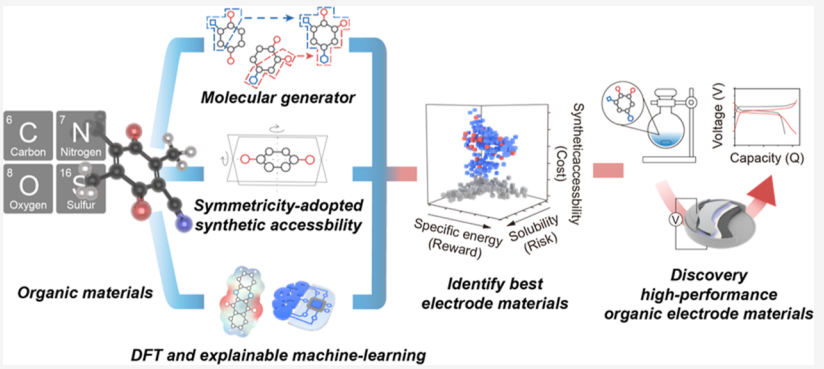
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

Zero-Shot Discovery of High-Performance, Low-Cost Organic Battery Materials Using Machine Learning

Jaehyun Park,[§] Farshud Sorourifar,[§] Madhav R. Muthyala, Abigail M. Houser, Madison Tuttle, Joel A. Paulson,* and Shiyu Zhang*

Cite This: J. Am. Chem. Soc. 2024, 146, 31230−31239
Read Online
ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Organic electrode materials (OEMs), composed of abundant elements such as carbon, nitrogen, and oxygen, offer sustainable alternatives to conventional electrode materials that depend on finite metal resources. The vast structural diversity of organic compounds provides a virtually unlimited design space; however, exploring this space through Edisonian trial-and-error approaches is costly and time-consuming. In this work, we develop a new framework, SPARKLE, that combines computational chemistry, molecular generation, and machine learning to achieve zero-shot predictions of OEMs that simultaneously balance reward (specific energy), risk (solubility), and cost (synthesizability). We demonstrate that SPARKLE significantly outperforms alternative black-box machine learning algorithms on interpolation and extrapolation tasks. By deploying SPARKLE over a design space of more than 670,000 organic compounds, we identified ≈5000 novel OEM candidates. Twenty-seven of them were synthesized and fabricated into coin-cell batteries for experimental testing. Among SPARKLE-discovered OEMs, 62.9% exceeded benchmark performance metrics, representing a 3-fold improvement over OEMs selected by human intuition alone (20.8% based on six years of prior lab experience). The top-performing OEMs among the 27 candidates exhibit specific energy and cycling stability that surpass the state-of-the-art while being synthesizable at a fraction of the cost.

INTRODUCTION

Organic electrode materials (OEMs) are promising components for energy storage in rechargeable batteries due to their high structural diversity and synthetic tunability.[1−10] Replacing expensive, unsustainably sourced redox-active metals with abundant carbon-based OEMs could substantially lower energy storage costs, as they can potentially be derived from biological and petroleum sources. However, the virtually unlimited design space of OEMs presents a significant challenge in identifying candidates with optimal stability, solubility, and redox potentials. Modifying functional groups, both in identity and position, can substantially alter several critical and interrelated factors, including intermolecular interactions, electronic structures, and solid-state packing. These changes can profoundly influence the performance metrics of OEMs in highly nonintuitive ways. Consequently, existing discovery processes for new OEMs still largely depend on Edisonian trial-and-error synthesis and electrochemical testing, which are inherently costly and time-intensive.[11] To overcome this challenge, one can rely on molecular "proxy property" values that strongly correlate to device-scale performance. Redox potential,[12−15] solubility,[12,15−18] and structural stability[19−22] are examples of OEM proxy properties that have been previously studied in the context of battery applications.

