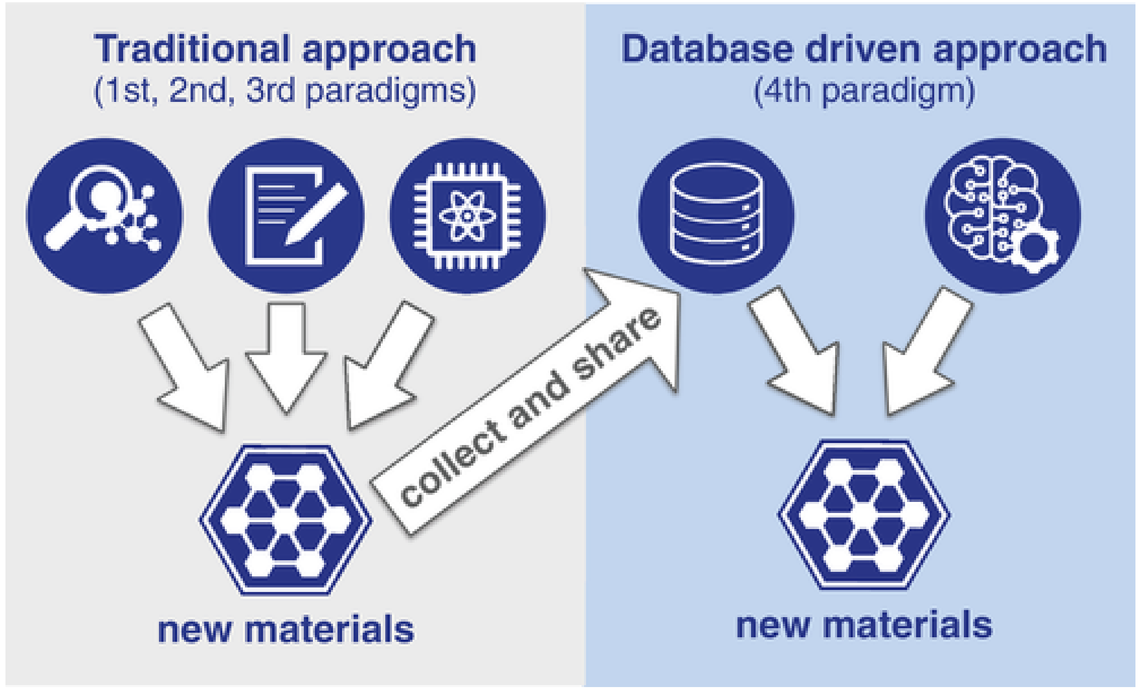
Accelerating Materials & Molecular Discovery Using Artificial Intelligence and Machine Learning



Himanen et a. 2019 Advanced Science

MISPR
materials informatics for structure property relationships

MDPropTools

Nav Nidhi Rajput

navnidhi.rajput@stonybrook.edu

Materials Science and Chemical Engineering

Stony Brook University

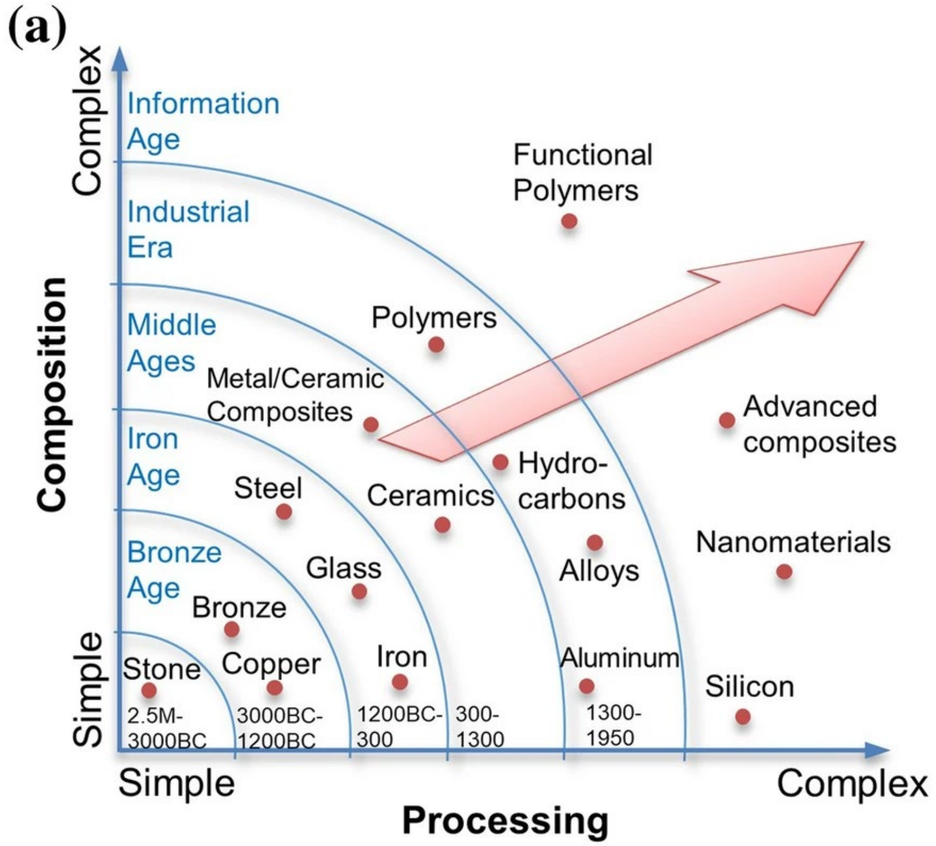
Stony Brook, NY

Stony Brook University

MoIMD
MOLECULAR SIMULATIONS FOR MATERIALS DESIGN

Challenges in Materials Discovery

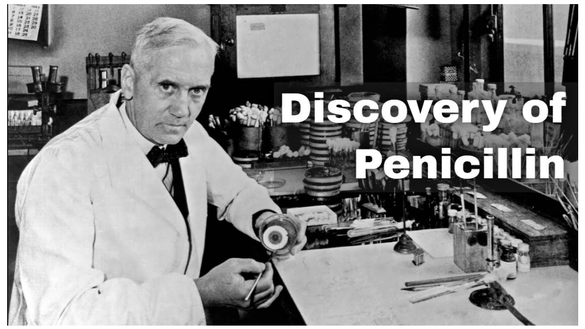
- Materials and molecules are back-bone of society



