Foundation Models Enabling Multi-Scale Battery Materials Discovery: From Molecules To Devices

Vidushi Sharma
IBM Almaden Research
San Jose, CA, USA
vidushis@ibm.com

Andy Tek
IBM Almaden Research
San Jose, CA, USA
atek@us.ibm.com

Maxwell Giammona
IBM Almaden Research
San Jose, CA, USA
Maxwell.Giammona@ibm.com

Murtaza Zohair
IBM Research Almaden
San Jose, CA, USA
mzohair@ibm.com

Nathaniel Park
IBM Research Almaden
San Jose, CA, USA
npark@us.ibm.com

Tim Erdmann
IBM Research Almaden
San Jose, CA, USA
tim.erdmann@ibm.com

Linda Sundberg
IBM Almaden Research
San Jose, CA, USA
lindas@us.ibm.com

Eduardo Soares
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
eduardo.soares@ibm.com

Khanh Nguyen
IBM Almaden Research
San Jose, CA, USA
khanh.vinh.nguyen@ibm.com

Young-Hye Na
IBM Almaden Research
San Jose, CA, USA
yna@us.ibm.com

Emilio Ashton Vital Brazil
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
evital@br.ibm.com

Abstract

Recent years have seen fast emergence and adoption of chemical foundation models in computational material science for property prediction and generation tasks that are focused mostly on small molecules or crystals. Despite these paradigm shifts, integration of newly discovered materials in real world devices continues to be a challenge due to design problems. New candidate material must be optimized to achieve compatibility with other components in the system and deliver the target performance. Chemical foundation model benchmarks must evaluate their scope in predicting macro scale outcomes that are the result of chemical interactions in multi-variate design space. This study evaluates performance of chemical foundation models that are pre-trained primarily with SMILES of small molecules, in extrapolating learning from molecules to material design challenges across multiple length scale in batteries. Ten prediction models are trained covering molecular properties, formulations performance, and battery device measurement. Material representations from several foundation models are compared and their performance is benchmarked against conventional molecular representations such as Morgan Fingerprints. The study further examines their capacity to generalize to out-of-distribution cases by quantifying prediction errors for novel material designs that differ substantially from the training data. Finally, interpretability of the trained predictors is assessed by correlating actual outcomes and predictions to the chemical moieties in the datasets, with the aim of enabling researchers to interpret design rules in chemical space where model has high confidence.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

With evolving technologies and world economy demands, the field of material discovery has remained strongly relevant. Recently, this field has acquired critical importance as new sustainable materials are sought to overcome limitations of current material systems (1). Battery technologies are one strong societally relevant area of research where the scope of known materials appears to be exhausted, and new materials that can deliver high capacities, fast charging and longer cycle stability are continuously sought to meet future demands (2; 3). Despite shifts in material research paradigms from slow, labor-intensive experiments, to faster data-driven models (4; 1), it remains challenging to integrate new materials in real world devices. This is due to several reasons: (i) most computational models including simulations and machine learning (ML) can be used to determine intrinsic properties of materials based on their chemical structure, but lack in extrapolating their outcome to meso or macro scale phenomenon (5); (ii) device performance is governed by complex interactions among several constituent materials, presenting vast multivariate design space difficult to screen or optimize (6); (iii) limited data availability for extrinsic characteristics such as temperature and concentration dependence of multi-constituent properties (7). While ML models accelerate several prediction, generative and optimization problems in material science, the field continues to face challenges stemming from opaque nature of the model's decision making, impractical proposed chemical structures, scarcity of quality datasets and inability to generalize out-of-distribution (OOD) (8).

Foundation models (FMs) have emerged as promising models to overcome some aforementioned challenges of data scarcity and generalization. These are a class of large language models (LLMs), that are pre-trained on a textual or multi-modal representations of materials in open-source databases like PubChem and ZINC through self-supervised learning (9; 10). Studies have demonstrated that embedding space of these transformer models segregates chemically relevant features of molecules making them a suitable general-purpose tool for material science research. These base models can be utilized to perform specific functions based on smaller labeled datasets with fine-tuning or transfer learning (11). FMs are rapidly evolving, and their adoption in different application areas is on the rise (12). Large portion of studies report their use in property prediction and inverse design of small molecules or crystals (11). Prior studies also evaluate their scope in predicting performance metrics for formulations (mixtures of more than two molecules in certain compositions) based on electrolyte-performance experimental datasets curated from literature. Results demonstrate best prediction accuracies from foundation models in comparison to other data-driven models (13; 14). The research on representing advanced material systems such as formulations, composites and devices to learning models is currently in nascent stages due to less understood chemical phenomenon and lack of quality datasets. Prior studies on formulation datasets present strong evidence that foundation models can extrapolate molecular features to multi-constituent properties.

In this work, we evaluate the capability of chemical FMs pre-trained with molecular representation SMILES (15), to predict material properties and performance resulting from interplay of complex chemical phenomenon at macroscale. We take battery electrolytes as an example where electrolyte engineering has emerged as a promising approach to improve battery performance metrics such as columbic efficiency (CE), cycle life and capacity. To achieve this, electrolytes are carefully designed based on the individual properties of constituent molecules, their collective performance as formulation and their compatibility with other battery components such as electrodes, separator and current collector. Electrolyte Genome initiative in 2015 accelerated electrolyte discovery cycle for new emerging battery chemistries by integrating computational workflows with experimentation (16). High-throughput screening enabled selection of candidate molecules meeting threshold values for HOMO-LUMO energy levels, toxicity and electrochemical stability. Once down-selection is done, laborious experimentation is required to find their right combination for a functional electrolyte formulation (17). Here, data availability is a primary roadblock in adoption of ML models since public datasets are inconsistent and industrial datasets are propriety (18). Thus, models that can be efficient with scarce datasets are desired in the domain.

We use FMs to map electrolyte formulations along with device variables to key performance indicators at multiple length scale in batteries as illustrated in Figure 1. In particular,

- We target prediction of key properties that are considered in electrolyte discovery such as molecular properties, formulation performance, manufacturability, surface contact characteristics and device performance. FMs are used to generate input features for these

multivariate battery datasets and predictive capability is compared with standard molecular representations like Morgan Fingerprints (MF) (19).

- We evaluate out-of-distribution (OOD) capability of prediction models for multi-variate battery datasets.
- Next, extrapolation capability of the models to new material designs is estimated based on the semantic similarity between train and test data. This presents a method to approximate errors in model predictions across new material landscape.
- We investigate interpretability of FM-based predictors and evaluate their promise in inferencing new material design rules.

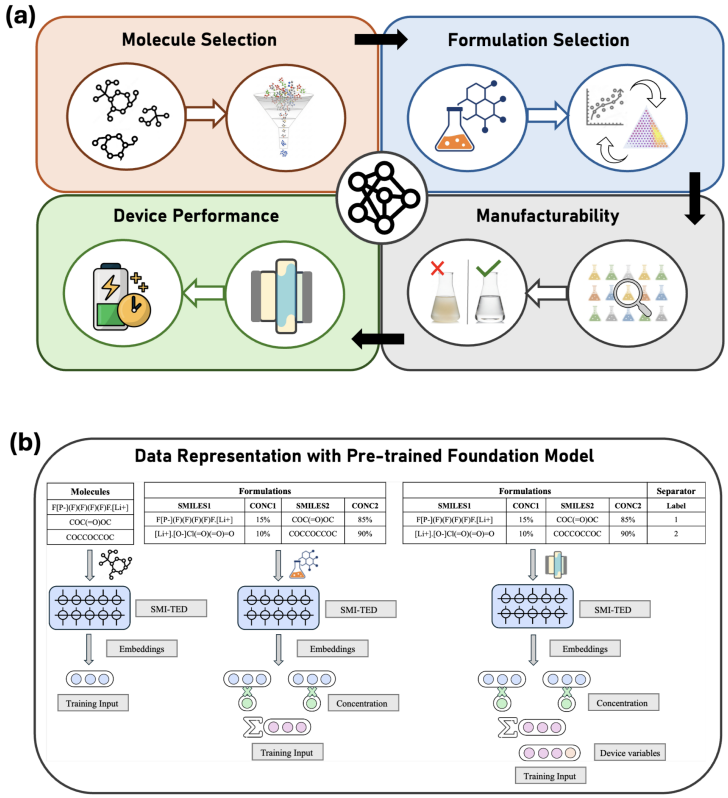

Figure 1: (a) Scheme illustrating electrolyte design problems at multiple scales. (b) Schematic summarizing the data representation for material design using pre-trained foundation models for molecules.

