SCIENCE CHINA
Chemistry

•COMMUNICATIONS•

November 2024 Vol.67 No.11: 3681–3687
https://doi.org/10.1007/s11426-024-2163-1

A universal strategy of multi-objective active learning to accelerate the discovery of organic electrode molecules

Jiayi Du[1,2], Jun Guo[1,3], Wei Liu[1], Ziwei Li[1], Gang Huang[1,2*] & Xinbo Zhang[1,2*]

[1]*State Key Laboratory of Rare Earth Resource Utilization, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, China;*
[2]*School of Applied Chemistry and Engineering, University of Science and Technology of China, Hefei 230026, China;*
[3]*School of Materials Science and Engineering, Changchun University of Science and Technology, Changchun 130022, China*

Received April 10, 2024; accepted June 20, 2024; published online September 29, 2024

Organic electrode molecules hold significant potential as the next generation of cathode materials for Li-ion batteries. In this study, we have introduced a multi-objective active learning framework that leverages Bayesian optimization and non-dominated sorting genetic algorithms-II. This framework enables the selection of organic molecules characterized by high theoretical energy density and low gap (LUMO-HOMO) (LUMO, lowest unoccupied molecular orbital; HOMO, highest occupied molecular orbital). Remarkably, after only two cycles of active learning, the determination of coefficient can reach 0.962 for theoretical energy density and 0.920 for the gap with a modest dataset of 300 molecules, showcasing superior predictive capabilities. The 2,3,5,6-tetrafluorocyclohexa-2,5-diene-1,4-dione, selected by non-dominated sorting genetic algorithms-II, has been successfully applied to Li-ion batteries as cathode materials, demonstrating a high capacity of 288 mAh g^{-1} and a long cycle life of 1,000 cycles. This outcome underscores the high reliability of our framework. Furthermore, we have also validated the universality and transferability of our framework by applying it to two additional databases, the QM9 and OMEAD. When the training dataset of the model includes at least 500 molecules, the determination of coefficient essentially reaches approximately 0.900 for four targets: gap, reduction potential, LUMO, and HOMO. Therefore, the universal framework in our work provides innovative insights applicable to other domains to expedite the screening process for target materials.

organic electrode molecules, Li-ion batteries, active learning, multi-objective Bayesian optimization

Citation: Du J, Guo J, Liu W, Li Z, Huang G, Zhang X. A universal strategy of multi-objective active learning to accelerate the discovery of organic electrode molecules. *Sci China Chem*, 2024, 67: 3681–3687, https://doi.org/10.1007/s11426-024-2163-1

With the utilization of fossil fuels, environmental issues have attracted a great deal of concern [1]. The storage and conversion of clean energy in the replacement of fossil fuel is supported by many countries, companies, and institutes in the world [2,3]. Hence, Li-ion batteries (LIBs) as a kind of energy storage device for clean energy are widely applied to mobile electronic devices and electric vehicles [4,5]. However, conventional LIBs' cathode materials, such as $LiFePO_4$, $LiCoO_2$, and $LiMn_2O_4$, have exposed themselves disadvantages like low capacity and high cost, which cannot gradually fulfill the booming demand of society [5,6]. Especially, these inorganic materials mainly originate from ores rather than renewable resources [7]. Therefore, developing new cathode materials is extremely urgent [4,8,9]. Organic electrode molecules (OEMs) have been centered on recently due to their distinctive characteristics [4,5,10]. Firstly, the primary constituents of OEMs are earth-abundant elements, such as carbon, oxygen, hydrogen, and nitrogen, which render them readily available [6,7,11]. Secondly, by manipulating the quantity of active functional groups in conjunction with other inactive components of the OEMs, the molecules can be effortlessly designed with high capacity

*Corresponding authors (email: ghuang@ciac.ac.cn; xbzhang@ciac.ac.cn)

and voltage [6,11–13]. In the meantime, the stability of OEMs should be taken into consideration on the drawing board. However, currently, most OEMs are explored and exploited by trial-and-error experiments, making it difficult to explore more molecules in the enormous chemical space [14,15].

In recent years, big data combined with machine learning (ML), regarded as the 'fourth paradigm of science' [16], have played more and more significant roles in chemistry and material fields [17,18]. In particular, lots of research about cathode materials [15,19], solid-state electrolytes [20,21], and other related energy storage and conversion fields [22,23] combined with ML boomingly emerge. The active learning (AL), a subfield of ML, has been also applied into electrocatalysts [24,25], redox flow batteries [26], and organic synthesis [27] owing to its distinctive merits [28]. Specifically, AL has the capability to acquire as many high-quality samples as possible by labeling a minimal number of samples from the unlabeled space [28]. This implies that the optimal molecules or other materials within the chemical space can be synthesized or calculated using only a limited number of experiments and calculations. For example, Rao *et al*. [29] constructed an AL framework that was capable of generating high-entropy alloy chemical space. In every cycle of AL, they only synthesized three samples recommended by the AL to obtain those high-entropy alloys with low thermal expansion coefficients. Lu and co-workers [30] proposed an AL framework with margin sampling to select two-dimensional ferromagnets with high Curie temperatures. In addition, Bayesian optimization combined with Gaussian process regression (GPR) is also a common practice in AL [25,26,31,32]. Furthermore, the multi-objective active learning (MOAL) has been adopted to screen organic conductors [32], redox active molecules [26], and molecular photoswitches [33]. MOAL has the capability to simultaneously select multi-property molecules, thus reducing screening time compared with the sequential step-by-step screening with multi-property molecules. Nevertheless, a persistent issue arises when selecting molecules with one desirable property, which may be at the expense of another desirable property [27]. This serves as the stumbling blocks of MOAL.

