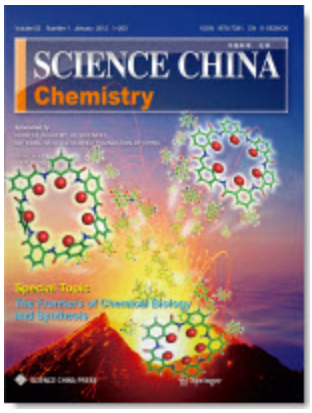


A universal strategy of multi-objective active learning to accelerate the discovery of organic electrode molecules

Journal:	*SCIENCE CHINA Chemistry*
Manuscript ID	SCC-2024-0422.R1
Manuscript Type:	Article
Date Submitted by the Author:	17-Jun-2024
Complete List of Authors:	Zhang, Xinbo; Chang Chun Institute of Applied Chemistry Chinese Academy of Sciences
Keywords:	Organic electrode molecules, Li-ion battery, active learning, multi-objective Bayesian optimization
Speciality:	Physical Chemistry

https://engine.scichina.com/doi/10.1007/s11426-024-2163-1
http://chem.scichina.com/english

SCIENCE CHINA
Chemistry

• Communications •

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 ***; accepted ***; published online ***

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). Remarkably, after only two cycles of active learning, the determination of coefficient can reach 0.962 for theoretical energy density and 0.920 for 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, demonstrating a high capacity of 288 mAh/g and a long cycle life of 1000 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. Our framework provides innovative insights applicable to other domains to expedite the screening process for suitable materials.

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

Citation: Du JY, Guo J, Liu W, Li ZW, Huang G, Zhang XB. A universal strategy of multi-objective active learning to accelerate the discovery of organic electrode molecules. Sci China Chem, ***, doi: ***

With the utilization of fossil fuel, environment 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 devices for clean energy are widely applied into 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].

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

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 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], has played more and more significant roles in chemistry and material fields [17,18]. In particular, lots of researches 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.* constructed an AL framework that was capable of generating high-entropy alloy chemical space [29]. In every cycle of AL, they only synthesized three samples recommended by the AL to obtain those high-entropy alloys with low thermal expansion coefficient. Lu and his co-workers proposed an AL framework with margin sampling to select two-dimensional ferromagnets with high Curie temperature [30]. In addition, Bayesian optimization combined with Gaussian process regression (GPR) is also in a common practice in AL [25,26,31,32]. What's more, the multi-objective active learning (MOAL) has been adopted to screen organic conductor [32], redox active molecules [26], and molecular photoswitches [33]. MOAL has the capability to simultaneously select multi-property molecules, thus reducing screening time compared to the sequential step-by-step screening with multi-property targets. 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 gap is the main task of MOAL. Moreover, the MOAL framework is comprised of machine learning model that combines convolution neural network with GPR, and Pareto front that is achieved by NSGA-II [34]. Furthermore, to gather high-performance potential OEMs, we have carried out 10 cycles of MOAL. 1100 molecules (including initial 100 molecules) have been selected and 4400 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 1100 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.

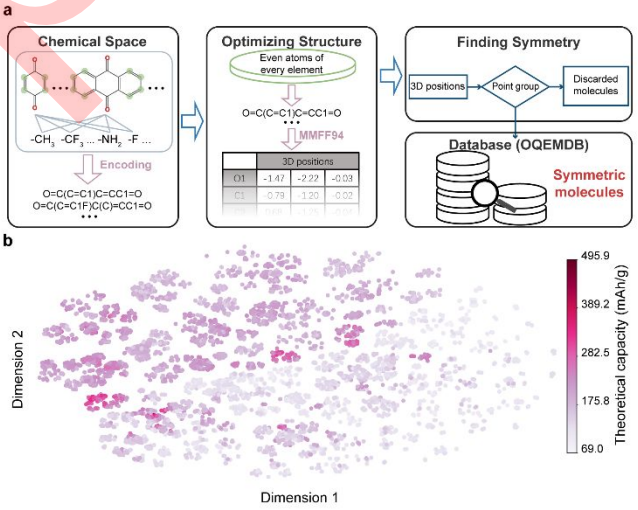


Figure 1 The creation and visualization of OQEMDB. (a) Schematic of the building process of quinone database. (b) Visualization of the quinone database by t-SNE. The data points are colored according to the theoretical capacity.

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 (Figure S1 and S2), 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 Supporting Information (SI) Note 1. 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 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 27463 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 represents these molecules in the course of visualization [41] (SI Note 2). 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 TED as well as low gap from our constructed database for LIB's cathode applications. Notably, the low gap implied fast intramolecular charge transfer [42]. But 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.

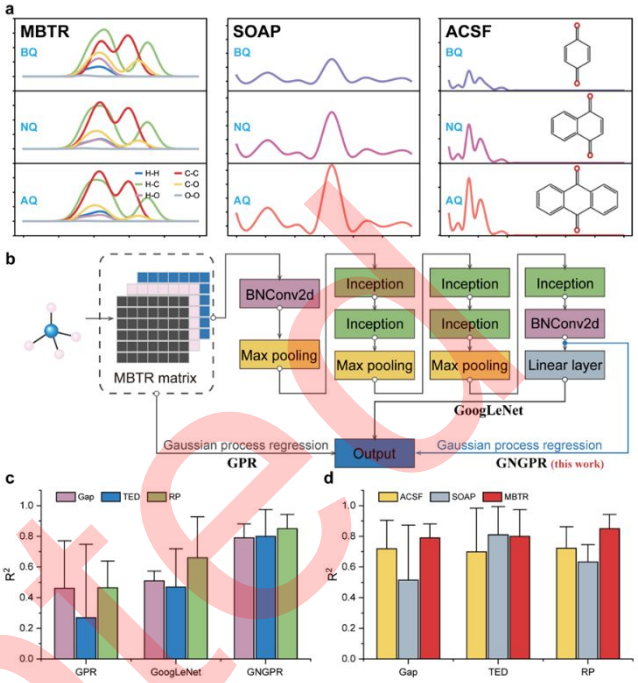


Figure 2 (a). The representations of BQ, NQ, and AQ based on the three descriptors, MBTR, SOAP, and ACSF. (b) The three models of GPR, GoogLeNet, and GNGPR. (c) The evaluations of GPR, GoogLeNet, and GNGPR with MBTR. (d) The evaluations of GNGPR with three descriptors, ACSF, SOAP, and MBTR.

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 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 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 could be found in SI Note 3, Table S1, and the reference [43].

Subsequently, an AL model (GNGPR) using GoogLeNet neural network (SI Note 4, Table S2, and Table S3) was combined with GPR (Figure 2b, SI Note 4, Equation S1) 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. Firstly, 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. Secondly, the output from the penultimate layer of 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 (Equation S5) is equal to the RP (SI Note 5, Equation S2, and Equation S3) multiplied by the TC (Equation S4), and the gap is equal to the LUMO minus HOMO. It is worthy of mentioning that the RP is usually served 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 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 exhibites superior predictive accuracy compared to 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 MBTR descriptor is better competent for MOAL.

Multi-objective Bayesian Optimization framework (MOBO), consisted of Bayesian Optimization (SI Note 6) and NSGA-II, was harnessed to build the MOAL loop (Figure 3a). Expected Improvement (EI, Equation 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 be also updated automatically. Particularly, the training and test sets were always split into an 8:2 ratio during the MOAL loop. 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, and so on. Furthermore, from Figure 3b and 3c, 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

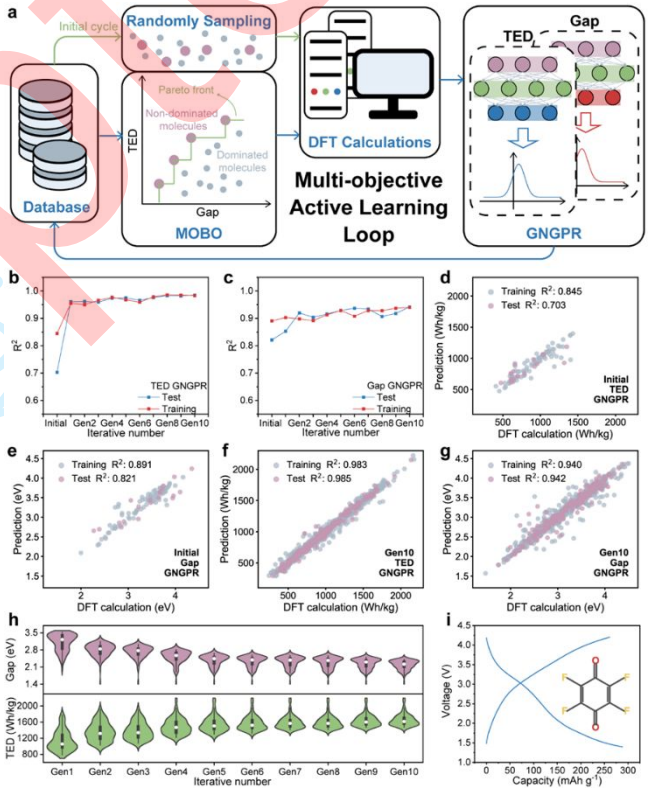


Figure 3 The results of MOAL. (a) Schematic of the MOAL process. The R^2 of each cycle in the MOAL for TED (b) and gap (c). The predicted results of the initial training dataset for TED (d) and gap (e). The predicted results of the training dataset of Gen10 for TED (f) and gap (g). (h) The distribution 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.

in Figure S8-13. The loss values almost remain unchanged after 50 epochs (Figure S8 and S9), indicating that 150 epochs of training for GoogLeNet are adequate. Concurrently,

the performance of GoogLeNet for TED and gap continues to improves during the MOAL (Figure 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 of high-performance molecules have been selected by our MOAL. Consequently, the MOAL could terminate 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 (SI Note 7), it achieves a capacity of 288 mAh/g at 0.1 A/g (Figure 3i) and a cycle lifetime of 1000 cyles at 1 A/g (Figure S14), suggesting the good ability of experimental guidance from the MOAL.

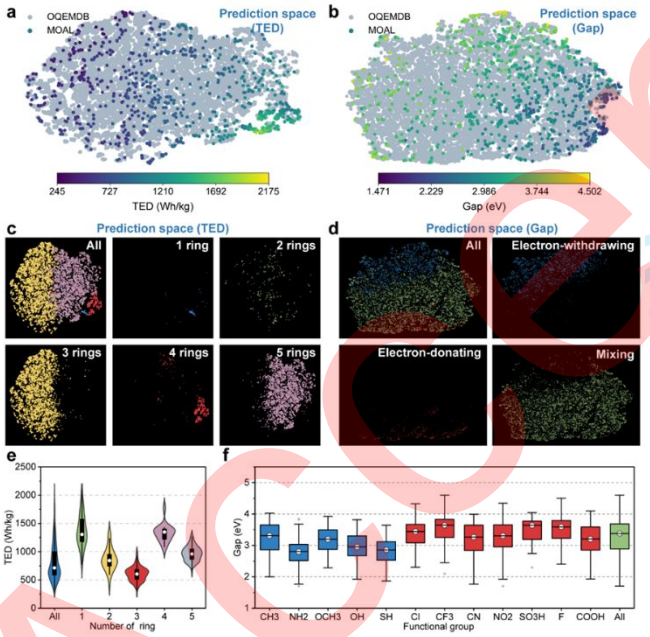


Figure 4 The visualization results of MOAL. The visualization of the prediction space from GoogLeNet for TED (a) and gap (b) by t-SNE. The data points of the MOAL colored by TED and gap, respectively. (c) The data points colored with their number of rings in the prediction space of TED. (d) The data 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.

Although the MOAL model displays excellent predicting ability, a significant limitation of the neural network models is their lack of interpretability, particularly in the applications where the transparency of decision-making is a crucial requirement. To this end, a two-dimensional plane constituted by the output vectors originated 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 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 though in the prediction space of TED and gap (Figure 4a and 4b), the 1100 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 and functional groups. For TED, it is evident that the regions with varying number of rings in the prediction space exhibit distinct characteristics (Figure 4c). This primarily pertains to the TC (Equation 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 the 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 left. 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 are indeed lower than the gap 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 this model also can be competent for other type 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 comprised of more than 26000 various molecules (referred as 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 the 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, 1000, 3000, and 5000) 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 1000 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 1000 to 5000 molecules, while there may be a slight increase in model performance, it is not significant. Specifically, Figure 5c-5e and Figure S18 present the predicted outcomes of QM9 and OMEAD, respectively, with a training size of 4000 and a test size of 1000. 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.