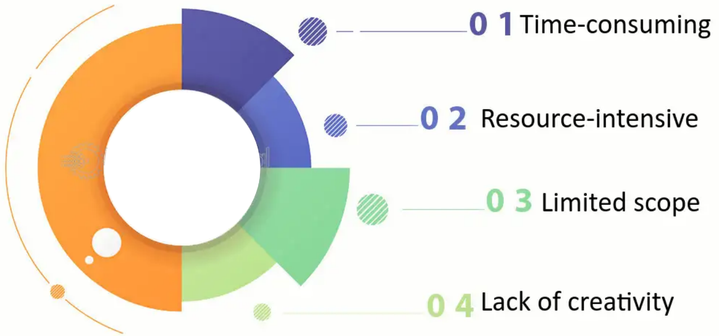
Wagner J Mater Sci. 2021

- Materials discovery: empirical, uneconomical, inefficient

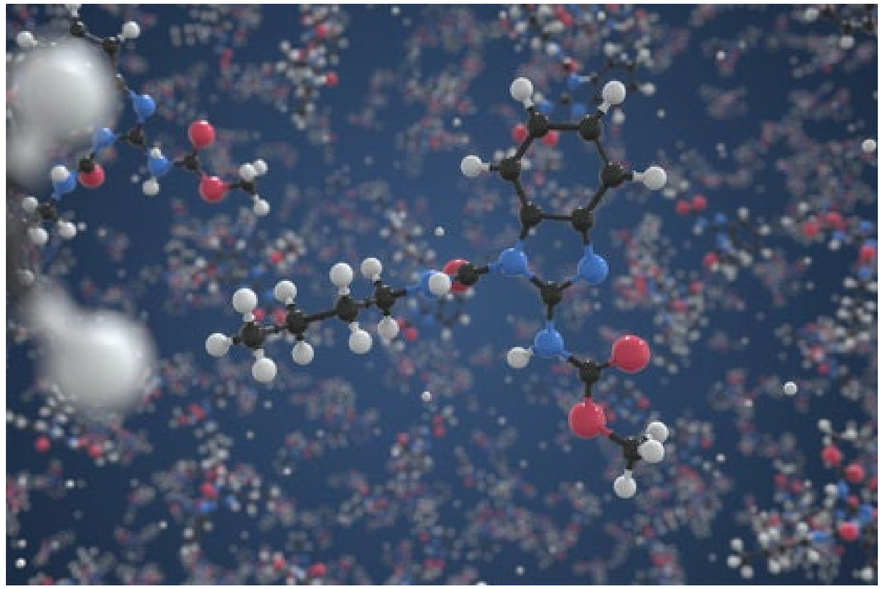
Serendipity

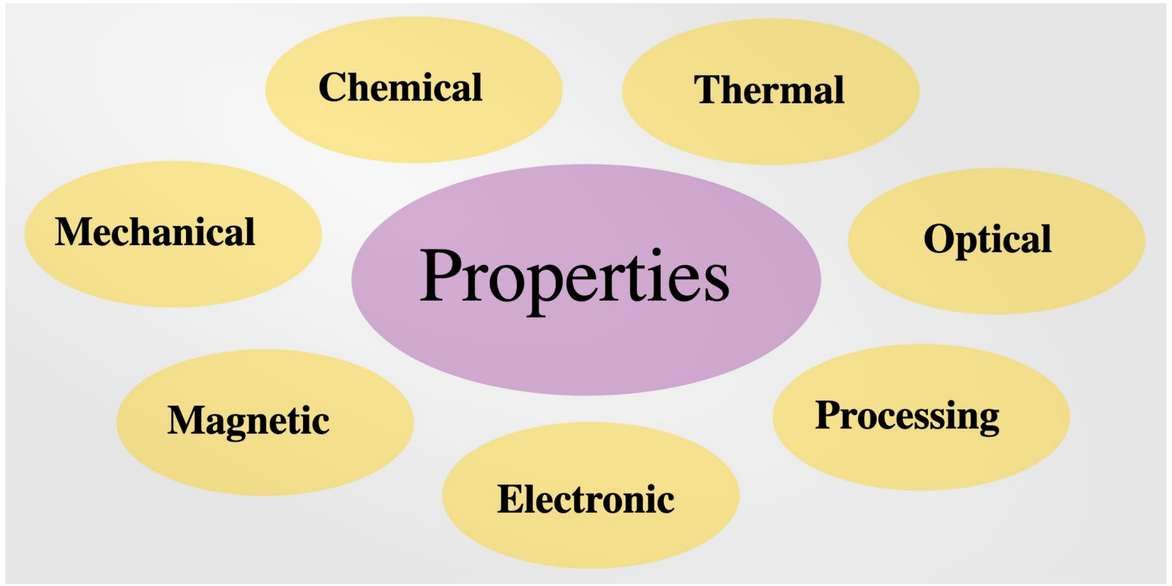


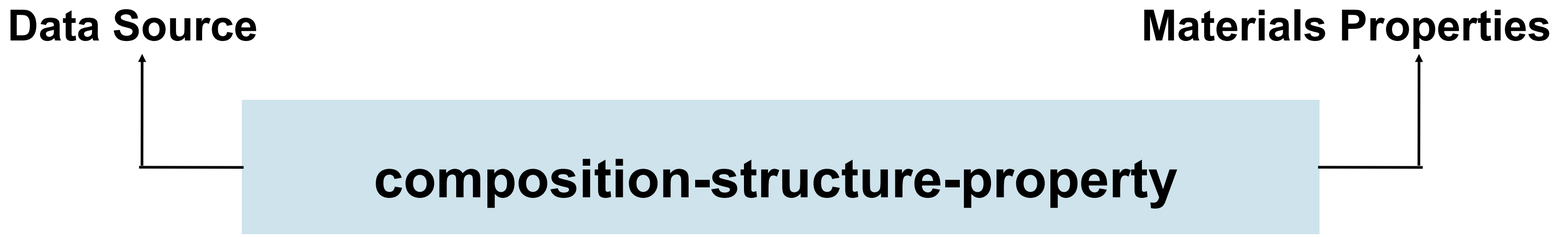
The Limitations of Trial and Error



Data-driven Next-generation Materials Discovery



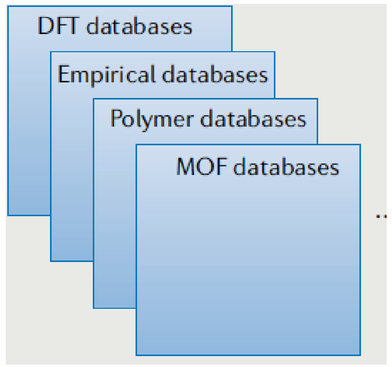




- Accelerating materials discovery requires:
  - Data by exploring relevant composition from a large compositional space
  - Improved understanding of composition-structure-processing property
  - Accurate knowledge of material response across multiple length scales

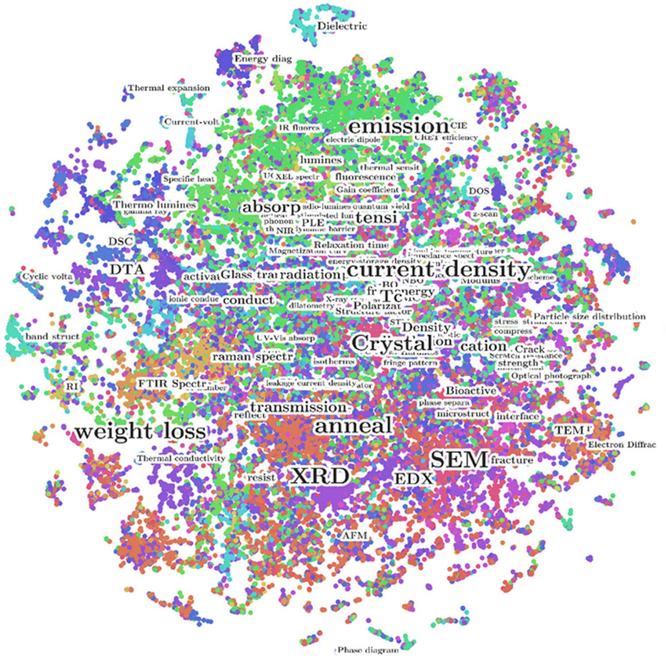
Data-driven Next-generation Materials Discovery

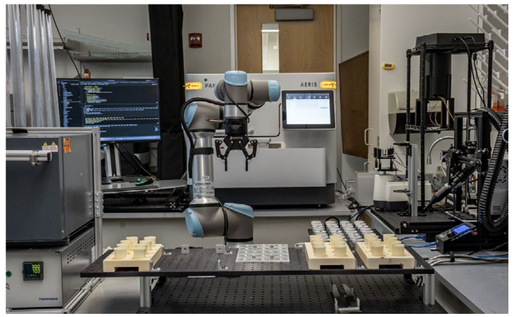
Databases



Materials Prediction

(Machine learning incorporating domain knowledge)





Data Driven Materials Discovery

Robotic Synthesis

(Self-driving laboratories for synthesis and characterization)

Text mining

(Extracting structured and unstructured data from text and images)

Venugopal et al. International Journal of Applied Glass Science, 2021

Data-driven Efforts Start with *Data*

Name	Material types	Source	No. of entries	Access
NIST ICSD[231]	Inorganic	Empirical	210,000	License
Pauling File[232]	Inorganic	Empirical	156,274	Open
PoLyInfo[233]	Polymers	Empirical	334,738	Open
Cambridge Structural Database[234]	Organic, MOFs	Empirical	>1 million	Open/license
MatWeb[235]	Inorganic, organic	Empirical	135,000	License
Total Metals[236]	Metals	Empirical	350,000	License
INTERGLAD[237]	Glasses	Empirical	350,000	License
Mindat[238]	Minerals	Empirical	5,500	Open
ASM Databases & Handbooks[239]	Alloys	Empirical	–	License
American Mineralogist Crystal Structure Database[240]	Minerals	Empirical	–	Open
ChemSpider[243]	Organic	Empirical, computational	81 million	Open

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

Existing Open-Source Software Tools for Materials Applications: What is missing?

MATERIALS PROJECT

AFLOW
Automatic - FLOW for Materials Discovery

NOMAD

OQMD
The Open Quantum Materials Database

JARVIS
ML
Tools
Accuracy & precision
Derived Apps
FF
DFT & Beyond

High-throughput multi scale simulations and machine learning infrastructure for liquids and solid-liquid interfaces

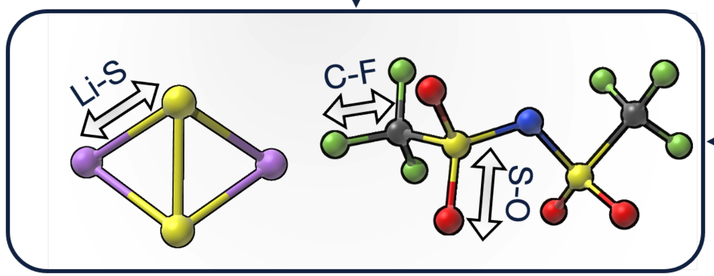
Promising materials

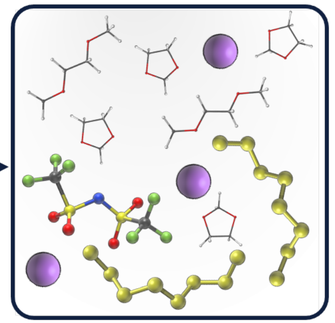
Database

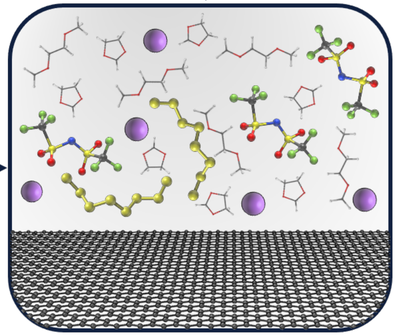
Bond dissociation energy
Electrochemical stability
Formation energy

Solvation structure
Ionic conductivity
Viscosity

Molecule distribution
Diffusion mechanism
Number density







DFT

2-3 nm (10^2 atoms)

MD simulations (Electrolyte)

