Capturing Formulation Design of Battery Electrolytes with Chemical Large Language Model

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

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

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

Renato Cerqueira
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
rcerq@br.ibm.com

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

Abstract

Recent progress in large transformers-based foundation models have demonstrated impressive capabilities in mastering complex chemical language representations. These models show promise in learning task-agnostic chemical language representations through a two-step process: pre-training on extensive unlabeled corpora and fine-tuning on specific downstream tasks. By utilizing self-supervised learning capabilities, foundation models have significantly reduced the reliance on labeled data and task-specific features, streamlining data acquisition and pushing the boundaries of chemical language representation. However, their practical implementation in further downstream tasks is still in its early stages and largely limited to sequencing problems. The proposed multimodal approach using MoLFormer, a chemical large language model, aims to demonstrate the capabilities of transformer based models to non-sequencing applications such as capturing design space of liquid formulations. Multimodal MoLFormer utilizes the extensive chemical information learned in pre-training from unlabeled corpora for predicting performance of battery electrolytes and showcases superior performance compared to state-of-the-art algorithms. The potential of foundation models in designing mixed material systems such as liquid formulations presents a groundbreaking opportunity to accelerate the discovery and optimization of new materials and formulations across various industries.

1 Introduction

In recent times, the field of large transformers-based foundation models has made remarkable progress, showcasing their impressive capacity to master complex chemical language representations [1, 2, 3]. This machine learning approach has become widely adopted for accurately predicting molecular properties due to its efficiency and ability to represent essential molecular features [4]. The said achievement of large chemical language models is just a tip of an iceberg that represents their immense scope and potential [5]. The applications of such models need to be further explored beyond the domain of molecules, towards more complex design spaces such as formulations.

Recent advancements in these models have shown great promise in learning task-agnostic chemical language representations through a two-step process: pre-training on extensive unlabeled corpora and fine-tuning on specific downstream tasks [6, 7, 8]. The significance of this achievement cannot be overstated [9]. By utilizing self-supervised learning capabilities, these foundation models have

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

effectively reduced the dependence on labeled data and task-specific features [7, 10], streamlining the once laborious data acquisition process and propelling the field of chemical language representation to new heights [11, 12, 13]. Although pre-trained Language Models (LMs) have shown promise in predicting molecular properties [14, 15, 16], their practical implementation on further downstream tasks is still in its early stages [17] and largely limited to sequencing problems such as predicting sequences of proteins [18], polymers [19] and chemical reactions [20]. The question remains: Could LMs be used to represent molecular systems with unstructured and random interactions like the ones observed in liquid formulations?

Liquid formulations are a big part of several industrial sectors like pharmaceuticals, automotive materials, and food science [21]. Current strategies to design formulations rely on high-throughput virtual screening that expedites the search for individual compounds but falls short in guiding the complete design of materials' formulations[22]. Battery liquid electrolytes are economically relevant examples of a formulation system, where comprehending and optimizing the interdependencies of constituent solvents and salts is of paramount importance for device performance [23]. Despite the exponential growth of the energy storage field in the last two decades, the cycling stability of current battery technologies continues to remain in question [24]. Electrolyte engineering has emerged to be a promising approach to improve the cycling efficiency of next generation batteries, and remains generally an experimentally driven process. The major bottleneck in adopting machine learning methods for electrolyte design discovery and optimization is the non-generalizability of the battery-specific datasets available in the literature and the expensive data acquisition process [25].

LMs can play a crucial role in bridging this gap, as their self-supervised capabilities align perfectly with scenarios where data availability is critical [26, 27]. In this paper, we propose a multimodal approach built upon the recently introduced LM MoLFormer to predict the performance of battery electrolyte formulations. Unlike recently introduced chemical LMs that predict properties based on a single molecule identifier (SMILES [16, 28], SELFIES [29, 30], etc. [31]), our approach takes up to six SMILES molecules as input, representing the constituents of the formulations along with their respective molar percentages, thus capturing the composition of the design space (see Fig. 1). To assess the effectiveness of our approach, we utilize a Li/Cu half-cell dataset from a previous study [32], which serves as a benchmark for our proposed methodology. Our approach has demonstrated superior performance compared to state-of-the-art algorithms, eliminating the need for laborious human feature engineering processes and extensive experimental data acquisition.

Our work demonstrates the potential of foundation models for the design of mixed material systems. By leveraging the power of machine learning, we can accelerate the discovery and optimization of new materials and formulations, with the potential to revolutionize a wide range of industries.

