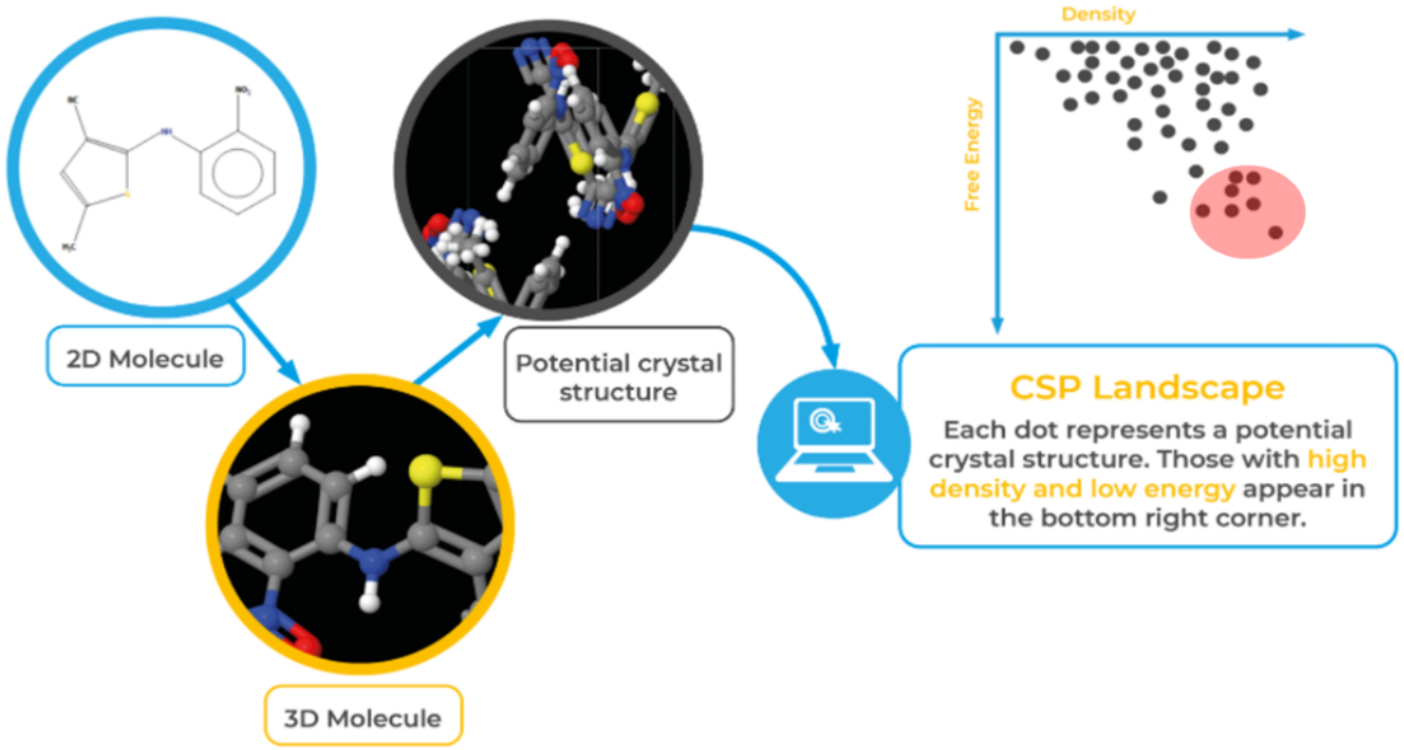
Pd Catalyzed Reactions



Generative Modeling

Anstine & Isayev, JACS, 2023

Crystal Structure Prediction (CSP): Overview



Challenges:

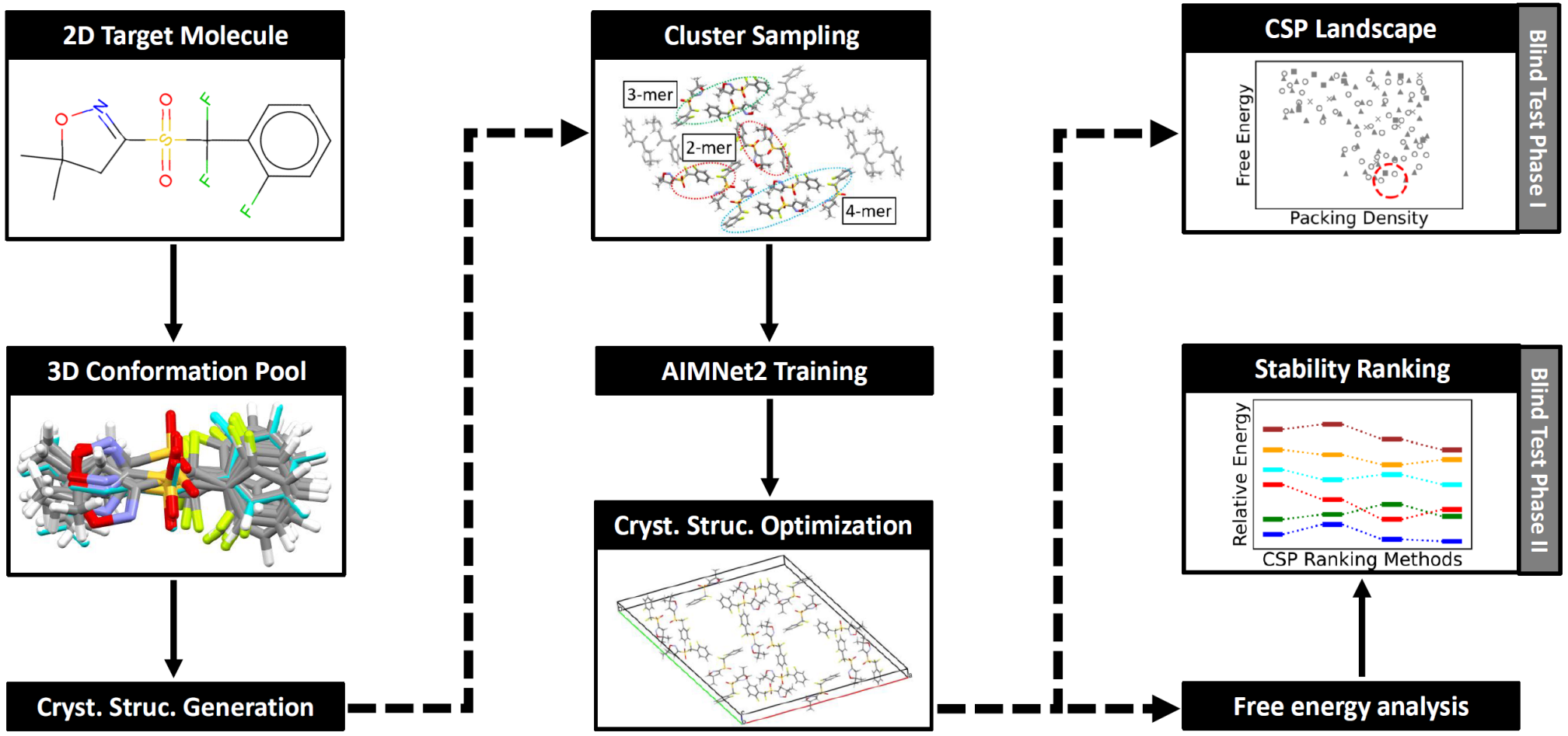
- Multiple molecule conformations
- Multiple crystal packing possibilities
- Predicted structures are not always observed
- Expensive
- Polymorphs
- Computational limitations

Benefits:

- Risk avoidance
- Manufacturing improvements
- Patent breaking/protection
- New materials discovery

Nayal KS, O'Connor D, Zubatyuk R, Anstine DM, Yang Y, Tom R, et al. Efficient Molecular Crystal Structure Prediction and Stability Assessment with AIMNet2 Neural Network Potentials. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-ksn4n

AIMNet2 CSP Finetunig Workflow



Nayal KS, O'Connor D, Zubatyuk R, Anstine DM, Yang Y, Tom R, et al. Efficient Molecular Crystal Structure Prediction and Stability Assessment with AIMNet2 Neural Network Potentials. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-ksn4n

7th CSP Blind Test Challenge Results

Group number	Group members	Target compounds attempted [total number of polymorphs]	Success (%)
1	Adjiman*, Pantelides*, Bowskill, Sugden, Sanders de Almada, Konstantinopoulos, Zhang	4 [9]	33
3	Boese*, List, Strasser, Hoja, Braun	3 [7]	43
5	Day*, Taylor, Butler	6 [13]	54
6	van Eijck*	7 [14]	14
8	Hofmann*, Kuleshova, Pilia	4 [8]	13
10	Jin*, Yang, L. Tan, Chang, Sun, X. Shi, C.Liu, Yue, Fu, Lin, Y.Zhou, Z.Liu, Zeng, Li, B. Shi, T. Zhou, Greenwell, Bellucci, Sekharan	7 [14]	86
11	Johnson*, Otero-de-la-Roza*, Clarke, Rumson, Mayo, A. J. A. Price	1 [1]	0
12	Jose*, Ramteke	4 [10]	0
13	Khakimov*, Pivina	3 [6]	0
16	Marom*, Isayev*, Bier, Hutchinson, Nayal, O'Connor, Tom, Zubatyuk	3 [6]	67
17	Matsui*, Shinohara	1 [1]	0
18	Mohamed*, Dhokale, Saeed, Alkhidir, Almehairbi	4 [10]	20
19	Muddana*, Jain, Darden, Skillman	6 [13]	23
20	Neumann*, Anelli, Woollam, Abraham, Dietrich, Firaha, Helfferich, Y. M. Liu, Mattei, Sasikumar, Tkatchenko, van de Streek	7 [14]	93
21	Obata*, Goto*, Utsumi, Ikabata, Okuwaki, Fukuzawa, Nakayama, Yonemochi	5 [11]	18
22	Oganov*, Maryewski, Momenzadeh Abardeh, Bahrami, Salimi	7 [14]	0
23	Pickard*, Cheng, Brandenburg	1 [1]	0
24	S. L. Price*, L. S. Price, Guo, Francia, Salvalaglio, Ding	7 [14]	50
25	Shang*, Z.-P. Liu	6 [11]	27
26	Szalewicz*, Ishaque, Nikhar, Podeszwa, Rogal, Vogt-Maranto	1 [4]	0
27	Tuckerman*, Szalewicz*, Bhardwaj, Chan, Hong, Ishaque, Jing, Melkumov, Nikhar, Podeszwa, Rehman, Rogal, Song, Vogt-Maranto	6 [10]	0
28	Q. Zhu*, Hu	6 [11]	0