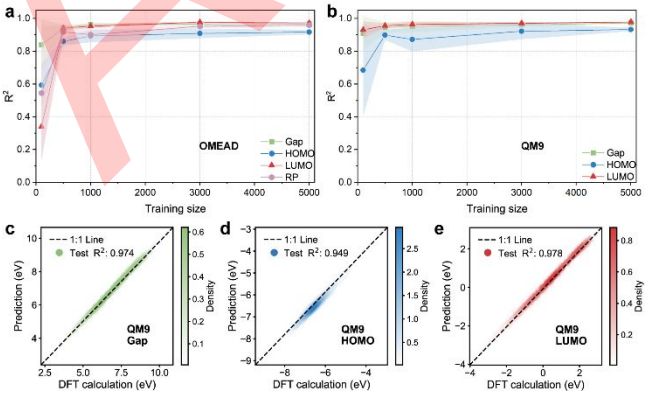


Figure 5 The training results of GNGPR for different targets based on OMEAD (a) and QM9 (b) databases by 5-fold cross-validation under different training size. The predicted results of gap (c), HOMO (d), LUMO (e) for OMEAD using GNGPR. The training set and test set are split into 4:1.

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 formulas to fit the targets. Accordingly, a total of 93 descriptors were meticulously 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 Table S7 and S8.

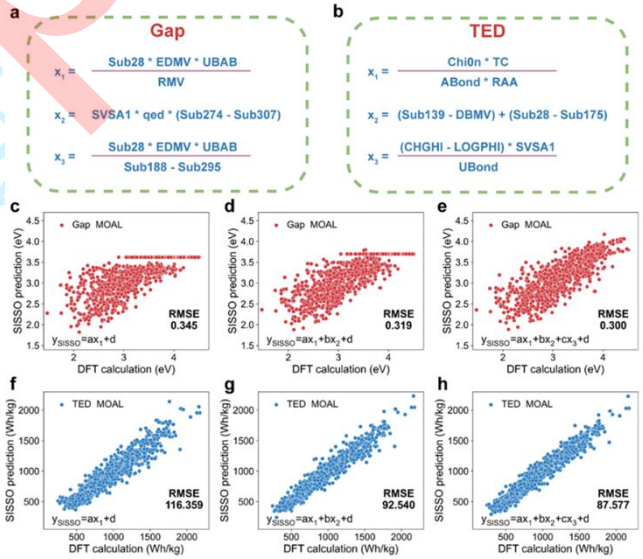


Figure 6 The descriptors and analysis of SISSO. The three descriptors generated by SISSO for gap (a) and TED (b). The predicted results of SISSO formulas with one (c), two (d), and three (e) descriptors for gap of MOAL. The predicted results of SISSO formulas with one (f), two (g), and three (h) descriptors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

As for 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-6e 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 Equation S5, wherein the formula *Chi0n/(ABond*RAA)* acts as the RP. This illustrates that the SISSO algorithm can discover formulas 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) and gap (0.209 eV) emerge, 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 and 3c). In brief, the explicit mathematical formulas have 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 and a long cycle life of 1000 cycles. Consequently, this framework demonstrates the outstanding capability of MOAL to assist 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 database. 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 guidance of designing high-performance OEMs.

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

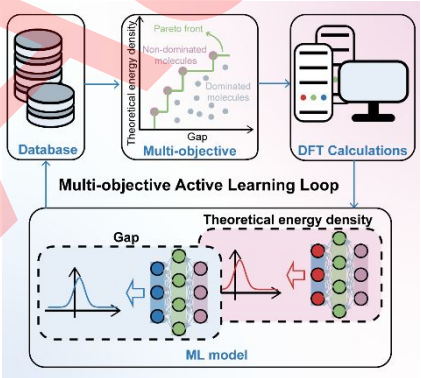
Conflict of interest The authors declare that they have no conflict of interest.

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 Y, Meng Q, Xu P, Zhao C, Chen J, Lu H, Kong X, Wang J, Chen K, Huang G, Zhang X, Su Y, Xiao Y, Chou S, 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 Y, Zhang Q, Chen J, Wan L. *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 Stor 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, Young P, 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 P-Y, Li Z, Gupta BB, Chen X, Wang X. *ACM Comput Surv*, 2021, 54: 1-40
29 Rao Z, Tung P, Xie R, Wei Y, Zhang H, Ferrari A, Klaver TPC, Körmann F, Sukumar PT, Silva AKd, 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 Comput*, 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 C, Filippo, 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

Table of Contents graphic



Supporting Information

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*

SI Note 1. Creating database

The database was generated by RDKit [1] from which the "AllChem" module was used to add 12 functional groups to the 12 quinone molecules (Figure S1). There were four reaction templates written in SMARTS code according to the environment of carbon atom from the 12 quinone molecules (Figure S2).

https://engine.scichina.com/doi/10.1007/s11426-024-2163-1
http://chem.scichina.com/english

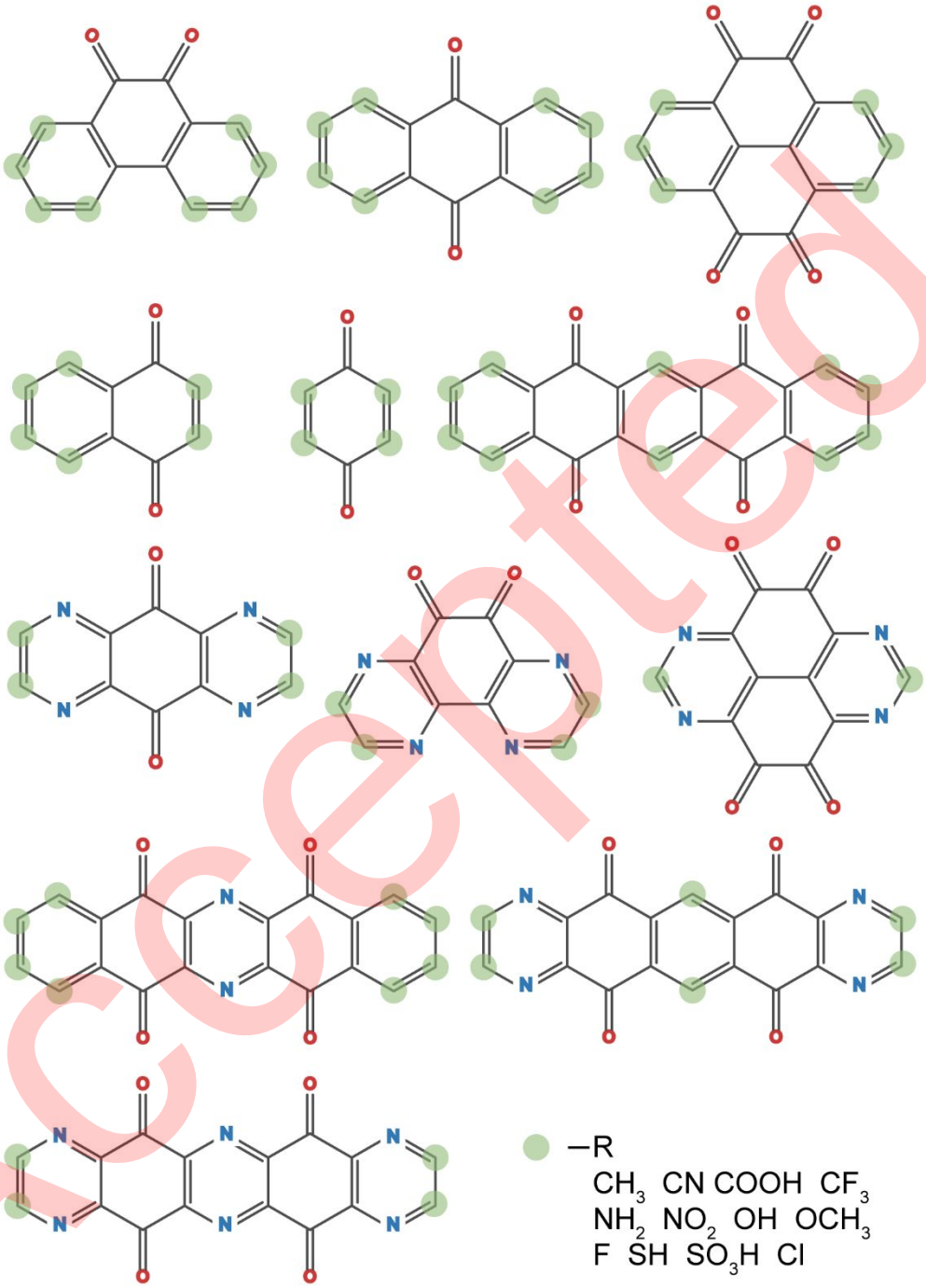


Figure S1. The chemical space composed of 12 quinone molecules and 12 functional groups. The green dots are substitutional positions of functional groups.

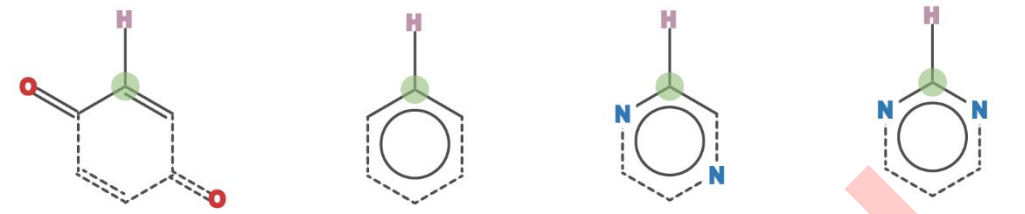


Figure S2. The four reaction templates of carbon atom whose H atom are replaced with 12 functional groups. The four strings are the SMARTS code to describe the four types of carbon atoms. The uppercase letters represent the non-aromatic atoms, and the lowercase letters represent the aromatic atoms. The number after colon is the ordinal number rather than the stoichiometric number. The green carbon atom is the reacted carbon atom whose H atom is substituted.

SI Note 2. Visualization of database

To show the OQEMDB, firstly, Morgan fingerprints [2] were extracted from the SMILES of 27463 quinone molecules by RDKit. And the *radius* and *nBits*, the parameters of Morgan fingerprints, were set to 2 and 2048, respectively. The t-Distribution Stochastic Neighbor Embedding (t-SNE) was implemented by scikit-learn [3].

https://engine.scichina.com/doi/10.1007/s11426-024-2163-1
http://chem.scichina.com/english

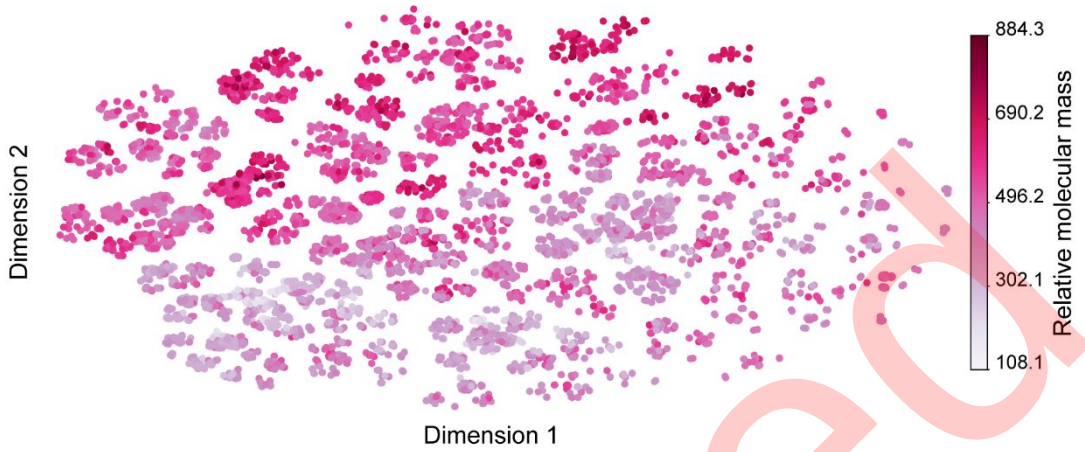


Figure S3. Visualization of the quinone database by t-SNE and the data points colored according to the relative molecular mass.

SI Note 3 Descriptors

In this work, three descriptors that were utilized to represent the organic molecules were many-body tensor representation (MBTR), Smooth Overlap of Atomic Orbitals (SOAP), and Atom-centered Symmetry Functions (ACSF). The three descriptors were generated by a python package, Dscribe [4].

Table S1. The detailed parameters of MBTR, ACSF, and SOAP.