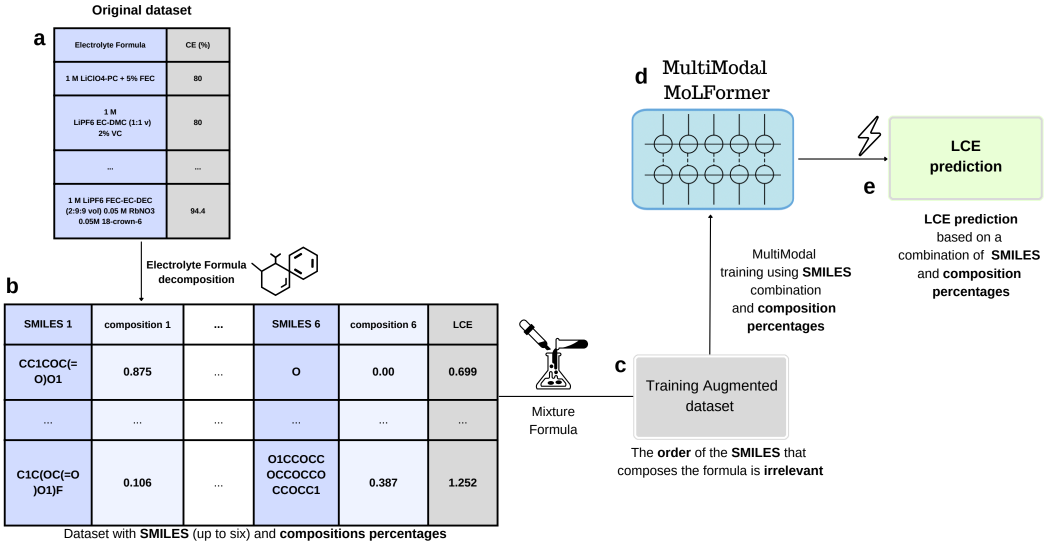
2 Method

Our approach builds upon the foundation of MoLFormer [7], a cutting-edge transformer-based model widely used for chemical language representations. MoLFormer is a large-scale masked language model that processes inputs through a series of blocks, alternating between self-attention and feed-forward connections.

The self-attention mechanism of MoLFormer allows the model to construct complex representations by incorporating contextual information from across the input sequence. By transforming the sequence features into query (q), key (k), and value (v) representations, attention mechanisms can weigh the importance of different elements within the sequence. This ability to capture informative relationships between tokens makes MoLFormer a powerful tool for predicting molecular properties.

To further enhance performance, recent studies have demonstrated the benefits of incorporating relative position embeddings between tokens [14]. MoLFormer optimizes relative encoding by using a modified version of the RoFormer [33] attention mechanism. This involves position-dependent rotations (R_m) of the query and keys at position m. These rotations can be efficiently implemented as pointwise multiplications, ensuring that the computational complexity remains manageable (as shown in Eq (1)).

$$Attention_m(Q, K, V) = \frac{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle v_n}{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle} \tag{1}$$



Electrolyte Formula	CE (%)
1 M LiClO4-PC + 5% FEC	80
1 M LiPF6 EC-DMC (1:1 v) 2% VC	80
...	...
1 M LiPF6 FEC-EC-DEC (2:9:9 vol) 0.05 M RbNO3 0.05M 18-crown-6	94.4

SMILES 1	composition 1	...	SMILES 6	composition 6	LCE
CC1COC(=O)O1	0.875	...	O	0.00	0.699
...	...	...	...	...	...
C1C(OC(=O)O1)F	0.106	...	O1CCOCCOCCOCCOCCOCC1	0.387	1.252

Figure 1: The figure illustrates the general architecture of the learning process of the MultiModal-MoLFormer. (**a**) Electrolyte formulation dataset. Illustrated by the example, the sequence which comprises electrolytes formulations along with their Coulombic Efficiency. **b** Description of electrolyte formulations as input. Each formulation is composed of a sequence of up to six SMILES along with their fraction of molar percentages and respective logarithmic Coulombic Efficiency as performance label. **c** Dataset augmentation in order to enrich the training of the model. **d** Training of the proposed MultiModal-MoLFormer approach with the sequence of SMILES and compositions percentages. **e** Prediction of the Coulombic efficiency property based on mixture formulation (SMILES and compositions).

In Eq (1), $Attention_m(Q, K, V)$ denotes the attention operation with queries (Q), keys (K), and values (V) at position m. The operation computes weighted sums of the value representations (v_n) based on the similarity of the transformed query ($\varphi(R_m q_m)$) and key ($\varphi(R_n k_n)$) representations. The relative position embeddings introduced through the rotations (R_m) allow the model to effectively capture positional information, leading to improved performance in molecular property predictions.

By leveraging the capabilities of MoLFormer and enhancing it with relative position embeddings, our approach offers an advanced and efficient solution for predicting complex molecular properties, providing valuable insights for various chemical applications.