To date, accurately predicting these and related proxy OEM properties remains a challenging task. This has motivated the use of computational methods, like density functional theory (DFT), to increase the rate at which OEM property data can

Received: August 23, 2024
Revised: October 18, 2024
Accepted: October 21, 2024
Published: November 1, 2024

https://doi.org/10.1021/jacs.4c11663
J. Am. Chem. Soc. 2024, 146, 31230−31239

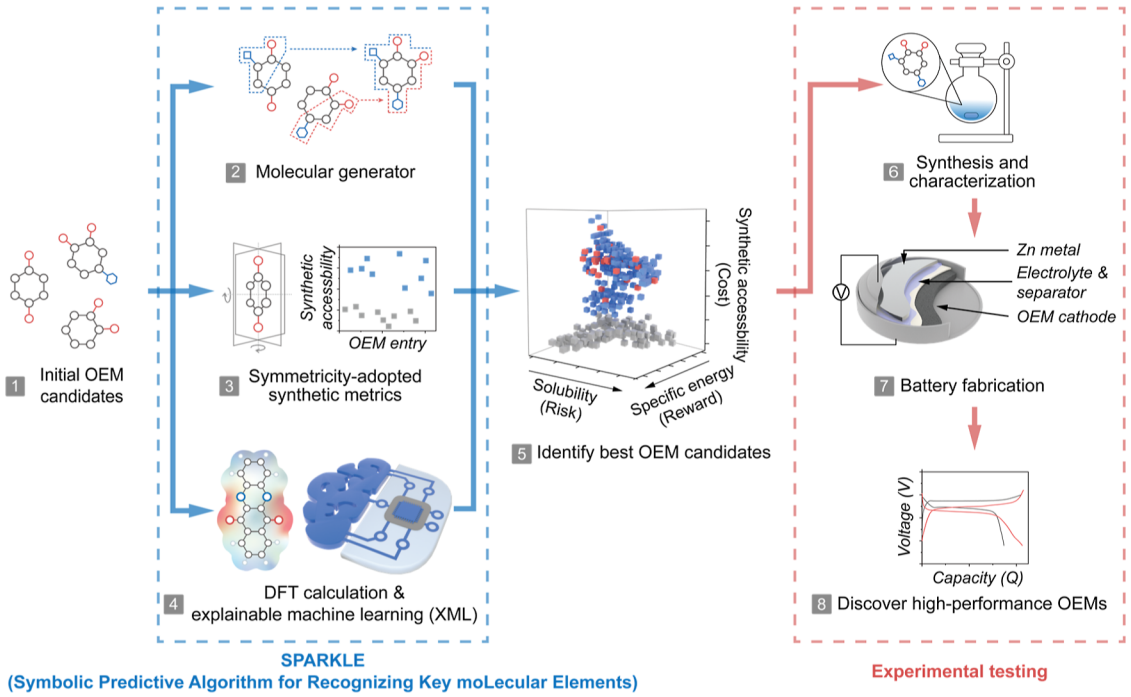


Figure 1. Proposed machine-learning-based workflow for discovery of high-performance OEMs. (left to right) (1) We start with an initial set of OEM candidates that are fed to the proposed SPARKLE (symbolic predictive algorithm for recognizing key molecular elements) method consisting of three parts, (2) a generative AI that expands the initial molecules to a much larger set of candidates, (3) an easy-to-calculate synthetic cost measure, and (4) a machine learning algorithm that uses small amounts of DFT data to learn simple symbolic models for specific energy (e) and solubility (ΔG_{sol}). (5) Multiobjective optimization is used to discover OEM candidates' trade-off between specific energy, solubility, and cost. (6) The top-performing OEM candidates are synthesized, (7) fabricated into aqueous zinc-ion batteries (AZIBs), and (8) tested under over 500 charge−discharge cycles to measure realistic device-level performance metrics as a final screening step.

be collected.[12,13,15−19] However, the computational cost of these DFT calculations is very high, making it difficult or impossible to scale to large design spaces with many millions of OEM candidates. As such, many ML methods have been investigated to construct cheap-to-evaluate surrogate models to replace the expensive DFT calculations,[11,20,23−29] including neural networks (NN),[30,31] random forests (RF),[32,33] Gaussian processes,[34] and support vector machines.[35] However, these traditional ML methods are susceptible to overfitting in the low data limit, such that they can require vast training data sets to ensure reliable prediction accuracy. Furthermore, due to their black-box nature, these models also commonly suffer from poor extrapolation performance, transferability, and human interpretability, which significantly limits their ability to facilitate novel scientific insights and discovery.