In the present work, we have developed a MOAL framework to swiftly and automatically identify those multi-properties OEMs from our created database (OQEMDB) containing 27463 quinone molecules. The theoretical energy density (TED) and gap are the two key properties of OEMs in LIBs. Thus, the selection of OEMs with high TED and low gaps is the main task of MOAL. Moreover, the MOAL framework is composed of a machine learning model that combines convolution neural network with GPR, and Pareto front that is achieved by non-dominated sorting genetic algorithms-II (NSGA-II) [34]. Furthermore, to gather high-performance potential OEMs, we have carried out 10 cycles of MOAL. The 1,100 molecules (including initial 100 molecules) have been selected and 4,400 tasks have been performed by density functional theory (DFT) calculations. The 2,3,5,6-tetrafluorocyclohexa-2,5-diene-1,4-dione (TFDD) determined by NSGA-II from the top 100 molecules among the 1,100 candidates has been applied into LIBs as high-perofrmance cathode materials. The successful implementation of our framework provides new insight to screen out OEMs or other electrode materials with designable multi-properties. Importantly, explicit mathematical formulas for both TED and gap have been sought to further quantify the structure-property relationship between these two targets and OEMs.

Quinone molecules have been a lot of traction in the exploration of new OEMs [35–37]. To dig out as many potential quinone molecules as possible, a chemical space was designed that involved a random combination of 12 quinone molecules and 12 functional groups (Figures S1 and S2, Supporting Information online), and these generated molecules were stored in the form of SMIELS (simplified molecular input line entry system) [38]. More details about the construction of our database were shown in Note 1 in the Supporting Information online. It should be mentioned that the application of a brute-force functional group substitution method could potentially generate an extensive chemical space, in which numerous molecules might either be non-existent or unsuitable for OEMs. According to our previous chemical experience and knowledge, most OEMs have symmetrical structures. Hence, serial methods were utilized to lessen the chemical space while symmetric molecules were identified (Figure 1a). Initially, the molecule would be chosen if the atomic numbers of every element within it were even. Then, these selected molecules underwent the optimization using the MMFF94 [39] method from RDKit to determine their three-dimensional positions. Nonetheless, due to the intricate structures, some molecules could not be optimized efficiently, and were consequently discarded. Following this, the SYVA [40], a software designed for calculating point group of molecules, was employed to screen out the symmetric molecules. Those molecules with the C_1 point group were removed. At last, a database containing 27,463 quinone molecules (OQEMDB) was constructed. Moreover, the t-distribution stochastic neighbor embedding (t-SNE) method was harnessed to visualize the database. Morgan fingerprints were introduced to represent these molecules in the course of visualization [41] (Note 2 in the Supporting Information online). Intriguingly, the theoretical capacity (TC) of molecules gradually increases from lower right to upper left in the visualization space (Figure 1b), and similarly, the relative molecular mass (RMM) gradually increases from down to up (Figure S3).

Our goals were to screen out those molecules with high

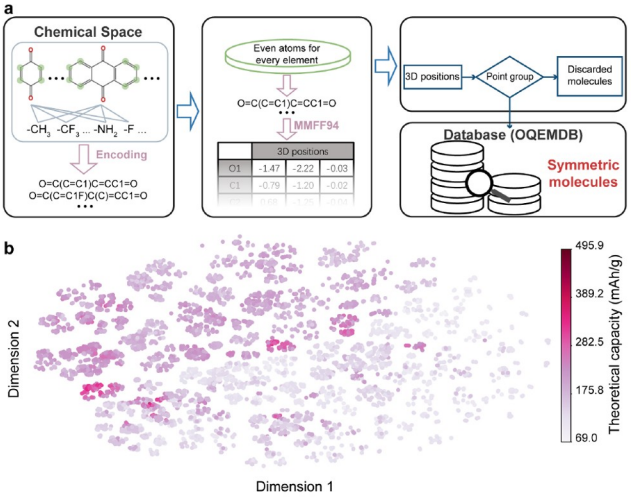

Figure 1 (Color online) The creation and visualization of OQEMDB. (a) Schematic of the building process of quinone database. (b) Visualization of the quinone database by t-SNE. The data points are colored according to the theoretical capacity.

TED as well as low gaps from our constructed database for LIB's cathode applications. Notably, the low gap implied fast intramolecular charge transfer [42]. However, it was impractical to adopt high-throughput experiments or calculations across all molecules in the database. Consequently, ML was deemed suitable for addressing this issue.