2 Datasets and Foundation Models

Data availability is a major enabler for artificial intelligence (AI) workflows aiming for material discovery and design. To discover new material design that meets the performance goals, series of data driven predictors must be realized to allow material identification, characterization and optimization for achieving compatibility with the device. For present study, we utilize several battery datasets and performance indicators that are used across multiple length scale for electrolyte development. Most datasets are curated from literature and some are experimentally generated in the laboratory. Dataset details are summarized in Supplementary Materials section A.1, while present section differentiates FMs evaluated.

There is a plethora of pre-trained transformer models in literature that are used for specific downstream scientific tasks (20; 10; 21; 22; 23). Particularly in the domain of chemistry and material science, sequence prediction, molecular property prediction and chemical description generation are a few tasks that are used in benchmarking FM. In this work, we aim to evaluate scope of FMs pre-trained

on molecular representations in addressing material design challenges across multiple length scale in batteries. Comparative analyses were performed across multiple FM to elucidate the extent to which model performance and generalization behaviors are influenced by differences in pretraining modalities.

SMI-TED: SMI-TED (SMILES Transformer Encoder Decoder) is an open-source chemical FM developed by IBM Research (10). This model has acquired a deep understanding of molecular structural representations through self-supervised pre-training on a vast dataset containing string representation (SMILES) of 91 million molecules, corresponding to 4 billion molecular tokens. Model has been previously validated to surpass the performance of conventional data-driven alternatives in downstream tasks.

MolT5: MolT5 (Molecular T5) is another open sourced chemical FM that is pre-trained with 100 million SMILES along with 33,000 natural language description of molecules (23). By correlating SMILES sequences to textual description of functionalities, the model has shown remarkable capabilities in manipulating molecules for discovery tasks.

Galactica: Galactica is a large language model developed for general scientific tasks by Meta AI (22). The model is trained on large corpus of scientific literature, natural sequences of proteins and 2 million chemical strings (SMILES). The inclusion of broad data makes is a reliable model for general scientific tasks such as equation probing, citation prediction, reasoning, etc.

GraphMVP: GraphMVP is a graphs based pre-trained model that formulates a multi-view self-supervised learning, integrating both 2D molecular graphs and rich 3D spatial arrangements of atoms (24). The GraphMVP learning framework allows its encoder to integrate topological and geometric information within a unified embedding space. It is worth noting that GraphMVP uses much smaller graph/conformer datasets in representation learning.

Morgan Fingerprints: As a benchmark, MF are employed as an established molecular descriptor (19). MF are highly effective for predicting molecular properties in ML models because they efficiently capture the substructural features of a molecule (25). By representing a molecule as a fixed-length binary vector, they encode the presence or absence of specific circular substructures and each atom's chemical environments. The resulting numerical representation is both computationally efficient and chemically intuitive, making it an ideal input for various learning algorithms, which can then identify complex patterns and relationships that are predictive of a molecule's behavior.

For downstream tasks, transfer learning approach is adopted to retain chemical information from the pre-trained model as molecular embeddings, and map these to the output label using a regressor model such as feed forward neural networks (NN). It is noted that fine-tuning the pre-trained FM containing several million parameters with labeled datasets can be computationally expensive. Furthermore, fine-tuning current state-of-the-art FM is not expandable to the string representations of formulations used in ref(14) as these are vastly different from the molecule representations models were pre-trained on. Meanwhile, transfer learning approach is relatively robust and deliver consistently reliable results (see Table S1). Therefore, embeddings from the FMs and MF are used to represent individual molecules in the battery datasets. Derived molecular embeddings are aggregated into a system representation based on their composition, and additional design variables in the dataset such as separator, temperature and cathode loading (indicated in Figure 1b). Details of feature engineering for appropriate representation of molecules, formulations and devices are described in A.4. For each prediction task, feed forward neural network (NN) architectures are optimized and trained using FM-derived and aggregated features (described in A.5). NNs were trained using five independent 80%-20% train–test splits, and prediction errors were quantified using the mean absolute error (MAE) metric.

3 Results and Discussion

3.1 Model performance

We use FMs that recognize SMILES modality for training electrolyte design predictors due to ease of chemical data representation and their demonstrated best performance in predicting molecular properties in several benchmark datasets (10). Prediction results for 10 battery datasets are summarized in Table 1 for FMs and MF. Tabulated are the average MAE across 5 random train-test splits for all models. Results show that SMI-TED and MolT5 based representations outperform MF in 7 out of

10 datasets. Meanwhile predictive capability of Galactica and GraphMVP is observed to be the lowest in all 10 datasets. Particularly for molecular properties, where several prior studies have backed that 2048 bits of MF are more predictive than domain-intuitive features (25), results in Table1 indicate SMI-TED outperforms MF. SMI-TED demonstrates notable computational efficiency despite using significantly smaller feature vector size (768). This efficacy of SMI-TED embeddings testifies that learnt representations encode more comprehensive set of structural features that are meaningful and comprehensive.

In the context of more complex systems, such as formulations, we observed a systematic divergence in model performance based on data size. On datasets characterized by a large volume of data, such as solubility (3300 data) and IC (18,000 data), MF outperform all FM in the present evaluation, categorizing miscible and immiscible electrolytes with 93.77% accuracy, and predicting log IC with MAE 0.0629, surpassing previously best reported results in ref (14). This outcome is consistent with the design of conventional ML methods that are optimized for large-scale data problems. MF's enhanced performance on these datasets suggests that the fundamental properties like IC and solubility are more contingent on specific functional groups in the system that are captured precisely by MF. This finding presents a critical consideration for the future development of foundation models.

SMI-TED and MolT5 demonstrated clear and consistent advantage over MF in low data regimes (100 to 200 data points), achieving superior predictive accuracy and robustness across these challenging multiscale problems. Particularly MolT5, having pre-trained on largest corpus of molecular data (100 Million SMILES), has the lowest prediction errors for contact angle (MAE 12.944 Degrees) and LiI capacity (MAE 22.408 mAh/g) datasets, and is second to MF for solubility (93.65% Accuracy) and IC (log IC MAE 0.0722) prediction. SMI-TED demonstrates next best predictive capability among FMs, reporting low prediction errors for all formulation datasets and outperforming all models for CE dataset (6; 3). These results highlight applicability of FM pretrained with molecules alone to multi-variate material design problems. Possible interpretation is that macroscale outcomes, such as electrolyte performance, are dictated by hierarchical interactions between chemical moieties. Ion aggregates and solvation substructures are examples of chemical moiety interactions responsible for charge-discharge kinetics in battery electrolytes. Models such as MolT5 and SMI-TED successfully predicts these macroscale outcomes due to having rich chemical vocabulary comprising of thousands of unique chemical tokens or moieties as reported in ref(10). Hence, latent space of SMILES-based FM is enriched with basic understanding of the chemical space formed by the combinations of chemical moieties in molecules (10). The downstream training utilizing aggregated formulation embeddings vs performance label is useful to correlate chemical moieties and compositions to the label, enabling multi-scale learning (see Figure 2). This knowledge transfer is particularly useful in low data regimes. Li-ICl Capacity data is a singular instance where MF outperforms FMs despite low data regime, highlighting FMs are likely not suitable for datasets lacking chemical variability.

