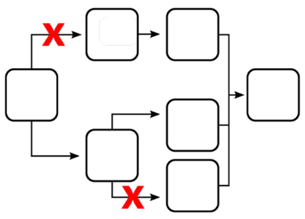
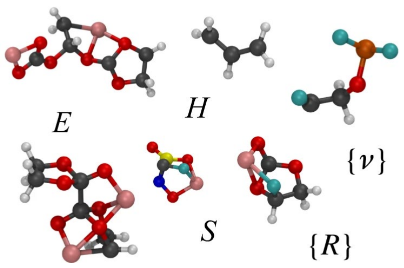
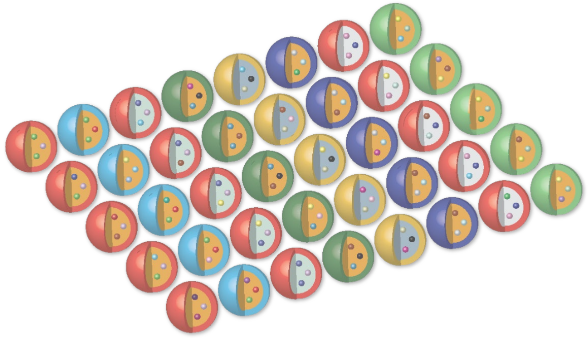
High-Throughput DFT and Monte Carlo for Reaction Networks and Machine Learning

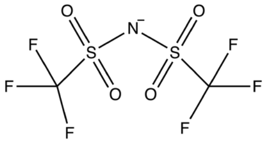




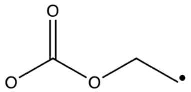
Samuel M. Blau
Research Scientist
Lawrence Berkeley Lab



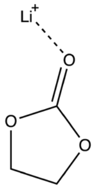
High-Throughput Molecular DFT Data Generation



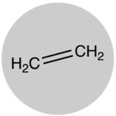
Charged molecules



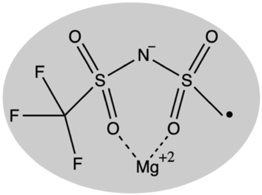
Radical molecules



Metal-coordinated molecules



Solvated molecules



Additional complexity

High-Throughput Molecular DFT Data Generation

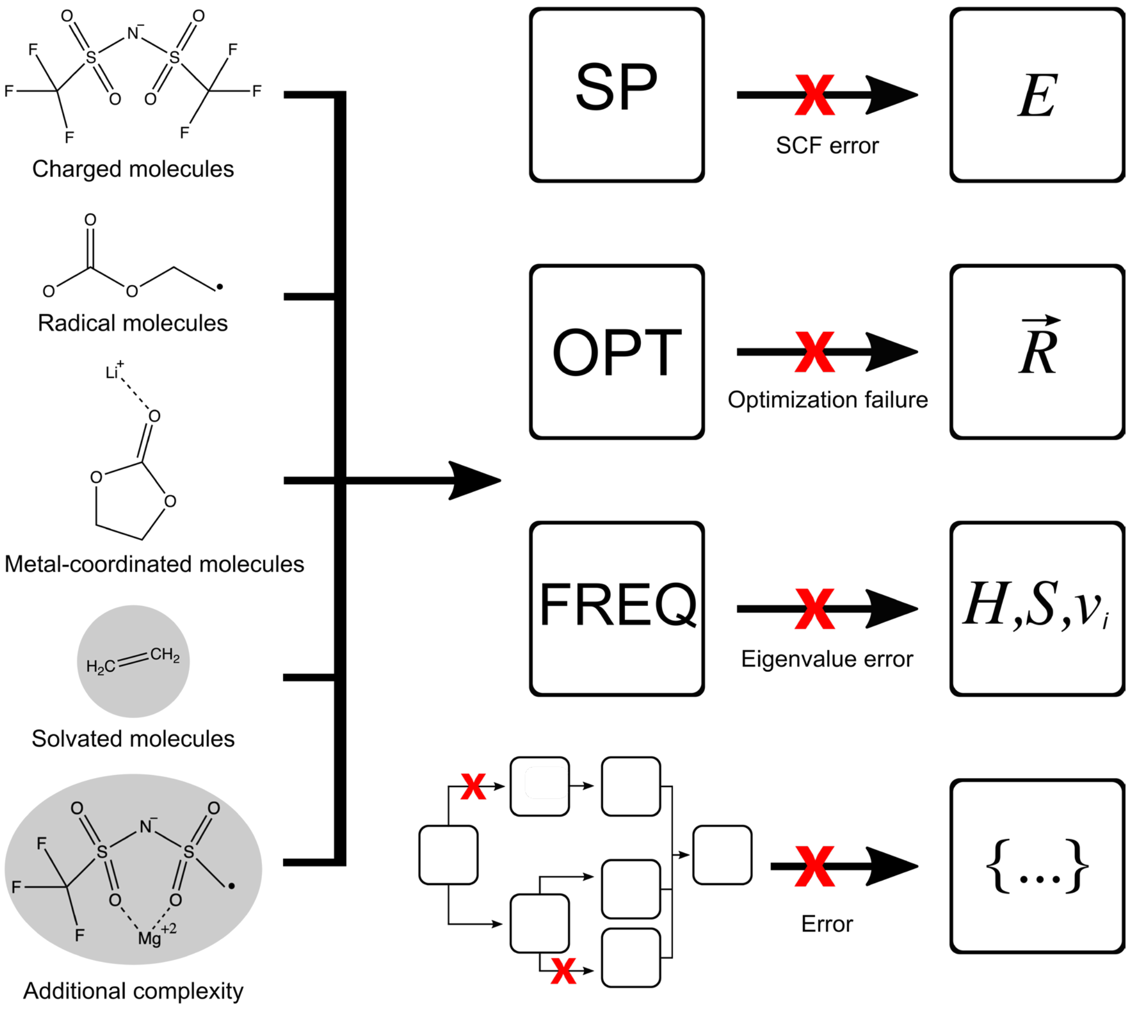
pymatgen

Custodian

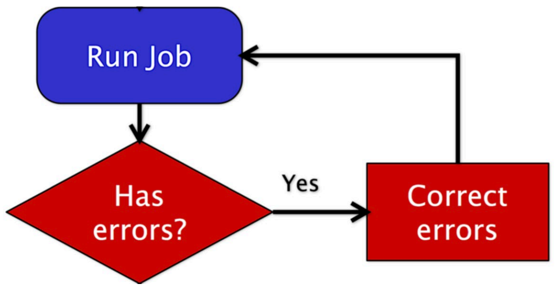
FireWorks

atomate

Q-CHEM
A QUANTUM LEAP INTO THE FUTURE OF CHEMISTRY



Typically, <75% success



With on-the-fly error correction:

>98% success

S. M. Blau*, E. Spotte-Smith*, B. Wood, S. Dwaraknath, K. A. Persson, *ChemRxiv* 2020

High-Throughput Molecular DFT Workflow Infrastructure

pymatgen

`pymatgen.io.qchem.inputs`:
- Input file generation & parsing

`pymatgen.io.qchem.outputs`:
- Output file parsing

`pymatgen.io.qchem.sets`:
- Default parameters

Custodian

`custodian.qchem.jobs`:
- Interface to Q-Chem
- Frequency-Flattening Optimization (FFOpt)

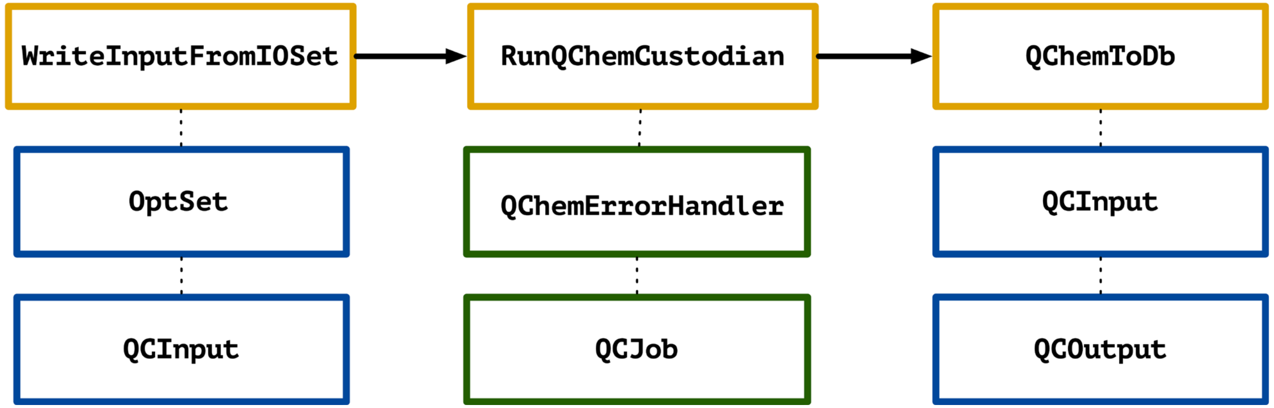
`custodian.qchem.handlers`:
- Automatic error handling

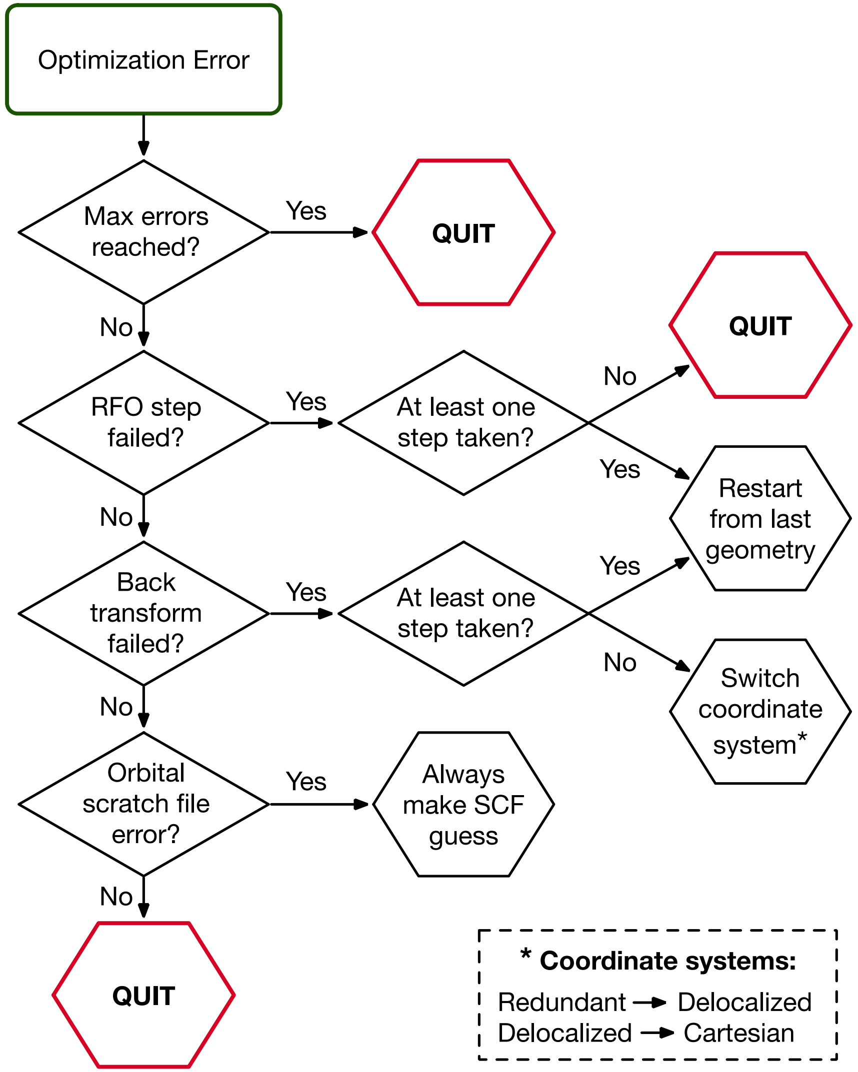
atomate

`atomate.qchem.firetasks`:
- Generating input files
- Executing with error handling
- Inserting data into DB

`atomate.qchem.fireworks`:
- Calculations, including FFOpt

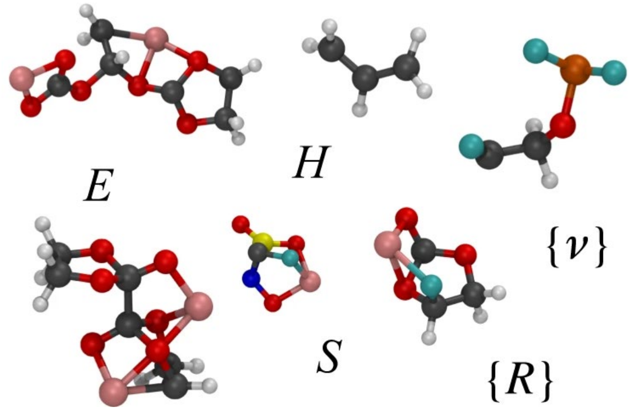
`OptimizeFW:`





S. M. Blau*, E. Spotte-Smith*, B. Wood, S. Dwaraknath, K. A. Persson, *ChemRxiv* 2020

We Use Workflows to Generate Unique Simulated Datasets

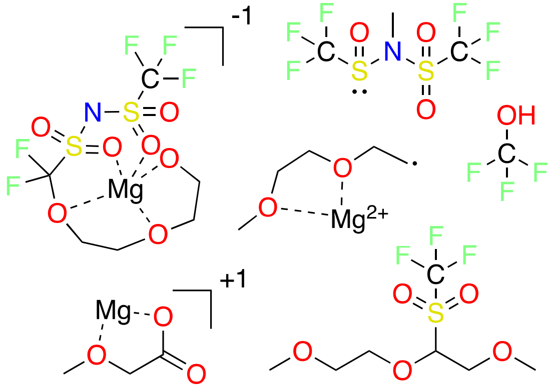


LIBE

Lithium-Ion Battery Electrolyte
17,190 molecules

E. W. C. Spotte-Smith*, **S. M. Blau***, et al., *Sci. Data* 2021

ωB97X-V/def2-TZVPPD/SMD



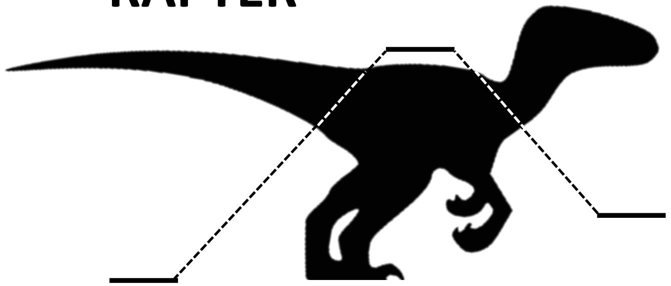
MADEIRA

MAgnesium **D**ataset of **E**lectrolyte and **I**nterphase **R**e**A**gents:
11,502 molecules

E. W. C. Spotte-Smith, **S. M. Blau**, et al., *JACS (accepted)*

ωB97X-V/def2-TZVPPD/SMD

RAPTER

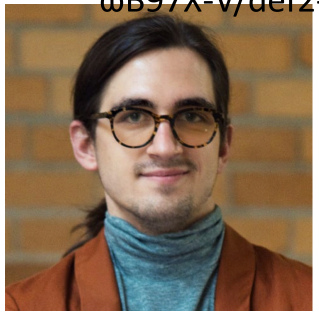


Re**A**ctants, **P**roducts, and **T**ransition-states of **E**lementary **R**eactions:
>15,000 complex reactions

E. W. C. Spotte-Smith, **S. M. Blau**, et al., *In preparation*

ωB97X-D/def2-SVPD/PCM

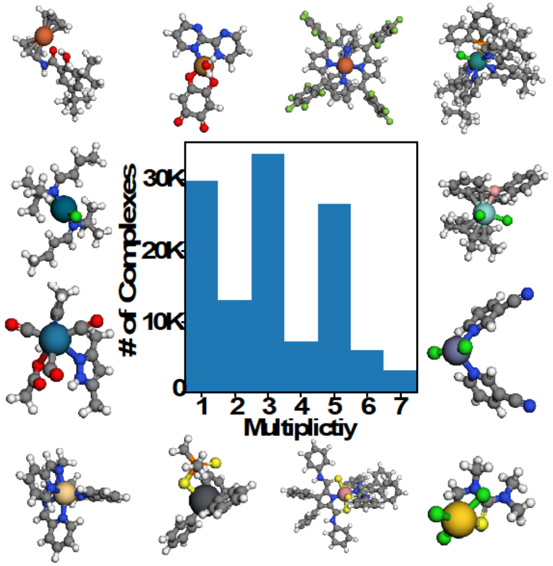
Collaborators:

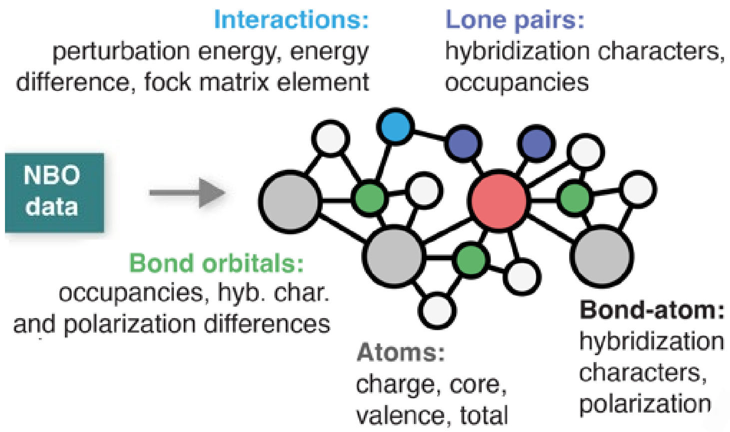


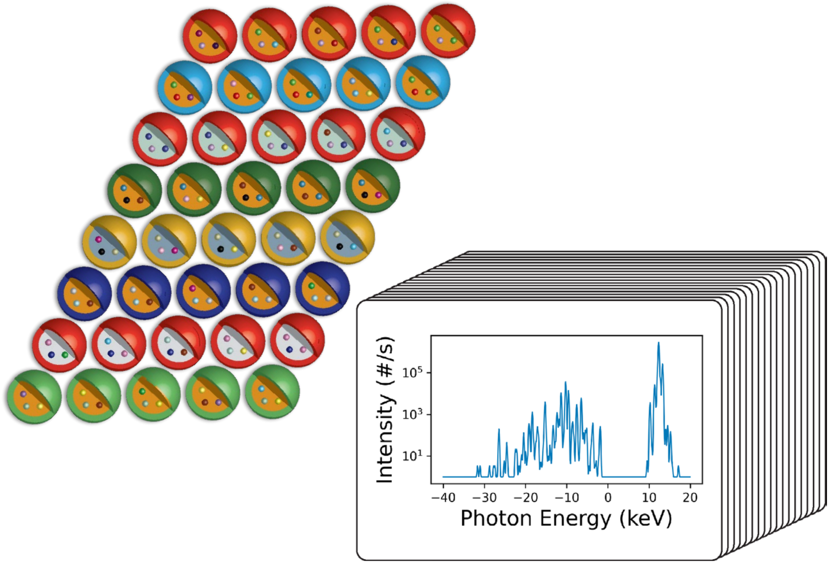
Evan Spotte-Smith

Kristin Persson

We Use Workflows to Generate Unique Simulated Datasets







ESCoMMS

Electronic **S**tructure of **Co**mplexes with **M**etals of **M**any **S**pins: **>140,000** complexes

ωB97M-V/def2-SVPD

ORIONS

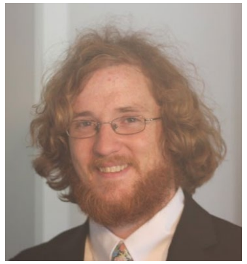
ORbital **I**nteractions of **O**rga**N**ic **S**pecies: **>230,000** molecules

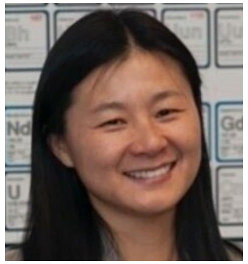
D. Boiko et al., *ChemRxiv* 2022

SUNSET

Simulated **U**pconverting **N**anoparticle **S**pectra for **E**missions **T**uning: **>6,000** spectra (kMC, not DFT)

Michael Taylor





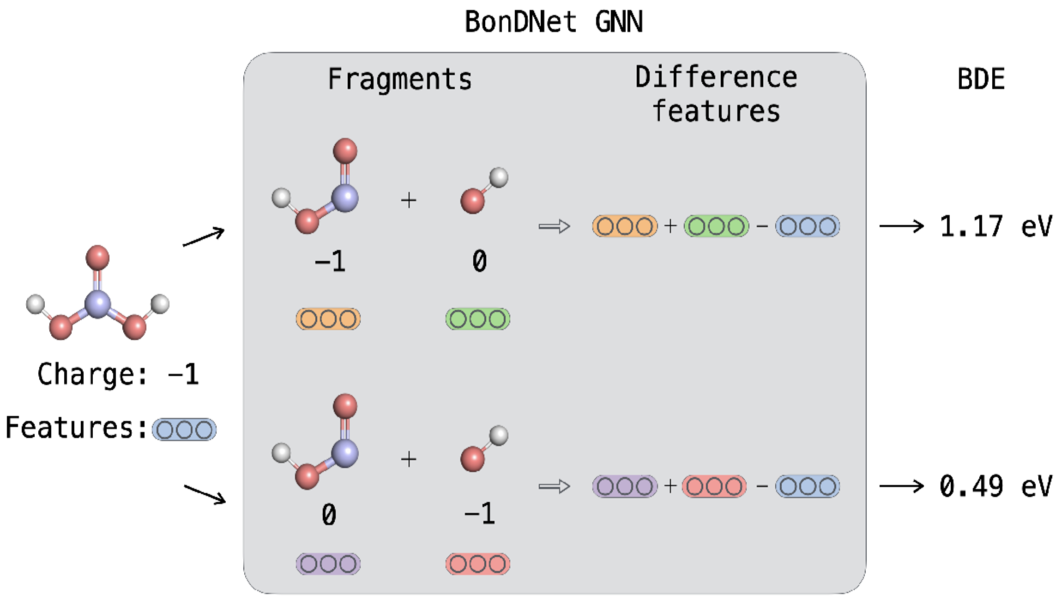
Ping Yang

Gabe Gomes

Eric Sivonxay

Emory Chan

Machine Learning Atop Our DFT Datasets



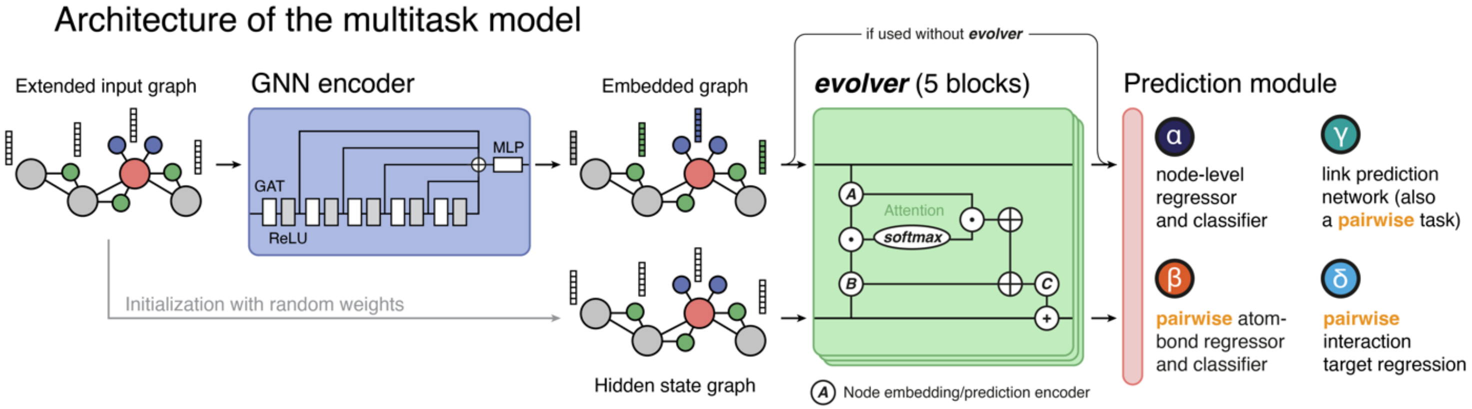
M. Wen, **S. M. Blau**, E. Spotte-Smith, S. Dwaraknath, K. A. Persson, *Chem. Sci.* 2021