- Achieved **highest success rate** of all **academic teams**
- One of only 3 teams that generated Target XXVII with RMSE<0.5
- Target XXIX structure with Z'=3 was not generated
- High energy polymorph of Target XXXI was not generated by any team

A seventh blind test of crystal structure prediction has been organized by the Cambridge Crystallographic Data Centre. The results are presented in two parts, with this second part focusing on methods for ranking crystal structures in order of stability. The exercise involved standardized sets of structures seeded from a range of structure generation methods. Participants from 22 groups applied several periodic DFT-D methods, machine learned potentials, force fields derived from empirical data or quantum chemical calculations, and various combinations of the above. In addition, one non-energy-based scoring function was used. Results showed that periodic DFT-D methods overall agreed with experimental data within expected error margins, while one machine learned model, applying system-specific AIMnet potentials, agreed with experiment in many cases demonstrating promise as an efficient alternative to DFT-based methods. For target XXXII, a consensus was reached across periodic DFT methods, with consistently high predicted energies of experimental forms relative to the global minimum (above 4 kJ mol^{-1} at both low and ambient temperatures) suggesting a more stable polymorph is likely not yet observed. The calculation of free energies at ambient temperatures offered improvement of predictions only in some cases (for targets XXVII and XXXI). Several avenues for future research have been suggested, highlighting the need for greater efficiency considering the vast amounts of resources utilized in many cases.

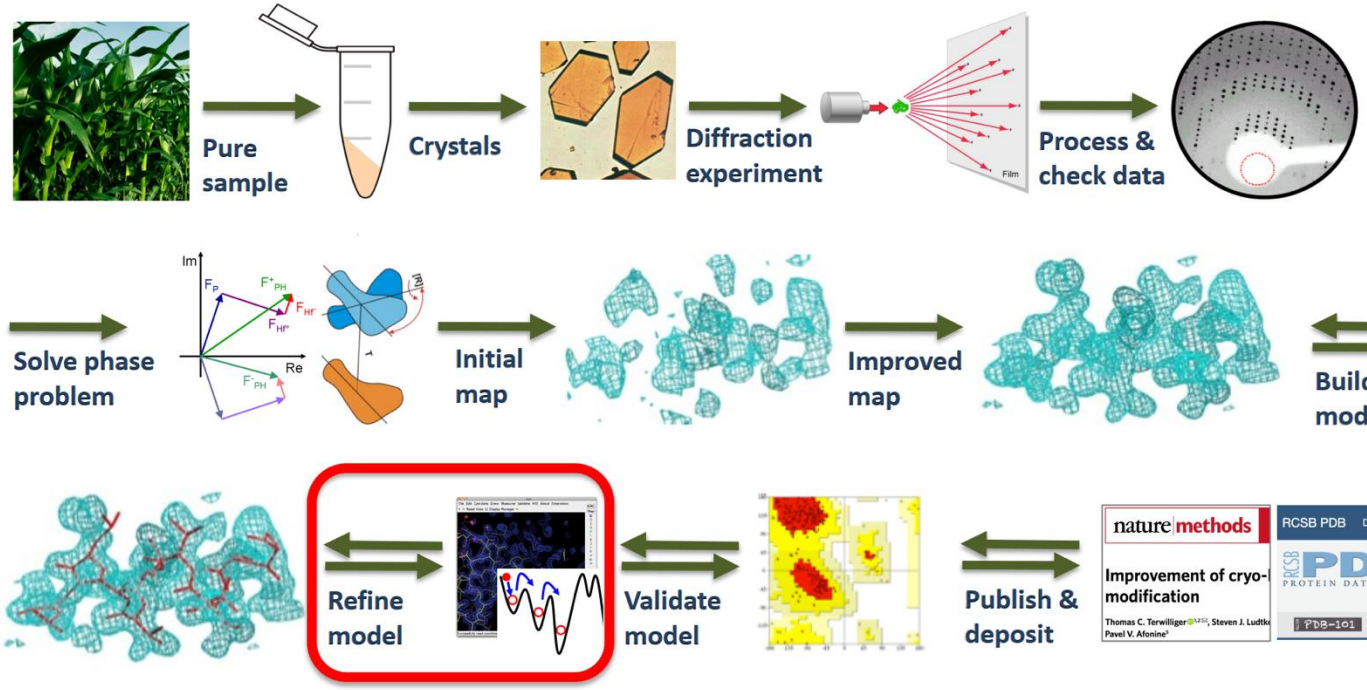
1. Introduction

1.1. Background

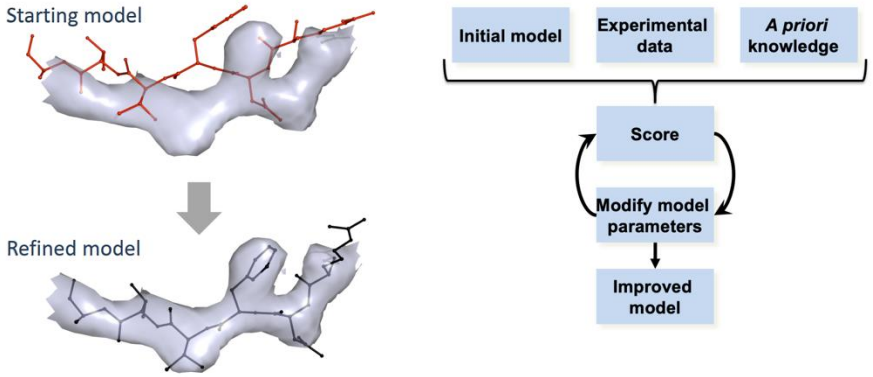
The Cambridge Crystallographic Data Centre (CCDC) has been organizing a set of blind tests to assess the predictive ability of existing methods for molecular crystal

Refinement of protein crystal structure with ML

Optimization process of fitting structural parameters to experimental data



***Priori* knowledge**



Fit atomic model to experimental data as good as possible while making sure the model makes physical and chemical sense

QM computations

https://github.com/qrefine/

Q|R

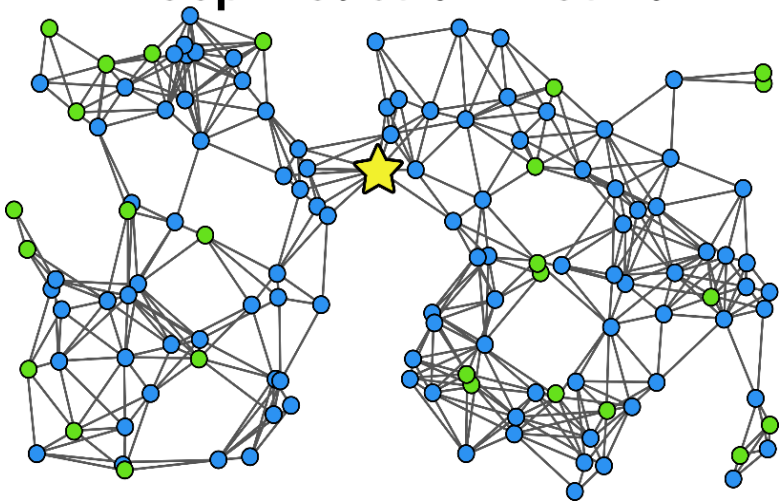
Phenix

R. Zubatyuk, et al. AQuaRef: Machine learning accelerated quantum refinement of protein structures. BioRxiv 2024.07.21.604493; DOI: https://doi.org/10.1101/2024.07.21.604493

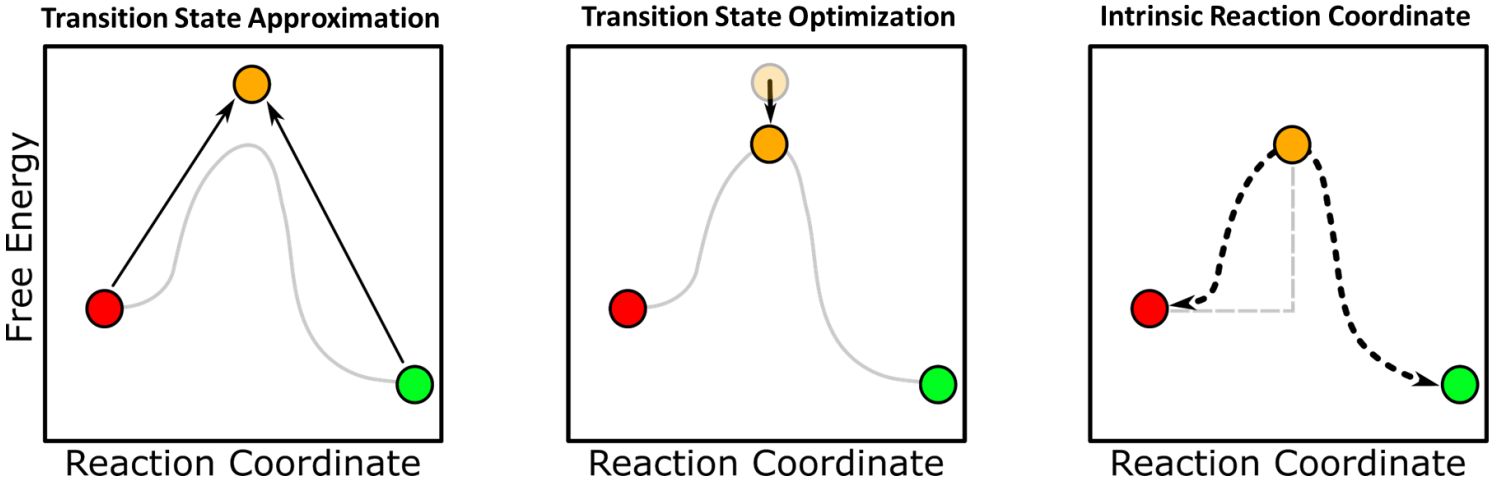
The Challenge of Computational Mechanistic Modeling

- Tasks such as synthesis planning or reaction engineering can require **traversal of deep reaction networks, which is laborious with QM methods**.
- **Goal:** Carry out mechanistic modeling tasks, e.g., transition optimization, efficiently on ML potential energy surfaces with minimal sacrifices to accuracy.