The main goal of this work is to introduce a multiobjective framework for the systematic design of high-performance, low-cost OEMs in battery applications that only rely on small amounts of proxy property data. The proposed framework, referred to as SPARKLE (symbolic predictive algorithm for recognizing key molecular elements) has three major components (Figure 1).

First, we used a state-of-the-art molecular generation method to construct a new library of over half of a million OEMs, many of which have not previously been synthesized in the lab. Second, we created a symmetry-aware synthetic accessibility score that provides an easy-to-calculate evaluation of the synthetic cost. Third, we developed a machine learning algorithm for molecular property prediction, such as solvation energy (risk) and specific energy (reward), that combines molecular descriptors with a symbolic regression approach to learn meaningful mathematical expressions from limited OEM-property data sets.[36,37]

The ML models identified by the proposed method significantly outperform popular black-box ML methods, including RF, NN, and sparse linear regression. SPARKLE fuses the predictions of solvation energy (risk), specific energy (reward), and cost (synthesizability) to identify trade-offs between these different objectives, making it, to our knowledge, the first fully multiobjective design approach for OEMs.

Utilizing SPARKLE on the library of generated OEMs, we identified ≈5000 promising novel candidates with an excellent balance between risk, reward, and cost. Twenty-seven of these OEMs were synthesized and experimentally tested in AZIBs. The combination of a low-cost OEM cathode with a zinc anode offers a safe, affordable, and environmentally friendly solution for renewable energy storage. Our initial search for OEMs focused on aqueous battery applications, since aqueous solvation energies are relatively easy to compute with DFT and have been widely used as a proxy for predicting aqueous solubility, making these systems more accessible for modeling.

We found that 17 of the 27 resulted in a galvanostatic charge−discharge profile that exceeds established benchmark conditions. This OEM selection success rate of 62.9% is more than three times higher than the collective success rate of 20.8% achieved by expert human selection in our laboratory over the past six years. The top-performing OEM candidate found by SPARKLE has specific energy and cycling stability that surpass the state-of-the-art in the literature while being synthesizable at a significantly lower cost. We highlight that these results were obtained in the absence of any experimental training data, which represents a fully "zero-shot discovery"