2.1 Tokenization process and vocabulary construction

The approach employs a tokenization process, as described in [6], to create its vocabulary. This tokenization process utilizes an extensive dataset comprising 1.1 billion molecules from PubChem and ZINC datasets, resulting in the generation of 2362 unique vocabulary tokens. These tokens are then used for fine-tuning or retraining the models, with a fixed embedding capacity and vocabulary size.

To optimize computation time and resource utilization, the sequence length is constrained to a range of 1 to 202 tokens, including special tokens. This decision is driven by the fact that over 99.4% of all 1.1 billion molecules in the dataset contain fewer than 202 tokens. By setting this limit, the model can effectively handle the vast majority of molecular structures while avoiding unnecessary computations for excessively long sequences.

The tokenization process and the limited sequence length enable the approach to efficiently process and represent molecular structures, making it feasible to scale the model to large datasets and achieve powerful predictive capabilities in various chemical applications.

2.2 Multimodal approach

In this section, we present the multimodal layer of our proposed approach, where SMILES notations embeddings are combined with their corresponding percentages of the formulations derived from the electrolyte process.

Let (x, y) denote a feature-target pair, where $x = (x_{\text{CL}}, x_{\text{proportions}})$. The x_{CL} represents all the features based on chemical language representations, and $x_{\text{proportions}}$ refers to the proportions of the compounds that compose the formulations. Each x_{CL} can have a sequence of tokens with a maximum length of 202 tokens, and each token has a 768-dimensional embedding. Fig. 2 illustrates the concatenation architecture of the proposed approach.

After the data transformation using Transformer layers (Fig. 2), the resulting embeddings per chemical language string are concatenated along with their corresponding percentages of the formulations derived from the electrolyte process, resulting in a learning vector of dimension $(d \times e + c)$. Here, d represents the dimension of the dataset, e is the size of the resulting embeddings, and c is the number of features that represent the formulations' percentages. This concatenated vector is then passed to a learning algorithm, specifically a Feed-Forward network with 2 fully connected layers, to calculate the loss function.

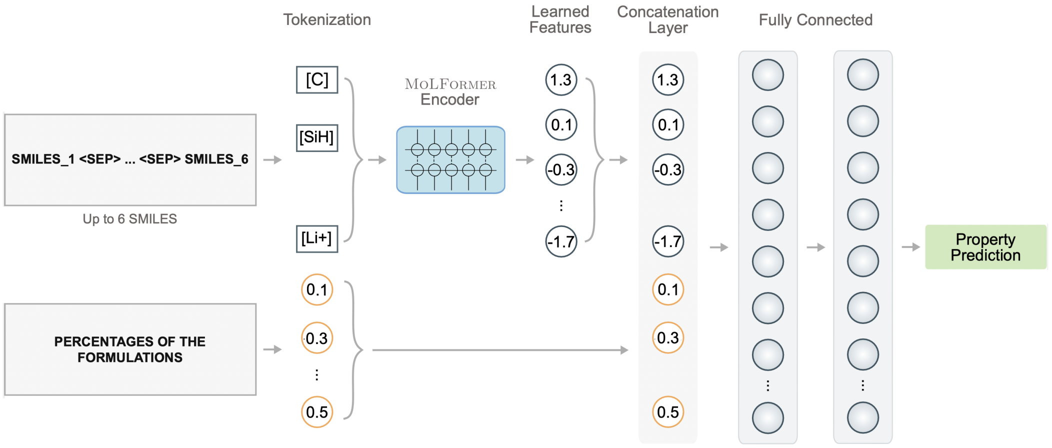


Figure 2: The figure illustrates the general architecture of the learning process of the MultiModal-MoLFormer. The concatenation layer is responsible to combine the SMILES embeddings along with the percentages of the formulations derived from the electrolyte process.

The method takes up to six SMILES as input, which are concatenated using a special token **<SEP>**. As the order of the SMILES does not matter in the formulation, during the training phase, we consider all possible permutations of the order in which the SMILES may appear to the system. This facilitates the learning process, as the algorithm is based on a Transformers architecture, where the order of tokens does matter to the system.

