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

Typically, <75% success

With on-the-fly error correction: >98% success

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

**OptimizeFW:**
- OptSet
- QCInput
- WriteInputFromIOSet
- RunQChemCustodian
- QCJob
- QCInput
- QCOutput
- QChemToDb

Optimization Error
- Max errors reached?
  - Yes QUIT
  - No
    - RFO step failed?
      - Yes QUIT
      - No
        - Back transform failed?
          - Yes QUIT
          - At least one step taken?
            - Yes QUIT
            - No
              - Switch coordinate system*
                - Yes
                  - Always make SCF guess
                - No QUIT
              - At least one step taken?
                - Yes QUIT
                - No
                  - Restart from last geometry
                    - Yes QUIT
                    - No
                      - Orbitals scratch file error?
                        - Yes QUIT
                        - No QUIT

* Coordinate systems:
  - Redundant → Delocalized
  - Delocalized → Cartesian

We Use Workflows to Generate Unique Simulated Datasets

LIBE
Lithium-Ion Battery Electrolyte
17,190 molecules

MADEIRA
MAgnesium Dataset of Electrolyte and Interphase ReAgents:
11,502 molecules
E. W. C. Spotte-Smith, S. M. Blau, et al., JACS (accepted)

RAPTER
ReActants, Products, and Transition-states of Elementary Reactions:
>15,000 complex reactions
E. W. C. Spotte-Smith, S. M. Blau, et al., In preparation

ωB97X-V/def2-TZVPPD/SMD
ωB97X-V/def2-TZVPPD/SMD
ωB97X-D/def2-SVPD/PCM

Collaborators:
Evan Spotte-Smith  Kristin Persson
We Use Workflows to Generate Unique Simulated Datasets

ESCoMMS
Electronic Structure of Complexes with Metals of Many Spins:
>140,000 complexes

ORIONS
ORbital Interactions of OrgaNic Species:
>230,000 molecules

SUNSET
Simulated Upconverting Nanoparticle Spectra for Emissions Tuning:
>6,000 spectra (kMC, not DFT)

Michael Taylor
Ping Yang
Gabe Gomes
Eric Sivonxay
Emory Chan

D. Boiko et al., ChemRxiv 2022
Machine Learning Atop Our DFT Datasets

An orbital-based representation for accurate quantum machine learning

“...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...”
Introduction to Chemical Reaction Networks (CRNs)

△ = Unstable intermediate

★ = Stable product

Initial species

Chemical reaction network (CRN)
Introduction to Chemical Reaction Networks (CRNs)

= Unstable intermediate  = Stable product

Initial species

Chemical reaction network (CRN)
Introduction to Chemical Reaction Networks (CRNs)

△ = Unstable intermediate  ★ = Stable product

Chemical reaction network (CRN)

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
- $\Delta G$ 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

Principal Molecules

Solvent molecules

SEI products

Salts

Lithium-Ion Battery Electrolyte
17,190 molecules

Atomic Molecules

Molecular Fragments

Selective Recombinant Molecules

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

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

Our solution: filters

Cyclic biradical scission

CO insertion

NH₃ elimination

Prescriptive templates are not well-suited to electron-driven chemistry
High-Performance Reaction Generation:

\[ S_{\text{init}} \]

**Input:** initial species

\[ \text{LIBE-CHOLi} = 8904 \text{ species} \]

\[ S_{\text{filtered}} \quad R_{\text{filtered}} \]

**Output:** species, reactions that compose network
High-Performance Reaction Generation: HiPRGen

**Input:** initial species $S_{init}$

LIBE-CHOLi = 8904 species

1. Filter species

- $H^+$ KEEP
- $OH$ KEEP
- $OOO$ KEEP
- $OOO$ DISCARD
- $OOO$ KEEP
- $OOO$ KEEP

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

After filtering = 5193 species
High-Performance Reaction Generation: HiPRGen

**S\text{\textsubscript{init}}**

**Input:** initial species

LIBE-CHOLi = 8904 species

1. Filter species

- \( H^+ \) KEEP  
- \( OH \) KEEP  
- \( SO_3 \) KEEP  
- \( Li^0 \) KEEP

- DISCARD

- ...  