Deep Reaction Network

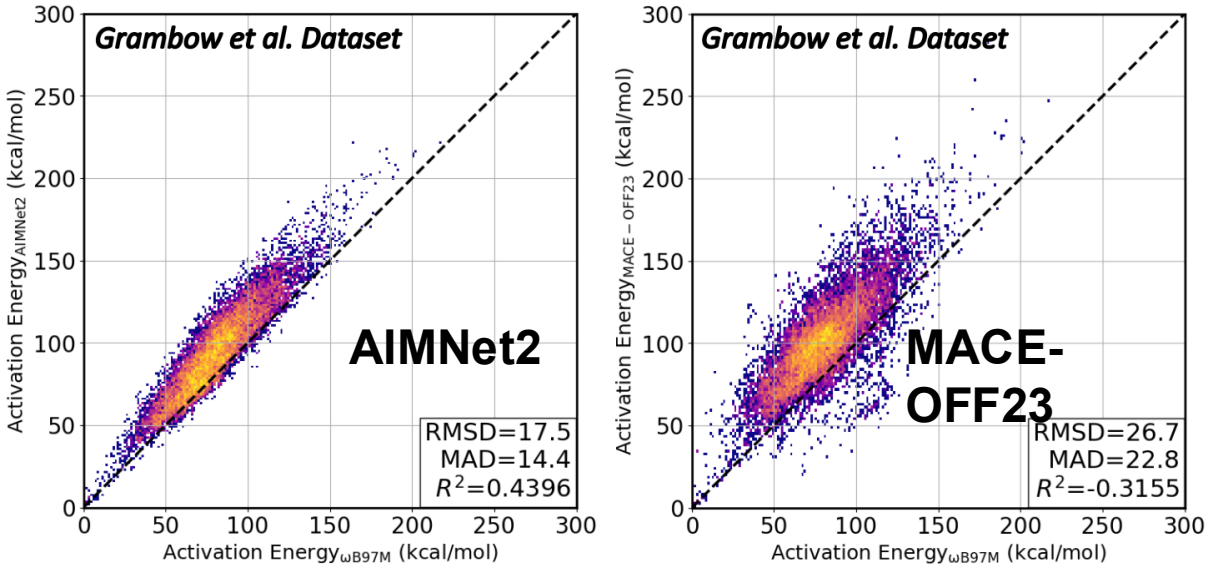


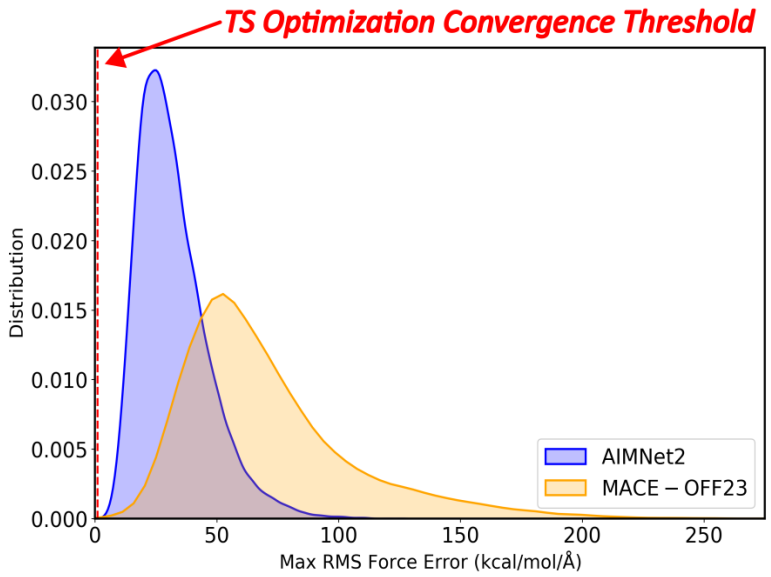
Synthesis Planning
Stressor-Specific Deconstruction
Total Synthesis
Biochemical Pathways
...



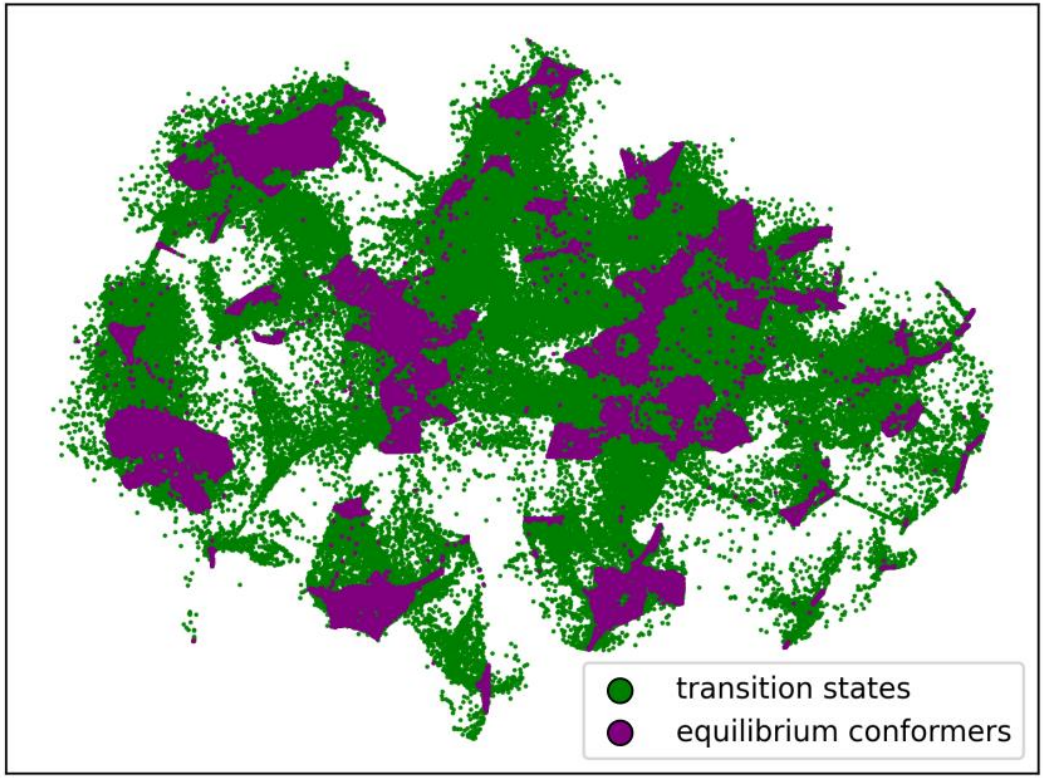
approximate time required per reaction with DFT: hour(s) – several days