Descriptors	Parameters	Shape of matrix
MBTR	species = ["H","C","N","O","S", "Cl","F"], geometry ={ "function": "inverse_distance"}, grid = {"min":-0.5, "max":1.5, "n":100, "sigma":0.1}, weighting = {"function": "exp", "scale": 0.5, "threshold": 1e-3,}, periodic = False, flatten = False, sparse = False, normalization = "l2"	(7, 7, 100)
SOAP	species = ["H","C","N","O","S","Cl","F"], periodic = False, rcut = 6, nmax = 1, lmax = 1	(60, 56)
ACSF	species = ["H","C","N","O","S","Cl","F"], rcut = 6.0, g2_params = [[1,1],[1,2],[1,3]], g4_params = [[1,1,1],[1,2,1],[1,1,-1],[1,2,-1]]	(60, 140)

SI Note 4 Active learning model

As mentioned in the main body, the active learning model was composed of GoogLeNet neural network and Gaussian Process Regression.

GoogLeNet was a convolutional neural network (CNN) proposed by Szegedy et al. In our GoogLeNet model, it was achieved by Pytorch 1.9.0 package [5] and mainly consisted of BNConv2d block, Inception block (Figure S4), max pooling layer, and a Linear layer. The *ReLU* served as the active function, and we chose *SmoothL1Loss* as the loss function. The architectures of GoogLeNet for MBTR, ACSF, and SOAP were listed in Table S2.

As for BNConv2d and Linear layer, the former number in parentheses was the neurons of input, and the latter was the neurons of output. Generally, after the operation of pooling, the number of neurons maintained unchanged. Particularly, before Linear layer, the multi-dimension matrixes should be converted into 1-D matrix (*nn.Flatten*()). The length of 1-D matrix was the number of neurons of input for Linear layer. And as for Inception block composed of 4 channels (Figure 2b), the first number (i) in parentheses was the neurons of input, the middle number (m) was the neurons of output for first layer, and BNConv2d and the neurons of input for second layer BNConv2d. The last number (o) was the neurons of output for second layer. Different from BNConv2d, the neurons of output of Inception was equal to m+3o rather than o. (Table S2)

Gaussian Process Regression (GPR) was often used for Bayesian optimization surrogate model. The radial basis function (RBF) kernel combined with constant kernel was trained GPR model. The RBF kernel was given by:

$$k_{RBF}(x, x') = exp\left(-\frac{d(x, x')^2}{2l^2}\right) \quad \text{Equation S1}$$

where d was the Euclidean distance and l was the length scale of RBF kernel. The constant kernel was used to scale the magnitude of RBF kernel by multiplication.

https://engine.scichina.com/doi/10.1007/s11426-024-2163-1
http://chem.scichina.com/english

Table S2. The neurons of input and output for every building block of GoogLeNet.

Building block	MBTR	ACSF	SOAP
BNConv2d	(7, 50)	(1, 50)	(1, 50)
Max pooling layer	----		
Inception	(50, 24, 32)	(50, 24, 32)	(50, 24, 32)
Inception	(120, 32, 48)	(120, 32, 48)	(120, 32, 48)
Max pooling layer	----		
Inception	(176, 16, 24)	(176, 16, 24)	(176, 16, 24)
Inception	(88, 10, 16)	(88, 10, 16)	(88, 10, 16)
Max pooling layer	----		
Inception	(58, 8, 10)	(58, 8, 10)	(58, 8, 10)
BNConv2d	(38, 1)	(38, 1)	(38, 1)
nn.Flatten()	----		
Linear layer	(15, 1)	(48, 1)	(36, 1)

Table S3. The distribution of neurons of Inception block.

Channel 1	Channel 2	Channel 3	Channel 4
(i, m)	(i, m)	(i, m)	(i, o)
	(m, o)	(m, o)	Max pooling layer
Neurons of output of Inception		$m+3o$	

https://engine.scichina.com/doi/10.1007/s11426-024-2163-1
http://chem.scichina.com/english

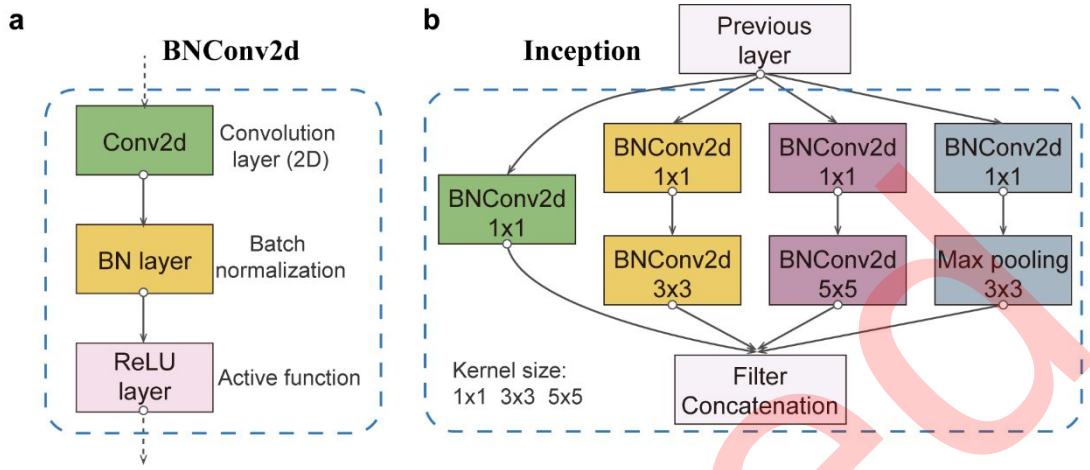


Figure S4. The building blocks of GoogLeNet and GNGPR.

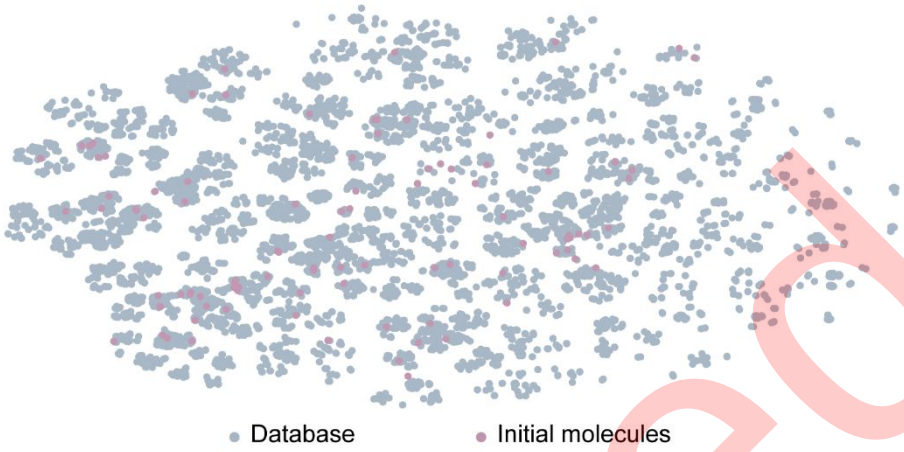


Figure S5. The initial selected 100 molecules in the database.

SI Note 5. The calculation of theoretical capacity, reduction potential, and theoretical energy density

Reduction potential: By a thermodynamic cycle (Figure S6), the four states of one organic molecule (R) were calculated by density of functional Theory (DFT). They were R in gas phase (R(gas)), R in solvation phase (R(sol)), R with one negative charge in gas phase (R^-(gas)), and R with one negative charge in solvation phase (R^-(sol)). Then, the reduction potential E^{red} was obtained through Equation S2 and Equation S3[6, 7].

All DFT calculations were performed by Gaussian 16 software[8], and the organic molecules were optimized at the level of b3lyp/6-31+(d, p). SMD solvation model were also considered during calculations. A constant 1.4 V is derived from the absolute potential of hydrogen electrode (-4.44 V) and the potential of Li/Li+, -3.04 V vs. SHE.

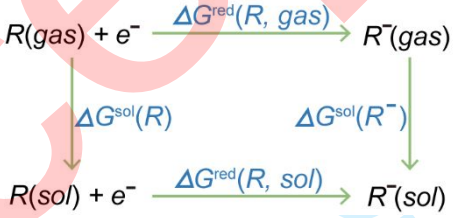


Figure S6. The thermodynamic cycle of calculating the reduction potential.

$$\Delta G^{red}(R, sol) = \Delta G^{red}(R, gas) + \Delta G^{sol}(R^-) - \Delta G^{sol}(R) \quad \text{Equation S2}$$

$$E^{red} = -\frac{\Delta G^{red}(R,sol)}{nF} - 1.4\ \text{V} \quad \text{Equation S3}$$

Theoretical capacity (TC): Here, we only count the carbonyl as the active site in the skeleton of aromatic rings rather than function groups such as -COOH, -NO_2, and SO_3H. Therefore, the theoretical capacity (TC, mAh g^{-1}) can be obtained by Equation S4[9].

$$\text{TC} = \frac{nF}{3.6M_w} \quad \text{Equation S4}$$

Where n is the number of active sites of the investigated organic molecule, F is the Faraday constant (C mol^{-1}), and M_w is the molar mass (g mol^{-1}) of the investigated organic molecule.

Theoretical energy density (TED): The energy density is an important metrics for batteries. Here we have adopted TED (Wh kg^{-1}) as one of the metrics of OEMs, which can be calculated by:

$$\text{TED} = \text{TC} \times \text{RP} = \frac{nF}{3.6M_w} \times \text{RP} \qquad \text{Equation S5}$$

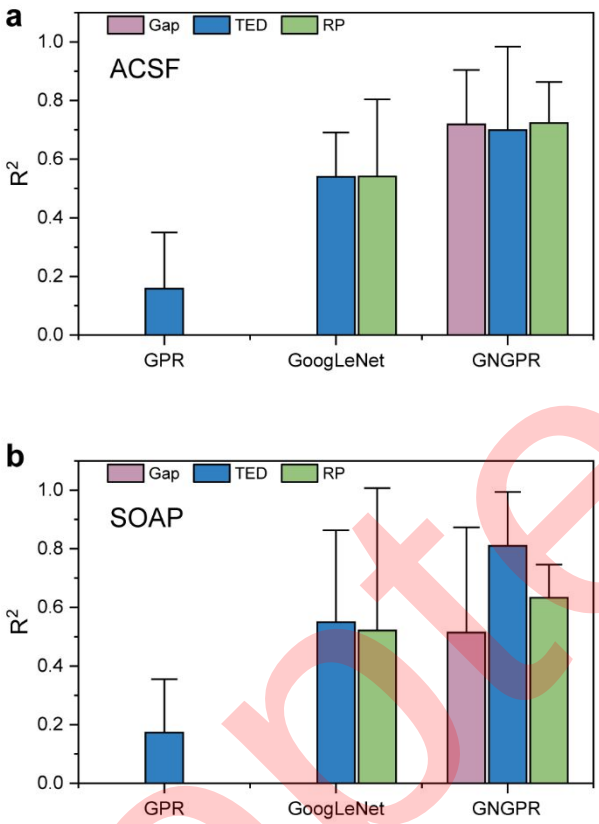


Figure S7. (a) The evaluations of GPR, GoogLeNet, and GNGPR with ACSF. (b) The evaluations of GPR, GoogLeNet, and GNGPR with SOAP.

SI Note 6. Bayesian optimization (BO)

In this work, BO was employed to pick out the potential good candidates from the predicting results of GNGPR (especially GPR part). The expected improvement (EI) as the acquisition function of BO was independently calculated for TED and Gap. The EI was given by:

$$\mathrm{EI} = \begin{cases} \mu(x) - f(x^+)\Phi(Z) + \sigma(x)\phi(Z) & \sigma(x) > 0 \\ 0 & \sigma(x) > 0 \end{cases} \quad \text{Equation S6}$$

$$Z = \frac{\mu(x) - f(x^+) - \xi}{\sigma(x)} \quad \text{Equation S7}$$

where $\mu(x)$ and $\sigma(x)$ were the mean and standard deviation values of GPR-predicted results. $f(x^+)$ was the best value of candidate, which was the max value for TED or minimum value for Gap in current training set. $\Phi(Z)$ and $\phi(Z)$ represented the cumulative distribution function and the probability density function, respectively. ξ was a constant value and here it was set to 0.01.