Prior to the development of a machine learning model, it was imperative to select suitable descriptors that aptly represented organic molecules. In this study, three descriptors were utilized to accurately depict organic molecules: the many-body tensor representation (MBTR), atom-centered symmetry functions (ACSF), and smooth overlap of atomic orbitals (SOAP). It was convenient to extract these descriptors by the python package DScribe [43] without tedious manual selection [19,26,44,45]. Specifically, the MBTR describes the interaction of each element in one molecule, while the SOAP and ACSF record the sum of the local environment information of every atom in one molecule. For example, the benzoquinone (BQ), 1,4-naphthoquinone (NQ), and 9,10-anthraquinone (AQ) are similar to each other. Thus, their representations from the three descriptors look alike (Figure 2a), but the intensity of peaks is slightly different according to the corresponding molecules. More details can be found in Note 3, Table S1 (Supporting Information online), and the reference [43].

Subsequently, an AL model (GNGPR) using GoogLeNet neural network (Note 4, Tables S2 and S3) was combined with GPR (Figure 2b, Note 4, Eq. S1 in the Supporting Information online) to screen out those OEMs with good performance. The basic constructions of BNConv2d and Inception are shown in Figure S4. Notably, in the training process of GNGPR, it was divided into two stages. First, the GoogLeNet underwent 150 cycles of training. After 100 cycles, the parameters of GoogLeNet model were reserved at the cycle that yielded the lowest value of the loss function. Second, the output from the penultimate layer of the reserved GoogLeNet model would serve as the input for GPR model, and followingly, the predicted results were obtained via training the GPR model. Additionally, 100 molecules were selected randomly as the initial training dataset for GNGPR model (Figure S5). The TED (Eq. S5) is equal to the RP (Note 5, Eqs. S2 and S3) multiplied by the TC (Eq. S4), and the gap is equal to the lowest unoccupied molecular orbital (LUMO) minus highest occupied molecular orbital (HOMO). It is worthy of mentioning that the RP usually serves as the target of OEMs in previous reports [14,15,45]. Here, we have replaced the RP with TED, attributing to the availability of TC to directly evaluate the energy density of LIBs. However, it is still necessary that the RP should be obtained through four states of molecules by DFT calculations (Figure S6).

The determination of coefficient (R^2) was utilized to assess the efficacy of GNGPR model for TED and gap. In the meantime, the multi-objective merit of our proposed model was also further displayed by its training with RP. Through five-fold cross-validation, the R^2 of gap, TED, and RP model based on the MBTR descriptor are found to be 0.790±0.091, 0.800±0.175, and 0.850±0.093, respectively (Figure 2c). In comparison, the R^2 values for the GPR model predicting all three targets are consistently below 0.5. However, the R^2 of the GoogLeNet model, which only predicts the RP, exceeds 0.6. Notably, it is below 0 observed for ACSF when the gap of the initial dataset has been predicted by both the GPR and GoogLeNet models (Figure S7a). Similarly, this trend is also observed for SOAP (Figure S7b). Nevertheless, the GNGPR still behaves well. The aforementioned data indicates that the GNGPR model exhibits superior predictive accuracy compared with the other two models. Furthermore, compared with ACSF and SOAP, the model employing MBTR on the three targets demonstrates greater robustness and improved performance (Figure 2d). As a result, the GNGPR model combined with the MBTR descriptor is better competent for MOAL.

Multi-objective Bayesian optimization framework (MOBO), composed of Bayesian optimization (Note 6) and NSGA-II, was harnessed to build the MOAL loop (Figure 3a). Expected improvement (EI; Eqs. S6 and S7) served as the acquisition function, which is the metric of good candidates. In each iteration cycle, the EI of all molecules except the previously chosen molecules would be calculated by individual GNGPR for TED and gap. Then, 100 molecules would be chosen based on the Pareto front achieved by NSGA-II and calculated by DFT in the next cycle. The dataset used for training GNGPR would also be updated automatically. Particularly, the training and test sets were always split into an 8:2 ratio during the MOAL loop.

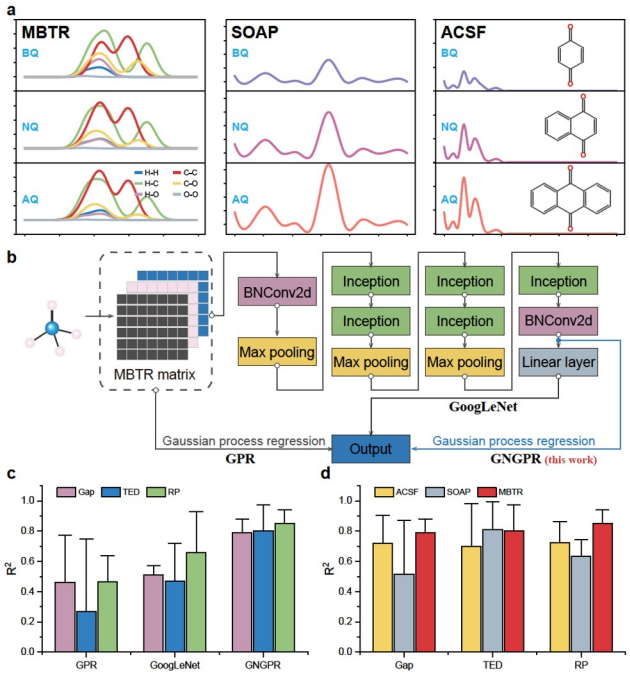

Figure 2 (Color online) (a) The representations of BQ, NQ, and AQ based on the three descriptors, MBTR, SOAP, and ACSF. (b) The three models of GPR, GoogLeNet, and GNGPR. (c) The evaluations of GPR, GoogLeNet, and GNGPR with MBTR. (d) The evaluations of GNGPR with three descriptors, ACSF, SOAP, and MBTR.