General Potentials Do Not Generalize to Reactions





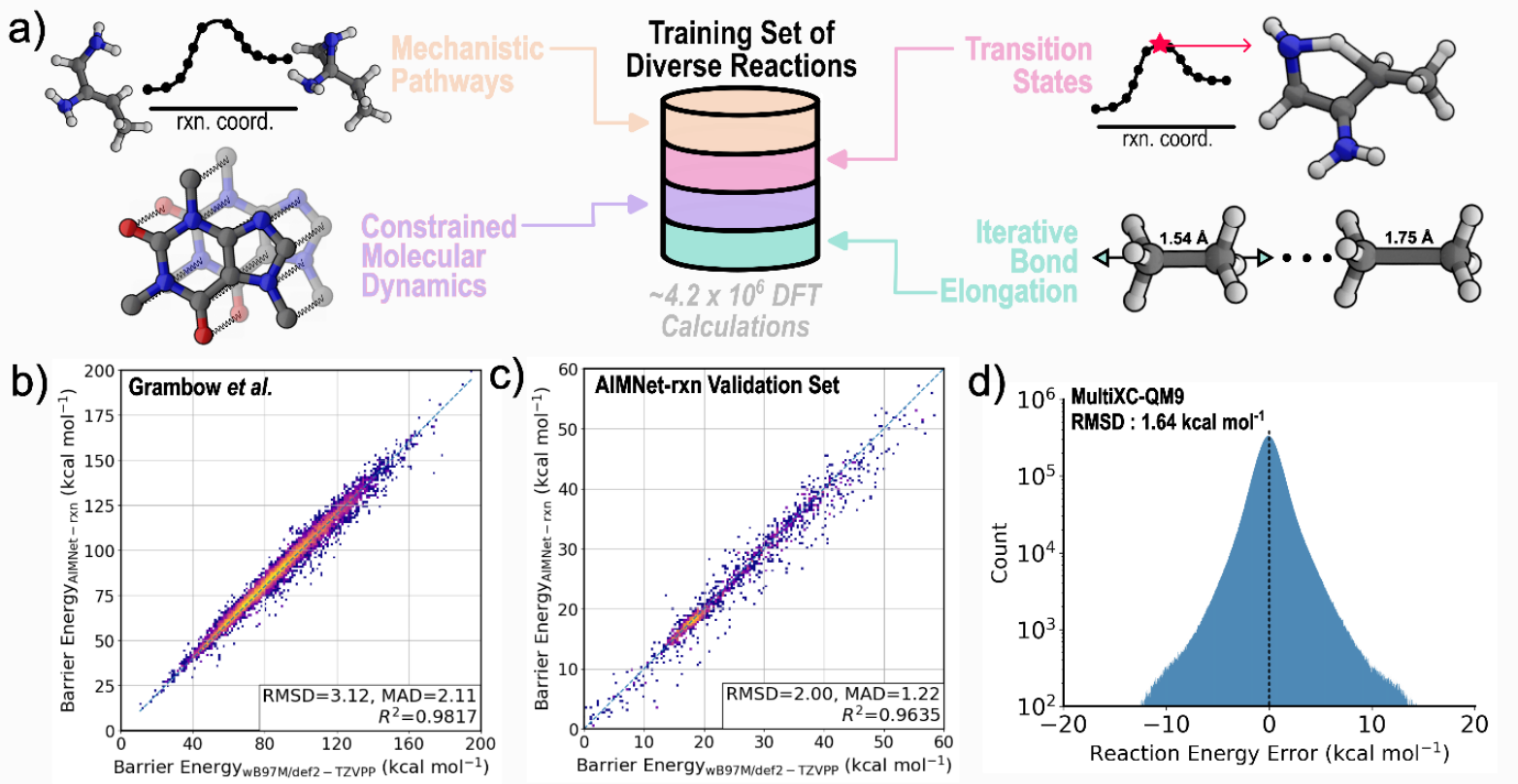
- **<u>Atomic Environments found in reactive systems are distinct</u>** from (near-)equilibrium conformers.



**t-SNE plot of learned carbon atomic environment descriptor vectors.*

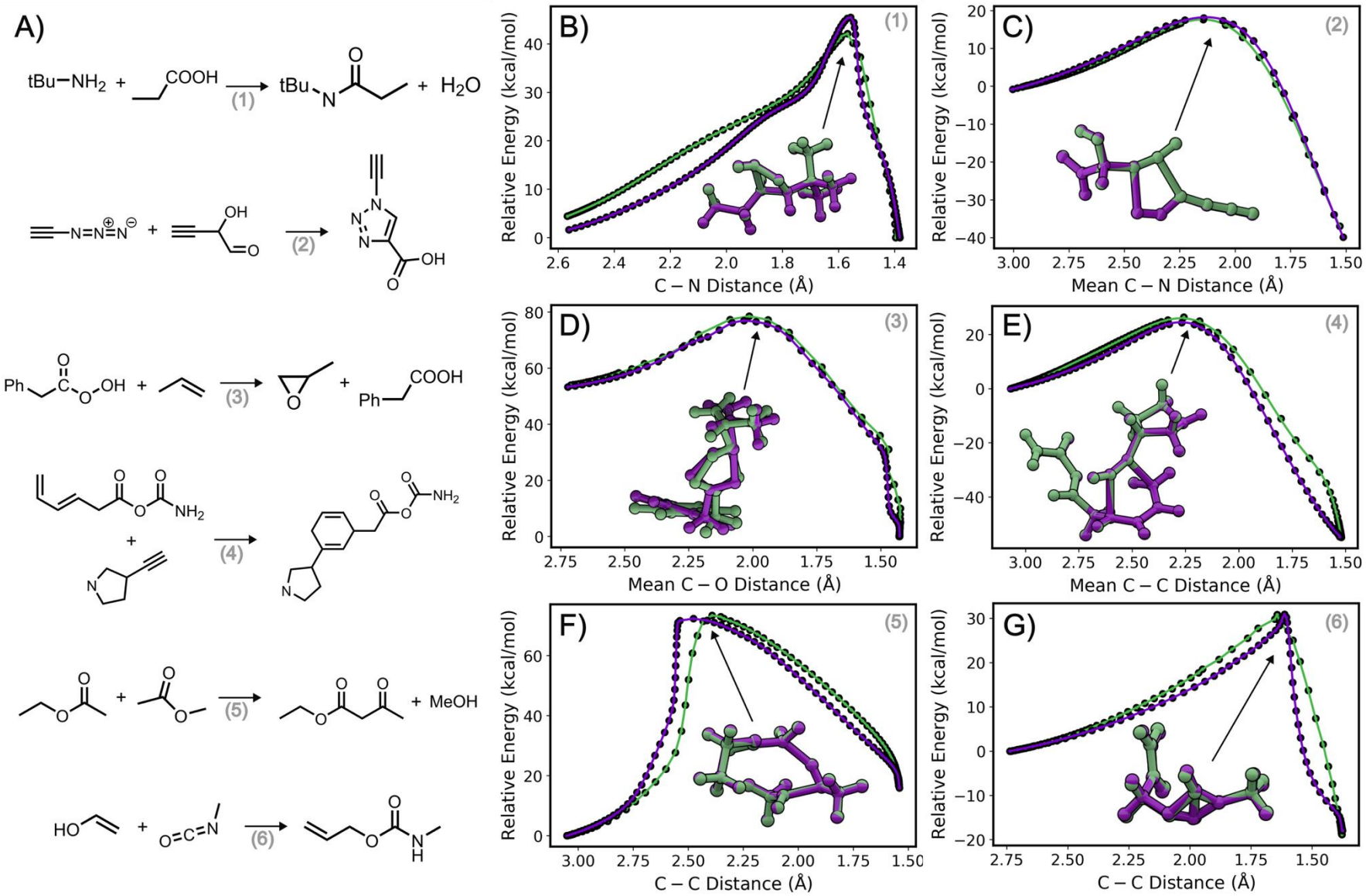
AIMNet2-rxn: A General MLIP For Reaction Modeling

It is a task-specific **model for general mechanistic study of organic reactions with ~hybrid DFT accuracy.**



Anstine DM, Zhao Q, Zubatiuk R, Zhang S, Singla V, Nikitin F, et al. AIMNet2-rxn: A Machine Learned Potential for Generalized Reaction Modeling on a Millions-of-Pathways Scale. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-hpdmg

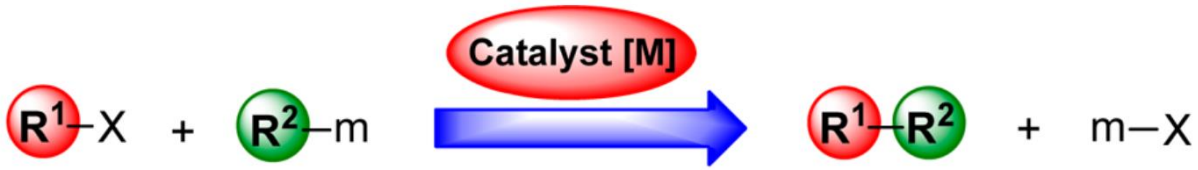
AIMNet2-rxn: One NNP, Many Reaction Mechanisms without Retraining



RxnAIMNet has been tested on Diels-Alder, triazole formation (click chemistry), combustion, tautomerization, hydrogen transfer, esterification, aldehyde and ketone formation, metathesis, ring-closing and ring-opening, amine protection-deprotection, …

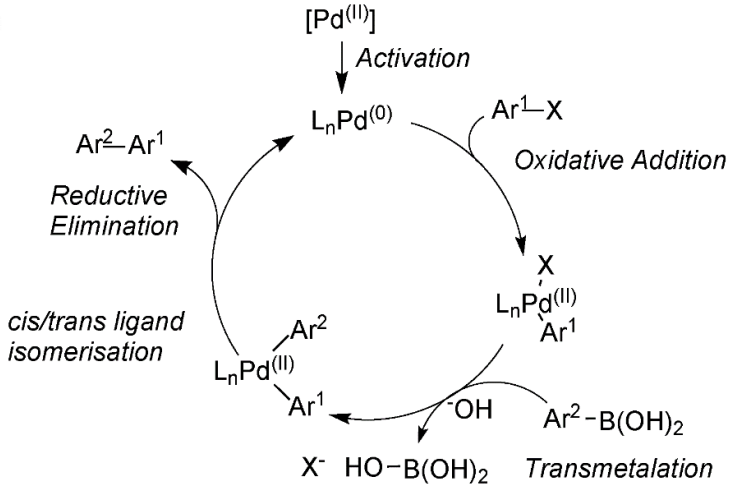
Anstine DM, Zhao Q, Zubatiuk R, Zhang S, Singla V, Nikitin F, et al. AIMNet2-rxn: A Machine Learned Potential for Generalized Reaction Modeling on a Millions-of-Pathways Scale. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-hpdmg