An orbital-based representation for accurate quantum machine learning

Cite as: J. Chem. Phys. 156, 114101 (2022); doi: 10.1063/5.0083301
Submitted: 23 December 2021 • Accepted: 24 February 2022 •
Published Online: 15 March 2022

Konstantin Karandashev[1,a] and O. Anatole von Lilienfeld[1,2,b]

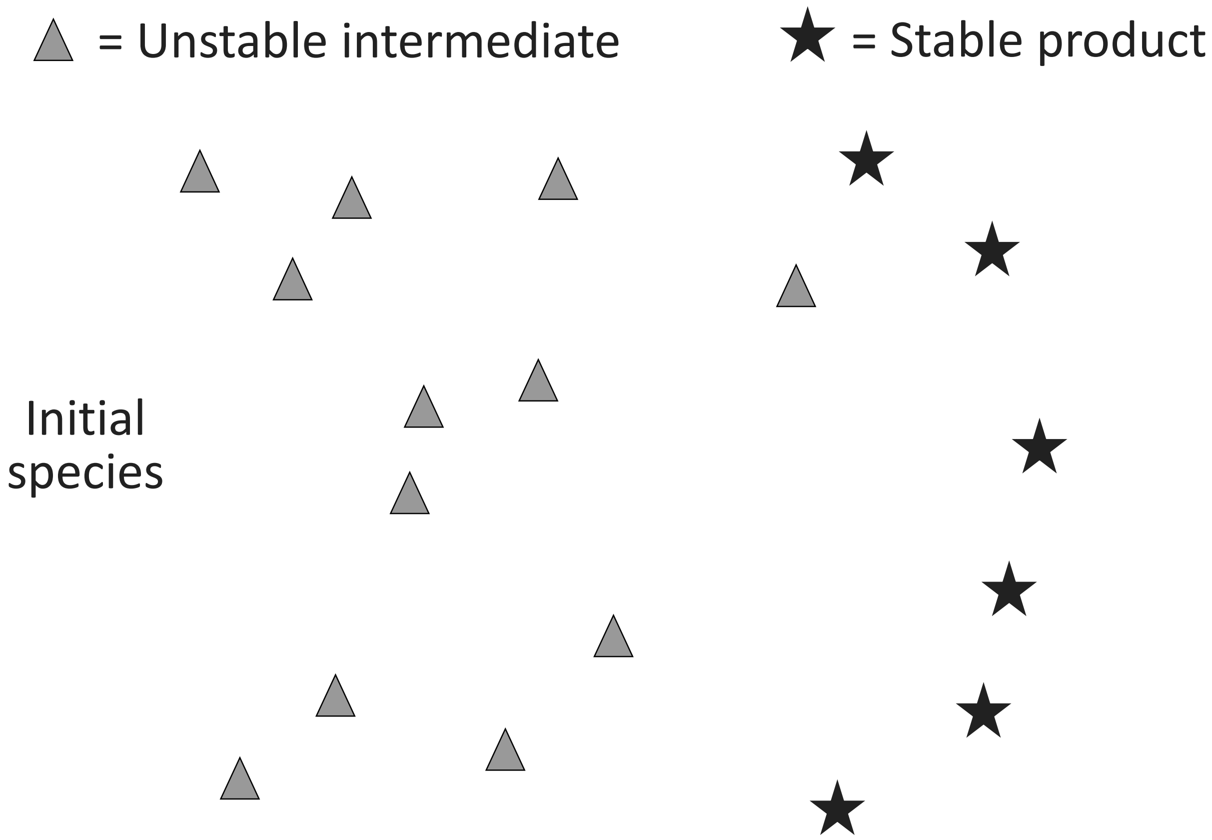
"The LIBE dataset is of particular interest… [because] it contains species of different charge and spin states, enabling us to test [our model]'s ability to process them…"



D. Boiko*, T. Reschützegger*, B. Sanchez-Lengeling, **S. M. Blau**, G. d. P. Gomes, *In preparation*

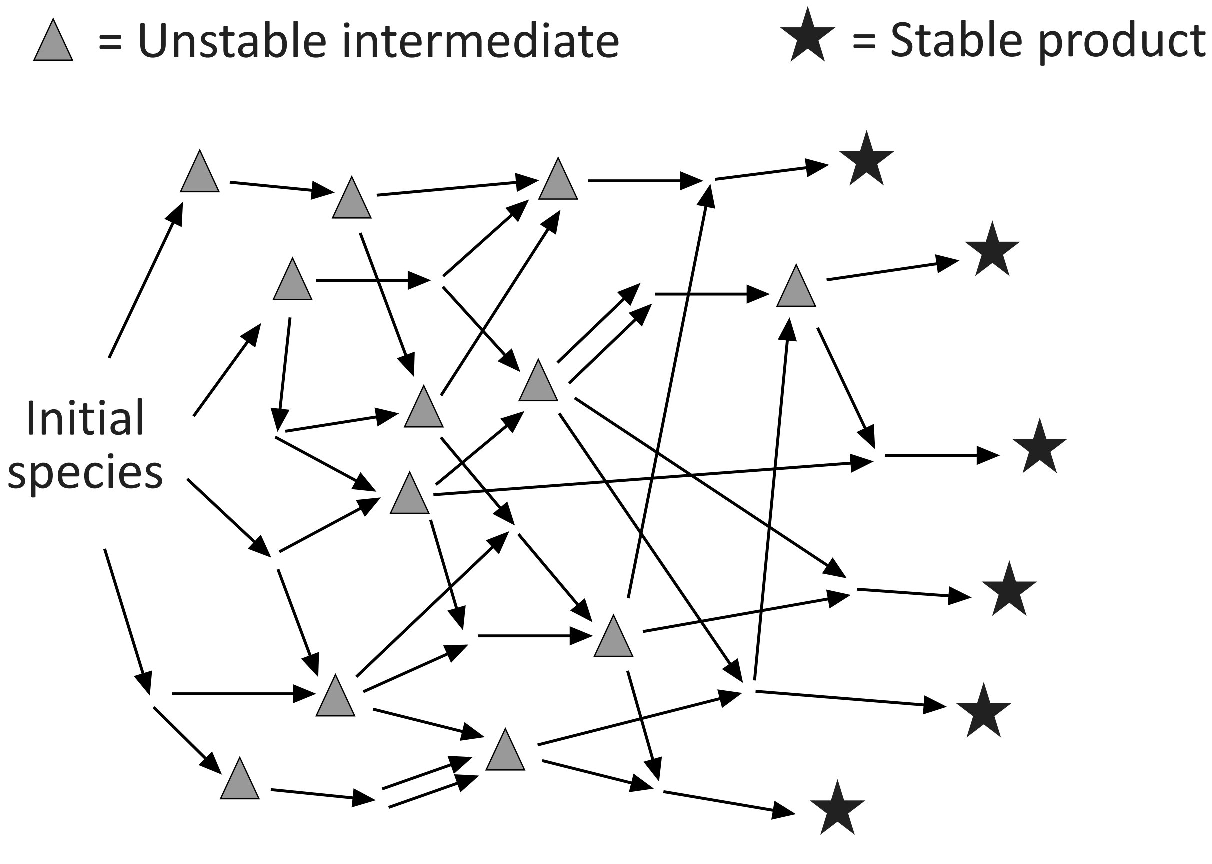
BERKELEY LAB

Introduction to Chemical Reaction Networks (CRNs)

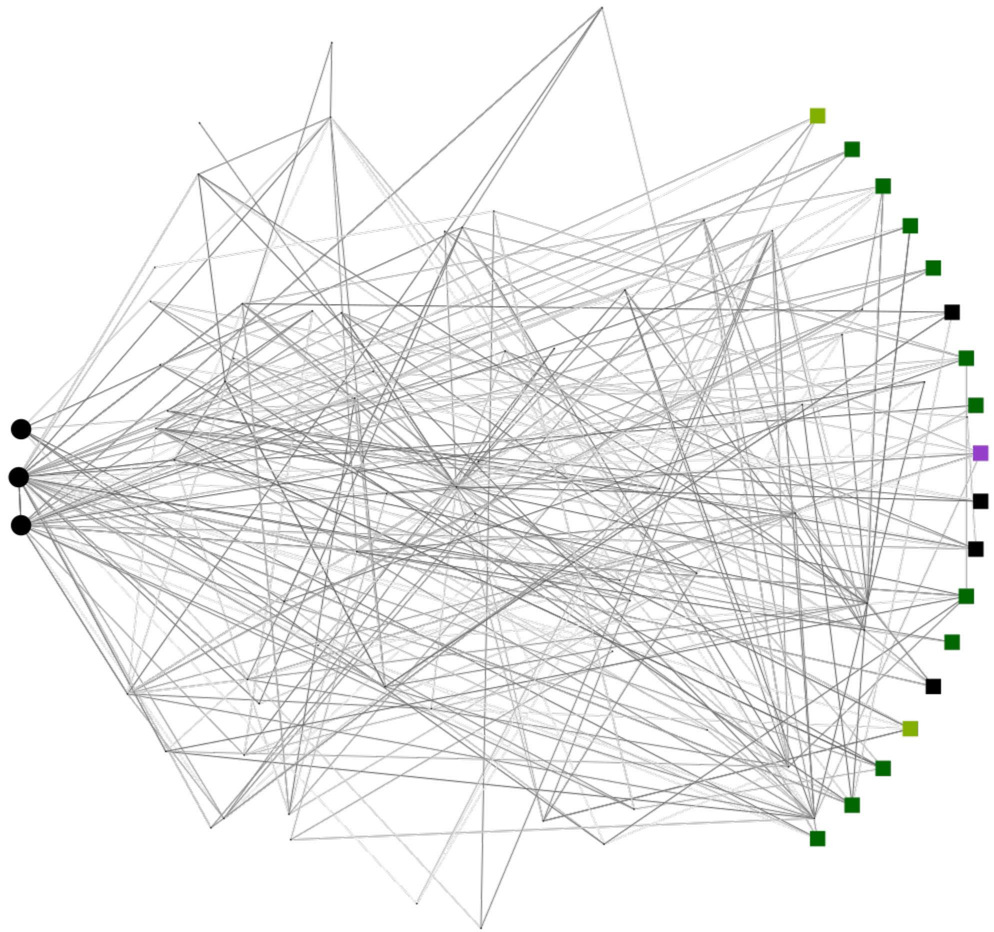


Chemical reaction network (CRN)

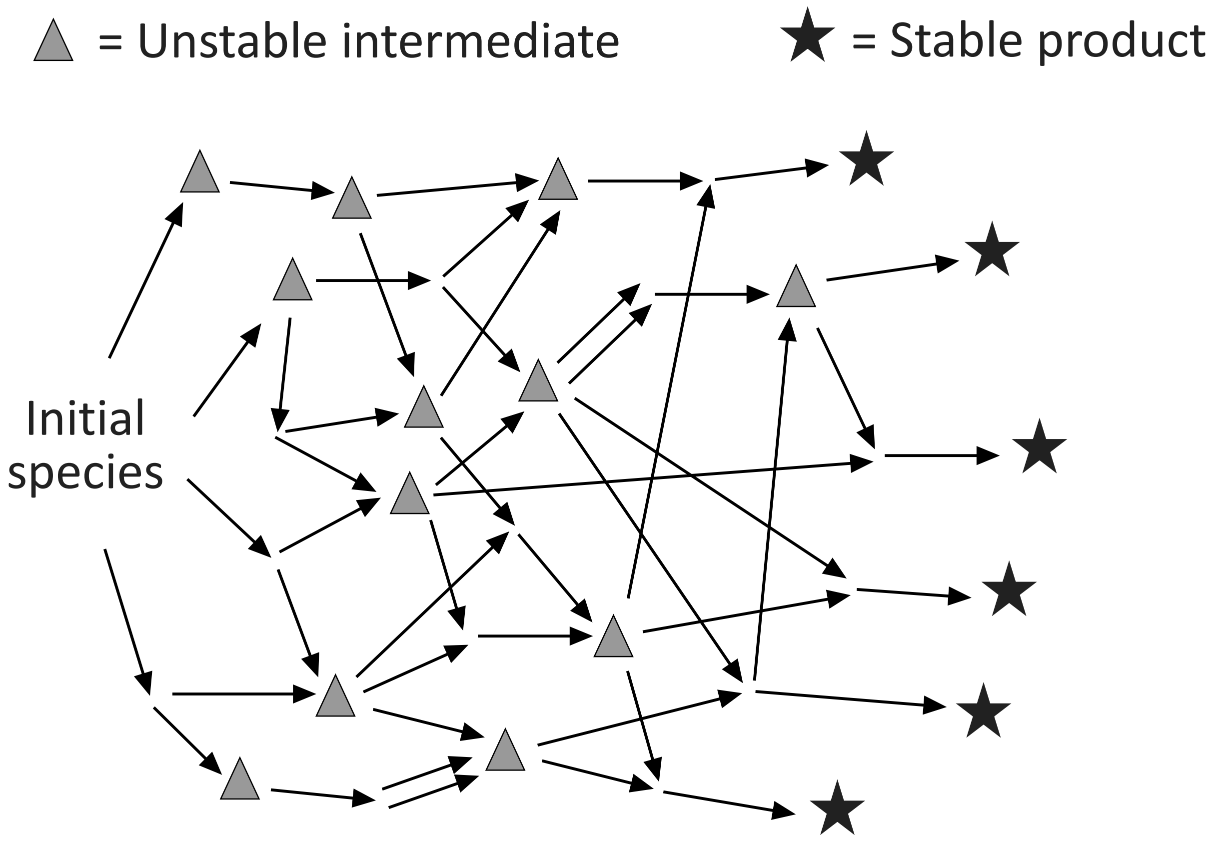
Introduction to Chemical Reaction Networks (CRNs)



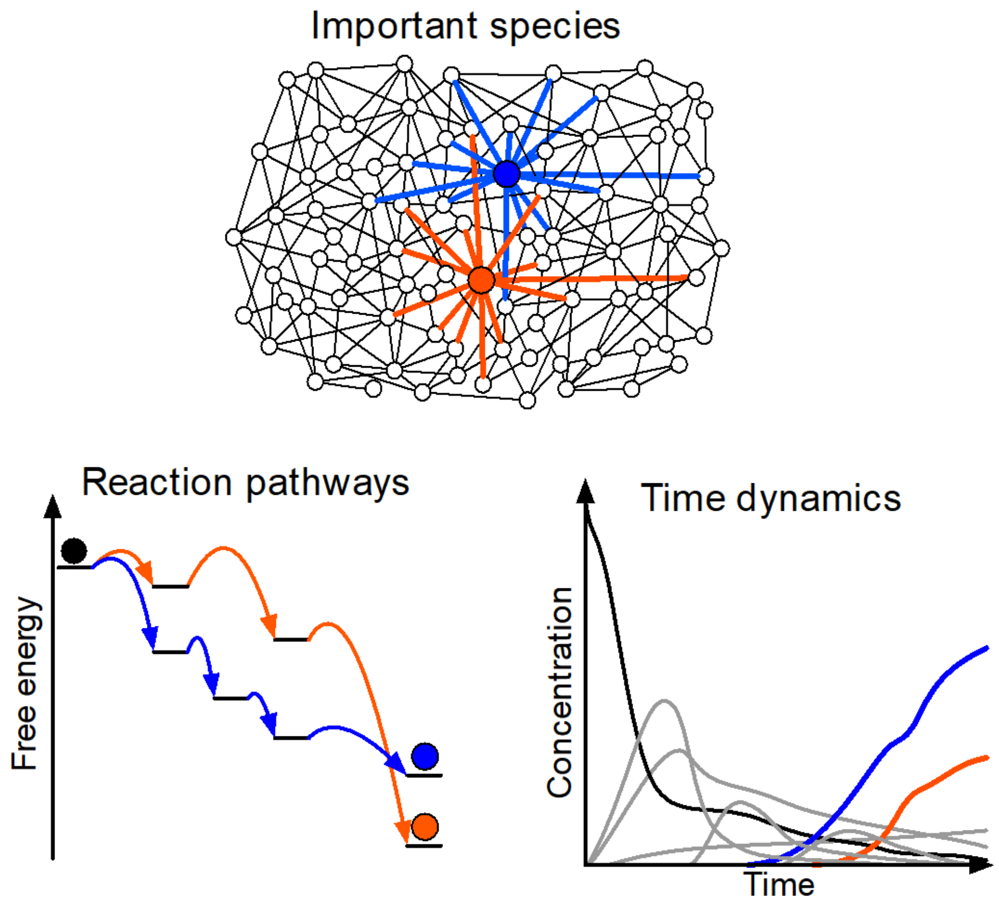
Chemical reaction network (CRN)



Introduction to Chemical Reaction Networks (CRNs)

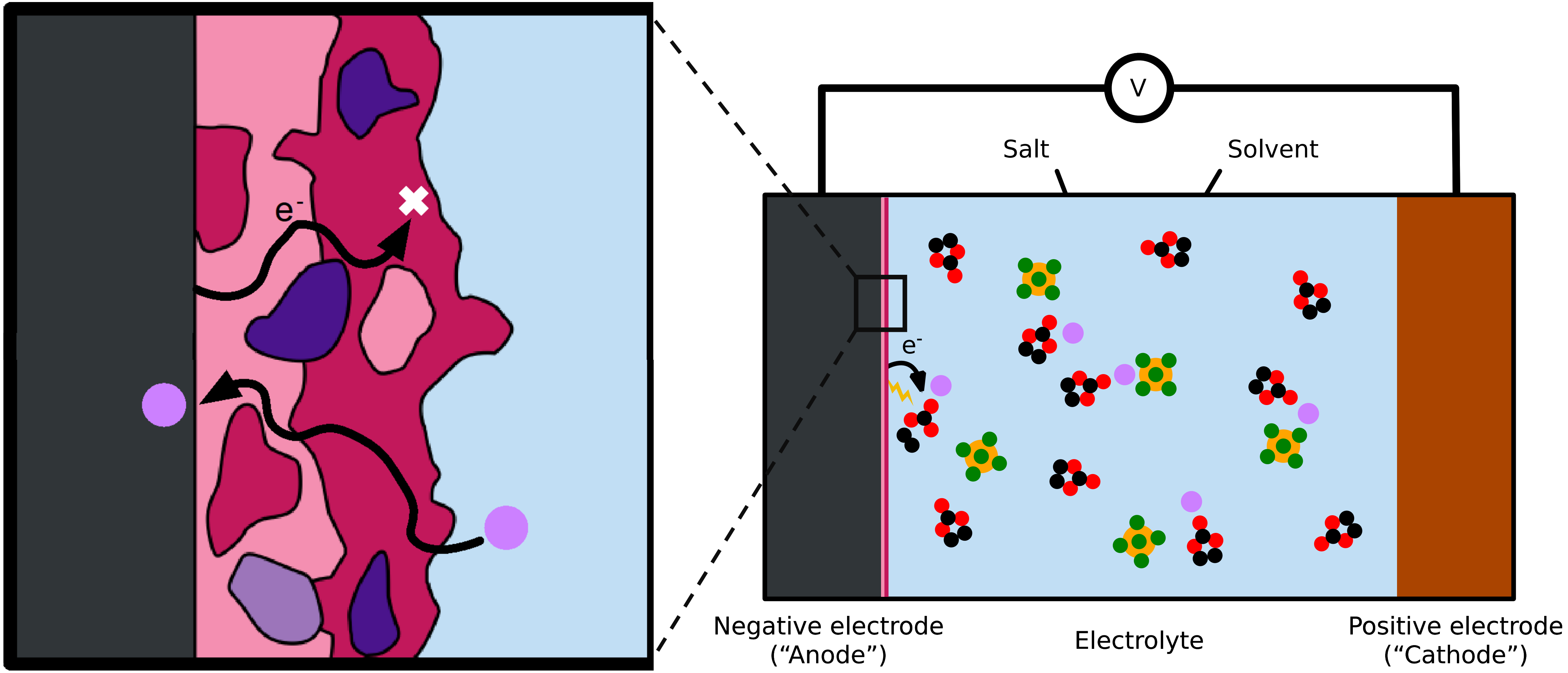


Chemical reaction network (CRN)

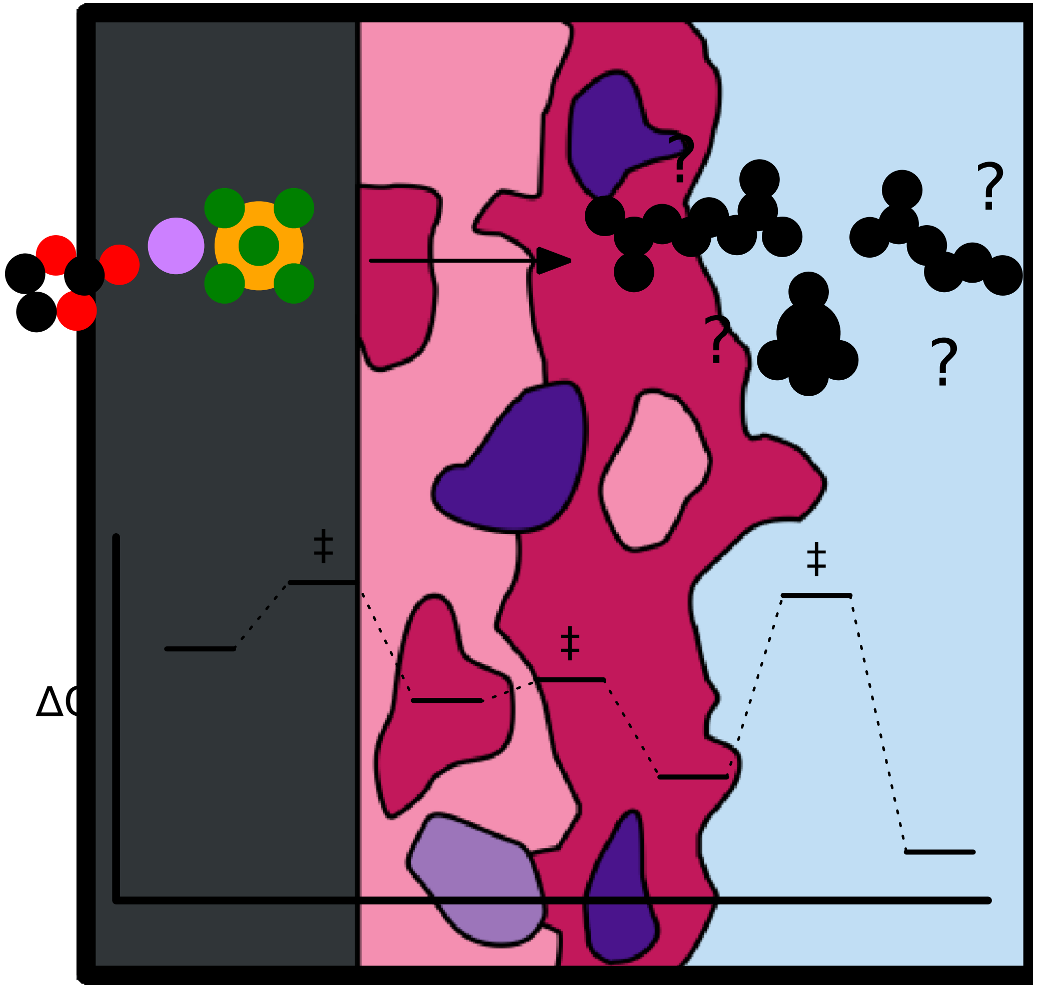


M. Wen, E. W. C. Spotte-Smith, **S. M. Blau**, M. J. McDermott, A. S. Krishnapriyan, K. A. Persson, *Nat. Comp. Sci.* 2023

Background: Solid-Electrolyte-Interphase Formation



Background: Solid-Electrolyte-Interphase Formation



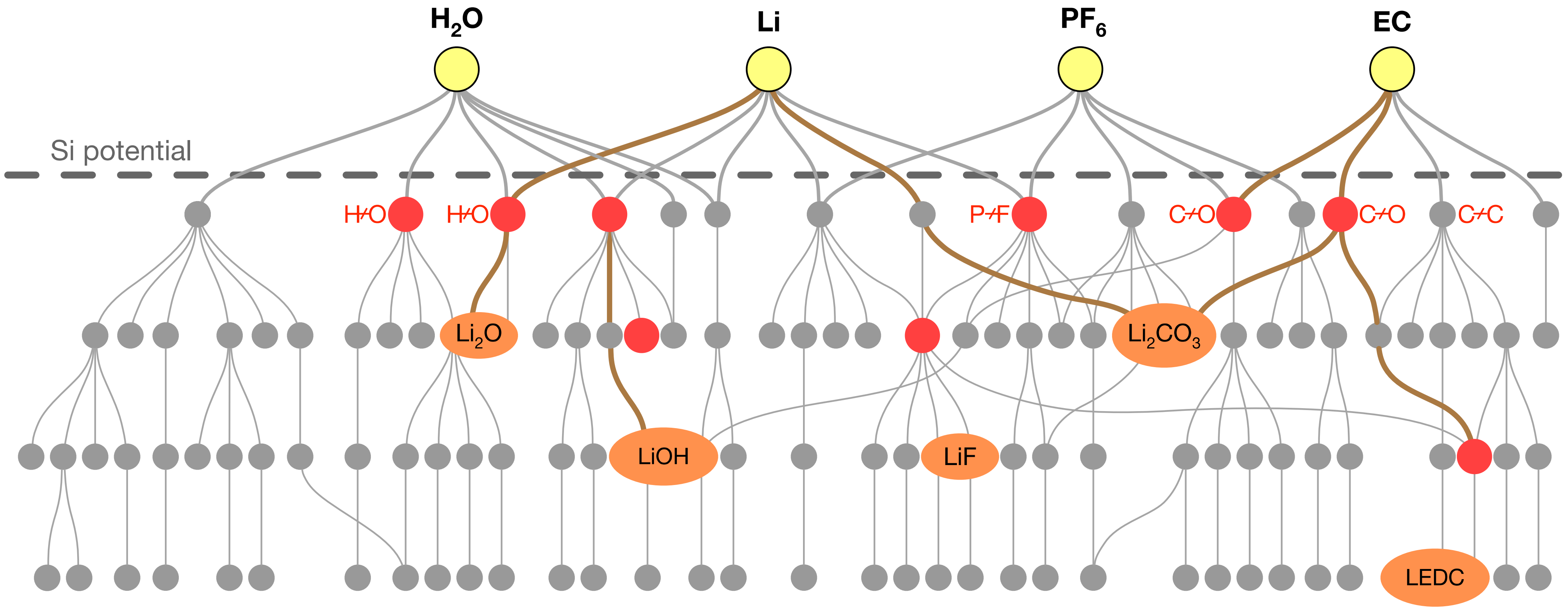
Goal: enable next-generation batteries by **controlling SEI formation**

Big questions:

1. What species form?
   - Identify products
2. How do those species form?
   - Reaction mechanisms
3. How do individual species, pathways compete and interact?