Results from MolT5 present additional interesting observations on multi-modal pre-training. Latent space of MolT5 is augmented with semantic understanding of molecular string representation, correlating molecule structures to specific functions (23). In Table 1, advantages of pretraining with multi-modal datasets is noted in multi-variate battery datasets but not in molecular datasets. Despite pre-training on largest SMILES corpus, predictive capability of MolT5 model is lower than SMI-TED for molecular properties, likely due to noted functional biases and scarcity of natural language datasets used during model development(23). Regardless, good predictive performance on multi-variate datasets underscore the critical importance of incorporating multi-modal data representations during the pretraining, enabling model to learn complex inter-dependencies and semantic nuances across datasets.

Poor performance of Galactica in predicting material properties underline limitations of high generality. Despite training on large corpus of scientific knowledge and 2 Million SMILES, model lacks sufficient specificity required to capture critical domain-relevant features. In lieu, GraphMVP also shows poor predictive power despite high specialization in molecular geometries. The model captures the 3-D topological and geometric features of molecules but lacks sufficient representational capacity to resolve finer substructural moieties and their inter-dependencies. Ultimately, the choice of representation is critical and must be determined by the nature of downstream task, quantity and the quality of the labeled dataset.

Table 1: Average mean absolute error (MAE) and prediction accuracy (%) for the battery datasets using embeddings from foundation models

Model ⇓	**Oxidation**	**Reduction**	**HOMO**	**LUMO**	**Solubility**	**IC**	**Contact Angle**	**LiI Capacity**	**CE**	**Li-ICl Capacity**
MAE Units →	eV	eV	eV	eV	Accuracy %	Log	$Degrees$	mAh/g	Log	mAh/g
SMI-TED	**0.2559**	**0.5825**	**0.4405**	**0.3663**	93.11	0.0910	16.243	22.449	**0.185**	47.93
MolT5	0.2679	1.7375	0.4451	0.3836	**93.65**	**0.0722**	**12.944**	**22.408**	0.188	37.57
Galactica	0.2714	0.7134	0.4802	0.4283	93.05	0.1035	23.982	25.011	0.225	39.570
GraphMVP	0.3355	0.6586	0.4987	0.4432	91.17	0.0939	22.099	29.051	0.209	42.451
MF	0.2594	0.5854	0.4580	0.3746	**93.77**	**0.0629**	17.815	28.990	0.223	**32.24**

3.2 Quantifying out-of-distribution performance

Formulations present multi-variate design space with infinite possibilities emerging from several million known compounds, their inestimable potential combinations, and composition variations. Given this, electrolyte design discovery becomes inherently an OOD problem as novel formulations will most likely be in unseen or unfamiliar data. Thus, evaluating OOD performance is crucial for ensuring the reliability and robustness of models. One can define OOD based on divergence between train-test sets with respect to either input distribution (chemical and composition space) or output distribution (property values). Presented OOD evaluation of FMs for formulation and device performance datasets spans both input and output distributions.

First, we start with most accepted OOD evaluation based on output distribution (26). We separate test sets based on tail ends of numerical outcome distribution, for instance, lower and upper end values of ionic conductivity, capacity, contact angle, etc. Tail-end distributions used as tests in 5 electrolyte regression datasets are highlighted in A.6. This distribution estimates extrapolation capabilities of the models beyond the training data. Results of OOD predictions are presented in Table 2 along with prediction uncertainty observed across 3 predictions. Both SMI-TED and MolT5 demonstrate best OOD prediction with each having lowest MAE in 2 out of 5 datasets. Both models also had high consistency in predicted outcomes as indicated by low uncertainty. Overall extrapolation across outcome values is promising for electrolyte datasets except for Li-ICl Capacity dataset where models perform poorly as seen in previous section.

Table 2: Mean absolute error (MAE) for out-of-distribution predictions using foundation models and Morgan Fingerprints

Model ⇓	**CE**	**Contact Angle**	**LiI Capacity**	**IC**	**Li-ICl Capacity**
MAE Units →	Log	$Degrees$	mAh/g	Log	mAh/g
SMI-TED	0.0548 ±0.04	13.5216 ±0.41	27.128 ±0.70	0.1938 ±0.01	109.21 ±0.95
MolT5	0.0819 ±0.00	14.0539 ±0.98	31.2229 ±1.61	0.1669 ±0.01	108.2197 ±0.93
Galactica	0.4635 ±0.39	31.4742 ±0.82	28.2692 ±14.38	0.2262 ±0.08	110.391 ±1.50
GraphMVP	2.7758 ±2.36	34.8031 ±1.88	7.9974 ±4.17	0.7429 ±0.04	108.6611 ±0.03
MF	0.1295 ±0.05	19.3304 ±1.26	29.5058 ±2.22	0.1717 ±0.03	114.3028 ±31.07

Next, ML models frequently show poor transferability across chemical spaces and fall short in predicting properties for materials outside their training scope (27). Generalizable base models like FM have seen increased adoption in the community for these reasons (27). Unlike small molecules, where property can be traced to substructures and chemical motifs (10), cause-effect in formulations-like materials are more complex and intertwined in multi-variate dynamic inter-dependencies (14). Therefore, the boundaries of OOD for dynamic multi-variate chemical space is needed to be explored in a focused study. In present study, we use chemical similarity as a metric for characterizing OOD

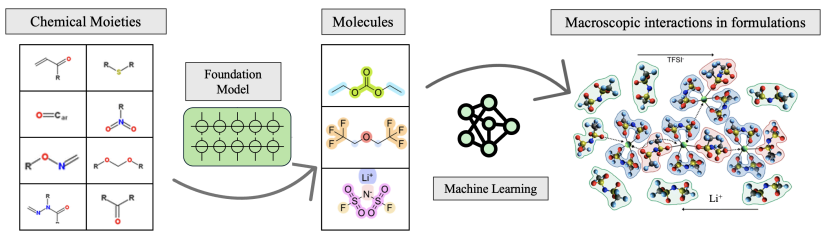

Figure 2: Multi-step training capturing complex chemical interactions at multiple length scale.

based on inputs. A chemical similarity score is employed as an approximation for how close test data is to training data in model's latent space, and is estimated by calculating maximum of average cosine similarity (normalized) of each test datapoint with all training samples. Upon evaluating the chemical similarity between embeddings of train-test sets for tail-end OOD evaluation in Table S3, we observe there is an inverse trend between chemical similarity of OOD train-test sets and prediction MAE from the models, suggesting model prediction errors are high for chemically disparate test sets. These results confirm chemical similarity can be a reliable metric to determine distance between test and train sets in model's latent space and characterize OOD.

This trend paves the way to ascertain reliability of a model when extrapolating to unexplored regions of the materials design space. By error estimation, we can systematically pinpoint regions where model lacks predictive capability, facilitating intelligent allocation of resources toward targeted experimental validation and data enrichment. We create several subsets of train-test data for battery across different length scale based on their relative distance in latent space of SMI-TED, given its reliable performance in both molecules and macroscale outcomes. These subsets were carefully curated to represent a different testing scenario than the ones used in the tail-end OOD evaluation such as distinct constituent count and chemicals. Relationship between semantic similarity between the input embeddings of train-test distributions (in red) across datasets is compared with prediction MAE for the respective train-test subset (in blue) in Figure 3. Trends confirm an inverse relationship between prediction MAE and semantic proximity of test data to the training samples, yielding a linear relationship $MAE = m.Similarity + c$ that estimates the approximate MAE of model predictions on new data points by quantifying their $Similarity$ to the model's training data. The slope (m) and intercept (c) for analyzed datasets are presented in Table S4. This approach enables systematic assessment of prediction uncertainty and confidence for new data, thereby supporting efficient screening in materials design and discovery.

