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

Anonymous Author(s)
Affiliation
Address
email

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 to attain the target performance. Chemical foundation model benchmarks must evaluate their scope in predicting macro scale outcomes that are the result of chemical interactions in multivariate design space. This study evaluates performance of chemical foundation model, pre-trained with 91 million SMILES of small molecules, in extrapolating learning from molecules to material design challenges across multiple length scale in batteries. The base model is fine-tuned using ten datasets covering molecular structures, formulations, and battery device measurements, and its performance is benchmarked against conventional molecular representations such as Morgan Fingerprints. The study further examines the model's capacity to generalize to out-of-distribution (OOD) cases by quantifying prediction errors for novel material designs that differ substantially from the training data. Finally, interpretability of the resulting models is assessed, with the aim of enabling researchers to apply them selectively for design interpretation within regions of chemical space where prediction confidence can be reasonably established.

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 continously 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

Submitted to AI for Science workshop (NeurIPS 2025).

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 for materials (FM4M) 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). 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 (10). Foundation models (FM) are rapidly evolving, and their adoption in different application areas is on the rise (11). Large portion of studies report their use in property prediction and inverse design of small molecules or crystals (10). 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 in comparison to other data-driven models (12; 13). 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. These results 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 a chemical foundation model pre-trained with molecular representations SMILES (14), to predict properties and performances of materials that are the result of 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 (15). 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 (16). Here, data availability is a primary roadblock in adoption of ML models since public datasets are inconsistent and industrial datasets are propriety (17). Thus, models that can be efficient with scarce datasets are desired in the domain.

We use FM4M 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. The results are compared with standard molecular representations like Morgan Fingerprints (MF) (18).
- We evaluate extrapolation capability of the trained models to new material designs based on the semantic similarity between train and test data. This presents a method to approximate errors and confidence in model predictions across new material landscape.
- We investigate interpretability of FM4M-based predictors and evaluate their promise in inferencing new material design rules.

2 Datasets

Data availability is a major enabler for artificial intelligence (AI) workflows aiming for material discovery and design. While 'material discovery' targets generating new candidates with specialized properties, 'material design' leans towards customization and optimization of candidates for compatibility with system or device to achieve target performance. Therefore, to meet the performance goals for respective application, series of data driven predictors must be realized to enable material identification, characterization and optimization for achieving compatibility with the device. Several datasets used in present study are curated from literature, while some are experimentally generated in the laboratory (see section Supplementary Materials for details).

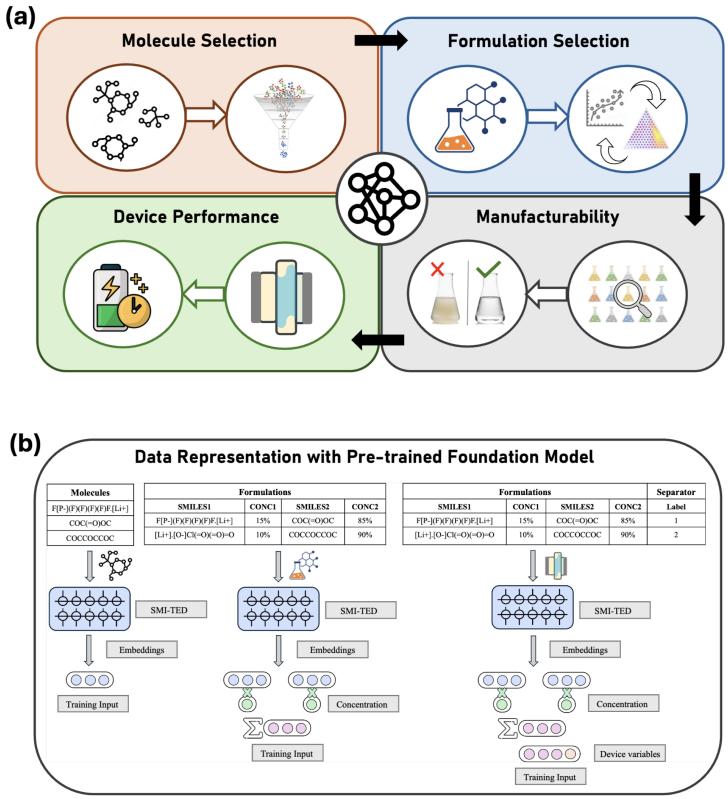

Figure 1: (a) Scheme illustrating electrolyte design problems at multiple scales. (b) Schematic summarizing the data representation for material design using pre-trained foundation models for molecules.

