A Journal of the Gesellschaft Deutscher Chemiker

Angewandte International Edition Chemie

GDCh

www.angewandte.org

Accepted Article

Title: A Knowledge–Data Dual-Driven Framework for Predicting the Molecular Properties of Rechargeable Battery Electrolytes

Authors: Yu-Chen Gao, Yu-Hang Yuan, Suozhi Huang, Nan Yao, Legeng Yu, Yao-Peng Chen, Qiang Zhang, and Xiang Chen

To be cited as: *Angew. Chem. Int. Ed.* **2024**, e202416506

Link to VoR: https://doi.org/10.1002/anie.202416506

WILEY VCH

10.1002/anie.202416506

WILEY VCH

RESEARCH ARTICLE

A Knowledge–Data Dual-Driven Framework for Predicting the Molecular Properties of Rechargeable Battery Electrolytes

Yu-Chen Gao,[a] Yu-Hang Yuan,[a] Suozhi Huang[b], Nan Yao,[a] Legeng Yu,[a] Yao-Peng Chen,[a] Qiang Zhang,[a] Xiang Chen*[a]

[a] Tsinghua Center for Green Chemical Engineering Electrification (CCEE), Beijing Key Laboratory of Green Chemical Reaction Engineering and Technology, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China
E-mail: xiangchen@mail.tsinghua.edu.cn

[b] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China

Abstract: Developing rechargeable batteries that operate within a wide temperature range and possess high safety has become necessary with increasing demands. Rapid and accurate assessment of the melting points (MPs), boiling points (BPs), and flash points (FPs) of electrolyte molecules is essential for expediting battery development. Herein, we introduce Knowledge-based electrolyte Property prediction Integration (KPI), a knowledge–data dual-driven framework for molecular property prediction of electrolytes. Initially, the KPI collects molecular structures and properties, and then automatically organizes them into structured datasets. Subsequently, interpretable machine learning further explores the structure–property relationships of molecules from a microscopic perspective. Finally, by embedding the discovered knowledge into property prediction models, the KPI achieved very low mean absolute errors of 10.4, 4.6, and 4.8 K for MP, BP, and FP predictions, respectively. The KPI reached state-of-the-art results in 18 out of 20 datasets. Utilizing molecular neighbor search and high-throughput screening, 15 and 14 promising molecules, with and without Chemical Abstracts Service Registry Number, respectively, were predicted for wide-temperature-range and high-safety batteries. The KPI not only accurately predicts molecular properties and deepens the understanding of structure–property relationships but also serves as an efficient framework for integrating artificial intelligence and domain knowledge.

Introduction

Rechargeable batteries such as lithium-ion batteries have been playing an increasingly important role in human daily life, including powering electric vehicles and portable electronic devices[1]. In addition to the need for high energy density, there are growing demands for extending the working temperatures of batteries due to harsh cold, desert regions, and specialized applications such as spacecraft, underground exploration, and sterilization of medical equipment[2, 3]. At high temperatures, the side reactions intensify, leading to rapid depletion of electrolytes and active materials, and potentially triggering thermal runaway in batteries[4]. On the contrary, the reaction kinetics decreases dramatically at low temperatures, which can cause the formation of lithium dendrites on graphite anodes and further induce safety hazards[5]. Therefore, developing wide-temperature-range batteries significantly heightened safety considerations[6].

The working temperature range and the safety performance of batteries are highly dependent on the physicochemical properties of the electrolytes[7, 8]. For example, Wang et al.[9] regarded a low melting point (MP) but a moderate boiling point (BP) as the primary criterion for designing electrolytes for lithium batteries. The authors reasonably designed 4.5 V $LiNi_{0.8}Mn_{0.1}Co_{0.1}O_2$ || graphite full cells with an areal capacity of 2.5 mAh cm^{-2}, which can effectively operate over a wide temperature range from −60 to 60°C. Yamada et al.[10] developed a fire-extinguishing concentrated $LiN(SO_2F)_2$/trimethyl phosphate (TMP) electrolyte by thoroughly considering its BP and flash point (FP). This electrolyte enabled the graphite anode to survive 1,000 cycles at a C/5 rate. However, previous electrolyte innovations have been mainly based on an experimental trial-and-error approach because of limited experimental data on electrolyte physicochemical properties[11]. This time-consuming approach faces grand challenges when seeking wide-temperature-range and safe electrolytes that simultaneously require a low MP, a high BP, and a high FP.

With the rapid development of high-performance computing and data-driven technologies, theoretical prediction and high-throughput screening have become favorable approaches to accelerate electrolyte design[12]. Many methods have been developed to predict the MP, BP, and FP of electrolytes. For instance, early estimations of these molecular properties employed specific mathematical functions with several fitted physical parameters, but these methods typically required the introduction of additional properties and simplifying assumptions for computations, resulting in low applicability[13]. To broaden the generalizability of property predictions, the group contribution (GC) method assumes that the contribution of individual functional groups is consistent across molecules, allowing for rapid acquisition of molecular properties through linear summation[14]. However, the accuracy of the GC method is often very limited due to the lack of considering the interaction between different groups[15]. With the rise of computational chemistry and materials methods, many electrolyte properties such as viscosity and dielectric constant can be accurately predicted by molecular dynamics simulations and statistical methods[16]. Simulating phase transitions typically requires prolonged simulations and significant computational resources[17]. The difficulty in accurately capturing equilibrium states affects the prediction accuracy. Furthermore, for MPs, simulating the temperature at which lattice

10.1002/anie.202416506

decomposition occurs is important, but obtaining the crystal structure poses challenges[18]. Even for FP, it is necessary to simulate the temperature at which the substance evaporates and forms a flammable mixture with air, which involves great complexity. With the rapid development of machine learning, the quantitative structure–property relationship (QSPR) method has rapidly evolved and achieved success in drug design[19]. However, this method is still largely dependent on the precise and manual construction of molecular descriptors[20]. Consequently, previous methods for predicting molecular properties have poor generalizability, are restricted to certain systems or single properties, and require substantial manual intervention. For the rapid and precise acquisition of MPs, BPs, and FPs, novel methods are urgently required to achieve generalized and automated property prediction.

Herein, we developed a Knowledge-based electrolyte Property prediction Integration (KPI) framework for a universal, automated, and interpretable prediction of electrolyte molecular properties, including MPs, BPs, and FPs. The framework is capable of automatically gathering data from public datasets, followed by data filtering, analysis, and visual representation. Moreover, it utilizes interpretable machine learning to extract chemical knowledge from the data, thereby enhancing the understanding of molecular properties. To further refine the prediction accuracy of the model, knowledge at different levels, *i.e.*, atoms, bonds, and molecules, is embedded into the model under the precise set of the controller. Compared with predictions based on only feature descriptors, the KPI framework demonstrated a significant improvement in prediction accuracy, reducing the mean absolute error (MAE) by 51.9%, 68.2%, and 55.5% for MP, BP, and FP predictions, respectively. When compared with unmodified deep learning models, the reductions were 6.7%, 14.7%, and 17.8% for MP, BP, and FP predictions, respectively. Impressively, the KPI framework surpasses state-of-the-art (SOTA) results in 18 out of 20 baseline datasets. The KPI framework demonstrated excellent talents in designing promising electrolyte molecules for wide-temperature-range and high-safety batteries. Fifteen and fourteen molecules, with and without Chemical Abstracts Service Registry Number (CAS ID), respectively, were predicted through molecular neighbor search and high-throughput screening.

Results and Discussion

Overview of the KPI framework

The KPI framework mainly consists of three modules (Fig. 1): (a) data organization and statistical analysis, (b) interpretability and knowledge discovery, and (c) knowledge-based molecular property prediction. The framework emphasizes the crucial role of knowledge across its entire process. The knowledge discovery and knowledge embedding are organically integrated to form a closed loop, which ensures a high prediction accuracy of MPs, BPs, and FPs. Each module of the KPI framework is introduced in detail as follows.

