Nanoscale

PAPER

View Article Online
View Journal | View Issue

Cite this: *Nanoscale*, 2026, **18**, 2613

Prediction of the low-temperature properties of electrolyte solvents for lithium-ion batteries *via* machine learning

Jiechen Guo,[a,b] Yifan Chai,[a,b] Cancan Hong,[a,b] Hao Liu,[a,b] Lijing Xie,[a] Tianle Wang,[a,b] Jingpeng Chen,[a] Ge Song,[a] Zonglin Yi*[a] and Fangyuan Su *[a]

Received 18th September 2025,
Accepted 22nd December 2025

DOI: 10.1039/d5nr03942h

rsc.li/nanoscale

Electrolytes with low melting points (MPs), high boiling points (BPs), and high dielectric constants (ε) can effectively mitigate performance degradation in lithium-ion batteries (LIBs) under low-temperature conditions. However, the lack of systematic experimental data on electrolyte properties poses a significant challenge to traditional design approaches. To address this limitation, we developed a machine learning workflow that integrates data acquisition using large language models, model construction, and interpretability analysis, aiming to predict key molecular properties, with a focus on MPs, BPs and ε. We constructed a multi-source database, LiElectroDB, that contains over 150 000 electrolyte molecules relevant to LIBs. The prediction models demonstrate strong performance across all three properties, achieving an R^2 of 0.8864 and a root mean square error (RMSE) of 23.3 K for the MP, a coefficient of determination (R^2) of 0.9608 and an RMSE of 14.3 K for the BP using the XGBoost algorithm, and an R^2 of 0.8718 and a RMSE of 6.7 for ε using an artificial neural network. To further uncover structure–property relationships, t-SNE and SHAP are employed to analyze the molecular features contributing to thermal behavior at a microscopic level. Finally, by integrating molecular neighborhood search with high-throughput screening, nine candidate molecules are identified as promising low-temperature electrolytes for LIBs. This work provides an efficient and generalizable framework for the design of low-temperature electrolytes in LIBs.

Introduction

Lithium-ion batteries (LIBs) have been widely applied as promising electrochemical energy storage materials due to their high energy density, long cycle life, and the absence of the memory effect.[1,2] Unfortunately, the practical application of LIBs in extremely cold environments is limited by their unsatisfactory energy retention at low temperature, owing to the deterioration of the electrolytes, which severely hinders the migration of lithium ions.[3,4] The optimization of electrolytes is considered a crucial strategy to enhance the low-temperature performance of LIBs.[5–7] The ideal low-temperature electrolyte solvents should exhibit the following properties: (i) low melting points (MPs) to enhance system fluidity;[8] (ii) high dielectric constants (ε) to promote the dissolution of lithium salts;[9,10] and (iii) medium-to-high boiling points (BPs) to suppress the consumption of electrolytes and prevent thermal runaway.[11] However, the lack of systematic experimental data on electrolyte properties poses a significant challenge to trial-and-error design approaches.

Recently, machine learning (ML) has emerged as a powerful tool to accelerate material design and property prediction. In the field of LIBs, high-throughput experimentation and data-driven modelling enable the development of key components, such as cathodes,[12,13] anodes,[14,15] and electrolytes.[16–18] The application of molecular simulation and ML to screen ideal electrolytes facilitates the advancement of next-generation low-temperature LIBs. In recent years, various ML algorithms have been employed—such as random forests, support vector machines, and neural networks—to predict key electrolyte properties including ionic conductivity,[18–20] the electrochemical stability window,[21–23] and viscosity.[24–26] However, these purely data-driven approaches often suffer from limited generalization ability and the lack of interpretability. To address these limitations, researchers have proposed the knowledge–data dual-driven Knowledge-based Property prediction Integration (KPI) framework to predict the critical thermophysical properties of electrolyte molecules including melting points (MPs), boiling points (BPs), and flash points (FPs).[27] KPI is specifically designed for applications requiring a wide temperature range and high-safety battery operation. KPI integrates expert domain knowledge with data-driven learning. This integration not only enhances predictive accuracy but also significantly improves model interpretability and generalization.

[a]*Shanxi Key Laboratory of Carbon Materials, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan, 030001, China. E-mail: sufangyuan@scxicc.ac.cn*
[b]*University of Chinese Academy of Sciences, Beijing 100049, China*

Leveraging molecular neighbourhood search and high-throughput virtual screening, KPI successfully identified 29 promising electrolyte candidates with desirable safety and thermal properties, demonstrating its effectiveness in accelerating electrolyte discovery. However, due to the significant differences in the datasets used across various studies, the existing models still exhibit insufficient generalization capabilities across different datasets or under specific operating conditions, such as low-temperature environments (Fig. 1a).

In this work, we propose a modelling method that integrates multi-level chemical knowledge to achieve highly precise prediction of the key physical properties of low-temperature electrolytes, including MPs, BPs and ε. Firstly, we constructed a LIB electrolyte database (LiElectroDB), a structurally diverse electrolyte database containing more than 150 000 molecules, assembled through multi-source data integration and LLM-assisted extraction. To fulfil the modelling requirements of different properties, XGBoost[28] (XGB) and neural network-based modelling strategies are designed to uncover the relationships between molecular structures and target properties. By embedding multi-level chemical knowledge including chemical composition, structural characteristics, and electronic descriptors, the models not only achieve high predictive accuracy but also exhibit enhanced interpretability. Additionally, by combining molecular neighbourhood search with high-throughput screening strategies, nine promising molecules are successfully identified as low-temperature electrolyte candidates for LIBs. This workflow not only estab-