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, redox potentials and solvation energy. While there is plethora of labeled dataset available in literature for these properties (19; 20; 15), there are inconsistencies between these datasets due to differences in the calculation methods. To avoid these inconsistencies, 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 (20).

Manufacturability: Shortlisted solvents and salts are combined in certain compositions to form electrolyte formulations. These formulations must be completely soluble to enable ion transport and manufacturing. 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. Yet, prediction of miscibility during electrolyte discovery faces technical challenges due to limited knowledge on physical properties and phase behavior of non-aqueous solutions. While aqueous solubility remains widely reported, literature on solubility of non-aqueous electrolytes remains lacking on two fronts (i) empirical observations, and (ii) multi-constituent mixtures. Here we use 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. Solubility metric considered across the literature have been numerous and inconsistent (21). To prevent these limitations and simplify data ingestion in prediction model, we do not target any specific parameter but a binary classification of soluble (1) or insoluble (0). This simplification of solubility metric to (0) or (1) enabled inclusion of widespread electrolyte datasets. The combined 3300 datapoints contained rich diversity of salts, solvents and electrolyte mixtures.

Formulation Property: Another crucial property to consider during electrolyte design is ionic conductivity (IC). Salts within an electrolyte dissociate into anions and cations. These dissociated ions form solvation structures to facilitate transport of charge ions 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; 13). 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 these 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. Data generated from these studies, however, is often specific to a particular system and lacks the generalizability of fundamental properties like solubility and IC. Nevertheless, data collected from evaluation of one similar system can be used to develop ML model to automate future screening and assessment of electrolytes. We use one such in-house generated empirical dataset of electrolyte formulation and their contact angle on four different separators to assess surface wettability of electrolytes. A dataset of 119 experiments is constructed using the electrolyte constituents, their respective concentrations, the experimentally measured contact angle, and a separator label. Four different Celgard separators were included in the dataset.

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 our 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 125 datapoints focuses on capacity metric for an interhalogen conversion (Li-ICl) battery, incorporating variations in cathode loading, separator type, and electrolyte composition (17).

3 Data Representation

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, we utilize SMI-TED (SMILES Transformer Encoder Decoder), an open-source chemical foundation model developed by IBM Research (9). This model has acquired a deep understanding of molecular structural representations through self-supervised pre-training on a vast dataset of 91 million molecules and has been previously validated to surpass the performance of conventional data-driven alternatives in downstream tasks.

Molecules: SMI-TED encoder is used to derive numerical embeddings of molecules present in the target datasets similar to previous studies (9; 22).

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 SMI-TED 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 summarized in Figure 1. There are more than one method to aggregate formulation descriptor (17; 23; 12). 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: 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. For best material representation, a foundation model (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 (17). 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 SMI-TED–derived molecular embeddings or formulation descriptor (see section Supplementary Materials for details). As a benchmark, Morgan fingerprints(MF) were employed as an established molecular descriptor (18).

4 Results and Discussion

4.1 Material representation and model performance

We use a SMILES-based foundation model for training electrolyte design predictors due to their demonstrated best performance against several benchmark models (9). SMI-TED takes string representation of material as an input. Fine-tuning the pre-trained SMI-TED encoder with labeled datasets can be computationally expensive considering FM are relatively large models with over several million parameters. The most efficient approach is 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 NN, XGBoost or random forest. This transfer learning approach is relatively robust and deliver comparative results in predicting molecular properties, such as reduction potential and oxidation potential, as indicated in Table 2. Moreover, fine-tuning SMI-TED is not expandable to the datasets targeting formulations as the string representations of formulations used in ref(13) are vastly different from the molecule representations SMI-TED was pre-trained on. Consequently, it is noted that fine-tuning 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 2 for IC dataset where finetuning achieves MSE 0.155 and transfer learning combined by NN regressor achieved MSE 0.025. Thus, transfer learning approach was used to train all datasets.