AIMD, by-hand DFT investigations: limited insight

A Data-Driven Approach to Understanding Reactivity



- Rational enumeration of possible species, reactions
- ΔG of each reaction in isolation via HT molecular DFT
- Network analysis: novel mechanistic insight
- Workflows necessary for data generation

High-Throughput Molecular DFT Data Generation

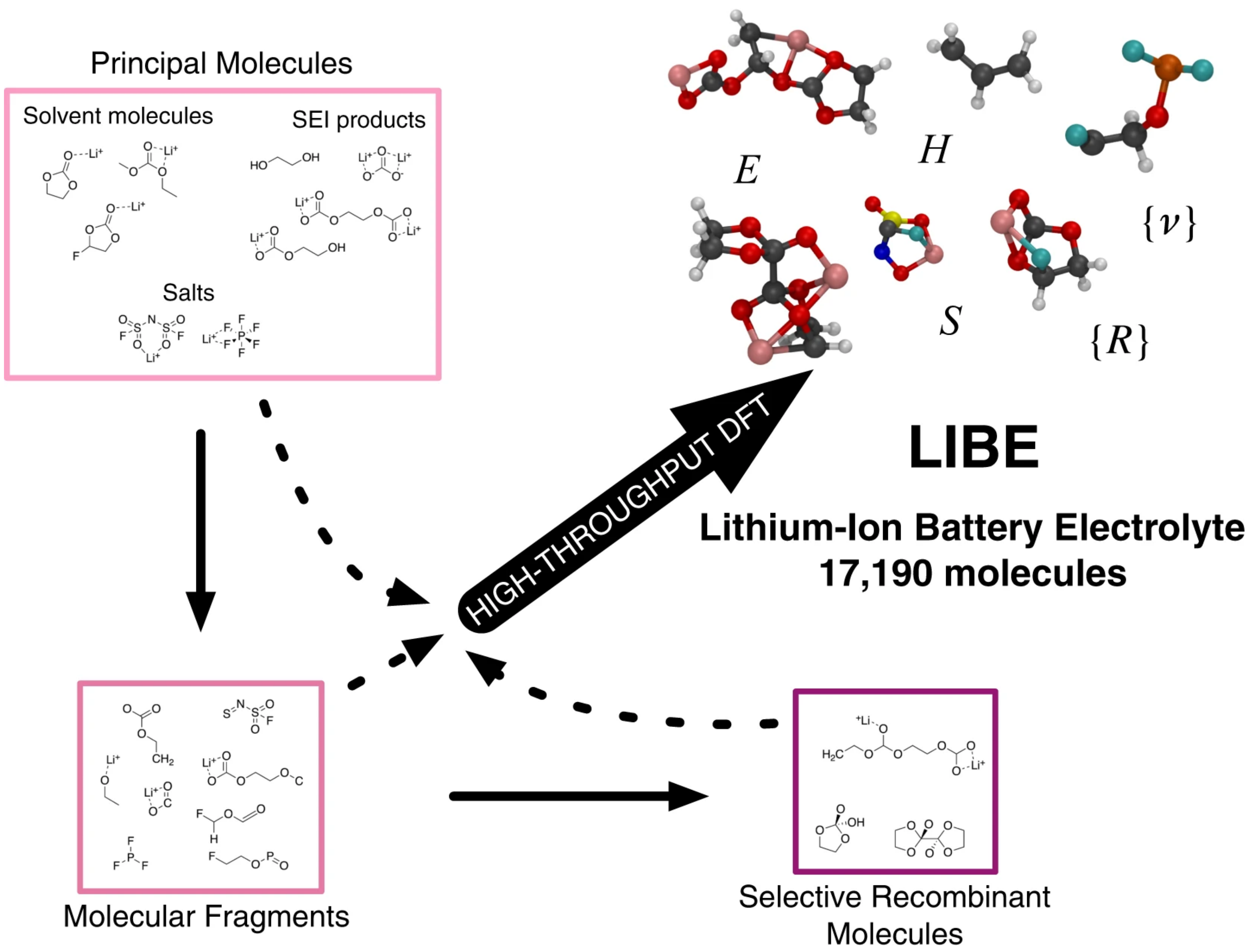
pymatgen

Custodian

FireWorks

atomate

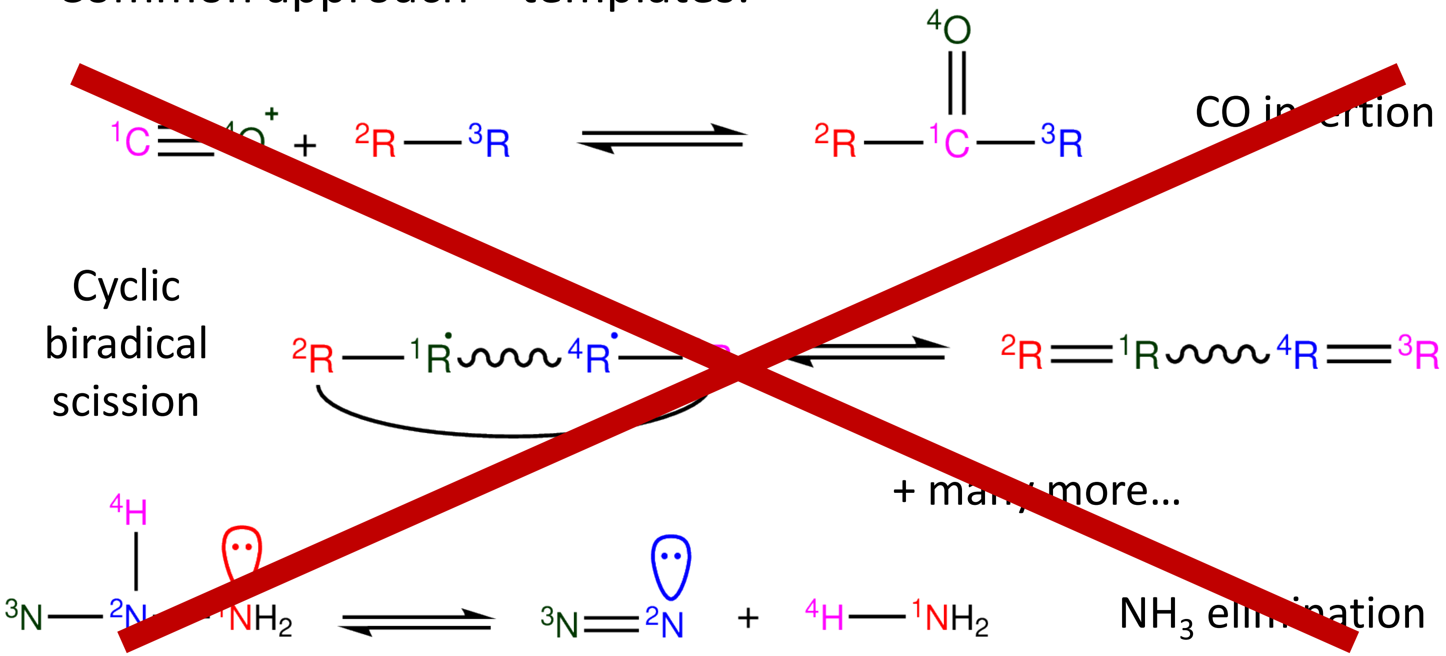
Q-CHEM
A QUANTUM LEAP INTO THE FUTURE OF CHEMISTRY



E. W. C. Spotte-Smith*, **S. M. Blau***, X. Xie, H. D. Patel, M. Wen, B. Wood, S. Dwaraknath, K. A. Persson, *Sci. Data* 2021

The Challenge of Reaction Generation

- Given e.g. 10k species, how to enumerate connecting reactions?
- Common approach – templates:



- Prescriptive templates are not well-suited to electron-driven chemistry

Our solution: filters

Goals:

- Minimize prescriptive constraints in order to *facilitate discovery*
- Want all reactions that:
  - Are likely to be single-step
  - May be kinetically viable
- Enable automated kinetic refinement
- Resolve complex competition

High-Performance Reaction Generation: HiPRGen

$\mathbf{S}_{init}$

Input: initial species

LIBE-CHOLi = 8904 species

$\mathbf{S}_{filtered}$ $\mathbf{R}_{filtered}$

Output: species, reactions that compose network

D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

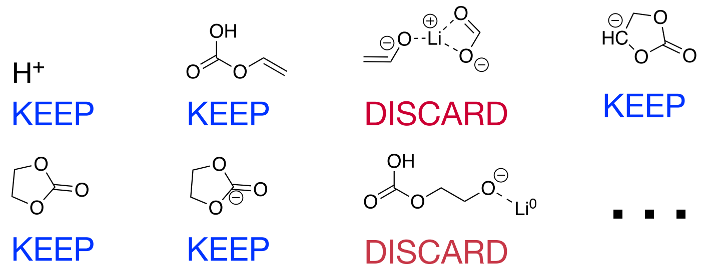
High-Performance Reaction Generation: HiPRGen

$\mathbf{S}_{init}$

Input: initial species

LIBE-CHOLi = 8904 species

1. Filter species



- Metal-centric complexes
- Li^0-containing species

After filtering = 5193 species

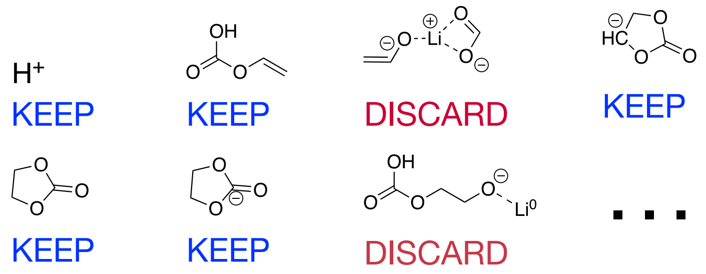
High-Performance Reaction Generation: HiPRGen

$\mathbf{S}_{init}$

Input: initial species

LIBE-CHOLi = 8904 species

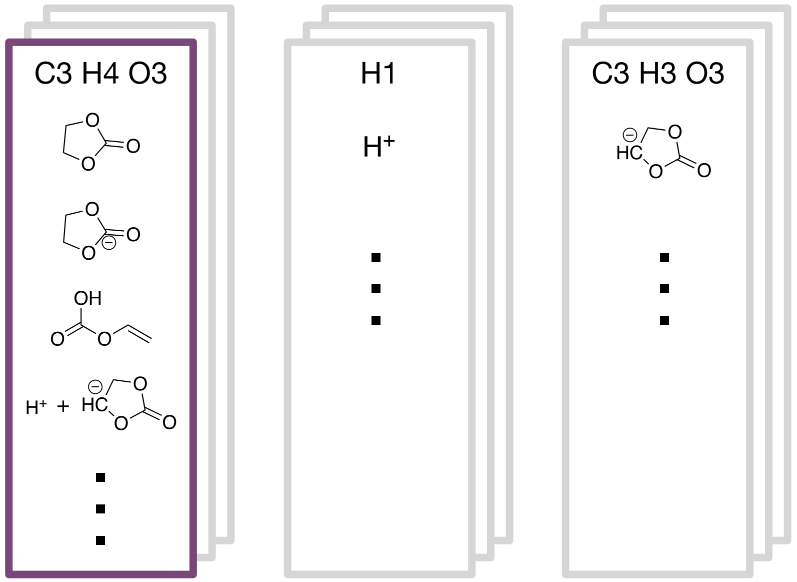
1. Filter species



- Metal-centric complexes
- Li^0-containing species

After filtering = 5193 species

2. Bucket Species by Composition



D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

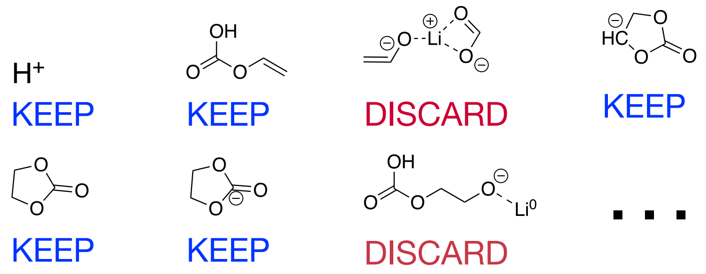
High-Performance Reaction Generation: HiPRGen

S_{init}

Input: initial species

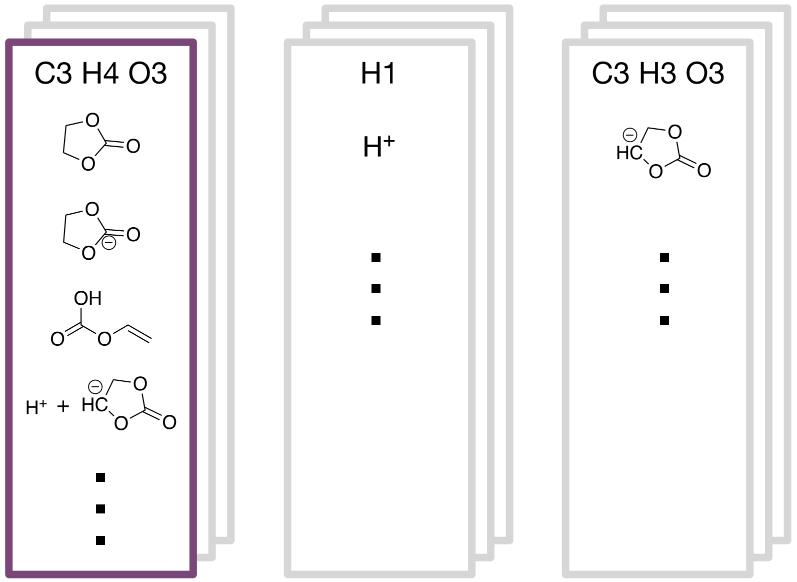
LIBE-CHOLi = 8904 species

1. Filter species

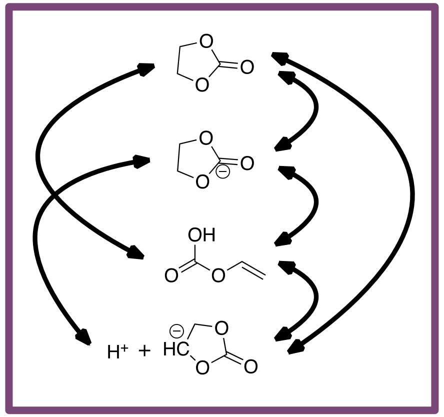


After filtering = 5193 species

2. Bucket Species by Composition



3. Generate reactions by stoichiometry



> 176 billion rxns

D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

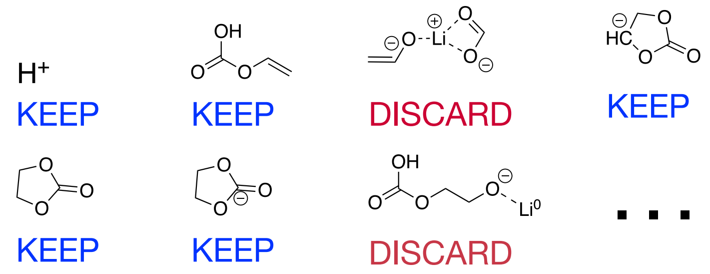
High-Performance Reaction Generation: HiPRGen

S_{init}

Input: initial species

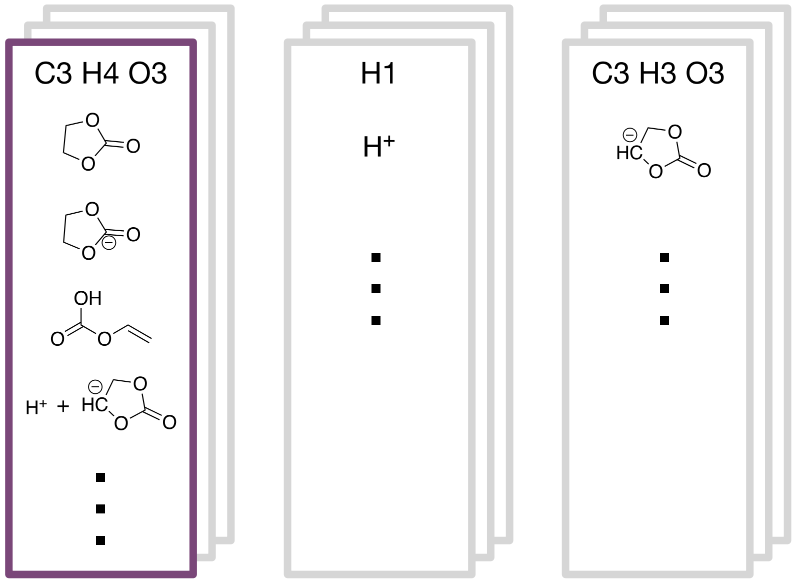
LIBE-CHOLi = 8904 species

1. Filter species

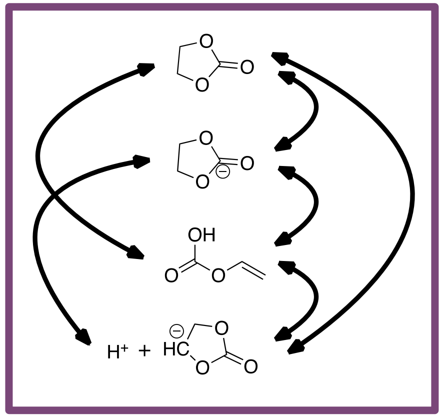


After filtering = 5193 species

2. Bucket Species by Composition

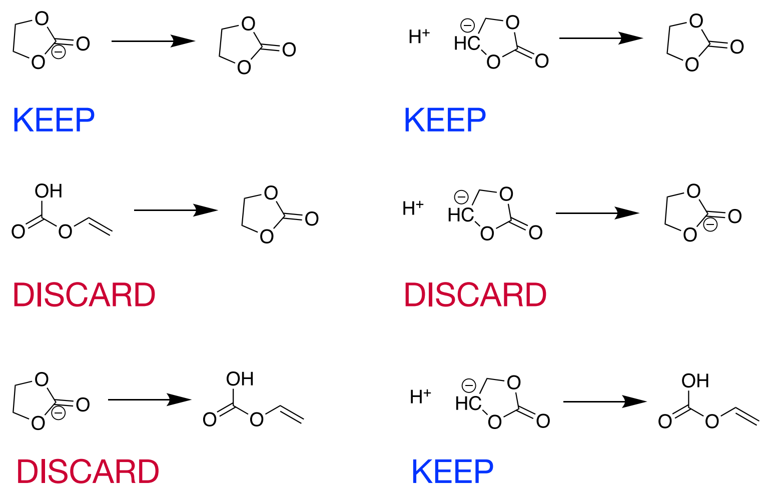


3. Generate reactions by stoichiometry



> 176 billion rxns

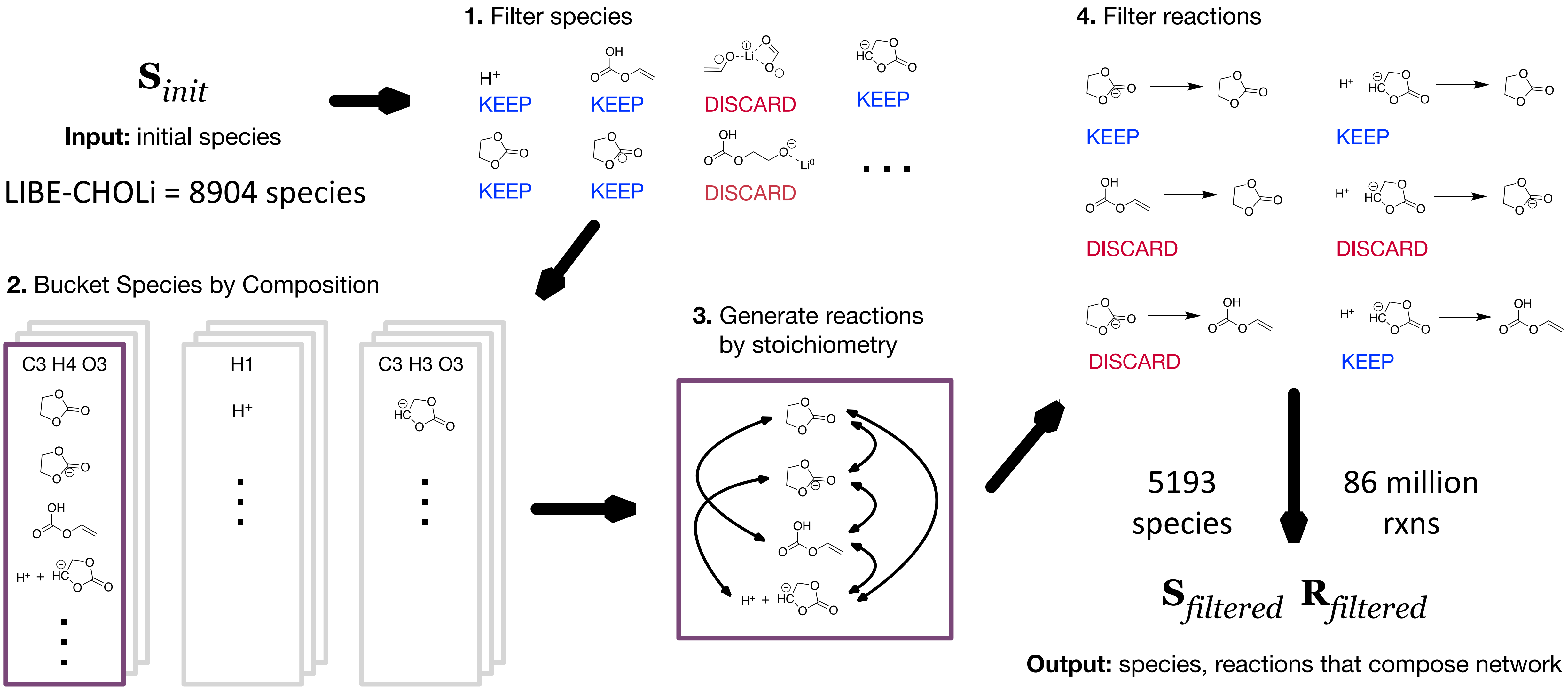
4. Filter reactions



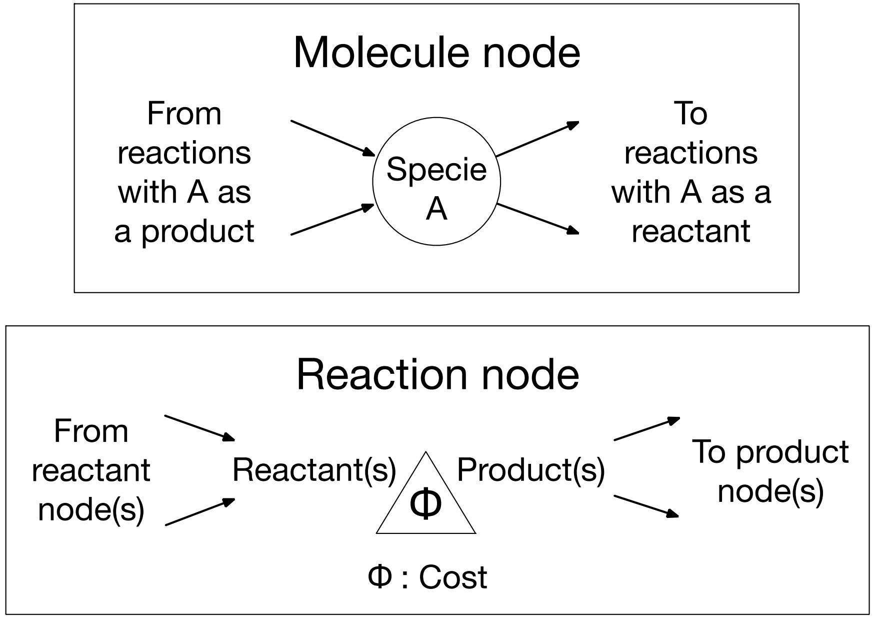
- Too many bonds changing
- Bond change + redox
- Coordination + covalent bond change

