STM JOURNALS

Journal of
Polymer & Composites

ISSN: 2321-2810 (Online)
ISSN: 2321-8525 (Print)
Volume 14, Special Issue 3, 2026
DOI (Journal): 10.37591/JoPC

https://journals.stmjournals.com/jopc

Research JoPC

Predicting Dielectric Constants of Polymers Using Molecular Structural Descriptors and Explainable Machine Learning: A Data-Driven Approach

C. Sridhathan[1], R. Arangasamy[2], Geetha Prahalad[3], Mohandass G.[4,*], N.M.G. Kumar[5], M. Ram Prasad Reddy[6]

Abstract

Accurate prediction of dielectric constants in polymeric materials is fundamental to the rational design of advanced electronic components, energy storage capacitors, flexible substrates, and high-frequency communication circuits. Conventional approaches to identifying suitable polymer dielectrics rely on extensive experimental synthesis and characterisation, which are both time-consuming and resource-intensive. In this work, an explainable machine learning framework is developed to predict the dielectric constant of polymers directly from molecular structural descriptors derived from first-principles calculations. The study exclusively utilises the Khazana Polymer Dataset — a publicly available, CC0-licensed repository of 1,073 polymers with density functional theory (DFT)-computed dielectric constants, band gaps, and atomisation energies — to construct and rigorously validate the predictive model. Eight structural descriptors including band gap, atomisation energy, unit cell volume, atom count per repeat unit, electronegativity variance, oxygen-to-carbon ratio, halogen presence, and metal-element flag are extracted from CIF crystal structure files using the pymatgen library. A gradient boosting regression (GBR) model optimised via five-fold cross-validated grid search achieves R^2 = 0.931, RMSE = 0.43, and MAPE = 4.7% on the held-out test set, outperforming random forest, support vector regression, and linear regression baselines across all metrics. SHAP (Shapley Additive Explanations) feature attribution identifies band gap and atomisation energy as the dominant predictors of dielectric behaviour, providing physically interpretable and chemically actionable structure–property design rules. The proposed framework offers a scalable, fully reproducible pathway for large-scale polymer dielectric screening and data-driven inverse design.

***Author for Correspondence**
Mohandass G.

[1]Professor, Department of Electronics and Communication Engineering, KCG College of Technology, Karapakkam, Chennai, Tamil Nadu, India
[2,4]Professor, Department of Biomedical Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India
[3]Professor, Department of Electronics and Communication Engineering, School of Engineering, Mohan Babu University, Tirupati, Andhra Pradesh, India
[5]Professor, Department of Electronics and Communication Engineering, School of Engineering, Mohan Babu University, Tirupati, Andhra Pradesh, India
[6]Professor, Department of Electrical and Electronics Engineering, Aditya College of Engineering, Madanapalle, Andhra Pradesh, India

Received Date: March 20, 2026
Accepted Date: May 02, 2026
Published Date: May 06, 2026

Citation: C. Sridhathan, R. Arangasamy, Geetha Prahalad, Mohandass G., N.M.G. Kumar, M. Ram Prasad Reddy. Predicting Dielectric Constants of Polymers Using Molecular Structural Descriptors and Explainable Machine Learning: A Data-Driven Approach. Journal of Polymer & Composites. 2026; 14(Special Issue 3): S297–S304p.

Keywords: Polymer dielectrics, dielectric constant prediction, molecular structural descriptors, explainable machine learning, gradient boosting regression, SHAP, Khazana dataset, structure–property relationships, band gap, polymer informatics

INTRODUCTION

Polymer dielectrics constitute a cornerstone of modern electronics, enabling advances in capacitive energy storage, flexible substrates, high-frequency communication circuits, and insulating coatings for power systems [1]. The dielectric constant — or relative permittivity — is the primary

descriptor governing how effectively a polymer stores electrical energy and how it interacts with electromagnetic fields [2]. Identifying polymers with specific dielectric constants for targeted applications has traditionally required extensive experimental synthesis and characterisation, a process that is both resource-intensive and time-consuming.