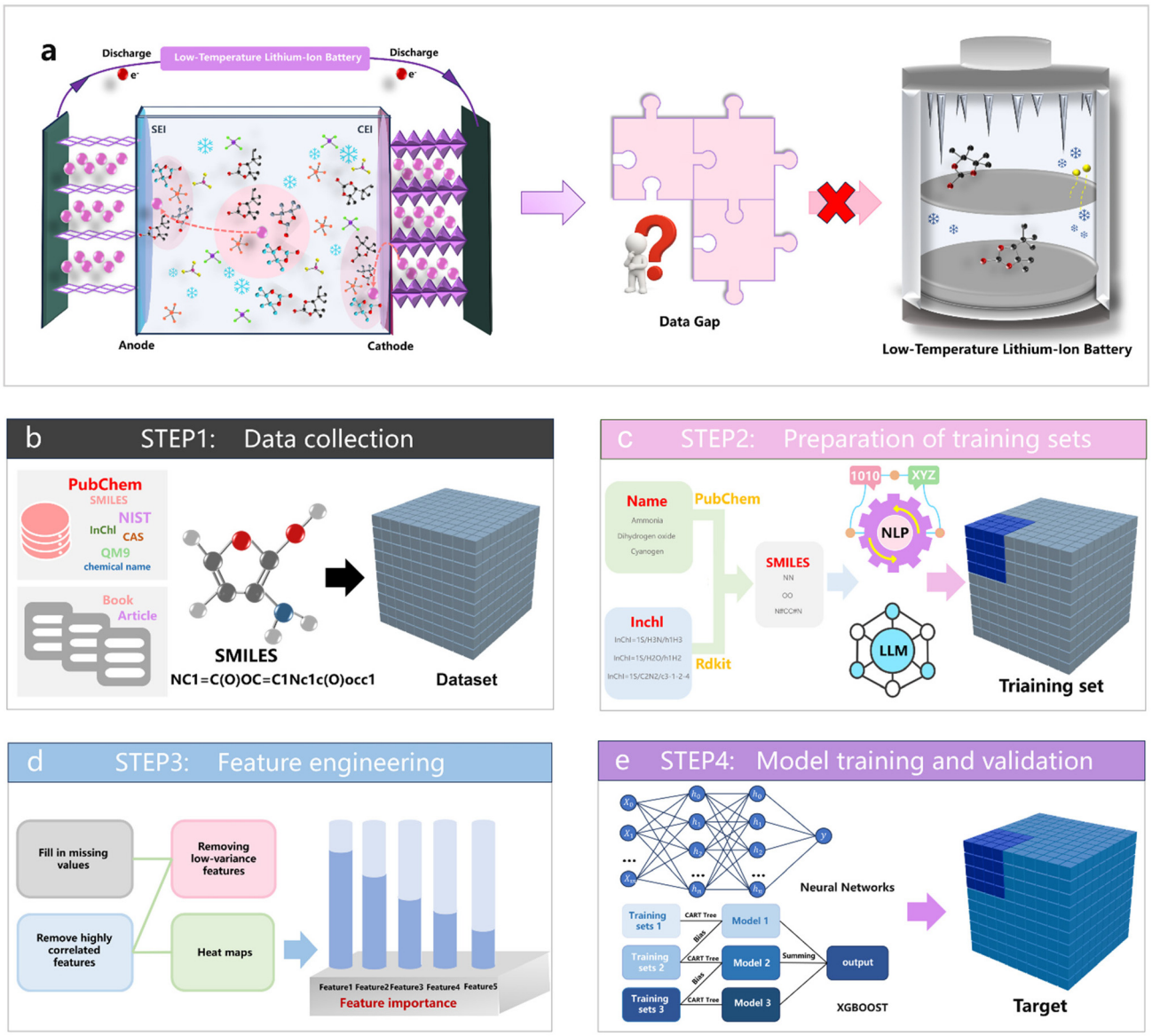

Fig. 1 Overall workflow of machine learning. (a) Background. (b) Data collection. (c) Preparation of training sets. (d) Feature engineering. (e) Model training and validation.

lishes a novel and scalable paradigm for electrolyte discovery but also demonstrates significant potential to accelerate molecular screening, reduce experimental costs, and provide actionable insights to guide future electrolyte design.

Results and discussion

Workflow overview

This study presents a framework for molecular property prediction, supported by the construction of a high-quality molecular database, LiElectroDB, integrated from multiple data sources. LiElectroDB is built with QM9 serving as the core dataset and is further supplemented with literature data and external databases, such as PubChem[29] and the National Institute of Standards and Technology (NIST)[30] (Fig. 1b). Additionally, an LLM (GPT-4-turbo) is utilized to assist in data acquisition (Fig. 1c), leading to a multi-source dataset containing over 150 000 unique molecules. To ensure model stability and generalization, strategies such as data augmentation and feature engineering are introduced (Fig. 1d). Several ML models based on various algorithms are systematically evaluated. Through five-fold cross-validation, the performance of each model is compared on the same task. The XGB-based model is ultimately selected for predicting MPs and BPs, while an artificial neural network (ANN)[31] is used for predicting ε (Fig. 1e). t-Distributed stochastic neighbour embedding (t-SNE)[32] and Shapley additive explanations (SHAP) analyses are further applied to provide interpretability. By integrating molecular neighbourhood exploration with high-throughput screening, nine candidate molecules are identified as promising low-temperature electrolytes for LIB applications. This workflow covers data construction, preprocessing, model optimization, and result interpretation, with the goal of offering reliable data support and interpretable modelling strategies for molecular property prediction.

Dataset construction

The study begins with the QM9[33] dataset, which contains *ab initio* properties of approximately 134 000 molecules, and is further extended to include common organic solvents and representative inorganic compounds based on literature data[23,34–37] and external databases. Simplified molecular input line entry system (SMILES)[38] strings are used as unique identifiers, and the PubChem API is employed to standardize molecular structures and eliminate duplicates. LLMs are also used to assist in extracting data from textual sources, with all retrieved information manually verified. This hybrid approach enables efficient scaling while maintaining data accuracy. Ultimately, the multi-source fusion database LiElectroDB is constructed (Fig. 1b), providing a structurally diverse foundation for downstream modelling.

To build a high-quality training set, we obtain property parameters such as MP, BP, and ε from authoritative sources, including the NIST and PubChem. During data collection, SMILES strings serve as a unified primary key to retrieve structured chemical information *via* the PubChem API. An LLM (GPT-4-turbo) is further applied to extract physical properties —MPs, BPs, and ε—for molecules in LiElectroDB (Fig. 1c). All extracted data are manually annotated to ensure accuracy and serve as the basis for subsequent model training and evaluation. Information from SMILES strings is integrated as the primary input, and three types of molecular characteristics are additionally extracted using the RDKit[39] in Python 3.11.7, including chemical composition, molecular structure, and electronic structure (Fig. 2a and Tables S1–S3).

To ensure physical plausibility, chemical diversity, and model stability, we restrict the molecular weight to 0–300 and the number of heavy atoms to 0–30 (Fig. S1). The final dataset includes 1251 MP data points, 1502 BP data points, and 895 ε data points. The resulting chemical space primarily consists of organic molecules containing C, H, O, N, S, and P, with diverse linear, cyclic, and aromatic scaffolds and a broad distribution of polar functional groups (ethers, carbonyls, amines, nitriles, sulfones, and alcohols). The property ranges—MP (50–450 K), BP (300–650 K), and ε (1–60)—are well aligned with electrolyte-relevant regimes, supporting reliable model training within the applicability domain.

We note that certain sparsely sampled regions—such as highly fluorinated systems, large polycyclic aromatics, and molecules containing fewer common heteroatoms—carry higher predictive uncertainty. Molecules with extremely high polarity or very limited conformational flexibility are likewise under-represented. These constitute the limitations of our mode. Overall, the curated structural and property diversity provides a robust basis for generalizable and physically meaningful predictions across conventional organic electrolyte candidates.

Kernel Density Estimation (KDE)[40] analysis is conducted on the numerical distributions of MP, BP, and ε to verify the rational division and representativeness of the dataset. The MP and BP datasets are randomly divided into training and test sets at a ratio of 4 : 1, while the ε dataset is randomly divided into training, validation, and test sets at a ratio of 8 : 1 : 1 (Fig. S2–S4). The distribution curves of each subset are generally consistent with the overall dataset, showing good uniformity. These differences are slightly reflected in the smoothness and local peaks of the curves, but the overall differences are minimal, indicating that the data partitioning process effectively retains the representativeness and diversity of the original data and avoids oversampling or bias. This consistency is essential in data modelling and model evaluation, ensuring that model training, validation, and testing processes are conducted under similar distribution conditions, thereby enhancing the robustness and generalization ability of the model.

Feature engineering

A systematic feature preprocessing pipeline is implemented, which mainly includes three steps: missing value imputation, removal of low-variance features, and elimination of highly correlated features, aiming to enhance training efficiency and model prediction performance.