D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

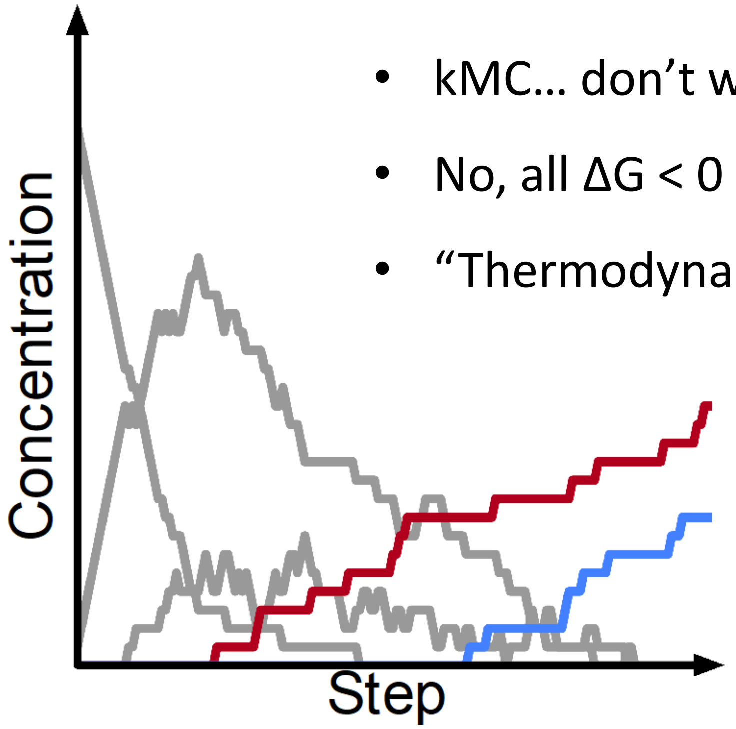
High-Performance Reaction Generation: HiPRGen



Reaction Network Analysis: Graphs vs Kinetic Monte Carlo



S. M. Blau, H. D. Patel, E. Spotte-Smith, X. Xie, S. Dwaraknath, K. A. Persson, *Chem. Sci.* 2021



- kMC… don't we need kinetics?
- No, all ΔG < 0 rxns, all same rate
- "Thermodynamically bounded"

- No concept of system state / concentrations
- Pathfinding to a given species scales as $O(N^2)$
- **Must know target of interest a priori**

- Need initial state, evolve full system stepwise
- Stochastic sampling scales as O(logN) + parallelizable
- **Target prediction from full system exploration…?**

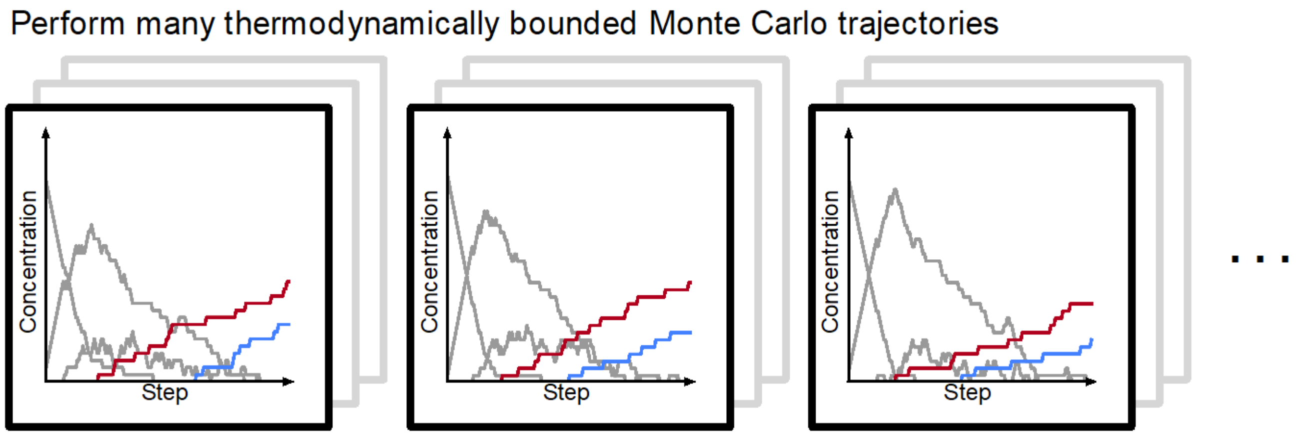
Reaction Network Monte Carlo: RNMC

Inputs:

$\mathbf{S}_{filtered}$

$\mathbf{R}_{filtered}$

$[x_i, x_j, \ldots]_0$



- 30 of each x_i
- All ΔG < 0: can run to completion
- 100k trajectories

D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

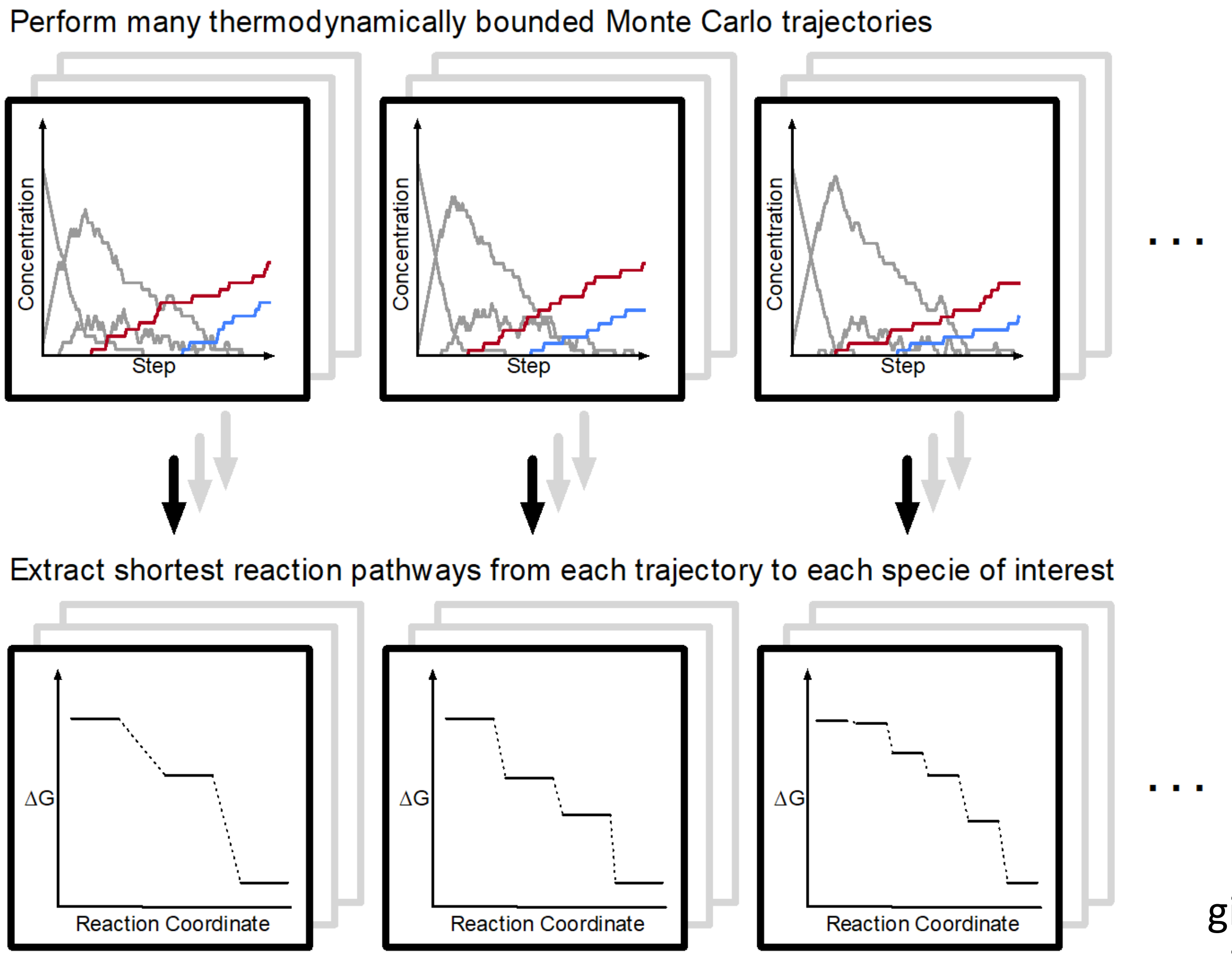
Reaction Network Monte Carlo: RNMC

Inputs:

$\mathbf{S}_{filtered}$

$\mathbf{R}_{filtered}$

$[\mathrm{x}_i, \mathrm{x}_j, \ldots]_0$



- 30 of each x_i
- All ΔG < 0: can run to completion
- 100k trajectories

Can do pathfinding on up to approx. 300 million reactions

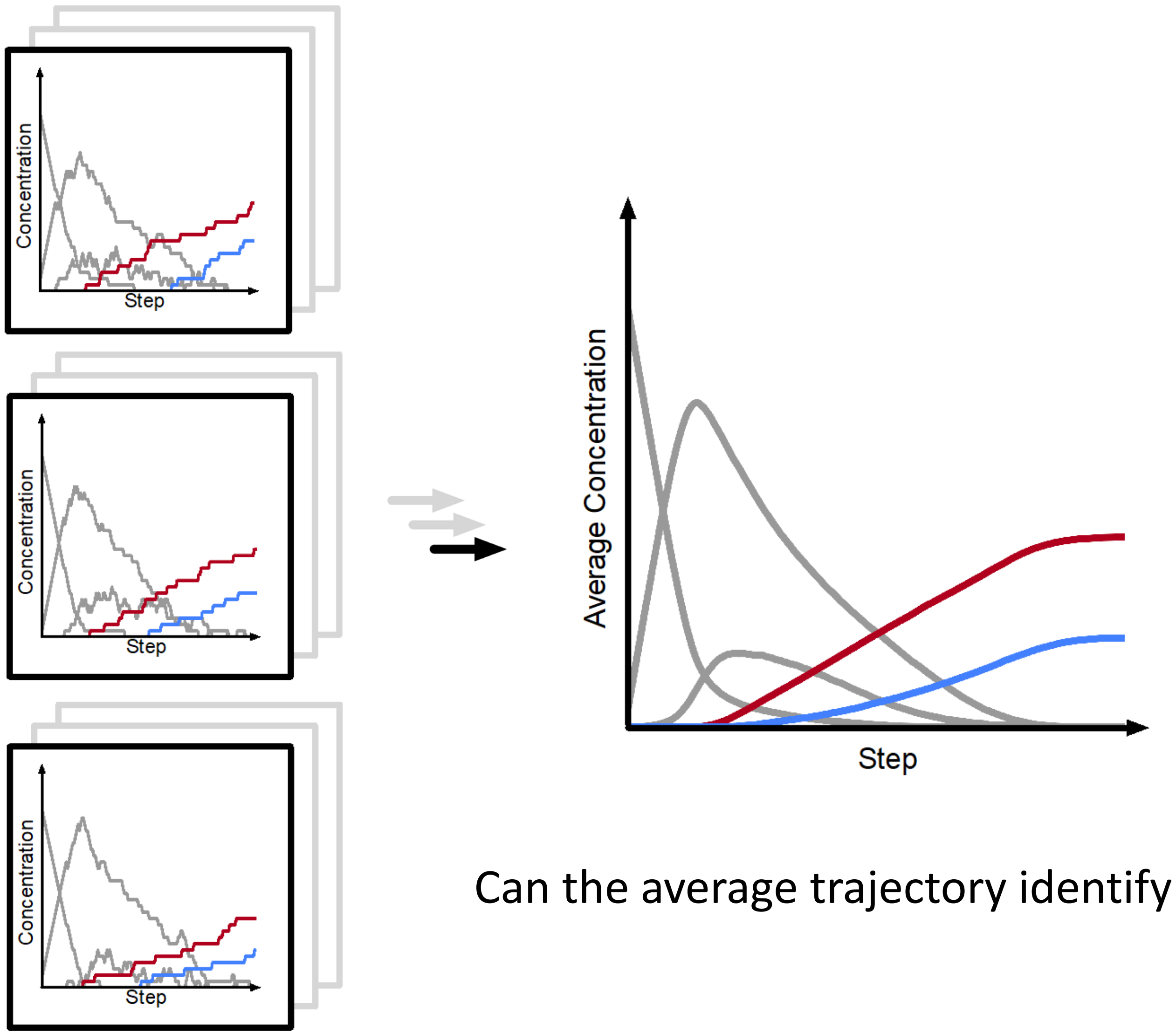
Group and order by cost
Φ ≈ # of reactions

github.com/BlauGroup/HiPRGen
github.com/BlauGroup/RNMC

D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

BERKELEY LAB

Converging RNMC and Identifying Network Products

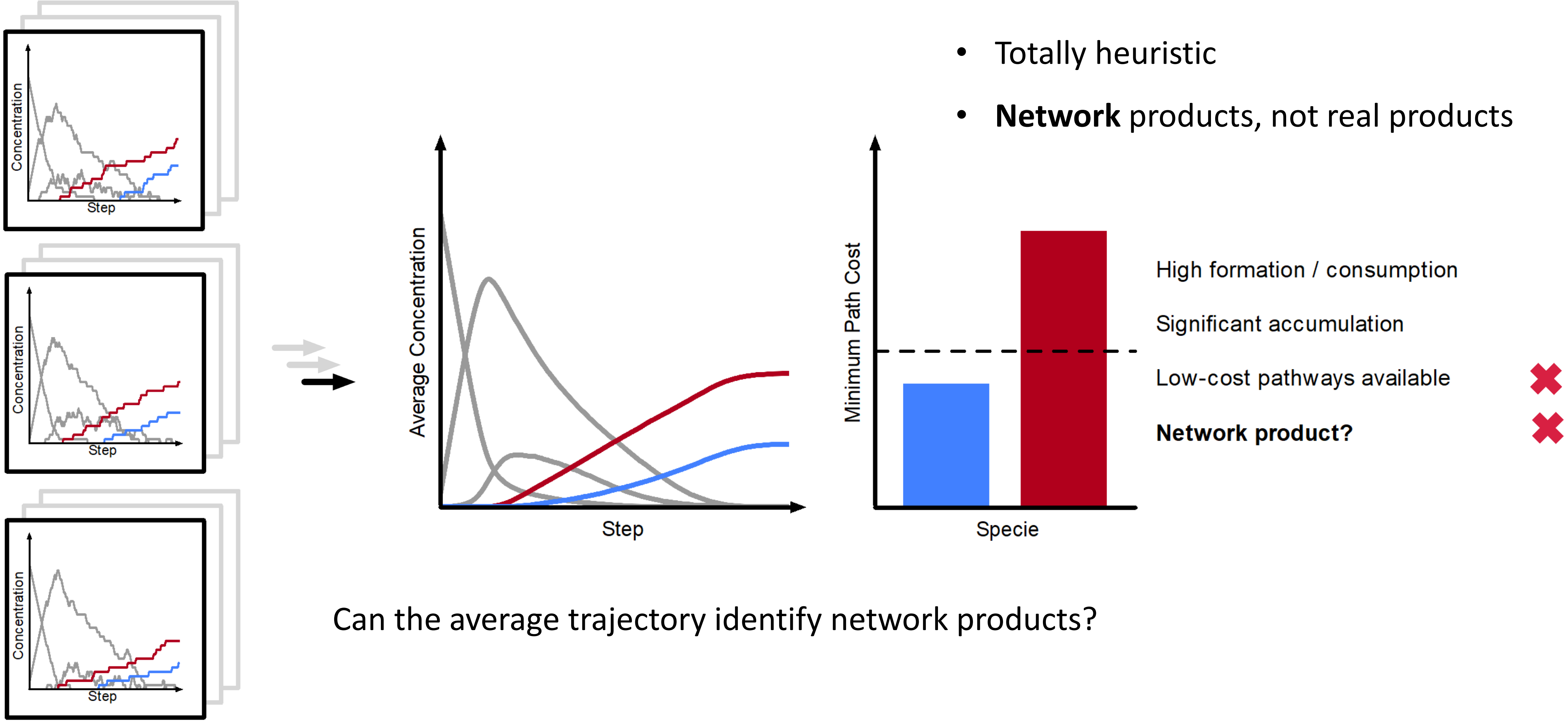


Can the average trajectory identify network products?

D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

Converging RNMC and Identifying Network Products

- Totally heuristic
- **Network** products, not real products



Can the average trajectory identify network products?

D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

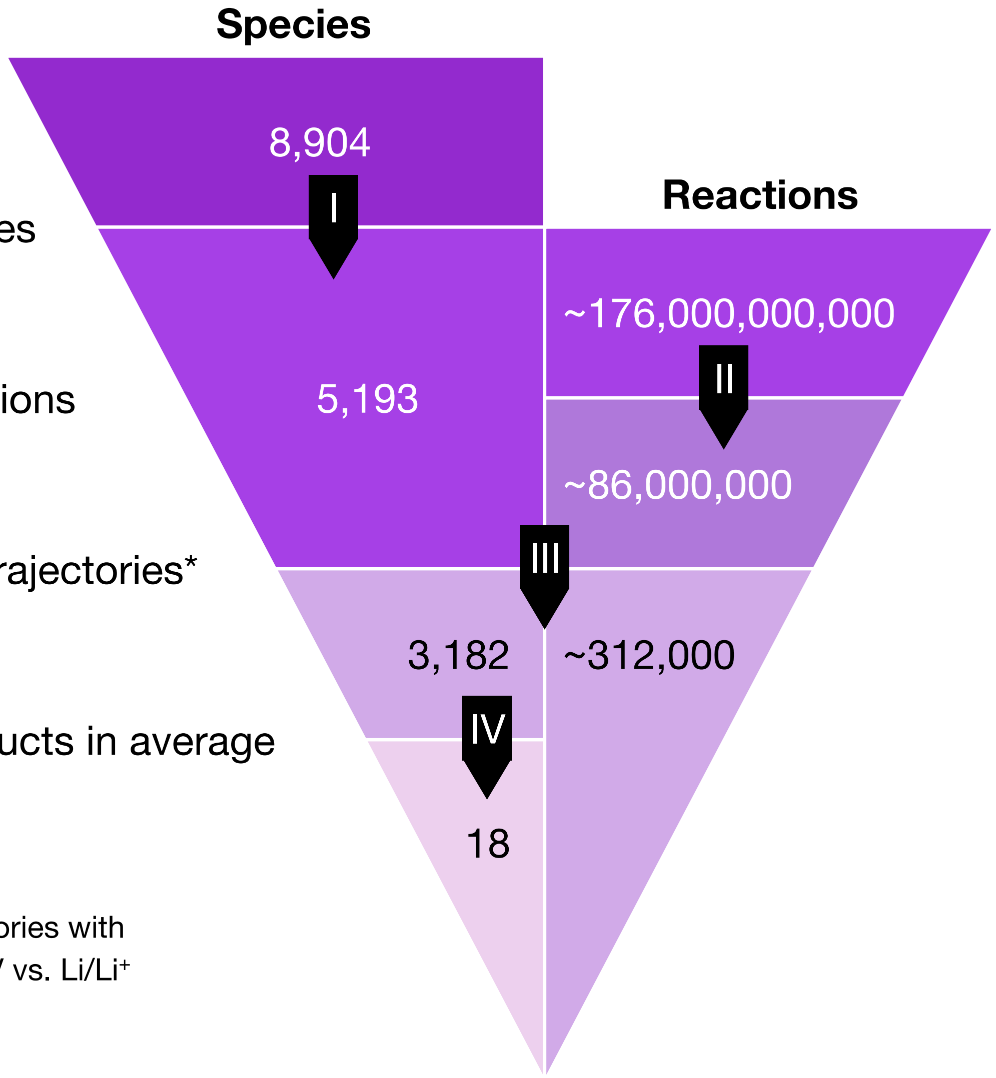
Building Up and Picking Apart Complexity

I: Filter species

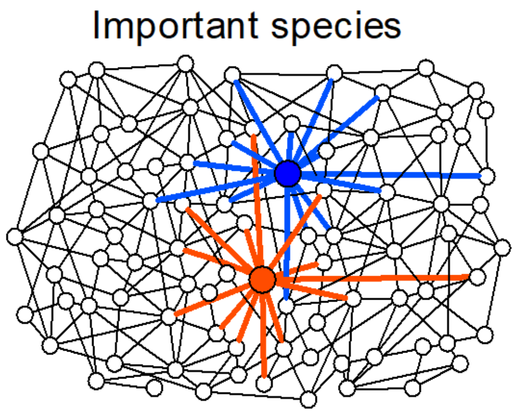
II: Filter reactions

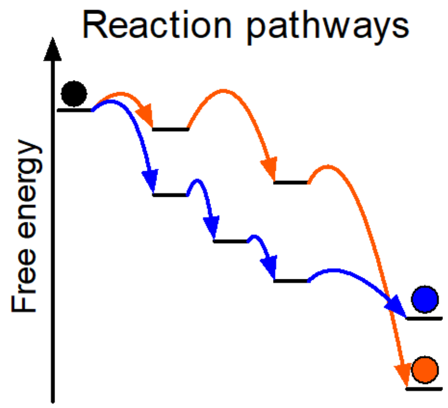
III: Run MC trajectories*

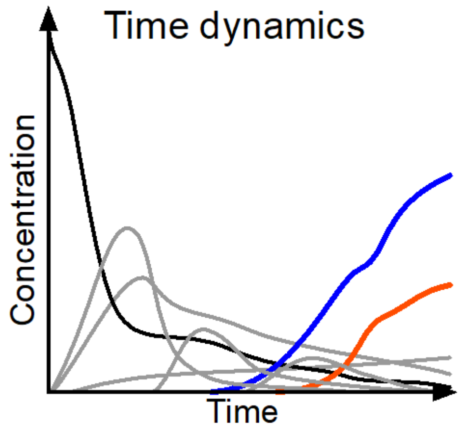
IV: Find products in average
trajectory



* 100,000 trajectories with
Li^+, EC, CO_2 at 0V vs. Li/Li^+







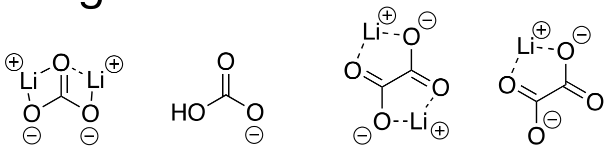
D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