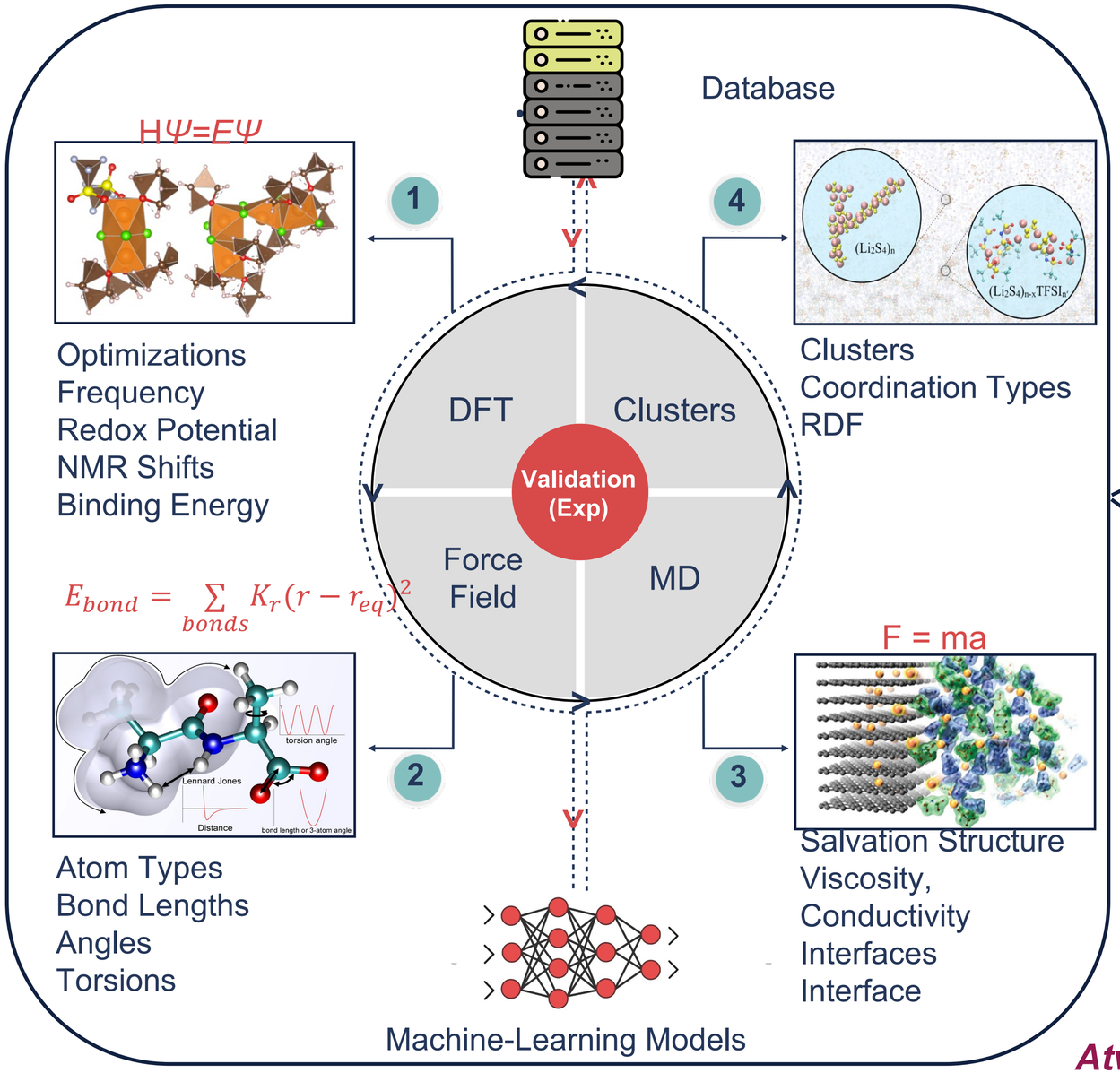
10 nm (10^5 atoms)

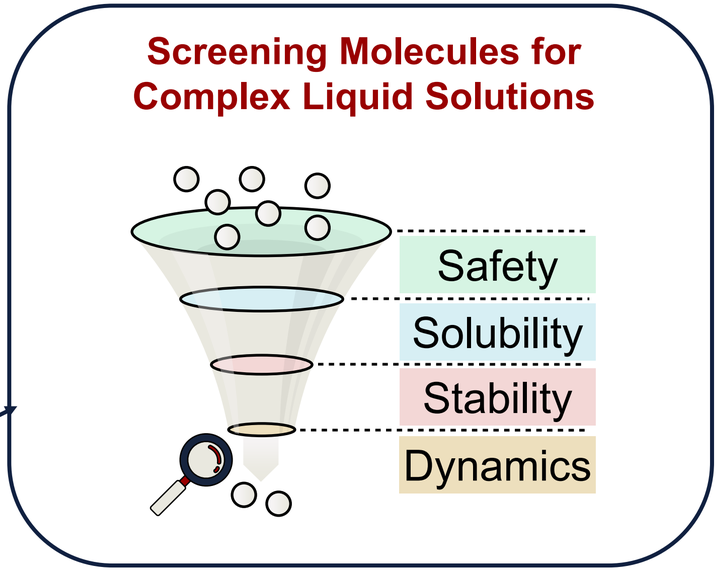
MD simulations (Interface)

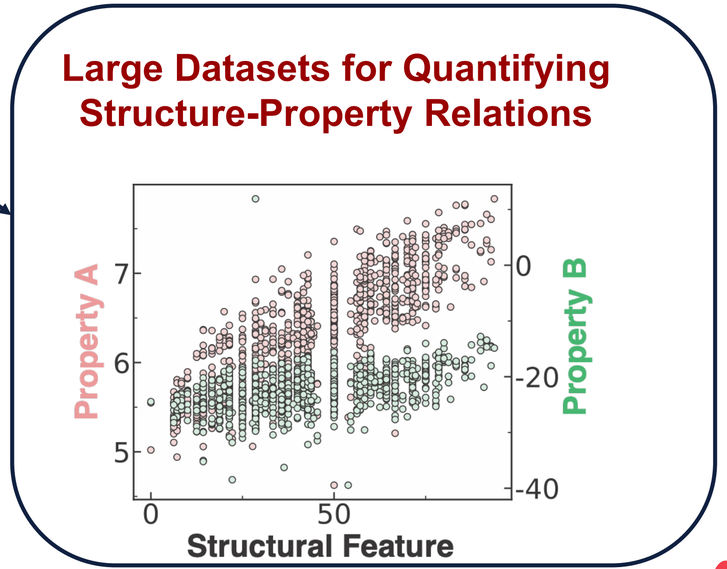
10 nm (10^5 atoms)

MISPR Materials Informatics for Structure-Property-Relationship

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





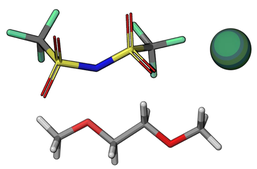


Atwi et al Scientific Reports, 2022

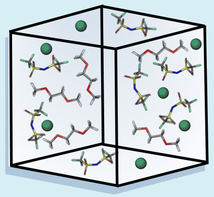
Several DFT-based and Classical Molecular Dynamics Simulations based workflows