https://doi.org/10.1021/jacs.4c11663
J. Am. Chem. Soc. 2024, 146, 31230−31239

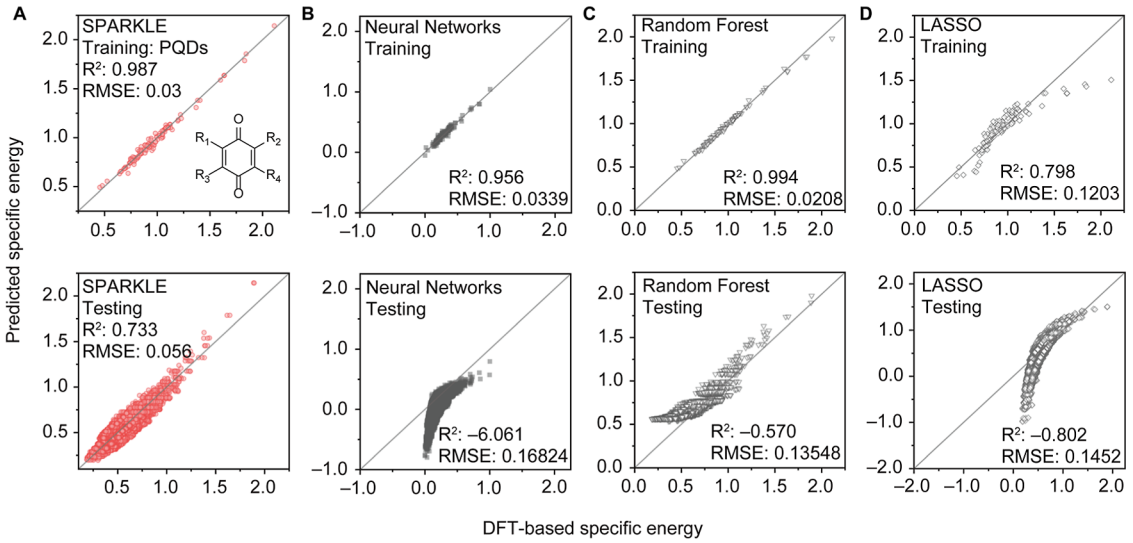


Figure 2. Training and testing performance of SPARKLE versus-discovered and conventional machine learning models for specific energy prediction. The training performance (top) and testing performance (bottom) achieved by (A) SPARKLE, (B) NN, (C) RF, and (D) LASSO linear regression models for specific energy. The training set consists of only 100 paraquinones while the testing set consists of over 103,000 quinones. We see that most models achieve a good training coefficient for determination (R^2) and root mean squared error; however, SPARKLE is the only one that achieves good prediction performance on the test set. Since the test set is more than 1000 times the size of the training set and includes many non-paraquinones, this demonstrates the strong generalization properties afforded by SPARKLE (much more robust to overfitting than conventional black-box machine learning methods).

process that is empowered by the extrapolation capabilities of the machine learning models.[38−41]

RESULTS AND DISCUSSION

Generating a Large Library of Novel Organic Electrode Materials. As highlighted in Figure 1, our approach uses an initial set of OEMs to generate a sufficiently large design space for exploration. A library of ≈140,000 quinone-derived molecules was generated in a previous study from a combination of programmatic substitution on various aromatic backbones and intuition-driven suggestions from chemists.[42] Although many candidates with low redox potentials were identified, molecules with high redox potential, critical for cathode materials of interest in this work, were much rarer. Therefore, this library, which is one of the largest existing collections of possible OEMs, is not suitable for our current work. The first key step of our proposed method is to enlarge this initial library using molecular generation methods.

To ensure the generated molecules are more likely to be synthesizable, we first perform a filtering step that retains only the fully oxidized para- and meta-quinones, resulting in a seed library of ≈103,000 molecules that serves as the seeding data for the generator.

Many generative artificial intelligence methods have been explored for molecular generation; here, we focus on the recently proposed fast assembly of SMILES Fragments (FASMIFRA) method due to its robustness and efficiency.[43] Applying FASMIFRA to the filtered library of ≈103,000 molecules, we generated a new, larger library of ≈670,000 molecules, increasing the design space size by more than a factor of 6. Figure S1 shows the generated library has a broader range of properties, such as molecular weight, resulting from the creation of molecules outside of the training set. As we show later, this feature is useful for designing previously unseen OEMs.

Developing a ML Model Using Compressed Sensing. Given the large design space, we need to execute the modeling and inference process outlined in Figure 1 by selecting the proxy molecular property values of interest. In accordance with prior studies, we focus on specific energy and solvation energy that, respectively, influence the energy density and viability of a battery made with an OEM cathode. Specific energy acts as our reward metric since larger values are likely to improve battery performance, while solvation energy acts as our risk metric since larger values are expected to lose electroactive material due to its dissolution. Specific energy e (units of W h kg^{-1}) can be directly computed from redox potential E° (units of V) and molecular weight M_w (units of g) as follows

$$e = \frac{n \times F \times (E^{\circ} - E_{\text{anode}})}{3600 \times M_w} \quad (1)$$

where n is the number of electrons (equal to 2 assuming a two-electron reduction process), F is the Faraday constant (units of C mol^{-1}), and E_{anode} is the redox potential of the zinc anode (units of V). The E° value can be estimated using DFT for quinones following the procedure introduced by Pineda Flores et al.[12] We use a similar DFT procedure for calculating the solvation-free energy ΔG_{solv}. Further details on these calculations are provided in the Supporting Information.

Computing e and ΔG_{solv} for all ≈670,000 molecules in our newly generated library would be too computationally expensive. Therefore, we explore an alternative data-driven modeling approach that, unlike many popular black-box modeling methods, can perform well in the low data regime (Figure S2).