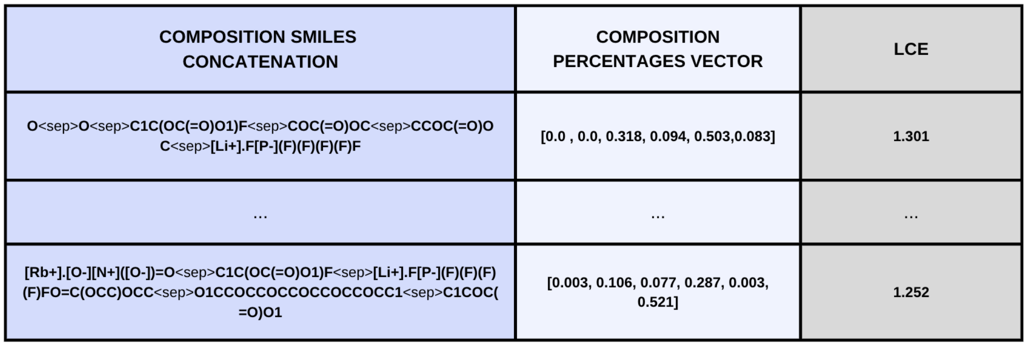
The multimodal approach successfully fuses SMILES notations and formulation percentages, enabling accurate predictions of performance metrics. By effectively representing the complex relationships between chemical components and performance metrics, our MultiModal-MoLFormer method achieves improved predictive performance compared to traditional approaches.

In summary, our proposed approach showcases the significance of leveraging multimodal information to enhance the understanding and prediction of complex systems in battery electrolyte formulations. The seamless integration of SMILES notations and formulation percentages contributes to the advancement of computational materials discovery, bridging the gap between material discovery and development, and offering a valuable tool to expedite the exploration of new materials with enhanced properties for diverse applications.

2.3 Dataset augmentation

When dealing with formulations, it is essential to note that the sequence of SMILES representations holds no significance [34]. Consequently, the arrangement of data within the dataset can be strategically permuted, as depicted in Fig. 3, to effectively enhance data augmentation. This augmentation approach serves the purpose of bolstering the performance of the proposed model, thereby contributing to its overall effectiveness [35]. An additional crucial point to emphasize is that these formulations have the potential to encompass up to six distinct SMILES. In cases where all six SMILES are not fully specified, any vacant spaces are automatically filled with "O", while the corresponding contribution to the composition percentage is then designated as 0.0 [36]. It is also important to note that the special token **<SEP>** is ignored by the proposed approach [37]. This systematic approach ensures consistency and accuracy in handling incomplete SMILES within the compositions. By incorporating the data augmentation process, the training dataset underwent a substantial expansion, transitioning from 147 compositions to 27,266 compositions. This notable augmentation was attained through a sequence of steps, commencing with the permutation procedure and culminating in the removal of duplicates. This intricate transformation effectively paved the way for the dataset's substantial growth. It is important to highlight that the test dataset is not augmented in order to preserve the fair evaluation of the algorithms.

Transformers-based approaches as the one proposed here, greatly benefit from larger datasets [38], as they encompass a more extensive range of text sequences [39]. The incorporation of these expanded datasets contributes significantly to refining the comprehension and predictive abilities of transformers. This refinement ultimately leads to improved performance across a wide array of tasks and applications. The increased diversity in sequence composition empowers transformers to enhance their performance by capturing a broader spectrum of linguistic patterns, nuances, and contextual intricacies [40]. In summary, the utilization of larger datasets empowers transformers to enhance their understanding and predictive capacities, thereby resulting in elevated performance levels across diverse tasks and applications.

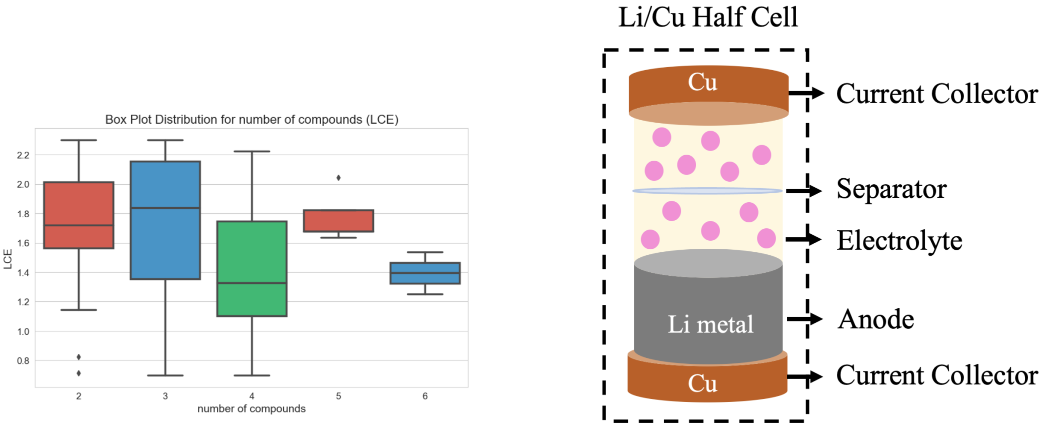


COMPOSITION SMILES CONCATENATION	COMPOSITION PERCENTAGES VECTOR	LCE
O<sep>O<sep>C1C(OC(=O)O1)F<sep>COC(=O)OC<sep>CCOC(=O)OC<sep>[Li+].F[P-](F)(F)(F)(F)F	[0.0 , 0.0, 0.318, 0.094, 0.503,0.083]	1.301
...	...	...
[Rb+].[O-][N+]([O-])=O<sep>C1C(OC(=O)O1)F<sep>[Li+].F[P-](F)(F)(F)(F)FO=C(OCC)OCC<sep>O1CCOCCOCCOCCOCCOCC1<sep>C1COC(=O)O1	[0.003, 0.106, 0.077, 0.287, 0.003, 0.521]	1.252