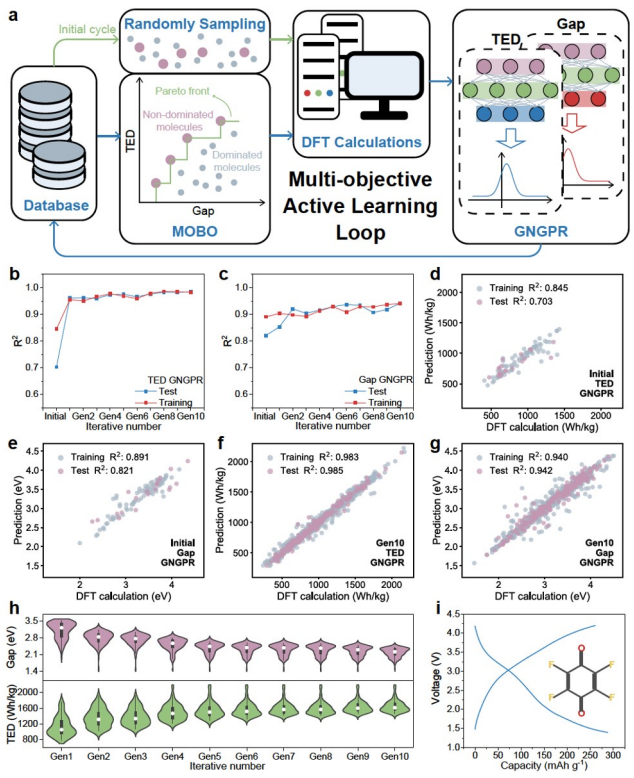

Figure 3 (Color online) The results of MOAL. (a) Schematic of the MOAL process. The R^2 of each cycle in the MOAL for TED (b) and gap (c). The predicted results of the initial training dataset for TED (d) and gap (e). The predicted results of the training dataset of Gen10 for TED (f) and gap (g). (h) The distribution of the top 100 molecules in terms of TED and gap after every cycle of MOAL. The white point is the median, and the black rectangle represents the quarter to three quarters for TED and gap distribution in the violin diagram. (i) The full charge and discharge curves of TFDD at 0.1 A g^{-1}.

Moreover, Gen1 denoted the molecules generated from the initial 100 molecules (Initial). Similarly, Gen2 represented the molecules generated from the updated dataset, including both Gen1 and Initial. Furthermore, from Figure 3b, c, the R^2 scores keep increasing over time, which suggests that the MOAL can optimize itself during the training process. The R^2 values of the initial test set are only 0.703 for TED (Figure 3d) and 0.821 for gap (Figure 3e). Nonetheless, at Gen10, the GNGPR model can achieve high scores of 0.985 for TED (Figure 3f) and 0.942 for gap (Figure 3g). Additionally, the training results of GoogLeNet and GNGPR during the MOAL are also shown in Figures S8–S13. The loss values almost remain unchanged after 50 epochs (Figures S8 and S9), indicating that 150 epochs of training for GoogLeNet are adequate. Concurrently, the performance of GoogLeNet for TED and the gap continues to be improved during the MOAL (Figures S10 and S11). Moreover, the top 100 molecules for TED and gap are immediately updated every loop (Figure 3h). Importantly, since Gen6, both TED and gap have maintained their violin-like shape with minimal alterations, indicating that most high-performance molecules have been selected by our MOAL. Consequently, the MOAL could be terminated after Gen10. Table S4 displays the top 100 molecules selected by NSGA-II, which both have high TED and low gap. Considering the molecular synthesizability and stability, the TFDD has been selected for further experimental verification. When implemented to the cathode of LIBs (Note 7), it achieves a capacity of 288 mAh g^{-1} at 0.1 A g^{-1} (Figure 3i) and a cycle lifetime of 1,000 cycles at 1 A g^{-1} (Figure S14), suggesting the good ability of experimental guidance from the MOAL.

Although the MOAL model displays excellent predicting ability, a significant limitation of the neural network models is their lack of interpretability, particularly in applications where the transparency of decision-making is a crucial requirement. To this end, a two-dimensional plane constituted by the output vectors originating from the previous layer of linear layer of GoogLeNet (Figure 2b) using the t-SNE has been constructed. Figure S15 illustrates the training space of TED and the gap. For TED, the data points gradually increase from the upper left to the lower right. Meanwhile, the molecules with higher TED tend to gather together. The gap is also in a similar status. Particularly, even in the prediction space of TED and gap (Figure 4a, b), the 1,100 molecules also showcase the similar distribution to the training space, demonstrating that the MOAL can achieve high accuracy to search those molecules with higher TED or lower gap.

