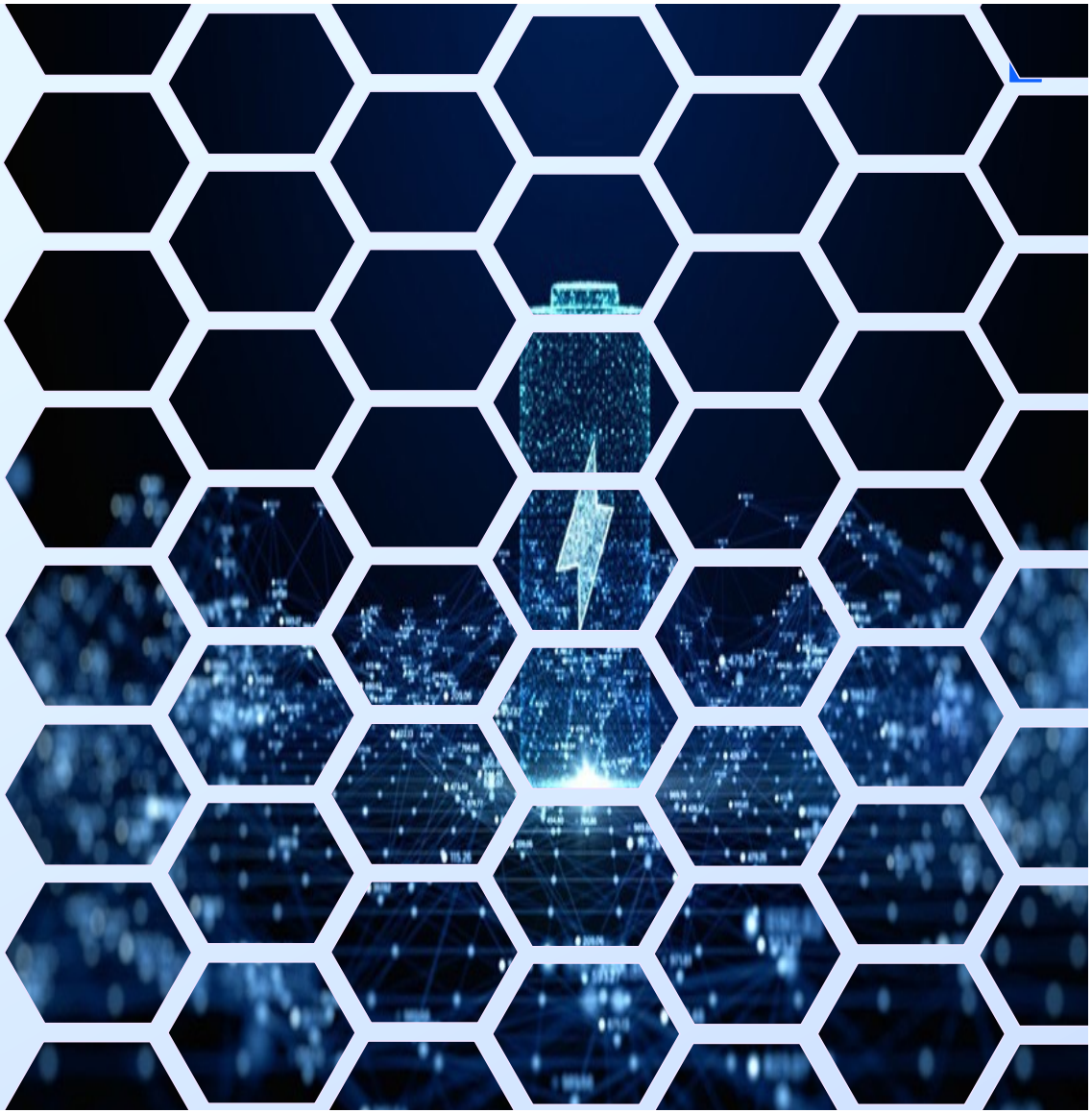
Chemical Foundation Models for Complex Materials

Vidushi Sharma
Staff Research Scientist,
IBM Almaden Research Center
San Jose, CA, USA