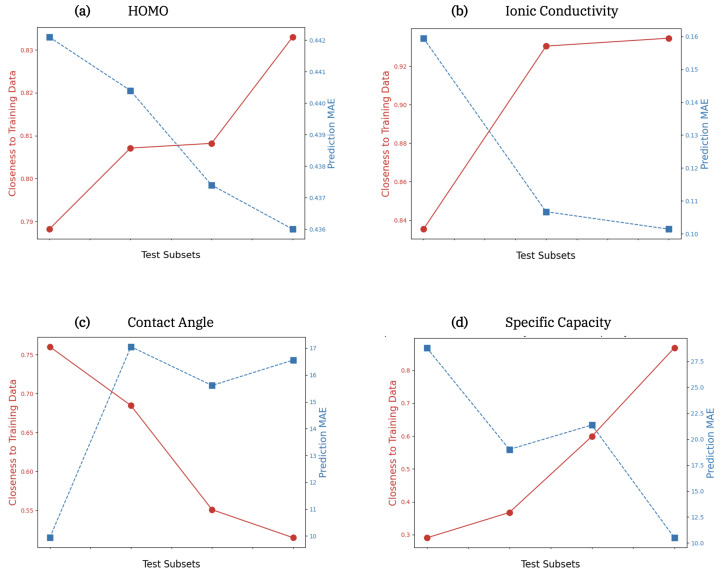

Figure 3: Relationship between prediction MAE (in blue) and chemical similarity (in red) between train and test datasets.

3.3 Interpretability

A widely embraced strategy in materials discovery involves interpreting chemical data into useful knowledge and chemical insights, uncovering conclusive design rules and trends for decision making (28; 29). The efficacy of this approach is maximized when it leverages accurate empirical data or highly reliable model-generated outputs spanning the intended design landscape. However, interpretability is frequently hindered by the intrinsic opacity of AI models, which predominantly operate as "black boxes" with internal mechanisms that remain inaccessible to researchers. This challenge is further exacerbated as training pipeline grow in complexity, for instance, input features are derived from transformer model and post processed before the training (18). Quantifying model

uncertainty in new material regions can facilitate users in identifying scope of the model. However, application of these models to uncover material design rules for interpretability remains a persistent challenge.

We propose a method to evaluate interpretability of FM derived predictors by investigating correlation of performance outcomes with chemical moieties in the datasets and compare trends in train and test subsets. First, a list of several potential chemical substructures and their SMARTS (SMILES Arbitrary Target Specification) string is devised (30). Over 550 chemical substructures are defined including general and specific moieties. For instance, amine is a general functional group of material containing Nitrogen atom with lone pair of electrons, and specific derivatives for the same include aromatic amine, heterocyclic amine, tertiary amine etc. Chemical moieties in molecules are identified by matching SMARTS and presence of every moiety is indicated by a bit in a fixed length vector. This vector is taken as molecular fingerprints and aggregated for constituents in each formulation by composition scaling and addition to represent concentration of each chemical moiety in a formulation. We adopt Spearman's correlation coefficient (SCC) (31) to determine strength and direction of monotonic relationship between chemical moieties in the dataset and the outcome performance. The analysis provides meaningful insights towards the positive or negative influence of a chemical moeity in the formulation towards the outcome. Analysis is performed for data used in training and test set to correlate moieties to actual outcomes. Simultaneously, the analysis is also extended to the outcomes predicted by the models based on SMI-TED representation for the very same test set. Figure 4 illustrates these correlations in three formulation datasets CE, LiI capacity and IC.

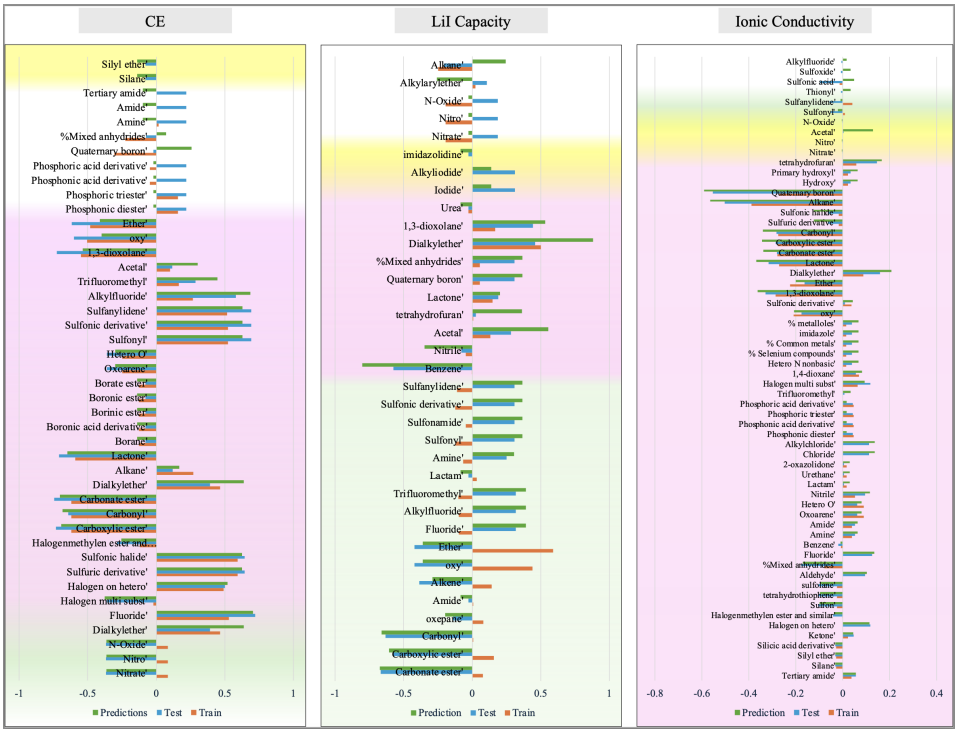

Figure 4: Correlation of chemical functional groups in formulations with performance in train (orange) - test (blue) dataset, compared with correlation to the predicted outcomes (green) in test data.

Comparison of correlation analysis for model prediction outcomes and actual performance within test sets is meant to demonstrate the capability of model in deriving sound chemical insights across unseen datapoints. Particularly in Figure 4, examples highlighted in green illustrate cases where the correlations in the training and test datasets were opposite, and the model correctly predicted the opposing trends. Instances highlighted in yellow represent scenarios where the model accurately identified chemical trends for the outcome, despite these trends being absent from the training data. Cases highlighted in pink show perfect alignment among all three correlations. The remaining instances in white indicate correlations that the foundation model misinterpreted. This analysis

reveals the chemical insights misunderstood by the model and allows users to selectively apply these models for design interpretation and discovery within a chemical space where confidence is justified.

4 Conclusion

In this work, we evaluate the scope of foundation models in addressing material design challenges across multiple length scale in batteries: molecules, formulations and device. Multiple foundation models are used to derive multi-variate representations of datasets by combining molecular representations with other variables such as compositions, temperature, electrode and separator variations. Results show FMs pre-trained with large corpus of SMILES modality, such as SMI-TED and MolT5, can be used to extrapolate learning from moiety-level interactions to macroscopic outcomes like specific capacity, surface characteristics, and battery performance using scarce datasets. These models are particularly useful in low data regimes where conventional molecular representations such as Morgan Fingerprints are found to be limiting. It is also observed that pre-training on multi-modal data representations has the scope to achieve superior performance in multi-variate material design space. The study also presents a method to analyze model's ability to generalize out-of-distribution and quantify model prediction errors across new material designs based on chemical similarity between train-test sets. SMILES-based models demonstrated reliable out-of-distribution performance trends. However, it is noted that out-of-distribution criterion for dynamic multi-variate chemical space needs further comprehensive investigation. Lastly, we demonstrate an approach to identify chemical space where model confidence is high by correlating actual outcomes and predicted outcomes to the chemical moieties in the datasets. The approach allows dependable material design interpretation from the model for discovery tasks.

References

[1] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, "Scaling deep learning for materials discovery," *Nature*, vol. 624, no. 7990, pp. 80–85, 2023.

[2] J. Datta, A. Nadimpally, N. Koratkar, and D. Datta, "Generative ai for discovering porous oxide materials for next-generation energy storage," *Cell Reports Physical Science*, 2025.

[3] S. C. Kim, S. T. Oyakhire, C. Athanitis, J. Wang, Z. Zhang, W. Zhang, D. T. Boyle, M. S. Kim, Z. Yu, X. Gao *et al.*, "Data-driven electrolyte design for lithium metal anodes," *Proceedings of the National Academy of Sciences*, vol. 120, no. 10, p. e2214357120, 2023.