Data-driven approaches powered by machine learning (ML) have emerged as transformative alternatives for materials property prediction, enabling rapid screening of large compositional and structural spaces [3]. In the domain of polymer dielectrics, ML strategies have been applied to predict dielectric constants from chemical fingerprints [4], molecular dynamics trajectories [5], and DFT-computed electronic descriptors [6]. The foundational work of Mannodi-Kanakkithodi et al. [7] demonstrated that motif-based fingerprinting combined with kernel ridge regression can accurately predict DFT-computed dielectric constants across hundreds of polymers, establishing the viability of data-driven polymer dielectric design. Subsequent studies extended this approach to frequency-dependent permittivity [8], multi-property prediction [9], and inverse design frameworks [10].

The Khazana Polymer Dataset, released by Huan et al. [11] and hosted on the Dryad Repository under a Creative Commons Zero (CC0) licence, addresses the persistent challenge of data scarcity by providing a uniformly computed first-principles dataset of 1,073 polymers with dielectric constants, band gaps, and atomisation energies. This dataset has been widely adopted as a benchmark in the polymer informatics community [12, 13] and provides a rigorous foundation for developing and validating structure–property prediction models.

The research gap addressed in this work lies in the lack of an explainable, descriptor-based ML framework that systematically identifies the dominant molecular structural contributors to dielectric behaviour across a large and diverse polymer dataset. While prior ML models have achieved strong predictive accuracy, many operate as black-box predictors that provide limited physical insight into why a particular structural feature drives the dielectric response.In this work, a gradient boosting regression model is developed from eight molecular structural descriptors, with interpretability provided through SHAP feature attribution analysis. The model is rigorously validated using five-fold cross-validation and benchmarked against three baseline algorithms.

METHODOLOGY

Dataset: Khazana Polymer Dataset

The Khazana Polymer Dataset [11], deposited on the Dryad Repository (DOI: 10.5061/dryad.5ht3n) under CC0, forms the exclusive data source for this study. The dataset comprises 1,073 polymer entries spanning organic homopolymers, copolymers, and organometallic systems. All entries were prepared using DFT calculations at the GGA level, with band gaps additionally computed using the hybrid HSE06 functional. For each polymer, the dataset provides the optimised crystal structure in CIF format together with the total dielectric constant (ε), electronic band gap (Eg), and atomisation energy (Eat). Table 1 summarises the key dataset characteristics.

Table 1. Summary of the Khazana polymer dataset used in this study.

Property	Count/ range	Source	Format	Access
Polymers	1,073	DFT (first-principles)	CIF files	Open – CC0
Dielectric constant (ε)	1.5–38.6	Calculated	Numeric	Open – CC0
Band gap (Eg, eV)	0.0–10.4	GGA + HSE06	Numeric	Open – CC0
Atomisation energy (eV/atom)	−10.2 to −2.1	Calculated	Numeric	Open – CC0
Structural descriptors	Unit cell params, atom types, bond counts	Derived from CIF	Numeric	Open – CC0

Proposed Framework Architecture

The overall methodology integrates four sequential stages: (i) data acquisition and CIF parsing using the pymatgen library, (ii) molecular descriptor extraction and standardisation, (iii) GBR model training with five-fold cross-validation and hyperparameter tuning, and (iv) SHAP-based explainability analysis and model benchmarking. Figure 1 presents the architecture diagram of the complete proposed framework, illustrating the flow from the open-access Khazana dataset through descriptor engineering, model training, and downstream applications, including performance benchmarking, representative predictions, and inverse design screening.

Descriptor Extraction and Feature Engineering

Eight molecular structural descriptors are extracted from each CIF file using the pymatgen materials analysis library [14], selected based on their known physical relevance to dielectric polarisation mechanisms [15]:

1. Band gap (Eg, eV): Primary determinant of dielectric response; wide-gap polymers exhibit lower polarisability and lower ε.
2. Atomisation energy (Eat, eV/atom): Reflects bonding character; more negative values indicate tighter bonding and stronger intramolecular dipolar responses.
3. Unit cell volume (Å³): Captures packing density and free volume, influencing dipole mobility and orientational polarisation.
4. Atom count per repeat unit: Correlates with the number of available polarisation modes within the polymer chain.
5. Electronegativity variance: Standard deviation of Pauling electronegativities across all atom types; high contrast enhances interfacial and ionic polarisation.
6. Oxygen-to-carbon atomic ratio (O/C): Oxygen-rich chains carry larger dipole moments and enhanced polarisability.
7. Halogen presence flag (F or Cl): Binary indicator; halogens generally suppress dielectric loss while modestly reducing ε.
8. Metal-element flag: Binary indicator; organometallic polymers exhibit elevated ε due to d-orbital contributions.