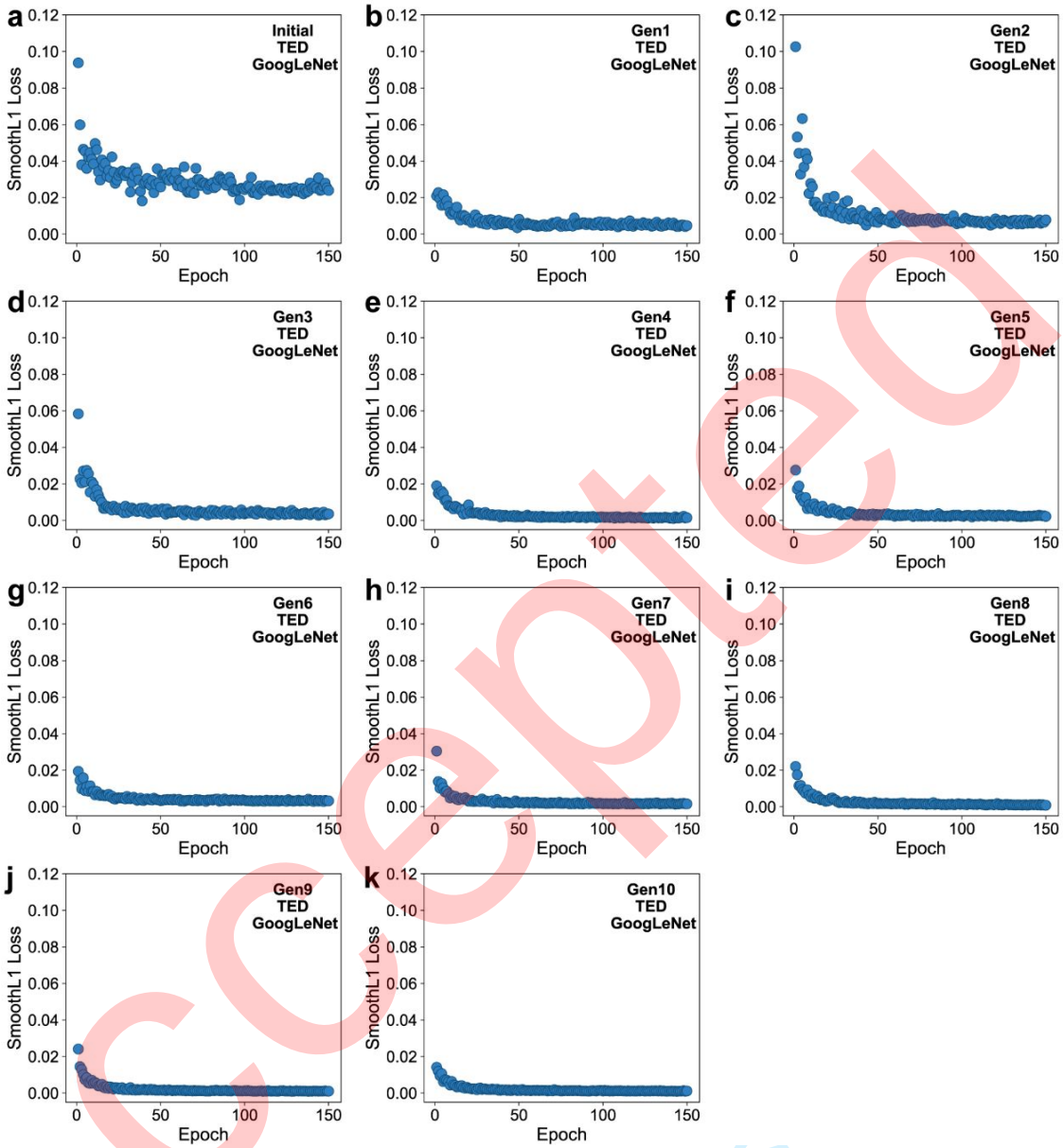


Figure S8. The SmoothL1 Loss as the loss function of the number of epochs for TED during MOAL.

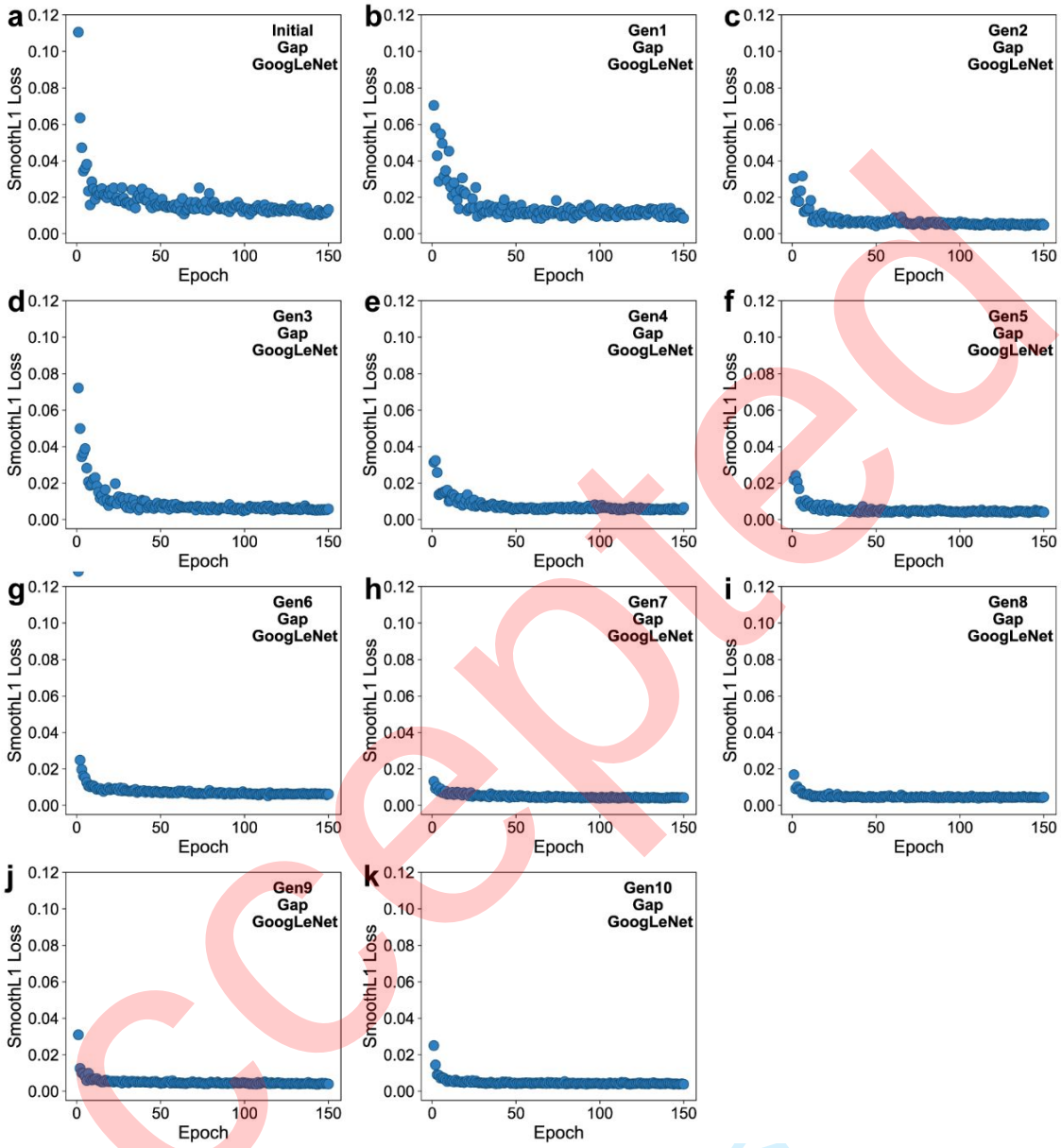


Figure S9. The SmoothL1 Loss as the loss function of the number of epochs for Gap during MOAL.

After 50 epochs of training GoogLeNet models (Figure S8, S9), the values of loss function experience barely decrease. Therefore, it is appropriate that the parameters of GoogLeNet are saved after 100 epochs.

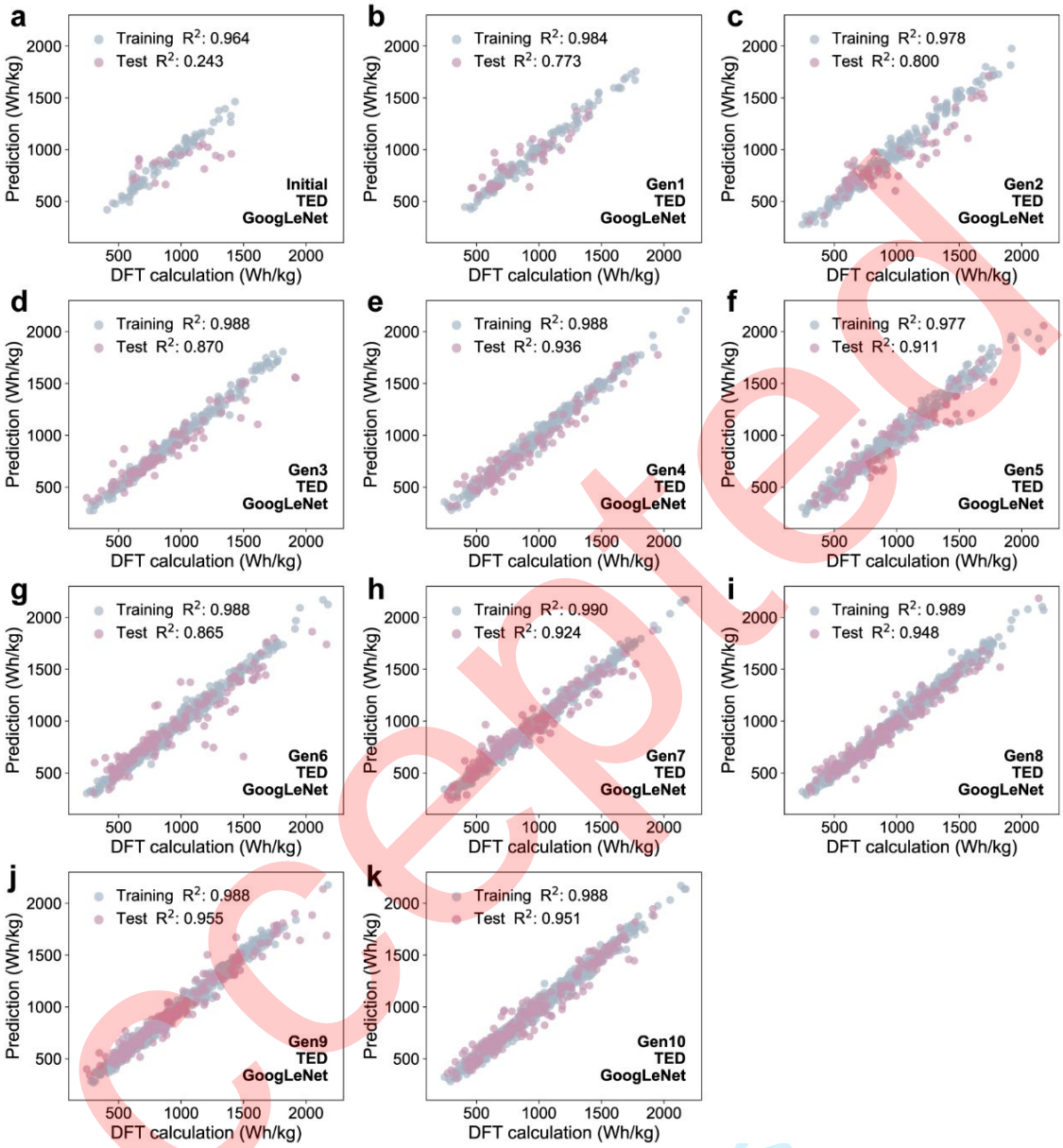


Figure S10. The performance of GoogLeNet for TED during MOAL.