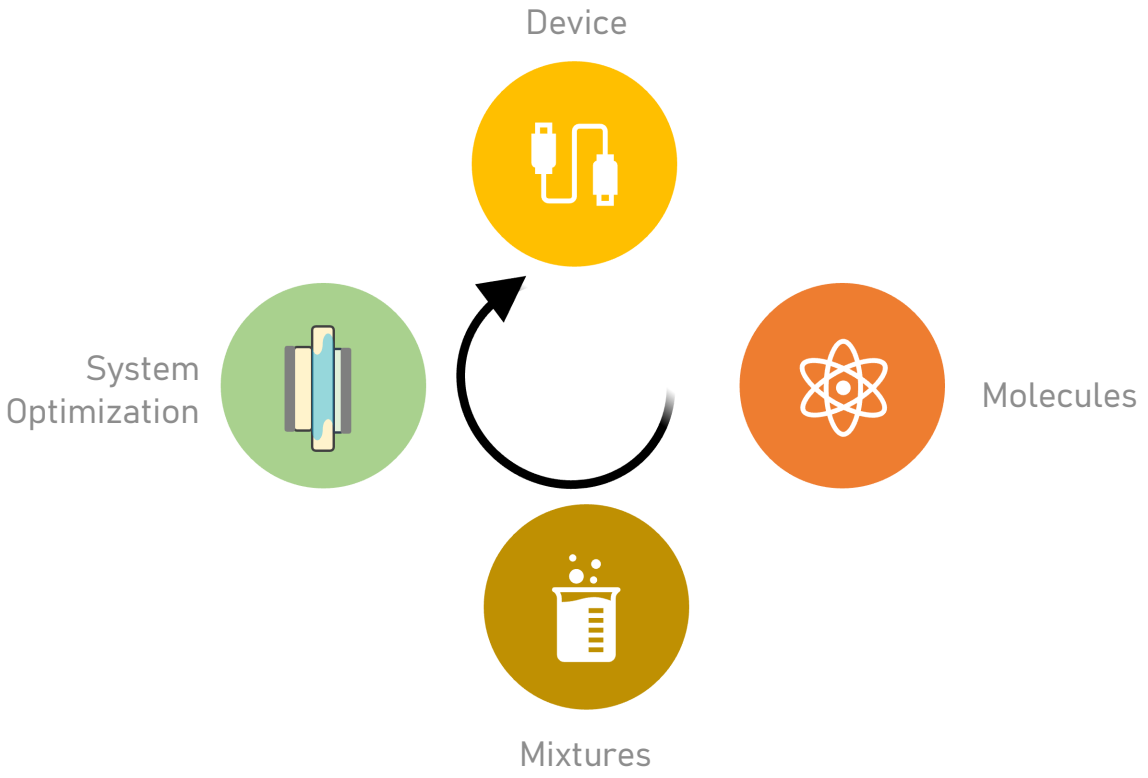
Energy Storage (Battery)

Global battery demand is expected to grow 16.7 % annually to $424B in 2030, driven by EVs and renewable energy
Urgent need for materials innovation

At IBM Research, we use AI-based workflows & models to discover sustainable battery materials
Data collection, model fine-tuning, and experimental validation

Our AI models map material structures, compositions, and performance
Prediction and optimization of complex materials (e.g., electrolytes)

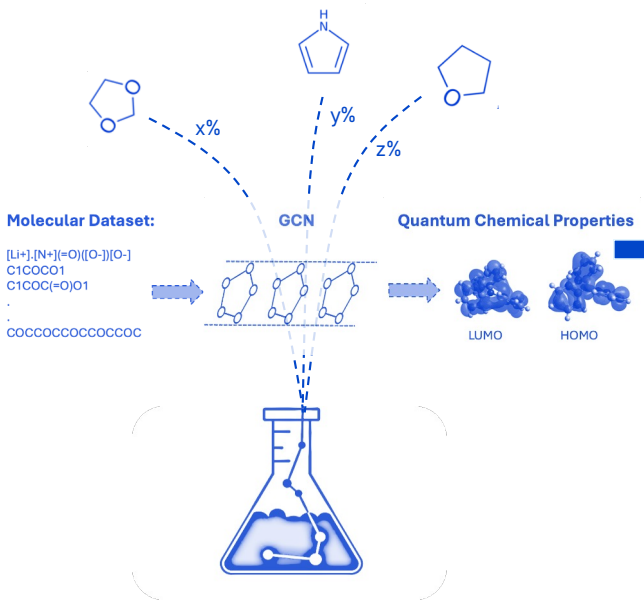
AI-Driven Battery Material Discovery and Optimization