AIMNet2-Pd: Pd-catalyzed C-C cross-coupling reaction

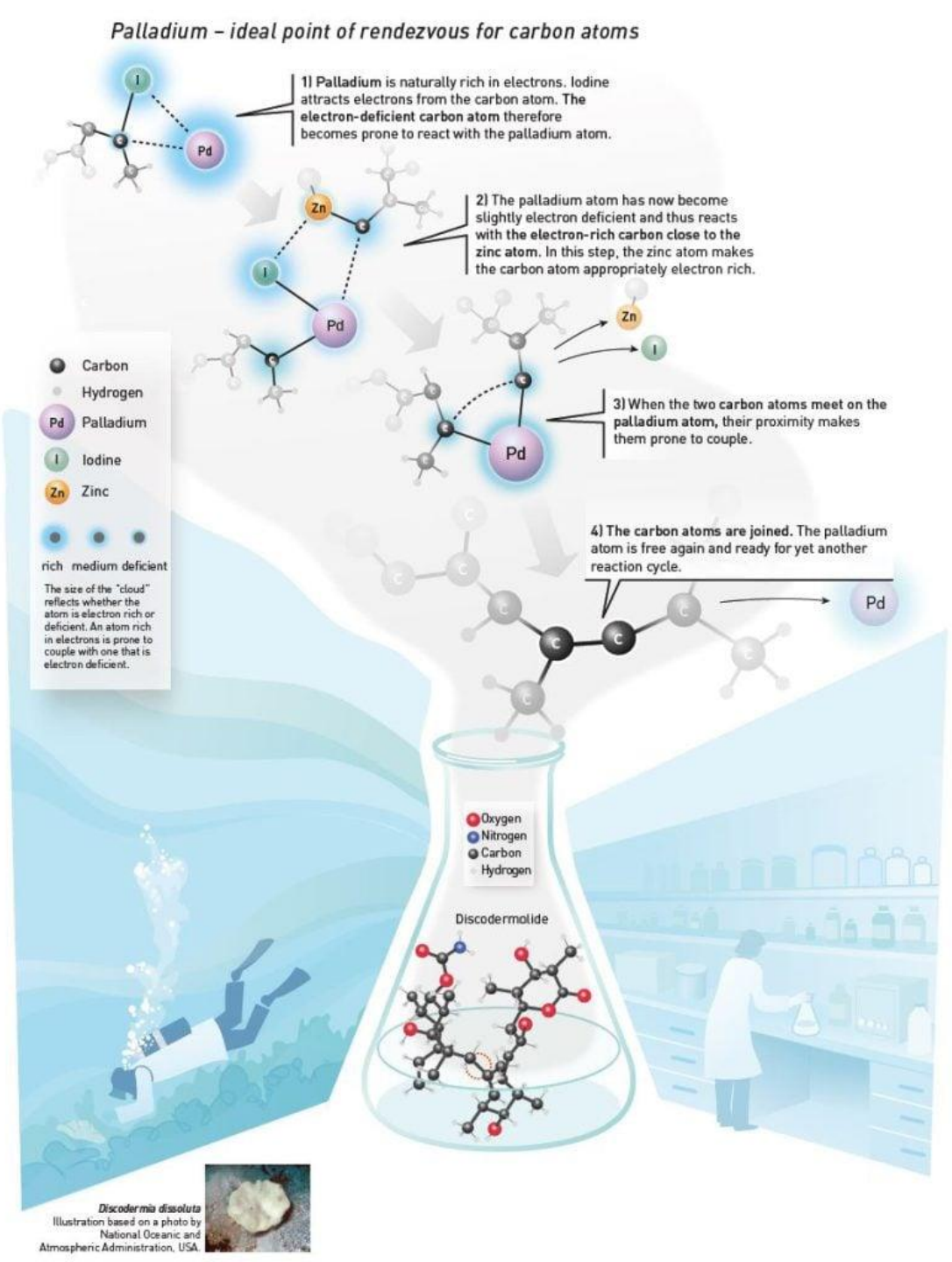


m= B (Suzuki-Miyaura)
Sn (Stille)
Zn (Negishi)
Si (Hiyama) ...

[M]= Fe, Rh, Ni, Pd ...
X= I, Br, Cl, OTf ...



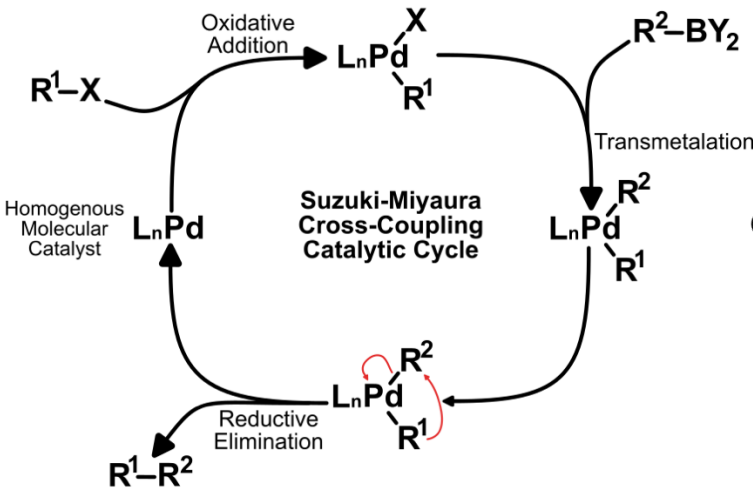
AIMNet2-Pd is the 1st MLIP with **broad applicability to reactive Pd-based metal-organic molecules**

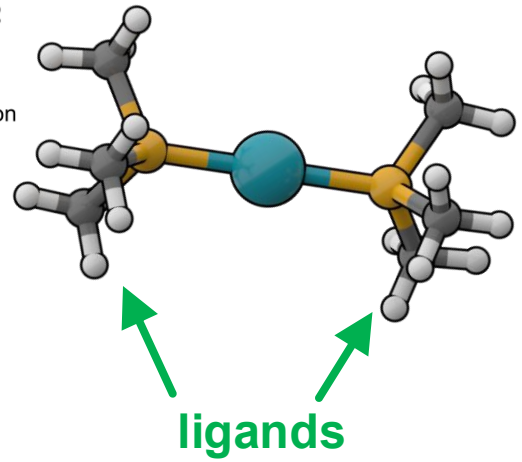


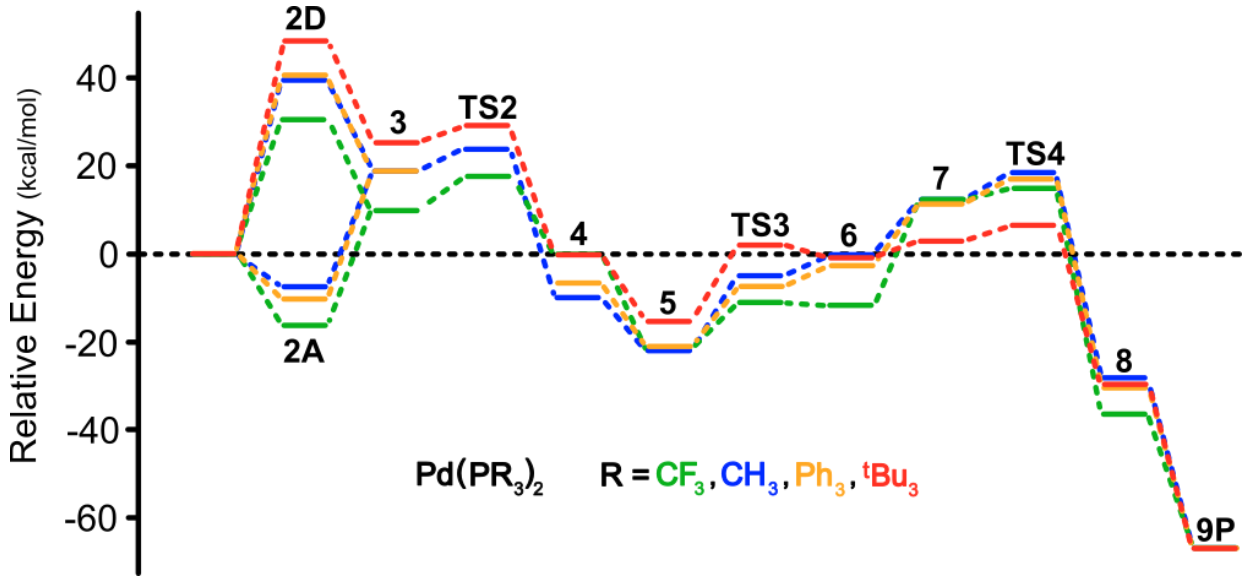
AIMNet2-Pd for Pd-based Homogenous Catalysis

- The homogenous catalyst **market in the U.S. is a multi billion-dollar industry[1].**
- Identifying **selective, fast, high-yielding catalytic pathways** can be transformative to, for example, pharmaceuticals, polymer synthesis, environmental science, petroleum engineering…