Figure 3: The figure elucidates the permuted SMILES of each formulation constituents, alongside a breakdown of the constituent percentages of each SMILE within the formulation. This breakdown is represented in the context of the input vector, providing a comprehensive understanding of the composition dynamics.

3 Electrolyte Formulation Dataset

To evaluate this proposed methodology, we assessed its efficacy in tackling a challenging task. We considered the Li/Cu half cell-based electrolyte formulations and their respective Coulombic Efficiencies (CE). This dataset was carefully curated from literature to encompass a wide range of electrolyte variations in terms of constituent molecules and compositions [32]. CE is a crucial metric for assessing battery performance that represents the ratio of discharge to charge capacity [41]. Maintaining a high CE is essential to ensure optimal battery function. However, over time, batteries can experience CE loss that is primarily caused by electrolyte and electrode decomposition [42]. The CE values have been converted to their logarithmic (LCE) by [32] to numerically amplify the change in output with respect to the electrolytes. This transformation allows for a more sensitive and accurate comparison of the performance of different electrolytes. The dataset is composed of 147 electrolyte formulations for training purposes and 13 electrolyte formulations for model evaluation. The box plots illustrated by Fig. 4a provide a clear visualization of the distribution of LCE outputs for the training data based on the count of formulation constituents. The plot showcases essential statistical insights, revealing the spread and central tendencies of the data. The formulations in the dataset consist of 2 to 6 electrolyte components in each, represented as a simplified molecular-input line-entry system (SMILES) notation. The presented approach allowed us to gain valuable insights into the potential applicability of LMs in device-level predictive applications.



(a) Distribution of performance label vs formulant count

(b) Lithium -Copper half cell schematic

Figure 4: (**a**)Box plots depicting electrolyte formulation vs LCE data as the function of number of formulants in the formulations. Each box plot is constructed with attention to detail, where the central line represents the median value of the LCE outputs, indicating the middle point of the data distribution. The colored box represents the interquartile range (IQR), encompassing the 25^{th} to 75^{th} percentile of the LCE values. This range provides valuable information about the variability and spread of the majority of data points.(**b**) Schematic representation of Lithium-Copper half cell that is conventionally used to study stability and cyclability of electrolyte formulations over Lithium metal anode.

4 Performance prediction of electrolyte formulations

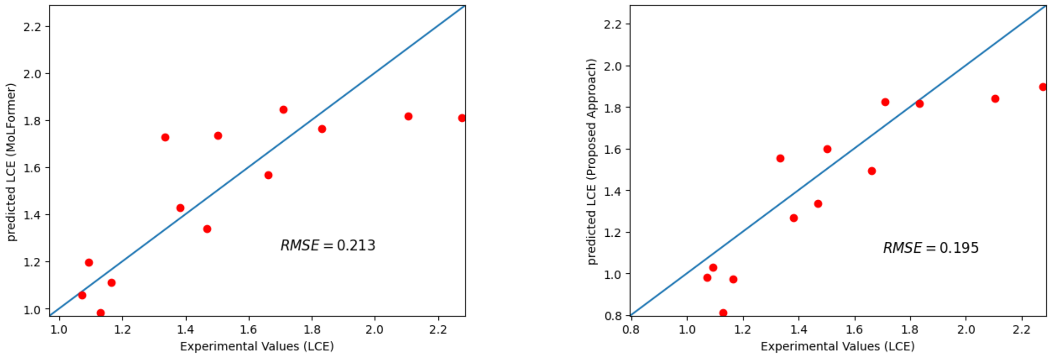
Here, we present a comparative analysis of LCE predictions using the Li/Cu half-cell dataset from two approaches: Molformer and MultiModal-MoLFormer. Table 1 summarizes the LCE predictions for the test dataset from MoLFormer [7] and proposed MultiModal-MoLFormer. The performance of each algorithm is assessed using the root mean squared error (RMSE) metric, which quantifies the prediction errors. Figure 5 illustrates the parity plot prediction of LCE for all the considered algorithms.

Each row in the table displays the predicted LCE values for individual electrolyte formulations. The numerical results indicate that our method achieves significantly lower prediction errors, as reflected in the RMSE calculation. Morever, as illustrated by Fig. 6, the residual plot demonstrates that the proposed approach has stable predictions.