Predicted Battery Network Products: 36 out of 5139

[EC, Li^+] and [EC, Li^+, CO_2] at 0V and +0.5V vs. Li/Li^+

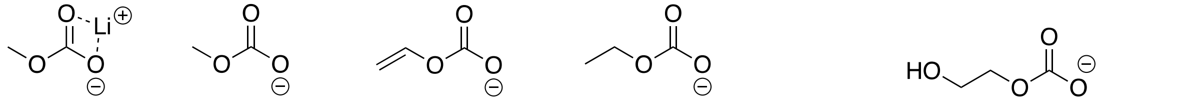
Small molecules/gases

Inorganics

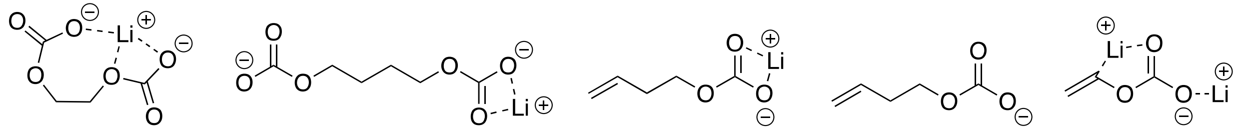


lithium carbonate lithium oxalate

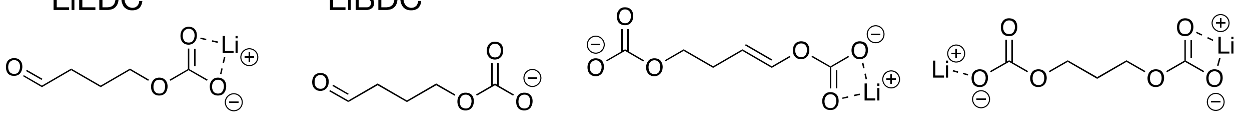
Alkyl carbonates



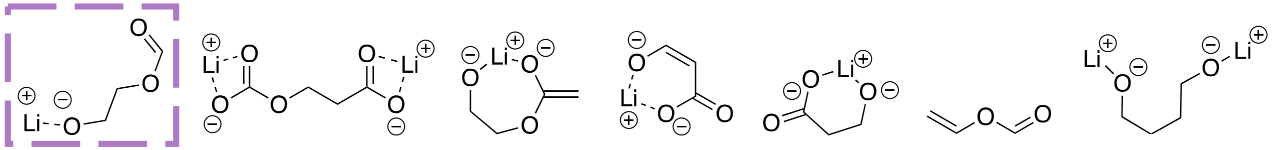
LMC vinyl carbonate ethylene monocarbonate



$LiEDC^-$ $LiBDC^-$

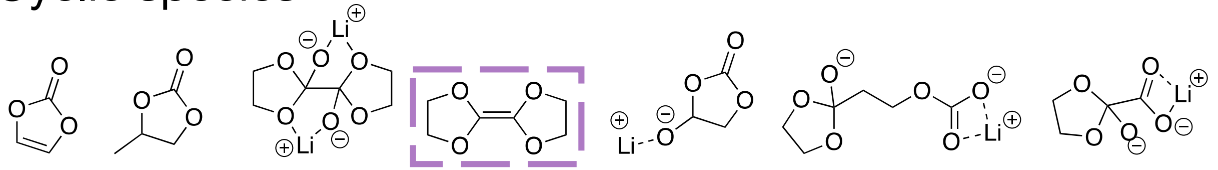


Carboxylates, esters, and oxides



LFEO

Cyclic species

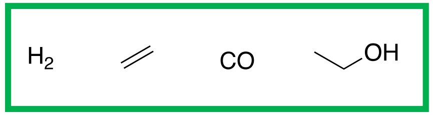


bi-dioxolylidene

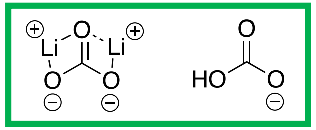
D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

Predicted Battery Network Products: 36 out of 5139

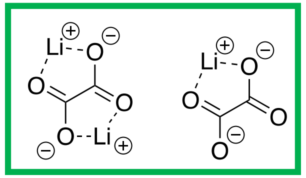
Small molecules/gases



Inorganics



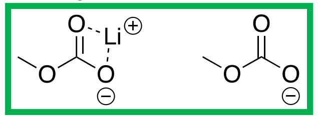
lithium carbonate



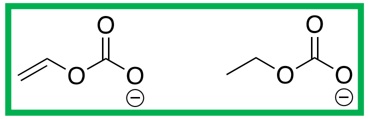
lithium oxalate

[EC, Li^+] and [EC, Li^+, CO_2] at 0V and +0.5V vs. Li/Li^+

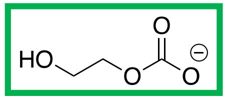
Alkyl carbonates



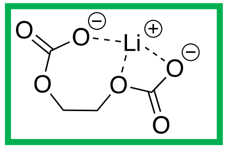
LMC



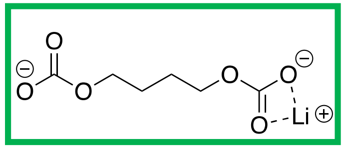
vinyl carbonate



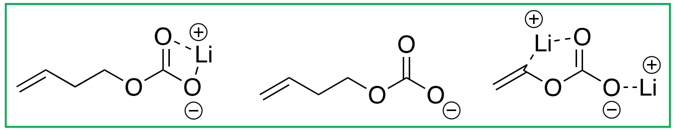
ethylene monocarbonate

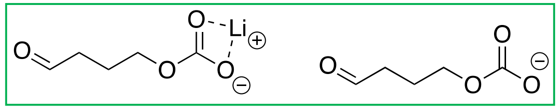


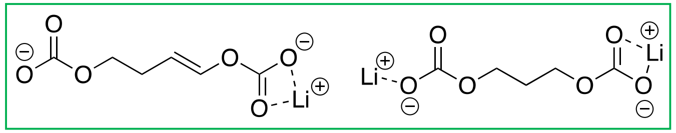
$LiEDC^-$



$LiBDC^-$

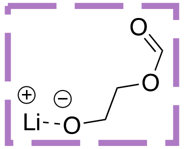




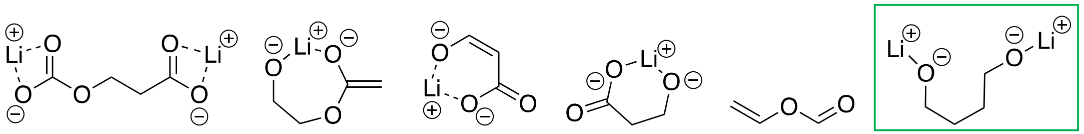


- Recovered nearly all observed or proposed molecular SEI components
- Only thermodynamics – unexpectedly effective!
- So about those particularly weird molecules…

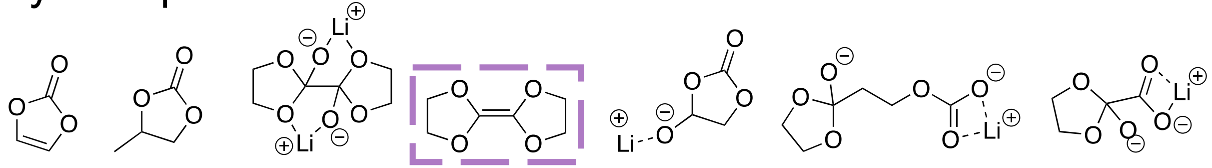
Carboxylates, esters, and oxides



LFEO



Cyclic species



bi-dioxolylidene

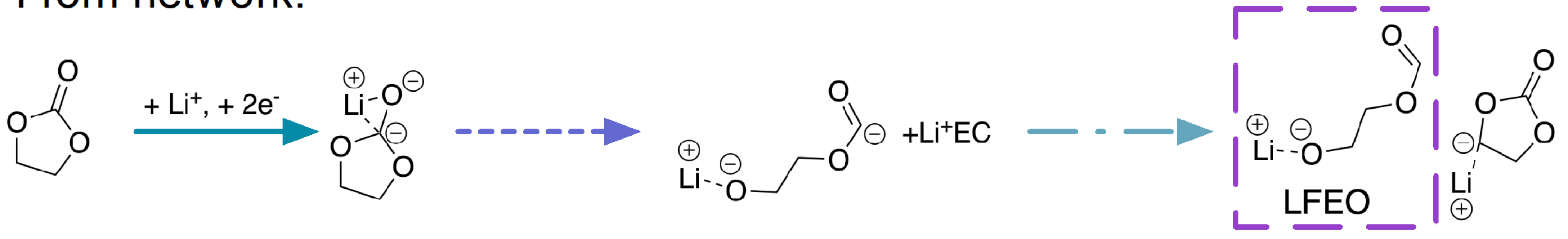
D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

BERKELEY LAB

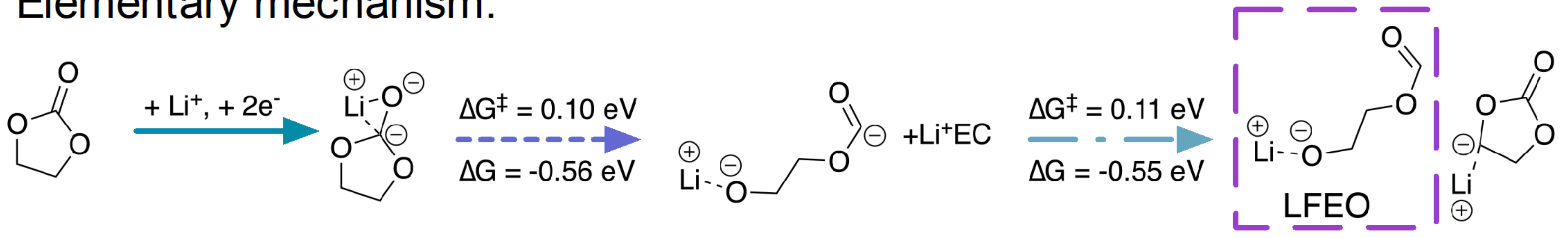
Network Path to Refined Mechanism: LFEO

Applied semi-automated TS procedure to 15 shortest thermo. paths – 12th shortest with [Li^+, EC] at 0V vs Li/Li^+:

From network:

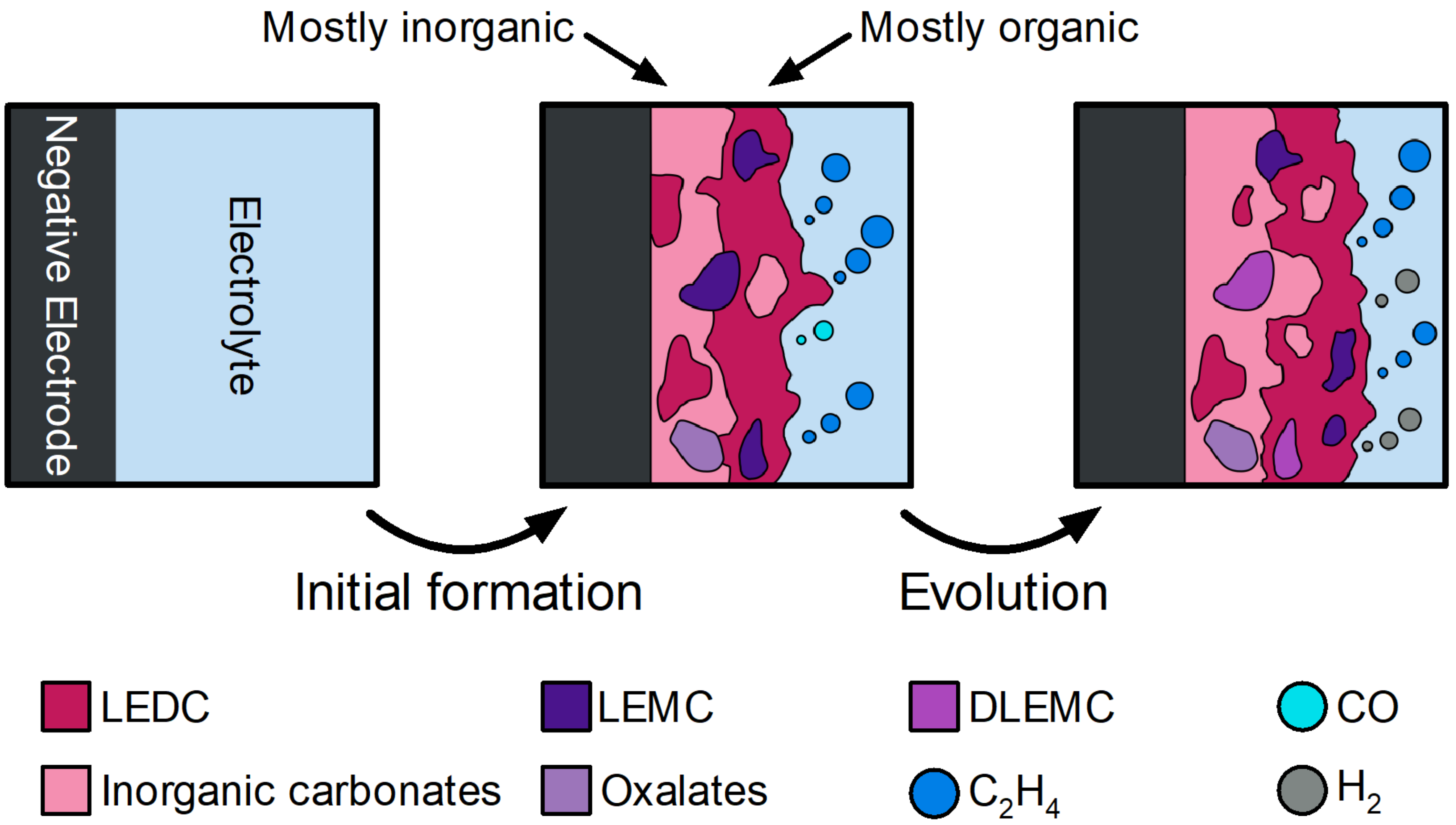


Elementary mechanism:



D. Barter*, E. W. C. Spotte-Smith*, N. Redkar, S. Dwaraknath, K. A. Persson, **S. M. Blau**, *Dig. Disc.* 2023

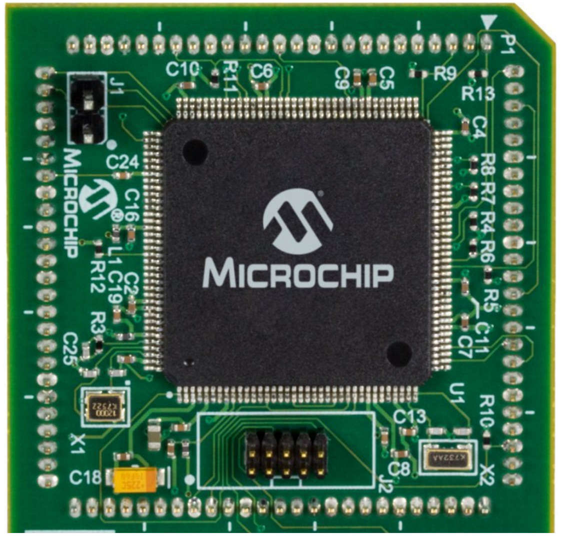
Mechanistic Model of SEI Formation Derived from CRN



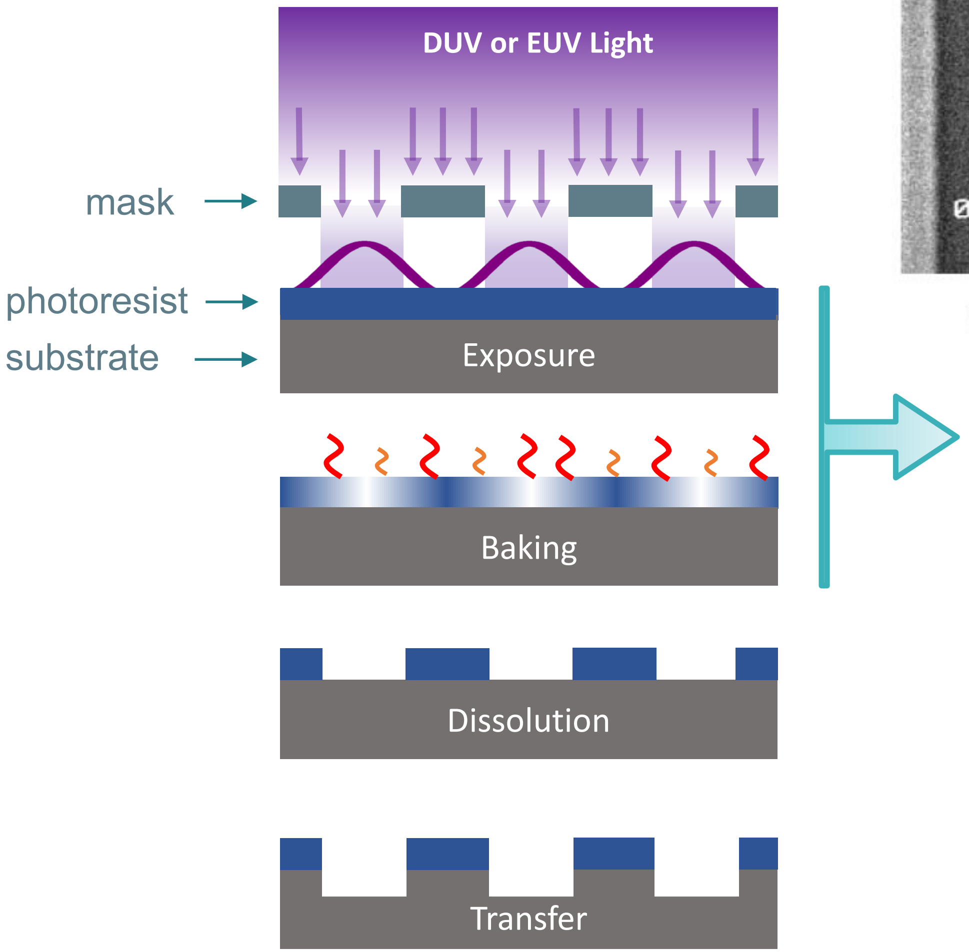
- Pathways derived from CRN, semi-automated $\Delta G^{\ddagger}$ calcs
- Recovered bi-layer SEI from first principles for first time
- **Is this approach limited to just SEI formation? No!**

E. Spotte-Smith*, R. Kam*, D. Barter, X. Xie, T. Hou, S. Dwaraknath, **S. M. Blau**, K. A. Persson, *ACS Energy Lett.* 2022

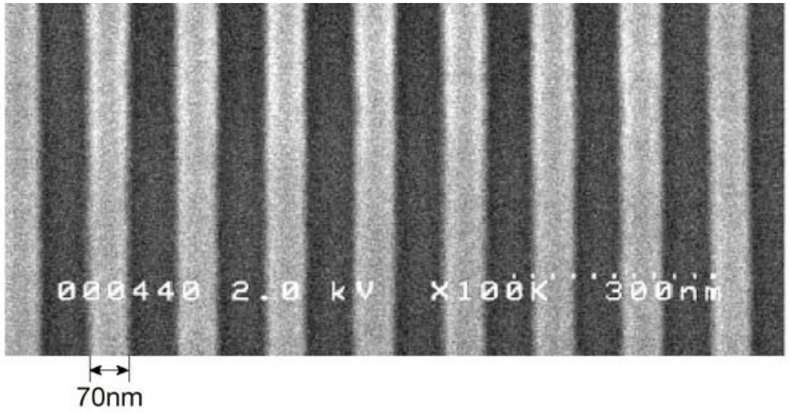
Background: Nanoscale Patterning with Photolithography



Background: Nanoscale Patterning with Photolithography

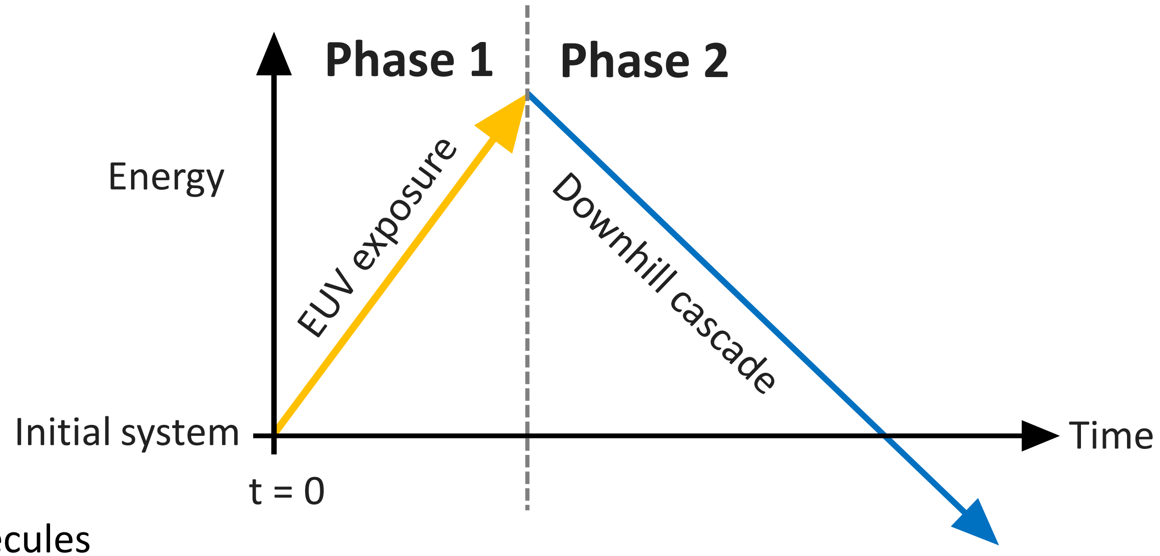


Chemical reactions cause solubility switch

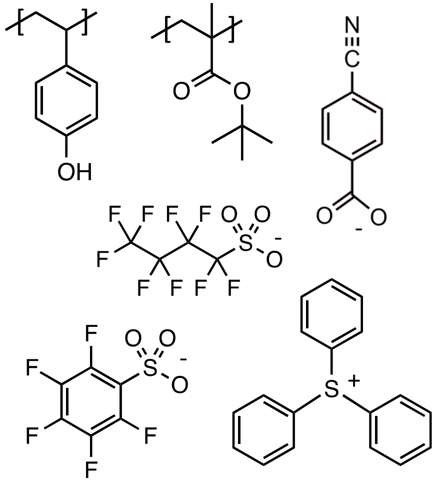


- 1994 to 2017: “deep” UV, 248 nm – 134 nm light
  - 5 eV – 9 eV photons
  - **Selective** resonant photochemistry
- Want smaller patterns? Need shorter wavelength!
- 2018 to now: “extreme” UV, 13.5 nm light
  - 92 eV photons
  - Stochastic photoionization yields **poorly understood** radical ion reaction cascade

EUV Lithography Reaction Network Construction



Principal molecules



Fragment → 108 species → Recombine → 3367 species → HiPRGen
DFT opt. DFT opt.

pymatgen FireWorks Custodian atomate

Phase 1: EUV exposure

Keep reactions that are:

- Electron detachment (all ΔG > 0)
- Electron attachment (all ΔG < 0)
- ΔG > 0 one-bond fragmentation
- ΔG > 0 H^+ or H^0 transfer

→ 185,929 reactions

Phase 2: Post-exposure cascade

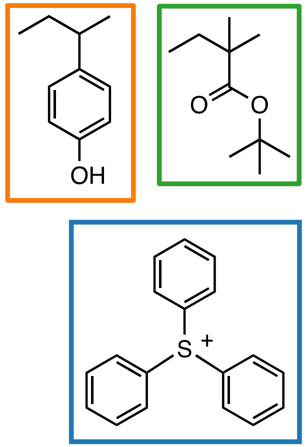
Remove reactions with:

- ΔG > 0
- Unbalanced redox
- >2 covalent bonds changing
- Sterically hindered reaction center