Battery Electrolyte Discovery

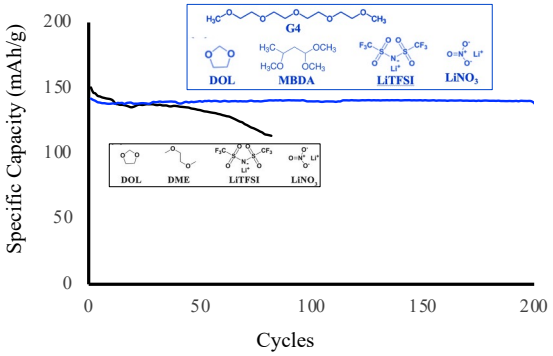
Formulation Graphs

-For Mapping Constituent Structures and Compositions to Battery Performance with experimental datasets such as Ionic Conductivity, Columbic Efficiency, Capacity



Sharma et al., Formulation Graphs for Mapping Structure-Composition of Battery Electrolytes to Device Performance, *J. Chem. Inf. Model.* 2023, 63, 22, 6998–7010

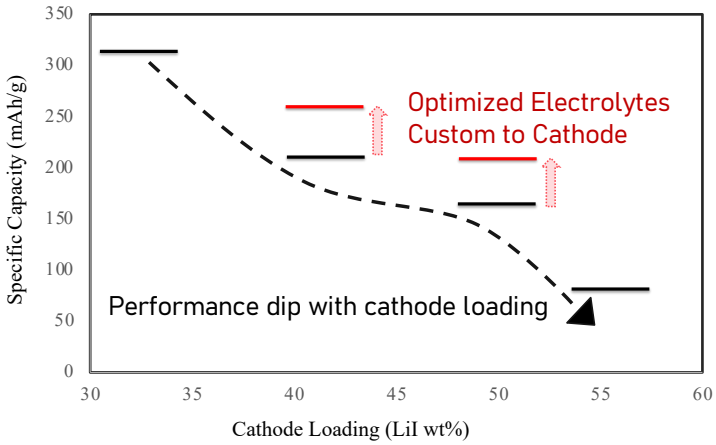
Electrolyte Design for Enhance Cycle Life



- Electrolyte formulation vs battery capacity datasets from 100 coin-cells based on Lithium-iodine conversion battery, used to screen 4000+ solvent candidates in formulation design to derive at a solvent system imparting cycling stability in 14 days.
- ~5000% speed-up as compared to manual evaluation of EPA solvent database

Elmegreen et. al, MDLab: AI frameworks for carbon capture and battery materials. *Frontiers in Environmental Science*, 2023, *11*, p.1204690.