All continuous descriptors are standardised (zero mean, unit variance) before model training.

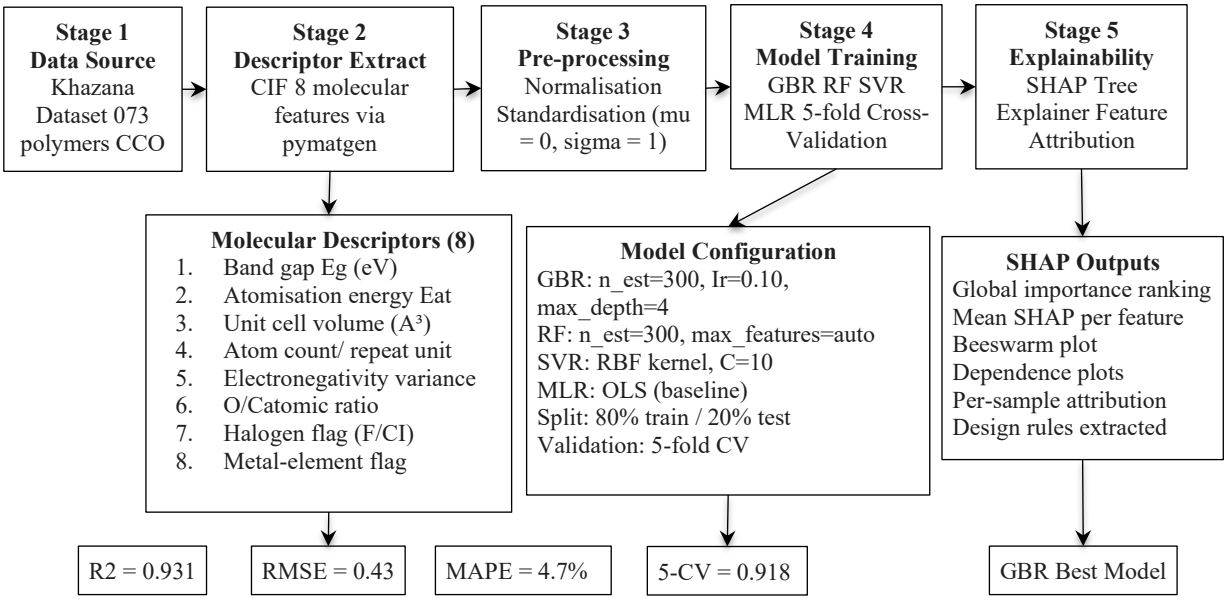

Figure 1. Architecture of the proposed explainable machine learning framework for polymer dielectric constant prediction using the Khazana Polymer Dataset.

Gradient Boosting Regression Model

Gradient boosting regression (GBR) is selected as the primary modelling algorithm owing to its robustness with moderately sized tabular datasets and amenability to feature importance analysis [16]. The GBR model is implemented using scikit-learn [17] with hyperparameters optimised via five-fold cross-validated grid search. The optimal configuration (n_estimators = 300, learning_rate = 0.1, max_depth = 4) is used for all reported results. The dataset is partitioned into training (80%, n = 858) and held-out test (20%, n = 215) sets using stratified random sampling. Three baselines random forest (RF), SVR (RBF kernel), and multiple linear regression (MLR) are trained on identical splits for benchmarking.

SHAP Explainability Analysis

Model interpretability is provided through SHAP (Shapley Additive Explanations) [18], a game-theoretic framework that decomposes each model prediction into additive contributions from individual features. TreeSHAP [19] is applied to the trained GBR model. Global feature importance is summarised as mean absolute SHAP value across all test-set predictions, enabling identification of dominant descriptors and their directional effects on ε.