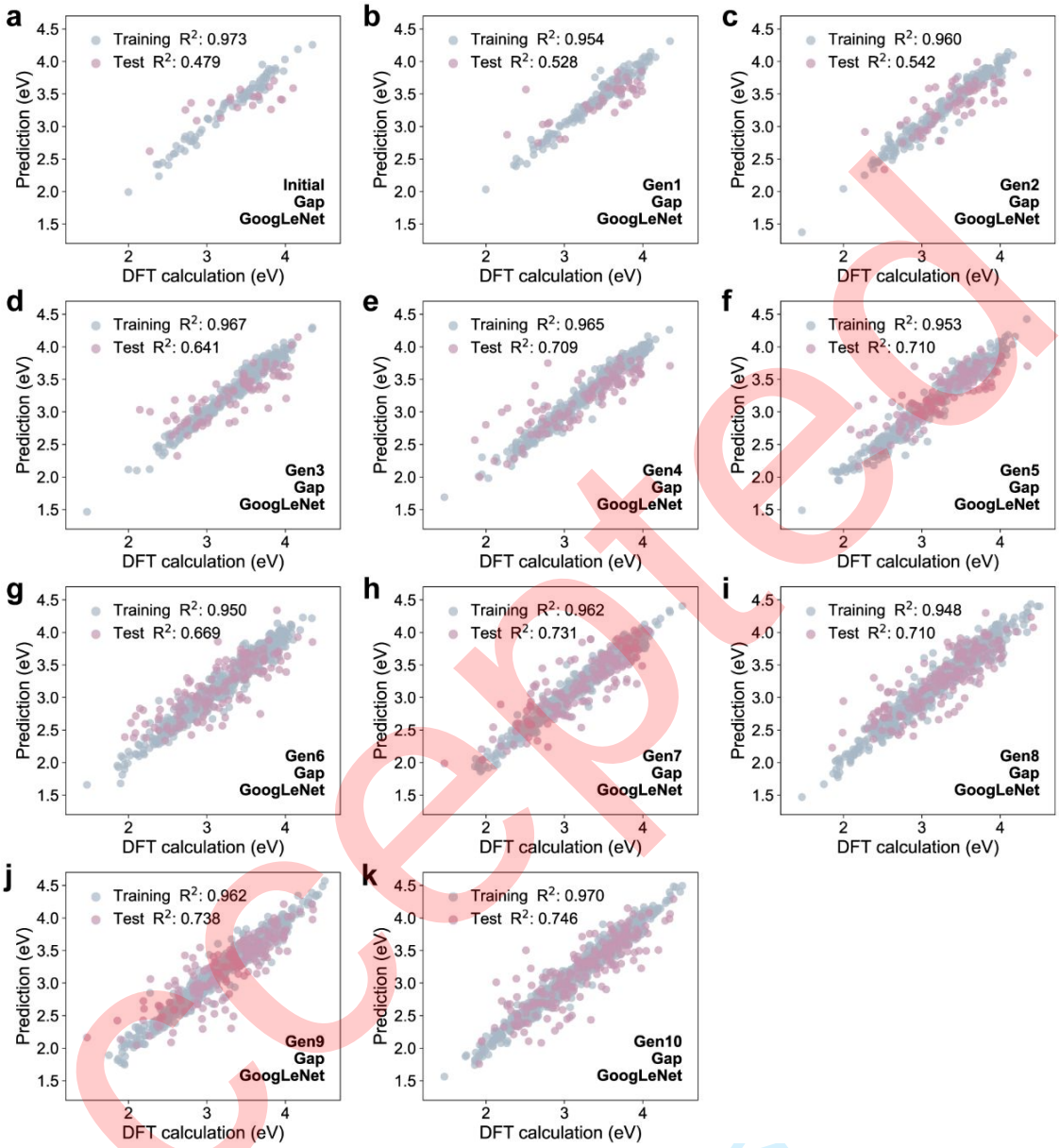


Figure S11. The performance of GoogLeNet for gap during MOAL.

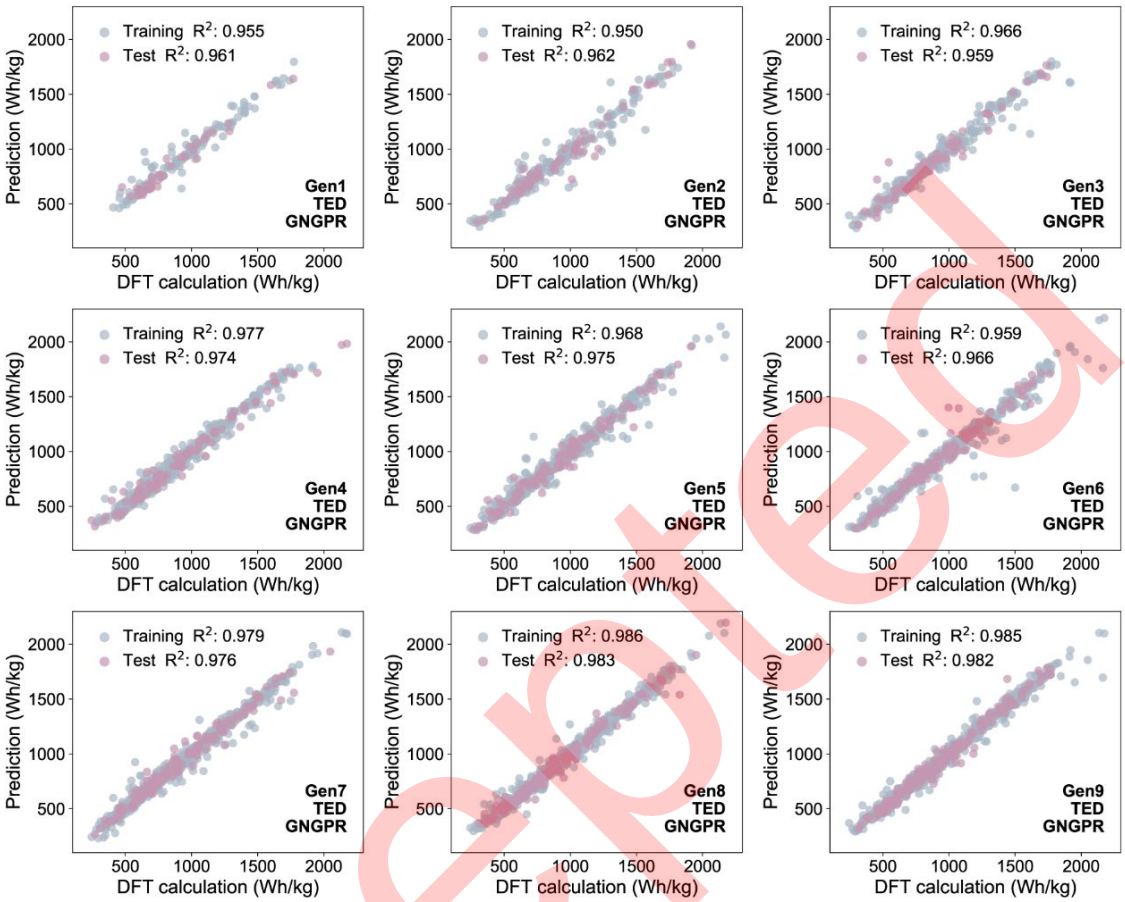


Figure S12. The performance of GNGPR for predicting the TED during MOAL except the Initial and Gen10.

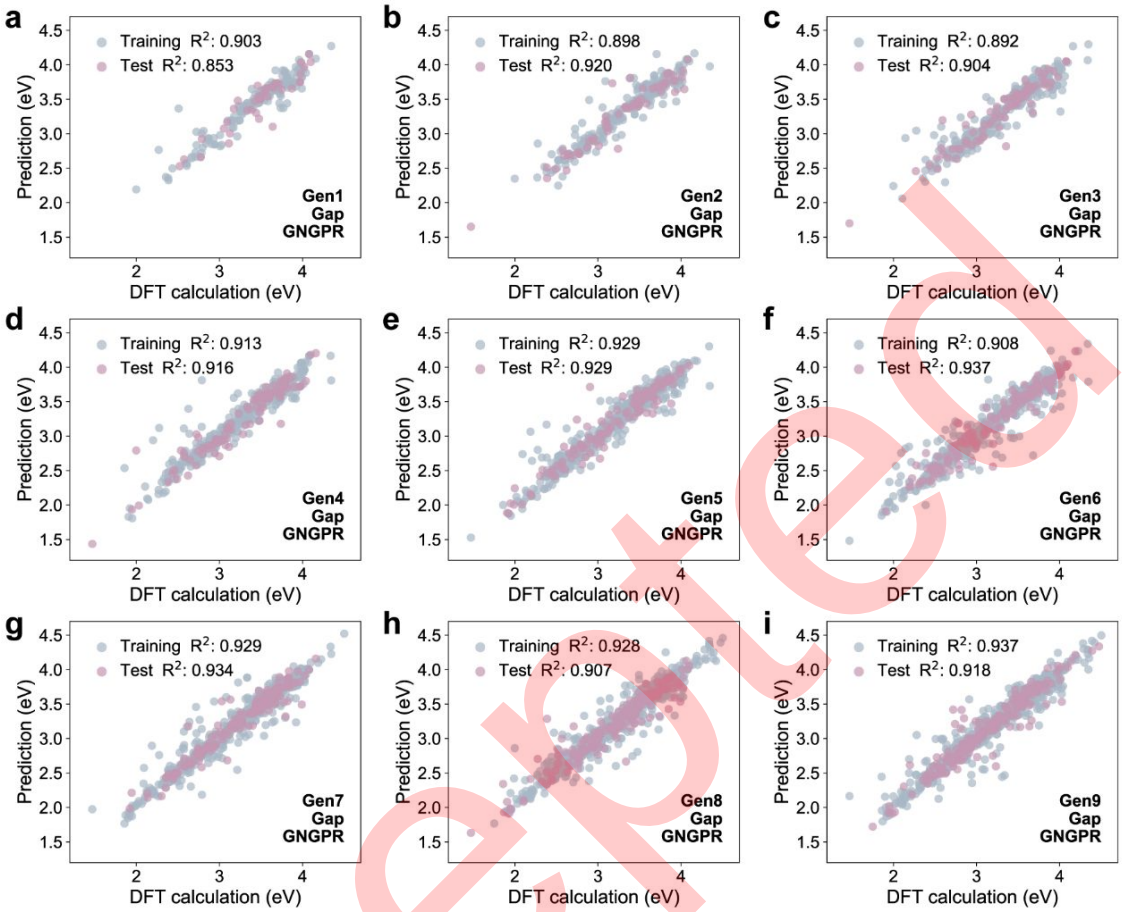


Figure S13. The performance of GNGPR for predicting the gap during MOAL except the Initial and Gen10.

Table S4. The top 100 molecules chosen according to the Pareto Front (NSGA-II). The ID number 2495 is TFDD.

ID	SMILES
1	O=c1c(=O)c2cc(F)cc3c2-c2c1cc(F)cc2c(=O)c3=O
59	N#Cc1cc(S)c2c3c1c(=O)c(=O)c1c(S)cc(C#N)c(c1-3)c(=O)c2=O
65	N#Cc1c(N)cc2c3c1c(=O)c(=O)c1cc(N)c(C#N)c(c1-3)c(=O)c2=O
91	N#Cc1cc(S)c2c3c1c(=O)c(=O)c1c(C#N)cc(S)c(c1-3)c(=O)c2=O
152	N#Cc1cc(C#N)c2c3c1c(=O)c(=O)c1c(C#N)cc(C#N)c(c1-3)c(=O)c2=O
158	N#Cc1cc2c3c(c1N)c(=O)c(=O)c1cc(C#N)c(N)c(c1-3)c(=O)c2=O
165	N#Cc1c(S)cc2c3c1c(=O)c(=O)c1cc(S)c(C#N)c(c1-3)c(=O)c2=O
191	N#Cc1cc2c3c(c1C#N)c(=O)c(=O)c1c(C#N)c(C#N)cc(c1-3)c(=O)c2=O
253	O=c1c(=O)c2c(Cl)cc(Cl)c3c2-c2c1c(S)cc(S)c2c(=O)c3=O
308	N#Cc1cc(C#N)c2c3c1c(=O)c(=O)c1c(S)cc(S)c(c1-3)c(=O)c2=O
335	O=c1c(=O)c2c(S)cc(Cl)c3c2-c2c1c(S)cc(Cl)c2c(=O)c3=O
342	O=c1c(=O)c2c(O)c(S)cc3c2-c2c1cc(S)c(O)c2c(=O)c3=O
353	Nc1nc2c3c(n1)c(=O)c(=O)c1nc(N)nc(c1-3)c(=O)c2=O
630	Cc1c(C)c(N)c2c(=O)c3cc4c(=O)c5ccccc5c(=O)c4cc3c(=O)c2c1N
1135	Nc1cc2c(=O)c3nc4c(=O)c5c(F)ccc(F)c5c(=O)c4nc3c(=O)c2cc1N
2445	O=C(C=C1)C=CC1=O
2449	O=C1C=C([N+](=O)[O-])C(=O)C=C1N(=O)=O
2453	N#CC1=C(C#N)C(=O)C=CC1=O
2454	O=C1C=CC(=O)C([N+](=O)[O-])=C1N(=O)=O
2455	N#CC1=CC(=O)C=C(C#N)C1=O
2456	O=C1C=CC(=O)C(F)=C1F
2458	O=C1C=C(O)C(=O)C(O)=C1
2461	O=C1C=C(F)C(=O)C(F)=C1
2470	O=C1C(O)=C(O)C(=O)C(O)=C1O
2471	O=C1C(S)=C(S)C(=O)C([N+](=O)[O-])=C1[N+](=O)[O-]
2480	N#CC1=C(C#N)C(=O)C(O)=C(O)C1=O
2495	**O=C1C(F)=C(F)C(=O)C(F)=C1F**
2523	N#CC1=C(C#N)C(=O)C([N+](=O)[O-])=C(N(=O)=O)C1=O
2524	N#CC1=C(F)C(=O)C(F)=C(C#N)C1=O
2536	N#CC1=C(C#N)C(=O)C(C#N)=C(C#N)C1=O
2539	N#CC1=C([N+](=O)[O-])C(=O)C(C#N)=C([N+](=O)[O-])C1=O
2554	N#CC1=C(C#N)C(=O)C([N+](=O)[O-])=C([N+](=O)[O-])C1=O
2563	O=C1C(O)=C(O)C(=O)C(F)=C1F
2566	COC1=C(OC)C(=O)C([N+](=O)[O-])=C([N+](=O)[O-])C1=O
2568	N#CC1=C(C#N)C(=O)C(S)=C(S)C1=O
2659	Nc1c2c(c(N)c([N+](=O)[O-])c1[N+](=O)[O-])C(=O)C=CC2=O
2719	NC1=C(N)C(=O)c2cc([N+](=O)[O-])c(N(=O)=O)cc2C1=O