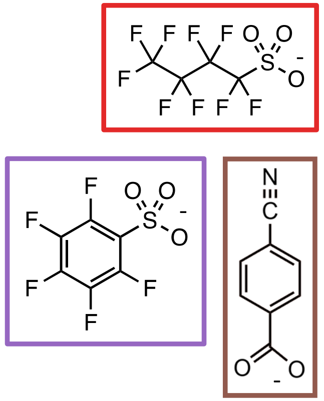
→ 2,776,867 reactions

J. Milton, F. Houle, **S. M. Blau**, *In preparation*

EUV Lithography Reaction Network Analysis

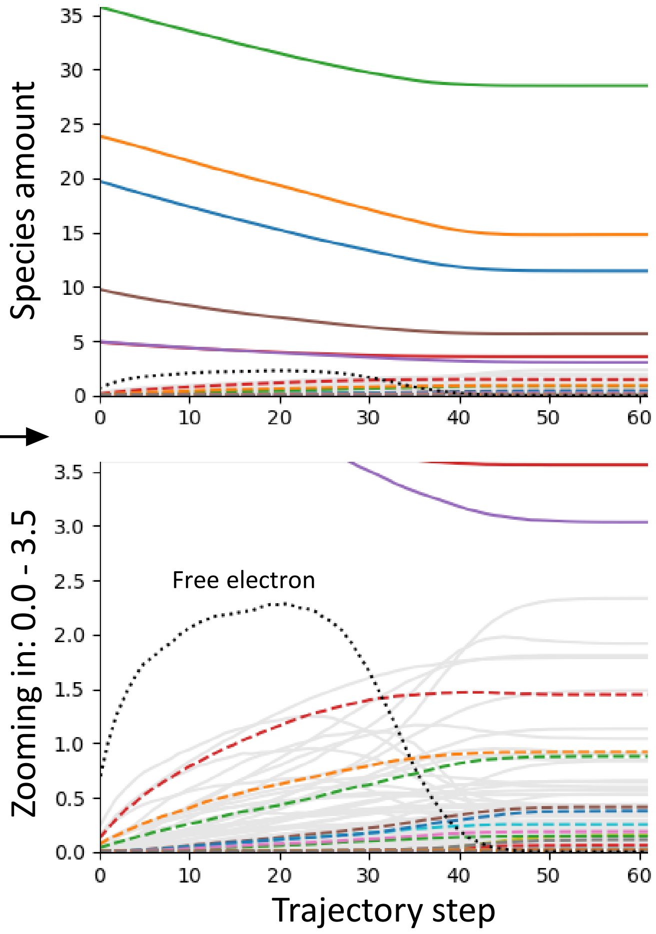


Initial state



RNMC

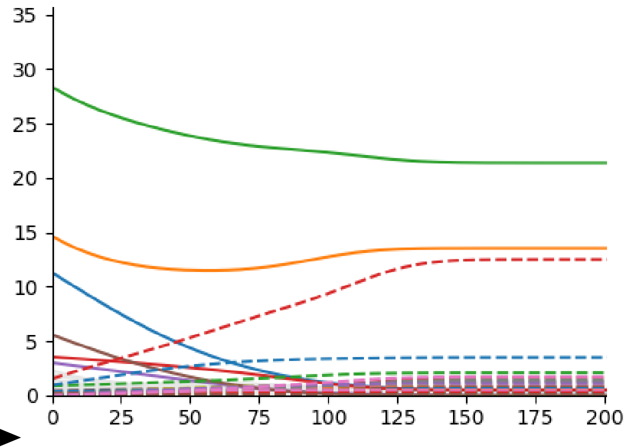
Phase 1: EUV exposure

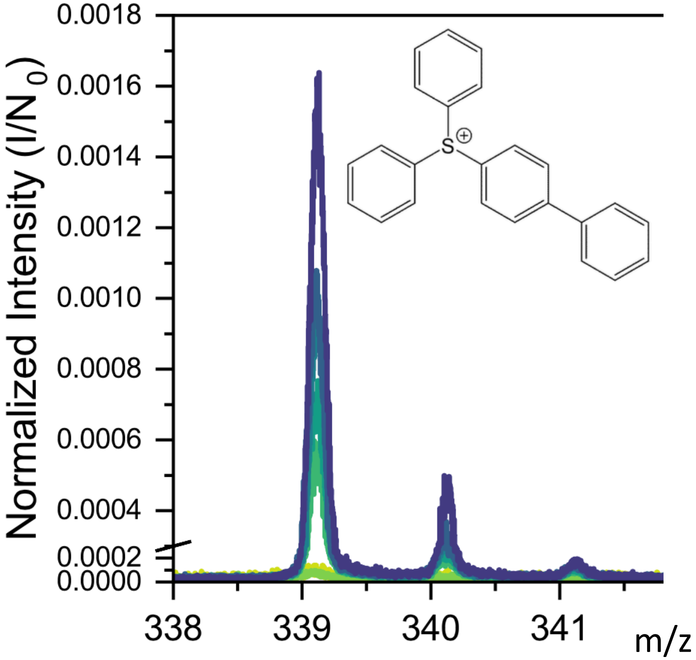


- +92 eV “energy budget”
- Explicit free electron species

System state

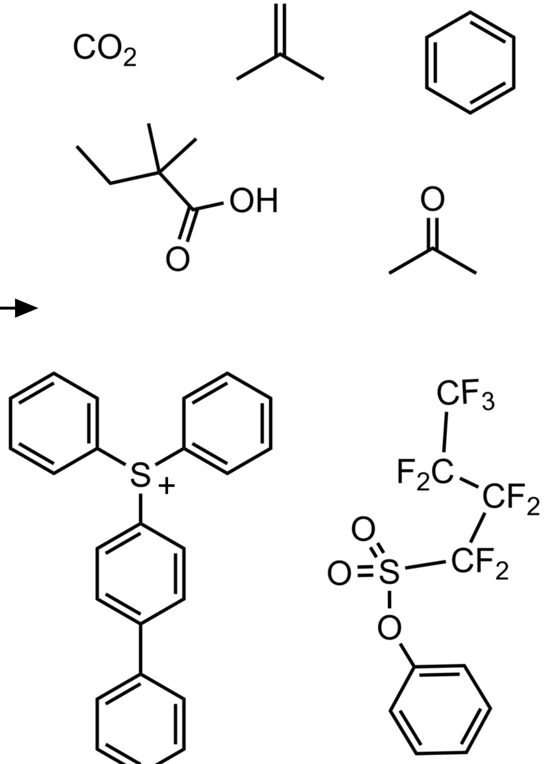
Phase 2: Post-exposure cascade





Cruz et al. *Proc. SPIE* 2022

Network products include:



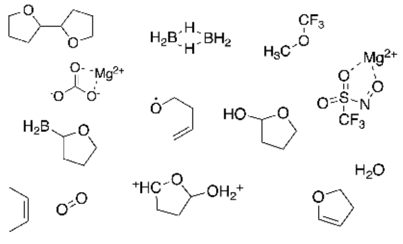
Recap: The Steps of Building and Analyzing a CRN

1. Species generation

Principal molecules → Molecular fragments → Recombinant molecules

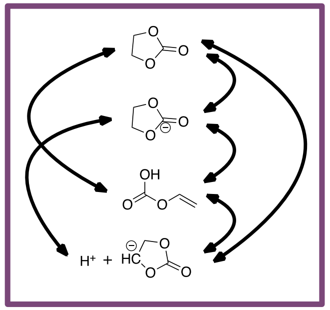
Species Enumeration

High-throughput DFT

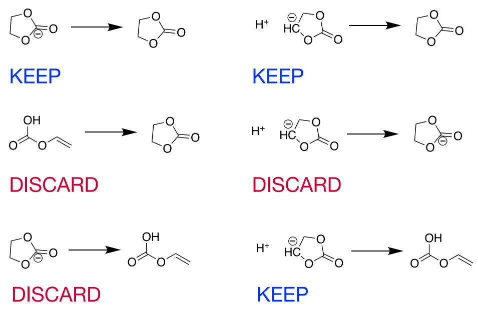


2. Reaction generation

Generate reactions by stoichiometry

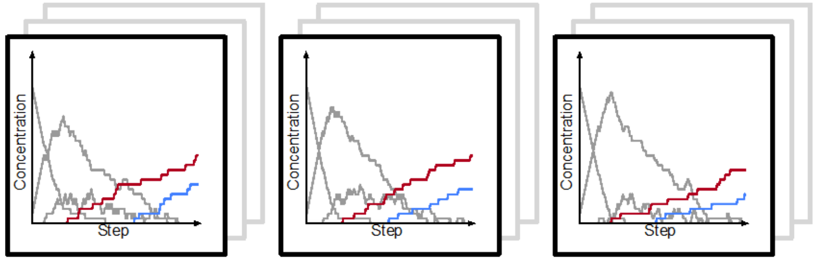


Filter reactions

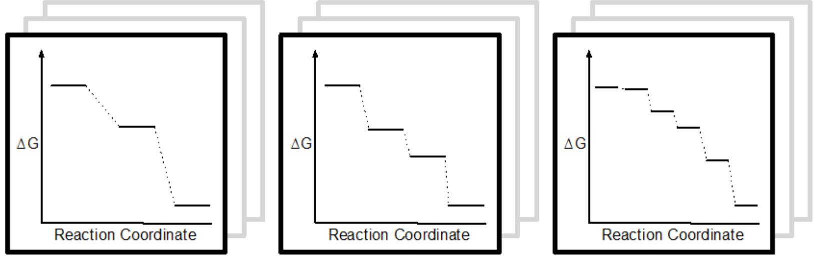


3. Pathway sampling

Perform many thermodynamically bounded Monte Carlo trajectories

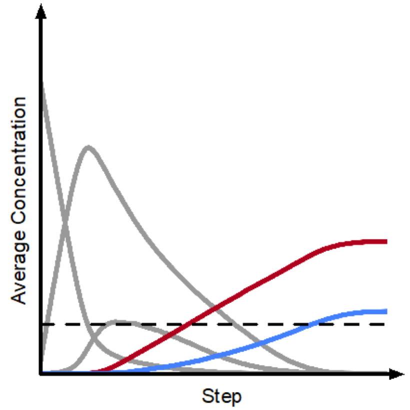


Extract shortest reaction pathways from each trajectory to each specie of interest



Transition state calcs
Build kinetic models

4. Identify Products



High formation / consumption

Significant accumulation

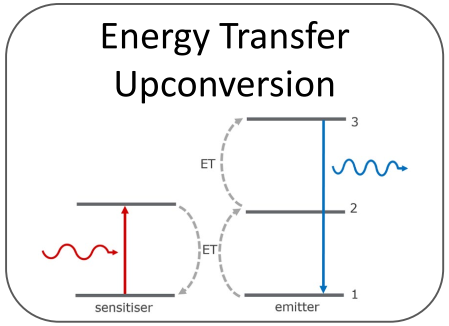
Low-cost pathways available ✖

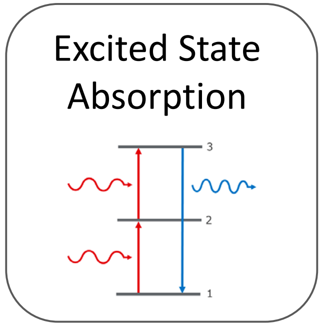
Network product? ✖

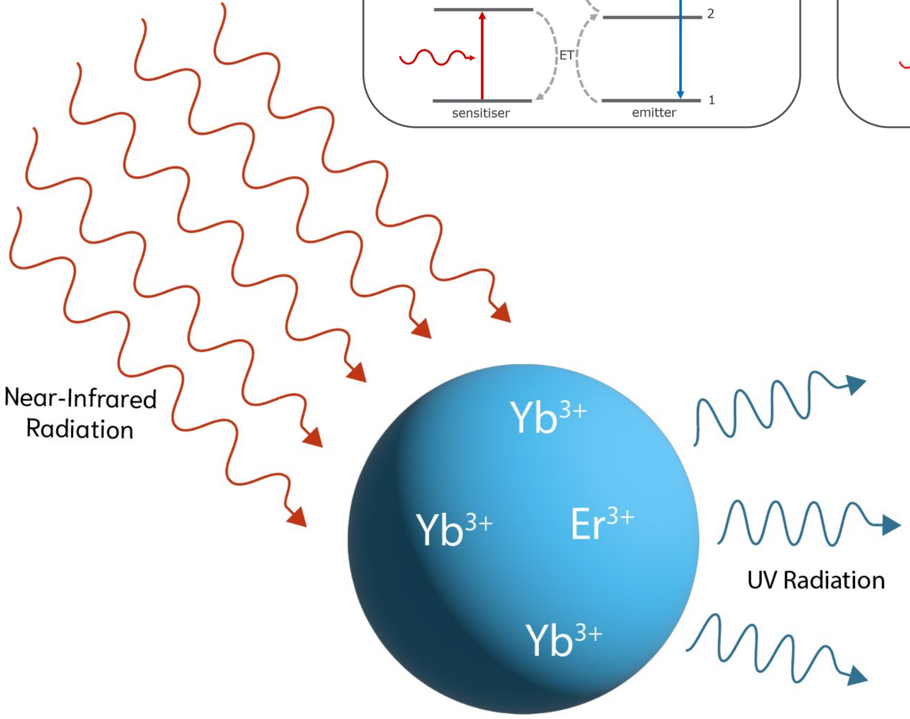
Discover novel important species and pathways

Under development: ML-assisted network expansion

Background: Upconverting Nanoparticles (UCNPs)

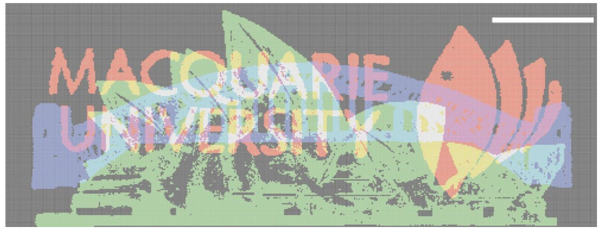






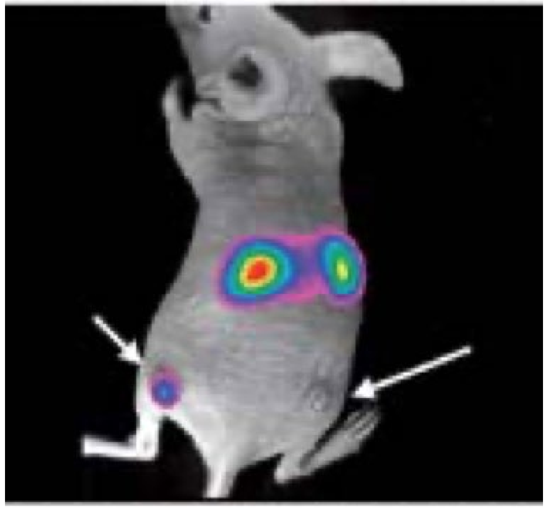
$NaYF_4$:Yb^{3+},Er^{3+}

Security Printing



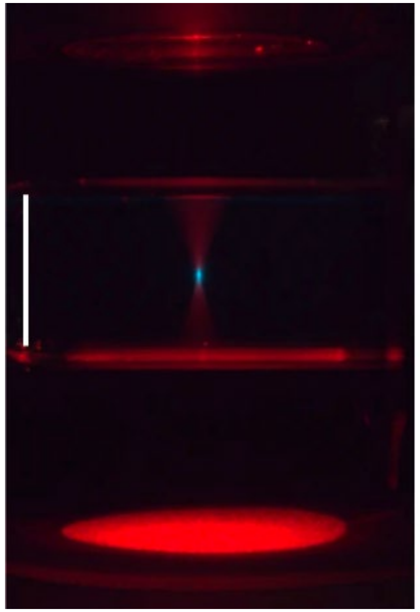
Lu et al. *Nat. Photon.* 2014

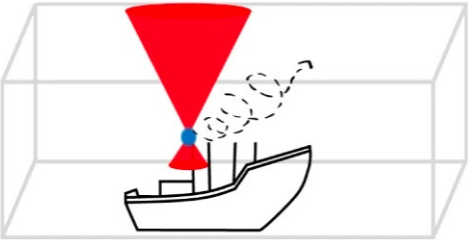
Bio-imaging



Xiong et al, *Anal. Chem.* 2009

3D Printing



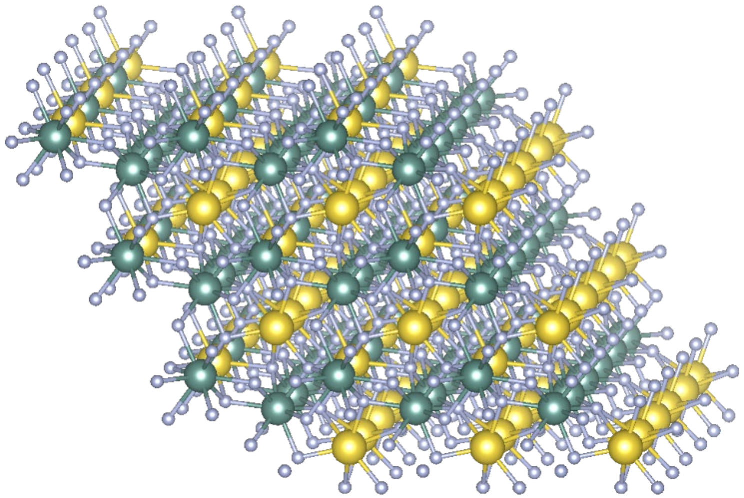


Sanders et al, *Nature* 2022

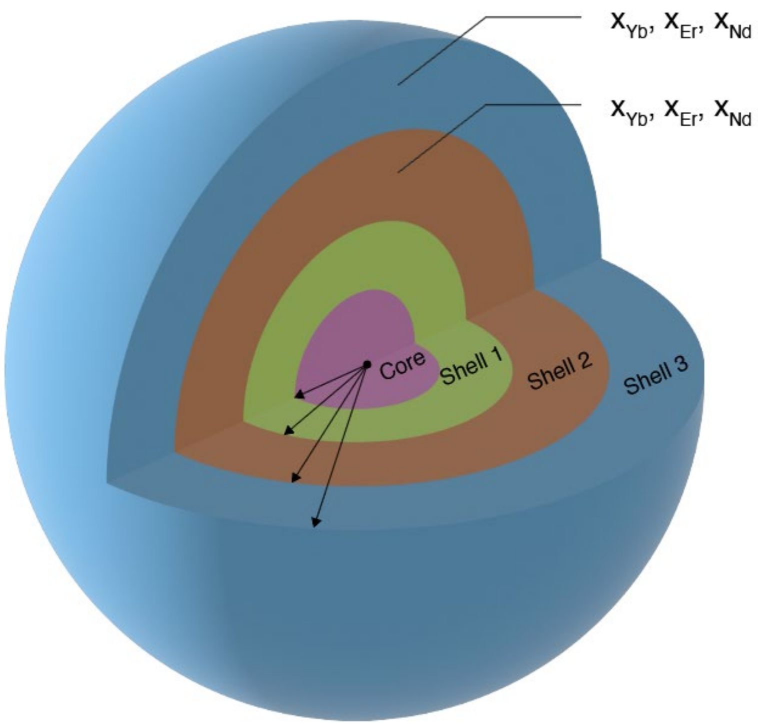
BERKELEY LAB

UCNP Doping and Heterostructure

Host Material: $NaYF_4$



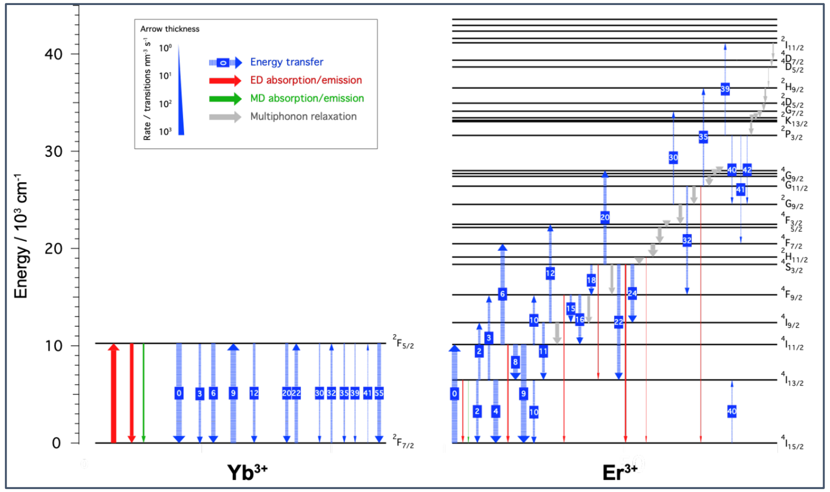
Dope Y^{3+} sites with Ln^{3+}



58	59	60	61	62	63	64	65	66	67	68	69	70
Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb

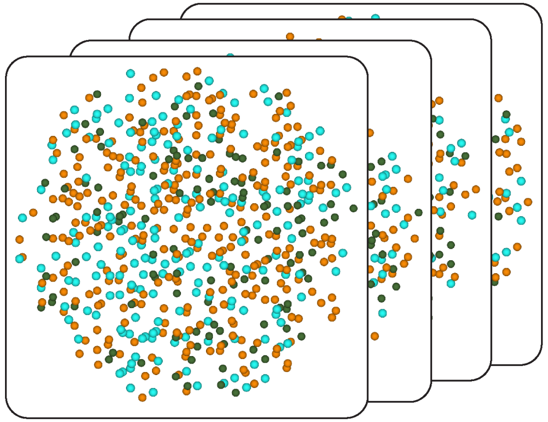
UCNP Photophysics Can Be Simulated With kMC

Transition rate constants



Teitelboim et al. *J. Phys. Chem. C* 2019

Generate ensemble of randomly doped structures

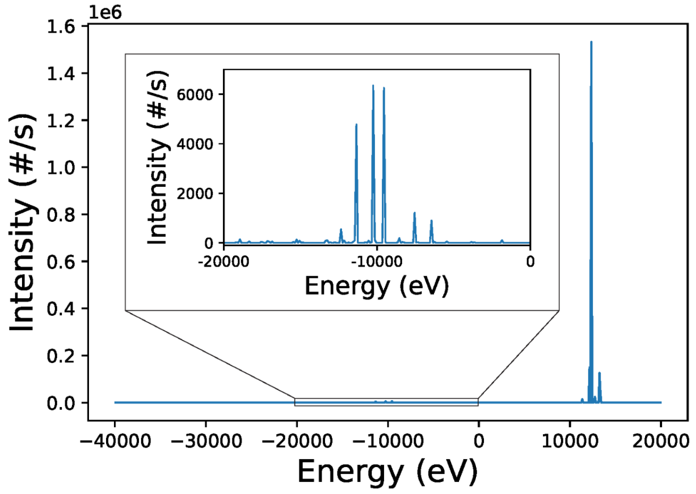


Energy Transfer
Kinetic Monte Carlo (kMC)

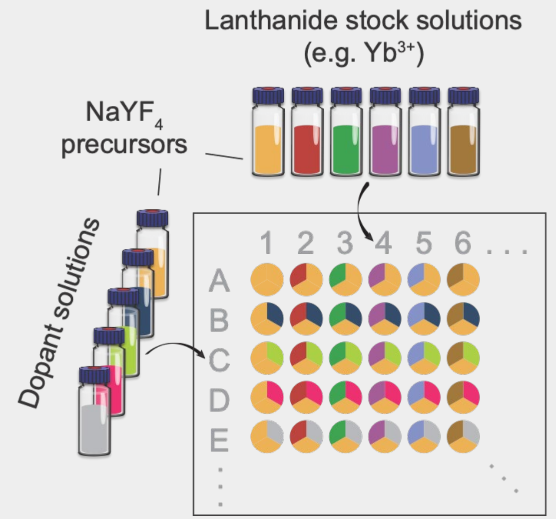
github.com/BlauGroup/NanoParticleTools
github.com/BlauGroup/NPMC