The data organization and statistical analysis module can extract molecular structures and corresponding properties from public databases and reported papers (Fig. 1a). After filtering by implicit prior knowledge, the data is automatically organized into structured tables. Then, the module analyzes the descriptive statistics of the collected molecules and visualizes the dataset to represent the chemical space. Herein, the MPs, BPs, and FPs of battery electrolyte molecules are especially focused on, but the module can handle other molecule discovery tasks according to their specific scenario requirements.

The interpretability and knowledge discovery module further transforms structured molecular data into feature-based molecular information and employs the Shapley additive explanations (SHAP) algorithm for knowledge discovery (Fig. 1b). The knowledge generated from the data-driven approach correlates molecular structure and properties, affording crucial references for designing new electrolyte molecules. Furthermore, the knowledge acquired by this module is fed into the next module as a prior for the deep learning models.

The knowledge-based molecular property prediction module integrates molecular structure with knowledge, leveraging deep learning models for predicting molecular properties (Fig. 1c). The model takes a large amount of molecular structure data as input, and the acquired knowledge is embedded as atomic, bonding, and molecular representations according to different purity levels and flow rates. This approach fully considers the interaction between molecular structure information and chemical knowledge, enabling the model to acquire additional knowledge based on data-driven model training.

Data organization and statistical analysis

The KPI framework utilizes the application programming interfaces (APIs) to automatically gather and format data from reported papers and public databases into binary data sets of molecular structures and corresponding properties (Fig. 2a). The molecular structures are represented using simplified molecular input line entry system (SMILES), and the properties include MPs, BPs, and FPs. Considering the molecular property distribution of routine electrolytes (Supplementary Fig. S1), KPI restricts the molecular weight (Molwt) within a range from 0 to 600 and the number of heavy atoms (#Heavy) from 0 to 30. Besides, molecular elements are restricted to hydrogen, carbon, nitrogen, oxygen, fluorine, silicon (Si), phosphorus (P), sulfur, chlorine, bromine, and iodine. Ultimately, the acquired database consists of 4,235 molecules for MPs, 4,153 for BPs, and 3,504 for FPs, with property ranges of 85.4–633.2, 182.5–692.2, and 190.1–538.5 K, respectively (Fig. 2b).

The distribution of these properties generally follows a normal distribution, which reflects the adequacy of the sample size and the completeness of data collection as per the central limit theorem and is advantageous for statistical inference in modeling prediction. The distributions of Molwt and #Heavy are concentrated at small values, aligning well with the characteristics of molecules widely used in electrolytes. Since both MPs and BPs are inherently affected by intermolecular forces, there is a certain correlation between them (Spearman's rank correlation coefficient $\rho_{Spearman}$ = 0.81, Supplementary Fig. S2). Furthermore,

10.1002/anie.202416506

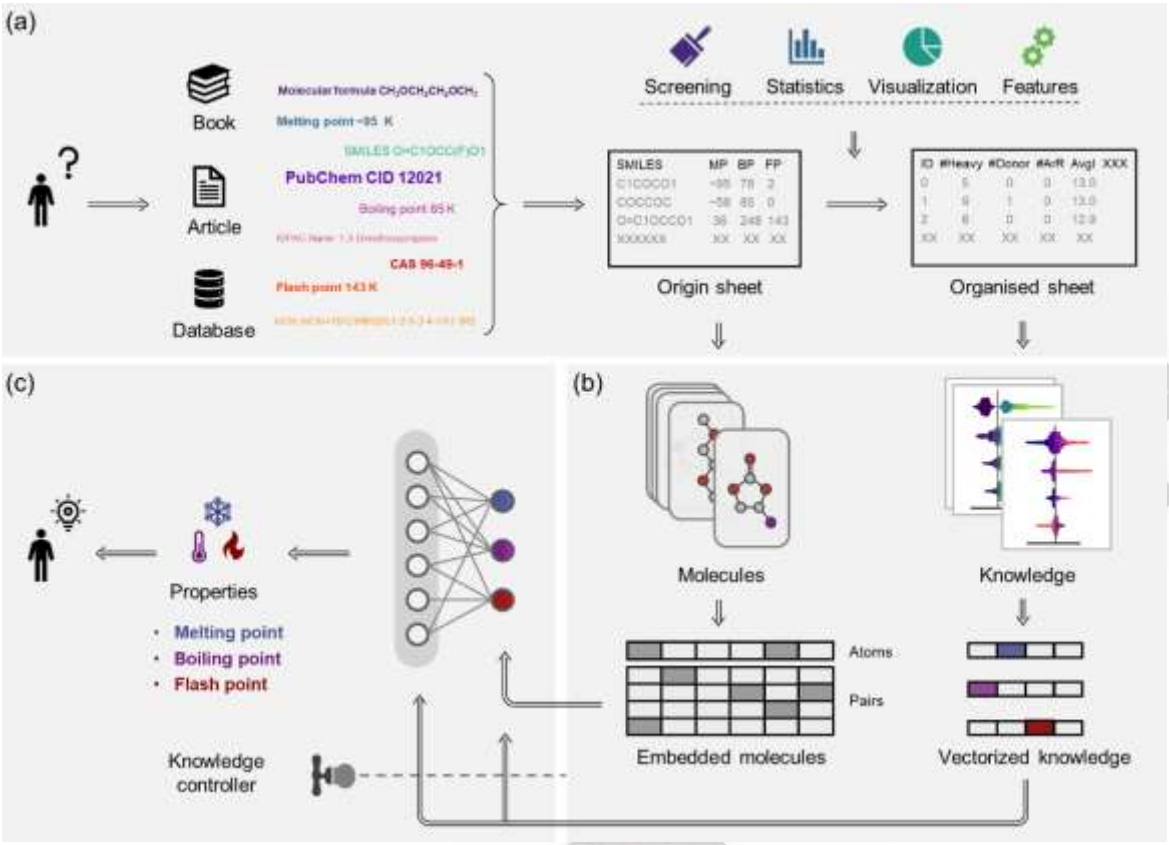

Figure 1. Overview of the Knowledge-based electrolyte Property prediction Integration (KPI) framework. (a) The data organization and statistical analysis module. For molecular properties of interest to researchers, such as melting points (MPs), boiling points (BPs), and flash points (FPs), the module can quickly retrieve data from databases via application programming interfaces (APIs) or collect it from papers or books. The model then cleanses this data and conducts statistical analysis to initially acquire macroscopic knowledge. Subsequently, the model performs cluster analysis and molecular neighbor searches, obtaining a visual representation of the molecular database and delineating the chemical space of molecules of interest. (b) The interpretability and knowledge discovery module. The module automatically extracts molecular descriptors from the simplified molecular input line entry system (SMILES) to obtain feature vectors. Utilizing the Shapley additive explanations (SHAP) for interpretable analysis, the module explores the impact of features on specified properties, ranks the importance of features, and identifies their relationships, thereby analyzing the factors influencing molecular properties from a microscopic perspective. (c) The knowledge-based molecular property prediction module. The module embeds molecular structures along with the knowledge acquired from the previous module into a deep learning model, adjusting the purity and flow of the embedded knowledge via a knowledge controller to optimize the model. Based on the trained model, it then provides feedback on the properties of molecules of interest to the researchers.

the trend in FPs generally aligns with that of BPs ($\rho_{Spearman}$ = 0.97, Supplementary Fig. S3), because BPs are related to the evaporation tendency, therefore directly influencing FPs.

To further visualize the distribution of the molecules, the KPI framework encodes molecules using the molecular access system (MACCS) and employs the t-distributed stochastic neighbor embedding (t-SNE) clustering method, colored according to corresponding properties (Fig. 2c). The clustering diagram exhibits distinct regions, suggesting the ability to identify potential electrolyte molecules under specific property zones. Molecules with excellent low-temperature performance, such as 1,3-dioxolane (DOL)[21] and ethyl acetate (EA)[22], high-temperature performance like fluobenzene[23] and dimethyl 2,5-dioxahexanedioate[24], as well as safety-oriented compounds like TMP[10] and N,N-dimethylacetamide[25] can be identified using the clustering maps (Supplementary Table S1). Since molecular fingerprints encapsulate structural information, clustering analysis holds promise for rapidly delineating feasible molecular spaces through targeting molecules.

