Accelerating Materials & Molecular Discovery Using Artificial Intelligence and Machine Learning

Nav Nidhi Rajput
navnidhi.rajput@stonybrook.edu
Materials Science and Chemical Engineering
Stony Brook University
Stony Brook, NY
Challenges in Materials Discovery

- Materials and molecules are back-bone of society
- Materials discovery: empirical, uneconomical, inefficient

Serendipity

The Limitations of Trial and Error

- Time-consuming
- Resource-intensive
- Limited scope
- Lack of creativity
Data-driven Next-generation Materials Discovery

Accelerating materials discovery requires:

- Data by exploring relevant composition from a large compositional space
- Improved understanding of composition-structure-processing property
- Accurate knowledge of material response across multiple length scales
Data-driven Next-generation Materials Discovery

Databases
- DFT databases
- Empirical databases
- Polymer databases
- MOF databases

Materials Prediction
(Machine learning incorporating domain knowledge)

Robotic Synthesis
(Self-driving laboratories for synthesis and characterization)

Data Driven Materials Discovery

Text mining
(Extracting structured and unstructured data from text and images)

Data-driven Efforts Start with *Data*

<table>
<thead>
<tr>
<th>Name</th>
<th>Material types</th>
<th>Source</th>
<th>No. of entries</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIST ICSD</td>
<td>Inorganic</td>
<td>Empirical</td>
<td>210,000</td>
<td>License</td>
</tr>
<tr>
<td>Pauling File</td>
<td>Inorganic</td>
<td>Empirical</td>
<td>156,274</td>
<td>Open</td>
</tr>
<tr>
<td>PoLyInfo</td>
<td>Polymers</td>
<td>Empirical</td>
<td>334,738</td>
<td>Open</td>
</tr>
<tr>
<td>Cambridge Structural Database</td>
<td>Organic, MOFs</td>
<td>Empirical</td>
<td>&gt;1 million</td>
<td>Open/license</td>
</tr>
<tr>
<td>MatWeb</td>
<td>Inorganic, organic</td>
<td>Empirical</td>
<td>135,000</td>
<td>License</td>
</tr>
<tr>
<td>Total Metals</td>
<td>Metals</td>
<td>Empirical</td>
<td>350,000</td>
<td>License</td>
</tr>
<tr>
<td>INTERGLAD</td>
<td>Glasses</td>
<td>Empirical</td>
<td>350,000</td>
<td>License</td>
</tr>
<tr>
<td>Mindat</td>
<td>Minerals</td>
<td>Empirical</td>
<td>5,500</td>
<td>Open</td>
</tr>
<tr>
<td>ASM Databases &amp; Handbooks</td>
<td>Alloys</td>
<td>Empirical</td>
<td>–</td>
<td>License</td>
</tr>
<tr>
<td>American Mineralogist Crystal Structure Database</td>
<td>Minerals</td>
<td>Empirical</td>
<td>–</td>
<td>Open</td>
</tr>
<tr>
<td>ChemSpider</td>
<td>Organic</td>
<td>Empirical, computational</td>
<td>81 million</td>
<td>Open</td>
</tr>
</tbody>
</table>

Changes in data management policies

FAIR (findable, accessible, interoperable and reusable) data principles provide guidelines for scientific data management

Batra et al. Machine Learning Reviews, 2021
High-throughput multi scale simulations and machine learning infrastructure for liquids and solid-liquid interfaces

Existing Open-Source Software Tools for Materials Applications: What is missing?
Materials Informatics for Structure-Property-Relationship

An Open-Source High-Throughput Multi-Scale Infrastructure for Materials Design

Validation

1. Optimizations
   - Frequency
   - Redox Potential
   - NMR Shifts
   - Binding Energy

2. Force Field

3. Salvation Structure
   - Viscosity,
   - Conductivity
   - Interfacial
   - Interfaces

4. Machine-Learning Models
   - Atom Types
   - Bond Lengths
   - Angles
   - Torsions

DFT

Clusters

Clustering
- Coordination Types
- RDF

Database

Large Datasets for Quantifying Structure-Property Relations

Screening Molecules for Complex Liquid Solutions

E_bond = \sum_{bonds} K_r(r - r_{eq})^2

\[ H\Psi = \mathcal{F}\Psi \]

\[ F = ma \]