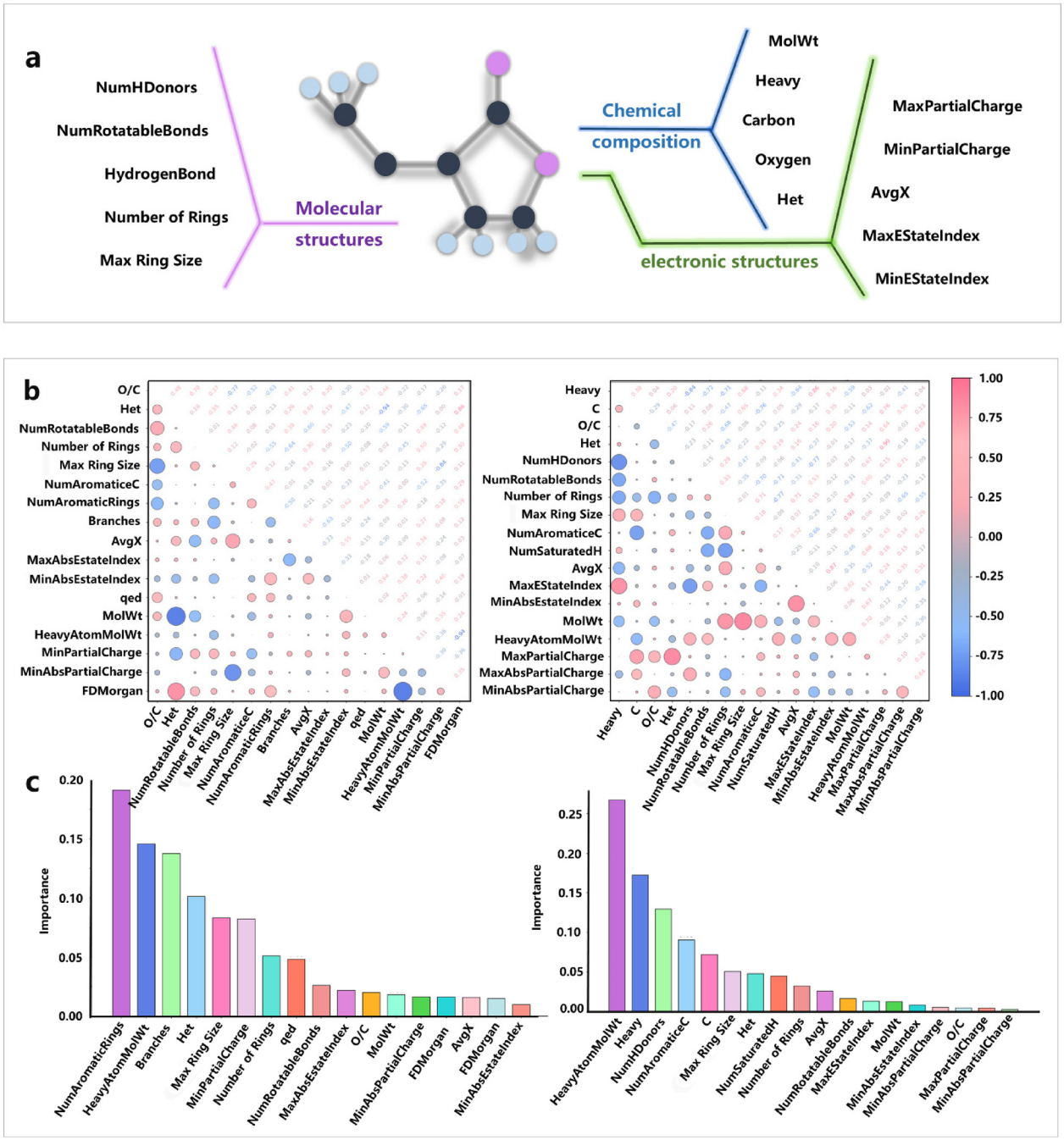

Fig. 2 Feature engineering. (a) Feature extraction. (b) Pearson correlation heatmaps. (c) Feature importance ranking based on XGBoost.

First, missing values in the dataset are detected and processed. Since missing values are inevitable in real-world data collection, using the data directly for modelling affects the accuracy and stability of the model. Therefore, the mean imputation method is applied to fill missing values in numerical features. Specifically, the SimpleImputer class from the sklearn.impute module[41] is employed to replace missing values in each column with the mean of the available values.

Next, low-variance feature selection is performed on the completed dataset. In modelling, low-variance features typically fluctuate minimally across samples and provide limited explanatory power for the target variable, potentially introducing redundant information. Based on this, the VarianceThreshold class from the sklearn.feature_selection module is applied to remove features with variance below 0.01. This step helps to reduce feature dimensionality and improve both training speed and model generalization.

Furthermore, to address potential multicollinearity among features, we computed the Pearson correlation coefficients for all feature pairs and visualized the results using a correlation

heatmap (Fig. S5–S10). Feature pairs with absolute correlation coefficients exceeding a predefined threshold were considered strongly correlated. In such cases, only one feature from each pair was retained for subsequent modelling. This procedure was adopted to prevent highly correlated features from jointly influencing the model, which could lead to unstable parameter estimates and potential overfitting. Fig. 2b shows Pearson correlation heatmaps of the engineered features.[42]

Finally, feature importance is assessed using a preliminary XGB model (Fig. S11–S13). Importance scores are derived from the internal gain-based metric, which quantifies the total information gain a feature contributes across all tree splits during model construction. Features with low importance scores are removed to reduce dimensionality and mitigate noise, except in the case of the ANN model, where such importance-based filtering is not directly applicable due to its non-tree-based architecture. Based on the above-mentioned feature engineering process, a representative subset of core features is selected as the final feature set (Fig. 2c), resulting in 17 features for melting-point prediction, 18 features for boiling-point prediction, and 61 features for dielectric constant prediction.

Model training and validation

During the training process, all input features were standardized using the StandardScaler from the sklearn.preprocessing module. Eight ML regression models are evaluated, including Decision Tree Regressor (DT),[43] Bagging Regressor (BG),[44] Random Forest Regressor (RF),[45] Extra Trees Regressor (ET),[46] AdaBoost Regressor (ADBR),[47] Gradient Boosting Regressor (GBR),[48] XGB, and ANN (Fig. S14–S40). All models underwent hyperparameter tuning *via* grid search. Using 5-fold cross-validation, we systematically compared the performance of each model on different prediction tasks (MPs, BPs, and ε). The main evaluation metrics include the root mean square error (RMSE) and the coefficient of determination (R^2). In the MP and BP prediction tasks, tree-based ensemble models, particularly XGB, are well adapted to such combined feature sets. They exhibit robustness to feature scaling variations, effective handling of nonlinear interactions, and stable performance against noisy or partially correlated descriptors. Consistently, XGBoost achieved the lowest RMSE and highest R^2 among all tree-based models in MP and BP cross-validation (Tables 1 and 2), validating its suitability for these thermophysical properties.

The XGB regression model is optimized to improve the accuracy of MP and BP predictions. Key hyperparameters are fine-tuned using grid search combined with five-fold cross-validation (Fig. S11 and S20). Model performance is evaluated using cross-validated RMSE and R^2 scores. The MP model achieves an R^2 of 0.8868 and an RMSE of 23.3 K under five-fold cross-validation (Table 1). The BP model further improves upon this, reaching an R^2 of 0.9608 and an RMSE of 14.3 K (Table 2). ε depends more strongly on electronic descriptors—including dipole-related features, partial charge distributions, and energy-state indices—which introduce complex higher-order nonlinear relationships. The ANN exhibits superior capability in capturing such intricate nonlinear mappings within high-dimensional spaces, thus achieving the highest predictive accuracy among all evaluated models; specifically, the ANN model achieves an R^2 of 0.8863 and an RMSE of 6.7 (Table 3). Upon implementing the Keras framework, the network adopts a simple yet effective architecture with two fully connected hidden layers and a single output node. The input layer receives a 61-dimensional feature vector generated from molecular features, covering the electronic properties, molecular structure, and chemical composition. Each hidden layer contains 300 neurons with ReLU activation to enhance non-linear representation and mitigate vanishing gradients. To improve