In addition, the MOAL is capable of extracting chemical information from other insights, such as the number of rings

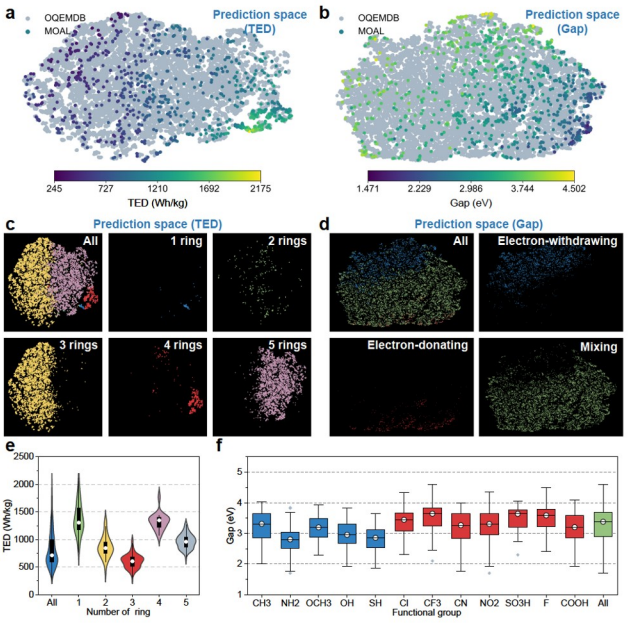

Figure 4 (Color online) The visualization results of MOAL. The visualization of the prediction space from GoogLeNet for TED (a) and gap (b) by t-SNE. The data points of the MOAL colored by TED and gap, respectively. (c) The data points colored with their number of rings in the prediction space of TED. (d) The data points colored with the functional group types (electron-donating, electron-withdrawing, and mixing groups) in the prediction space of gap. (e) The distribution of the number of rings according to the TED of MOAL. (f) The distribution of functional groups according to the gap of MOAL.

and functional groups. For TED, it is evident that the regions with varying numbers of rings in the prediction space exhibit distinct characteristics (Figure 4c). It primarily pertains to the TC (Eq. S4), wherein molecules possessing three rings exhibit the lowest TC (Table S5). Consequently, these molecules are situated in the leftmost low-TED region. Conversely, the pyrene-4,5,9,10-tetraone and benzoquinone derivatives are positioned in the bottom right high-TED region. The distinct distributions clearly demonstrate that the MAOL can discern the key factors influencing TED. Specifically, the distribution of TED across various rings in the MOAL database serves to validate the authenticity of the prediction space (Figure 4e). However, it is unfortunate that no discernible pattern exists for the distribution of molecules within the gap prediction space based on the number of rings (Figure S16). Furthermore, the functional groups are categorized into electron-withdrawing groups (–CN, –COOH, $–CF_3$, $–NO_2$, –F, –Cl, and $–SO_3H$) and electron-donating groups ($–CH_3$, $–NH_2$, –OH, $–OCH_3$, and –SH). Following this classification, the molecules are further classified into three categories: those exclusively containing electron-withdrawing groups, those exclusively containing electron-donating groups, and those containing both types of groups (mixing groups) (Figure 4d and Figure S17). In particular, in the prediction space of gap, it is obvious that the electron-withdrawing groups are situated on the upper left, while the electron-donating groups are located on the lower right. This suggests that the electron-withdrawing groups could increase the gap, whereas the electron-donating groups reduce the gap. Meanwhile, the gap of those molecules containing $–NH_2$, –OH, and –SH is indeed lower than that of those molecules containing –F, $–CF_3$, and $–SO_3H$ (Figure 4f). Consequently, the visualization through the t-SNE underscores the robust learning capability of our MOAL in relation to various targets.

As described above, the GNGPR model has played a significant role in our work. Apart from the quinone OEMs, we anticipate that this model can also be competent for other types of OEMs or other systems. Therefore, another two databases were chosen to test the universality of GNGPR. One of the two databases was created by Carvalho *et al.* [15], which was composed of more than 26,000 molecules (referred OMEAD) and their physical and chemical properties, such as HOMO, LUMO, RP, and oxidation potential. Another database was the QM9 dataset which contained more than 133 thousand molecules with molecular geometric, energetic, electronic, and thermodynamic properties. For OMEAD, the model was trained based on the gap, HOMO, LUMO, and RP with different dataset sizes (100, 500, 1,000, 3,000, and 5,000) using 5-fold cross validation (Figure 5a). For QM9, a similar processing procedure to OMEAD was followed, except no RP dataset in the QM9 (Figure 5b). According to the training outcomes, when the training size is set at 100, only for gap, the R^2 of test set can reach the value of 0.839±0.158 (OMEAD) and 0.909±0.105 (QM9). The R^2 of other targets falls below 0.8, with the exception of LUMO (QM9). However, when the training size is increased to 500, all models exhibit satisfactory performance, most of which surpass 0.9, except for HOMO. From 100 to 1,000 molecules, the improvement of GNGPR for all the targets in both databases is similar to our MOAL process, meaning that our model is universal and worth popularizing. Conceivably, for 1,000 to 5,000 molecules, while there may be a slight increase in model performance, it is not significant. Specifically, Figure 5c–e and Figure S18 present the predicted outcomes of QM9 and OMEAD, respectively, with a training size of 4,000 and a test size of 1,000. The R^2 scores for all targets, excluding HOMO, exceed 0.95. However, the R^2 scores for the HOMO model stand at 0.949 for QM9 and 0.925 for OMEAD, thereby demonstrating the robust performance of GNGPR.