Atwi et al Scientific Reports, 2022
Several DFT-based and Classical Molecular Dynamics Simulations based workflows

- NMR Chemical Shift
  - Technical Validation
  - Individual species
  - MD Simulations

- Electrochemical and Chemical Stability

- Structural and Dynamical Properties
  - ML predicted Diffusion Coefficient
  - Binding Energy

Atwi et al Scientific Reports, 2022
Workflow for Machine Learning Model Development

1. Data Extraction
   - Use a natural language processing pipeline for tokenizing, tagging, and parsing literature
   - Database: electrolyte formulations and Coulombic Efficiency (CE)

2. Feature Engineering
   - Derive structural and DFT properties of molecular components in the database
   - Properties included are based on our-prior physics-based studies

3. Model Development
   - Benchmark machine learning (ML) models with different architectures
   - Select and fine-tune model parameters for better prediction capabilities

4. Model Validation
   - Validate and evaluate the model performance on validation and test datasets, respectively
   - Interpret the model output by identifying features with high-impact on predicted properties

In collaboration with Prof. Haibin Ling (CS, SBU) and experimental team at Pacific Northwest National Laboratory
Knowledge discovery from spectroscopy literature

In collaboration with Prof. Haibin Ling (CS, SBU) and experimental team at Pacific Northwest National Laboratory

1. Manual Annotation + Dataset Creation
   - 200 battery papers
   - Electrolyte compositions
     - electrolyte: "1 M LiTFSI in DOL/DME, cation: Li^+ (lithium), anion: TFSI^−, solvents: DOL, DME", abbreviations: Li^+, lithium, DOL...
     - temperature: "25 °C, ...
   - Spectra collection (PNG)

2. BatteryBert Model Adaptation
   - Original BatteryBert
     - Single sentences or paragraphs
   - Modified BatteryBert
     - Entire paper text (From XML/HTML)
   - Training dataset
   - Testing dataset

3. BatteryBert Model Fine-Tuning
   - Pre-trained BatteryBert model
   - Training on annotated dataset with new hyper-parameters
   - More stable output

4. LineFormer Model Fine-Tuning
   - Model training for subplot handling

5. Fine-Tuned Models Evaluation
   - Reliability in extracting new compositions and spectrum data using evaluation metrics (Precision, recall, F1, etc.)
   - New experiments to fill missing gaps
   - Output dataset from fine-tuned models
Major Challenges Persists

- Constant flux of high-fidelity data generated in a consistent and systematic manner
- Benchmark datasets are necessary for consistent testing of new algorithms
- Interpretability of ML models for outlier remains a major challenge
- Adequate training of the current and next generation of materials scientists on AI and ML methods is needed to ensure the effective and appropriate utilization of these tools
- Interdisciplinary collaboration
Scientific NLP

Extracting information from literature using Natural language processing

- The majority of scientific knowledge about materials is scattered across the text, figures, and tables of millions of academic research papers
- Create software tools for auto-generating materials database

<table>
<thead>
<tr>
<th>Entity recognition toolkits</th>
<th>Information capable of extracting</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChemDataExtractor(^3)</td>
<td>Chemicals Tables</td>
</tr>
<tr>
<td>ChemicalTagger(^3)</td>
<td>Chemicals Quantities Synthesis actions and conditions</td>
</tr>
<tr>
<td>Chem Spot 2.0(^{14,79})</td>
<td>Chemicals</td>
</tr>
<tr>
<td>BANNER-CHEMDNER(^{27})</td>
<td>Chemicals Bio-relevant entities</td>
</tr>
<tr>
<td>ChemXSeer(^{10}) and TableSeer(^{31})</td>
<td>Chemicals Tables</td>
</tr>
<tr>
<td>OSCAR(^4)</td>
<td>Chemicals Reaction names Bio-relevant entities</td>
</tr>
<tr>
<td>LeadMine(^{82})</td>
<td>Chemicals Named reactions Bio-relevant entities</td>
</tr>
<tr>
<td>tmChem(^{31})</td>
<td>Chemicals</td>
</tr>
</tbody>
</table>
FAIR (findable, accessible, interoperable, and reusable) Datasets