Table 1 Performance comparison of regression models for MP prediction

	Training set		Test set		Test set (cross-validation)	
	R^2	RMSE (K)	R^2	RMSE (K)	R^2	RMSE (K)
DT	0.9028	21.60	0.7805	32.36	0.7664	33.46
BG	0.9368	17.41	0.8498	26.77	0.8402	27.67
RF	0.9795	9.92	0.8722	24.69	0.8642	25.51
ET	0.9530	15.02	0.8467	27.04	0.8432	27.41
ADBR	0.7329	30.46	0.7617	28.00	0.7017	37.81
GBR	0.9814	9.45	0.8546	26.34	0.8611	25.19
XGB	0.9795	9.91	0.8856	23.41	0.8864	23.30
ANN	0.9447	11.83	0.8193	22.90	0.8463	19.84

Table 2 Performance comparison of regression models for BP prediction

	Training set		Test set		Test set (cross-validation)	
	R^2	RMSE (K)	R^2	RMSE (K)	R^2	RMSE (K)
DT	0.9581	15.10	0.8998	21.54	0.8635	26.85
BG	0.9695	12.89	0.9316	17.79	0.9187	20.73
RF	0.9896	7.51	0.9461	15.80	0.9359	18.40
ET	0.9815	10.05	0.9403	16.62	0.9379	18.1094
ADBR	0.8077	27.52	0.7552	28.10	0.7938	33.03
GBR	0.9813	10.08	0.9440	16.10	0.9398	18.19
XGB	0.9787	10.78	0.9550	14.43	0.9608	14.30
ANN	0.9751	11.48	0.9323	18.46	0.9513	15.90

Table 3 Performance comparison of regression models for ε prediction

	Training set		Test set		Test set (cross-validation)	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
DT	0.5330	13.18	0.4715	13.14	0.5007	13.46
BG	0.8704	6.9415	0.7022	9.87	0.7421	9.67
RF	0.9687	3.41	0.7285	9.42	0.7826	8.88
ET	0.9725	3.20	0.5927	11.54	0.7662	9.21
ADBR	0.5878	12.38	0.3345	12.43	0.3471	15.12
GBR	0.9856	2.31	0.7538	8.97	0.7904	8.17
XGB	0.9090	5.81	0.6743	10.32	0.7428	8.95
ANN	0.9779	2.85	0.8870	7.91	0.8718	6.70

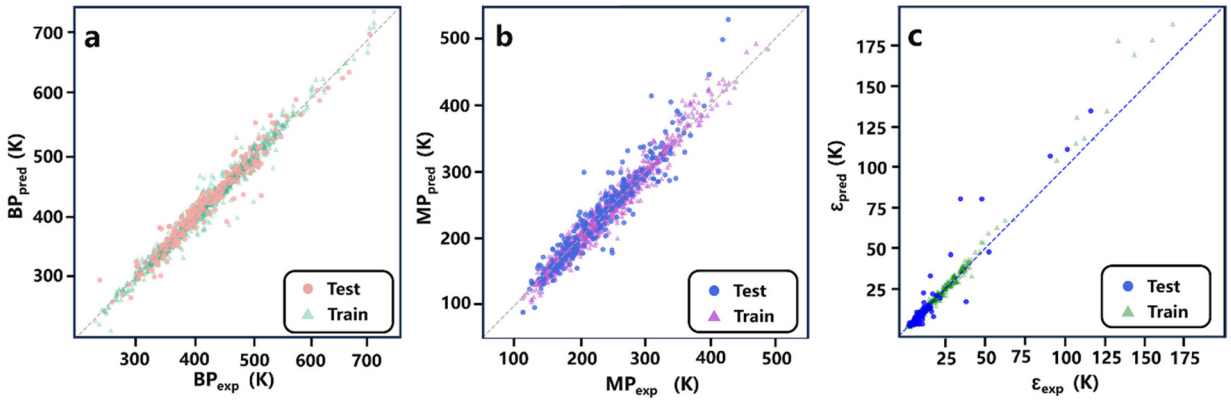

Fig. 3 Predicted *versus* actual values for (a) the MP, (b) the BP using the XGBoost algorithm, and (c) ε using an artificial neural network.

generalisation ability and prevent overfitting, each hidden layer is followed by a 10% dropout layer. Weight parameters are initialized using a random normal distribution, and L2 regularization (λ = 0.001) is applied to limit model complexity. The output layer consists of a single neuron. The model is trained using the MSE loss function and optimized with the Adam algorithm to ensure stable and efficient convergence. The training and validation loss curves show rapid initial convergence and sustained stability during training epochs, with minimal differences between training and validation losses (Fig. S39).

To further benchmark the model performance, a classical Group Contribution (GC) baseline was implemented using a Joback-type correlation. A validation subset of 47 structurally diverse molecules—covering linear, cyclic, aromatic, and heteroatom-containing species (O, N, S, and halogens), as well as functional groups such as ethers and carbonyls—was selected from the curated database. For each molecule, GC-estimated MP and BP were computed and compared against the corresponding experimental measurements and ML predictions. Across all three properties, the ML models consistently achieved significantly lower RMSE and MAE relative to the GC baseline, demonstrating their superior accuracy and generalizability (Tables S7 and S8) (Fig. 3).

Structure–property visualization

The t-SNE algorithm is employed for dimensionality reduction and visualization analysis, focusing on BPs, MPs, and ε. Initially, molecular structures are transformed into fingerprint vectors through numerical encoding, enabling the quantitative characterization of structural information. Subsequently, molecular structure labels (aromatic, linear, and cyclic) are incorporated into the analytical framework to guide the clustering process. Furthermore, violin plots are utilized to evaluate the distribution characteristics of different structural categories across various physicochemical dimensions. This comprehensive strategy not only facilitates an in-depth exploration of the intrinsic relationships between molecular structure and physicochemical properties but also effectively assesses the discriminative ability of molecular features for different structural classes.

In the t-SNE plot based on MPs and BPs (Fig. 4a and b), distinct clustering patterns emerge across the three structural types. Aromatic compounds form a compact cluster in the left region, reflecting the rigidity and symmetry conferred by aromatic rings and conjugated systems, which contribute to higher and more consistent MPs (BPs).[49,50] Linear compounds are dispersed across the right side of the plot with weak clustering, attributed to variations in chain length, branching, and functional groups, resulting in widely scattered thermal properties.[51,52] Cyclic compounds occupy the lower-central region, showing intermediate clustering behaviour. Although lacking aromatic stabilization, their ring-induced rigidity still imparts some packing regularity. It can also be observed that the boundaries between the three structural types are less distinct.

Aromatic compounds exhibit broader dispersion and overlap with cyclic compounds due to the influence of polar substituents (*e.g.*, hydroxyl and carboxyl groups), which enhance hydrogen bonding and increase the variability of MPs and BPs (Fig. 4a and b). Linear compounds remain on the right with increased density but fuzzy borders, reflecting the complex nonlinear interactions among chain length, polarity, and branching.[53] These observations suggest that while structure–property associations are evident, MPs and BPs offer limited discriminative power for structural classification in certain cases. Violin plots further support the observed clustering patterns, revealing significant differences in MP and BP distributions across structural classes (Fig. S41 and S42). Aromatic compounds exhibit narrow and high-centered distributions, reflecting the inherent rigidity and symmetry of their conjugated ring systems. In contrast, linear compounds show broad and multimodal distributions, indicative of substantial structural diversity and corresponding variability in thermal properties. Cyclic compounds exhibit intermediate behaviour in both spread and central tendency. Overall, the distribution of MPs and BPs is closely linked to molecular polarity and functional group composition. Strongly polar compounds (*e.g.*,