Our model starts by representing molecules as a collection of molecular descriptors (i.e., outcomes of a well-defined mathematical procedure that transforms a symbolic representation of a molecule into a numerical quantity; Figure S2A). We focus on $D \sim 1800$ descriptors computable by the open-source Mordred package,[44] including a broad mix of theoretical

https://doi.org/10.1021/jacs.4c11663
J. Am. Chem. Soc. 2024, 146, 31230−31239

descriptors derived from count, fragment, fingerprint, and graph invariant operations. We then perform a filtering and expansion step over the descriptor vector $\boldsymbol{x} \in \mathbb{R}^D$ to create a very high-dimensional feature space consisting only of simple mathematical expressions (Figure S2B).

Next, we use a sure independence screening method to reduce the large D-dimensional space to the top 100 descriptors and apply three levels of recursion to generate a final feature space with more than 109 possible expressions.[45,46] Lastly, we apply a compressed sensing method, SISSO,[36,47] to identify the optimal expression given the existing training data (Figure S2C−E). A complete mathematical description of our proposed ML method is provided in the Supporting Information (Supporting Information Section 3). The corresponding code used to generate all results in this paper can be found on GitHub at this link: https://github.com/PaulsonLab/SPARKLE.

Predicting Specific Energy and Solvation Energy Using ML. Next, we deployed the proposed ML method to learn interpretable models for specific energy and solvation energy. Since our goal is to learn from as little data as possible, we start with DFT calculations for a small subset (≈100) of the design space.

Our ML procedure found the following simple model for their specific energy

$$\hat{e} = 1717.16 \times \left(\frac{x_{\mathrm{BE}}}{x_{\mathrm{BP}} \times (x_{\mathrm{MB1}} + x_{\mathrm{MB2}})} \right) + 0.027413 \tag{2}$$

where $\hat{e}$ is the predicted specific energy and x_{BE}, x_{BP}, x_{MB1}, and x_{MB2} are easy-to-calculate Mordred descriptors (see Supporting Information Section 4 for more information). In essence, the x_{BE} and x_{BP} terms provide measures of electronegativity and polarizability, respectively, while x_{MB1} and x_{MB2} are related to molecular weight.

Our model shows a major advantage over alternative black-box machine learning methods [NN,[30,31] RF,[32,33] and linear regression with L1 regularization (LASSO)[48,49]] in terms of generalization performance (Figure 2). Considering the DFT predictions of specific energy on the ≈103,000 quinones as the test set,[42] our ML model achieves a test R^2 value of 0.733 while the NN, RF, and LASSO models all end up with negative R^2 values (implying heavily biased models, Figure 2 bottom).

It is worth noting that this level of prediction performance was obtained using only ≈100 data points (less than 0.1% of the test data set), suggesting the discovered model (eq 2) is robust to the overfitting of the training data. The reduced level of overfitting is well beyond what is achievable with traditional black-box machine learning, which is a consequence of the sparsity enforced by SPARKLE. We perform further analysis of the robustness and reproducibility of the SPARKLE algorithm to different random splits of the training data in Supporting Information Section 4 and Figures S4 and S5. We find that, given ≈100 randomly selected training points, SPARKLE identifies a structurally unique model that has very low prediction error, indicating that the results shown in Figure 2 are insensitive to the specific train/test split.

We believe sparsity is a key factor that allows SPARKLE to outperform other methods. More complex models often have many degrees of freedom, leading to high variance. Additionally, SPARKLE's ability to engineer new features enables it to learn complex models with relatively few trainable parameters, a major reason for its better performance than LASSO. Although LASSO also induces sparsity, it strictly enforces linear relationships with respect to the base features, limiting its ability to capture nonlinear interactions.