NMR Chemical Shift

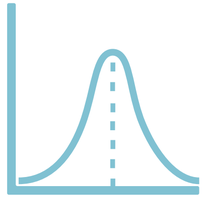
Individual species



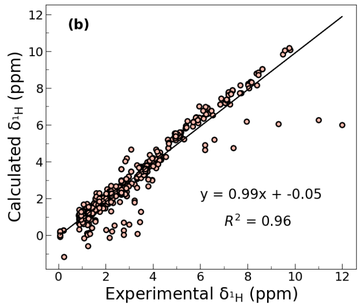
MD Simulations



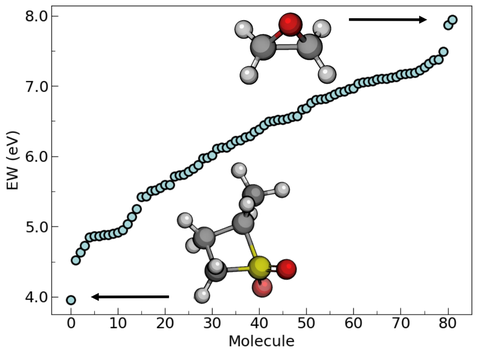
Technical Validation



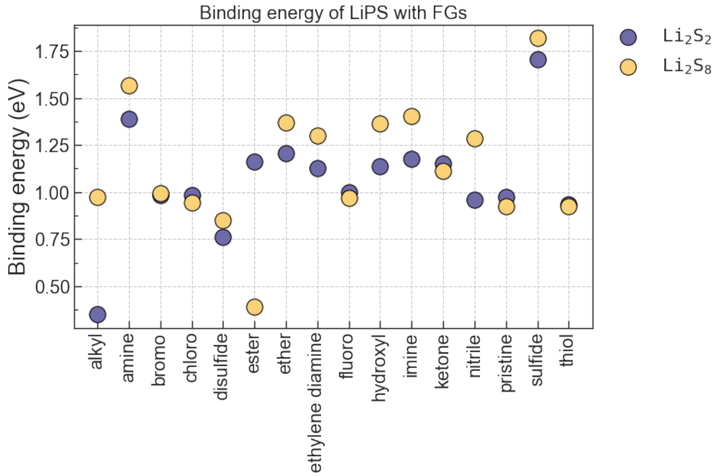
NMR Chemical Shift



Electrochemical and Chemical Stability

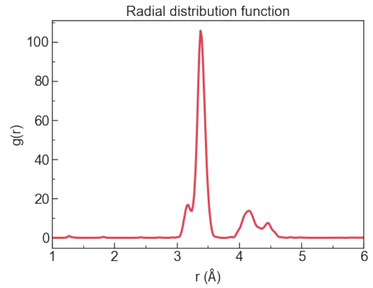


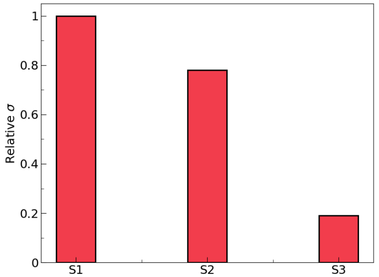
Binding Energy



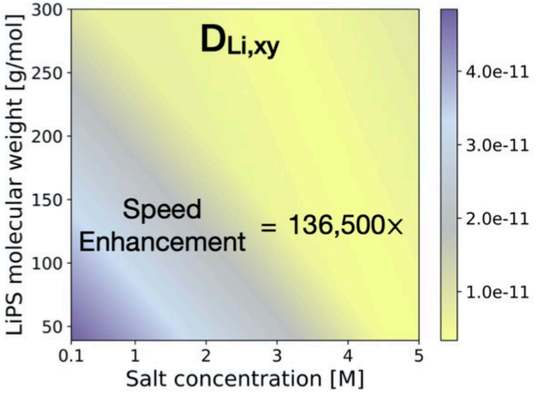
MDPropTools

Structural and Dynamical Properties





ML predicted Diffusion Coefficient



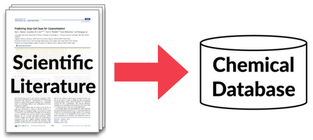
Atwi et al Scientific Reports, 2022

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

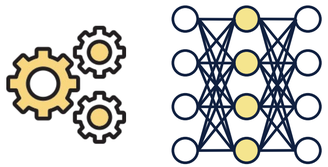
- Derive structural and DFT properties of molecular components in the database
- Properties included are based on our-prior physics-based studies

Fluorination %
C/O ratio
Steric effect
Inorganic/Organic
Binding energy …

3

Model Development

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

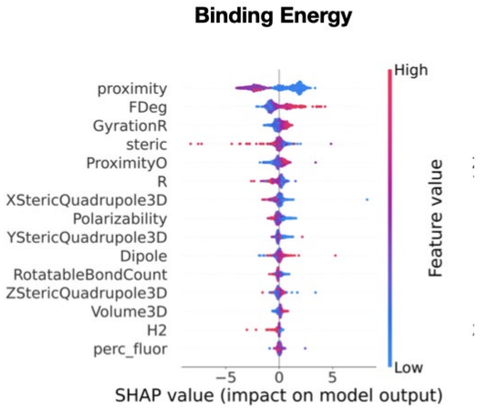


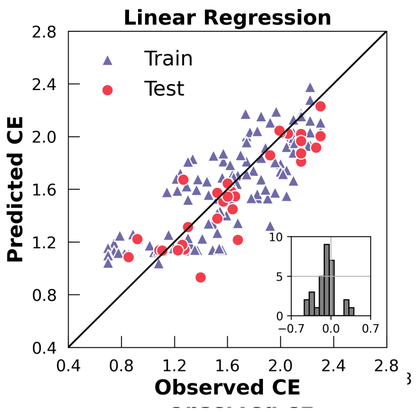
4

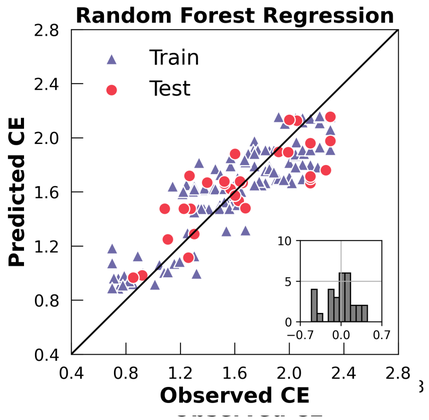
Model Validation

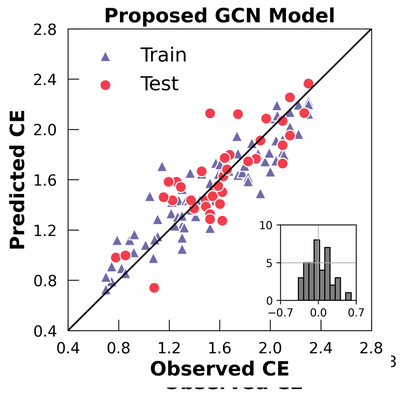
- Validate and evaluate the model performance on validation and test datasets, respectively
- Interpret the model output by identifying features with high-impact on predicted properties