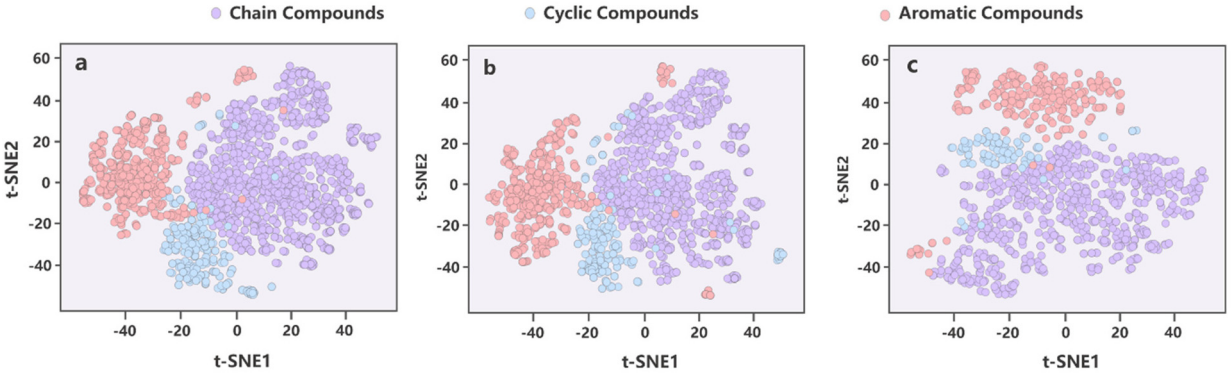

Fig. 4 t-SNE-based visualization of molecular distributions for the MP (a), BP (b), and ε (c).

carboxylic acids and phenols) tend to exhibit high MPs and BPs due to enhanced intermolecular interactions. Compounds of moderate polarity (*e.g.*, aldehydes and ketones) occupy an intermediate range, while nonpolar molecules (*e.g.*, hydrocarbons and ethers) cluster at the lower end of the spectrum. The inclusion of heteroatoms such as nitrogen and sulfur further amplifies polarity differences, broadening the overall property distribution.

In contrast, as for ε, the clustering distinction among molecular structural types is significantly weakened (Fig. 4c). The overall distribution is highly scattered, with substantial overlap across aromatic, linear, and cyclic compounds, and without the emergence of distinct boundaries. Aromatic compounds are loosely scattered in the upper region of the map, while linear and cyclic structures run throughout the entire region. This distribution reflects that ε is primarily influenced by electronic structure, especially charge distribution, conjugation effects, and polar substituents rather than by the molecular backbone or topology.[54–56] Linear compounds typically contain a wide range of polar functional groups and exhibit a particularly broad spread in ε, ranging from nonpolar alkanes ($\varepsilon \approx 1$) to highly polar amines and carboxylic acids ($\varepsilon > 30$). Cyclic compounds exhibit similar diffuse distributions, falling between the two extremes and contributing to the overall overlap. Violin plots further confirm this trend: oxygen-rich, highly polar compounds (*e.g.*, alcohols, carboxylic acids, and phenols) consistently show elevated ε (30–50), while nonpolar species such as aromatics and alkanes cluster in the low-permittivity region (ε = 1–3). Heteroatom-containing compounds (*e.g.*, amides and sulfones) exhibit wide variability due to their diverse polar characteristics (Fig. S43). These results indicate that ε primarily captures the electronic responsiveness of molecules, with limited correlation to structural symmetry or geometry.

Interpretability and knowledge discovery

We use the SHAP method to systematically analyse feature importance, aiming to enhance the interpretability of molecular MP and BP prediction models, as shown in Fig. 5a and b. Molecular features are divided into three categories: chemical composition, structural features, and electronic properties. The importance of all molecular features is analysed, and the features are visualized. The vertical axis reflects the distribution of molecules for each feature, while the horizontal axis indicates the contribution of feature values to the prediction results, enabling both local and global interpretations of the complex nonlinear models.

In the MP model (Fig. 5a), chemical composition features play a dominant role with heavy-atom molecular weight (HeavyAtomMolWt) emerging as the single most influential feature (mean |SHAP| = 0.26), while the impact of the number of heteroatoms (Het, 0.10) and the ratio of oxygen atoms to carbon atoms (O/C, 0.04) is also pronounced. Critically, the O/C ratio shows structure-dependent behaviour: in highly polar, oxygen-rich scaffolds, it tends to increase the MP, whereas in other contexts, it may exert a negative influence. For BP prediction (Fig. 5b), chemical-composition features exert an even more decisive influence: three of the five most predictive variables encode molecular size and elemental constitution, namely the number of heavy atoms (Heavy, 0.20), hydrogen-bond donor count (0.18), and HeavyAtomMolWt (0.15). Elevated HeavyAtomMolWt values are associated with positive SHAP contributions, implying that larger and heavier molecules are assigned higher predicted BPs. The analogous behaviour of hydrogen bond donor counts underscores the pivotal role of hydrogen bonding in enhancing cohesive interactions. This comparative analysis reveals that the BP is governed primarily by global molecular properties, whereas the MP is more sensitively modulated by specific functional group interactions.

In MP prediction (Fig. 5a), structural features account for ≈38.5% of explanatory power. The maximum number of atoms of the ring (Max Ring Size, 0.09) and the number of aromatic rings (NumAromaticRings, 0.08) confer moderate positive contributions, suggesting that extended ring systems and aromaticity enhance structural rigidity and thus the MP. The number of rotatable bonds (NumRotatableBonds, 0.05) shows a weaker negative contribution, consistent with the notion that melting involves the disruption of crystal packing, which is a process less sensitive to conformational flexibility than vaporization.