[4] E. O. Pyzer-Knapp, J. W. Pitera, P. W. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith, and A. Curioni, "Accelerating materials discovery using artificial intelligence, high performance computing and robotics," *npj Computational Materials*, vol. 8, no. 1, p. 84, 2022.

[5] J. Li, K. Lim, H. Yang, Z. Ren, S. Raghavan, P.-Y. Chen, T. Buonassisi, and X. Wang, "Ai applications through the whole life cycle of material discovery," *Matter*, vol. 3, no. 2, pp. 393–432, 2020.

[6] V. Sharma, M. Giammona, D. Zubarev, A. Tek, K. Nugyuen, L. Sundberg, D. Congiu, and Y.-H. La, "Formulation graphs for mapping structure-composition of battery electrolytes to device performance," *Journal of Chemical Information and Modeling*, vol. 63, no. 22, pp. 6998–7010, 2023, pMID: 37948621. [Online]. Available: https://doi.org/10.1021/acs.jcim.3c01030

[7] P. de Blasio, J. Elsborg, T. Vegge, E. Flores, and A. Bhowmik, "Calisol-23: Experimental electrolyte conductivity data for various li-salts and solvent combinations," *Scientific Data*, vol. 11, no. 1, p. 750, 2024.

[8] A. K. Cheetham and R. Seshadri, "Artificial intelligence driving materials discovery? perspective on the article: Scaling deep learning for materials discovery," *Chemistry of Materials*, vol. 36, no. 8, pp. 3490–3495, 2024.

[9] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, "Large-scale chemical language representations capture molecular structure and properties," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1256–1264, 2022.

[10] E. Soares, E. Vital Brazil, V. Shirasuna, D. Zubarev, R. Cerqueira, and K. Schmidt, "An open-source family of large encoder-decoder foundation models for chemistry," *Communications Chemistry*, vol. 8, no. 1, p. 193, 2025.

[11] J. Choi, G. Nam, J. Choi, and Y. Jung, "A perspective on foundation models in chemistry," *JACS Au*, vol. 5, no. 4, pp. 1499–1518, 2025.

[12] E. O. Pyzer-Knapp, M. Manica, P. Staar, L. Morin, P. Ruch, T. Laino, J. R. Smith, and A. Curioni, "Foundation models for materials discovery–current state and future directions," *Npj Computational Materials*, vol. 11, no. 1, p. 61, 2025.

[13] I. Priyadarsini, V. Sharma, S. Takeda, A. Kishimoto, L. Hamada, and H. Shinohara, "Improving performance prediction of electrolyte formulations with transformer-based molecular representation model," in *ICML'24 Workshop ML for Life and Material Science: From Theory to Industry Applications*.

[14] M. Zohair, V. Sharma, E. A. Soares, K. Nguyen, M. Giammona, L. Sundberg, A. Tek, E. A. Vital, and Y.-H. La, "Chemical foundation model-guided design of high ionic conductivity electrolyte formulations," *npj Computational Materials*, vol. 11, no. 1, p. 283, 2025.

[15] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, pp. 31–36, 1988.

[16] L. Cheng, R. S. Assary, X. Qu, A. Jain, S. P. Ong, N. N. Rajput, K. Persson, and L. A. Curtiss, "Accelerating electrolyte discovery for energy storage with high-throughput screening," *The journal of physical chemistry letters*, vol. 6, no. 2, pp. 283–291, 2015.

[17] A. Benayad, D. Diddens, A. Heuer, A. N. Krishnamoorthy, M. Maiti, F. L. Cras, M. Legallais, F. Rahmanian, Y. Shin, H. Stein *et al.*, "High-throughput experimentation and computational freeway lanes for accelerated battery electrolyte and interface development research," *Advanced Energy Materials*, vol. 12, no. 17, p. 2102678, 2022.

[18] V. Sharma, A. Tek, K. Nguyen, M. Giammona, M. Zohair, L. Sundberg, and Y.-H. La, "Improving electrolyte performance for target cathode loading using an interpretable data-driven approach," *Cell Reports Physical Science*, vol. 6, no. 1, 2025.

[19] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[20] J. Pan, "Large language model for molecular chemistry," *Nature Computational Science*, vol. 3, no. 1, pp. 5–5, 2023.

[21] J. Ross, B. Belgodere, S. C. Hoffman, V. Chenthamarakshan, J. Navratil, Y. Mroueh, and P. Das, "Gp-molformer: A foundation model for molecular generation," *Digital Discovery*, 2025.

[22] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," *arXiv preprint arXiv:2211.09085*, 2022.

[23] C. Edwards, T. Lai, K. Ros, G. Honke, K. Cho, and H. Ji, "Translation between molecules and natural language," *arXiv preprint arXiv:2204.11817*, 2022.

[24] S. Liu, H. Wang, W. Liu, J. Lasenby, H. Guo, and J. Tang, "Pre-training molecular graph representation with 3d geometry," *arXiv preprint arXiv:2110.07728*, 2021.

[25] H. Zhou and J. Skolnick, "Utility of the morgan fingerprint in structure-based virtual ligand screening," *The Journal of Physical Chemistry B*, vol. 128, no. 22, pp. 5363–5370, 2024.

[26] E. R. Antoniuk, S. Zaman, T. Ben-Nun, P. Li, J. Diffenderfer, B. Demirci, O. Smolenski, T. Hsu, A. M. Hiszpanski, K. Chiu *et al.*, "Boom: Benchmarking out-of-distribution molecular property predictions of machine learning models," *arXiv preprint arXiv:2505.01912*, 2025.

[27] M. A. Skinnider, R. G. Stacey, D. S. Wishart, and L. J. Foster, "Chemical language models enable navigation in sparsely populated chemical space," *Nature Machine Intelligence*, vol. 3, no. 9, pp. 759–770, 2021.

[28] H. Choubisa, P. Todorović, J. M. Pina, D. H. Parmar, Z. Li, O. Voznyy, I. Tamblyn, and E. H. Sargent, "Interpretable discovery of semiconductors with machine learning," *NPJ Computational Materials*, vol. 9, no. 1, p. 117, 2023.

[29] J. Dean, M. Scheffler, T. A. Purcell, S. V. Barabash, R. Bhowmik, and T. Bazhirov, "Interpretable machine learning for materials design," *Journal of Materials Research*, vol. 38, no. 20, pp. 4477–4496, 2023.

[30] X. Liu, S. Swaminathan, D. Zubarev, B. Ransom, N. Park, K. Schmidt, and H. Zhao, "Accfg: Accurate functional group extraction and molecular structure comparison," *Journal of Chemical Information and Modeling*, 2025.

[31] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation," *Anesthesia & analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[32] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," *Scientific data*, vol. 1, no. 1, pp. 1–7, 2014.

[33] R. Duke, V. Bhat, P. Sornberger, S. A. Odom, and C. Risko, "Towards a comprehensive data infrastructure for redox-active organic molecules targeting non-aqueous redox flow batteries," *Digital Discovery*, vol. 2, no. 4, pp. 1152–1162, 2023.

[34] I. Priyadarsini, S. Takeda, L. Hamada, E. V. Brazil, E. Soares, and H. Shinohara, "Self-bart: A transformer-based molecular representation model using selfies," *arXiv preprint arXiv:2410.12348*, 2024.

[35] H. Zhang, T. Lai, J. Chen, A. Manthiram, J. M. Rondinelli, and W. Chen, "Learning molecular mixture property using chemistry-aware graph neural network," *PRX Energy*, vol. 3, no. 2, p. 023006, 2024.

A Supplementary Material

A.1 Electrolyte Datasets

Molecule screening: Battery electrolytes can comprise of one or more organic solvent, and one or more salt, which facilitate Li+ ion transport between electrodes and electrode surface conditioning to prevent unwanted degrading side reactions. Each electrolyte component plays a crucial role in this ecosystem and is therefore selectively picked based on certain properties like HOMO-LUMO levels and redox potentials. While there is plethora of labeled dataset available in literature for these properties (32; 33; 16), we use a data from a singular source to train and evaluate model's performance, i.e., D3TaLES, a database of DFT simulated properties of 40,000 organic molecules for battery systems (33).