- Metal-centric complexes
- \( Li^0 \)-containing species

2. Bucket Species by Composition

**After filtering = 5193 species**

High-Performance Reaction Generation:

$S_{init}$

**Input:** initial species

LIBE-CHOLi = 8904 species

1. Filter species

2. Bucket Species by Composition

3. Generate reactions by stoichiometry

After filtering = 5193 species

$\geq 176$ billion rxns

HiPRGen

High-Performance Reaction Generation:

$S_{\text{init}}$

Input: initial species

LIBE-CHOLi = 8904 species

1. Filter species

H$^+$ KEEP

H$^+$ KEEP

DISCARD

KEEP

2. Bucket Species by Composition

C3 H4 O3

H1

C3 H3 O3

H$^+$ + HC$_2$O$_2$

3. Generate reactions by stoichiometry

After filtering = 5193 species

4. Filter reactions

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

High-Performance Reaction Generation:

\[ S_{\text{init}} \]

**Input:** initial species

\[ \text{LIBE-CHOLi} = 8904 \text{ species} \]

1. Filter species

\[ \begin{align*}
\text{KEEP} & & \text{KEEP} & & \text{DISCARD} & & \text{KEEP} \\
\text{KEEP} & & \text{KEEP} & & \text{DISCARD} & & \cdots
\end{align*} \]

2. Bucket Species by Composition

<table>
<thead>
<tr>
<th>C3 H4 O3</th>
<th>H1</th>
<th>C3 H3 O3</th>
<th>H+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Generate reactions by stoichiometry

4. Filter reactions

\[ \begin{align*}
\text{KEEP} & & \text{KEEP} \\
\text{DISCARD} & & \text{DISCARD} \\
\text{DISCARD} & & \text{KEEP}
\end{align*} \]

**Output:** species, reactions that compose network

\[ S_{\text{filtered}} \quad R_{\text{filtered}} \]

Reaction Network Analysis: Graphs vs Kinetic Monte Carlo

- 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 $O(\log N)$ + parallelizable
- Target prediction from full system exploration...?
Reaction Network Monte Carlo:

Inputs:

\[ S_{\text{filtered}} \]

\[ R_{\text{filtered}} \]

\[ [x_i, x_j, …]_0 \]

• 30 of each \( x_i \)
• All \( \Delta G < 0 \): can run to completion
• 100k trajectories

Reaction Network Monte Carlo:

- 30 of each $x_i$
- All $\Delta G < 0$: can run to completion
- 100k trajectories

Can do pathfinding on up to approx. 300 million reactions

Converging RNMC and Identifying Network Products

Can the average trajectory identify network products?
Converging RNMC and Identifying Network Products

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

Can the average trajectory identify network products?
Building Up and Picking Apart Complexity

Species

I: Filter species
- 8,904

II: Filter reactions
- 5,193
- ~176,000,000,000
- ~86,000,000

III: Run MC trajectories*
- 3,182
- ~312,000

IV: Find products in average trajectory
- 18

* 100,000 trajectories with Li⁺, EC, CO₂ at 0V vs. Li/Li⁺

Predicted Battery Network Products: 36 out of 5139

Small molecules/gases

Inorganics

\[ \text{H}_2 \quad \text{CO} \quad \text{OH} \]

\[ \text{lithium carbonate} \quad \text{lithium oxalate} \]

[EC, Li\(^+\)] and [EC, Li\(^+\), CO\(_2\)] at 0V and +0.5V vs. Li/Li\(^+\)

Alkyl carbonates

\[ \text{LMC} \quad \text{vinyl carbonate} \quad \text{ethylene monocarbonate} \]

\[ \text{LiEDC}^- \quad \text{LiBDC}^- \]

Carboxylates, esters, and oxides

\[ \text{LFCO} \]

Cyclic species

\[ \text{bi-dioxoylidene} \]

---

Predicted Battery Network Products: 36 out of 5139

[EC, Li⁺] and [EC, Li⁺, CO₂] at 0V and +0.5V vs. Li/Li⁺

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

Small molecules/gases

\[
\begin{align*}
H₂ & \rightarrow CO \rightarrow OH \\
\text{lithium carbonate} & \rightarrow \text{lithium oxalate}
\end{align*}
\]

Inorganics

Alkyl carbonates

\[
\begin{align*}
\text{LMC} & \rightarrow \text{vinyl carbonate} \rightarrow \text{ethylene monocarbonate} \\
\text{LIEDC⁻} & \rightarrow \text{LiBDC⁻} \\
\end{align*}
\]

Carboxylates, esters, and oxides

\[
\begin{align*}
\text{LFEO} \\
\end{align*}
\]

