Active Learning via Bayesian Optimization for Discovery of Energy Storage Materials

Hieu A. Doan
Garvit Agarwal

Molecular Materials Group
Materials Science Division

Doan, Agarwal, Qian, Counihan, Rodríguez-López, Moore, & Assary. (2020). https://doi.org/10.1021/acs.chemmater.0c00768
Redox Flow Battery (RFB) as a Stationary Energy Storage System

Advantages:
- Separation of energy and power capacity
- Integration of intermittent renewable energy sources into the electrical grid
- Non-aqueous RFBs potentially yield high energy density and lower cost

Kowalski, Su, Milshtein, Brushett (2016) Current Opinion in Chemical Engineering
Redox Flow Battery (RFB) as a Stationary Energy Storage System

Advantages:
- Separation of energy and power capacity
- Integration of intermittent renewable energy sources into the electrical grid
- Non-aqueous RFBs potentially yield high energy density and lower cost

Desirable HBEs’ properties:
- Suitable redox potential windows
- High solubility
- Ease of synthesis
- Electrochemical reversibility

Kowalski, Su, Milshein, Brushett (2016) Current Opinion in Chemical Engineering
The Challenges

1. Large space of molecular candidates

HBE scaffold

\[
\begin{align*}
R_1 &= \text{-Me, -Et, -Pr, -Ph, -CN, -Eth, -COMe, -C(Me)Me, -CCOMe} \\
R_2 &= \text{-N(Me)_3^+, -COMe, -Et, -OCMe, N(Me)_2, -NO_2, -C(=O), -Pr, -C(Me)Me, -CCOMe, -C(Me)OMe} \\
R_3 &= \text{-N(Me)_3^+, -Me, -COMe, -Br, -C(=O), -OEt, -Pr, -C(Me)Me, -C(COMe), -C(Me)(OMe)} \\
R_4 - R_5 &= \text{-N(Me)_3^+, -OMe, -COMe, -Br, -N(Me)Me, -Et, -OEt, -NO_2, -C(=O), -Pr, -C(Me)Me, -C(COMe), -C(Me)(OMe)}
\end{align*}
\]

\[
[R3]C1=CC=C(C([R4])([R5])C([R2])-O)[R1])C=C1
\]

> \(10^5\) molecules

2. Expensive/time-consuming synthesis and characterization
The Opportunities

1. Density Functional Theory (DFT) calculations starting from SMILES

\[
E^\text{ox} = \frac{\Delta G^\circ_{\text{ox}}}{nF} - 4.28 \text{ V}
\]

2. Train machine learning (ML) models using DFT-computed \( E^{\text{ox}} \)

Need ML models to not only make accurate predictions but also guide the selection of training data.

Active learning/ Bayesian optimization

(Surrogate model + Acquisition function)
Gaussian Process Regression (GPR) as a surrogate model

In a nutshell: Predict properties/outputs based on feature/input differences (distances)

Covariance is calculated as a function of (feature) distances, e.g.

\[ K(x_1, x_2) = \exp \left[ -\frac{1}{2} \frac{(x_1 - x_2)^2}{l^2} \right] \]
Use of GPR in Active learning/Bayesian optimization

GPR-predicted $\mu(x)$ and $\sigma(x)$ enables Active Learning/Bayesian optimization

"Exploitation": follow $\mu(x)_{\text{max}}$

“Exploration”: follow $\sigma(x)_{\text{max}}$
Acquisition function

Acquisition function formulates an optimal strategy toward an objective by guiding the next evaluation.

"Exploitation": follow $\mu(x)_{\text{max}}$  

“Exploration”: follow $\sigma(x)_{\text{max}}$
Upper Confidence Bound (UCB)

\[ UCB(x) = \mu(x) + \xi \cdot \sigma(x) \]

\[ x_{next} = \arg\max(UCB(x)) \]

\[ \mu(x): \text{Mean} \]
\[ \sigma(x): \text{Standard deviation} \]
Probability of improvement (PI) and Expected improvement (EI)

**Probability of Improvement (PI)**

\[ PI(x) = P\left(y'_x \geq y_{\text{currentBest}} + \xi\right) \]

\[ = \Phi(Z) \]

\[ Z = \frac{\mu(x) - y_{\text{currentBest}} - \xi}{\sigma(x)} \]

\[ x_{\text{next}} = \arg\max(PI(x)) \]

**Expected Improvement (EI)**

\[ EI(x) = \begin{cases} 
(\mu(x) - y_{\text{currentBest}} - \xi)\Phi(Z) + \sigma(x)\phi(x), & \sigma(x) > 0 \\
0, & \sigma(x) = 0 
\end{cases} \]

\[ x_{\text{next}} = \arg\max(EI(x)) \]

- \(\mu(x)\): Mean
- \(\sigma(x)\): Standard deviation
- \(\Phi\): Cumulative Distribution Function
- \(\phi\): Probability Density Function
Problem definition for Bayesian optimization

**Input**

Molecule candidates (SMILES library)

“SMILES.csv”

**Output/Objective**

Molecules w/ desired oxidation potential

RDKit

Feature vectors

E.g.

<table>
<thead>
<tr>
<th>MW</th>
<th># of C</th>
<th># of aromatic rings</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>179</td>
<td>11</td>
<td>1</td>
<td>…</td>
</tr>
</tbody>
</table>

“features.csv” 125-component vectors
Bayesian optimization scheme

**Objective:** Given 1000 HBE molecules, find one with maximized $E^{ox}$ within $N$ evaluations

**Step 1**
- **Surrogate Model Training**
  - Gaussian Process Regression
  - 10 random SMILES with computed $E^{ox}$

**Step 2**
- **Acquisition Function Evaluation**
  - E.g.: Expected-improvement
  - Next SMILES to test

**Step 3**
- **$E^{ox}$ Calculation**
  - DFT simulation

Initial training data

- HBE Library
  - SMILES strings
- "SMILES.csv"
- “features.csv”
  - Features
    - MW, logP, TPSA…
    - Principle Component Analysis

Predicted $\mu$ and $\sigma$

Code adapted from https://github.com/rajak7/Bayesian_Optimization_Material_design
Launch the nanoHUB tool

From your browser go to this link: https://nanohub.org/resources/bayesopt

Click to start the notebook