Furthermore, the KPI conducted preliminary macroscopic statistical analyses on the data categorized into hydrocarbons (HC), oxygen-containing (OC), and other heteroatom-containing (OHC) molecules to explore the impact of functional groups on molecular properties (Supplementary Fig. S4–S6 and Supplementary Table S2). HC have an average MP of 224.3 K, which is around 100 K lower than those of molecules in the other two categories (339.2 and 332.4 K for OC and OHC, respectively. Supplementary Table S3). The low average MP of HC can be attributed to their low Molwt and #Heavy (the average of Molwt and #Heavy is 164.7 and 12.0, respectively) and their limited ability to form hydrogen bonds compared with OC and OHC (the average of the number of hydrogen bond donors (#Donor) is 0. Supplementary Fig. S7). Among OC, ethers and esters, with low average MPs of 291.3 and 326.6 K, respectively, are potential categories for low-temperature electrolytes (Supplementary Table S3). Additionally, Si and P elements can reduce the average of MPs to 240.4 and 304.2 K, respectively. Although P-containing molecules have a lower average BP than the whole (406.9 K *vs.*

10.1002/anie.202416506

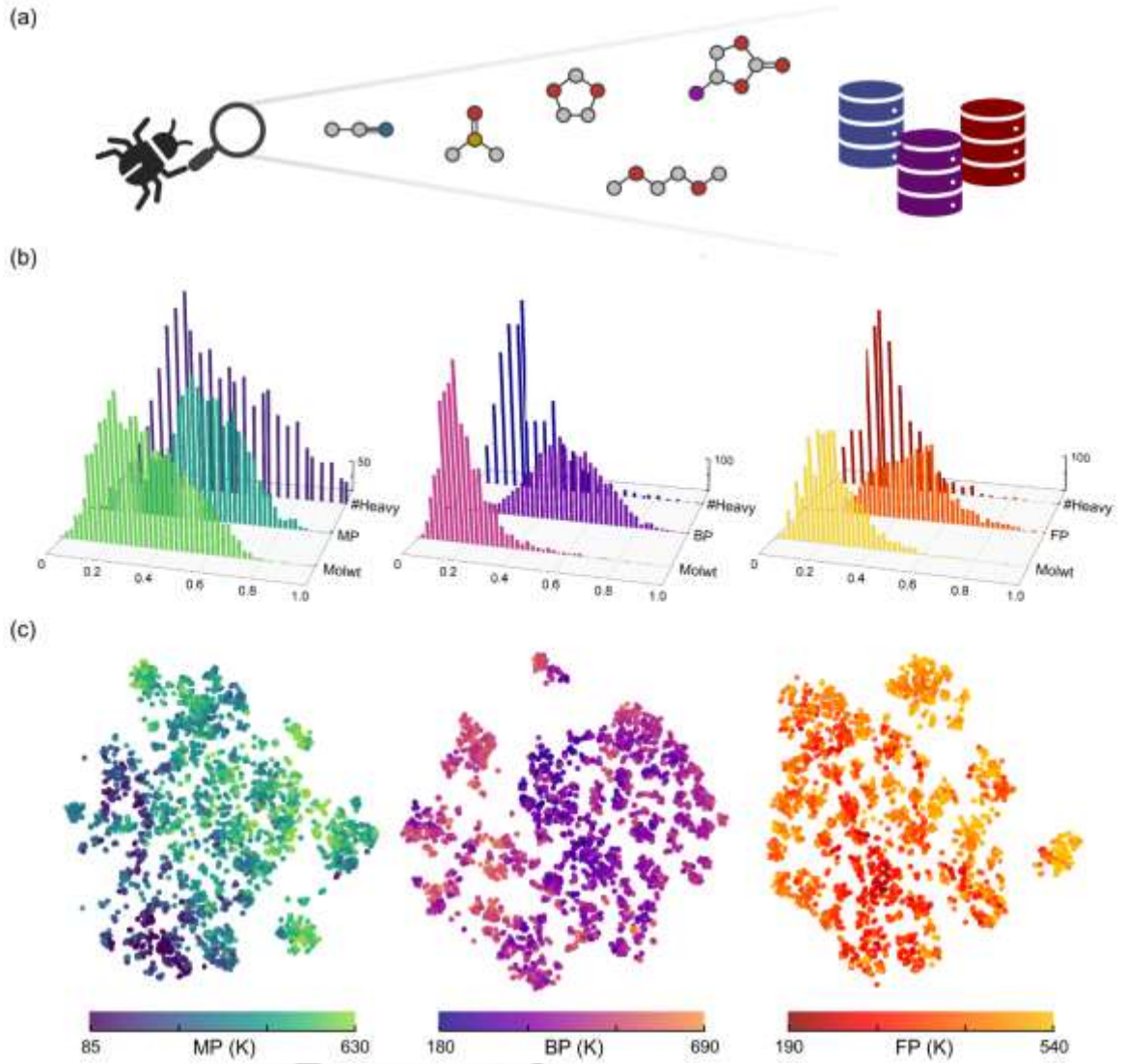

Figure 2. Data collection, statistical analysis, and visualization of MPs, BPs, and FPs. (a) The schematic of data collection. Batch and rapid data acquisition are implemented using APIs, with a focus on ensuring that common electrolyte molecules are included in our constructed database. (b) Frequency distribution graphs of data for MPs (left), BPs (medium), and FPs (right). The graphs include these properties as well as molecular weight (Molwt) and the number of heavy atoms (#Heavy), with these attributes normalized to a range from 0 to 1 using minimum and maximum normalization. The scale in the vertical direction represents the number of molecules. (c) Visualization analysis of data for MPs (left), BPs (medium), and FPs (right). Molecules are represented using the molecular access system (MACCS), and cluster analysis is conducted using the t-distributed stochastic neighbor embedding (t-SNE) algorithm. Molecules are visualized in the graph, with each data point representing a molecule and the color gradient from dark to light indicating property values increasing from low to high.

442.1 K), their FPs are relatively higher (386.4 K *vs.* 350.3 K) (Supplementary Table S4 and S5). P-containing molecules tend to decompose to produce PO_2• and HPO_2• radicals and the radicals can capture H• and OH · radicals generated during combustion, which explains why flame retardants usually include P elements.

Interpretability and knowledge discovery

Besides functional groups, the KPI framework further automatically extracts 64-dimensional molecular descriptors, including the number of atoms, the nature of bonds, functional groups, and electronic properties (Fig. 3a and Supplementary Table S6–S9). Subsequently, the heat map is used to examine the correlations between the features (Supplementary Fig. S8–S10), and SHAP is utilized to analyze the impact of each molecular feature on MPs, BPs, and FPs. The importance of all molecular features is analyzed, and the top ten features are visualized (Fig. 3b and Supplementary Table S10). The vertical display reflects the distribution of molecules for each feature, and the horizontal display shows the contribution of feature values to the prediction outcomes. Overall, MPs, BPs, and FPs are primarily influenced by the number of atoms in the molecules and the nature of the bonds, while electronic properties predominantly affect the BPs and FPs of the molecules.