2731	Nc1cc2c(cc1N)C(=O)C([N+](=O)[O-])=C([N+](=O)[O-])C2=O
2764	N#Cc1c(F)c(S)c2c3c1c(=O)c(=O)c1c(S)c(F)c(C#N)c(c1-3)c(=O)c2=O
2781	N#Cc1c(N)c2c3c(c1C#N)c(=O)c(=O)c1c(C#N)c(C#N)c(N)c(c1-3)c(=O)c2=O
2811	Nc1c(S)c2c3c(c1S)c(=O)c(=O)c1c(S)c(N)c(S)c(c1-3)c(=O)c2=O
2858	N#Cc1c(F)c(C#N)c2c3c1c(=O)c(=O)c1c(C#N)c(F)c(C#N)c(c1-3)c(=O)c2=O
2888	Nc1c(F)c(S)c2c3c1c(=O)c(=O)c1c(N)c(F)c(S)c(c1-3)c(=O)c2=O
2901	O=c1c(=O)c2c(S)c([N+](=O)[O-])c(S)c3c2-c2c1c(O)c(F)c(O)c2c(=O)c3=O
2934	N#Cc1c(O)c(C(=O)O)c2c3c1c(=O)c(=O)c1c(C#N)c(O)c(C(=O)O)c(c1-3)c(=O)c2=O
2937	N#Cc1c(S)c2c3c(c1S)c(=O)c(=O)c1c(C#N)c(C#N)c(C#N)c(c1-3)c(=O)c2=O
2940	Nc1c([N+](=O)[O-])c2c3c(c1[N+](=O)[O-])c(=O)c(=O)c1c([N+](=O)[O-])c(N)c([N+](=O)[O-])c(c1-3)c(=O)c2=O
2946	Nc1c(F)c(S)c2c3c1c(=O)c(=O)c1c(S)c(F)c(N)c(c1-3)c(=O)c2=O
2981	O=c1c(=O)c2c(S)c(S)c(Cl)c3c2-c2c1c(S)c(S)c(Cl)c2c(=O)c3=O
3018	Cc1c(N)c(N)c2c3c1c(=O)c(=O)c1c(C)c(N)c(N)c(c1-3)c(=O)c2=O
3023	N#Cc1c(N)c([N+](=O)[O-])c2c3c1c(=O)c(=O)c1c([N+](=O)[O-])c(N)c(C#N)c(c1-3)c(=O)c2=O
3050	N#Cc1c(S)c2c3c(c1Cl)c(=O)c(=O)c1c(Cl)c(C#N)c(S)c(c1-3)c(=O)c2=O
3055	Nc1c(Cl)c(S)c2c3c1c(=O)c(=O)c1c(S)c(Cl)c(N)c(c1-3)c(=O)c2=O
3059	O=c1c(=O)c2c(Cl)c(S)c(S)c3c2-c2c1c(S)c(S)c(Cl)c2c(=O)c3=O
3061	Nc1c(C(=O)O)c2c3c(c1[N+](=O)[O-])c(=O)c(=O)c1c(C(=O)O)c(N)c([N+](=O)[O-])c(c1-3)c(=O)c2=O
3068	Nc1c(O)c(O)c2c3c1c(=O)c(=O)c1c(O)c(O)c(N)c(c1-3)c(=O)c2=O
3106	O=c1c(=O)c2c(S)c(O)c(O)c3c2-c2c1c(O)c(O)c(S)c2c(=O)c3=O
3129	Nc1c(S)c2c3c(c1[N+](=O)[O-])c(=O)c(=O)c1c(S)c(N)c([N+](=O)[O-])c(c1-3)c(=O)c2=O
3200	N#Cc1c(F)c2c3c(c1F)c(=O)c(=O)c1c(S)c(C#N)c(S)c(c1-3)c(=O)c2=O
3230	N#Cc1c(O)c(C#N)c2c3c1c(=O)c(=O)c1c(C#N)c(O)c(C#N)c(c1-3)c(=O)c2=O
3233	O=c1c(=O)c2c(F)c(F)c(S)c3c2-c2c1c(F)c(F)c(S)c2c(=O)c3=O
3240	Cc1c(C#N)c(S)c2c3c1c(=O)c(=O)c1c(C)c(C#N)c(S)c(c1-3)c(=O)c2=O
3245	N#Cc1c(N)c(F)c2c3c1c(=O)c(=O)c1c(C#N)c(N)c(F)c(c1-3)c(=O)c2=O
3273	Nc1c(S)c2c3c(c1Cl)c(=O)c(=O)c1c(Cl)c(N)c(S)c(c1-3)c(=O)c2=O
3298	Nc1c(F)c2c3c(c1S)c(=O)c(=O)c1c(F)c(N)c(S)c(c1-3)c(=O)c2=O
3367	Nc1c(N(=O)=O)c2c3c(c1[N+](=O)[O-])c(=O)c(=O)c1c([N+](=O)[O-])c(N)c([N+](=O)[O-])c(c1-3)c(=O)c2=O
3369	COc1c(C#N)c(S)c2c3c1c(=O)c(=O)c1c(OC)c(C#N)c(S)c(c1-3)c(=O)c2=O
3372	O=c1c(=O)c2c(S)c(F)c(S)c3c2-c2c1c(S)c(F)c(S)c2c(=O)c3=O
3394	N#Cc1c(F)c2c3c(c1S)c(=O)c(=O)c1c(F)c(C#N)c(S)c(c1-3)c(=O)c2=O
3404	N#Cc1c(S)c([N+](=O)[O-])c2c3c1c(=O)c(=O)c1c(C#N)c(S)c([N+](=O)[O-])c(c1-3)c(=O)c2=O
3409	N#Cc1c(N)c2c3c(c1C#N)c(=O)c(=O)c1c(N)c(C#N)c(C#N)c(c1-3)c(=O)c2=O

3500	Cc1c(N)c(N)c2c3c1c(=O)c(=O)c1c(N)c(N)c(C)c(c1-3)c(=O)c2=O
3550	N#Cc1c(S)c(S)c2c3c1c(=O)c(=O)c1c(S)c(S)c(C#N)c(c1-3)c(=O)c2=O
3565	N#Cc1c(F)c(C#N)c2c3c1c(=O)c(=O)c1c(C#N)c(Cl)c(C#N)c(c1-3)c(=O)c2=O
3572	N#Cc1c(C#N)c2c3c(c1C#N)c(=O)c(=O)c1c(C#N)c(C#N)c(C#N)c(c1-3)c(=O)c2=O
3601	N#Cc1c(N)c([N+](=O)[O-])c2c3c1c(=O)c(=O)c1c(C#N)c(N)c([N+](=O)[O-])c(c1-3)c(=O)c2=O
3628	N#Cc1c(N)c(C#N)c2c3c1c(=O)c(=O)c1c(C#N)c(N)c(C#N)c(c1-3)c(=O)c2=O
3656	N#Cc1c(S)c2c3c(c1C#N)c(=O)c(=O)c1c(S)c(C#N)c(C#N)c(c1-3)c(=O)c2=O
3671	O=c1c(=O)c2c(S)c(Cl)c(S)c3c2-c2c1c(S)c(F)c(S)c2c(=O)c3=O
3673	N#Cc1c(N)c2c3c(c1N)c(=O)c(=O)c1c(F)c(C#N)c(F)c(c1-3)c(=O)c2=O
3676	Nc1c(O)c2c3c(c1[N+](=O)[O-])c(=O)c(=O)c1c(O)c(N)c([N+](=O)[O-])c(c1-3)c(=O)c2=O
5845	N#Cc1c(C#N)c(F)c2c(=O)c3cc4c(=O)c5c(N)c(C(=O)O)c(C(=O)O)c(N)c5c(=O)c4cc3c(=O)c2c1F
6009	N#Cc1c(F)c(F)c(C#N)c2c(=O)c3cc4c(=O)c5c(N)c(C(=O)O)c(C(=O)O)c(N)c5c(=O)c4cc3c(=O)c12
10072	Nc1c(O)c(O)c(N)c2c(=O)c3nc4c(=O)c5cc(F)c(F)cc5c(=O)c4nc3c(=O)c12
10343	Nc1c(N)c(O)c2c(=O)c3nc4c(=O)c5c(N)ccc(N)c5c(=O)c4nc3c(=O)c2c1O
10743	N#Cc1c(C#N)c(N)c2c(=O)c3nc4c(=O)c5cc(C(=O)O)c(C(=O)O)cc5c(=O)c4nc3c(=O)c2c1N
10934	Nc1c(N)c(N)c2c(=O)c3nc4c(=O)c5cc(S)c(S)cc5c(=O)c4nc3c(=O)c2c1N
11412	Nc1c(O)c(O)c(N)c2c(=O)c3nc4c(=O)c5c(F)c(F)c(F)c(F)c5c(=O)c4nc3c(=O)c12
11719	Nc1c(N)c(N)c2c(=O)c3nc4c(=O)c5c(O)c(C(F)(F)F)c(C(F)(F)F)c(O)c5c(=O)c4nc3c(=O)c2c1N
12268	Nc1c([N+](=O)[O-])c(O)c(O)c2c(=O)c3nc4c(=O)c5c(O)c(O)c([N+](=O)[O-])c(N)c5c(=O)c4nc3c(=O)c12
12859	N#Cc1c([N+](=O)[O-])c(N)c2c(=O)c3nc4c(=O)c5c(O)c(C#N)c([N+](=O)[O-])c(N)c5c(=O)c4nc3c(=O)c2c1O
12910	Nc1c(N(=O)=O)c([N+](=O)[O-])c(N)c2c(=O)c3nc4c(=O)c5c(O)c(F)c(F)c(O)c5c(=O)c4nc3c(=O)c12
13814	Nc1c([N+](=O)[O-])c(S)c2c(=O)c3nc4c(=O)c5c(S)c(N)c([N+](=O)[O-])c(S)c5c(=O)c4nc3c(=O)c2c1S
13834	N#Cc1c(N(=O)=O)c(N)c2c(=O)c3nc4c(=O)c5c(N)c(C#N)c([N+](=O)[O-])c(N)c5c(=O)c4nc3c(=O)c2c1N
17413	Nc1c2c(c(N)c([N+](=O)[O-])c1[N+](=O)[O-])C(=O)c1c(C(=O)O)c(C(=O)O)c(C(=O)O)c(C(=O)O)c1C2=O
18577	Nc1c(O)c(O)c(N)c2c1C(=O)c1c(c([N+](=O)[O-])c([N+](=O)[O-])c([N+](=O)[O-])c1[N+](=O)[O-])C2=O
19463	Nc1c2c(c(N)c([N+](=O)[O-])c1[N+](=O)[O-])C(=O)c1c(O)c(O)c(O)c(O)c1C2=O
27353	Nc1c2c(c(N)c([N+](=O)[O-])c1[N+](=O)[O-])C(=O)C(O)=C(O)C2=O
27397	Nc1c2c(c(N)c([N+](=O)[O-])c1[N+](=O)[O-])C(=O)C(Cl)=C(Cl)C2=O
27406	Nc1c2c(c(N)c([N+](=O)[O-])c1[N+](=O)[O-])C(=O)C(F)=C(F)C2=O

SI Note 7. Experiment

Synthesis of 2,3,5,6-tetrafluorocyclohexa-2,5-diene-1,4-dione (TFDD)

2,3,5,6-tetrafluorohydroquinone (10 mmol) was added to an aqueous solution of ceric ammonium nitrate (21.1 mmol) in water and stirred for 4 h at room temperature. Then, the mixture was extracted with diethyl ether. After purification with a silica gel column, the solution was removed to obtain the yellow solid TFDD.