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

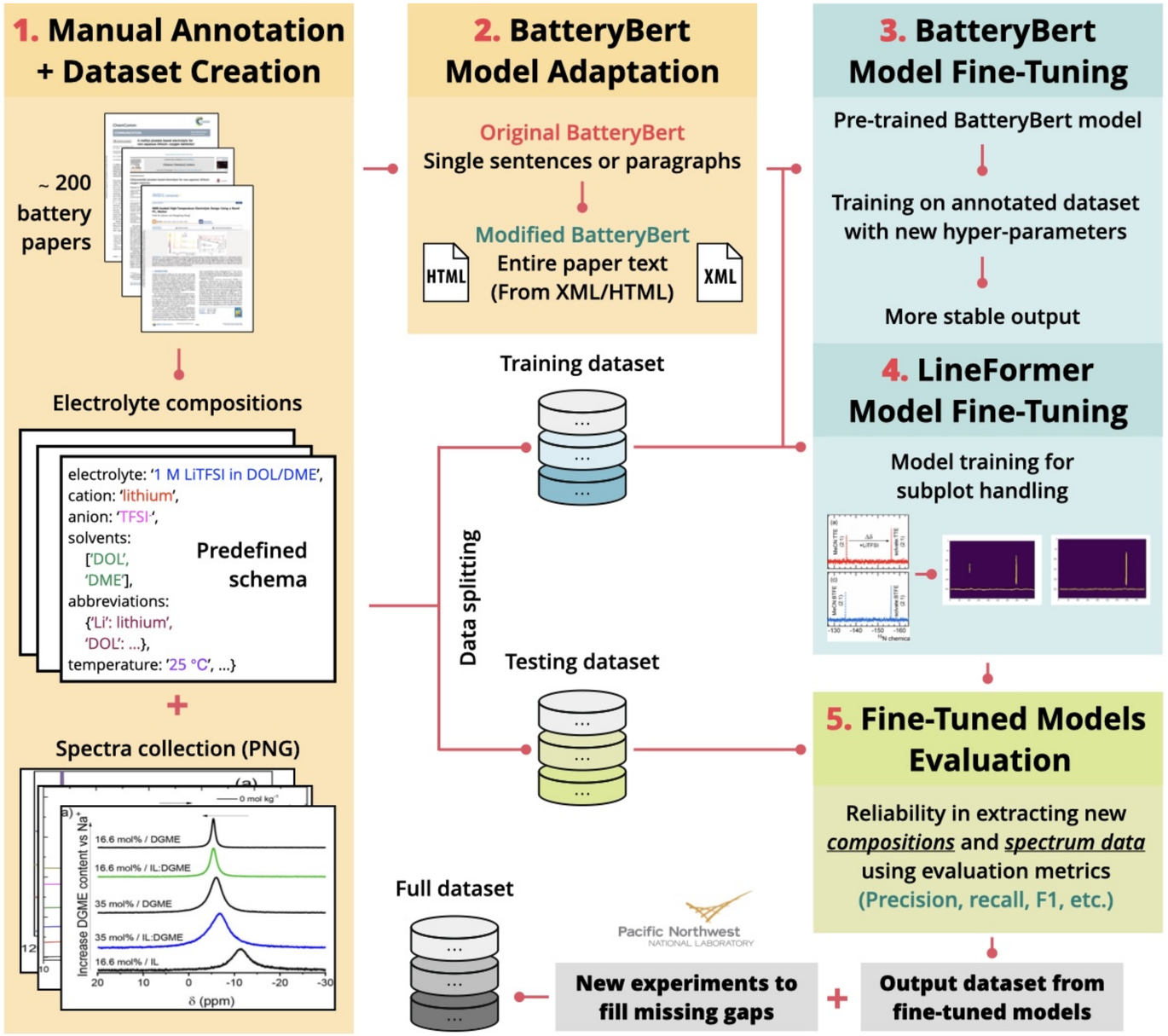






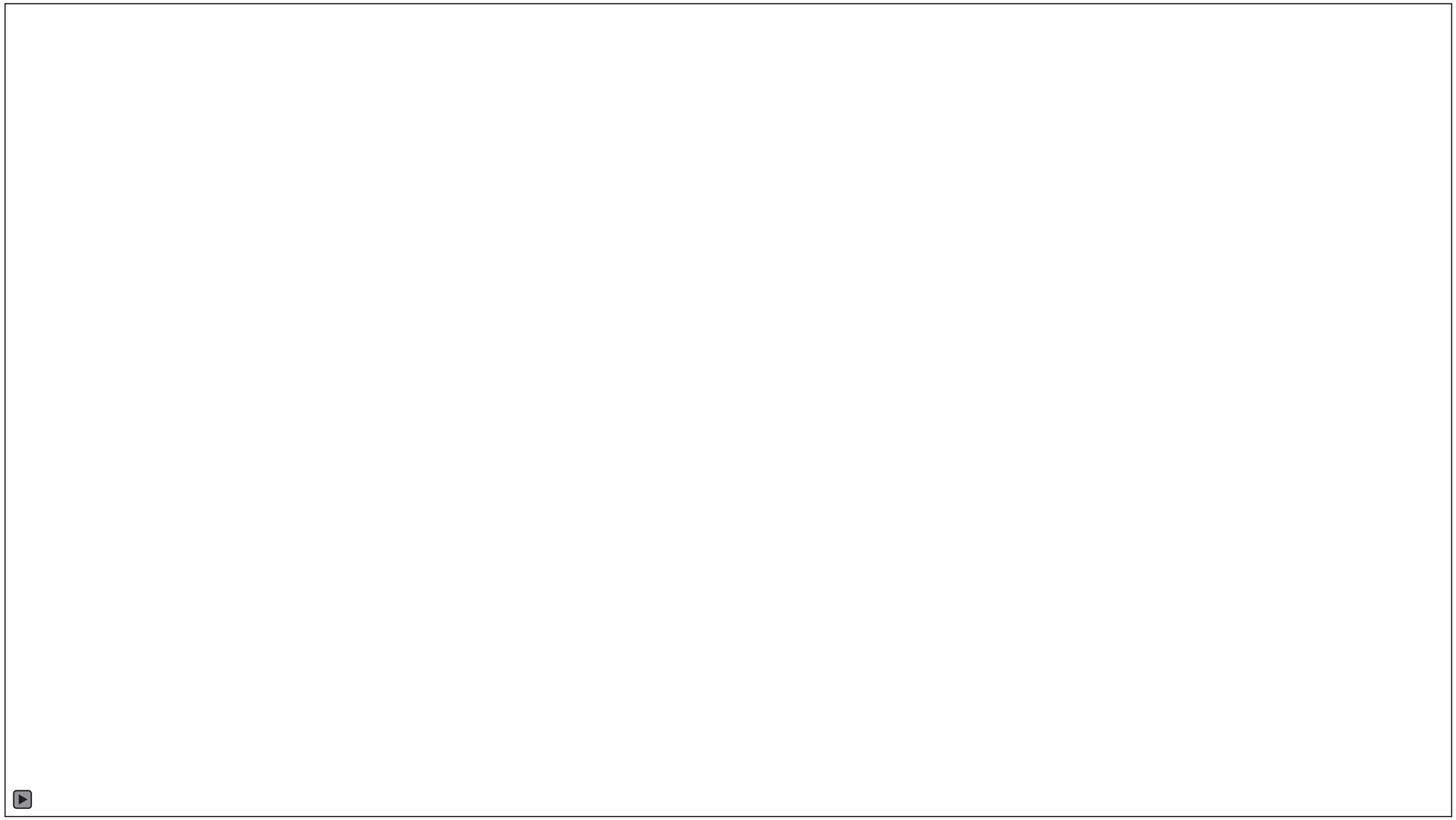


Knowledge discovery from spectroscopy literature



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

Autonomous Materials Laboratories



Hitosugi & Shimizu Lab

Major Challenges Persists

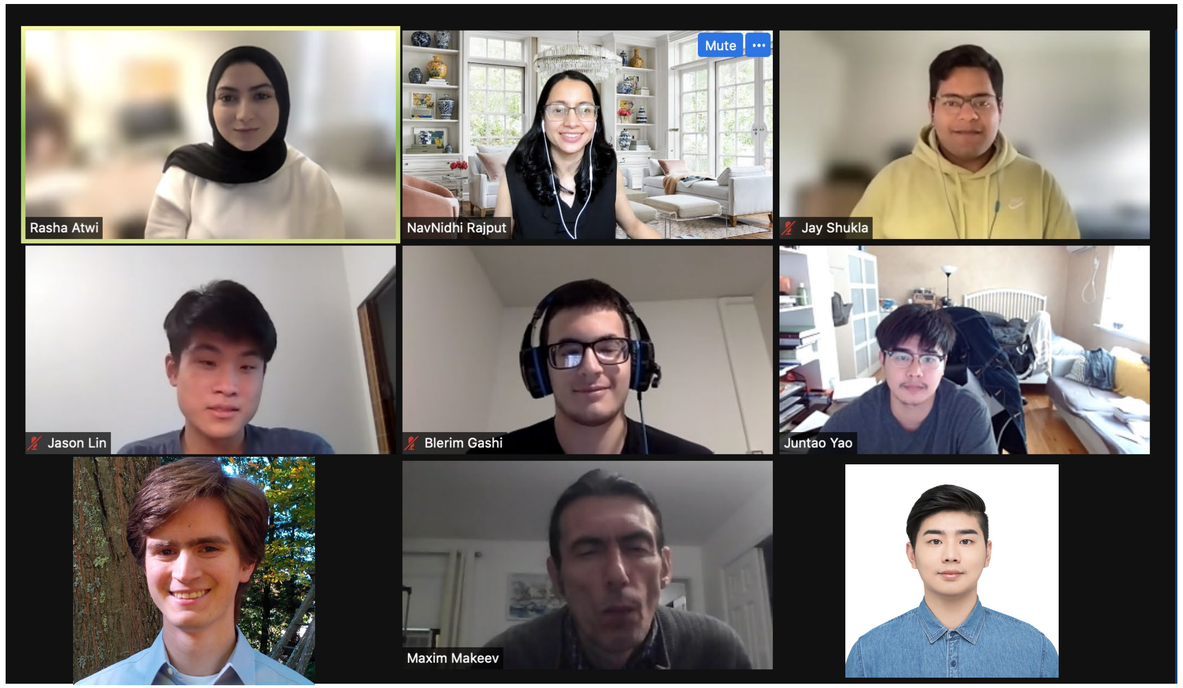
- Constant flux of high-fidelity data generated in a consistent and systematic manner
- Benchmark datasets are necessary for consistent testing of new algorithms
- Interpretability of ML models for outlier remains a major challenge
- Adequate training of the current and next generation of materials scientists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

