Improving Performance Prediction of Electrolyte Formulations with Transformer-based Molecular Representation Model

Indra Priyadarsini S, Vidushi Sharma, Seiji Takeda, Akihiro Kishimoto, Lisa Hamada, and Hajime Shinohara
IBM Research – Tokyo

Introduction

- Electrolytes are critical in many fields, including energy storage (batteries), fuel cells, sensors, and electrochemical devices.
- In most practical applications such as electrolyte formulations in batteries, individual molecules are a part of multi-constituent system, that require capturing all individual constituents and their complex interactions to precisely predict the property or performance of the system
- Existing approaches lacks generalizability across different formulations and formulation constituents or the relation between the composition and the respective formulants cannot be guaranteed.
- In this paper we introduce a **transformer-based approach** suitable for **multi-constituent systems** such as battery electrolyte formulations.
- We propose a suitable approach to effectively capture the representation of electrolyte components, proportionate to their composition in the electrolyte formulation, to improve the performance of property prediction of electrolytes.

Electrolyte Formulation Dataset

- We evaluate the performance of the proposed approach on two datasets - **Li—Cu half cell** and **Li—I full cell** in the prediction of coulombic efficiency and specific capacities, respectively, given the electrolyte formulation.
- Li-Cu Half Cell dataset contains 147 entries of liquid electrolyte formulations along with their respective molar percentage and coulombic efficiency.
- The Li—I Full-Cell battery dataset was experimentally obtained for Li-I battery coin cells with cycling tests at 1mA/cm2 and contains 125 entries of electrolyte formulations.
- Each electrolyte formulation entry comprises of 2 to 6 electrolyte components

Electrolyte Formulation 1

Constituent	Composition
COCCC#N	0.56
O = C1N(C)CCCN1C	0.42
[Li+].[N+](= O)([O−])[O−]	0.02

Electrolyte Formulation 2

Constituent	Composition
C1C(OC(= O)O1)F	0.106
C1COC(= O)O1	0.522
O = C(OCC)OCC	0.287
[Li+].F[P−](F)(F)(F)(F)F	0.077
O1CCOCCOCCOCCOCCOCC1	0.008

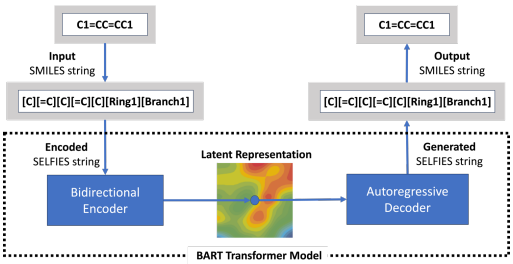
Electrolyte Formulation N

Constituent	Composition
...	...
...	...

Model and Schematic

Pretraining – BART model

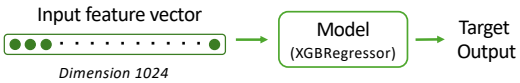
BART model pre-trained with SELFIES of drug-like small molecules from **ZINC-22** and **PubChem** dataset.



Finetuning

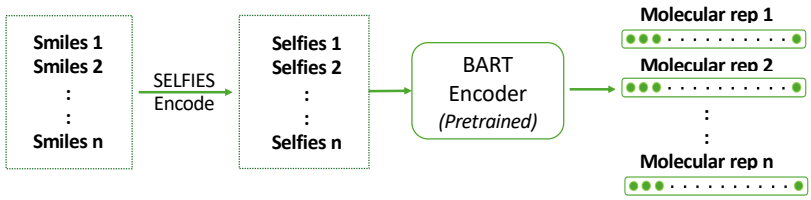
The input feature vector is constructed as a weighted linear combination of individual molecular representations obtained from the pretrained model

Input feature vector, $\mathbf{SA} = c_1\boldsymbol{r_1} + c_2\boldsymbol{r_2} + \ldots + c_n\boldsymbol{r_n}$

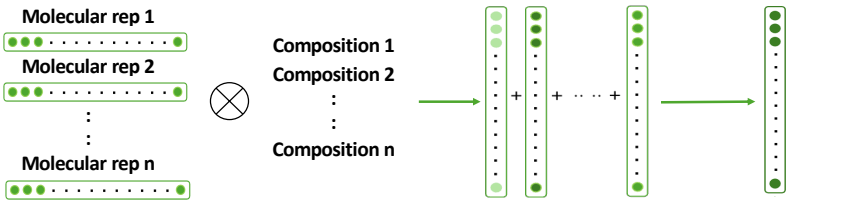


Feature Vector Construction

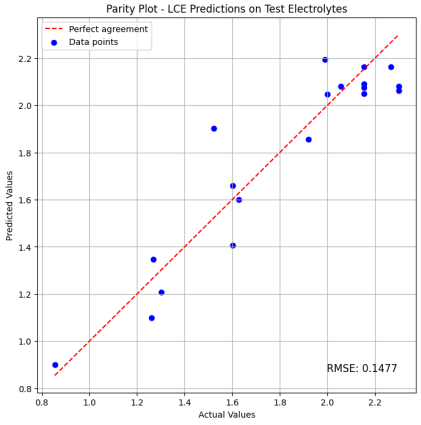
Step 1: Get molecular representations of the constituents of the formulation

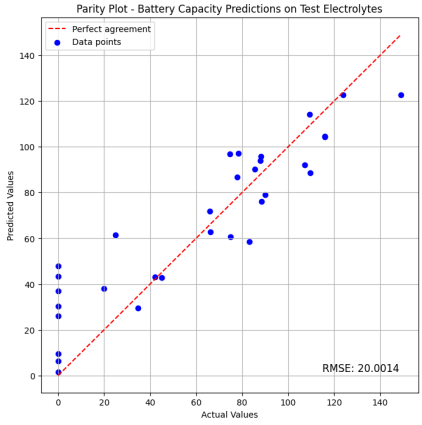


Step 2: Scale the molecular representations by composition and add



Results





Method	RMSE
Linear regression (Kim et al., 2023)	0.585
Random forest (Kim et al., 2023)	0.577
Boosting (Kim et al., 2023)	0.587
Bagging (Kim et al., 2023)	0.583
F-GCN TL (Sharma et al., 2023)	0.389
MoLFormer (Soares et al., 2024)	0.213
MM-MoLFormer (Soares et al., 2024)	0.195
BART-SA	**0.148**

Table 1. Comparison of RMSE of LCE prediction task

Method	RMSE
F-GCN no-TL (Sharma et al., 2023)	39.823
F-GCN TL (Sharma et al., 2023)	20.495
BART-SA	**20.001**

Table 2. Comparison of RMSE of Specific Capacity prediction task

References

1. Krenn, Mario, et al. "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation." Machine Learning: Science and Technology 1.4 (2020): 045024.
2. Kim, S. C., et al. Data-driven electrolyte design for lithium metal anodes. Proceedings of the National Academy of Sciences, 120(10):e2214357120, 2023.
3. Soares, E ., et al. Capturing formulation design of battery electrolytes with chemical large language model. 2024.
4. Sharma, V., et al. Formulation graphs for mapping structure-composition of battery electrolytes to device performance. Journal of Chemical Information and Modeling, 63(22):6998–7010, 2023.