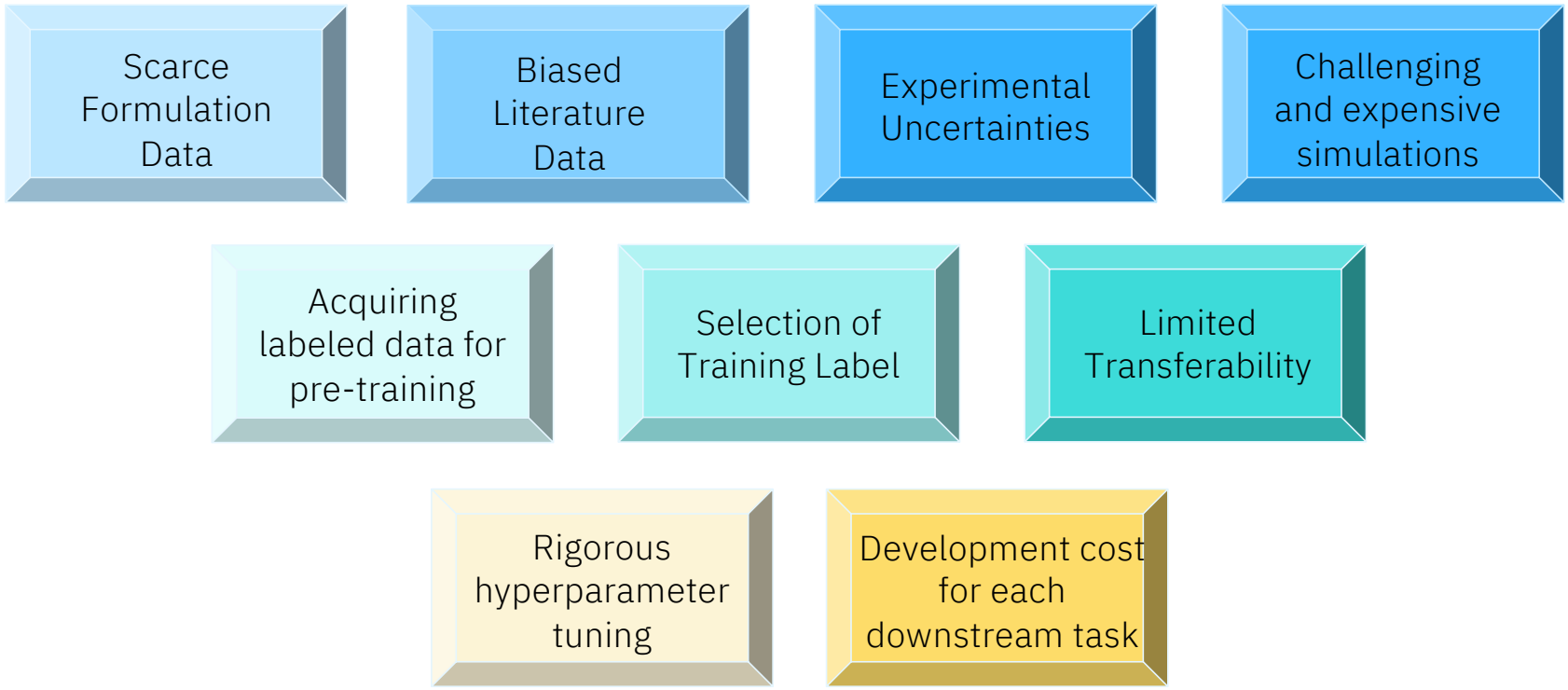
Electrolyte Design for Improving Capacity at Higher Cathode Loading



- Higher loading of active electrode materials is desired in batteries for enhanced energy density and cost efficiency.
- But, increasing active material loading in electrodes can cause significant performance depreciation due to internal resistance, shuttling, and parasitic side reactions
- Electrolyte optimization custom to target cathode loadings, based on experimental datasets of Li-ICl ($4e^-$) battery

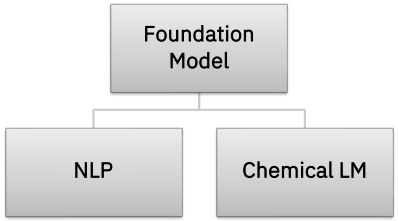
Manuscript in Review.

Existing Challenges

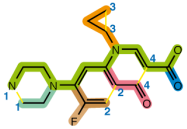


Foundation Models for Chemistry and Material Science

What are Chemical Foundation Models?

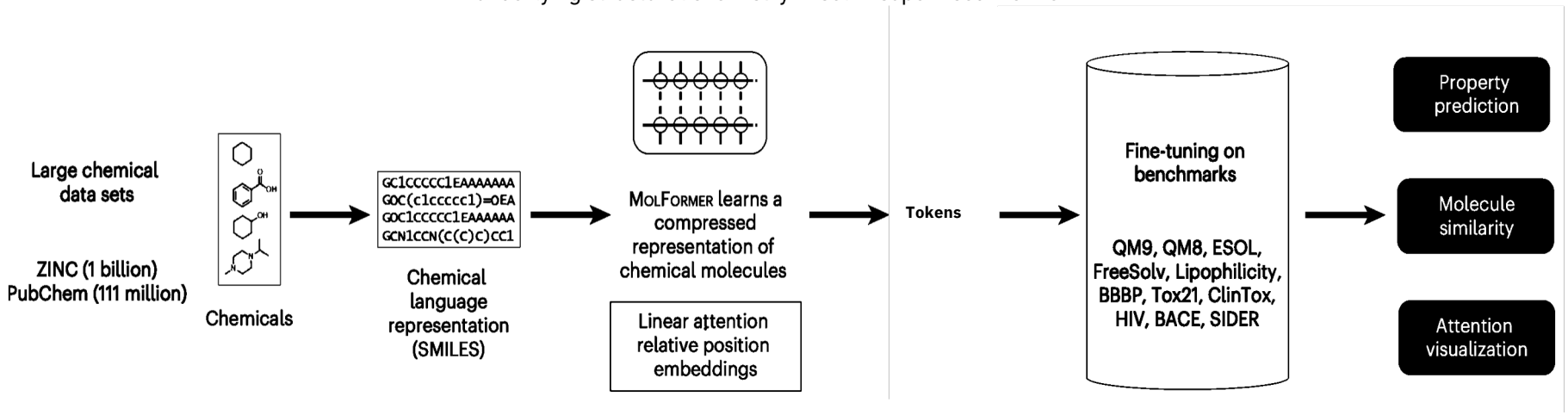


Chemical Language- Simplified molecular-input line-entry system (SMILES)