Manufacturability: Screened solvents and salts are combined in certain compositions to form electrolyte formulations. These formulations must be completely miscible (or soluble) to enable ion transport and manufacturing. We curate a heterogeneous dataset containing solubility information of single salt-single solvent mixtures, single salt-multi solvent formulations, and multi salt- multi solvent electrolytes, enabling development of a generalized model for electrolyte miscibility prediction. Refer to A.2 for details on electrolyte solubility data generation. For inclusion of heterogeneous datasets, we simplify approach to binary classification indicating insoluble (0) or soluble (1). The combined 3,300 dataset contained rich diversity of salts, solvents and electrolyte mixtures.

Formulation property: Another crucial property to consider during electrolyte design is ionic conductivity (IC). The salts dissociated into ions within an electrolyte form solvation structures that facilitate transport of charge between two electrodes and are responsible for battery's charge-discharge kinetics. For IC, we use 18,000 reported empirical values of electrolyte formulations at different temperatures in published literature (7; 14). The dataset constitutes diverse set of solvents and salts.

Surface contact characterization: An electrolyte interfaces with multiple internal components within a battery, including electrodes, separators, and current collectors. Consequently, optimizing the surface interactions between the electrolyte formulation and various device constituents is crucial for achieving peak performance. Traditionally, such evaluations have relied on the empirical expertise of domain experts and expensive computational simulations. Nevertheless, data collected from evaluation of one similar system can be used to automate future screening and assessment of electrolytes. We use one such in-house generated empirical dataset of 119 electrolyte formulations and their contact angle on four different separators to predict surface contact angle of electrolytes (see A.3 for experimental details).

Device performance: The ultimate objective of developing a new battery electrolyte formulation is to achieve superior performance metrics, such as enhanced capacity, Coulombic Efficiency (CE), and cycle life. The public dissemination of such data is often limited, as its relevance is typically highly specific to a particular device configuration, thereby precluding its full adherence to FAIR (Findable, Accessible, Interoperable, and Reusable) data principles. To address this challenge, we leverage three distinct datasets from previous publications. The first dataset, derived from a study by Kim et al. (3), examines the relationship between electrolyte composition and CE across 150 datapoints. A second dataset containing 125 electrolytes, originally reported by Sharma et al. (6), explores the influence of electrolyte formulation on the specific capacity of a LiI conversion battery. Finally, the third dataset constituting 91 datapoints focuses on capacity metric for an interhalogen conversion (Li-ICl) battery, incorporating variations in cathode loading, separator type, and electrolyte compositions with fixed chemicals (18).

A.2 Solubility Data Collection

Complete electrolyte miscibility is desired in batteries for manufacturing to ensure that the electrolyte composition is consistent batch to batch and devoid of any phase separation for uniformity in battery performance at production scale. Therefore, it is essential to identify potentially miscible formulations from the vast combinatorial design space. Heterogeneous solubility dataset is generated through experimentation:

Single salt- single solvent solubility assessment: A dataset of binary system containing single salt and a single organic solvent was collected experimentally in the laboratory. The dataset spans five most popular electrolyte salts, LiNO3, LiFSI, LiBOB, LiFOB, and LiPF6, and up to fifty organic solvents. The experiments were conducted in an inert glovebox (Argon, < 0.1 ppm H2O and O2) and all salts were dried on a hotplate at 150 °C, except for LiFSI and LiPF6 , which were used as received due to their lower thermal stability. Solvents were dried over 3Å molecular sieves for at least 24 hours prior to use. An upper salt concentration limit of 2M was set during the data collection. Salts were weighed to make 2M solution and the respective organic solvent was then added to decrease the concentration by a 0.25M interval until the solutions were visually clear without any precipitation or undissolved materials. The salt-solvent combination was considered insoluble if the solution was not clear at 0.25M concentration.

Single salt- Multi solvent solubility assessment: The dataset has measurement of the highest molar concentration of single salt dissolved in mixture of organic solvents. The data was curated during the development of electrolyte for our prior study where four salts and four solvents were shortlisted for lithium metal battery electrolyte (18). The four salts, LiCl, LiNO3, LiTFSI and LiBOB, are individually dissolved in solvent formulations containing different compositions of ethylene carbonate, Tetraglyme, 1,3-Dimethyl-2-imidazolidinone and 1,3-Dioxolane. The solubility measurements were made as per the method described above.

Multi salt-multi solvent solubility assessment: Conventionally, functioning and high-performing electrolytes are published in literature (3; 18; 6). We also share a few "failed" non-miscible electrolytes in our previous works (18; 14). We curated 300 electrolyte formulations from these studies. Simplification of solubility metric to (0) or (1) enabled inclusion and test across widespread electrolyte dataset. The combined dataset contained rich diversity of salts, solvents and electrolyte mixtures.

Post processing: The solubility of single salt- single solvent pairs and single salt- multi solvent formulations were measured in terms of highest soluble molarity of the salt. To further add context to the solute molarity noted as metric in empirical dataset, data augmentation was done to interpolate solubility of target salt in each respective solvent system to include soluble(1) datapoints below highest soluble molarity, and insoluble(0) datapoints above recorded metric until the tested molarity. Next, the constituent moles in each formulation system were converted to molar percentage (mole%). Post data processing, there are 3300 electrolyte formulation vs solubility data that is used in the study.

A.3 Contact Angle Measurement Experiments

Electrolyte uptake by separator is an important parameter that determines ion transport and electrolyte performance. There are several separators in the commercial market based on constitution such as polymer and quartz. Within a single category like polymer separators, vast variations can be noted based in changes in polymer monomers and ratios. Electrolyte formulations are prepared inside an Ar-filled glove box (<1 ppm O2, <1 ppm H2O). Prior to mixing, solvents that are liquid at room temperature are dried using molecular sieves (Millipore Sigma, 3) and salts are dried on a hot plate at 100 °C. Electrolytes are mixed for 24 hrs prior to contact angle measurement. Contact angle measurements were conducted using an OCA video-based contact angle goniometer (FDS Future Digital Scientific Corporation) employing the sessile drop technique. Prior to measurement, the separator was carefully placed on a flat silicon wafer substrate to ensure a uniform surface. A 2L droplet of electrolyte was then dispensed onto the separator surface and allowed to equilibrate for 800ms. Image analysis was performed on a selected video frame by manually defining the baseline and applying an ellipse-fitting algorithm to achieve optimal conformity to the droplet profile. The reported static contact angles represent the average of 3–5 independent measurements. All procedures were carried out with minimal air exposure to preserve the integrity of the electrolyte and ensure reproducibility. A dataset of 119 experiments is created using the electrolyte constituents, their respective concentrations, the experimentally measured contact angle, and a separator label. There are four different Celgard separators in the dataset, identified by unique label (1-3).

A.4 Feature engineering

The application of data-driven models in material systems rely on the correct transformation of system into a numerical representation suitable for mathematical operations. Accordingly, the intricate description of a battery's formulation, which includes the identity of constituent molecules, their composition, and additional configuration parameters, must be systematically converted into a relevant numerical descriptor. For this purpose, pretrained FMs are used to acquire molecular representations which are then transformed to represent multi-scale systems as described below:

Molecules: FMs are used to derive numerical embeddings of molecules present in the target datasets similar to previous studies (10; 34).

Formulations: Three formulation datasets including solubility, CE and LiI battery capacity map electrolyte formulations to the outcome. Formulation inputs constitute multiple constituents per datapoint and their respective composition as mole percent ($mol\%$) in the mixture. Here, constituent molecules are transformed to FM embeddings, and are then scaled based on their $mol\%$ in the formulation to indicate their activity within the system. The scaled embeddings are aggregated to form a formulation descriptor by addition as also summarized in Figure 1. There are more than one method to aggregate formulation descriptor (18; 35; 13). Each method has its own merit and preferred use. We observe that scaled addition is most convenient aggregation as the resultant formulation descriptor size is invariant to the formulation constituent count. IC dataset contains temperature as an additional extrinsic variable that is concatenated with the formulation descriptor for training.