For MPs, the significance of bond features accounts for 48.9% (including #Donor, the number of rings (#Ring), the number of atoms on the maximum ring (#Nring), the number of double bonds (#R=R), and the number of rotatable bonds (#Rot)) among the top ten features, and the number of atoms accounts for 30.5%

10.1002/anie.202416506

RESEARCH ARTICLE

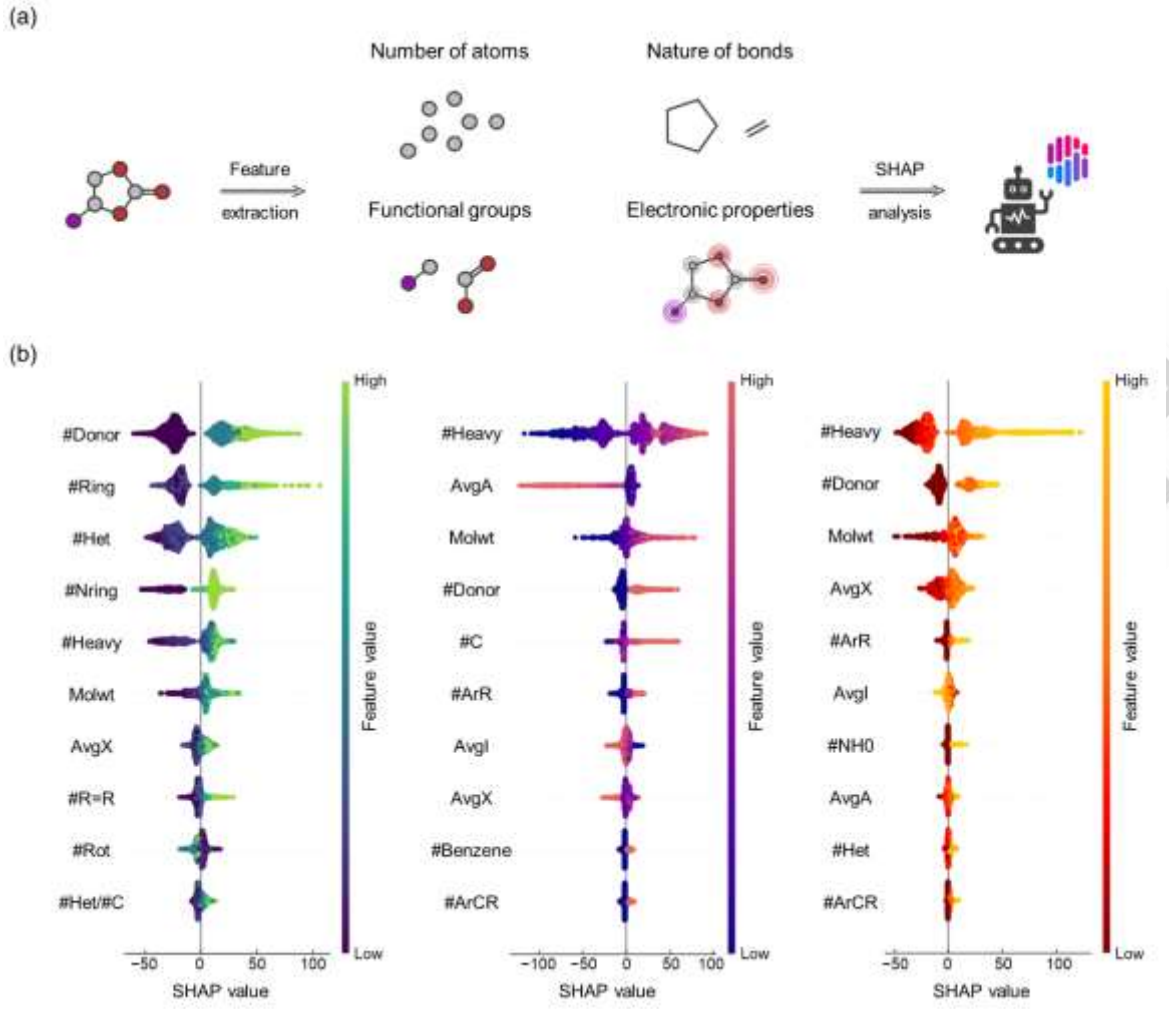

Figure 3. Feature extraction and interpretable analysis based on SHAP. (a) The schematic of molecular feature extraction. Utilizing the RDKit toolkit, the SMILES serves as the input. Extracted features include the number and mass of atoms, the nature of bonds, functional groups, and electronic properties. These features are subjected to interpretable analysis. (b) SHAP feature importance ranking for MPs (left), BPs (medium), and FPs (right). The 64-dimensional features extracted for each SMILES of molecule serve as the input, and analysis is performed using the Tree SHAP algorithm. The top ten most important features identified are visualized. The SHAP value indicates the contribution of each sample point to the performance. The color gradient from dark to light represents the increasing values of each feature.

(including the number of heteroatoms (#Het), #Heavy, Molwt, and the ratio of heteroatoms to carbon atoms (#Het/#C)), which is consistent with the discussion in the previous section. The interpretable machine learning can further help identify important factors. (Fig. 3b and Supplementary Table S11 and S12). Specifically, #Donor ranks first, with an importance share of 18.7%. Compared to molecules without any donor, those with one, two, and three donors exhibit an average MP increase of 65.6, 56.0, and 35.4 K, respectively (Supplementary Fig. S11 and Supplementary Table S13). The increase of MP can be explained by the formation of hydrogen bonds, which are generally stronger than typical van der Waals forces and enable molecules to form ordered network structures due to their directionality. However, there is no obvious change in the average MP when #Donor reaches 4, which is due to the saturation of the formation of hydrogen bonds between molecules.

Following #Donor, the importance of the #Ring in a molecule and the #Nring accounts for 14.6% and 11.4%, respectively (Fig. 3b and Supplementary Table S11). Compared to linear molecules, molecules containing a single ring increase the average MP by 77.7 K, and as #Ring increases, the average MP also rises progressively (Supplementary Fig. S12 and Supplementary Table S14). The increase of MP is due to the conformational constraints imposed by rings, which lead to tight molecular packing and thus enhance intermolecular forces. The #R=R shows a strong positive correlation with MP, and each additional double bond in a molecule increases the average MP by approximately 30 K (Supplementary Fig. S13 and Supplementary Table S15). This effect is attributed to the rigidity imparted by double bonds and their contribution to molecular conjugation. The negative correlation between the #Rot and MP also indicates that a reduction in molecular degrees of freedom and increased rigidity can raise MP, which is consistent with the findings related to the #R=R.

For BPs and FPs, the importance of the number of atoms in a molecule accounts for 48.6% and 47.3%, respectively, and electronic properties also play significant roles, contributing 14.5% and 13.4%, respectively (Fig. 3b and Supplementary Table