N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

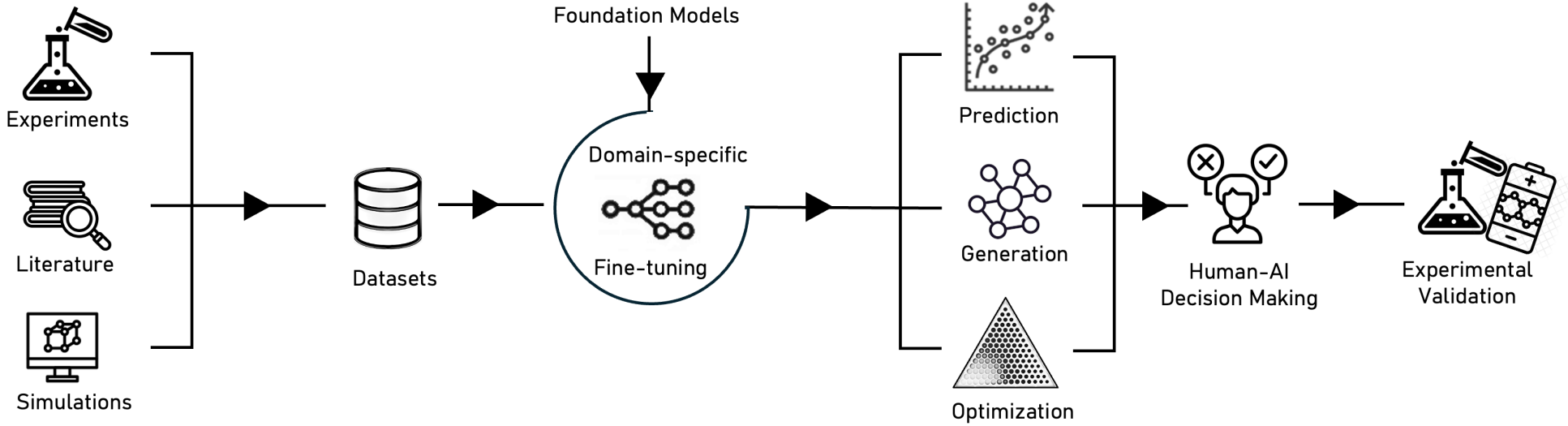
Transformer model trained with string representations of molecules such as SMILES and SELFIES, learn the underlying structural chemistry in self - supervised manner.



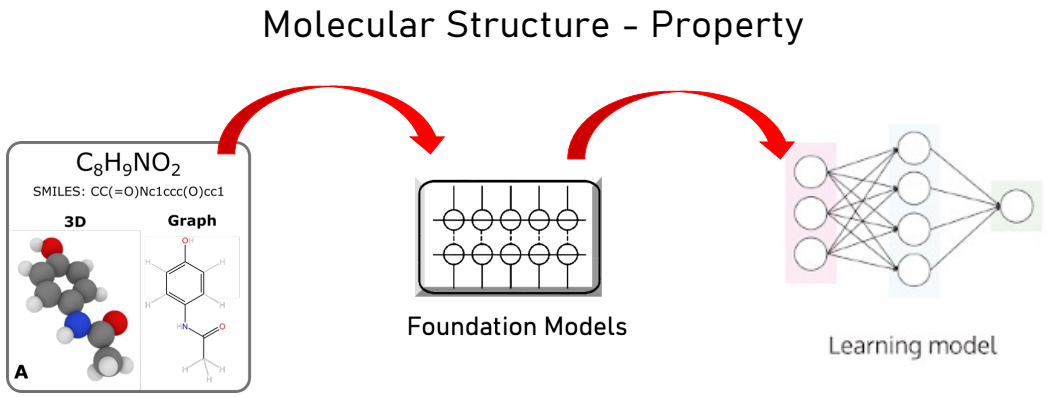
Ross. et al. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell 4, 1256–1264 (2022). https://doi.org/10.1038/s42256-022-00580-7

Fine-Tuning Foundation Models for Domain

General purpose Foundation Models for chemistry and material science are fine-tuned for domain specific tasks with labeled datasets derived from literature, simulations or experimentation.