Surface contact characterization: In present study, contact angle of electrolyte on several polymer-based separators are measured to assess their compatibility. For best representation, a FM for polymer can be

used. However, since present study is focused on assessing molecular FM, separator representation has been simplified by the use of labels. There are four polymer separators in the dataset labeled 0-3. These labels are concatenated with formulation representation analogous to temperature in IC dataset.

Device: Li-ICl battery dataset reports specific capacity of the battery with varying compositions of 8 electrolyte constituents for a range of active material loadings (30% to 60%) in cathode and varying separators (18). Electrolyte formulations are aggregated as defined for formulations and additional cell variables are concatenated to formulation descriptor as model inputs.

For each dataset, neural network (NN) architectures are individually optimized and trained using the derived dataset inputs. This feature engineering for representing molecules, formulations and devices was consistent across all FMs and MF.

A.5 Model Training

It is noted that fine-tuning FMs such as SMI-TED with string representation of formulations could result in relatively higher mean squared error (MSE) than the transfer learning approach where formulation descriptor aggregates pre-learned molecular embeddings scaled with the composition. MSE for both the approaches are compared in Table S1 for IC dataset where finetuning achieves MSE 0.155 and transfer learning combined by NN regressor achieved MSE 0.025.

Table S1: Mean squared error (MSE) for property prediction using SMI-TED

Dataset	MSE	
	Fine-tuning	**Transfer learning**
Reduction Potential	0.65	0.68
Oxidation Potential	0.13	0.14
Ionic Conductivity	0.155	0.025

Hyperparameter Tuning: Neural network (NN) architectures were individually optimized and trained using FM–derived molecular embeddings or formulation descriptor. NN with 2 or 3 hidden layers, with nodes 500-250-100 or 500-250, and activation function relu was found optimum. Model was trained with learning rate 0.0001, factoring 0.5 every 200 epochs of no reduction in loss function. The model was trained for maximum of 2500 epochs or until 200 iterations of no improvement in validation loss. Batch size was varied based on data size. For datasets < 200, batch size was kept 1, batch size was 12 for dataset <5000, and batch size of 32 was used for data >5000. Regression loss was measured using mean squared error (MSE) and mean absolute error (MAE) was the used metric. For binary classification of electrolyte solubility, binary cross entropy was the loss function and accuracy was the metric.

Table S2: Tuning neural network hyperparameters for SMI-TED predictors

Dataset	Hidden layers	Activation Function	MAE
LCE	500-250-100	relu	0.17
LCE	500-250	relu	0.16
LCE	500-250	sigmoid	0.32
LCE	500-250-100	sigmoid	0.32
LCE	500-250-250	relu	0.16
LCE	500-500	relu	0.17
LCE	250-100	relu	0.16
IC	500-250-100	relu	0.08
IC	500-250-100	sigmoid	0.22
IC	500-250	relu	0.09
IC	500-500	relu	0.10
IC	250-250-250	relu	0.08
IC	700-700	relu	0.11
IC	500-250-100-50	relu	0.08
HOMO	500-250-100	relu	0.43
HOMO	500-250-100	sigmoid	0.44
HOMO	500-250	relu	0.44
HOMO	250-100	relu	0.44
HOMO	500-500-500	relu	0.44
HOMO	250-250-250	relu	0.44

A.6 Out-of-distribution (OOD) evaluation

Two-fold OOD evaluation is done: (1) tail end evaluation based on numerical distribution of outcome labels, and (2) chemical design evaluation based on chemical similarity between train-test sets. For tail-end evaluation, test set are created from the training data to include lower and upper end values. In certain cases such as in Figure S3

and Figure S4, only one end of data was considered as the outcome label was highly biased towards the other end.

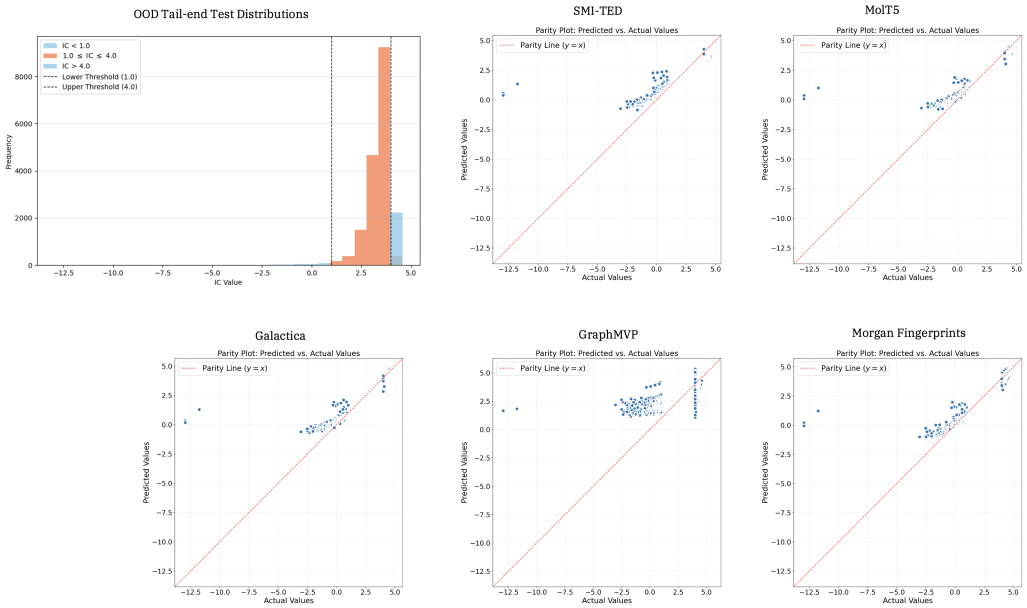

Figure S1: Tail-end OOD and parity plots for ionic conductivity test sets using benchmarking models.

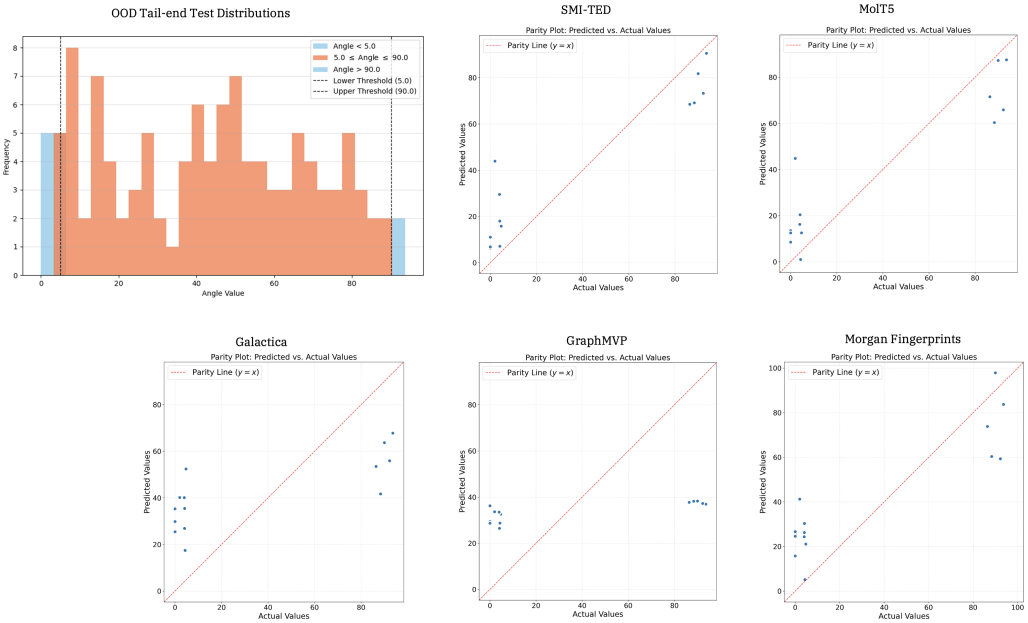

Figure S2: Tail-end OOD and parity plots for contact angle test sets using benchmarking models.