We repeated a similar procedure for solvation energy; however, we found the initial set of ≈100 molecules was insufficient for learning an accurate model. Instead, we opted for a training set of solvation energies of 3000 randomly selected molecules. The model discovered by our method using this training set can be expressed as follows

$$\Delta\hat{G}_{\mathrm{solv}} = -8.2725 \times 10^{-4} \times \left(\frac{x_{\mathrm{MB1}} \times (x_{\mathrm{MIDN}} + x_{\mathrm{GC3P}})}{x_{\mathrm{GC2SE}}} \right) - 9.31160 \times 10^{-2} \tag{3}$$

where $\Delta\hat{G}_{\mathrm{solv}}$ (units of kcal mol^{-1}) is the predicted solvation energy and x_{MB1}, x_{MIDN}, x_{GC3P}, and x_{GC2SE} are the relevant Mordred descriptors that are, respectively, related to molecular weight, number of nitrogen atoms, sigma electrons, and polarizability. The R^2 values obtained on the training and test sets are 0.832 and 0.827, respectively, again showing strong generalization performance on the held-out ≈100,000 quinones. More information on the Mordred descriptors and the training and testing performance can be found in the Supporting Information (Supporting Information Section 5 and Figures S6−S8).

Constructing an Effective Synthesizability Metric. Since the generated library of OEMs was made through random fragmentation and recombination of seed molecules, it is possible that many of them cannot be readily synthesized in the lab. This limits us from assessing their performance in the desired end-use application of solid-state batteries. To facilitate systematic search over the large-scale OEM library, we need to develop a synthesizability metric to serve as a proxy for the "cost" of device-level fabrication of a given molecule.

Although there has been significant recent work on the development of synthetic accessibility scores, such as SYBA,[50] SCScore,[51] and SAScore,[52] we found that these models have a tendency to recommend asymmetric molecules that require relatively complicated multistep synthesis procedures from costly reagents. This result is not particularly surprising given these scores were developed for pharmaceutical drug discovery applications, which places less emphasis on the cost of starting materials.

A key component of our work, however, is the discovery of affordable OEMs that can be readily produced from low-cost feedstock chemicals. Therefore, we propose a simple modification (or fine-tuning) of the existing SCScore that boosts the score for symmetric molecules. Not only is symmetry expected to reduce cost, as they typically require fewer, simpler synthetic steps, but it also is expected to improve the solid-state ordering of the molecules when they crystallize. This improved ordering increases the lattice energy and thus is anticipated to decrease the solubility of OEMs when exposed to an electrolyte.[16]

The proposed symmetry-adapted synthetic accessibility (SASA) score is expressed by the following mathematical relationship

$$\mathrm{SASA} = \frac{\mathrm{TA}}{\mathrm{SCScore} + \mathrm{UA}} \tag{4}$$

where TA refers to the total number of non-hydrogen atoms in an organic structure, SCScore refers to the metric introduced