Results are summarized in Table 3 for SMI-TED embeddings and Table 4 for MF. As tabulated in the respective tables, SMI-TED based predictors outperform MF in 7 out of 10 datasets. For molecular properties, SMI-TED is marginally better than MF. Several prior studies have backed that 2048 bits of MF are more predictive than domain-intuitive features for molecular properties (24). Meanwhile, SMI-TED demonstrates notable computational efficiency by achieving lower MAE to that of MF, 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. SMI-TED demonstrated a 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. Proposed approach reports lower prediction errors for LiI capacity and CE datasets, outperforming previously published models (6; 3) using the same datasets or their subsets. These results highlight applicability of foundation models to multivariate 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. SMI-TED successfully predicts these macroscale outcomes due to having rich chemical vocabulary comprising of 2988 unique chemical tokens or moieties. Hence, model latent

space is enriched with basic understanding of the chemical space formed by the combinations of chemical moieties in molecules (9). The fine-tuning step 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.

On datasets characterized by a large volume of data, such as solubility (3300 data) and IC (18,000 data), MF outperform SMI-TED embeddings in the present evaluation. 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 also 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. Nevertheless, SMI-TED approach still outperforms the array of ML approaches evaluated in literature for IC prediction as reported in ref (13). Another instance where MF outperforms SMI-TED despite low data regime is Li-ICl Capacity (MF MAE 32.24 mAh/g vs SMI-TED MAE 47.94 mAhg), highlighting present approach is not suitable for datasets lacking chemical variability. Ultimately, the choice of representation is a critical and must be determined by the nature of output label, quantity and the variability in the dataset, and the desired interpretability of the model.

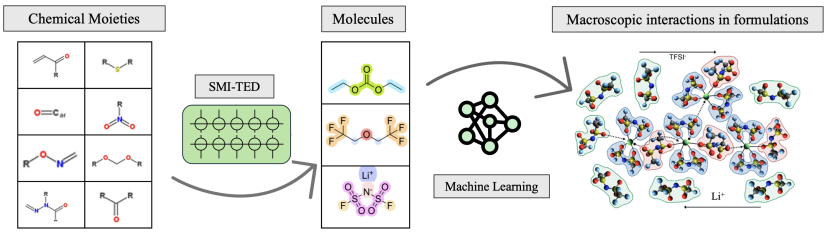

Figure 2: Multi-step training capturing complex chemical interactions at multiple scale.

4.2 Quantifying model uncertainty for out of distribution data

ML models frequently show poor transferability across chemical spaces and fall short in predicting properties for materials outside their training scope (25). Task-specific models trained on labeled data lack robustness when faced with new material classes. Improvements via transfer learning, domain adaptation, and embedding physics constraints are underway, but broad generalization remains elusive (6). Generalizable base models like foundation models have seen increased adoption in the community for these reasons (25). Latest works show discovery of new electrolyte formulations achieving high ionic conductivity (above 10 mS/cm) by screening a large generated formulation design space with a fine-tuned SMI-TED model(13). These results exhibit that 44% (7 in 16) of the electrolytes recommended by the model met the performance target during experimental validation. We observe there is further potential to ascertain the reliability of these models when extrapolating to unexplored regions of the materials design space. There are factors intrinsic to material design including scale and end-use application that inject additional complexity, fundamentally constraining the generalizability and reliability of OOD predictions in these contexts. This insight highlights the need for more nuanced evaluation strategies and tailored model development when extending AI methods to new regimes of materials science.

By incorporating uncertainty quantification into the model, we can systematically pinpoint regions where model lacks confidence. This capability is critical, as it allows for the intelligent allocation of resources toward targeted experimental validation and data enrichment, which is essential for improving the model's reliability and performance. We compared semantic similarity between the input embeddings of train-test distributions across several datasets in Figure 3. A similarity score (indicated in red) is employed as an approximation for how close test data is to training data, and is estimated by calculating maximum of average cosine similarity (normalized) of each test datapoint with all training samples. This metric is compared with prediction MAE for the respective train-test subset (in blue). These subsets were not random splits, but were instead carefully curated to represent a different testing scenario than the ones used in the previous section. Our evaluations confirm there is an inverse relationship between prediction MAE and semantic proximity of test data to the training samples.