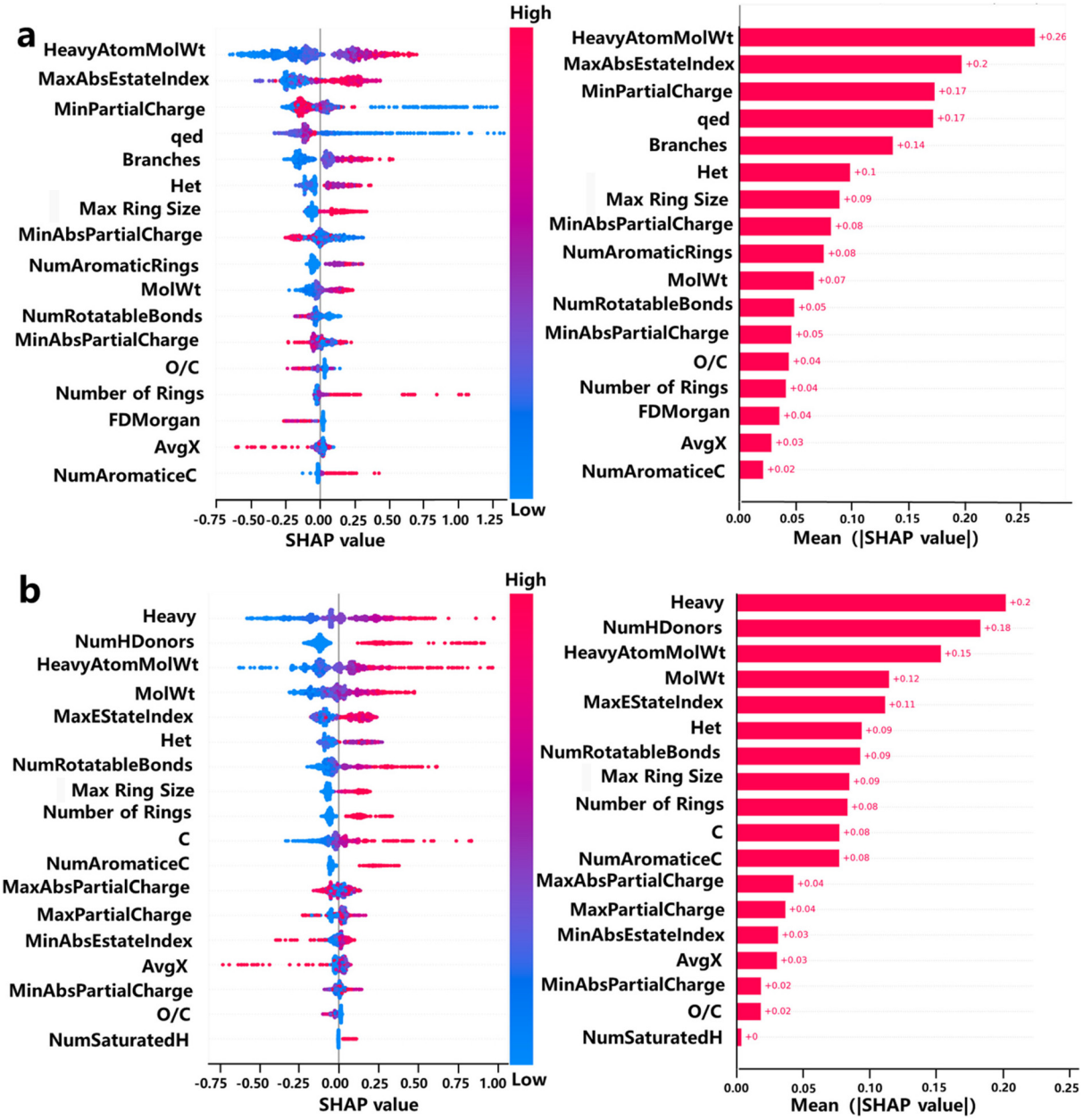

Fig. 5 Interpretation of the machine learning models for the MP (a) and the BP (b) using the SHAP algorithm.

For BP prediction (Fig. 5b), structural features contribute approximately 30–35% to the predictability. Here, increased NumRotatableBonds (0.09) yields more pronounced negative SHAP values, indicating that highly flexible molecules (*e.g.*, long-chain alkanes) more readily access conformational freedom, thereby lowering the BP. Conversely, Max Ring Size and NumAromaticRings exert positive contributions, attesting to the beneficial influence of molecular rigidity and π–π stacking on elevating the BP.

In MP prediction (Fig. 5a), electronic features account for ≈29% of explanatory power. The maximum value of the electron state exponent for all atoms (MaxEStateIndex, 0.20) and the minimum partial charge of an atom (MinPartialCharge, 0.17) emerge among the most influential features, underscoring the pronounced role of localized extreme charge sites in intensifying intermolecular electrostatic attraction. This dominant electronic contribution highlights the exceptional relevance of electronic properties to crystal stability, as optimal lattice packing necessitates precise electrostatic complementarity. By comparison, electronic structure features contribute modestly (≈19%) to global BP predictability while still furnishing critical local interpretability for polar architectures (Fig. 5b). Positive correlations observed for MaxEStateIndex and average electronegativity (AvgX) indicate that enhanced electron delocalization or polarity strengthens intermolecular electrostatic interactions, thereby elevating the BP.

These differential patterns work synergistically in the collective model interpretations. The MP model reveals a reordered importance result where chemical-composition features (≈32.5%) drive variations by modulating polarity and intermolecular interaction strength; electronic structure features collectively contribute ≈29%, attesting to their pervasive influence; and structural features provide secondary modulation (≈38.5%). This configuration indicates that MP changes are governed primarily by the intensity of intermolecular forces, jointly determined by molecular composition and electronic structures (Fig. 5a).

The BP model presents a fundamentally distinct hierarchy: chemical-composition variables dominate (≈58% of explained variance), primarily encoding molecular size and hydrogen bond capacity. Structural features contribute 23% with aromaticity and ring architecture being paramount, while electronic structural features contribute ≈19% (Fig. 5b). This contrast signifies that the BP is determined primarily by global molecular attributes, whereas the MP is governed by electronic and structural considerations.

In summary, based on the results of t-SNE clustering, violin plot distribution, and SHAP feature importance analysis, this study proposes the key structural features associated with the excellent low-temperature electrolyte performance. First, molecules with high conformational flexibility, as reflected by the increased NumRotatable bonds and the scattered t-SNE distribution of linear structures, achieved lower melting points (MPs) and boiling points (BPs). Second, moderately polar functional groups such as ethers and carbonyl units, can enhance the dielectric constant (ε) while avoiding excessive hydrogen bonding interactions that lead to elevated MPs/BPs—a characteristic validated by the SHAP contributions of the O/C ratio and MinPartialCharge. Third, electronic features such as distributed partial charges and moderate electronic state endpoints, captured by MaxEStateIndex and MinPartialCharge, facilitate low melting transitions by weakening structural stability. In contrast, rigid aromatic systems and macrocyclic systems, which cluster in the high MP/BP region in t-SNE plots and exhibit positive SHAP contributions, are generally detrimental to low-temperature electrolyte performance due to enhanced molecular rigidity and significant cohesive interactions. These comprehensive insights provide practical structural guidance for designing electrolyte molecules with improved low-temperature performance.

Electrolyte design and screening

In addition, we construct a low-dimensional molecular embedding map *via* clustering-based dimensionality reduction, enabling efficient neighbourhood search within high-dimensional chemical space. This map facilitates intuitive identification of molecular communities with similar structures and properties, enhancing both interpretability and candidate localization. For example, in our molecular space, DOL (1,3-dioxolane)[57,58] and DMS (dimethyl sulfite)[59] are grouped into the same region due to their low melting points and favourable low-temperature fluidity, both pointing to their suitability as components in low-temperature lithium battery electrolytes, as shown in Fig. 6a. Notably, in the vicinity of this cluster, we also identified molecules such as MeTHF (2-methyltetrahydrofuran) that exhibit comparable potential for low-temperature performance. This observation underscores the effectiveness of our method in identifying structurally and functionally similar candidate solvents. The clustering-neighbourhood strategy not only reveals underlying structural similarities but also supports rapid identification of property-related molecular clusters, laying a solid foundation for high-throughput screening guided discovery. Finally, leveraging our curated organic electrolyte database, we apply this strategy to screen molecules for low-temperature electrolyte applications. We define performance criteria of MP below −40 °C to ensure liquid stability, BP above 100 °C for thermal robustness, and ε ranging from 10 to 50 is essential to ensure adequate ionic conductivity. This is because ε is a major determinant of salt dissociation and ion-pair equilibrium—higher ε values enhance salt solubility, suppress ion-pair formation, and promote efficient ionic conduction. Moreover, ε partially reflects molecular polarity and solvation ability; thus, the upper limit was set at 50 to mitigate the impact of desolvation difficulties on capacity fading at low temperatures. Meanwhile, viscosity is another critical factor affecting ionic mobility. However, the lack of high-quality viscosity data with consistent temperature and measurement conditions may introduce noise into the model. Consequently, ε was employed as the