While the black-box model of GNGPR is adept at learning essential chemical information for molecules and possesses excellent predictive ability, it falls short in defining or quantifying the influencing factors of targets. Herein, the sure-independence-screening-and-sparsifying-operator (SISSO) algorithm [46] was utilized to quantify both physical and chemical descriptors to create a formula to fit the targets. Accordingly, a total of 93 descriptors were meticu-

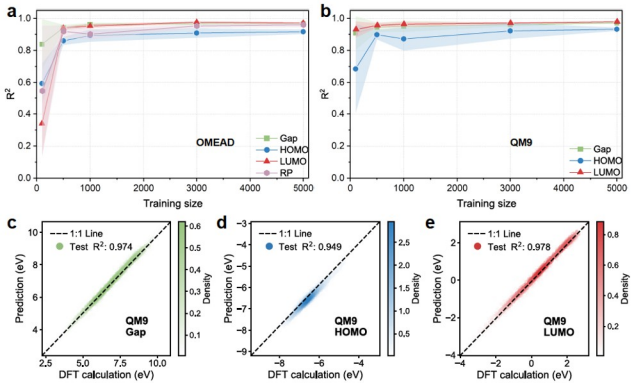

Figure 5 (Color online) The training results of GNGPR for different targets based on OMEAD (a) and QM9 (b) databases by 5-fold cross-validation under different training sizes. The predicted results of gap (c), HOMO (d), LUMO (e) for QM9 using GNGPR. The training set and test set are split into 4:1.

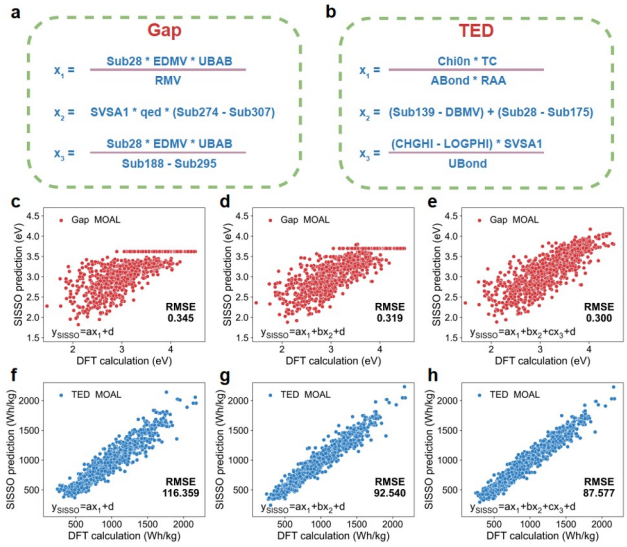

Figure 6 (Color online) The descriptors and analysis of SISSO. The three descriptors generated by SISSO for gap (a) and TED (b). The predicted results of SISSO formula with one (c), two (d), and three (e) descriptors for the gap of MOAL. The predicted results of SISSO formula with one (f), two (g), and three (h) descriptors for TED of MOAL.

lously selected and presented in Table S6, which were composed of the covalent radius and Pauli electronegative of atoms, molecular volume, type of bonds, functional groups, BCUT2D, and SlogP. These descriptors could function as the input for the SISSO algorithm, resulting in the generation of three distinct SISSO-descriptors (x_1, x_2, x_3) for gap (Figure 6a) and TED (Figure 6b), respectively. Linear regression models were employed to adjust the three SISSO-descriptors to fitting the gap and TED. The parameters '*a*', '*b*', '*c*', and '*d*' are detailed in Tables S7 and S8.