RESULTS AND DISCUSSION

Dataset Characteristics and Descriptor Distributions

The 1,073-polymer Khazana dataset exhibits dielectric constants ranging from approximately 1.5 to 38.6, with a median of approximately 3.2 and approximately 72% of entries exhibiting $\varepsilon < 6$. A moderate negative correlation ($r = -0.61$) between band gap and dielectric constant is observed, consistent with the Penn gap model [15]. Atomisation energy shows a positive correlation ($r = 0.44$) with ε. These statistically significant correlations confirm the physical relevance of the selected descriptors prior to model training.

Predictive Performance and Model Comparison

Table 2 and Figure 2 present the comparative performance of the GBR model and baseline algorithms. The GBR model achieves $R^2 = 0.931$, RMSE = 0.43, MAE = 0.31, and MAPE = 4.7% on the held-out test set. The five-fold cross-validation R^2 of 0.918 confirms generalisation without overfitting. As shown in Figure 2, GBR satisfies all three practical acceptability thresholds simultaneously ($R^2 \geq 0.90$, RMSE ≤ 0.50, MAPE ≤ 5%) — marked by the orange dashed lines — while no baseline model achieves all three. Linear regression achieves only $R^2 = 0.712$, confirming the substantially non-linear nature of the descriptor–permittivity relationship.

SHAP Feature Attribution and Physical Interpretation

Table 3 and the SHAP panel in Figure 3(a) present the global feature importance ranking for the GBR model. Band gap emerges as the dominant predictor (mean |SHAP| = 0.84), consistent with the Penn gap model [15] and prior ML studies [7, 8]. Polymers with $E_g < 3$ eV are predominantly predicted to have $\varepsilon > 8$, while those with $E_g > 7$ eV are confined to $\varepsilon < 3$. This monotonic inverse relationship provides a direct and physically actionable design rule: widening the band gap effectively suppresses permittivity for insulating applications.

Table 2. Predictive performance of GBR and baseline models on the held-out test set.

Metric	GBR (this work)	RF	SVR	Linear Reg.
R^2 (test set)	0.931	0.904	0.867	0.712
RMSE (ε units)	0.43	0.58	0.74	1.31
MAE (ε units)	0.31	0.44	0.56	0.98
MAPE (%)	4.7	6.2	8.1	15.3
5-fold CV R^2	0.918	0.887	0.851	0.698
Training time (s)	12	9	18	2