The Table 2 compares the predictive errors of proposed models with alternative formulation models. Specifically, the RMSE values for each algorithm are as follows: “F-GCN TL” achieves an RMSE of 0.389, and MoLFormer shows an RMSE of 0.213. In contrast, our proposed approach achieves the lowest RMSE of 0.195, demonstrating its superior predictive capability. F-GCN are formulation graph convolution networks that use graph representations for representing formulation constituents along with their compositions and overcome the limitations of small experimental dataset by pre-training graphs on labeled simulation data. Thus, ’TL’ denotes transfer learning in the F-GCN framework [36]. It is interesting to note here that despite skipping over important compositional information,

Table 1: Summary of logarithmic Coulombic efficiency (LCE) predictions from MoLFormer and Muldimodal-MoLFormer. The root mean squared error (RMSE) metric was used to measure the errors of the algorithms. Each row in the table displays the predicted LCE values for an individual electrolyte formulation.

Experimental values (LCE) [32]	MoLFormer	**Multimodal MoLFormer**
1.094	1.198	1.028
1.384	1.428	1.267
1.468	1.340	1.336
1.710	1.845	1.823
1.832	1.763	1.816
2.104	1.816	1.841
2.274	1.809	1.897
1.071	1.058	0.979
1.166	1.109	0.971
1.335	1.727	1.554
1.129	0.982	0.810
1.501	1.735	1.599
1.663	1.565	1.492
RMSE	0.213	**0.195**



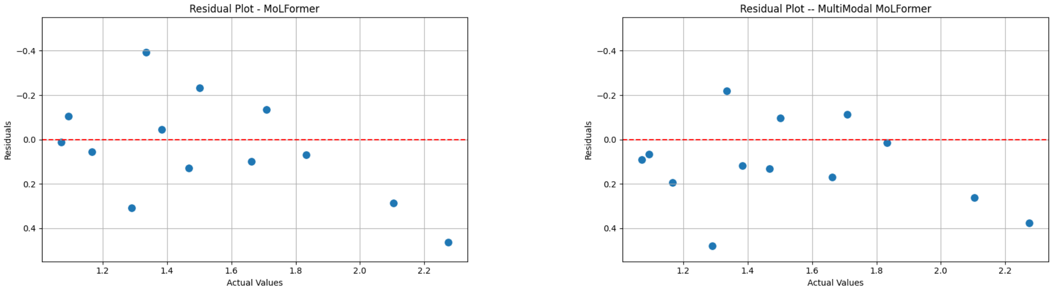
(a) Parity plot for MoLFormer predicted LCE values

(b) Parity plot for Multimodal-MoLFormer predicted LCE values

Figure 5: The figure depicts the performance of our multimodal chemical language approach, which leverages both SMILES and composition percentages of formulation constituents to predict performance metric LCE.

MolFormer still outperforms F-GCN TL model that captures both structure and compositions of formulation constituents. This could be attributed to the large-scale learning of underlying chemical information by MolFormer from unlabeled corpora which makes it highly generalizable and property-independent. Meanwhile, graphs depend upon initial vectorization and specific property to learn structure-based relational latent space. Not only acquiring labeled data to train graphs can be very costly, but it also renders the ability of such models to predict formulation properties dependent upon pre-training labels. MultiModal-MoLFormer manages to overcome the compositional negligence of MoLFormer in capturing formulation design and demonstrates the best predictive capability of any algorithm report to date. The table demonstrates the efficacy of our multimodal chemical language approach, which just utilizes SMILES and formulation percentages for improved prediction accuracy. By effectively incorporating essential chemical information, our proposed method better captures the complex interactions that influence LCE in Li/Cu half-cell batteries. This enhanced predictive power for a complex unstructured design space is crucial for electrolyte engineering and optimizing battery performance, thus driving advancements in battery technology.

In summary, Table 2 clearly shows that our proposed approach outperforms other state-of-the-art algorithms in predicting LCE values for the Li/Cu half-cell dataset. The lower RMSE value obtained



(a) Residual plots of the predicted LCE values by the MoLFormer approach

(b) Residual plots of the predicted LCE values by the multimodal approach proposed here

Figure 6: The residuals plots depicting Multimodal- MoLFormer to have more stable predictions for LCE values than the alternative method.

Table 2: Comparison of Coulombic Efficieny prediction from different algorithms.

Algorithm	**RMSE**
MultiModal-MoLFormer	**0.195**
MoLFormer	0.213
F-GCN TL [36]	0.389
Linear regression [32]	0.585
Random forest [32]	0.577
Boosting [32]	0.587
Bagging [32]	0.583

by our method underscores its potential to facilitate more reliable and efficient battery design, ultimately contributing to the development of sustainable and high-performance energy storage solutions.