Acknowledgments

MolMD
MOLECULAR SIMULATIONS FOR MATERIALS DESIGN

MISPR
materials informatics for structure property relationships

MDPropTools



Funding

LG Energy Solution

Office of Naval Research
ONR
Science & Technology

Research Computing Resources

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

XSEDE
Extreme Science and Engineering Discovery Environment

Collaborations

Stony Brook University

Pacific Northwest NATIONAL LABORATORY

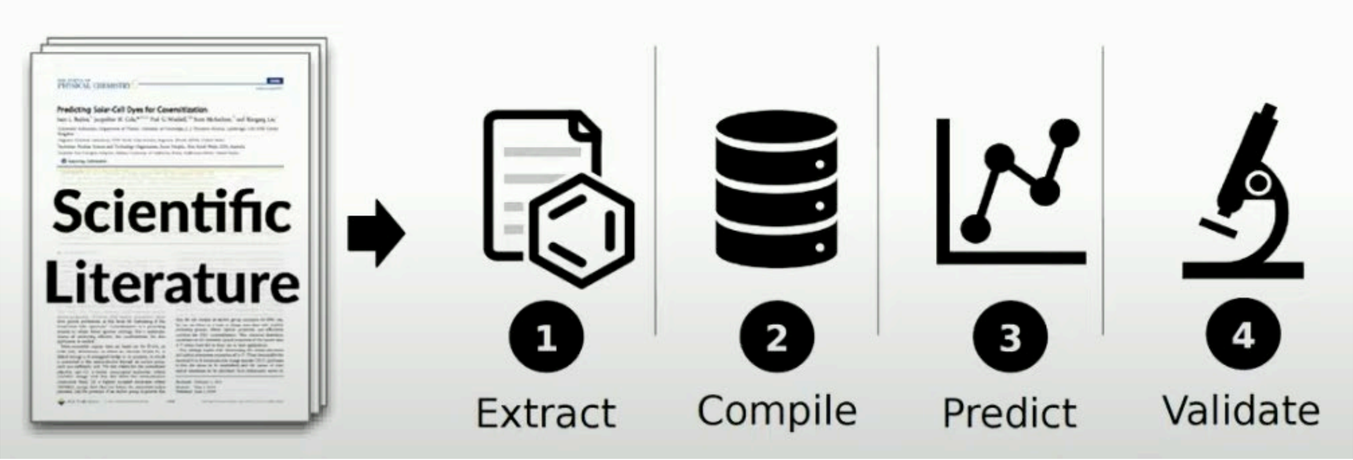
NYU

PURDUE UNIVERSITY

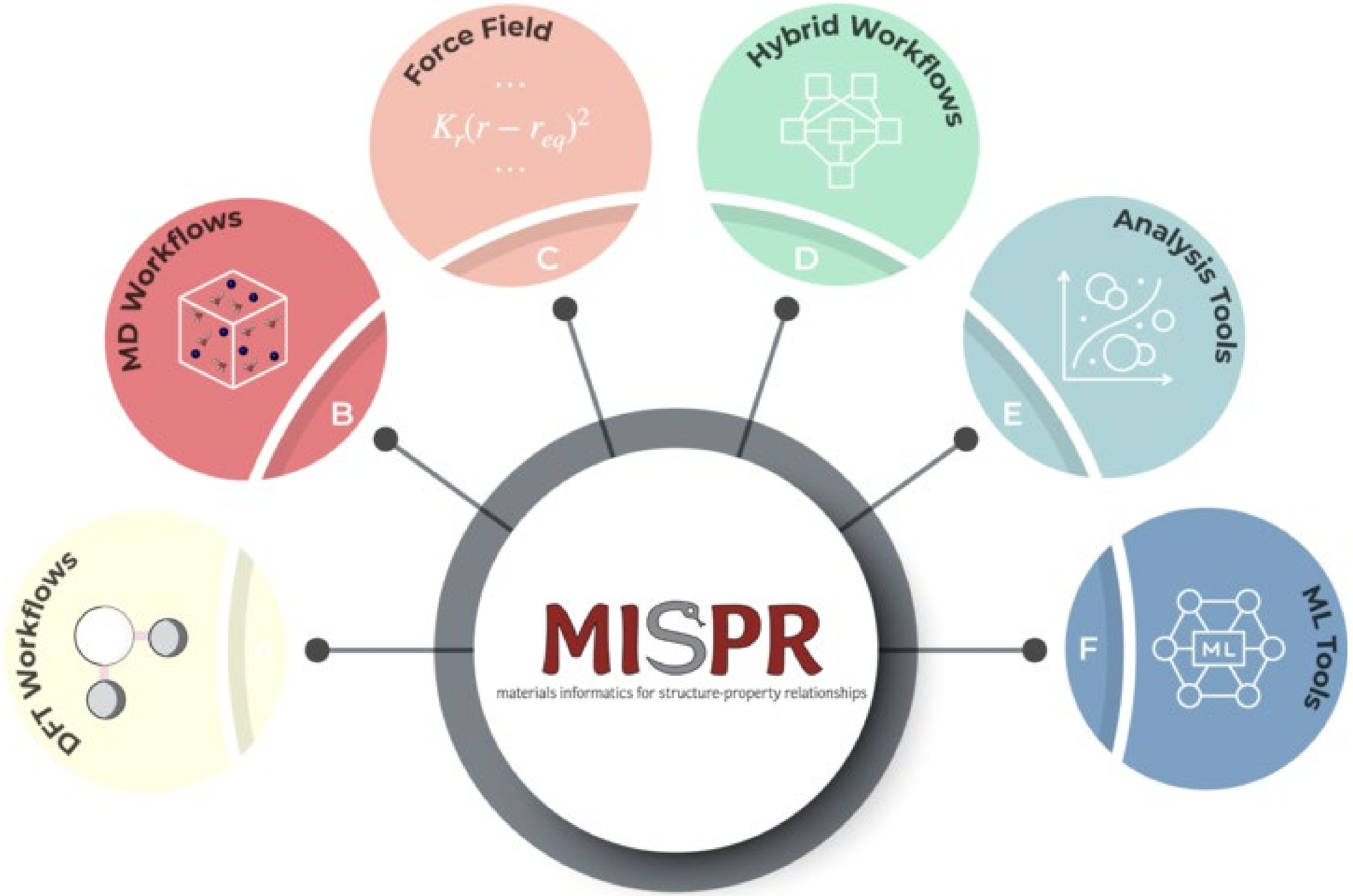
Scientific NLP

Extracting information from literature using Natural language processing

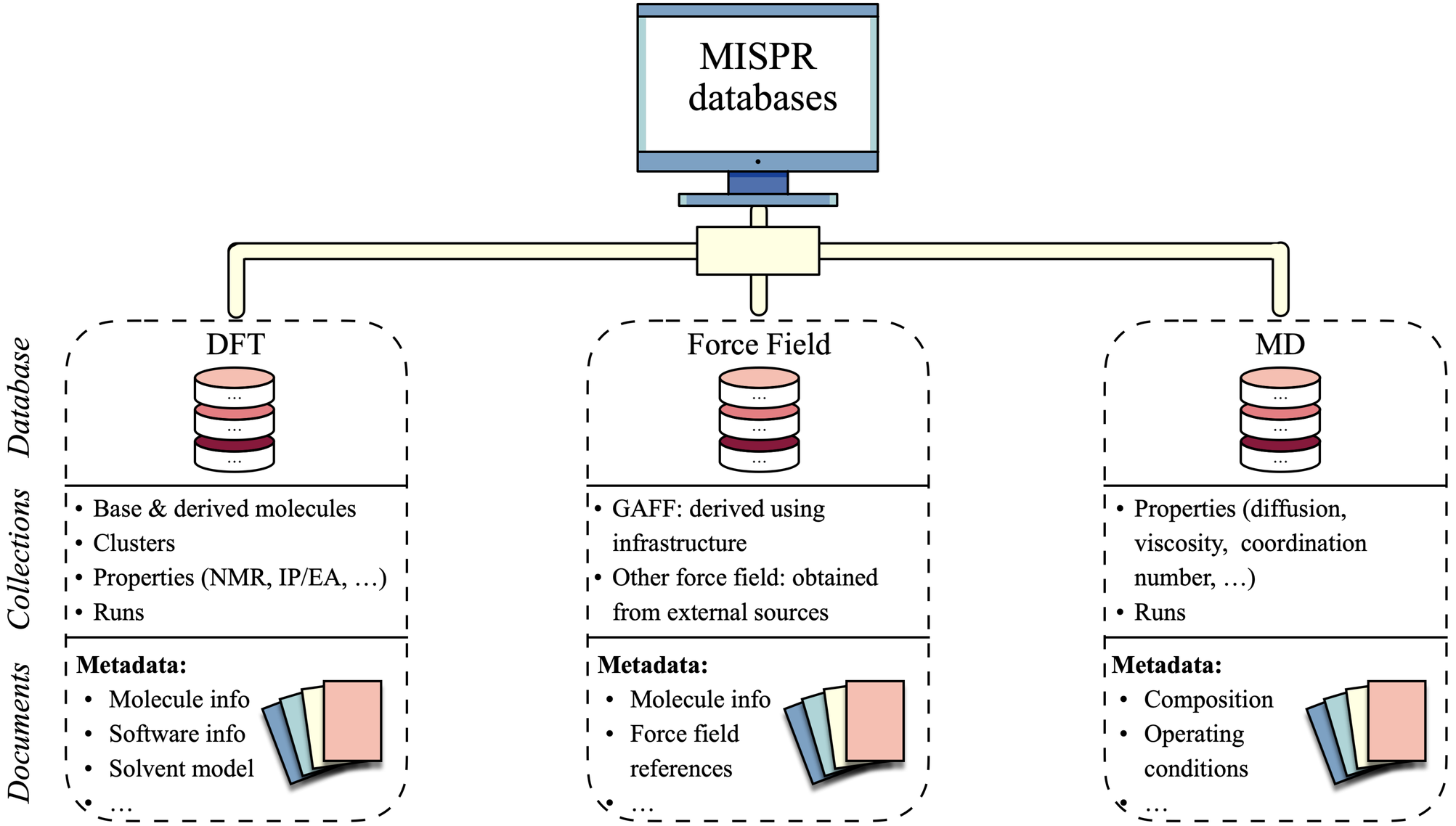
- The majority of scientific knowledge about materials is scattered across the text, figures, and tables of millions of academic research papers
- Create software tools for auto-generating materials database



Entity recognition toolkits	Information capable of extracting
ChemDataExtractor[33]	Chemicals Tables
ChemicalTagger[61]	Chemicals Quantities Synthesis actions and conditions
Chem Spot 2.0[14,79]	Chemicals
BANNER-CHEMDNER[27]	Chemicals Bio-relevant entities
ChemXSeer[80] and TableSeer[81]	Chemicals Tables
OSCAR4	Chemicals Reaction names Bio-relevant entities
LeadMine[82]	Chemicals Named reactions Bio-relevant entities
tmChem[31]	Chemicals

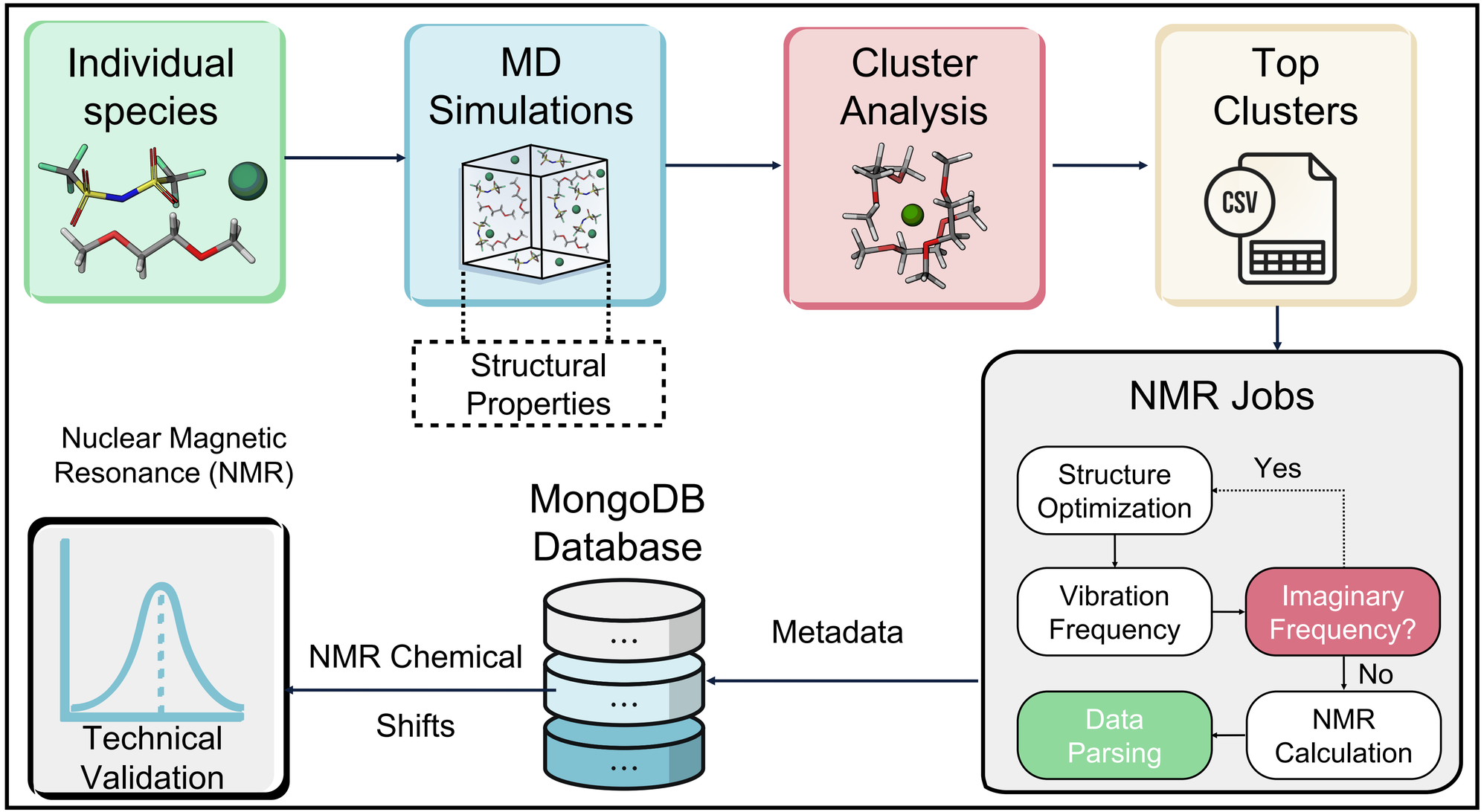

Force Field
...
$K_r(r - r_{eq})^2$
...
C
Hybrid Workflows
D
MD Workflows
B
Analysis Tools
E
DFT Workflows
A
MISPR
materials informatics for structure-property relationships
F
ML
ML Tools