10.1002/anie.202416506

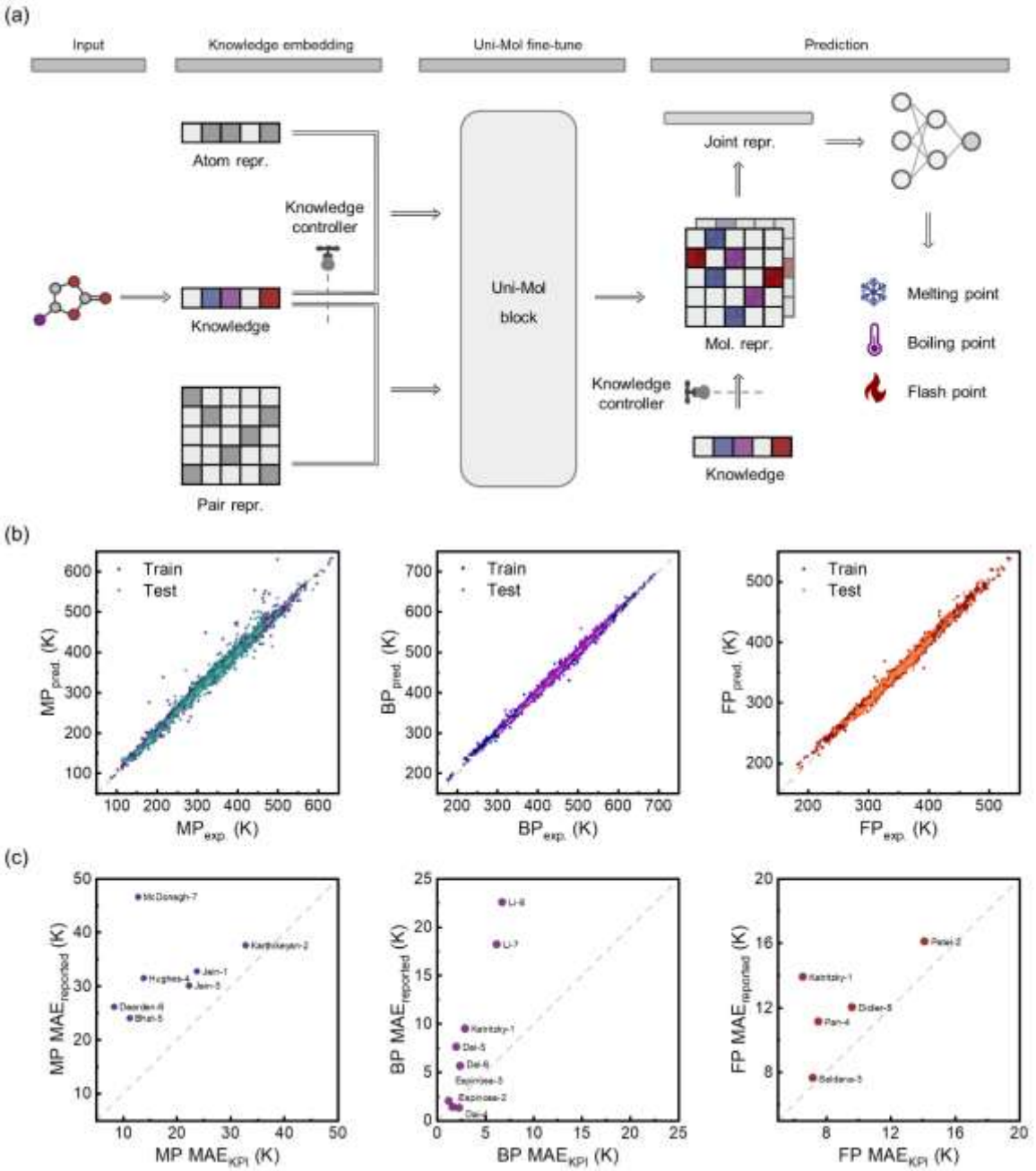

Figure 4. Knowledge-based learning model and model evaluation. (a) The schematic of the knowledge-based learning model. The models use molecular conformations as input, including encodings for atoms and bonds. Knowledge vectors are transformed and then embedded into the molecular representation through purity and flow controls. Subsequently, the model is fine-tuned based on the Uni-Mol model, and the output overall molecular representation is combined again with the knowledge. Finally, the combined representation is input into the prediction head for predicting target properties. (b) Prediction results for the state-of-the-art models of MPs (left), BPs (medium), and FPs (right). When knowledge vectors are embedded at levels of 20, 20, and 10 respectively, the predictions for MPs, BPs, and FPs are optimized. Each point represents a molecule, with colors differentiating between the training and testing sets. Points closer to the diagonal indicate better prediction accuracy. (c) Comparison of reported methods for MPs (left), BPs (medium), and FPs (right). Tests were conducted using the same datasets as those used in previously reported papers, examining 7 datasets for MPs, 8 for BPs, and 5 for FPs. Each point represents the comparison result of a dataset, with the name indicating the first author of the article and the number corresponding to its respective identifier. Complete dataset information and references can be found in Supplementary Table S26–S28. The points above the diagonal indicate that our method outperforms the methods previously reported in the paper.

S12). Compared to MPs, the importance analysis changes, which can be attributed to the transition from liquid to gas states in BPs and FPs, as opposed to the transition from solid to liquid states. In the liquid state, the intermolecular distances are larger than in the solid state, thereby reducing the influence of bond properties and amplifying the impact of the intrinsic molecular properties. A strong positive correlation is observed for Molwt and #Heavy concerning BPs and FPs, especially in HC molecules (Supplementary Fig. S14 and S15). Regarding bond nature, over 65% of molecules have no donor, and the maximum #Donor in a

molecule is three (Supplementary Fig. S16 and S17). Therefore, the effect of hydrogen bonds is less important on BPs and FPs compared with MPs, which can also be interpreted from the perspective of different phase transitions. Nonetheless, the #Donor still shows a positive correlation with both BPs and FPs, which is related to the ability of hydrogen bonds to increase intermolecular forces and structural stability.

Furthermore, a certain negative correlation is observed between BPs and the average electron affinity (AvgA), average first ionization energy (AvgI), and average electronegativity (AvgX) of molecules. However, these relationships are influenced by multiple factors and are difficult to discern from simple linear correlations (Supplementary Fig. S18). These features reflect the ease of electron gain or loss at the molecular level. Statistically, lower values of these properties indicate weaker atomic binding capacities for electrons, resulting in greater deformability and ease of polarization at the molecular level. Consequently, stronger intermolecular forces are present, leading to higher BPs.

The extracted knowledge facilitates a deep understanding of how molecular properties vary across different conditions. In low-temperature environments, for instance, the linear molecule dimethyl carbonate has a melting point of 273.6 K, which is lower than that of ethylene carbonate (309.5 K), despite their similar molecular weights. Under high-temperature conditions, the larger molecule adiponitrile exhibits a higher boiling point (568 K) compared to succinonitrile (538 K). However, the inherent complexity of molecular properties makes it challenging to predict them accurately using single structural descriptors alone, highlighting the necessity for precise predictive models.

Knowledge-based learning model and model evaluation

To build accurate models for predicting chemical properties, relying solely on data-driven approaches is insufficient. This is because many models lack fundamental chemical intuition or common sense, which can lead to gaps in understanding the underlying molecular behaviors. Data alone may capture patterns but fail to grasp the chemical principles that govern molecular properties. By fully utilizing the chemical knowledge discovered above, the KPI framework integrates this understanding into data-driven models, making them not only mathematically accurate and computationally efficient but also chemically reasonable. Specifically, the KPI framework integrates the above-discovered knowledge using Uni-Mol[26] as a foundation model for knowledge–data dual-driven molecular property prediction (Fig. 4a). Initially, molecules are input into the model in SMILES, where the molecular structure is converted into 11 different conformations, resulting in encoded representations of atoms and bonds. Knowledge from the previously mentioned module is then embedded into these representations, with the proportion of knowledge integrated into the molecular encoding regulated by two components: the knowledge purity controller and the knowledge flow controller (collectively referred to as knowledge controllers). The knowledge purity controller adjusts how many dimensions of the knowledge vector are concatenated with the molecular representation, while the knowledge flow controller determines the proportion of the selected knowledge vector to be integrated into the molecular representation. Then, the representations are processed through the transformer-based encoder, the knowledge is embedded into the fully encoded molecular vector to enhance the model's retention of knowledge. Finally, the complete molecular vector is input into the prediction head to determine the final properties.

The KPI initially sets the knowledge purity controller to the maximum, namely all the 64-dimensional knowledge vectors from the knowledge discovery module are fully encoded into the molecular representation. Simultaneously, the knowledge flow controller is set to auto-adjust, modulating the proportion of current knowledge embedded in the molecular representation in a learnable manner. The MP, BP, and FP databases were used for model training, with each dataset divided into training, validation, and test sets in a ratio of 8:1:1. Four-fold cross-validation was employed to minimize the impact of data splitting on model performance. Model parameters were continuously optimized throughout the training, with the best model corresponding to the validation set being saved. The final performance on the previously unseen test sets was evaluated using the coefficient of determination (R^2) and MAE. For MPs, BPs, and FPs, the average R^2 is 0.966, 0.988, and 0.982, and the average MAE is 11.3, 5.2, and 5.4 K, respectively (Supplementary Table S16–S19). Compared to the non-knowledge-embedded version, the performance of the model embedded with 64-dimensional knowledge improved by 5%–11%, preliminarily demonstrating the effectiveness of knowledge embedding.