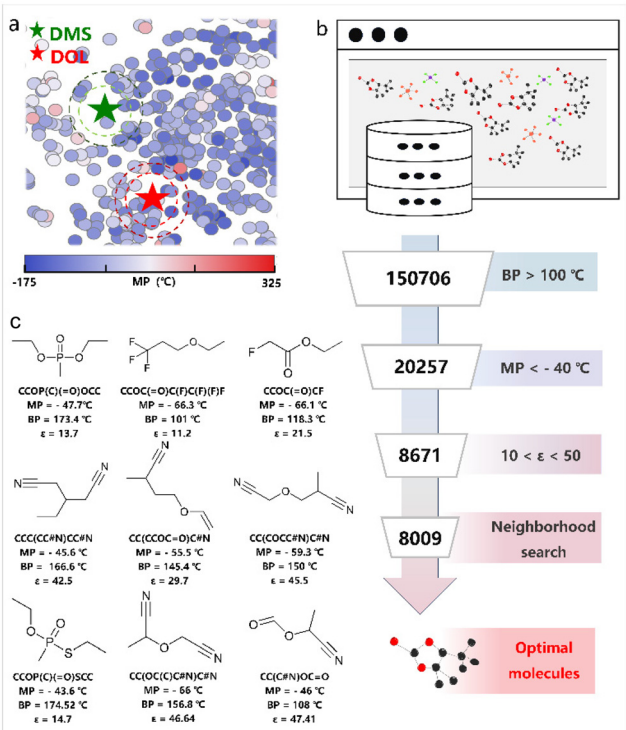

Fig. 6 Workflow for identifying low-temperature electrolyte candidates. (a) Local exploration around DOL and DMS. (b) High-throughput. (c) Nine candidate electrolyte molecules.

primary conductivity-related descriptor in this study, while viscosity will be incorporated in future work. Beyond the three criteria outlined above, molecules containing reactive hydrogen groups (*e.g.*, –OH and –COOH) were excluded to avoid uncontrolled side reactions. Based on these criteria, we identify nine candidate electrolyte molecules characterized by functional groups such as nitrile, fluorine, and phosphorus-containing moieties that confer favourable electrochemical properties and thermal stability (Fig. 6c). Notably, compounds bearing nitrile, phosphorus, or fluorine functionalities have already been employed as electrolyte solvents or additives in various high-performance battery systems.[60–63] These findings validate the effectiveness of the clustering-neighbourhood strategy and offer a feasible path for the rational design of advanced electrolytes.

To further evaluate their practical relevance, the nine screened candidate molecules were compared with existing experimental data. The deviations fall within the RMSE range of the model (Tables S4 and S5). We have also compared the candidate molecules with widely used low-temperature solvents, such as 1,3-dioxolane (DOL), 1,2-dimethoxyethane (DME), and 2-methyltetrahydrofuran (MeTHF). These molecules are generally consistent with the criteria we established, which confirms the rationality of our screening process and demonstrates the potential of the nine selected molecules as low-temperature electrolytes. Although this work focuses on large-scale ML-based screening, further validation is planned in future studies, focusing primarily on the candidate molecules with available CAS numbers. Beyond the primary screening properties, we have reported the HOMO and LUMO values of the nine electrolyte molecules (Table S4) and compared them with those of commercial low-temperature electrolyte molecules to gain further insights into their electrochemical stability (Table S5). The results confirm the adequacy of the electrochemical window for the nine screened molecules. In future studies, we intend to incorporate HOMO and LUMO descriptors into the large-scale screening workflow.

Conclusions

In this work, we established an integrated machine learning workflow that combines data acquisition, feature-model synergy, and interpretable analysis to enable accurate prediction of MPs, BPs, and ε. We constructed LiElectroDB, a comprehensive electrolyte property database encompassing 150 000 molecules from multiple sources. The structure–property relationships for MPs, BPs, and ε are systematically analysed. XGB models are employed for MP and BP prediction due to their effectiveness in feature selection and segmented fitting. We achieved R^2 values of 0.8868 and 0.9608 and RMSEs of 16.8 K and 9.15 K under five-fold cross-validation, respectively. In contrast, an ANN is adopted to model ε, which shows strong nonlinearity with respect to molecular polarity and electron distribution. The model achieves an R^2 of 0.8863 and an RMSE of 6.7. t-SNE visualization reveals that the distributions of the MP and BP are closely related to molecular polarity and functional group composition while ε primarily reflects electronic response characteristics. SHAP analysis further confirms that the BP depends on global molecular features such as size and hydrogen-bonding capacity, whereas the MP is influenced by intermolecular interactions. These insights not only validate the predictive models but also provide actionable guidance for rational electrolyte design. Finally, by combining molecular neighbourhood search with high-throughput screening, nine candidate molecules are identified as promising low-temperature electrolytes for lithium-ion batteries. This work establishes an efficient and generalizable framework for the rational design of advanced electrolytes under low-temperature conditions. Overall, this study contributes a novel data-driven framework that accelerates molecular screening, reduces experimental cost, and enables interpretable and generalizable design of advanced electrolytes under low-temperature conditions.

Author contributions

Jiechen Guo: investigation, validation, and writing – original draft. Yifan Chai: formal analysis and visualization. Cancan Hong: data curation and validation. Hao Liu: formal analysis and investigation. Lijing Xie: project administration. Tianle Wang: visualization. Jingpeng Chen: project administration. Ge Song: funding acquisition. Zonglin Yi: conceptualization and writing – review & editing. Fangyuan Su: supervision and funding acquisition.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors declare that the main data supporting the findings of this study are available within the paper and its associated supplementary information (SI). Supplementary information is available. See DOI: https://doi.org/10.1039/d5nr03942h.

All other relevant data are available from the corresponding author upon reasonable request.

Acknowledgements

This work was supported by the Preparation of Spherical Asphalt-based Hard Carbon and its Sodium Storage Mechanism (E3SWR4791Z), the Talent Projects for Outstanding Doctoral Students to Work in Shanxi Province, and the Fundamental Research Program of Shanxi Province (202403021222485 and 202403021222486).