These trends yield 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 5. This approach enables systematic assessment of prediction uncertainty and confidence for new data, thereby supporting efficient screening in materials design and discovery.

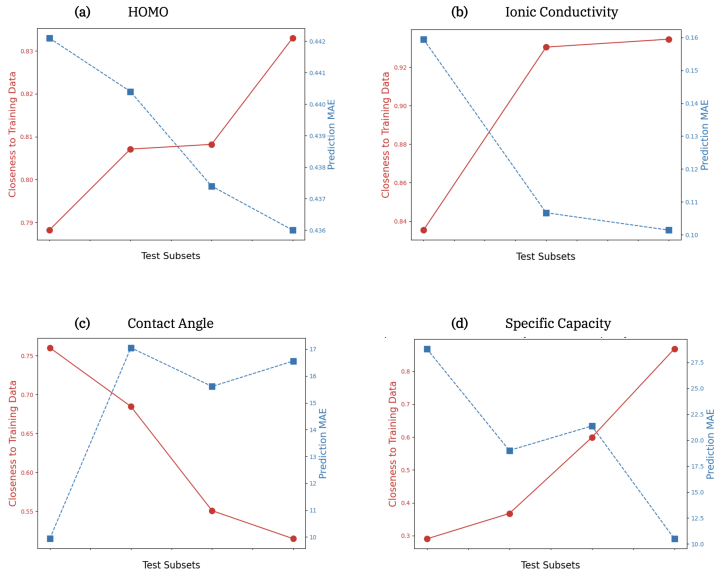

Figure 3: Relationship between prediction MAE (in blue) and chemical similarity (in red) between train and test datasets.

4.3 Foundation model 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 (26; 27). 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 (17). 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.

To evaluate interpretability of proposed foundation model derived predictors, we investigate 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 (28). 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) (29) 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 trained model for the very same test set. Figure 4 illustrates these correlations in three formulation datasets CE, LiI capacity and IC.

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.

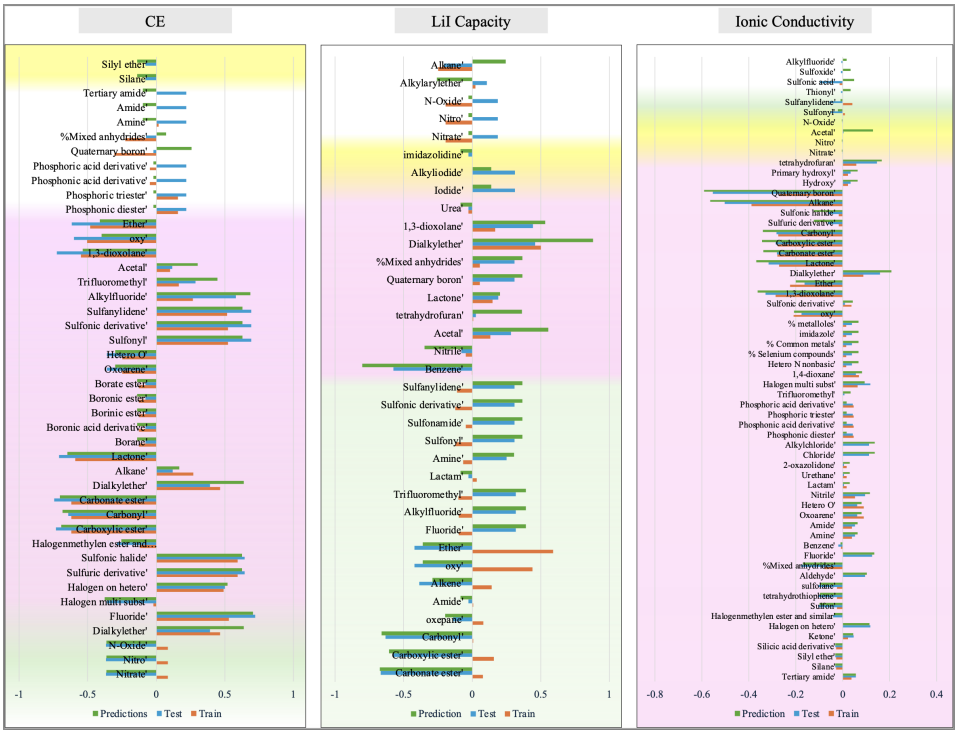

Figure 4: Correlation of chemical functional groups in formulations with performance in train (orange) - test (blue) dataset, compared with correlation to the predicted outcomes (green) in test data.