Fine-Tuning Foundation Models for Molecular Property Prediction



Performance of fine-tuned MOLFORMER and other supervised GNN with benchmarking datasets (represented by mean squared error (MSE))

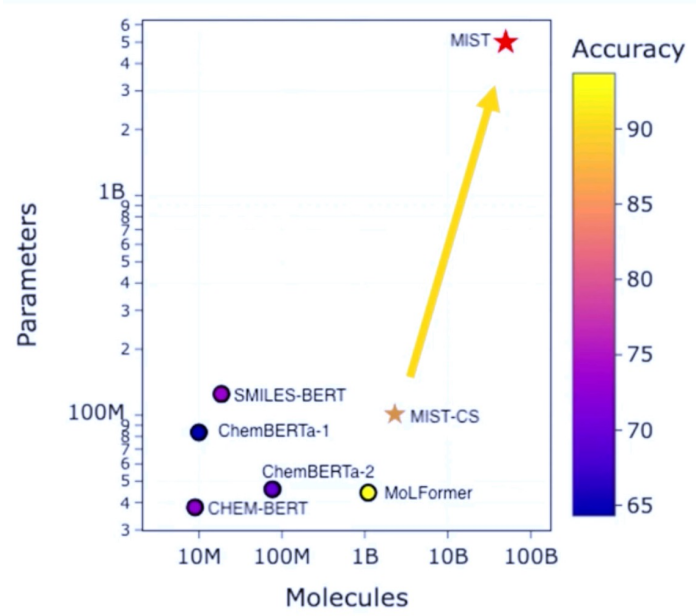
Models	QM9	QM8	ESOL	FreeSolv	Lipophilicity
GC	4.3536	0.0148	0.970	1.40	0.655
A-FP	2.6355	0.0282	0.5030	0.736	0.578
MPNN	3.1898	0.0143	0.58	1.150	0.7190
MOLFORMER-XL	1.5894	0.0102	0.2787	0.2308	0.5289

Ross. *et al.* Large-scale chemical language representations capture molecular structure and properties. *Nat Mach Intell* 4, 1256–1264 (2022). https://doi.org/10.1038/s42256-022-00580-7

Chemical Foundation Model

Scalability

Ongoing efforts in the community on scaling large foundation model with 10-40B molecular SMILES with aspirations to improve accuracies.

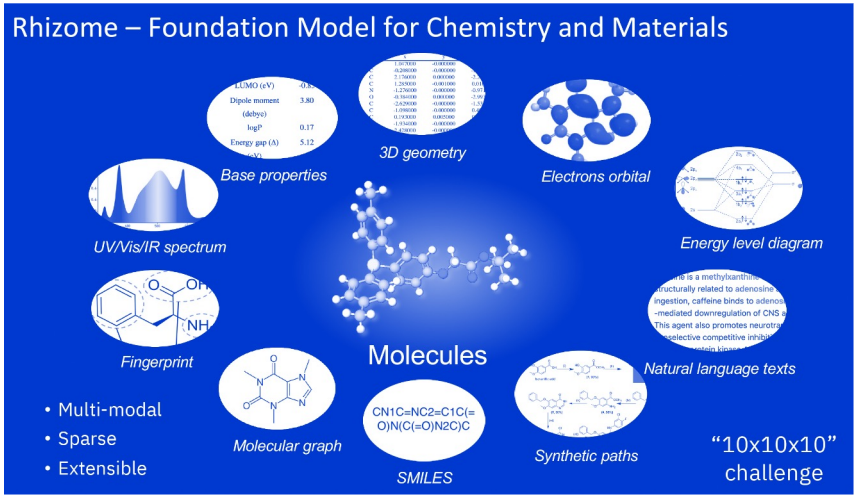


Venkat Viswanathan, Scientific Foundation Models , MICDE Conference April 2nd & 3rd, 2024

Multi-Modality

Large foundation model trained with multi-modal materials and chemistry data

- Enables multi-modal generation and property prediction for materials and chemistry tasks
- Supports inferencing and fine-tuning for domain specific downstream adaptation
- Billions of parameters x Billions samples x 10 modalities



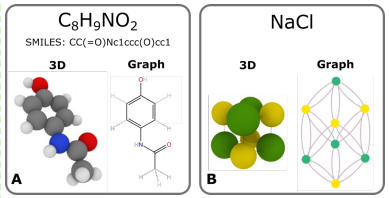
Takeda et. al, Multi-modal Foundation Model for Material Design. In *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*.