Electrochemical measurement

The cathodes were prepared with TFDD, KB conductive additive, and polytetrafluoroethylene (PTFE) with a weight ratio of 3:6:1 by water. The mixed slurry was fully ground by ball milling for 120 min, and then rolled into a film and pressed on a aluminum mesh current collector (10 mm) to obtain TFDD -based cathodes. The mass loading of TFDD was controlled at around 1.5 mg cm^{-2}. The electrochemical performance of TFDD //Li metal batteries were tested in 2025-type coin cells and the electrolyte was 1 M $LiPF_6$ in EC:DMC=1:1 Wt%. Galvanostatic charge–discharge cycling test was performed with a Land CT2001A battery testing system.

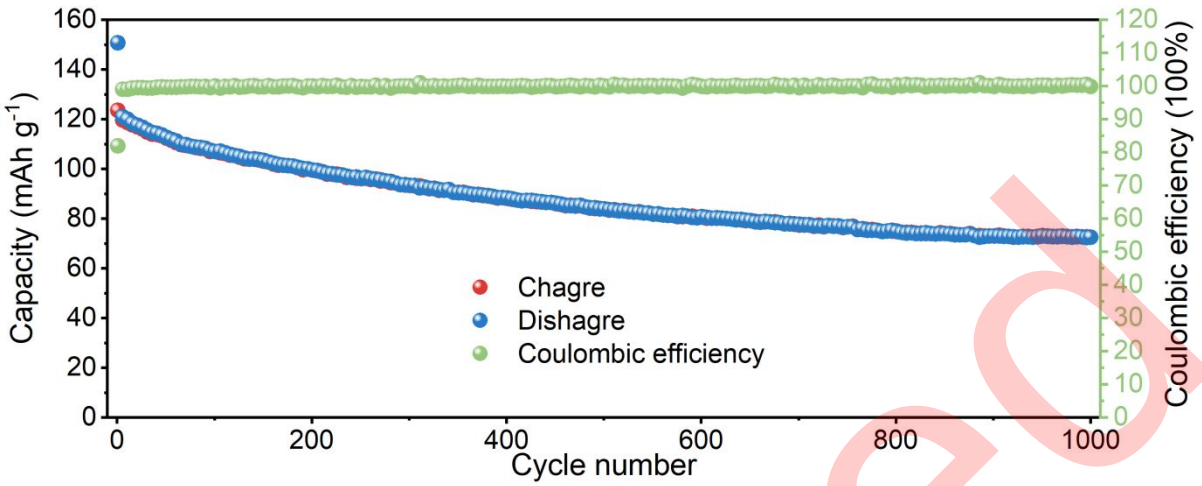


Figure S14. The long-term cycling performance of TFDD at 1 A/g.

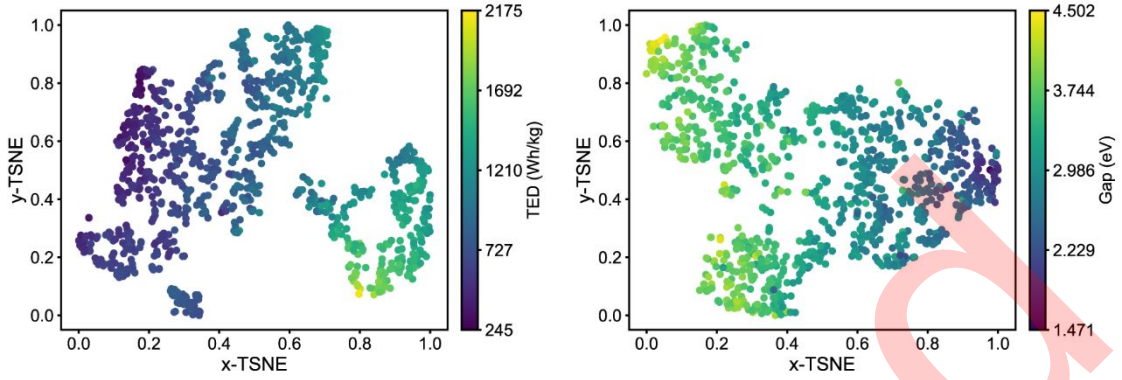
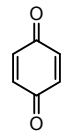
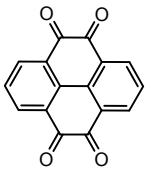
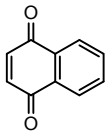
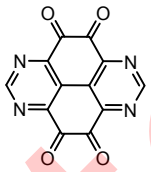
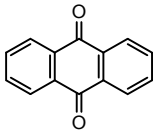
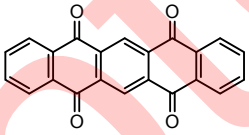
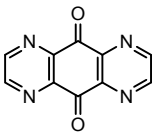
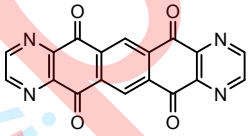
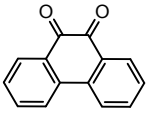
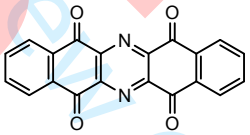
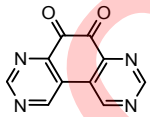
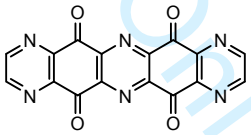


Figure S15. The visualization of training space from GoogLeNet for TED (a) and Gap (b) by t-SNE. The data points of MOAL are colored by TED and Gap, respectively.

Table S5. The theoretical capacities of 12 quinone molecules.

	Rings	RMW	TC		Rings	RMW	TC
	1	108.10	495.86		4	262.22	408.50
	2	158.16	338.91		4	266.17	402.77
	3	208.22	257.43		5	338.32	316.88
	3	212.77	251.93		5	342.27	313.22
	3	208.22	257.43		5	340.29	315.04
	3	212.77	251.93		5	344.25	311.42

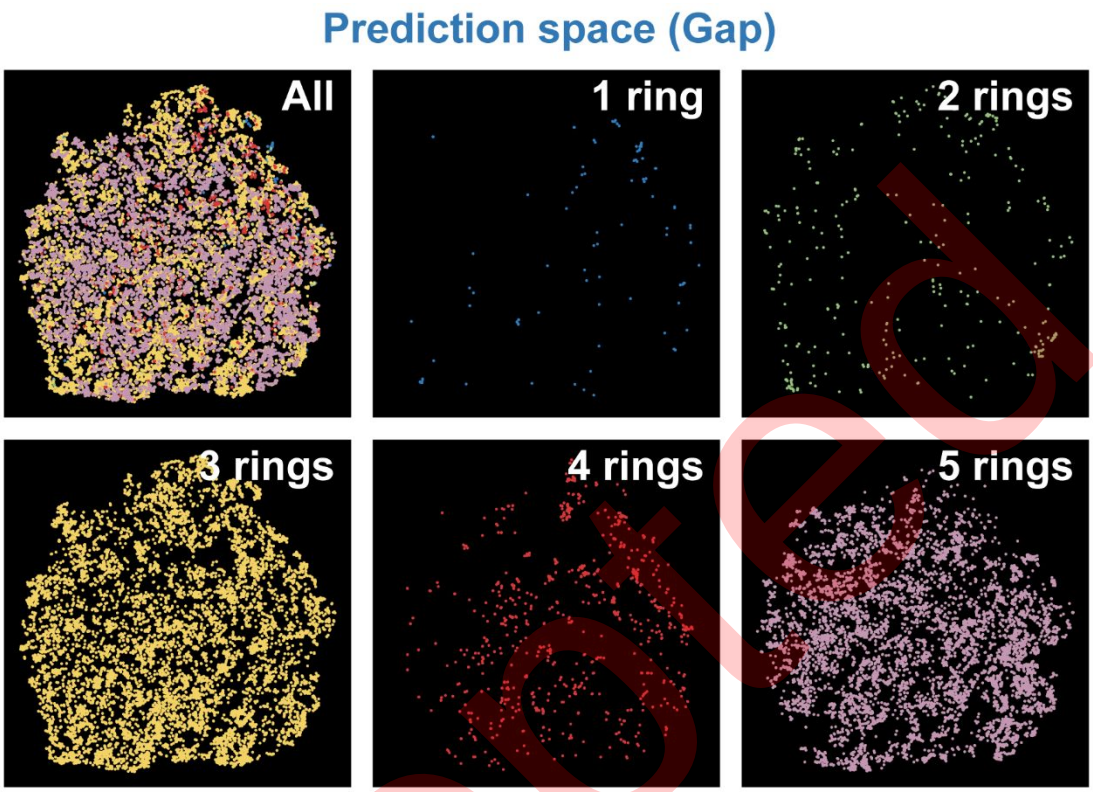


Figure S16. The data points are colored with their number of rings in the prediction space of Gap.

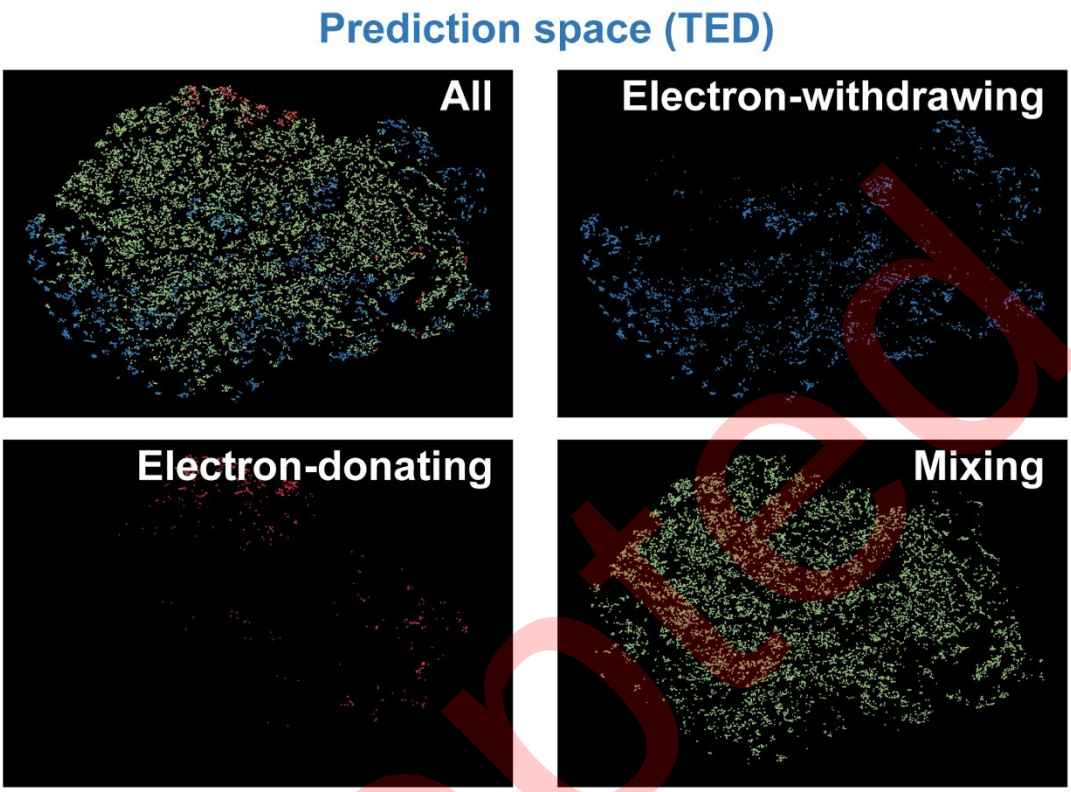


Figure S17. The data points are colored with the functional group types (electron-donating, electron-withdrawing, and mixing groups) in the prediction space of Gap.