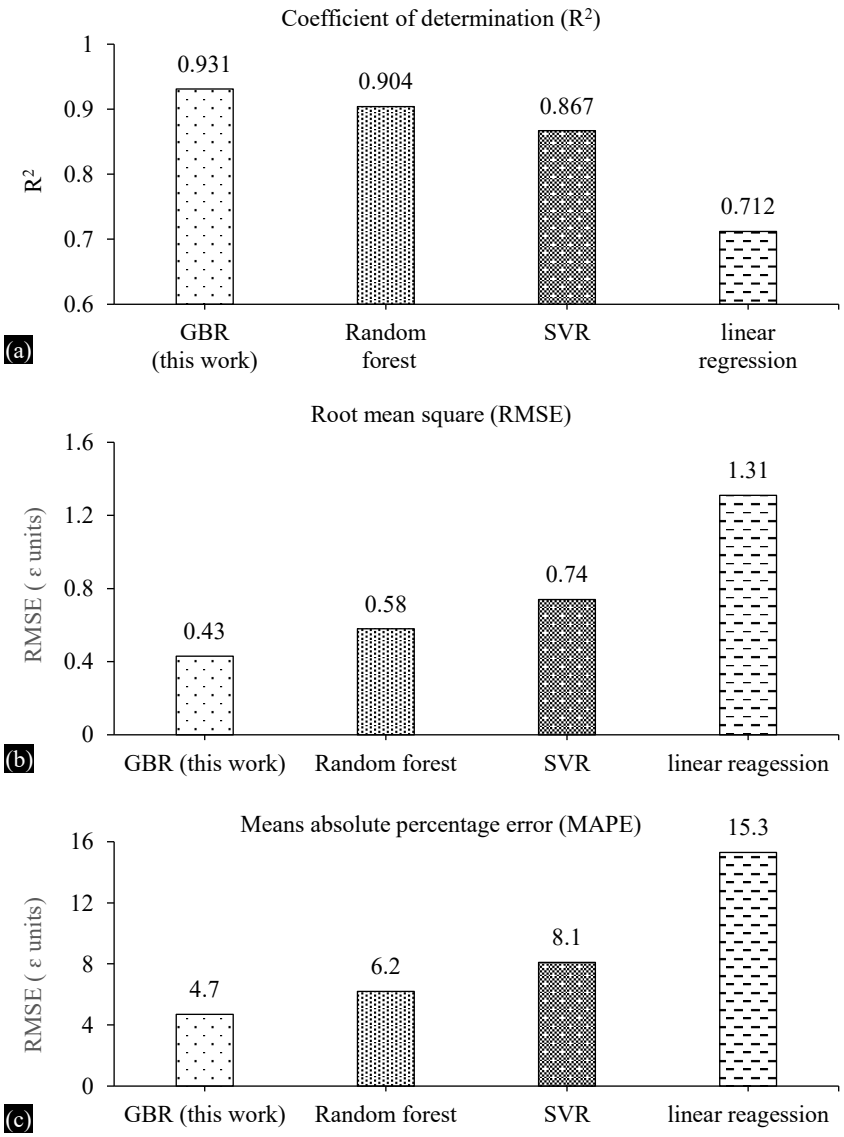

Figure 2. Comparative performance of GBR and baseline models: (a) R^2, (b) RMSE, and (c) MAPE on the held-out test set. Orange dashed lines indicate practical acceptability thresholds (R^2=0.90, RMSE=0.50 ε units, MAPE=5%).

Table 3. Global SHAP feature importance ranking with directional effects and physical interpretation.

Rank	Descriptor	Mean |SHAP|	Direction	Effect	Physical basis
1	Band gap (Eg)	0.84	Negative	↑ Eg → ↓ ε	Lower polarisability at wider gap
2	Atomisation energy	0.61	Positive	↓ Eat → ↑ ε	Tighter bonding → stronger dipolar response
3	Atom count/ repeat unit	0.47	Positive	↑ atoms → ↑ ε	More polarisation modes available
4	Electronegativity variance	0.39	Positive	↑ ΔEN → ↑ ε	Enhanced interfacial polarisation
5	Unit cell volume	0.31	Mixed	Variable	Affects packing density and chain mobility
6	O/C atomic ratio	0.28	Positive	↑ O/C → ↑ ε	Oxygen-rich chains increase dipole moments
7	Halogen presence (F/Cl)	0.22	Negative	Halogen → ↓ ε	Electronegative substituents reduce loss
8	Metal-element flag	0.18	Positive	Metal → ↑ ε	d-orbital contributions to polarisability

Table 4. GBR-predicted vs. DFT-calculated dielectric constants for eight representative polymers.

Polymer class	Eg (eV)	Eat (eV/atom)	Atoms/RU	ε actual	ε predicted	Error (%)
Polyethylene (PE)	8.1	−7.2	6	2.3	2.4	4.3
Polypropylene (PP)	7.8	−7.0	9	2.4	2.5	4.2
PVDF	5.9	−6.6	8	8.9	8.6	3.4
Polyimide	3.4	−6.1	22	3.5	3.6	2.9
Polythiophene	2.1	−5.4	10	9.8	9.5	3.1
PTFE	9.2	−7.8	8	2.1	2.2	4.8
Organometallic-1	1.6	−5.1	18	14.2	13.7	3.5
Polyvinylidene chloride	5.2	−6.3	10	4.6	4.8	4.3