Complex Material

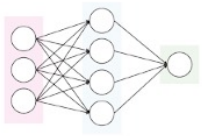
What are complex materials?

Materials

Molecules Crystals

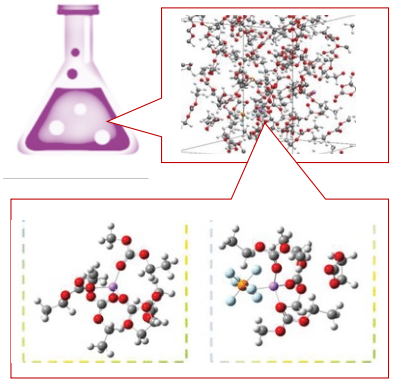


Structure - Property



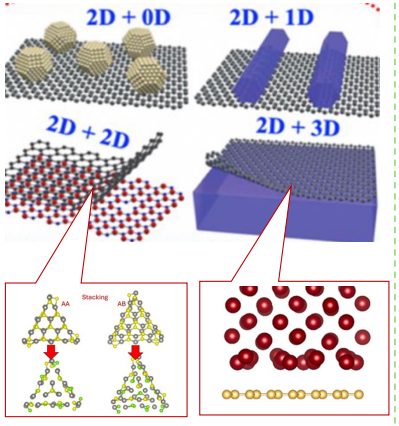
Learning model

Mixtures



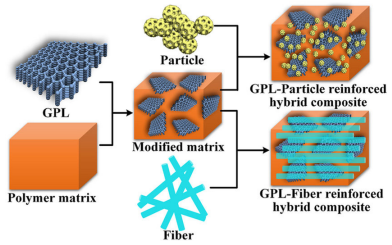
Complex dynamic interactions

Heterostructures



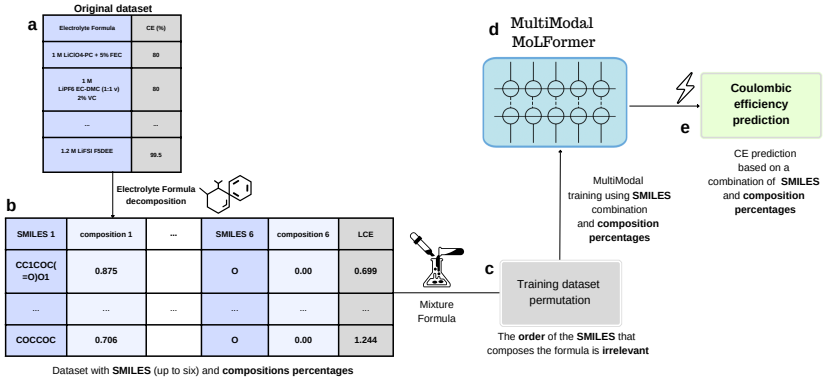
Complex Interfaces
-> New Properties

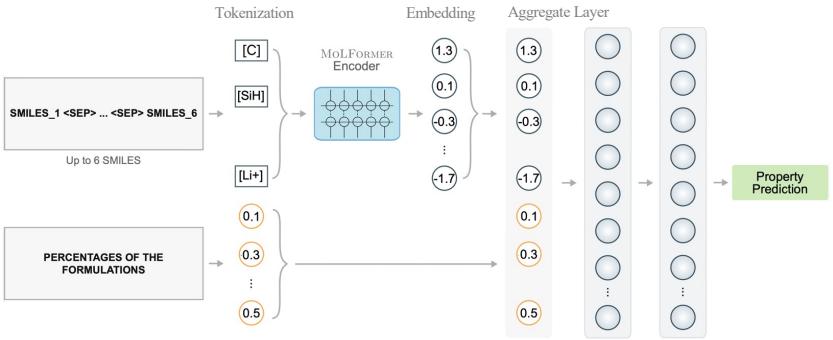
Composites



Different class of materials
-
Complex interactions
Unique Interfaces
Unique Properties
Challenging to model