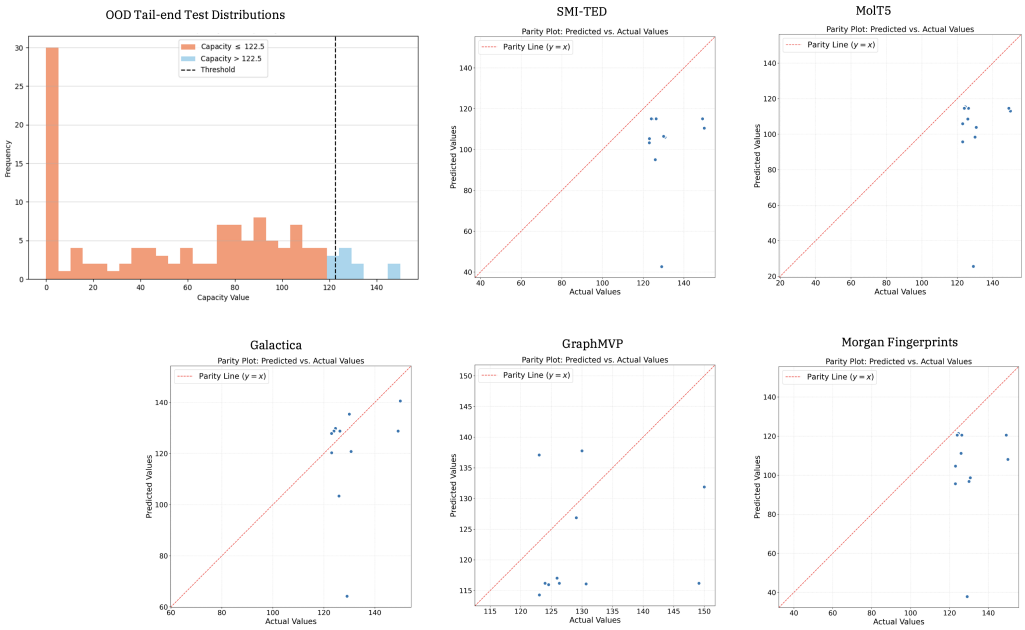

Figure S3: Tail-end OOD and parity plots for LiI capacity test sets using benchmarking models.

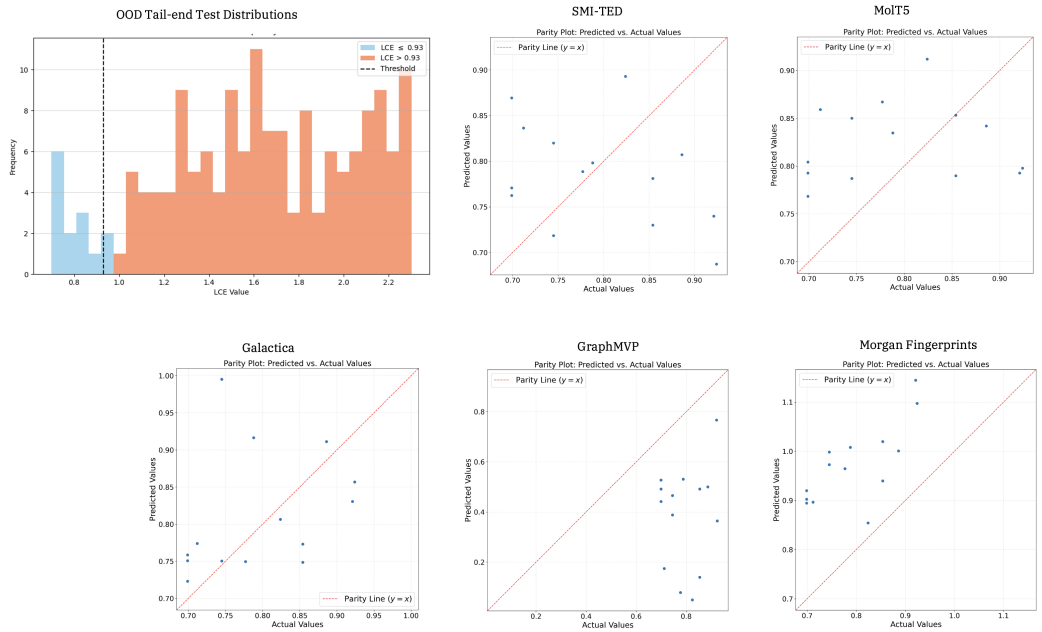

Figure S4: Tail-end OOD and parity plots for LCE test sets using benchmarking models.

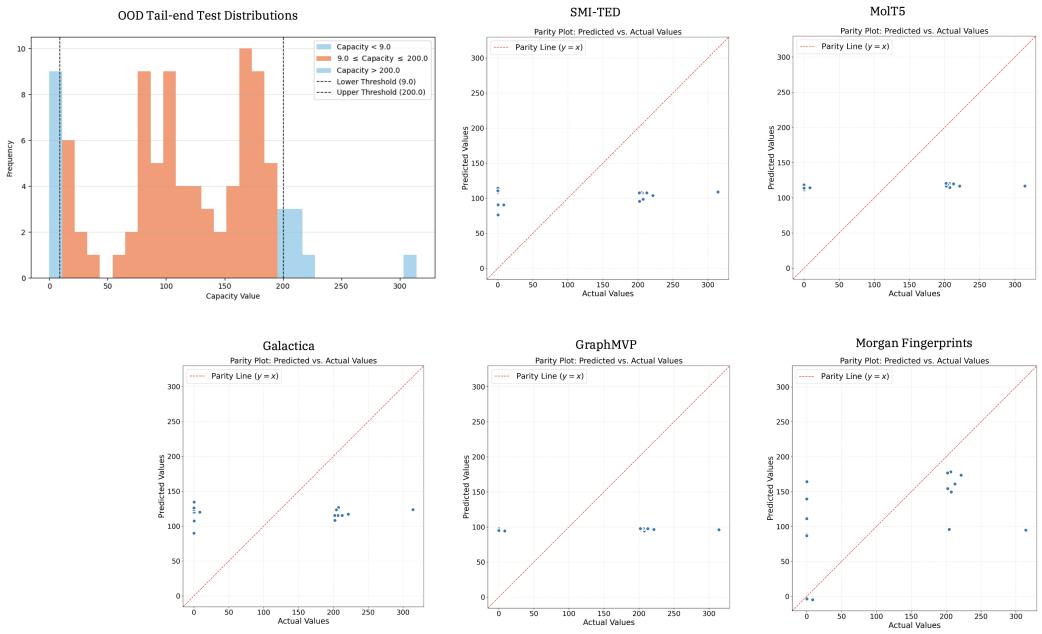

Figure S5: Tail-end OOD and parity plots for Li-ICl Capacity test sets using benchmarking models.

Table S3: Chemical similarity of out-of-distribution test datasets with training data using embeddings from foundation models and Morgan Fingerprints

Model	CE	Contact Angle	LiI Capacity	IC	Li-ICl Capacity
SMI-TED	**0.3324**	**0.6791**	0.2557	**0.9244**	0.6021
MolT5	0.2592	0.5472	0.1868	0.8209	0.641
Galactica	0.1925	0.6556	**0.4531**	0.9178	**0.681**
GraphMVP	0.0514	0.1099	0.0619	0.1814	0.0206
MF	0.2198	0.3281	0.1144	0.751	0.4748

Table S4: Parameters to estimate mean absolute error (MAE) in model prediction based on similarity between test-train data for SMI-TED

Datasets	**Slope**(m)	**Intercept**(c)
HOMO	-0.1602	0.5699
Ionic Conductivity	-0.5724	0.6377
Contact Angle	-19.6820	0.7601
Specific Capacity	-24.9776	33.2050