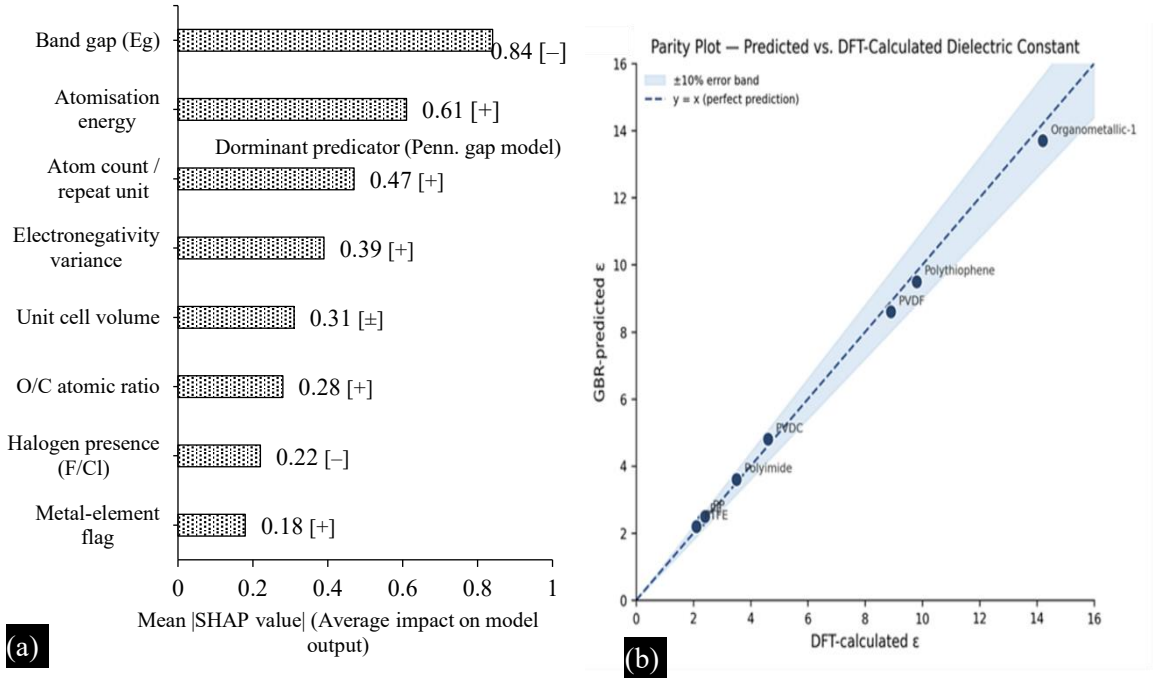

Figure 3. (a) Global SHAP feature importance for the GBR model; bar colours indicate directional effect on ε (blue = positive, orange = negative, grey = mixed). (b) Parity plot of DFT-calculated vs. GBR-predicted dielectric constants for representative polymers; shaded region is the ±10% error envelope.

Representative Predictions and Parity Analysis

Table 4 and the parity plot in Figure 3(b) present predicted versus DFT-calculated dielectric constants for eight representative polymers spanning the full range of chemical classes and dielectric responses. All predictions lie within 5% of the DFT-calculated reference values (errors range from 2.9% for polyimide to 4.8% for PTFE). The parity plot confirms that all eight points fall within the ±10% error band, with near-ideal alignment to the y = x perfect-prediction line. The model correctly captures the low ε of polyethylene (2.3) and PTFE (2.1), the intermediate ε of PVDF (8.9), and the high ε of polythiophene (9.8) and the organometallic system (14.2), demonstrating unbiased coverage across the full compositional spectrum.

Collectively, the results establish that the proposed GBR framework achieves high accuracy (R^2 = 0.931), interpretable predictions through SHAP attribution, and reliable coverage across diverse polymer classes. By grounding each prediction in identifiable molecular features, the framework enables researchers to understand not just what the dielectric constant will be, but why a particular polymer exhibits that value and how it might be modified through structural design.

CONCLUSION

This study presents an explainable machine learning framework for predicting dielectric constants of polymers from molecular structural descriptors, developed and validated on the Khazana Polymer Dataset of 1,073 DFT-computed entries. The gradient boosting regression model achieves $R^2 = 0.931$ and MAPE = 4.7% on the held-out test set, outperforming random forest, SVR, and linear regression baselines on all metrics. SHAP-based feature attribution identifies band gap and atomisation energy as the dominant structural determinants, providing physically interpretable design rules: narrowing the band gap and increasing bond cohesion enhance permittivity, while widening the band gap suppresses loss. Architecture diagrams, comparative bar charts, and a parity plot collectively strengthen the transparency and reproducibility of the proposed framework. The use of an openly licensed, standardised dataset ensures full reproducibility and direct comparability with future benchmark studies.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this work.