carbon-carbon bond formation via Pd-based molecular catalysts



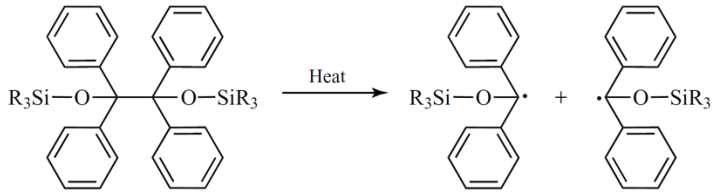


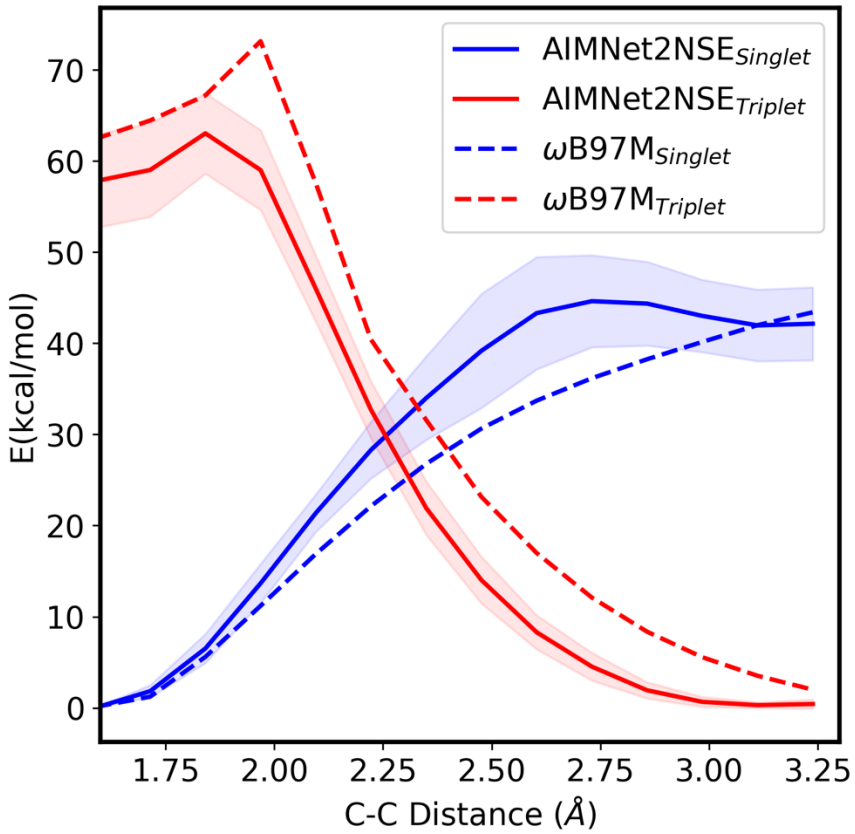


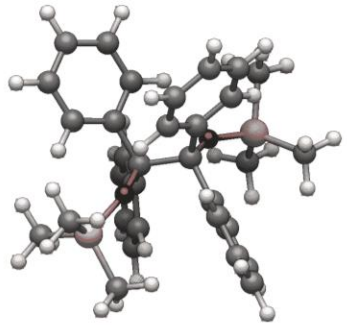
Anstine D, Zubatyuk R, et al. Transferable Machine Learning Interatomic Potential for Pd-Catalyzed Cross-Coupling Reactions. *ChemRxiv*. 2025; doi:10.26434/chemrxiv-2025-n36r6

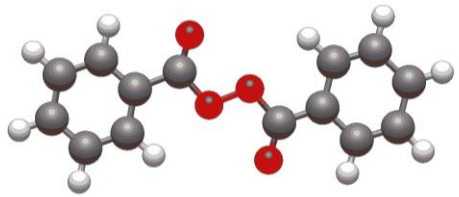
Radical Reaction Profiles

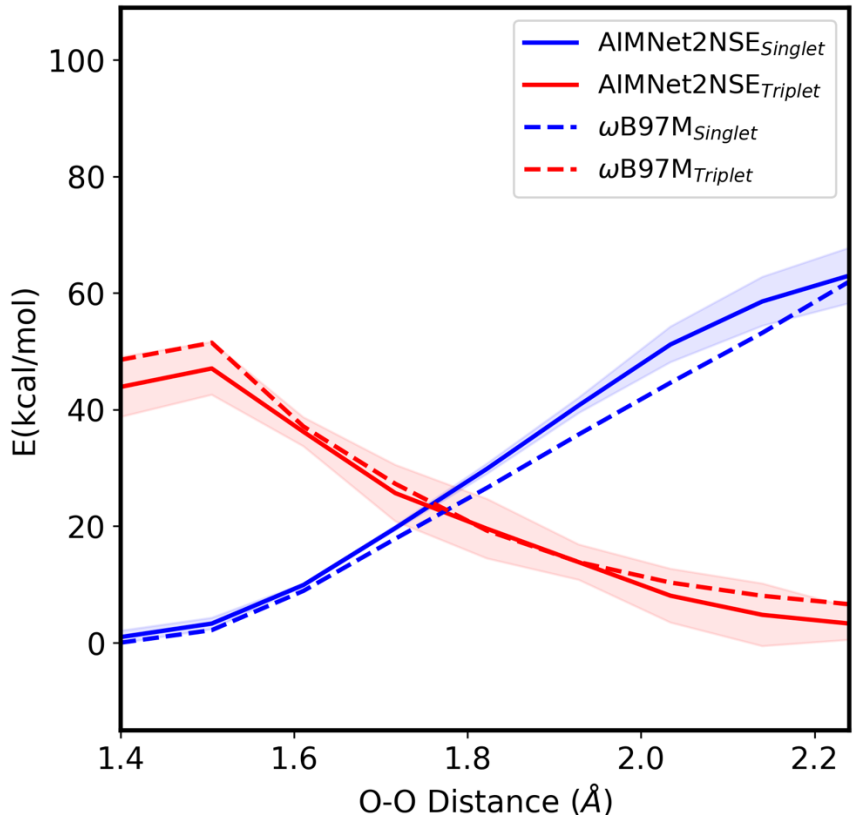
Silylated Benzopinacol

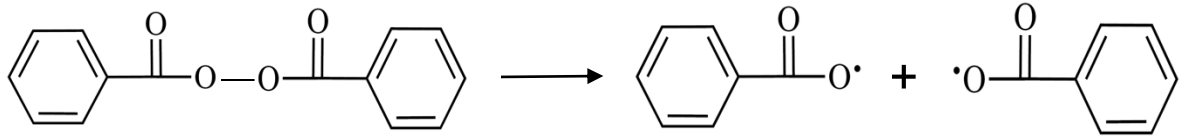








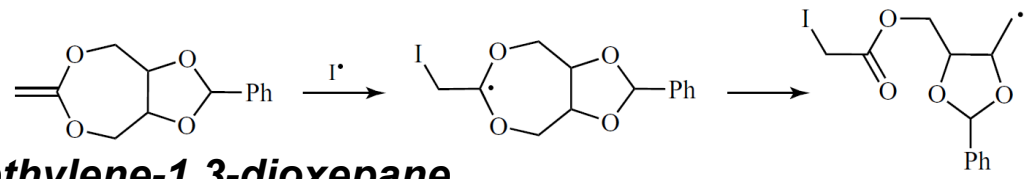




Dibenzoyl Peroxide

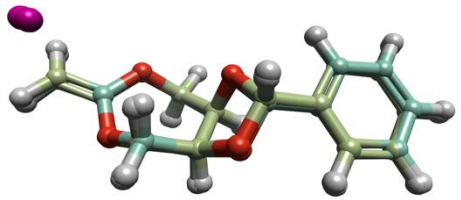
B. Kalita et al. Under review, 2025

Industrial Radical Polymerization Reactions



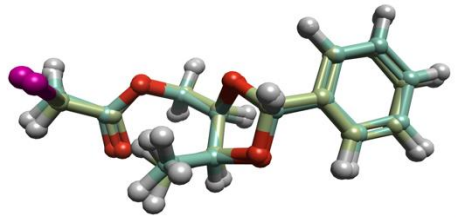
2-methylene-1,3-dioxepane

Reactant



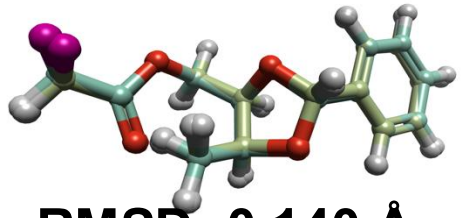
RMSD: 0.139 Å

Product

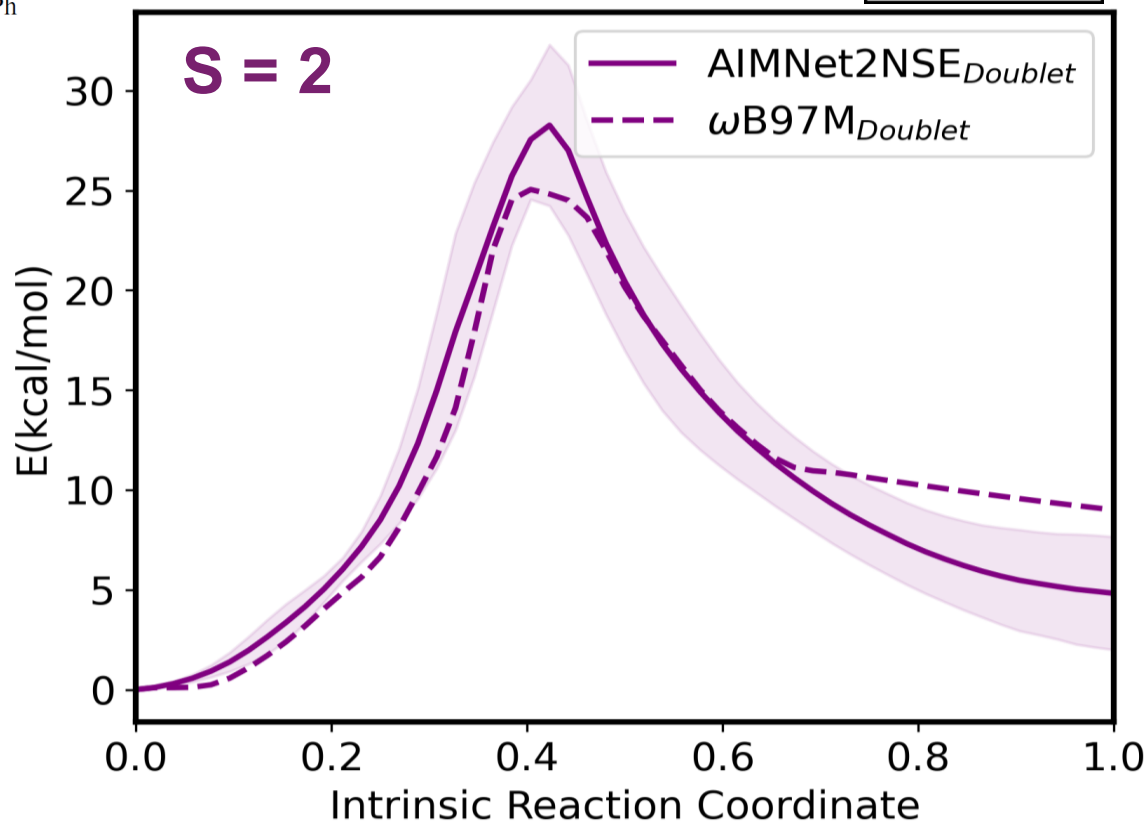


RMSD: 0.231 Å

Transition State



RMSD: 0.140 Å



B. Kalita et al. Under review, 2025

Nesvadba, P. Radical Polymerization in Industry. https://doi.org/10.1002/9781119953678.rad080