To improve the model performance, KPI finely tunes the knowledge purity controller to adjust the effective purity of the embedded knowledge. Features ranking higher in importance contribute more significantly to improving molecular prediction accuracy. As more features are embedded, the introduction of less relevant knowledge increases the learning burden to some extent, leading to a decrease in prediction accuracy. Building upon the best-performing baseline where all knowledge is embedded, the KPI activates the knowledge purity controller in increments of 10 features. For MP, BP, and FP predictions, the model achieves its best performance when the amount of embedded knowledge is 20, 20, and 10, with R^2 values of 0.974, 0.990, and 0.986, and MAE values of 10.5, 4.6, and 4.8 K, respectively (Fig. 4b and Supplementary Table S20). Compared to models without knowledge embedding, performance for MAEs improves by 6.7%, 14.7%, and 17.8% for MP, BP, and FP predictions, respectively. Even compared to a random forest model dependent only on knowledge, the improvements are 51.9%, 68.2%, and 55.5% for MP, BP, and FP predictions, respectively (Supplementary Table S21–S22). To further quantify the effectiveness of the embedded knowledge, ablation studies were conducted to assess the contribution of various components within the model. For the optimal models predicting MPs, BPs, and FPs, modules embedding knowledge about atoms, bonds, and the entire molecule components were removed individually or in a combination way. The results indicate that the removal of any component leads to a decrease in the model performance (Supplementary Table S23–S25).

To evaluate the generalization of the KPI framework, datasets used in reported methods were employed as baselines, involving

10.1002/anie.202416506

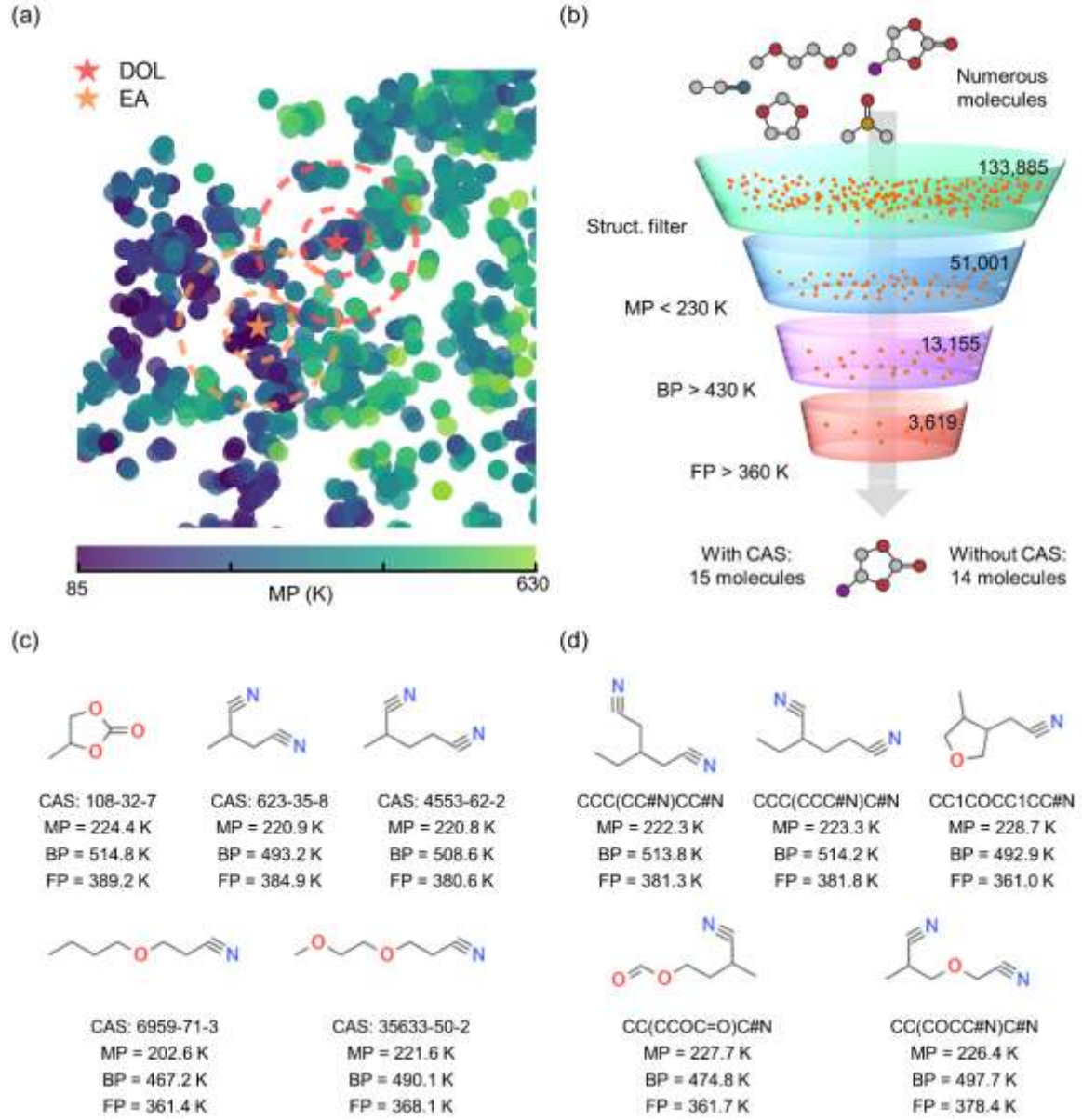

Figure 5. The KPI framework promotes advanced molecular discovery. (a) Molecular neighbor search identifies potentially valuable molecules. Using the well-known low-temperature performance of 1,3-dioxolane (DOL) and ethyl acetate (EA) as central points, primary and secondary neighborhoods are explored for potential molecules. The clustering diagram is a local enlargement of the MP from Fig. 1c. (b) High-throughput screening progressively narrows down the molecular chemical space. Starting with 133,885 molecules from QM9, three layers of filtering are applied, requiring an MP below 250 K, a BP above 450 K, and an FP above 380 K, ultimately yielding 15 molecules with Chemical Abstracts Service Registry Number (CAS ID) and 14 molecules without CAS ID. (c) Five molecules with CAS ID were obtained from the high-throughput screening. The skeletal formula, CAS ID, MP, BP, and FP of each molecule are present. The complete list of molecules can be found in Supplementary Fig. S20. (d) Five molecules without CAS ID were obtained from the high-throughput screening. The skeletal formula, SMILES, MP, BP, and FP of each molecule are present. The complete list of molecules can be found in Supplementary Fig. S21.

approaches such as GC, QSPR-based multivariate linear regression, and machine learning. The KPI framework achieves the SOTA results in 18 out of 20 datasets collected (Fig. 4c). For MP and FP predictions, the KPI consistently outperformed all other methods. However, for BP predictions, the two methods showed slightly higher performance, which can be attributed to the randomness brought by the small size of the test sets, consisting of 8 and 18 data points, respectively. Supplementary materials provide detailed information on the molecular species, scale, and methods (Supplementary Table S26–S28). Overall, the KPI framework exhibits robust predictive and generalization capabilities, enabling precise predictions for MPs, BPs, and FPs.

KPI-assisted electrolyte molecular design

The KPI framework is strongly supposed to accelerate the development of wide-temperature-range and high-safety electrolytes for advanced batteries. First, KPI can quickly lock onto potential molecular space through neighbor searches based on the dimensionality-reduced maps obtained from cluster analysis (Fig. 5a). For instance, in the primary neighborhood of

DOL, which is known for its good low-temperature performance, tetrahydrofuran[27] with a similar structure and a low MP can be discovered (Supplementary Table S1). More inspirationally, the KPI can continue to explore the secondary neighborhoods, and promising molecules such as tetraethyl orthosilicate and bis(2-ethylhexyl) hydrogen phosphate were found. Similarly, 46 and for high-throughput screening of electrolyte molecules (Fig. 5b). For instance, 15 molecules with CAS ID (Fig. 5c) and 14 molecules without CAS ID (Fig. 5d) were screened out from a large database, namely QM9[28], containing 133, 885 molecules when setting a selection criterion of MP smaller than 230 K, BP larger than 430 K, and FP larger than 360 K after the structural filtering (including removing molecules containing active hydrogen groups (–OH, –COOH) and restricting the Molwt < 600 and #Heavy < 30) (Supplementary Fig. S19, S20, and S21). Compared to the original dataset, the high-throughput screening reduced the pool by more than 4, 600 times, significantly accelerating the molecular discovery. Based on the screening results, we identified familiar molecules such as propylene carbonate (PC), which is known for its wide temperature range, high flash point, and strong solvating ability, originally adopted in the 1950s. Recently, Xie et al.[23] developed a PC-based electrolyte with a wide temperature range from −90 to 90°C by tuning non-solvating interactions. Additionally, nitrile-based electrolyte molecules were screened out. In fact, extensive research has focused on nitrile-based electrolytes[2, 29], including molecules with a single cyano group (acetonitrile[7], propionitrile[30], isobutyronitrile[31], valeronitrile[32]), molecules with multiple cyano groups (succinonitrile[33], adiponitrile[34]), and alkoxy nitrile compounds (3-methoxypropyionitrile[35]). The molecules we identified include homologous compounds that share structural similarities (differing by several -CH_2 groups) with these reported molecules, suggesting their potential in wide-temperature-range applications.