- Simulations require 10-150+ hours on one CPU core
  - Cannot be parallelized

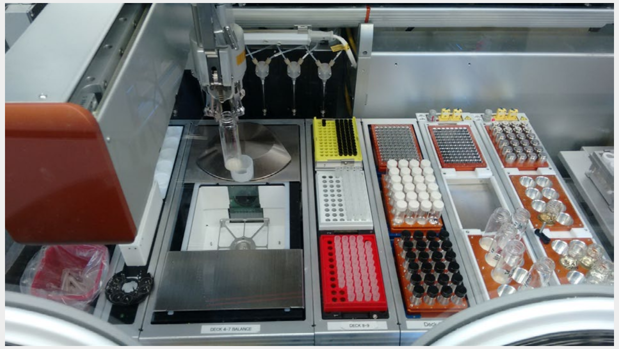
Spectra



Large Search Space Necessitates Intelligent Searching



Combinatorial/Robotic Synthesis



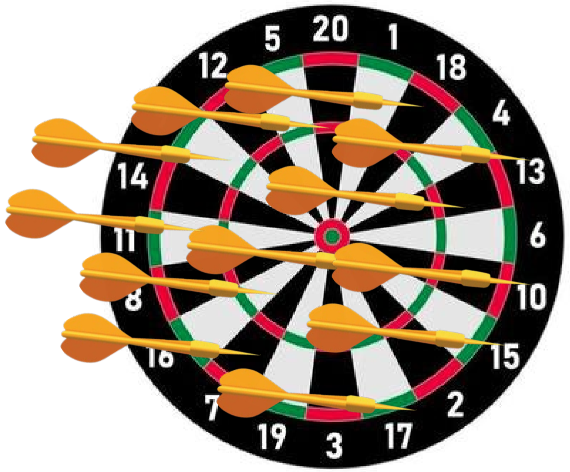
Consider a simple spherical nanoparticle:

- Chose up to 4 dopants (of 13 lanthanides)

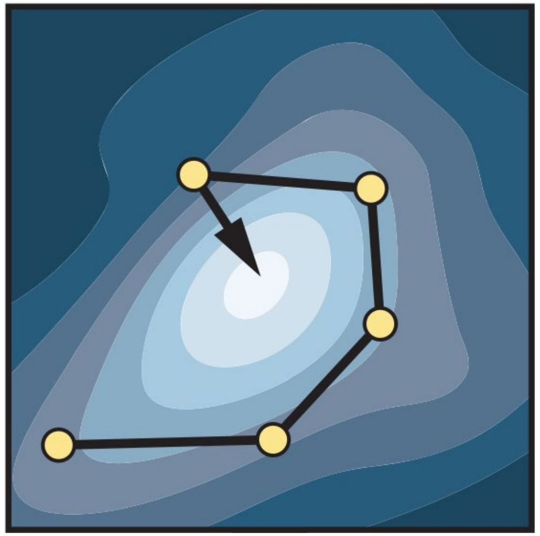
  1,093 combinations
- 3 Dopant concentrations - Low, Medium, High

  66,379 dopant configurations
- 5 particle sizes - 4, 6, 8, 10, & 12nm

265,516 nanoparticle configurations

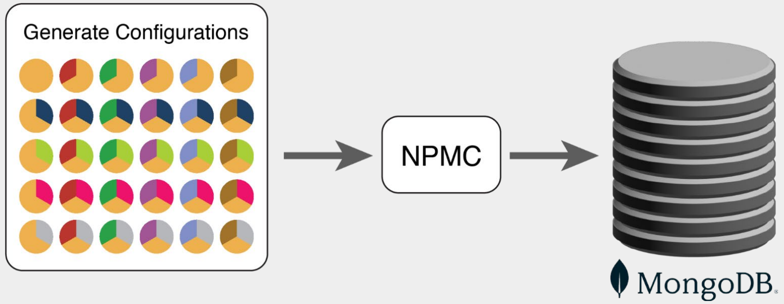


Inverse Design



Sanchez-Lengling et al. *Science* 2018

High-throughput search



Generating a Dataset for Machine Learning

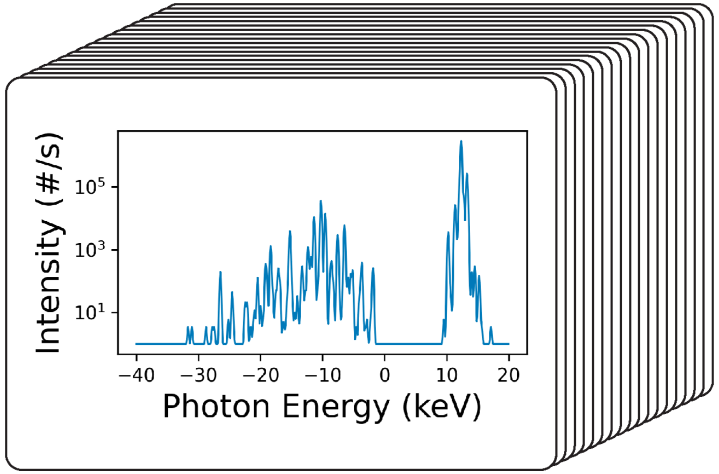
<u>IID Dataset</u>:

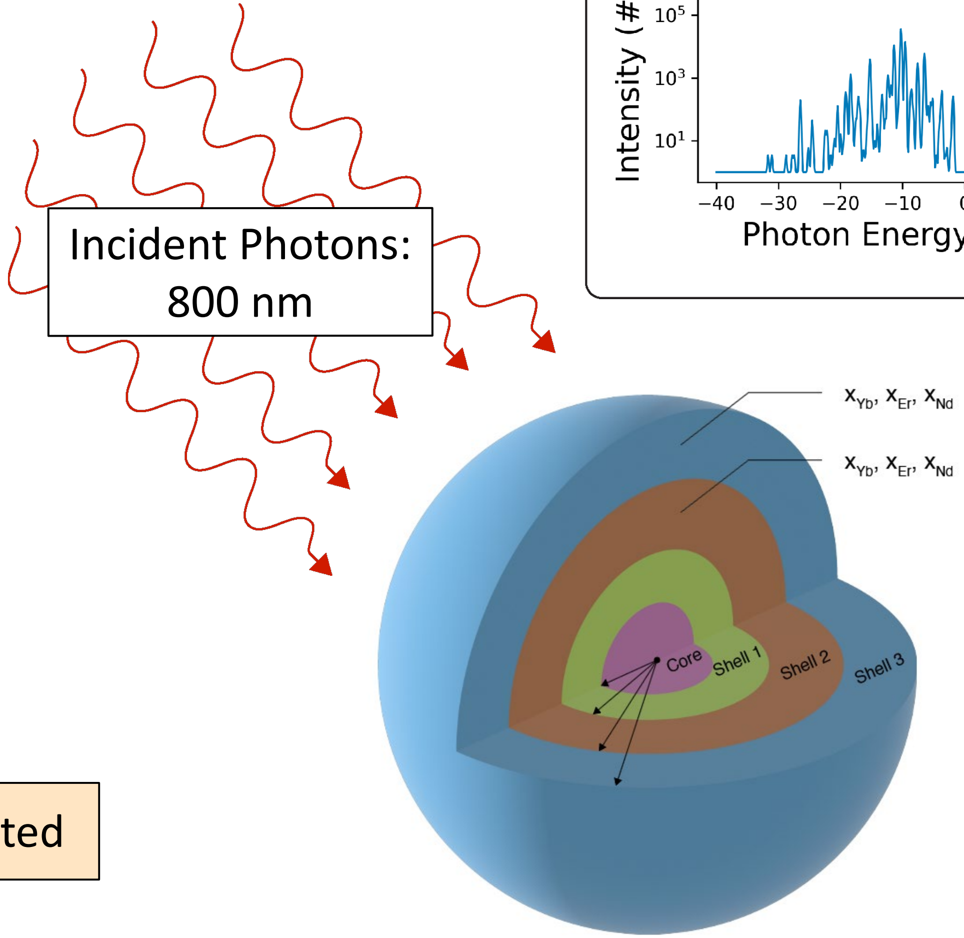
- Up to 8 nm diameter core
- Up to 3 shells
  - Each shell is 1-2.5 nm thick
- Consider only Yb, Er, and Nd dopants

<u>OOD Testing Dataset:</u>

- Up to 8 nm diameter core
- 4 shells
  - Each shell is 1-2.5 nm thick
- Consider only Yb, Er, and Nd dopants

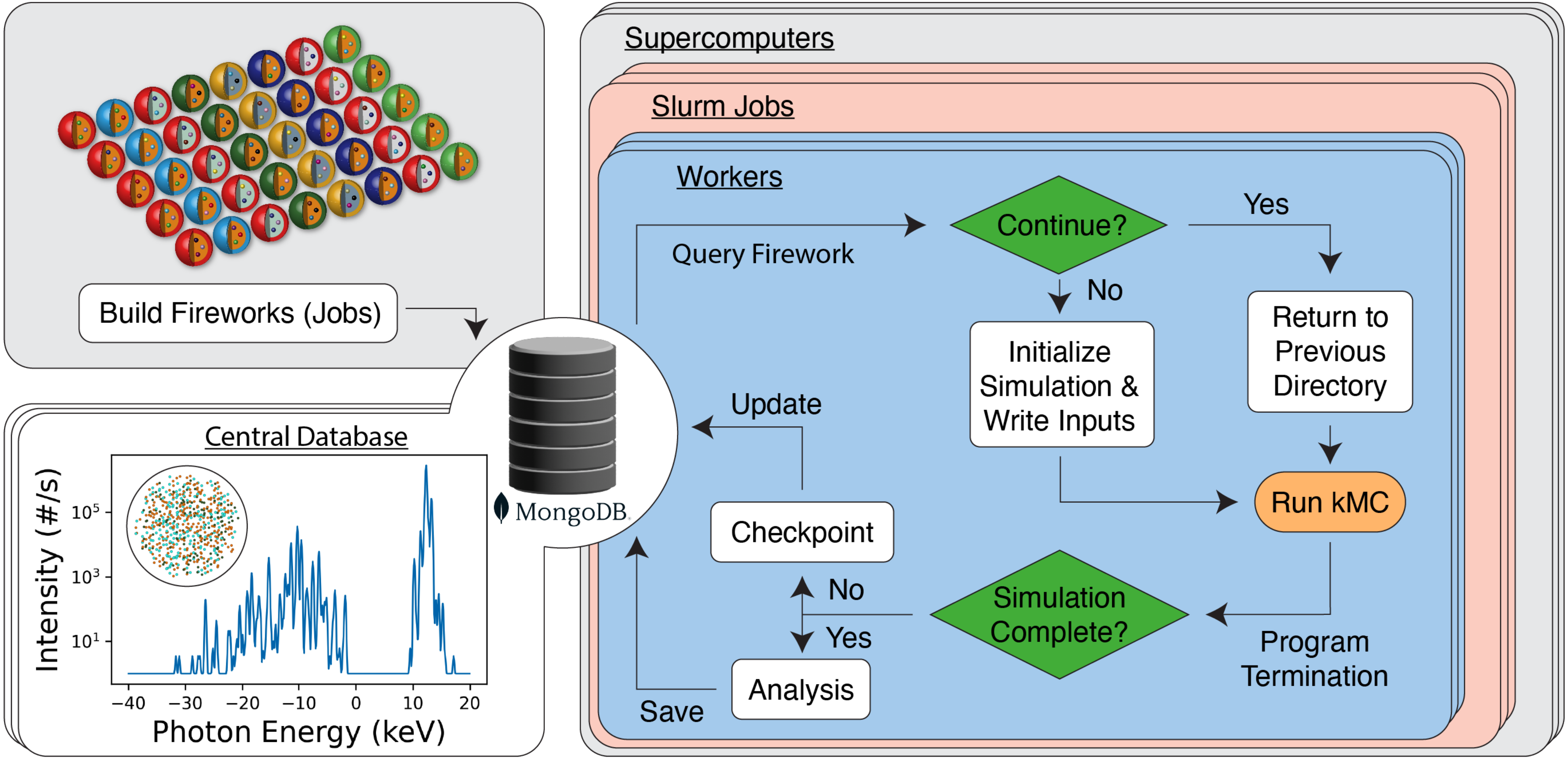
>6,000 nanoparticle configurations/spectra simulated





E. Sivonxay, E. Chan, **S. M. Blau**, *In preparation*

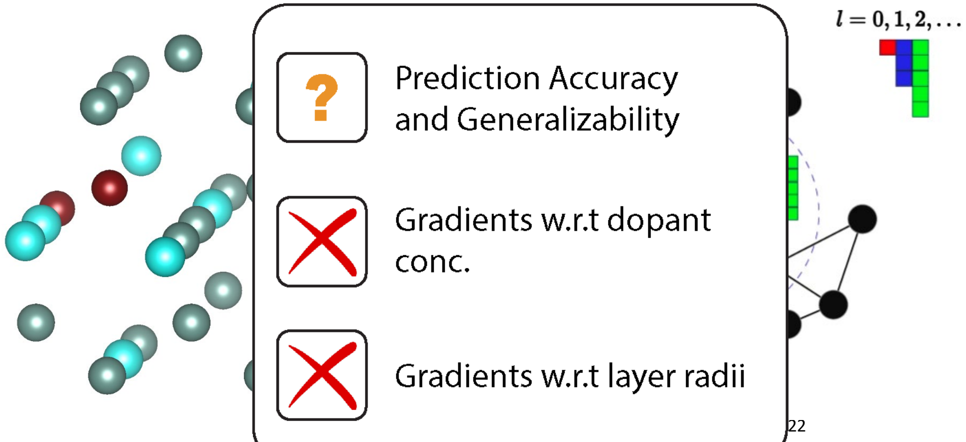
UCNP kinetic Monte Carlo Simulation Workflow



E. Sivonxay, E. Chan, **S. M. Blau**, *In preparation*

Representations of Nanoparticles for Machine Learning

Atomistic Representation



- ? Prediction Accuracy and Generalizability
- ✗ Gradients w.r.t dopant conc.
- ✗ Gradients w.r.t layer radii

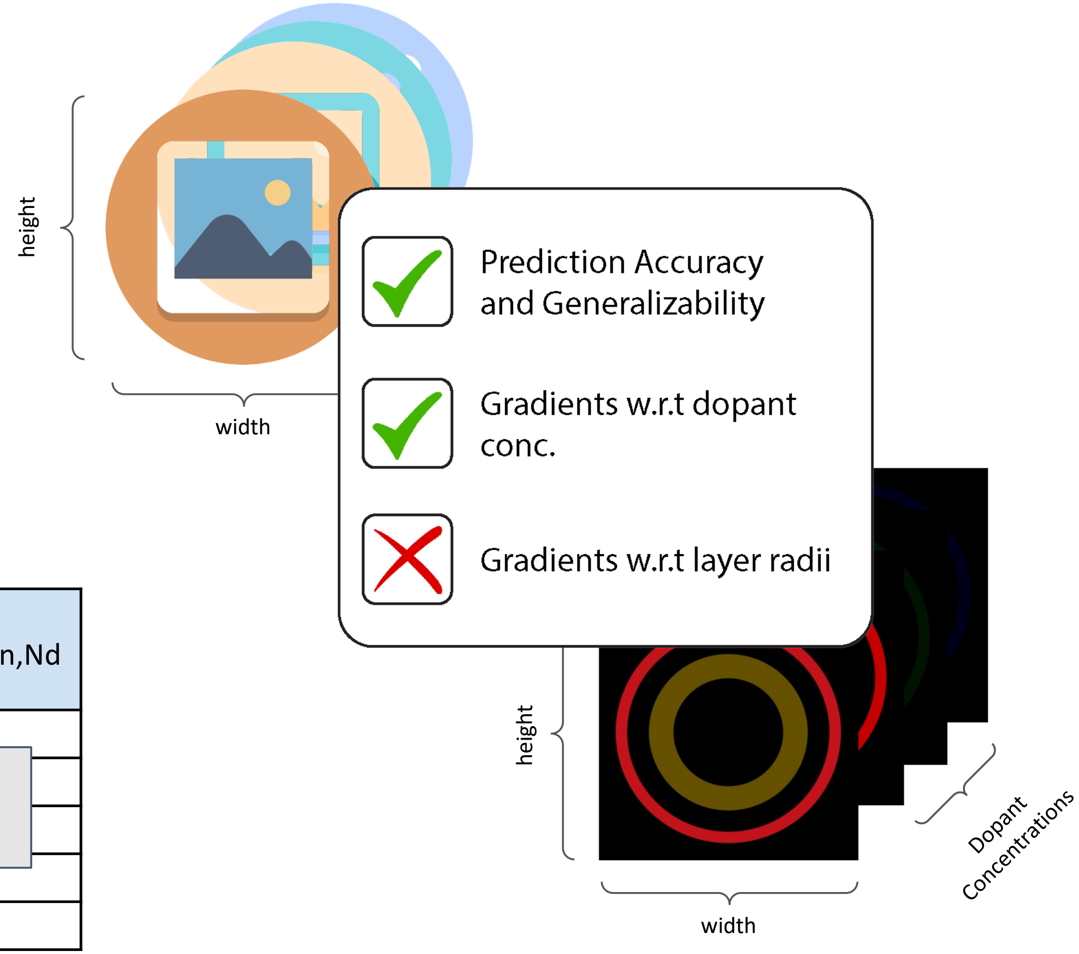
Tabular Representation

r_0	V_0	$x_{0,Yb}$		$x_{n,Yb}$	$x_{n,Er}$	$x_{n,Nd}$

Core feature | Layer features

- ✗ Prediction Accuracy and Generalizability
- ✓ Gradients w.r.t dopant conc.
- ✓ Gradients w.r.t layer radii

Image Representation

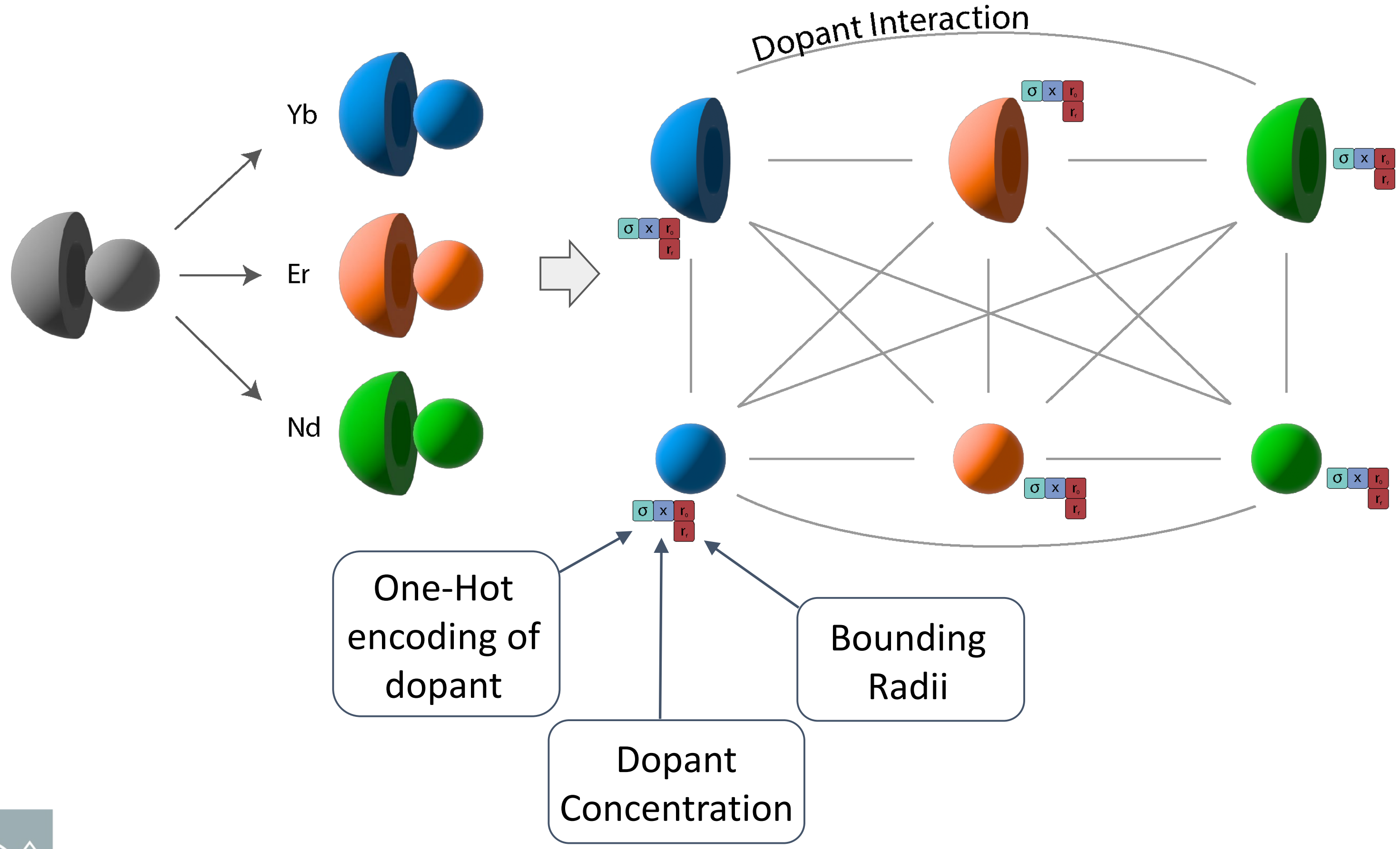


- ✓ Prediction Accuracy and Generalizability
- ✓ Gradients w.r.t dopant conc.
- ✗ Gradients w.r.t layer radii

E. Sivonxay, E. Chan, **S. M. Blau**, *In preparation*

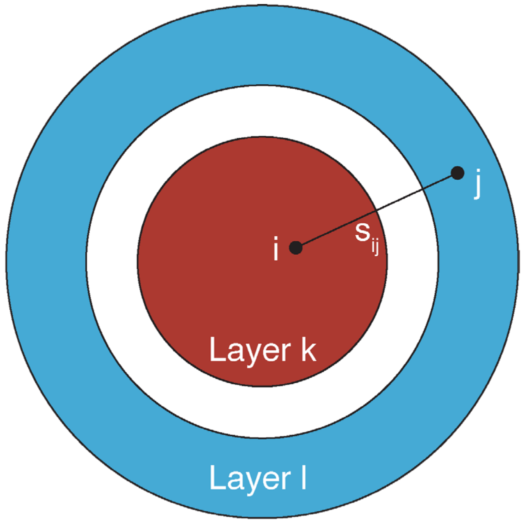
Developing a Physics-Infused Graph Representation

Dopant/Control Volume Graph



Edge feature - Integrated Interaction

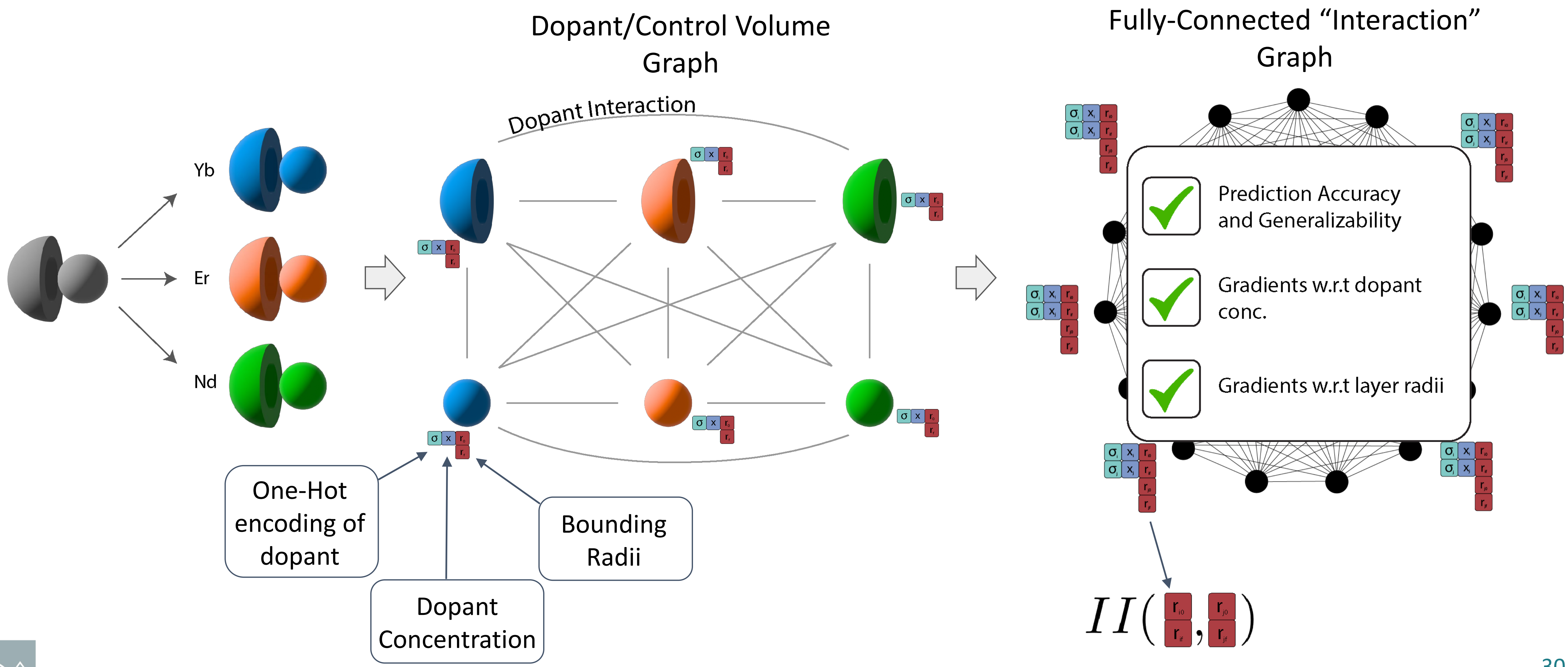
$$I_{ij} = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2 \frac{s_{ij}}{\sigma}^2}$$



$$II\left(\begin{matrix} r_{i0} \\ r_{if} \end{matrix}, \begin{matrix} r_{j0} \\ r_{jf} \end{matrix}\right) = \int\int I_{ij} dV_k dV_l$$

