A Large Encoder–Decoder Polymer-Based Foundation Model

Eduardo Soares[1], Nathaniel H. Park[2], Emilo Vital Brazil[1], Victor Shirasuna[3]

[1]IBM Research Brazil, Av. República do Chile, 330 Centro, Rio de Janeiro, RJ, 20031-170
[2]IBM Research Almaden, 650 Harry Rd, San Jose, CA, 95120
[3]IBM Research Brazil, Rua Tutóia, 1157 Vila Mariana, São Paulo, SP 04007-900

Introduction

The development and employment of generalizable foundation models is imperative for the accelerated development of new, high-performance polymeric materials. Few models to date have been demonstrated to be adept at prediction tasks across multiple property classes and polymer types. Here we demonstrate the capabilities of a new polymer foundation model trained on a new, serialized polymer graph (SPG) representation across numerous property prediction tasks.

Challenges with Polymer Data

— Statistical nature of polymers and properties

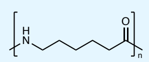
— Homopolymers and simple co-polymers dominate most available datasets

— Difficulty in represent microstructure, tacticity, and complex topologies

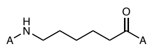
— Coacervates, mixed micelles, composites and other higher order structures are not well captured by existing representation systems

Polymer Data Representation

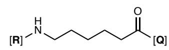
Homopolymer Representation



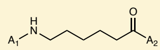
Nylon 6 (SMILES)
NCCCCCC(=O



PSMILES
NCCCCCC(=O)

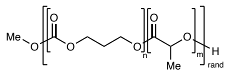


CMDL Polymer Graph v1[1]
<A|[Q]NCCCCCC(=O)[R]|A.Q -> A.R>

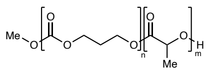


CMDL Polymer Graph v2 (SPG)
[*:1]NCCCCCC(=O)[*:2];2->1
This Work

Copolymer Representation

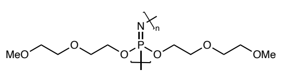


AB Block Copolymer
CO[*:1];O=C([*:2])OCCCO[*:3];O=C([*:4])C(C)O[*:5];
1->2;3->2;3->4;5->4

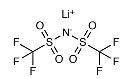


Statistical Copolymer
CO[*:1];O=C([*:2])OCCCO[*:3];O=C([*:4])C(C)O[*:5];
1->2;3->2;3->4;5->4;5->2

Polymer Formulations, Blends, & Mixtures



Polymer Component

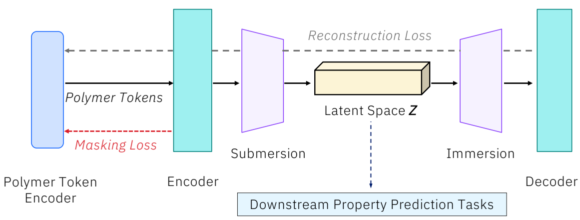


Salt Component

Polymer Electrolyte Formulation
[*:1]N=P([*:2])(OCCOCCOC)OCCOCCOC;2->1;O=S([N-]S(=O)(C(F)(F)F)=O)(C(F)(F)F)=O.[Li+]

Polymer Foundation Model

Model Architecture



Base Model Parameters

Hyperparameter	Value
Hidden Size	768
Attention Heads	12
Layers	12
Dropout	0.2
Normalization	LayerNorm
Vocabulary Size	2993
SMILES	91M
Mol Tokens	4T
Encoder Parameters	47M
Decoder Parameters	242M
Total Parameters	289M

Polymer Model Pre-training

— 1M polymer pre-training dataset represented as SPG

— Pretraining done over 150 epochs with a batch size of 256

— Two loss functions for pre-training

1) Token embeddings

2) Token reconstruction

Experiments

Benchmarking Dataset Performance

Dataset	Source	Metric	SOTA	SPG-SMI
Chain Bandgap	DFT	RMSE (↓)	**0.44**	0.49
Bulk Bandgap	DFT	RMSE (↓)	0.52	**0.32**
Electron Affinity	DFT	RMSE (↓)	**0.28**	0.29
Dielectric Constant	DFT	RMSE (↓)	0.52	**0.38**
Refractive Index	Exp.	RMSE (↓)	0.031	**0.021**
Conductivity–I	Exp.	MAE (↓)	1.00	**0.89**
Conductivity-II	Exp.	RMSE (↓)	0.61	**0.61**
CO_2 Permeability	Exp.	MAE (↓)	**0.29**	**0.29**
CH_4 Permeability	Exp.	MAE (↓)	0.37	**0.35**
N_2 Permeability	Exp.	MAE (↓)	0.38	**0.31**
CO_2/CH_4 Selectivity	Exp.	MAE (↓)	5.34	**4.71**
CO_2/N_2 Selectivity	Exp.	MAE (↓)	4.14	**3.89**
T_g -I (Polyimides)	Exp. & Syn.	MAE (↓)	24.4	**9.56**
T_g-II (Homopolymers)	Exp.	RMSE (↓)	**19.4**	27.7
T_d (50%)	Simulated	R^2 (↑)	0.92	**0.96**

Conclusions

— Evaluated new polymer foundation model on a variety of property prediction tasks

— Achieved state-of-the-art or near state-of-the-art performance on nearly all tasks

References

1. Park, N. H. *et al.* Artificial Intelligence Driven Design of Catalysts and Materials for Ring Opening Polymerization Using a Domain-Specific Language. *Nat Commun* **2023**, *14* (1), 3686.