The KPI framework boasts extensive versatility and achieved the SOTA models for predicting molecular MPs, BPs, and FPs. These property prediction models are not only crucial for the development of advanced electrolytes but also play a significant role in industries such as petroleum and fuels[36]. Besides, the KPI framework integrates data organization and knowledge discovery with property prediction, and supports the analysis of various molecular properties, providing powerful tools for high-throughput acquisition of molecular properties and discovery of domain knowledge. This knowledge–data-driven framework throughout the entire process sets a new paradigm for the developing deep learning models, combining knowledge discovery and embedding to significantly enhance the performance of artificial intelligence methods in practical applications.

The prediction MAE of MPs is consistently larger than that of BPs and FPs, which involve transitions between liquid and gas states and can be accurately predicted using only molecular features. However, MPs involve the transition from solid to liquid and the intermolecular interactions within the crystal are stronger than those in liquids. As a result, the current molecular descriptors can not adequately describe solid–liquid transition and crystallographic information as prior knowledge is necessary to improve the prediction accuracy of MPs in further model development.

242 molecules with similar structures were found in the primary and secondary neighborhoods of the targets when searching in the BP and FP clustering maps (Supplementary Table S1). Second, the KPI framework can accurately and rapidly predict molecular properties in a large space, providing a powerful tool

Conclusion

A knowledge–data dual-driven molecular property prediction framework, named KPI, has been developed to predict MPs, BPs, and FPs of electrolyte molecules, discover electrolyte chemistry knowledge, and assist in advanced electrolyte design. The KPI framework gathers molecular properties, performs initial screening, and organizes it into structured datasets. Through statistical descriptions and cluster analysis, the KPI framework preliminarily extracts macroscopic knowledge and simultaneously delineates feasible regions for the discovery of potential molecules. Interpretable machine learning further explores the structure–property relationship at the atomic level, indicating that MPs are primarily influenced by the nature of bonds, while BPs and FPs are mainly affected by the number and mass of atoms, as well as electronic properties. By embedding the discovered knowledge into the prediction models and finely tuning the purity and flow of knowledge, the KPI framework achieves a low MAE of 10.5, 4.6, and 4.8 K for MP, BP, and FP predictions, respectively, which reached the SOTA results in 18 of 20 datasets. The KPI framework demonstrated excellent talents in designing promising electrolyte molecules. By utilizing molecular neighbor search and high-throughput screening, 15 promising molecules with CAS ID and 14 without CAS ID were predicted for wide-temperature-range and high-safety batteries. The KPI framework significantly deepens the understanding of molecular structure–property relationships and accelerates the high-throughput screening of target molecules, thereby greatly advancing the development of wide-temperature-range and high-safety electrolytes.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (T2322015 and 22109086), National Key Research and Development Program (2021YFB2500300 and 2022YFB2404402), and Beijing Municipal Natural Science Foundation (Z200011), and the Tsinghua University Initiative Scientific Research Program. The authors acknowledged the support from Tsinghua National Laboratory for Information Science and Technology for theoretical simulations. We thank Rui Zhang, Zheng Li, Rui Tan, Ming-Kang Liu, and Shi-Qiu Yin for their helpful discussion.

Keywords: rechargeable battery electrolyte • machine learning • molecular property prediction • structure–property relationship • knowledge–data dual driven

[1] G. Harper, R. Sommerville, E. Kendrick, L. Driscoll, P. Slater, R. Stolkin, A. Walton, P. Christensen, O. Heidrich, S. Lambert, A. Abbott, K. Ryder, L. Gaines, P. Anderson, *Nature* **2019**, *575*, 75–86; C. P. Grey, D. S. Hall, *Nat. Commun.* **2020**, *11*, 6279; Z. Zhu, T. Jiang, M. Ali, Y. Meng, Y. Jin, Y. Cui, W. Chen, *Chem. Rev.* **2022**, *122*, 16610–16751.

[2] Y. Feng, L. Zhou, H. Ma, Z. Wu, Q. Zhao, H. Li, K. Zhang, J. Chen, *Energy Environ. Sci.* **2022**, *15*, 1711–1759.

[3] M.-T. F. Rodrigues, G. Babu, H. Gullapalli, K. Kalaga, F. N. Sayed, K. Kato, J. Joyner, P. M. Ajayan, *Nat. Energy* **2017**, *2*, 17108; M. C. Smart, B. V. Ratnakumar, R. C. Ewell, S. Surampudi, F. J. Puglia, R. Gitzendanner, *Electrochim. Acta* **2018**, *268*, 27–40.

[4] X. Lin, M. Salari, L. M. R. Arava, P. M. Ajayan, M. W. Grinstaff, *Chem. Soc. Rev.* **2016**, *45*, 5848–5887; K. Liu, Y. Liu, D. Lin, A. Pei, Y. Cui, *Sci. Adv.*, *4*, eaas9820.

[5] A. Gupta, A. Manthiram, *Adv. Energy Mater.* **2020**, *10*, 2001972; Z. Li, Y.-X. Yao, S. Sun, C.-B. Jin, N. Yao, C. Yan, Q. Zhang, *Angew. Chem. Int. Ed.* **2023**, *62*, e202303888; X.-B. Cheng, R. Zhang, C.-Z. Zhao, Q. Zhang, *Chem. Rev.* **2017**, *117*, 10403–10473.

[6] X. Fan, X. Ji, L. Chen, J. Chen, T. Deng, F. Han, J. Yue, N. Piao, R. Wang, X. Zhou, X. Xiao, L. Chen, C. Wang, *Nat. Energy* **2019**, *4*, 882–890; Y. Ou, P. Zhou, W. Hou, X. Ma, X. Song, S. Yan, Y. Lu, K. Liu, *J. Energy Chem.* **2024**, *94*, 360–392.

[7] Y. Yang, Y. Yin, D. M. Davies, M. Zhang, M. Mayer, Y. Zhang, E. S. Sablina, S. Wang, J. Z. Lee, O. Borodin, C. S. Rustomji, Y. S. Meng, *Energy Environ. Sci.* **2020**, *13*, 2209–2219.

[8] X. Fan, L. Chen, O. Borodin, X. Ji, J. Chen, S. Hou, T. Deng, J. Zheng, C. Yang, S.-C. Liou, K. Amine, K. Xu, C. Wang, *Nat. Nanotechnol.* **2018**, *13*, 715–722; Y. Yamada, J. Wang, S. Ko, E. Watanabe, A. Yamada, *Nat. Energy* **2019**, *4*, 269–280; K. Xu, *Chem. Rev.* **2014**, *114*, 11503–11618; Z. Li, L.-P. Hou, N. Yao, X.-Y. Li, Z.-X. Chen, X. Chen, X.-Q. Zhang, B.-Q. Li, Q. Zhang, *Angew. Chem. Int. Ed.* **2023**, *62*, e202309968; Z. Li, L. Yu, C.-X. Bi, X.-Y. Li, J. Ma, X. Chen, X.-Q. Zhang, A. Chen, H. Chen, Z. Zhang, L.-Z. Fan, B.-Q. Li, C. Tang, Q. Zhang, *SusMat* **2024**, *4*, e191; D. Zhang, L. Li, W. Zhang, M. Cao, H. Qiu, X. Ji, *Chinese. Chem. Lett.* **2023**, *34*, 107122.