As for the gap, three linear regression models were built, and they were $y_{SISSO}=ax_1+d$, $y_{SISSO}=ax_1+bx_2+d$, and $y_{SISSO}=ax_1+bx_2+cx_3+d$. Then the predicted results of MOAL are displayed in Figure 6c–e and Table S9. It is observed that as the number of SISSO-descriptors increases, there is a corresponding decrease in the value of the root mean square error (RMSE). The descriptor x_1 is associated with the electron-donating groups and unsaturated bonds. Compared with x_1, x_2 and x_3 augment the predictive accuracy of gap in low and high ranges. This enhancement can be attributed to the supplement of electron-withdrawing groups like *Sub188* (*Sub274–Sub307*) and the global properties like *qed*, *SVSA1*. In addition, the three SISSO-descriptors of TED also demonstrate superior predictive performance (Figure 6f–h). Specifically, x_1 is analogous to Eq. S5, wherein the formula *Chi0n*/(*ABond***RAA*) acts as the RP. This illustrates that the SISSO algorithm can discover formula with physical and chemical relevance. Furthermore, x_2 and x_3 enhance the correction of TED in terms of the functional groups, atomic charge, and molecular polarity, respectively. Ultimately, the SISSO algorithm is utilized to fit the predicted values of TED and gap of all molecules in OQEMDB by MOAL (Figure S19). Inconceivably, the smaller RMSE of TED (72.061 Wh kg^{-1}) and gap (0.209 eV) emerges, illustrating the excellent robustness of SISSO. Besides, the SISSO still remains good prediction performance with small size of molecules, such as 50, 100 and 500 molecules (Figure S20), while GNGPR model reaching stable predicting level needs more than or equal to 300 molecules (Figure 3b, c). In brief, the explicit mathematical formula has been sought to further quantify the structure-property relationship of gap and TED and facilitate the analysis of gap and TED.

In summary, we have proposed a MOAL framework to rapidly and accurately search multi-properties OEMs for LIBs. The MOAL framework constructed based on the neural network and GPR model could integrate the NSGA-II to harmonize the exploration-exploitation tradeoff in multi-objective optimizations. The application of TFDD selected by our MOAL into LIBs exhibits a high energy density of 774 Wh kg^{-1} and a long cycle life of 1,000 cycles. Consequently, this framework demonstrates the outstanding capability of MOAL to assist the experimental synthesis of high-performance OEMs and accelerate the discovery of new materials with requisite properties for batteries. Moreover, in the field of materials science, it frequently occurs that merely small datasets are accessible for a specific task related to the material discovery. However, MOAL can achieve high scores of 0.920 for gap and 0.962 for TED with only 300 molecules, thus offering an approach to coping with small datasets in material science. Importantly, another advantage is that our framework is universal and transferable by the validation of two additional databases. Last but not the least, we have also employed the SISSO algorithm as assistance to accurately establish predictive and physically interpretable formulas that link the functional groups and molecular charge descriptors with the TED and gap. This approach can contribute to the guidance of designing high-performance OEMs.

Acknowledgements This work was supported by the National Key R&D Program of China (2022YFB2402200), the National Natural Science Foundation of China (92372206, 52271140, 52171194), the Jilin Province Science and Technology Development Plan Funding Project (YDZJ202301ZYTS545), the National Natural Science Foundation of China Excellent Young Scientists (Overseas), and the Youth Innovation Promotion Association CAS (2020230).

Conflict of interest The authors declare no conflict of interest.

Supporting information The supporting information is available online at http://chem.scichina.com and http://link.springer.com/journal/11426. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