Automated Science – Self-Driving Experiments

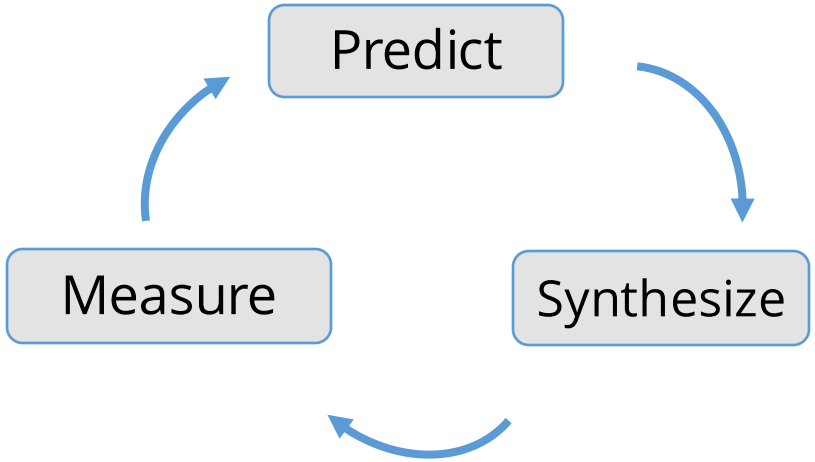
New way of doing science: from Edisonian to truly autonomous discovery