References

1 G. Zhu, K. Wen, W. Lv, X. Zhou, Y. Liang, F. Yang, Z. Chen, M. Zou, J. Li, Y. Zhang and W. He, *J. Power Sources*, 2015, **300**, 29–40.
2 W. Cai, Y.-X. Yao, G.-L. Zhu, C. Yan, L.-L. Jiang, C. He, J.-Q. Huang and Q. Zhang, *Chem. Soc. Rev.*, 2020, **49**, 3806–3833.
3 P. Mei, Y. Zhang and W. Zhang, *Nanoscale*, 2023, **15**, 987–997.
4 Y. Mo and X. Dong, *Next Energy*, 2024, **3**, 100115.
5 Q. Li, G. Liu, H. Cheng, Q. Sun, J. Zhang and J. Ming, *Chemistry*, 2021, **27**, 15842–15865.
6 Z. Wang and B. Zhang, *Energy Mater. Devices*, 2023, **1**, 9370003.
7 X. Fan and C. Wang, *Chem. Soc. Rev.*, 2021, **50**, 10486–10566.
8 J. Hou, M. Yang, D. Wang and J. Zhang, *Adv. Energy Mater.*, 2020, **10**, 1904152.
9 Q. Li, G. Liu, H. Cheng, Q. Sun, J. Zhang and J. Ming, *Chem. – Eur. J.*, 2021, **27**, 15842–15865.
10 W. Lv, C. Zhu, J. Chen, C. Ou, Q. Zhang and S. Zhong, *Chem. Eng. J.*, 2021, **418**, 129400.
11 X. Lin, M. Salari, L. M. R. Arava, P. M. Ajayan and M. W. Grinstaff, *Chem. Soc. Rev.*, 2016, **45**, 5848–5887.
12 N. Nicodemo, R. Di Rienzo, M. Lagnoni, A. Bertei and F. Baronti, *J. Energy Storage*, 2024, **99**, 113257.
13 X. Zhang, D. Mu, S. Lu, Y. Zhang, Y. Zhang, Z. Yang, Z. Zhao, B. Wu and F. Wu, *Energy Environ. Mater.*, 2024, **7**, e12744.
14 Y. Chai, Z. Yi, J. Guo, X. Guo, C. Hong, G. Song, Y. Fan, W. Li, X.-M. Li, L. Xie and F. Su, *Energy Storage Mater.*, 2025, **80**, 104435.
15 R. Li, W. Zhao, R. Li, C. Gan, L. Chen, Z. Wang and X. Yang, *J. Energy Chem.*, 2025, **106**, 44–62.
16 Y. Zhao, Z. Hu, Z. Zhao, X. Chen, S. Zhang, J. Gao and J. Luo, *J. Am. Chem. Soc.*, 2023, **145**, 22184–22193.
17 Y. Yang, N. Yao, Y.-C. Gao, X. Chen, Y.-X. Huang, S. Zhang, H.-B. Zhu, L. Xu, Y.-X. Yao, S.-J. Yang, Z. Liao, Z. Li, X.-F. Wen, P. Wu, T.-L. Song, J.-H. Yao, J.-K. Hu, C. Yan, J.-Q. Huang and Q. Zhang, *Angew. Chem., Int. Ed.*, 2025, **64**, e202505212.
18 Y. Wang, *npj Comput. Mater.*, 2025, **11**, 89.
19 R. Kumar, M. C. Vu, P. Ma and C. V. Amanchukwu, *Chem. Mater.*, 2025, **37**, 2720–2734.
20 Y. Ma, S. Han, Y. Sun, Z. Cui, P. Liu, X. Wang and Y. Wang, *J. Power Sources*, 2024, **604**, 234492.
21 Z. Yi, Y. Zhou, H. Liu, L. Li, Y. Zhao, J. Li, Y. Mao, F. Su and C.-M. Chen, *npj Comput. Mater.*, 2025, **11**, 7564–7574.
22 C. V. Prasshanth, A. K. Lakshminarayanan, B. Ramasubramanian and S. Ramakrishna, *Next Mater.*, 2024, **2**, 100145.
23 Y.-C. Gao, N. Yao, X. Chen, L. Yu, R. Zhang and Q. Zhang, *J. Am. Chem. Soc.*, 2023, **145**, 23764–23770.
24 S. Gong, Y. Zhang, Z. Mu, Z. Pu, H. Wang, X. Han, Z. Yu, M. Chen, T. Zheng, Z. Wang, L. Chen, Z. Yang, X. Wu, S. Shi, W. Gao, W. Yan and L. Xiang, *Nat. Mach. Intell.*, 2025, **7**, 543–552.
25 N. Yao, L. Yu, Z.-H. Fu, X. Shen, T.-Z. Hou, X. Liu, Y.-C. Gao, R. Zhang, C.-Z. Zhao, X. Chen and Q. Zhang, *Angew. Chem., Int. Ed.*, 2023, **62**, e202305331.
26 A. K. Chew, M. Sender, Z. Kaplan, A. Chandrasekaran, J. Chief Elk, A. R. Browning, H. S. Kwak, M. D. Halls and M. A. F. Afzal, *J. Cheminf.*, 2024, **16**, 31.
27 Y. C. Gao, Y. H. Yuan, S. Huang, N. Yao, L. Yu, Y. P. Chen, Q. Zhang and X. Chen, *Angew. Chem., Int. Ed.*, 2024, **64**, e202416506.
28 T. Chen and C. Guestrin, *ACM*, 2016, 785–794.
29 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton, *Nucleic Acids Res.*, 2023, **51**, D1388–D1395.
30 M. Linstrom and W. G. Mallard, NIST Chemistry WebBook, NIST Standard Reference Database Number 69, 2025, DOI: 10.18434/T4D303.
31 D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, 1986, **323**, 533–536.
32 V. D. M. Laurens and G. Hinton, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.
33 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.
34 J. A. Dean, *Lange's Handbook of Chemistry*, 1999.
35 J. Liu, S. Gong, H. Li and G. Liu, *Fuel*, 2022, **313**, 122712.
36 R. Li, J. M. Herreros, A. Tsolakis and W. Yang, *Fuel*, 2021, **304**, 121437.
37 F. Gharagheizi, S. A. Mirkhani, P. Ilani-Kashkouli, A. H. Mohammadi, D. Ramjugernath and D. Richon, *Fluid Phase Equilib.*, 2013, **354**, 250–258.
38 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
39 RDKit: Open-Source Cheminformatics, https://www.rdkit.org/.
40 J. S. Kim and C. Scott, *IEEE*, 2008, DOI: 10.48550/arXiv.1107.3133.
41 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
42 I. Guyon and A. Elisseeff, *J. Mach. Learn. Res.*, 2003, **3**, 1157–1182.
43 J. R. Quinlan, *Mach. Learn.*, 1986, 81–106.
44 L. Breiman, *Mach. Learn.*, 1996, **24**, 123–140.
45 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
46 P. Geurts, D. Ernst and L. Wehenkel, *Mach. Learn.*, 2006, **63**, 3–42.
47 Y. Freund and R. E. Schapire, *J. Comput. Syst. Sci.*, 1997, **55**, 119–139.
48 J. H. Friedman, *Ann. Stat.*, 2001, **29**, 1189–1232, 1144.
49 D. H. Lopez and S. H. Yalkowsky, *Pharm. Res.*, 2023, **40**, 2801–2815.
50 Y. L. Slovokhotov, I. S. Neretin and J. A. K. Howard, *New J. Chem.*, 2004, **28**, 967–979.
51 B. Admire, B. Lian and S. H. Yalkowsky, *Chemosphere*, 2015, **119**, 1436–1440.
52 J. C. Dearden, *Environ. Toxicol. Chem.*, 2003, **22**, 1696–1709.

53 M. S. Westwell, M. S. Searle, D. J. Wales and D. H. Williams, *J. Am. Chem. Soc.*, 1995, **117**, 5013–5015.
54 T. Meng, S. Yang, Y. Peng, P. Li, S. Ren, X. Yun and X. Hu, *Adv. Energy Mater.*, 2025, **15**, 2404009.
55 K. Bao, M. Wang, Y. Zheng, P. Wang, L. Yang, Y. Jin, H. Wu and B. Sun, *Nano Energy*, 2024, **120**, 109089.
56 J. Yu, W. Ren, C. Yu, Z. Wang, Y. Xie and J. Qiu, *Energy Environ. Mater.*, 2023, **6**, e12602.
57 Y. X. Yao, X. Chen, C. Yan, X. Q. Zhang, W. L. Cai, J. Q. Huang and Q. Zhang, *Angew. Chem., Int. Ed.*, 2020, **60**, 4090–4097.
58 Z. Ma, J. Chen, J. Vatamanu, O. Borodin, D. Bedrov, X. Zhou, W. Zhang, W. Li, K. Xu and L. Xing, *Energy Storage Mater.*, 2022, **45**, 903–910.
59 J. Liu, B. Yuan, N. He, L. Dong, D. Chen, S. Zhong, Y. Ji, J. Han, C. Yang, Y. Liu and W. He, *Energy Environ. Sci.*, 2023, **16**, 1024–1034.
60 N. D. Rodrigo, C. Jayawardana, L. Rynearson, E. Hu, X.-Q. Yang and B. L. Lucht, *J. Electrochem. Soc.*, 2022, **169**, 110504.
61 Y.-G. Cho, Y.-S. Kim, D.-G. Sung, M.-S. Seo and H.-K. Song, *Energy Environ. Sci.*, 2014, **7**, 1737–1743.
62 J. Chen, J. Vatamanu, L. Xing, O. Borodin, H. Chen, X. Guan, X. Liu, K. Xu and W. Li, *Adv. Energy Mater.*, 2020, **10**, 1902654.
63 H. Liang, Z. Ma, Y. Wang, F. Zhao, Z. Cao, L. Cavallo, Q. Li and J. Ming, *ACS Nano*, 2023, **17**, 18062–18073.