Cyclic species

\[
\text{bi-dioxolylidene}
\]

Network Path to Refined Mechanism: LFEO

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

From network:

Elementary mechanism:
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!
Background: Nanoscale Patterning with Photolithography
Background: Nanoscale Patterning with Photolithography

- **DUV or EUV Light**
- **Exposure**
- **Baking**
- **Dissolution**
- **Transfer**

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
Phase 1: EUV exposure

Keep reactions that are:
- Electron detachment (all $\Delta G > 0$)
- Electron attachment (all $\Delta G < 0$)
- $\Delta G > 0$ one-bond fragmentation
- $\Delta G > 0$ $H^+$ or $H^0$ transfer

Phase 2: Post-exposure cascade

Remove reactions with:
- $\Delta G > 0$
- Unbalanced redox
- >2 covalent bonds changing
- Sterically hindered reaction center

185,929 reactions

2,776,867 reactions
EUV Lithography Reaction Network Analysis

Phase 1: EUV exposure

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

Network products include:

CO₂

Phase 2: Post-exposure cascade

Species amount

System state

Trajectory step

Initial state

Zooming in: 0.0 - 3.5

Normalized Intensity (I/N₀)

m/z
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

- KEEP
- DISCARD

3. Pathway sampling

   - Perform many thermodynamically bounded Monte Carlo trajectories

   - Extract shortest reaction pathways from each trajectory to each species of interest

   - Transition state calcs
   - Build kinetic models

4. Identify Products

   - Discover novel important species and pathways

   - Under development: ML-assisted network expansion

   - High formation / consumption
   - Significant accumulation
   - Low-cost pathways available
   - Network product?
Background: Upconverting Nanoparticles (UCNPs)

- **Energy Transfer Upconversion**
- **Excited State Absorption**

**Near-Infrared Radiation**

**NaYF₄:Yb³⁺,Er³⁺**

**Security Printing**
- Lu et al., Nat. Photon. 2014

**Bio-imaging**
- Xiong et al., Anal. Chem. 2009

**3D Printing**
- Sanders et al., Nature 2022
UCNP Doping and Heterostructure

Host Material: NaYF$_4$

Dope Y$^{3+}$ sites with Ln$^{3+}$
UCNP Photophysics Can Be Simulated With kMC

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

Generate ensemble of randomly doped structures

Transition rate constants

Energy Transfer
Kinetic Monte Carlo (kMC)

github.com/BlauGroup/NanoParticleTools
github.com/BlauGroup/NPMC
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
Generating a Dataset for Machine Learning

IID Dataset:
- 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

OOD Testing Dataset:
- 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
UCNP kinetic Monte Carlo Simulation Workflow

Supercomputers

Workers

Slurm Jobs

Central Database

MongoDB

Intensity (#/s)

Photon Energy (keV)

Build Fireworks (Jobs)

Query Firework

Initialize Simulation & Write Inputs

CheckPoint

Simulation Complete?

Program Termination

Save

Analysis

Yes

No

Continue?

Return to Previous Directory

Run kMC

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

**Image Representation**
- Prediction Accuracy and Generalizability
- Gradients w.r.t dopant conc.
- Gradients w.r.t layer radii

**Tabular Representation**

<table>
<thead>
<tr>
<th>$r_0$</th>
<th>$V_0$</th>
<th>$x_{0,Yb}$</th>
<th>$x_{0,Er}$</th>
<th>$x_{0,Nd}$</th>
<th>$x_{n,Yb}$</th>
<th>$x_{n,Er}$</th>
<th>$x_{n,Nd}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Core features

Layer features

E. Sivonxay, E. Chan, S. M. Blau, *In preparation*
Developing a Physics-Infused Graph Representation

One-Hot encoding of dopant
Dopant Concentration
Bounding Radii
Dopant Interaction

Edge feature - Integrated Interaction

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

$$II(\mathbf{L}, \mathbf{R}) = \int \int I_{ij} dV_k dV_i$$
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

![Chart comparing validation and OOD test MSE for different representations.](chart.png)

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
Summary

HiPRGen

Network with up to ~200,000,000 reactions

RNMC

LFOE

Supercrystals

Berkeley Lab
Acknowledgements

Daniel Barter  Evan Spotte-Smith  Eric Sivonxay  Jacob Milton  Frances Houle  Kristin Persson  Emory Chan