5 Conclusion

In this study, we evaluated the scope of foundation models in addressing material design challenges across multiple length scale in batteries: molecules, formulations and device. Open source SMI-TED model is used to encode molecular representations, then combined with other variables such as compositions, temperature, electrode and separator variations. The results showcase models like SMI-TED can be used to extrapolate learning from moiety-level interactions to macroscopic outcomes such as specific capacity, surface characteristics, and battery performance using both simulation and scarce empirical datasets. Results highlight that foundation model outperform alternatives methods especially in low data regimes. The study also presents methods to analyze model's ability to generalize out of distribution (OOD) and quantify model prediction errors across new material designs that are dissimilar to the training datasets. Lastly, we evaluate interpretability of these models and suggest users to selectively apply these models for design interpretation and discovery within a chemical space where confidence is justified.

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] 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.

[10] 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.

[11] 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.

[12] 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*.

[13] 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," *arXiv preprint arXiv:2503.14878*, 2025.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] P. Llompart, C. Minoletti, S. Baybekov, D. Horvath, G. Marcou, and A. Varnek, "Will we ever be able to accurately predict solubility?" *Scientific Data*, vol. 11, no. 1, p. 303, 2024.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

A Supplementary Material

A.1 Solubility Data Collection

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 highest molar concentration of single salt dissolved in mixture of organic solvents. The four salts, LiCl, LiNO3, LiTFSI and LiBOB, are individually dissolved in solvent formulations containing different compositions of EC, G4, DMI and DOL. 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; 17; 6) along with a few "failed" non miscible electrolytes (17; 13). 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.

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 processings, there are 3300 electrolyte formulation vs solubility data that is used in the study.

A.2 Contact Angle Experiments

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 constructed 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.

A.3 Model Training

Neural network (NN) architectures were individually optimized and trained using SMI-TED–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 <5000, and for >5000 batch size of 32 was used. 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 1: 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.4 Foundation model performance for downstream tasks

For each dataset, NNs were trained using five independent 80%-20% train–test splits, and prediction errors were quantified using the mean absolute error (MAE) metric.

Table 2: 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

Table 3: Mean absolute error (MAE) and prediction accuracy (%) across multiple train-test splits for the battery datasets using SMI-TED embeddings

Split	**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
1	0.2519	0.5870	0.4421	0.3673	93.80	0.0986	16.556	22.078	0.170	50.66
2	0.2547	0.5881	0.4374	0.3661	94.11	0.0871	15.610	29.904	0.185	55.19
3	0.2560	0.5842	0.4460	0.3688	92.29	0.1001	17.053	19.501	0.204	43.87
4	0.2609	0.5795	0.4404	0.3645	93.35	0.0867	9.937	15.348	0.192	50.46
5	0.2564	0.5741	0.4367	0.3651	91.99	0.0812	22.063	25.417	0.175	39.60
Average	0.2559	0.5825	0.4405	0.3663	93.11	0.0910	16.243	22.449	0.185	47.93

Table 4: Mean absolute error (MAE) and prediction accuracy (%) across multiple train-test splits for the battery datasets using Morgan Fingerprints

Split	**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
1	0.2563	0.5895	0.4617	0.3781	93.80	0.0648	16.876	29.940	0.199	9.77
2	0.2598	0.5922	0.4552	0.3762	94.86	0.0673	18.521	34.550	0.228	49.87
3	0.2608	0.5885	0.4572	0.3734	93.05	0.0633	17.355	27.848	0.244	17.16
4	0.2638	0.5773	0.4576	0.3720	93.95	0.0598	12.889	15.513	0.199	46.42
5	0.2580	0.5798	0.4587	0.3737	93.20	0.0597	23.438	37.101	0.244	37.97
Average	0.2594	0.5854	0.4580	0.3746	93.77	0.0629	17.815	28.990	0.223	32.24

Table 5: Parameters to estimate mean absolute error (MAE) in model prediction based on similarity between test-train data

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