Representing Liquid Mixtures: Electrolyte Formulation





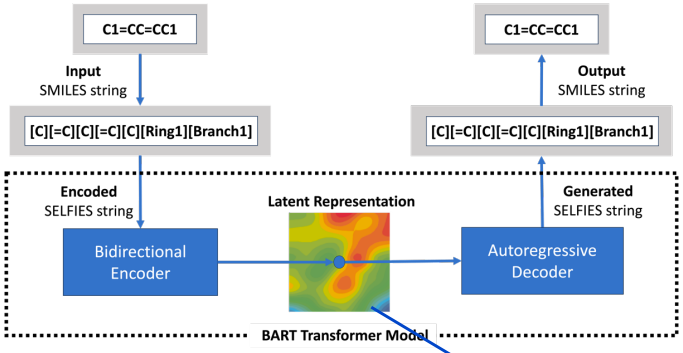
Methods	RMSE
MoLFormer (Combinatorial of SMILES + formulations)[1]	**0.195**
Formulation Graphs[2]	0.323
Machine Learning Models[3]	0.577

Predictive modeling to estimate the Log **Coulombic efficiency of electrolyte solutions.** This task involves utilizing a dataset of 152 samples, originally curated by Kim et al. in 2023[3]

[1]Soares, Eduardo, et al. "Capturing Formulation Design of Battery Electrolytes with Chemical Large Language Model." *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*. 2023.
[2]Sharma et al., Formulation Graphs for Mapping Structure-Composition of Battery Electrolytes to Device Performance, ACS JCIM, 10.1021/acs.jcim.3c01030
[3]KIM, Sang Cheol et al. Data-driven electrolyte design for lithium metal anodes. Proceedings of the National Academy of Sciences, v. 120, n. 10, p. e2214357120, 2023

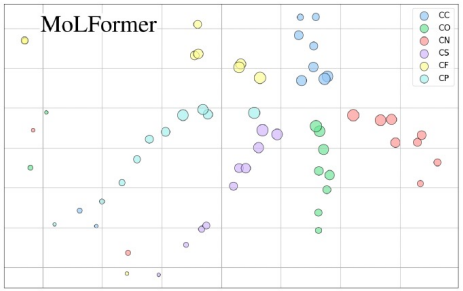
Representing Liquid Mixtures: Electrolyte Formulation

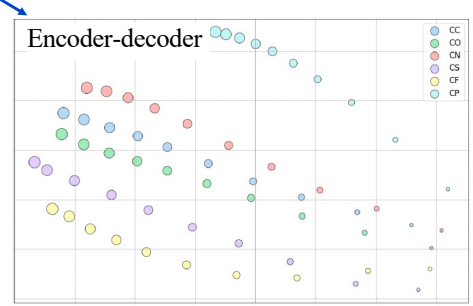
Decoding Latent Space



Encoder-decoder based foundation models recognized to outperform encoder-only MoLFormer

Models	RMSE in Log CE Prediction
Linear Regression	0.585
Random Forest	0.577
Bagging	0.583
Formulation Graphs (IBM)	0.389
SMILES MoLFormer(IBM-2023)	0.195
SELFIES BART (IBM-2024)	0.148





Experiments to investigate the structure of the latent space created by Foundation Models reveal the composability (linear mapping) of structural motifs in Encoder-Decoder architectures.

Priyadarsini et. al. 'Improving Performance Prediction of Electrolyte Formulations withTransformer-based Molecular Representation Model' in Machine Learning for Life and Material Science (ML4MS Workshop), ICML, July 26th , Austria

Summary

- Chemical foundation models can be fine-tuned for downstream tasks on non-sequencing complex mixed material systems.
- Generalizable models reduce development cost for downstream tasks and computational cost associated with acquiring labeled simulation datasets.

Open problems:

- Is scalability necessary for accuracy?
- How can the effects of pre-training FM with synthetic molecular SMILES percolate to the downstream tasks?

Acknowledgments

IBM Almaden Research Center, San Jose, CA

IBM Brazil

Emilio Vital Brazil

Renato Cerqueira

Eduardo Soares

IBM Tokyo

Indra Priyadarsini S

Energy Storage Team

(Left to Right) Murtaza Zohair, Max Giammona, Khan Nyuguen, Tim Erdmann, Young-Hye Na, Linda Sundberg, Vidushi Sharma, Andy Tek, Anthony Fong

Thank you for your attention!

Vidushi Sharma
(vidushis@ibm.com)