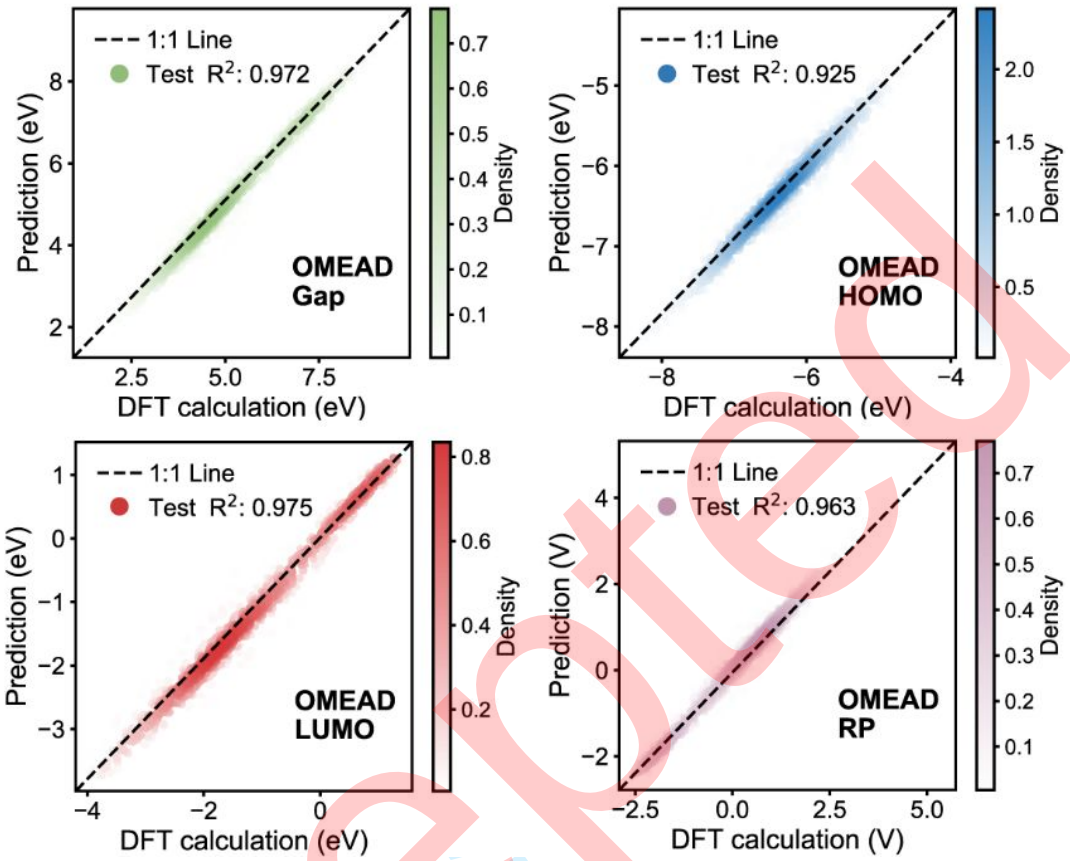


Figure S18. The predicted results of Gap, HOMO, LUMO, and RP for OMEAD using GNGPR.

Table S6. The selected descriptors for SISSO algorithm.

Descriptor	Description		Origination
SBond	Count of single bonds		----
DBond	Count of double bonds		----
TBond	Count of triple bonds		----
UBond	Count of unsaturated bonds		----
ABond	Count of all bonds		----
EW	Counts of electron-withdrawing groups		----
ED	Count of electron-donating groups		----
AAtom	Count of all atoms		----
MV	Molecule volume		RDKit
R	Atomic covalent radius		----
En	Pauli electronegativity		----
SBAB	SBond/ABond		----
DBAB	DBond/ABond		----
TBAB	TBond/ABond		----
UBAB	UBond/ABond		----
TC	Theoretical capacity		----
qed	Quantitative estimate of drug-likeness		RDKit
EnRAA	$(\sum_N^i(En_i/R_i) * A_i)/AAtom$		----
EnAA	$(\sum_N^i(En_i * A_i)/AAtom$		----
RAA	$(\sum_N^i R_i * A_i)/AAtom$		----
EnRMV	$(\sum_N^i(En_i/R_i) * A_i)/MV$		----
EnMV	$(\sum_N^i(En_i * A_i)/MV$		----
RMV	$(\sum_N^i R_i * A_i)/MV$		----
SBMV	SBond/MV		----
DBMV	DBond/MV		----
TBMV	TBond/MV		----
UBond/MV	UBond/MV		----
EWMV	EW/MV		----
EDMV	ED/MV		----
Sub1	SubFPC1/MV	SubFPC	Padel
Sub9	SubFPC9/MV		
Sub18	SubFPC18/MV		
Sub28	SubFPC28/MV		
Sub49	SubFPC49/MV		
Sub96	SubFPC96/MV		
Sub133	SubFPC133/MV		
Sub135	SubFPC135/MV		
Sub137	SubFPC137/MV		

Sub139	SubFPC139/MV		
Sub169	SubFPC169/MV		
Sub171	SubFPC171/MV		
Sub172	SubFPC172/MV		
Sub175	SubFPC175/MV		
Sub181	SubFPC181/MV		
Sub184	SubFPC184/MV		
Sub188	SubFPC188/MV		
Sub224	SubFPC224/MV		
Sub274	SubFPC274/MV		
Sub275	SubFPC275/MV		
Sub287	SubFPC287/MV		
Sub288	SubFPC288/MV		
Sub294	SubFPC294/MV		
Sub295	SubFPC295/MV		
Sub296	SubFPC296/MV		
Sub297	SubFPC297/MV		
Sub298	SubFPC298/MV		
Sub299	SubFPC299/MV		
Sub300	SubFPC300/MV		
Sub301	SubFPC301/MV		
Sub302	SubFPC302/MV		
Sub307	SubFPC307/MV		
MWHI	BCUT2D_MWHI	BCUT2D	RDKit
MWLOW	BCUT2D_MWLOW		
CHGHI	BCUT2D_CHGHI		
CHGLO	BCUT2D_CHGLO		
LOGPHI	BCUT2D_LOGPHI		
LOGPLOW	BCUT2D_LOGPLOW		
MRHI	BCUT2D_MRHI		
MRLOW	BCUT2D_MRLOW		
SVSA1	SlogP_VSA1	SlogP	RDKit
SVSA2	SlogP_VSA2		
SVSA3	SlogP_VSA3		
SVSA4	SlogP_VSA4		
SVSA5	SlogP_VSA5		
SVSA6	SlogP_VSA6		
SVSA7	SlogP_VSA7		
SVSA8	SlogP_VSA8		
SVSA9	SlogP_VSA9		
SVSA10	SlogP_VSA10		

SVSA11	SlogP_VSA11		
SVSA12	SlogP_VSA12		
Chi0	Rev. Comput. Chem. 2:367-422 (1991)		RDKit
Chi0n			
Chi0v			
Chi1			
Chi1n			
Chi1v			
Chi2n			
Chi2v			
Chi3n			
Chi3v			
Chi4n			
Chi4v			

Table S7. The fitting parameters of SISSO for Gap.

	$y_{SISSO}=ax_1+d$	$y_{SISSO}=ax_1+bx_2+d$	$y_{SISSO}=ax_1+bx_2+cx_3+d$
a	-10.11656	-9.46115	-8.90098
b	----	2.76411	2.84066
c	----	----	-0.13274
d	3.62012	3.69012	3.52160

Table S8. The fitting parameters of SISSO for TED.

	$y_{SISSO}=ax_1+d$	$y_{SISSO}=ax_1+bx_2+d$	$y_{SISSO}=ax_1+bx_2+cx_3+d$
a	11.60170	11.02244	10.74278
b	----	-8019.74503	-9203.03814
c	----	----	-462.45118
d	-71.35810	-123.79000	-126.68600

Table S9. The R^2 and mean absolute error (MAE) of SISSO for Gap and TED.

Database	Metrics	$y_{SISSO}=ax_1+d$		$y_{SISSO}=ax_1+bx_2+d$		$y_{SISSO}=ax_1+bx_2+cx_3+d$	
		Gap	TED	Gap	TED	Gap	TED
MOAL	R^2	0.592	0.897	0.651	0.935	0.690	0.942
	MAE	0.266	88.991	0.247	69.477	0.228	67.461
OQEMDB	R^2	0.679	0.872	0.693	0.897	0.731	0.898
	MAE	0.182	62.559	0.179	55.035	0.166	54.894

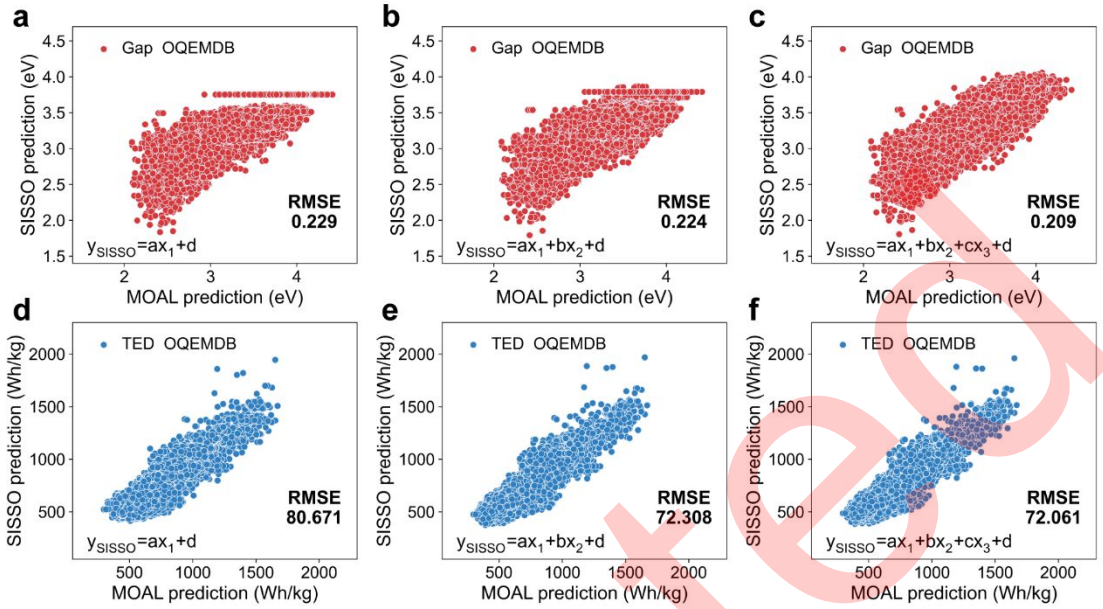


Figure S19. The predicted results of SISSO formulas with one (a), two (b), and three (c) descriptors for Gap in the prediction space of OQEMDB. The predicted results of SISSO formulas with one (d), two (e), and three (f) descriptors for TED in the prediction space of OQEMDB.

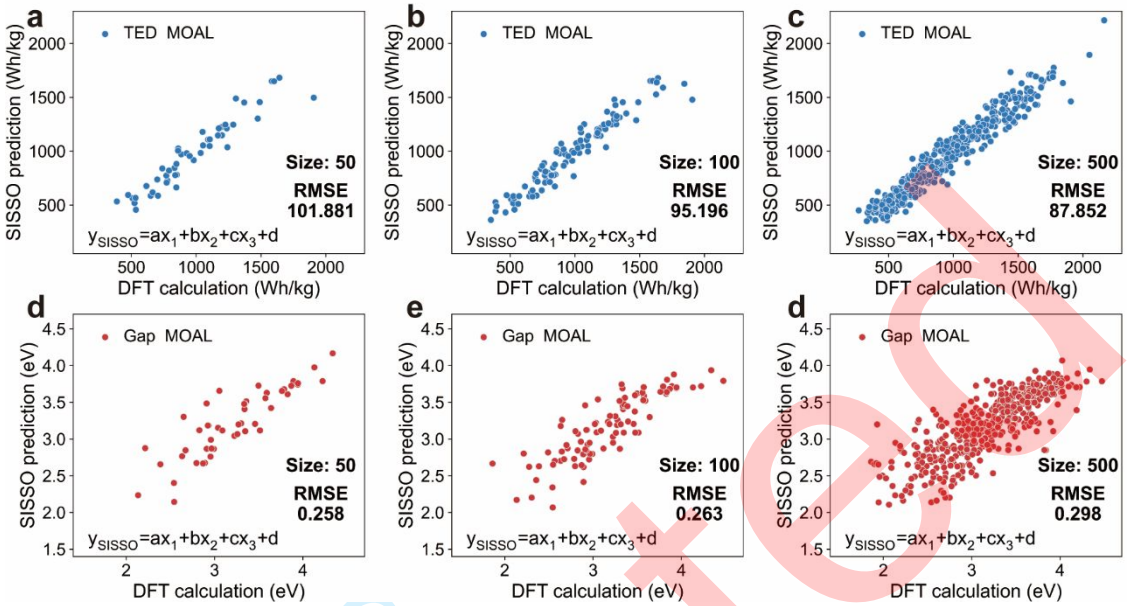


Figure S20. The predicted results of SISSO algorithm with the datasets of 50, 100, and 500 molecules.

References

1 https://www.rdkit.org, accessed on 2014

2 Rogers D, Hahn M. *J Chem Inf Model*, 2010, 50: 742-754

3 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Louppe G, Prettenhofer P, Weiss R, Weiss RJ, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay EJJMLR. *J Mach Learn Res*, 2011, 12: 2825-2830

4 Himanen L, Jäger MOJ, Morooka EV, Federici C, Filippo, Ranawat YS, Gao DZ, Rinke P, Foster AS. *Comput Phys Commun*, 2020, 247: 106949

5 https://pytorch.org/, accessed on 2018-04