[9] J. Xu, J. Zhang, T. P. Pollard, Q. Li, S. Tan, S. Hou, H. Wan, F. Chen, H. He, E. Hu, K. Xu, X.-Q. Yang, O. Borodin, C. Wang, *Nature* **2023**, *614*, 694–700.

[10] J. Wang, Y. Yamada, K. Sodeyama, E. Watanabe, K. Takada, Y. Tateyama, A. Yamada, *Nat. Energy* **2018**, *3*, 22–29.

[11] T. Lombardo, M. Duquesnoy, H. El-Bouysidy, F. Årén, A. Gallo-Bueno, P. B. Jørgensen, A. Bhowmik, A. Demortière, E. Ayerbe, F. Alcaide, M. Reynaud, J. Carrasco, A. Grimaud, C. Zhang, T. Vegge, P. Johansson, A. A. Franco, *Chem. Rev.* **2022**, *122*, 10899–10969; A. Benayad, D. Diddens, A. Heuer, A. N. Krishnamoorthy, M. Maiti, F. L. Cras, M. Legallais, F. Rahmanian, Y. Shin, H. Stein, M. Winter, C. Wölke, P. Yan, I. Cekic-Laskovic, *Adv. Energy Mater.* **2022**, *12*, 2102678.

[12] X. Chen, X. Liu, X. Shen, Q. Zhang, *Angew. Chem. Int. Ed.* **2021**, *60*, 24354–24366; Y.-C. Gao, N. Yao, X. Chen, L. Yu, R. Zhang, Q. Zhang, *J. Am. Chem. Soc.* **2023**, *145*, 23764–23770; R. Atwi, Y. Chen, K. S. Han, K. T. Mueller, V. Murugesan, N. N. Rajput, *Nat. Comput. Sci.* **2022**, *2*, 112–122; Y. Cao, A. Aspuru-Guzik, *Nat. Comput. Sci.* **2024**, *4*, 89–91.

[13] R. Abramowitz, S. H. Yalkowsky, *Pharmaceut. Res.* **1990**, *7*, 942–947; Z. Jianzhong, Z. Biao, Z. Suoqi, W. Renan, Y. Guanghua, *Ind. Eng. Chem. Res.* **1998**, *37*, 2059–2060.

[14] L. Constantinou, R. Gani, *AIChE J.* **1994**, *40*, 1697–1710; K. G. Joback, R. C. Reid, *Chem. Eng. Commun.* **1987**, *57*, 233–243.

[15] R. Gani, *Curr. Opin. Chem. Eng.* **2019**, *23*, 184–196; G. Di Nicola, M. Falone, M. Pierantozzi, R. Stryjek, *Ind. Eng. Chem. Res.* **2014**, *53*, 13804–13809.

[16] N. Yao, X. Chen, X. Shen, R. Zhang, Z.-H. Fu, X.-X. Ma, X.-Q. Zhang, B.-Q. Li, Q. Zhang, *Angew. Chem. Int. Ed.* **2021**, *60*, 21473–21478; N. Yao, L. Yu, Z.-H. Fu, X. Shen, T.-Z. Hou, X. Liu, Y.-C. Gao, R. Zhang, C.-Z. Zhao, X. Chen, Q. Zhang, *Angew. Chem. Int. Ed.* **2023**, *62*, e202305331; N. Yao, X. Chen, Z.-H. Fu, Q. Zhang, *Chem. Rev.* **2022**, *122*, 10970–11021.

[17] A. E. Sifain, B. M. Rice, S. H. Yalkowsky, B. C. Barnes, *J. Mol. Graph.* **2021**, *105*, 107848.

[18] P. M. Agrawal, B. M. Rice, D. L. Thompson, *J. Chem. Phys.* **2003**, *119*, 9617–9627; Y. Zhang, E. J. Maginn, *J. Chem. Phys.* **2012**, *136*, 144116.

[19] A. R. Katritzky, V. S. Lobanov, M. Karelson, *Chem. Soc. Rev.* **1995**, *24*, 279–287; A. Tropsha, O. Isayev, A. Varnek, G. Schneider, A. Cherkasov, *Nat. Rev. Drug Discov.* **2024**, *23*, 141–155; U. P. P, M. Suresh, F. T. Tolasa, E. Bonyah, *Sci. Rep.* **2024**, *14*, 13150.

[20] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Rev.* **1996**, *96*, 1027–1044; A. R. Katritzky, M. Kuanar, S. Slavov, C. D. Hall, M. Karelson, I. Kahn, D. A. Dobchev, *Chem. Rev.* **2010**, *110*, 5714–5789.

[21] C.-B. Jin, N. Yao, Y. Xiao, J. Xie, Z. Li, X. Chen, B.-Q. Li, X.-Q. Zhang, J.-Q. Huang, Q. Zhang, *Adv. Mater.* **2023**, *35*, 2208340.

[22] N. Zhang, T. Deng, S. Zhang, C. Wang, L. Chen, C. Wang, X. Fan, *Adv. Mater.* **2022**, *34*, 2107899.

[23] M. Qin, M. Liu, Z. Zeng, Q. Wu, Y. Wu, H. Zhang, S. Lei, S. Cheng, J. Xie, *Adv. Energy Mater.* **2022**, *12*, 2201801.

[24] T. Taskovic, A. Eldesoky, C. P. Aiken, J. R. Dahn, *J. Electrochem. Soc.* **2022**, *169*, 100547.

[25] Q. Wang, L. Jiang, Y. Yu, J. Sun, *Nano Energy* **2019**, *55*, 93–114.

[26] G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang, G. Ke, in *The Eleventh International Conference on Learning Representations*, **2023**.

[27] Y. Lin, Z. Yang, X. Zhang, Y. Liu, G. Hu, S. Chen, Y. Zhang, *Energy Storage Mater.* **2023**, *58*, 184–194.

[28] L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem Inf. Model.* **2012**, *52*, 2864–2875; R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *Sci. Data* **2014**, *1*, 140022.

[29] S. Tan, Z. Shadike, X. Cai, R. Lin, A. Kludze, O. Borodin, B. L. Lucht, C. Wang, E. Hu, K. Xu, X.-Q. Yang, *Electrochem. Energy Rev.* **2023**, *6*, 35.

[30] Y.-G. Cho, Y.-S. Kim, D.-G. Sung, M.-S. Seo, H.-K. Song, *Energy Environ. Sci.* **2014**, *7*, 1737–1743.

[31] L. Luo, K. Chen, H. Chen, H. Li, R. Cao, X. Feng, W. Chen, Y. Fang, Y. Cao, *Adv. Mater.* **2024**, *36*, 2308881.

[32] Z. Wang, Z. He, Z. Wang, J. Yang, K. Long, Z. Wu, G. Zhou, L. Mei, L. Chen, *Chem. Sci.* **2024**, *15*, 13768–13778.

[33] Q. Hou, P. Li, Y. Qi, Y. Wang, M. Huang, C. Shen, H. Xiang, N. Li, K. Xie, *ACS Energy Lett.* **2023**, *8*, 3649–3657.

[34] T. Zheng, J. Xiong, X. Shi, B. Zhu, Y.-J. Cheng, H. Zhao, Y. Xia, *Energy Storage Mater.* **2021**, *38*, 599–608.

[35] S. A. Langevin, M. M. McGuire, N. Q. Le, E. Ragasa, T. Hamann, G. Ferguson, C. Chung, J. Domenico, J. S. Ko, *J. Mater. Chem. A* **2022**, *10*, 19972–19983; T. Qin, H. Yang, L. Wang, W. Xue, N. Yao, Q. Li, X. Chen, X. Yang, X. Yu, Q. Zhang, H. Li, *Angew. Chem. Int. Ed.* **2024**, *63*, e202408902.

[36] H. Muller, N. A. Alawani, I. A. Naqvi, M. M. Al-Shammari, R. Y. Al-Bakor, F. M. Adam, *Fuel* **2024**, *358*, 130066; D. A. Saldana, L. Starck, P. Mougin, B. Rousseau, L. Pidol, N. Jeuland, B. Creton, *Energy & Fuels* **2011**, *25*, 3900–3908.

10.1002/anie.202416506

Table of Contents