1 Sun Q, Sun T, Du J, Li K, Xie H, Huang G, Zhang X. *Adv Mater*, 2023, 35: 2301088
2 Carley S, Konisky DM. *Nat Energy*, 2020, 5: 569–577
3 Xin S, Zhang X, Wang L, Yu H, Chang X, Zhao YM, Meng Q, Xu P, Zhao CZ, Chen J, Lu H, Kong X, Wang J, Chen K, Huang G, Zhang X, Su Y, Xiao Y, Chou SL, Zhang S, Guo Z, Du A, Cui G, Yang G, Zhao Q, Dong L, Zhou D, Kang F, Hong H, Zhi C, Yuan Z, Li X, Mo Y, Zhu Y, Yu D, Lei X, Zhao J, Wang J, Su D, Guo YG, Zhang Q, Chen J, Wan LJ. *Sci China Chem*, 2023, 67: 13–42
4 Kwon G, Ko Y, Kim Y, Kim K, Kang K. *Acc Chem Res*, 2021, 54: 4423–4433
5 Kim J, Kim Y, Yoo J, Kwon G, Ko Y, Kang K. *Nat Rev Mater*, 2022, 8: 54–70
6 Lu Y, Zhang Q, Li L, Niu Z, Chen J. *Chem*, 2018, 4: 2786–2813
7 Lu Y, Chen J. *Nat Rev Chem*, 2020, 4: 127–142
8 Hu Z, Zhao X, Li Z, Li S, Sun P, Wang G, Zhang Q, Liu J, Zhang L. *Adv Mater*, 2021, 33: 2104039
9 Cui H, Wang T, Huang Z, Liang G, Chen Z, Chen A, Wang D, Yang Q, Hong H, Fan J, Zhi C. *Angew Chem Int Ed*, 2022, 61: e202203453
10 Yang G, Zhu Y, Zhao Q, Hao Z, Lu Y, Zhao Q, Chen J. *Sci China Chem*, 2023, 67: 137–164
11 Lee S, Hong J, Kang K. *Adv Energy Mater*, 2020, 10: 2001445
12 Gan X, Song Z. *Sci China Chem*, 2023, 66: 3070–3104
13 Wu D, Xie Z, Zhou Z, Shen P, Chen Z. *J Mater Chem A*, 2015, 3: 19137–19143
14 Allam O, Kuramshin R, Stoichev Z, Cho BW, Lee SW, Jang SS. *Mater Today Energy*, 2020, 17: 100482
15 Carvalho RP, Marchiori CFN, Brandell D, Araujo CM. *Energy Storage Mater*, 2022, 44: 313–325
16 Agrawal A, Choudhary A. *APL Mater*, 2016, 4: 053208
17 Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. *Nature*, 2018, 559: 547–555
18 Ling C. *npj Comput Mater*, 2022, 8: 33
19 Park S, Park S, Park Y, Alfaruqi MH, Hwang JY, Kim J. *Energy Environ Sci*, 2021, 14: 5864–5874
20 He X, Bai Q, Liu Y, Nolan AM, Ling C, Mo Y. *Adv Energy Mater*, 2019, 9: 1902078
21 Ahmad Z, Xie T, Maheshwari C, Grossman JC, Viswanathan V. *ACS Cent Sci*, 2018, 4: 996–1006
22 Chen A, Zhang X, Zhou Z. *InfoMat*, 2020, 2: 553–576
23 Li X, Chen X, Bai Q, Mo Y, Zhu Y. *Sci China Chem*, 2023, 67: 276–290
24 Zhong M, Tran K, Min Y, Wang C, Wang Z, Dinh CT, De Luna P, Yu Z, Rasouli AS, Brodersen P, Sun S, Voznyy O, Tan CS, Askerka M, Che F, Liu M, Seifitokaldani A, Pang Y, Lo SC, Ip A, Ulissi Z, Sargent EH. *Nature*, 2020, 581: 178–183
25 Flores RA, Paolucci C, Winther KT, Jain A, Torres JAG, Aykol M, Montoya J, Nørskov JK, Bajdich M, Bligaard T. *Chem Mater*, 2020, 32: 5854–5863
26 Agarwal G, Doan HA, Robertson LA, Zhang L, Assary RS. *Chem Mater*, 2021, 33: 8133–8144
27 Torres JAG, Lau SH, Anchuri P, Stevens JM, Tabora JE, Li J, Borovika A, Adams RP, Doyle AG. *J Am Chem Soc*, 2022, 144: 19999–20007
28 Ren P, Xiao Y, Chang X, Huang PY, Li Z, Gupta BB, Chen X, Wang X. *ACM Comput Surv*, 2021, 54: 180
29 Rao Z, Tung PY, Xie R, Wei Y, Zhang H, Ferrari A, Klaver TPC, Körmann F, Sukumar PT, Kwiatkowski da Silva A, Chen Y, Li Z, Ponge D, Neugebauer J, Gutfleisch O, Bauer S, Raabe D. *Science*, 2022, 378: 78–85
30 Lu S, Zhou Q, Guo Y, Wang J. *Chem*, 2022, 8: 769–783
31 Xu S, Li J, Cai P, Liu X, Liu B, Wang X. *J Am Chem Soc*, 2021, 143: 19769–19777
32 Kunkel C, Margraf JT, Chen K, Oberhofer H, Reuter K. *Nat Commun*, 2021, 12: 2422
33 Griffiths RR, Greenfield JL, Thawani AR, Jamasb AR, Moss HB, Bourached A, Jones P, McCorkindale W, Aldrick AA, Fuchter MJ, Lee AA. *Chem Sci*, 2022, 13: 13541–13551
34 Deb K, Pratap A, Agarwal S, Meyarivan T. *IEEE Trans Evol Computat*, 2002, 6: 182–197
35 Lin Z, Shi HY, Lin L, Yang X, Wu W, Sun X. *Nat Commun*, 2021, 12: 4424
36 Hernández-Burgos K, Burkhardt SE, Rodríguez-Calero GG, Hennig RG, Abruña HD. *J Phys Chem C*, 2014, 118: 6046–6051
37 Kim KC, Liu T, Lee SW, Jang SS. *J Am Chem Soc*, 2016, 138: 2374–2382
38 Weininger D. *J Chem Inf Comput Sci*, 1988, 28: 31–36
39 Halgren TA. *J Comput Chem*, 1996, 17: 490–519
40 Gyevi-Nagy L, Tasi G. *Comput Phys Commun*, 2017, 215: 156–164
41 Rogers D, Hahn M. *J Chem Inf Model*, 2010, 50: 742–754
42 Ye F, Liu Q, Dong H, Guan K, Chen Z, Ju N, Hu L. *Angew Chem Int Ed*, 2022, 61: e202214244
43 Himanen L, Jäger MOJ, Morooka EV, Federici Canova F, Ranawat YS, Gao DZ, Rinke P, Foster AS. *Comput Phys Commun*, 2020, 247: 106949
44 Sendek AD, Yang Q, Cubuk ED, Duerloo KAN, Cui Y, Reed EJ. *Energy Environ Sci*, 2017, 10: 306–320
45 Xu S, Liang J, Yu Y, Liu R, Xu Y, Zhu X, Zhao Y. *J Phys Chem C*, 2021, 125: 21352–21358
46 Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M, Ghiringhelli LM. *Phys Rev Mater*, 2018, 2: 083802