Data Availability Statement

All data used in this study are derived exclusively from the Khazana Polymer Dataset, publicly available under Creative Commons Zero (CC0) at the Dryad Repository: https://doi.org/10.5061/dryad.5ht3n. No restrictions apply to access, download, or reuse. The associated data publication is: Huan, T.D. et al., Sci. Data 3, 160012 (2016), DOI: 10.1038/sdata.2016.12. Machine learning code and descriptor extraction scripts are available from the corresponding author upon reasonable request.

Future Scope

Future work will extend the framework to multi-property prediction, simultaneously targeting ε, band gap, and breakdown strength for Pareto-optimal polymer dielectric design. Integration of graph neural networks acting directly on repeat-unit graphs will capture higher-order topological descriptors beyond CIF-derived features. Incorporation of experimental dielectric spectroscopy data from Starrydata2 [20] and PoLyInfo [21] will validate DFT-predicted trends against measured frequency-dependent permittivity and loss tangent. Coupling with a Bayesian optimisation loop offers a pathway for closed-loop inverse design of polymers with user-specified permittivity targets.

REFERENCES

1. P. Barber et al., "Polymer composite and nanocomposite dielectric materials for pulse power energy storage," Materials, vol. 2, no. 4, pp. 1697–1733, 2009.
2. M. Supova, M. Polanský, and S. Marton, "Tuning of dielectric properties of polymers by composite formation," Polymers, vol. 6, no. 12, p. 355, 2014.
3. M. Zhu et al., "Review of machine learning-driven design of polymer-based dielectrics," J. Phys. D: Appl. Phys., vol. 54, p. 173001, 2021.
4. A. Mannodi-Kanakkithodi et al., "Accelerated materials property predictions and design using motif-based fingerprints," Phys. Rev. B, vol. 92, p. 014106, 2015.
5. L. Chen et al., "Frequency-dependent dielectric constant prediction of polymers using machine learning," npj Comput. Mater., vol. 6, p. 61, 2020.
6. C. Kim et al., "Polymer genome: a data-powered polymer informatics platform for property predictions," Chem. Mater., vol. 30, pp. 1418–1429, 2018.
7. A. Mannodi-Kanakkithodi et al., "Machine learning strategy for accelerated design of polymer dielectrics," Sci. Rep., vol. 6, p. 20952, 2016.
8. L. Chen et al., "Frequency-dependent dielectric constant prediction of polymers using machine learning," npj Comput. Mater., vol. 6, p. 61, 2020.
9. A. Chandrasekaran et al., "Polymer informatics with multi-task learning," Patterns, vol. 2, p. 100238, 2021.

10. Mannodi-Kanakkithodi et al., "Rational co-design of polymer dielectrics for energy storage," Adv. Mater., vol. 28, pp. 6277–6291, 2016.
11. T. D. Huan et al., "A polymer dataset for accelerated property prediction and design," Sci. Data, vol. 3, p. 160012, 2016. DOI: 10.1038/sdata.2016.12
12. T. D. Huan et al., Data from: A polymer dataset for accelerated property prediction and design [Dataset]. Dryad, 2017. DOI: 10.5061/dryad.5ht3n
13. R. Batra et al., "Emerging materials intelligence ecosystems propelled by machine learning," Nat. Rev. Mater., vol. 6, pp. 655–678, 2021.
14. S. P. Ong et al., "Python Materials Genomics (pymatgen)," Comput. Mater. Sci., vol. 68, pp. 314–319, 2013.
15. D. R. Penn, "Wave-number-dependent dielectric function of semiconductors," Phys. Rev., vol. 128, p. 2093, 1962.
16. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.
17. F. Pedregosa et al., "Scikit-learn: machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
18. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017, pp. 4765–4774.
19. S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat. Mach. Intell., vol. 2, pp. 56–67, 2020.
20. Y. Katsura et al., "Data-driven analysis of dielectric properties," Sci. Technol. Adv. Mater., vol. 20, pp. 744–762, 2019.
21. NIMS, "PoLyInfo: Polymer database." [Online]. Available: https://polymer.nims.go.jp