MISPR databases

- **DFT**
  - Base & derived molecules
  - Clusters
  - Properties (NMR, IP/EA, ...)
  - Runs
  - Metadata:
    - Molecule info
    - Software info
    - Solvent model

- **Force Field**
  - GAFF: derived using infrastructure
  - Other force field: obtained from external sources
  - Metadata:
    - Molecule info
    - Force field references

- **MD**
  - Properties (diffusion, viscosity, coordination number, ...)
  - Runs
  - Metadata:
    - Composition
    - Operating conditions
An Automated Solvation Structure Characterization Tool In MISPR

Critical for understanding structural properties of molecules and clusters ...

- Individual species
- MD Simulations
- Cluster Analysis
- Top Clusters

Nuclear Magnetic Resonance (NMR)

MongoDB Database

NMR Chemical Shifts

Technical Validation

Structure Optimization
Vibration Frequency
Imaginary Frequency?
Data Parsing
NMR Calculation

Atwi et al. Nature Computational Science, 2022
Discrepancies in literature regarding solvation structure of Mg(TFSI)$_2$ in DME

**SCXRD:** Structure for MgTFSI$_2$ single crystal, recrystallized from solution  
*Salama et al.* 2016

**NMR:** Number of bound DME per Mg$^{2+}$ at varying temperatures and concentrations  
*Ying et al.* 2020

**MD Simulations:** Coordination between Mg$^{2+}$ and other electrolyte components  
*Rajput et al.* 2015

**Kubisiak & Eilmes** 2018
Predicted $^{25}\text{Mg}$ NMR Chemical Shifts using the NMR Computational Protocol

Atwi et al. Nature Computational Science, 2022
High-throughput Capability of the Workflow: $^{13}$C and $^{1}$H NMR chemical shift of 100 molecules

- **DFT details:**
  - Solvent: chloroform
  - Solvation model: PCM
  - Level of theory: ωB97X/def2-TZVP

- Computed $^{13}$C and $^{1}$H chemical shifts deviate from unity (desired slope = 1) by 0.05 and 0.01 ppm, respectively

- High correlation coefficients are obtained

---

Structure of NMR document

- `id`: 
- `molecule`: 
- `module`: 
- `pymatgen.core.structure`
- `bclass`: `Molecule`
- `charge`: 0
- `spin_multiplicity`: 1
- `sites`: `[...]
- `smiles`: "O1CCOCOCOCOCOCOCOCOCOCOC1"
- `inch`: "InChI=1S/C12H24O6/c1-2-14-5-6-16-9-10-18-12-11-17-8-7-15-4-3-13-1/h1-12H2"
- `formula_alphabetical`: "C12 H24 O6"
- `chemsys`: "C-H-O"
- `energy`: -923.134
- `tensor`: 
- `type`: "O"
- `isotropic`: 293.7568
- `Anisotropy`: 46.9776
- `eigenvalues`: [...]
- `functional`: "wb97X"
- `basis`: "Def2TZVP"
- `phase": "solution"
- `solvent": "chloroform"
- `solvent_model": "pcm"
- `solvent_properties": null
- `tag": "htp-paper"
- `state": "successful"
- `wall_time": 8076.92
- `version": "0.0.1"
- `gauss_version": "ES64L-G16RevC.01"
- `last_updated": "2021-05-23T21:00:58.289Z"
- `run_ids": [...]

---

Atwi et al. Nature Computational Science, 2022
Computational Database of Electrolyte Properties

https://github.com/rashatwi/combat

DFT Properties
Electronic & thermodynamic properties

Li-S Database

Ensemble Properties
Structural & dynamical properties

Guiding strategies for the development of effective solvents for Li-S liquid electrolytes

Electrolytes composition

1 M LiTFSI, 0.25 M Li₂S₈, in:

Base: DOL/DME (1/1, v/v)
Variable: DOL/Co-solvent (1/1, v/v)