5 Conclusion

This paper introduces an innovative multimodal chemical language-based model, specifically designed to establish intricate relationships among the structural composition and performance of battery electrolytes within their formulation space. By seamlessly integrating both SMILES notations and the corresponding formulation percentages, our model provides a more comprehensive depiction of chemical interactions.

To validate the efficacy of our proposed approach, we conducted rigorous experiments using the Li/Cu half-cell data [32]. Our approach not only surpassed state-of-the-art methodologies in predicting formulation property i.e. logarithmic coulombic efficiency in this case, but also achieved an impressive RMSE of 0.195. Noteworthy is the fact that our method even outperformed the formulation graph model (F-GCN), known for its reliance on resource-intensive HUMO-LUMO features for pre-training. This compellingly suggests that our approach excels in both efficiency and effectiveness compared to existing methods, which rely on simpler input features such as SMILES notations and formulation percentages.

In contrast to other algorithms demanding resource-heavy features, our method relies exclusively on straightforward input features, rendering it more accessible and computationally streamlined. Moreover, the residual plot (refer to Figure 6) illustrates a narrower spread of residuals compared to the alternative formulation models. This significant finding implies that our proposed approach is more resilient against noise and outliers in the data. In simpler terms, our approach demonstrates heightened resistance to random data fluctuations.

It is paramount to underscore that our approach exhibits remarkable performance even when faced with challenging tasks, given the relatively modest size of the datasets employed in this study. Acquiring extensive datasets can be a resource-intensive endeavor, imposing constraints on data availability. Consequently, we advocate for further experimentation on larger datasets to holistically explore the model's capabilities, thus validating its scalability and robustness.

In conclusion, the proposed multimodal chemical language-based model showcases exceptional predictive prowess regarding logarithmic Coulombic efficiency. Its reliance on uncomplicated features underscores its practicality and efficiency. While acknowledging the prevailing limitations of our dataset, this work serves as a foundational stepping stone for future research in the realm of battery technology. We fervently encourage continued exploration on more expansive datasets to unlock the full potential of our model.

References

[1] F. Grisoni, "Chemical language models for de novo drug design: Challenges and opportunities," *Current Opinion in Structural Biology*, vol. 79, p. 102527, 2023.

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

[3] A. D. White, "The future of chemistry is language," *Nature Reviews Chemistry*, pp. 1–2, 2023.

[4] D. S. Wigh, J. M. Goodman, and A. A. Lapkin, "A review of molecular representation in the age of machine learning," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 12, no. 5, p. e1603, 2022.

[5] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[6] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction," *ACS central science*, vol. 5, no. 9, pp. 1572–1583, 2019.

[7] 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.

[8] M. Moret, I. Pachon Angona, L. Cotos, S. Yan, K. Atz, C. Brunner, M. Baumgartner, F. Grisoni, and G. Schneider, "Leveraging molecular structure and bioactivity with chemical language models for de novo drug design," *Nature Communications*, vol. 14, no. 1, p. 114, 2023.

[9] D. Flam-Shepherd, K. Zhu, and A. Aspuru-Guzik, "Language models can learn complex molecular distributions," *Nature Communications*, vol. 13, no. 1, p. 3293, 2022.

[10] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[11] 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.

[12] S. Huang and J. M. Cole, "Batterybert: A pretrained language model for battery database enhancement," *Journal of Chemical Information and Modeling*, vol. 62, no. 24, pp. 6365–6377, 2022.

[13] C. Qian, H. Tang, Z. Yang, H. Liang, and Y. Liu, "Can large language models empower molecular property prediction?" *arXiv preprint arXiv:2307.07443*, 2023.

[14] Y. Liu, R. Zhang, T. Li, J. Jiang, J. Ma, and P. Wang, "Molrope-bert: An enhanced molecular representation with rotary position embedding for molecular property prediction," *Journal of Molecular Graphics and Modelling*, vol. 118, p. 108344, 2023.

[15] H. Chen and J. Bajorath, "Designing highly potent compounds using a chemical language model," *Scientific Reports*, vol. 13, no. 1, p. 7412, 2023.

[16] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang, "Smiles-bert: large scale unsupervised pre-training for molecular property prediction," in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 2019, pp. 429–436.

[17] B. Sanchez-Lengeling and A. Aspuru-Guzik, “Inverse molecular design using machine learning: Generative models for matter engineering,” *Science*, vol. 361, no. 6400, pp. 360–365, 2018.

[18] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos Jr, C. Xiong, Z. Z. Sun, R. Socher *et al.*, “Large language models generate functional protein sequences across diverse families,” *Nature Biotechnology*, pp. 1–8, 2023.