FAIR (findable, accessible, interoperable, and reusable) Datasets



An Automated Solvation Structure Characterization Tool In MISPR

Critical for understanding structural properties of molecules and clusters …

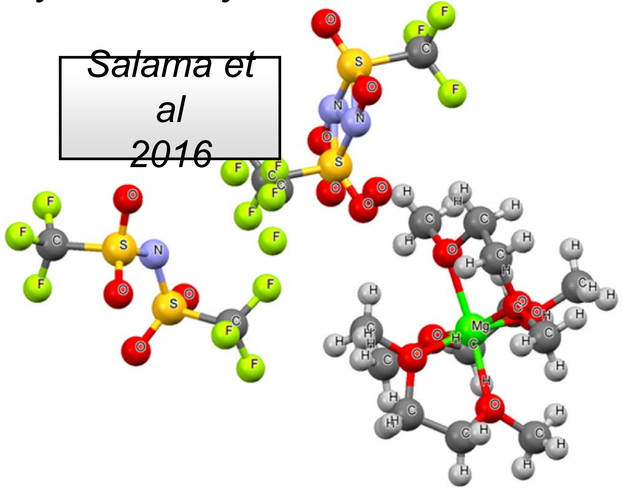


Atwi et al Nature Computational Science, 2022

Discrepancies in literature regarding solvation structure of $Mg(TFSI)_2$ in DME

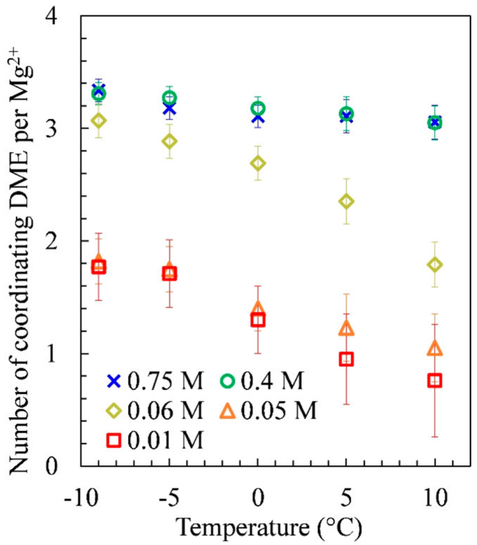
SCXRD: Structure for $MgTFSI_2$ single crystal, recrystallized from solution

Salama et al 2016



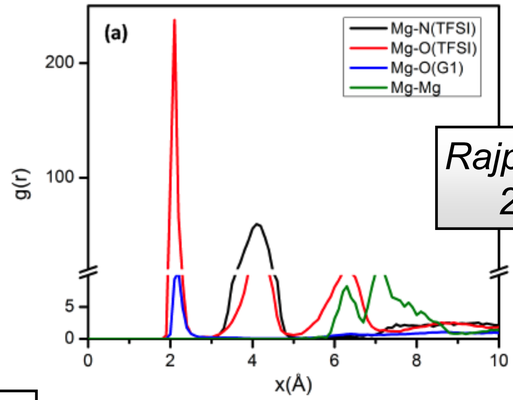
NMR: Number of bound DME per Mg^{2+} at varying temperatures and concentrations