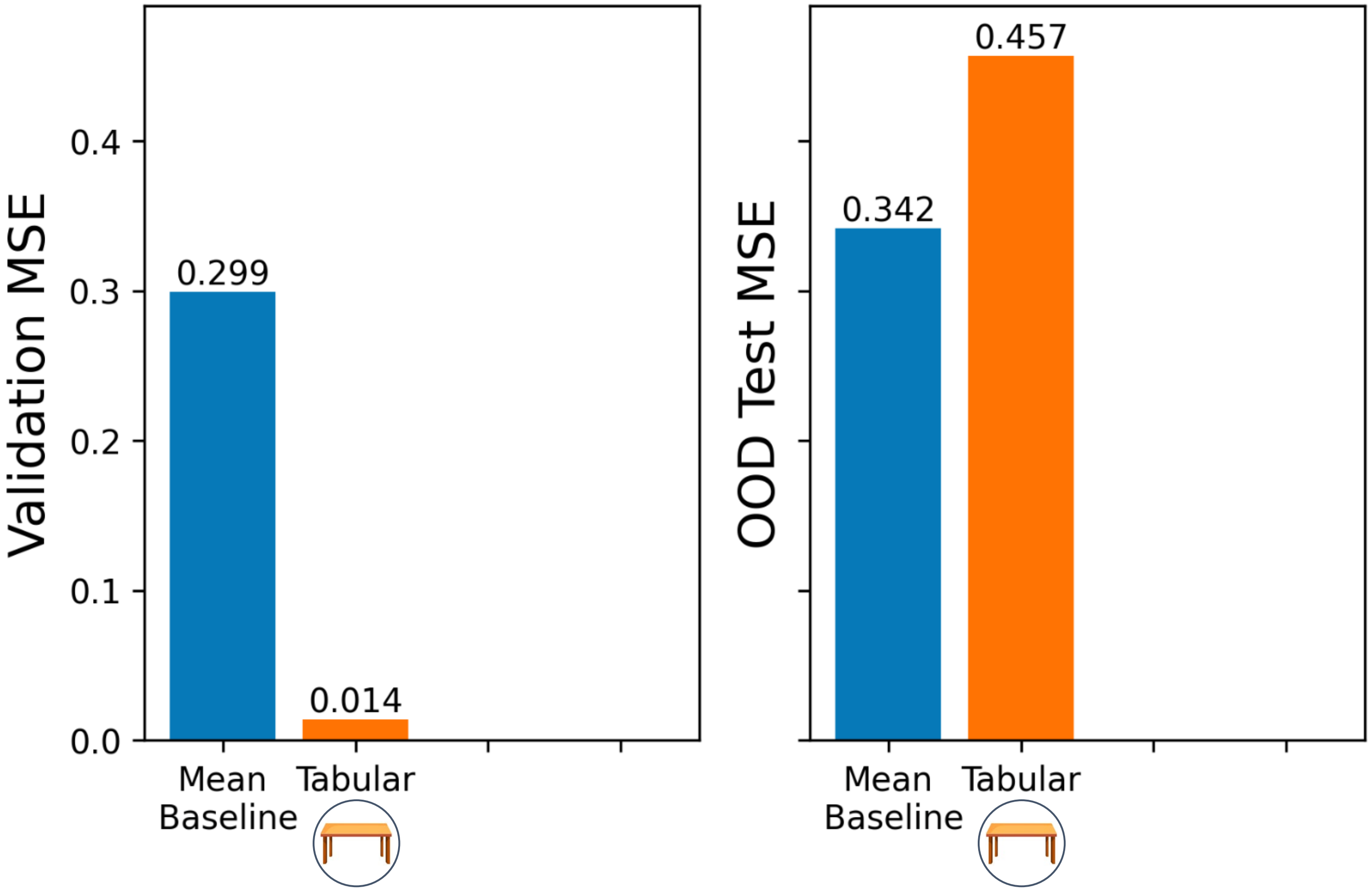
E. Sivonxay, E. Chan, **S. M. Blau**, *In preparation*

Developing a Physics-Infused Graph Representation



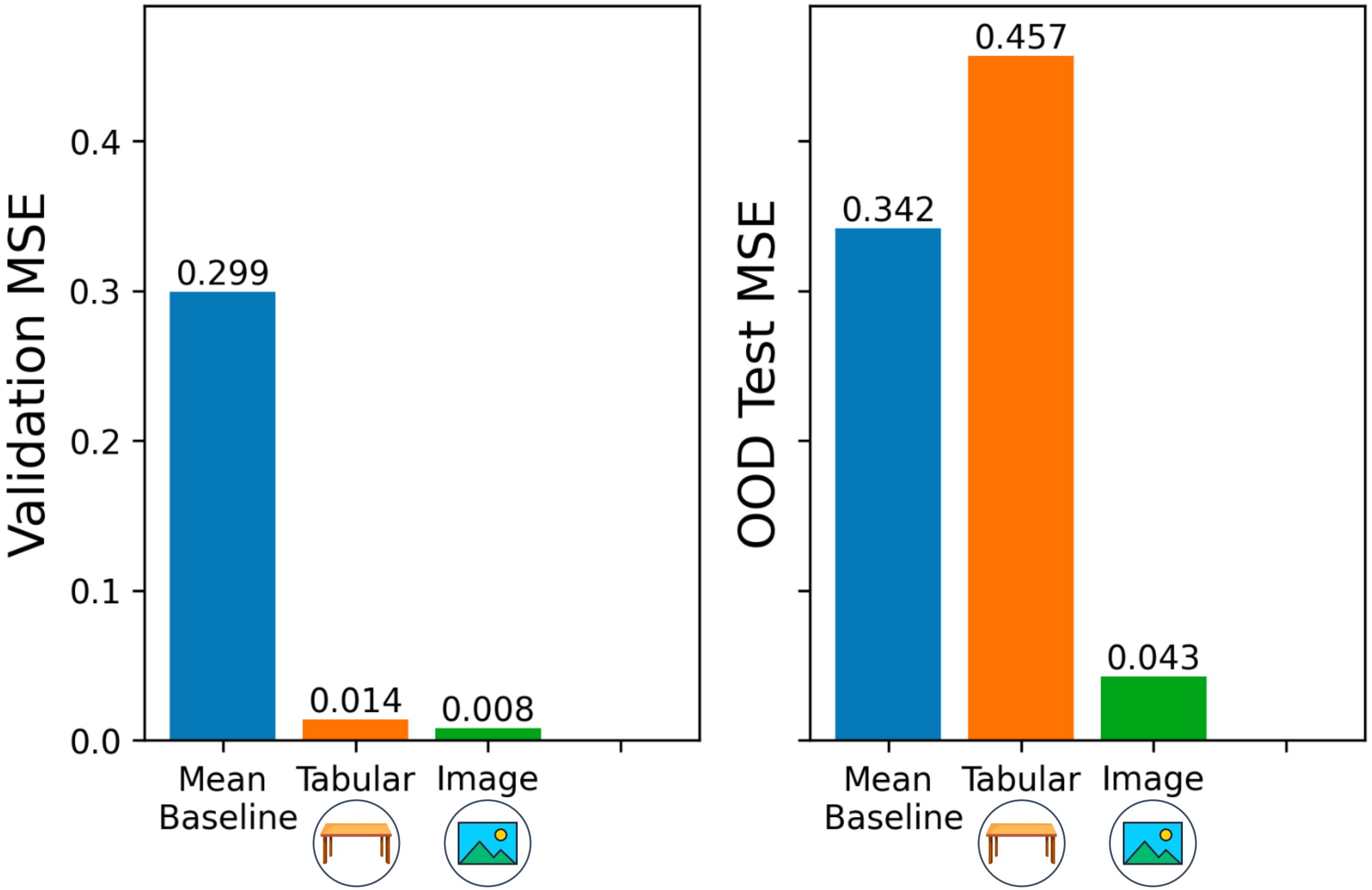
E. Sivonxay, E. Chan, **S. M. Blau**, *In preparation*

Comparing Tabular vs. Image vs. Graph Rep. Performance



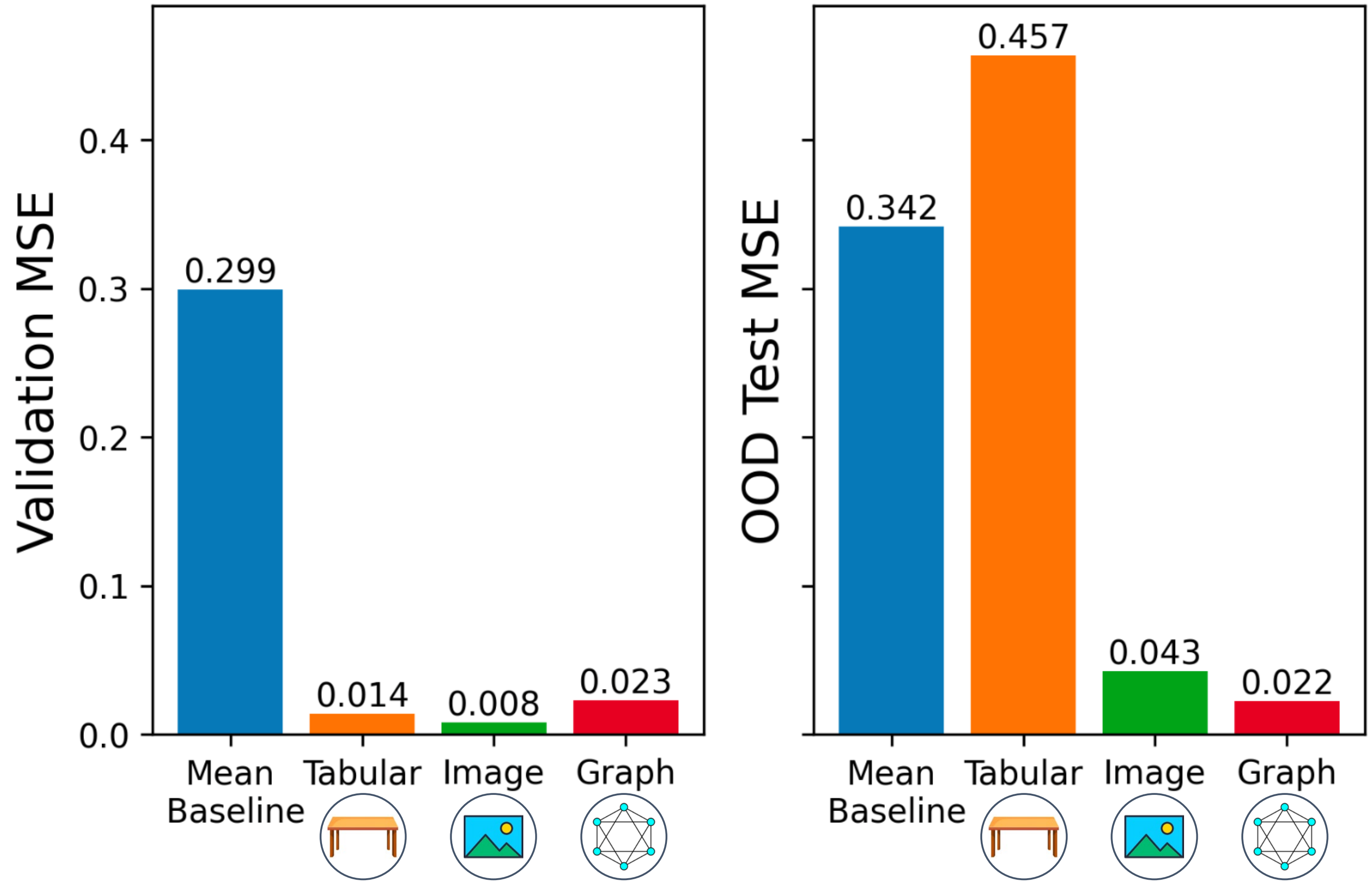
E. Sivonxay, E. Chan, **S. M. Blau**, *In preparation*

Comparing Tabular vs. Image vs. Graph Rep. Performance



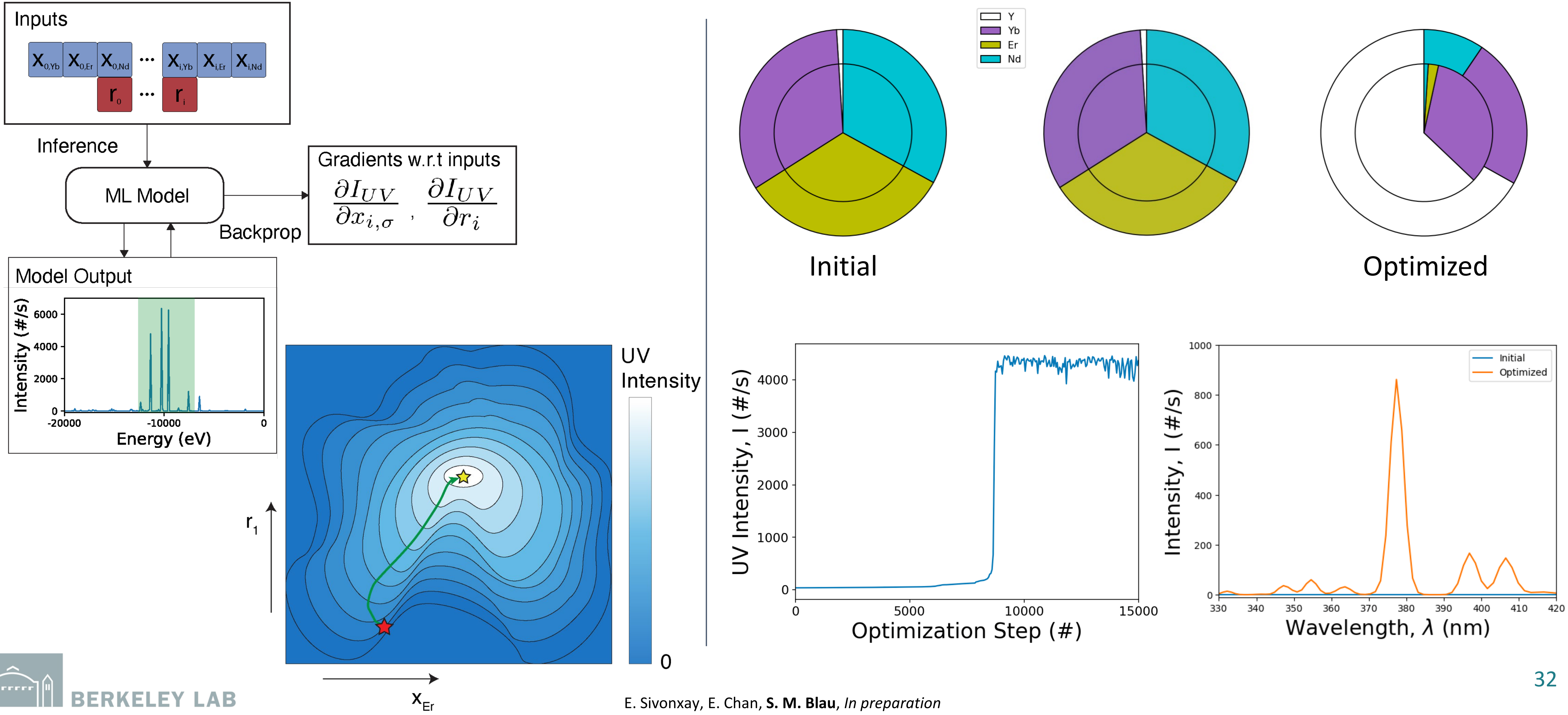
E. Sivonxay, E. Chan, **S. M. Blau**, *In preparation*

Comparing Tabular vs. Image vs. Graph Rep. Performance



E. Sivonxay, E. Chan, **S. M. Blau**, *In preparation*

Inverse Design of Nanoparticles Via Gradient Ascent



E. Sivonxay, E. Chan, **S. M. Blau**, *In preparation*

BERKELEY LAB

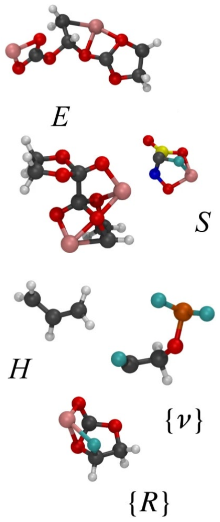
Summary

pymatgen

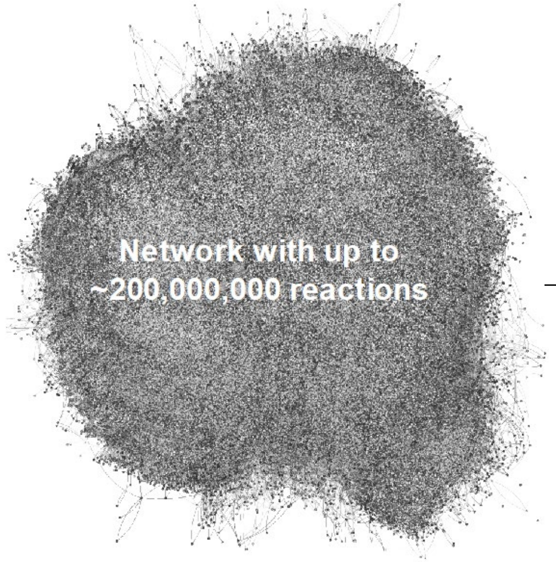
Custodian

FireWorks

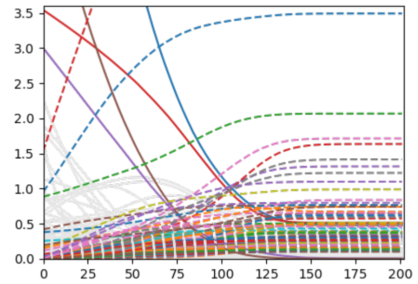
atomate

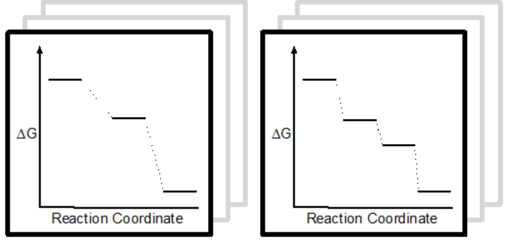


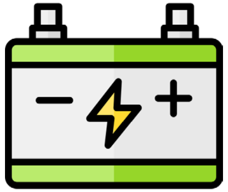
HiPRGen

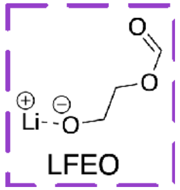


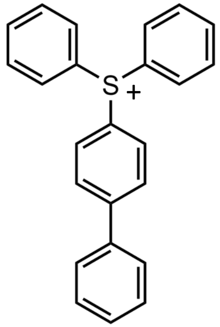
RNMC

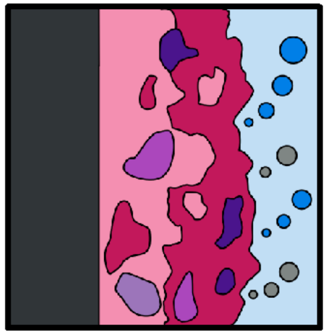


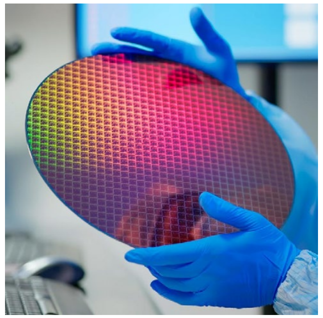


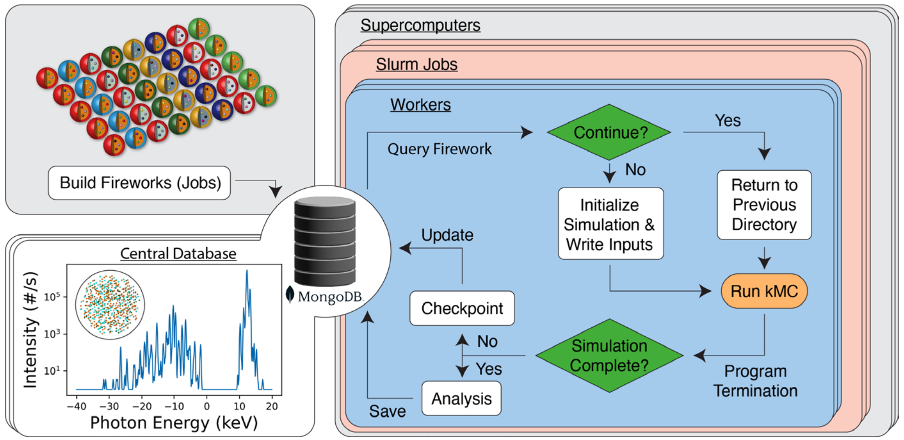


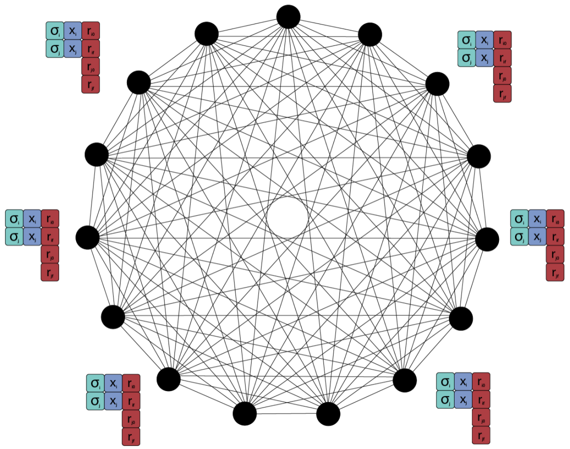


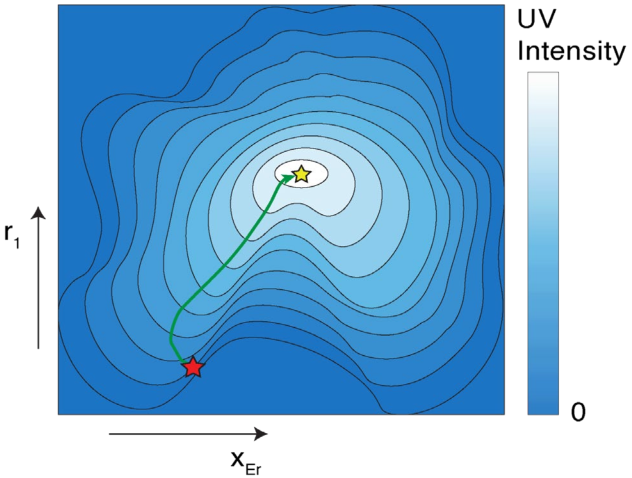


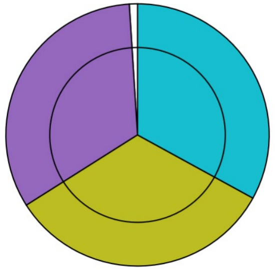












Acknowledgements


Daniel Barter


Evan Spotte-Smith


Eric Sivonxay


Jacob Milton


Frances Houle


Kristin Persson


Emory Chan

BERKELEY LAB
Lawrence Berkeley National Laboratory

JCESR

U.S. DEPARTMENT OF ENERGY
Office of Science

CHiPPS
Center for High Precision Patterning Science

DARPA

MOLECULAR FOUNDRY

Schrödinger

NERSC