[19] C. Xu, Y. Wang, and A. Barati Farimani, “Transpolymer: a transformer-based language model for polymer property predictions,” *npj Computational Materials*, vol. 9, no. 1, p. 64, 2023.

[20] A. M. Bran, S. Cox, A. D. White, and P. Schwaller, “Chemcrow: Augmenting large-language models with chemistry tools,” *arXiv preprint arXiv:2304.05376*, 2023.

[21] K. M. Herbert, H. E. Fowler, J. M. McCracken, K. R. Schlafmann, J. A. Koch, and T. J. White, “Synthesis and alignment of liquid crystalline elastomers,” *Nature Reviews Materials*, vol. 7, no. 1, pp. 23–38, 2022.

[22] M. Abolhasani and E. Kumacheva, “The rise of self-driving labs in chemical and materials sciences,” *Nature Synthesis*, pp. 1–10, 2023.

[23] J. Janek and W. G. Zeier, “A solid future for battery development,” *Nature energy*, vol. 1, no. 9, pp. 1–4, 2016.

[24] X. He, D. Bresser, S. Passerini, F. Baakes, U. Krewer, J. Lopez, C. T. Mallia, Y. Shao-Horn, I. Cekic-Laskovic, S. Wiemers-Meyer *et al.*, “The passivity of lithium electrodes in liquid electrolytes for secondary batteries,” *Nature Reviews Materials*, vol. 6, no. 11, pp. 1036–1052, 2021.

[25] S. Hu and C. Huang, “Machine-learning approaches for the discovery of electrolyte materials for solid-state lithium batteries,” *Batteries*, vol. 9, no. 4, p. 228, 2023.

[26] R. A. Patel and M. A. Webb, “Data-driven design of polymer-based biomaterials: high-throughput simulation, experimentation, and machine learning,” *ACS Applied Bio Materials*, 2023.

[27] E. V. Brazil, E. Soares, L. V. Real, L. Azevedo, V. Segura, L. Zerkowski, and R. Cerqueira, “Position paper on dataset engineering to accelerate science,” *arXiv preprint arXiv:2303.05545*, 2023.

[28] B. Winter, C. Winter, J. Schilling, and A. Bardow, “A smile is all you need: predicting limiting activity coefficients from smiles with natural language processing,” *Digital Discovery*, vol. 1, no. 6, pp. 859–869, 2022.

[29] A. Yüksel, E. Ulusoy, A. Ünlü, and T. Doğan, “Selformer: Molecular representation learning via selfies language models,” *Machine Learning: Science and Technology*, 2023.

[30] M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle, K. M. Jablonka *et al.*, “Selfies and the future of molecular string representations,” *Patterns*, vol. 3, no. 10, 2022.

[31] J. Born and M. Manica, “Regression transformer enables concurrent sequence regression and generation for molecular language modelling,” *Nature Machine Intelligence*, vol. 5, no. 4, pp. 432–444, 2023.

[32] 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.

[33] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, “Roformer: Enhanced transformer with rotary position embedding,” *arXiv preprint arXiv:2104.09864*, 2021.

[34] K. Xu, “Li-ion battery electrolytes,” *Nature Energy*, vol. 6, no. 7, pp. 763–763, 2021.

[35] C. Shorten, T. M. Khoshgoftaar, and B. Furht, “Text data augmentation for deep learning,” *Journal of big Data*, vol. 8, pp. 1–34, 2021.

[36] 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,” *arXiv preprint arXiv:2307.03811*, 2023.

[37] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, “What does bert look at? an analysis of bert’s attention,” *arXiv preprint arXiv:1906.04341*, 2019.

[38] T. Lin, Y. Wang, X. Liu, and X. Qiu, “A survey of transformers,” *AI Open*, 2022.

[39] M. C. Frank, “Baby steps in evaluating the capacities of large language models,” *Nature Reviews Psychology*, pp. 1–2, 2023.

[40] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, “Recent advances in natural language processing via large pre-trained language models: A survey,” *ACM Computing Surveys*, 2021.

[41] L. Sun, Y. Liu, J. Wu, R. Shao, R. Jiang, Z. Tie, and Z. Jin, “A review on recent advances for boosting initial coulombic efficiency of silicon anodic lithium ion batteries,” *Small*, vol. 18, no. 5, p. 2102894, 2022.

[42] W. Wu, C. Li, Z. Wang, H.-Y. Shi, Y. Song, X.-X. Liu, and X. Sun, “Electrode and electrolyte regulation to promote coulombic efficiency and cycling stability of aqueous zinc-iodine batteries,” *Chemical Engineering Journal*, vol. 428, p. 131283, 2022.