Ying et al 2020

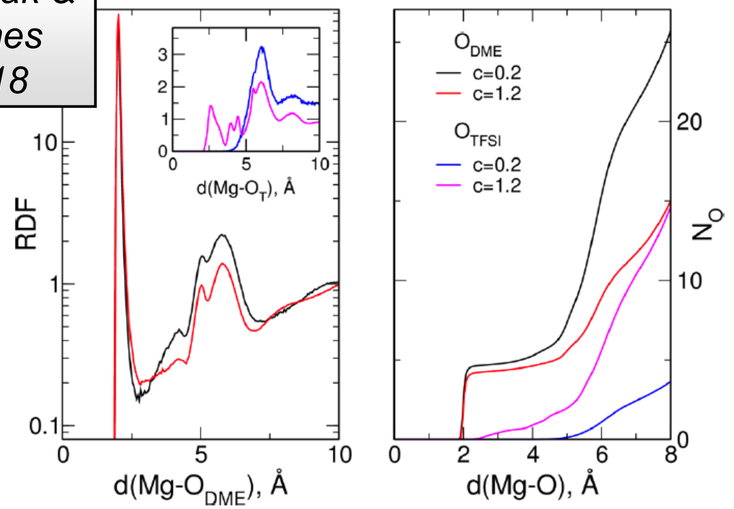


MD Simulations: Coordination between Mg^{2+} and other electrolyte components

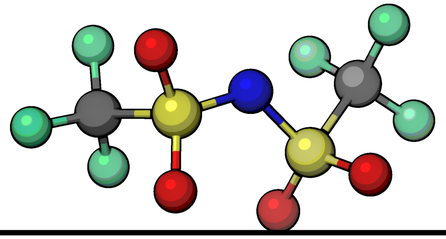
Rajput et al 2015



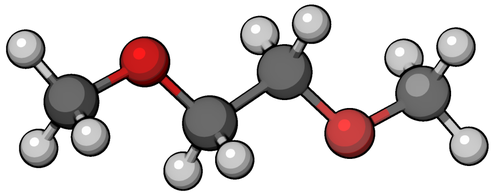
Kubisiak & Eilmes 2018



TFSI

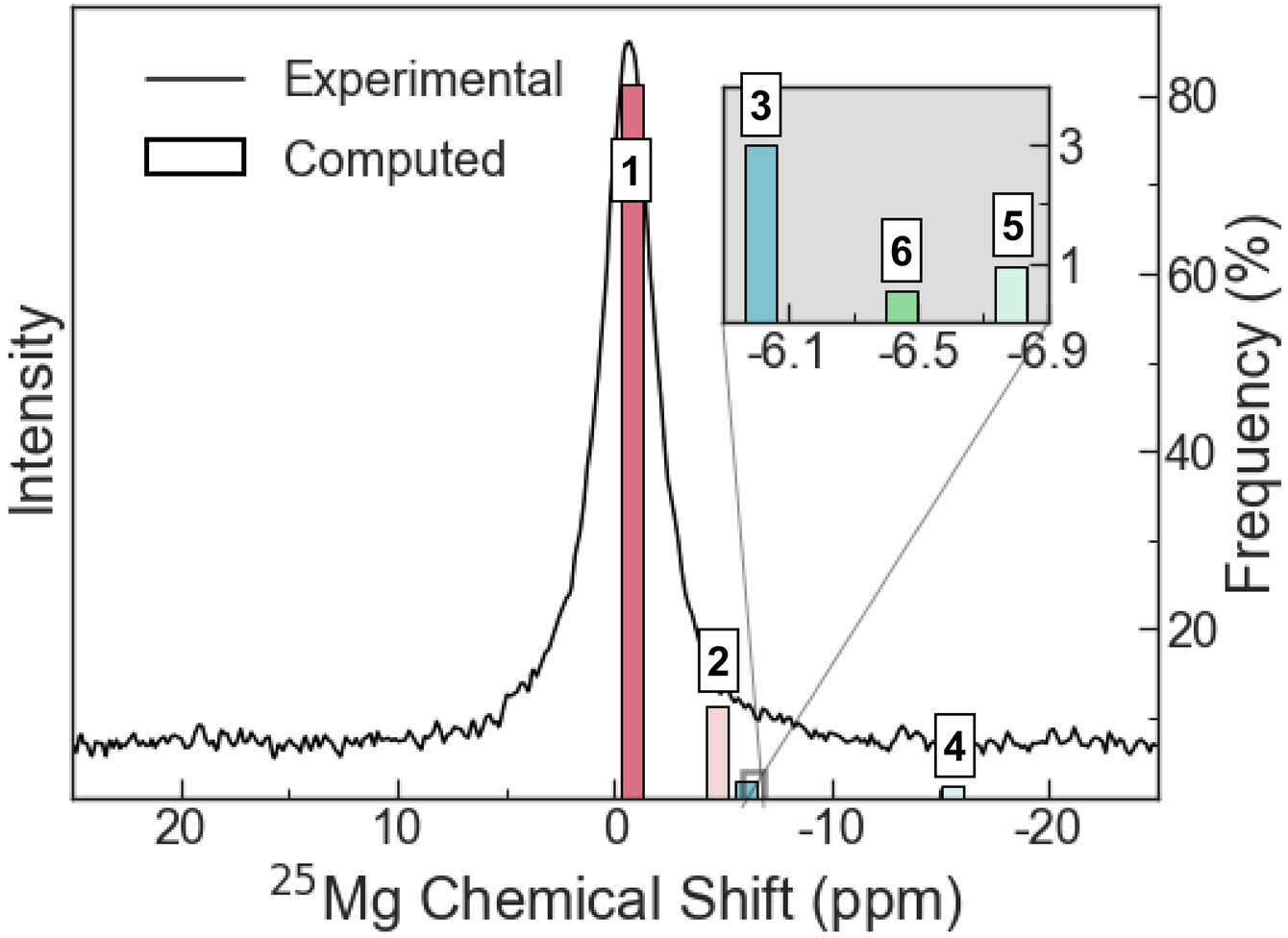


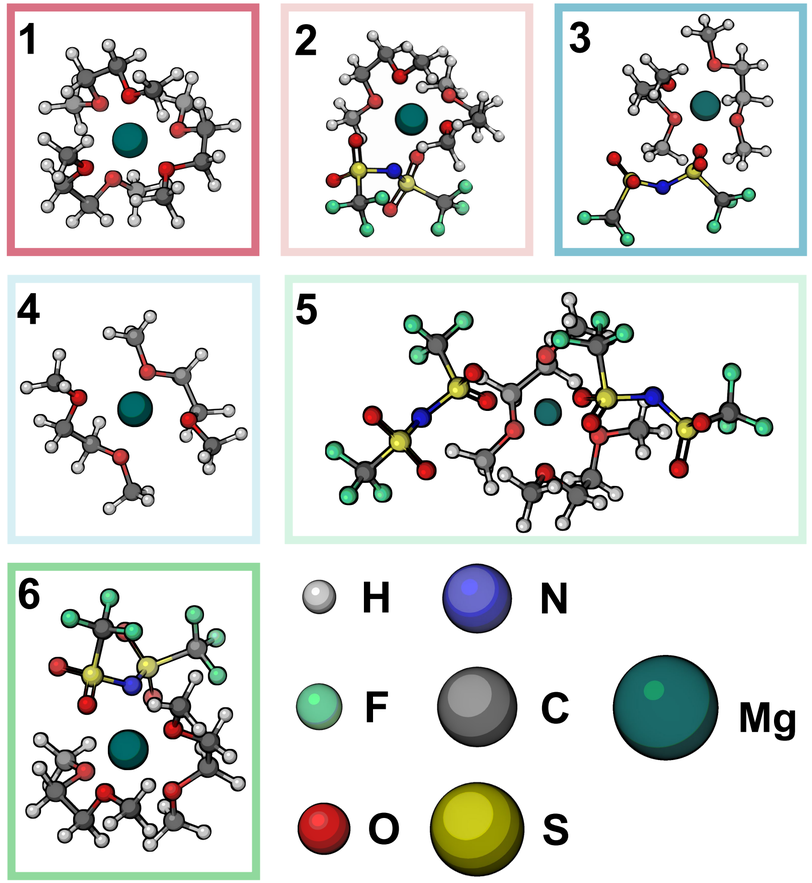
DME



Predicted ^{25}Mg NMR Chemical Shifts using the NMR Computational Protocol

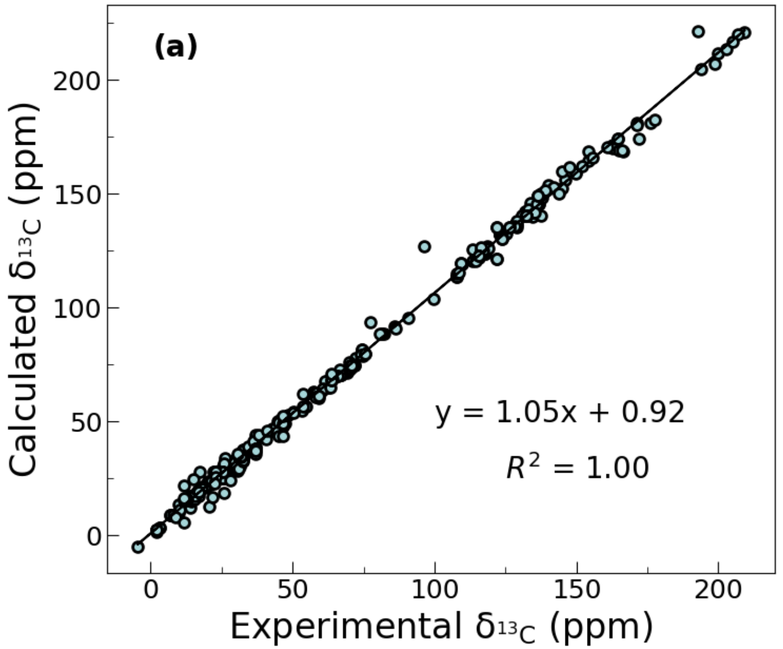
Rasha Atwi

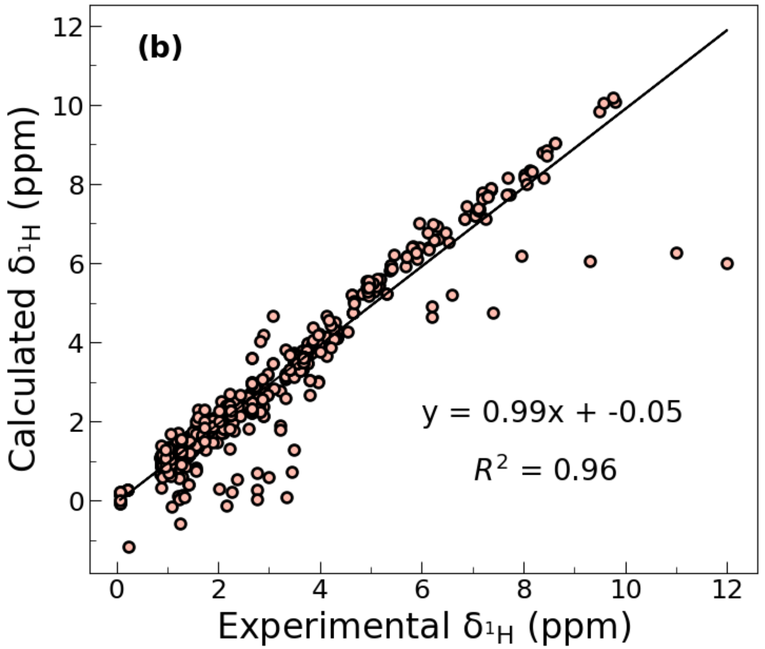




Atwi et al Nature Computational Science, 2022

High-throughput Capability of the Workflow: ^{13}C and ^{1}H NMR chemical shift of 100 molecules





Structure of NMR document

{"_id": {"$oid": "60aac28a6dec7edccfa2c88b"},
"molecule": {"@module": "pymatgen.core.structure",
"@class": "Molecule",
"charge": 0,
"spin_multiplicity": 1,
"sites": [...]},
"smiles": "O1CCOCCOCCOCCOCCOCC1",
"inchi": "InChI=1S/C12H24O6/c1-2-14-5-6-16-9-10-18-12-11-17-8-7-15-4-3-13-1/h1-12H2",
"formula_alphabetical": "C12 H24 O6",
"chemsys": "C-H-O",
"energy": -923.134,
"tensor": {"1": {"type": "O" ,
"Isotropic": 293.7568,
"Anisotropy": 46.0776,
"tensor": [[...],[...],[...]],
"eigenvalues": [..., ..., ...],
...},
"functional": "wB97X",
"basis": "Def2TZVP",
"phase": "solution",
"solvent": "chloroform",
"solvent_model": "pcm",
"solvent_properties": null,
"tag": "htp-paper",
"state": "successful",
"wall_time (s)": 8076.92,
"version": "0.0.1",
"gauss_version": "ES64L-G16RevC.01",
"last_updated": {"$date": "2021-05-23T21:00:58.269Z"},
"run_ids": [...]}

- **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

Atwi et al Nature Computational Science, 2022

Computational Database of Electrolyte Properties

https://github.com/rashatwi/combat

Paper ID: 3741952

MISPR

https://github.com/molmd/mispr

Package for HTP simulations

MDPropTools

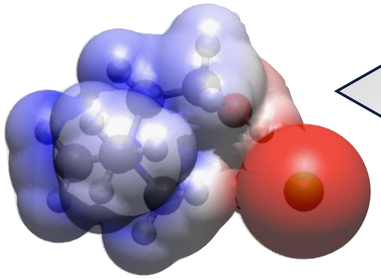
https://github.com/molmd/mdproptools

Package for MD analysis

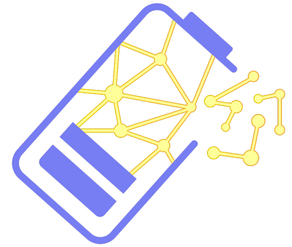
Code powering the simulations

DFT Properties

Electronic & thermodynamic properties



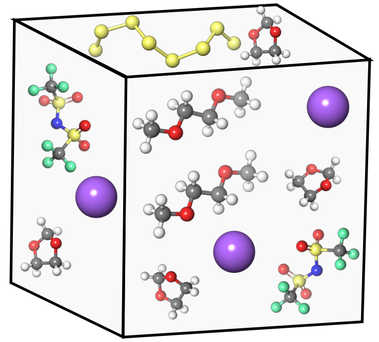
Li-S Database



ComBat

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 Li_2S_8, **in:**

Base: DOL/DME (1/1, v/v)

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