Generative Models

Prediction of structure based on desired properties

Machine Learning

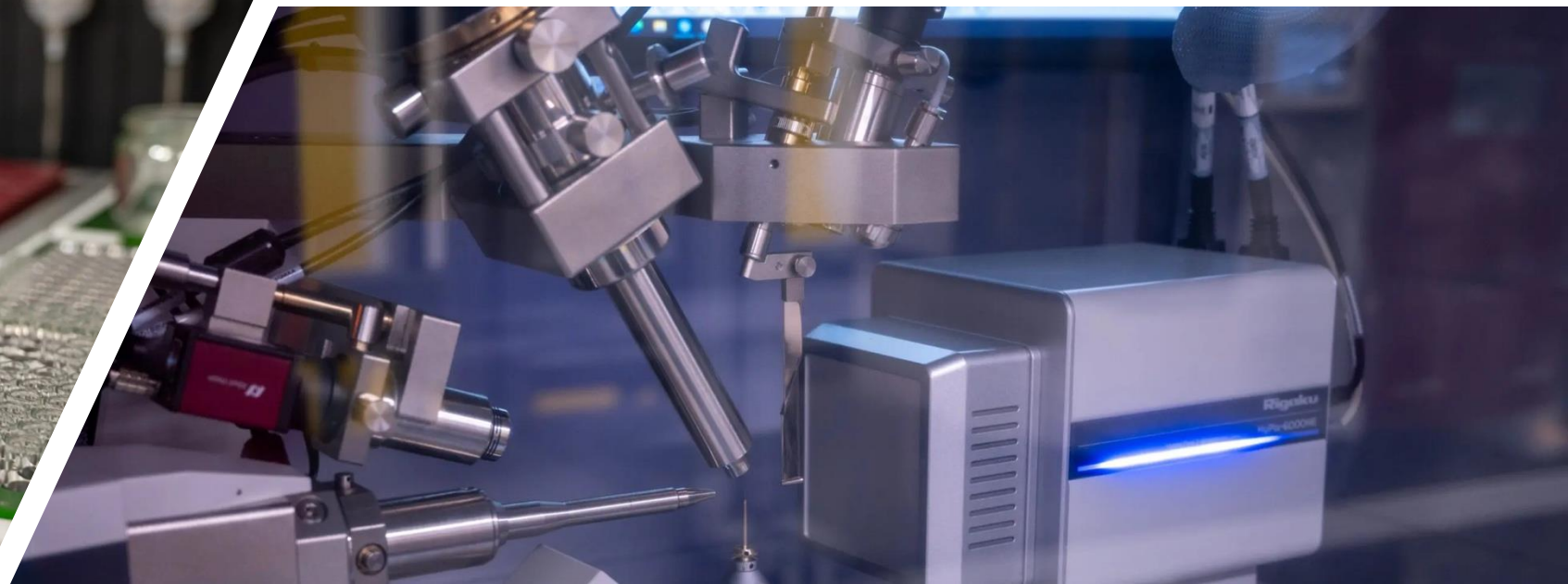
Optimizing experimental workflows



Cloud Lab

High throughput synthesis and characterization

CMU Bakery Square Lab



Instrumentation at CMU BKSQ Lab

Materials Storage and Handling

Sample Preparation

Workcell
Micro Liquid Handler

Separations & Mass Spectrometry

Organic Synthesis

Peptide Synthesizer

Property Measurements

Density

Spectroscopy

NMR

Bioassays

Plate readers
Electrophoresis
Biophysics

Microscopy & Crystallography

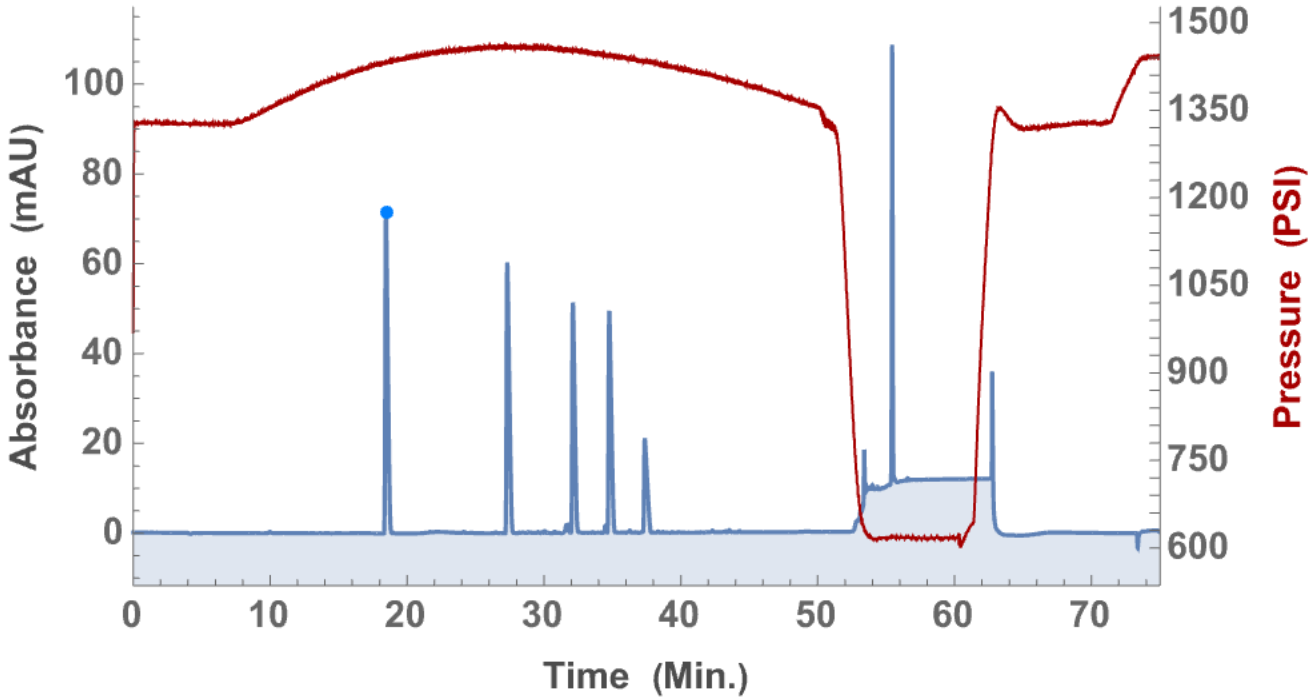
X-ray diffractometer

Cell Preparation

Bioreactors

In-brief: High-Performance Liquid Chromatography

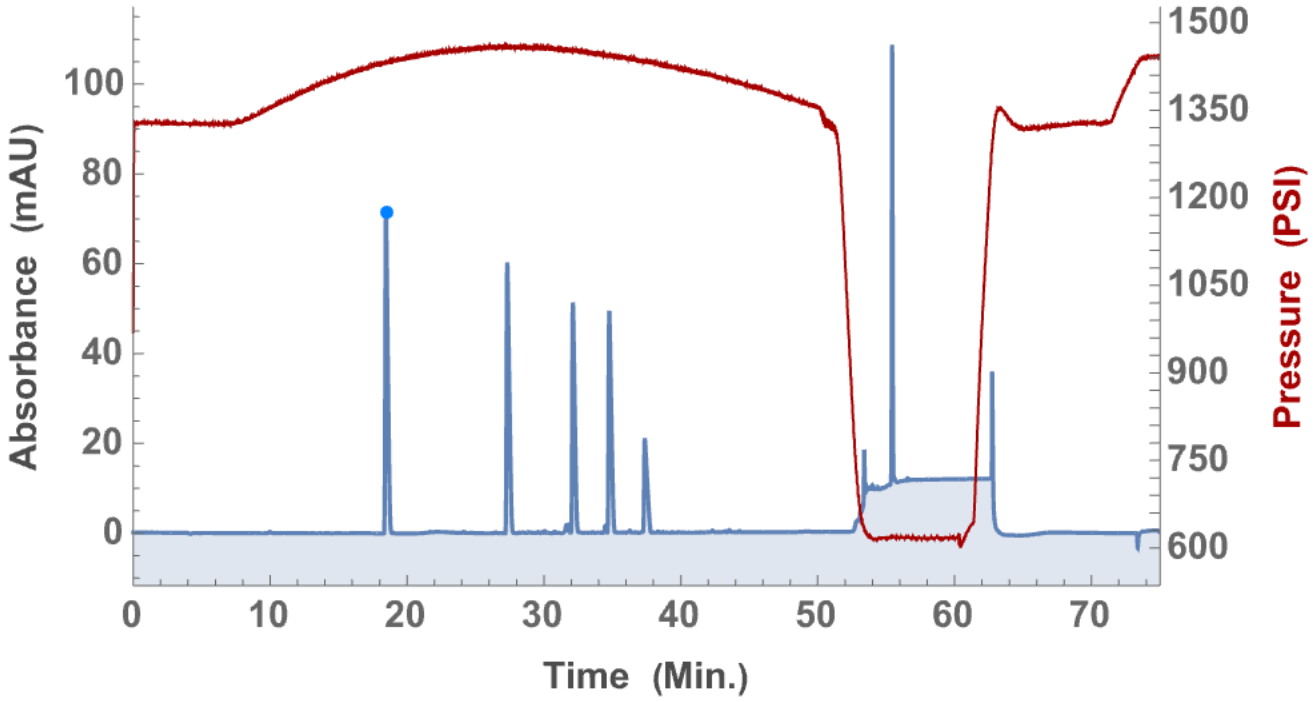
A general-purpose method for the detection and separation of chemical entities (small molecules, peptides, etc).



Gusev F, Kline BC, Quinn R, Xu A, Smith B, Frezza B, et al. Machine Learning anomaly detection of automated HPLC experiments in the Cloud Laboratory. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-7ggzl

In-brief: High-Performance Liquid Chromatography

A general-purpose method for the detection and separation of chemical entities (small molecules, peptides, etc).



define gradient method

```
{Object[Method, Gradient, "X-min gradient at Y and pH Z with _% ACN "]...}
```

submit experiment

```
ExperimentHPLC[injList,
               Gradient -> gradList,
               Instrument -> Model[Instrument, HPLC, ""],
               Column -> Object[Item, Column, "id:L8kPEjnLYbRW"],
               AbsorbanceWavelength -> (254 * Nanometer),
               AbsorbanceSamplingRate -> 20 Second^-1,
               BufferA -> Model[Sample, ""],
               BufferB -> Model[Sample, ""],
               BufferC -> Model[Sample, ""],
               BufferD -> Model[Sample, ""],
               FlowRate -> (1.0 * Milliliter/Minute),
               InjectionVolume -> (2 * Microliter)]
```

Gusev F, Kline BC, Quinn R, Xu A, Smith B, Frezza B, et al. Machine Learning anomaly detection of automated HPLC experiments in the Cloud Laboratory. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-7ggzl

Docs searcher Large Language Model (with Gabe Gomes, CMU)

Input prompt from scientist

Docs searcher

Planner

API usage information / Prompt-to-SLL

SLL information-flow back to Planner for downstream task

planner requested queries

```
analyze a complex mixture
to see what is in it
```

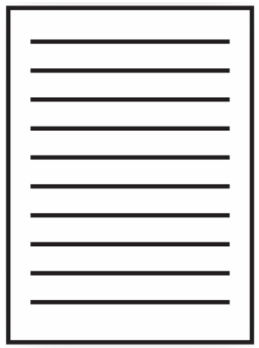
```
quantify proteins
in my sample
```

```
determine m/z ratio for
compounds in my sample
```

natural language prompts

+

Function Guide



extracted from webpage

Prompt-to-Function

```
ExperimentHPLC,
ExperimentIonChromatography
```

```
ExperimentTotalProteinQuantification,
ExperimentTotalProteinDetection
```

```
ExperimentMassSpectrometry,
ExperimentGCMS
```

function selection →

Running Experiments

Collection of functions used to remotely conduct experiments in an ECL facility.

plain text documentation

further investigation →

```
ExperimentHPLC[Samples] => Protocol
    Experimental Principles...
    Instrumentation...
    Experiment Options...
    Sample Parameters...
    ...
```

description summarization

LLM

code retention

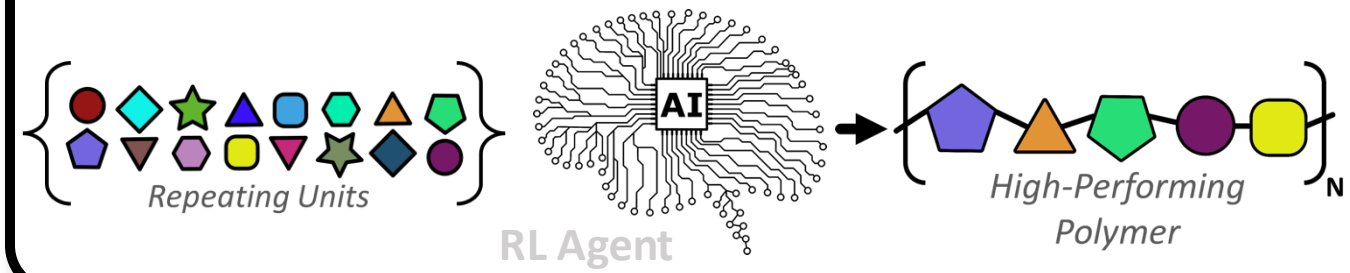
**not from webpage* required extensive work from the folks at ECL*

Automating Polymeric Materials Development

Discovering and processing *de novo* polymers can be treated as a RL problem(s).

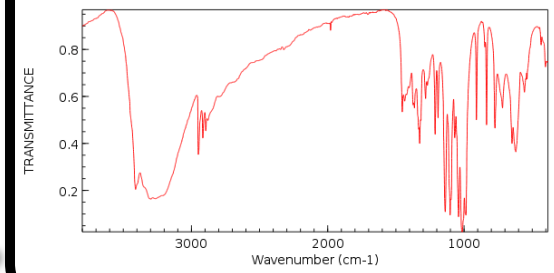
Select Polymer Composition from a Pool of Molecules

use value-based RL to select repeating unit compositions

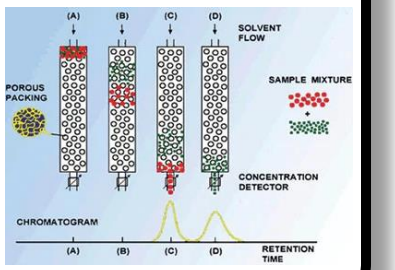


Exp. Synthesis and Characterization

IR Spectrum



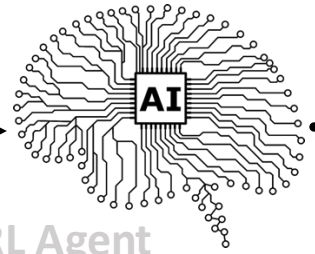
GPC



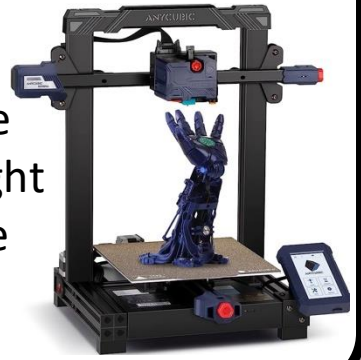
Select Printing Parameters Based on Polymer Characterization

use policy-based RL to select optimal printer parameters

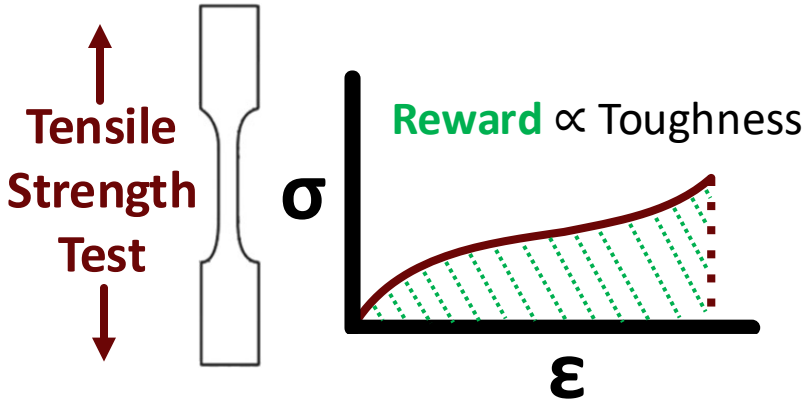
Composition
+
Characterization



Temperature
Sample Height
UV Exposure
•••



Exp. Polymer Property Evaluation



Rapp J et al. Design of Tough 3D Printable Elastomers with Human-in-the-Loop Reinforcement Learning. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-w1563

Use the AIMNet2:
Anstine D, Zubatyuk R, Isayev O.
Chem. Sci., 2025; DOI:
https://doi.org/10.1039/D4SC08572H

AIMNet2 implementation in Pytorch
Available at:
https://github.com/isayevlab/aimnetcentral

Plugins: ORCA 6.1, ASE, LAMMPS, OpenMM, SCM-ADF, AMBER(dev)

AIMNet2 Dataset:
https://doi.org/10.1184/R1/27629937.v2

Used in Government labs, companies etc.

Los Alamos NATIONAL LABORATORY EST. 1943

BERKELEY LAB Lawrence Berkeley National Laboratory

NIH National Institutes of Health

BASF We create chemistry

GOLD SCIENCE AMBASSADOR 2024 DASSAULT SYSTEMES

P&G

idorsia

Genentech

Pfizer