https://doi.org/10.1021/jacs.4c11663
J. Am. Chem. Soc. 2024, 146, 31230−31239

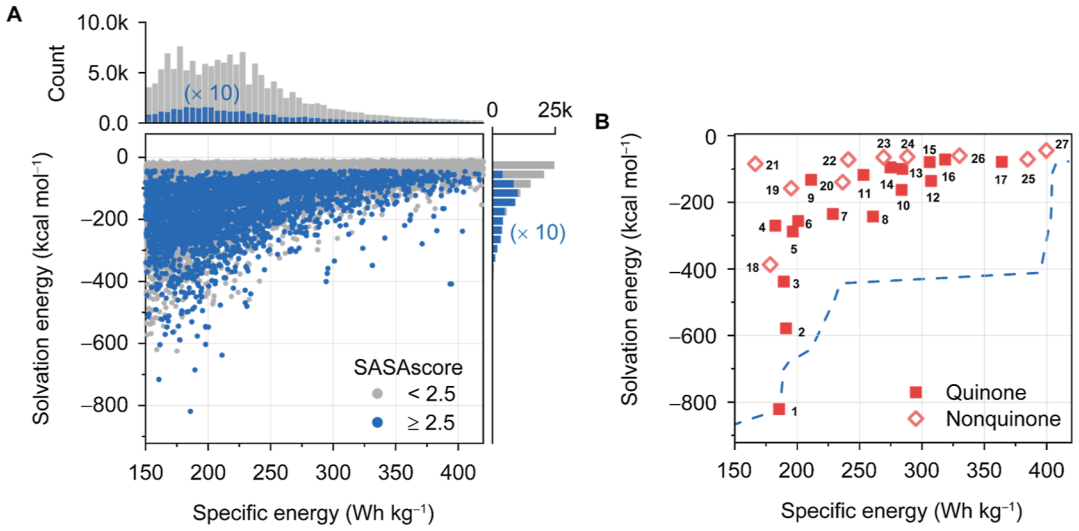


Figure 3. Deployment of SPARKLE to predict specific energy (reward), solvation energy (risk), and SASA (cost) of OEMs. (A) The plot shows initial 60,000 candidates from the 600 k design space. Two histograms indicate the distribution of the number of OEM candidates for hard-to-synthesize molecules with SASA < 2.5 (gray) and easy-to-synthesize molecules with SASA ≥ 2.5 (blue). The histogram for SASA ≥ 2.5 was magnified by ten times for visibility. (B) A plot of twenty-seven final OEM candidates selected for experimental testing. The rectangles indicate quinone-based OEMs while diamonds show nonquinone OEMs recommended by SPARKLE prediction.

in Coley et al.,[51] and UA refers to the number of chemically unique non-hydrogen atoms in an organic structure. Larger values of SASA are predicted to be easier to synthesize molecules.

To determine the effectiveness of SASA compared to SCScore, we selected a random pool of ≈400 molecules and ranked them from easiest to hardest to synthesize according to each of these scores. Four chemists then selected the top ≈30 easiest-to-synthesize molecules from this pool. The resulting synthesizable molecules pool had a 61.8% agreement with the ones selected by SASA score, which is much higher than state-of-the-art (21.8%). Further information on this validation process as well as how to efficiently compute the proposed SASA metric using the Python package RDKit is provided in Supporting Information (Supporting Information Section 6 and Figures S9−S12).

Discovery of New High-Performance, Low-Cost Electrode Materials. Given validated predictive models for the specific energy (reward), solvation energy (risk), and synthesizability (cost) of a molecule, along with the large, diverse design space consisting of over 670,000 compounds, we can now search for novel OEMs for various end-use applications (Figure S13). We focus on discovering OEMs suitable for AZIBs, where the combination of a low-cost OEM cathode with a zinc anode could serve as a safe, affordable, and environmentally friendly approach for renewable energy storage.

We take a multiobjective optimization approach to identify an optimal balance between the three key properties of interest (reward, risk, and cost). Since the generated library is much too large to sort through manually, we first implement a series of constraints to prioritize molecules that are more likely to exhibit high performance in each dimension. Specifically, we require SASA to be greater than 2.5, the specific energy to be greater than 175 W h kg^{-1}, and the solvation energy to be less than −20 kcal mol^{-1}. The SASA threshold γ was selected by showing random molecules from the filtered library with SASA $\geq \gamma$ to chemists; the smallest value of γ that resulted in most of the random molecules to be easy-to-synthesize was found to be 2.5. The specific energy and solvation energy thresholds were selected to be in line with the current state-of-the-art values reported in the literature.[53−57]

Figure 3A shows the two-dimensional distribution of predicted specific energy and solvation energy values for the large OEM library for compounds likely to be easy (blue) and hard (gray) to synthesize. Around 5000 of them exceed all three thresholds; a complete list of their SMILES string representations and their predicted property values are included in Supporting Information (Supporting Document S2). We manually explored this set of ≈5000 molecules to identify 27 promising OEMs deemed worthy of experimental device-level testing (Figure 3B). A manual selection of OEM candidates was performed, considering the viability of the synthesis, including (1) number of synthetic step, (2) relative ease of synthesis, (3) commercial availability, (4) cost of starting materials, and (5) expected yield. The detailed chemical structures of these candidates are provided in Figure S14 (labeled as numbers **1**−**27**). We prioritized searching for candidates near the predicted Pareto frontier between specific energy and solvation energy, wherein one cannot improve one property without hurting the other. One can effectively think of these as "optimal" molecules that achieve the best balance between risk and reward. In addition, we highlight that these 27 molecules have never been previously reported in the literature and were found completely in the absence of any electrochemical cycling data. It should be noted that the selected candidates may not fall within the optimal redox windows in aqueous environment to avoid side reactions such as oxygen/hydrogen evolution reactions; however, the generated OEMs were seeded from quinone-based molecules and thus have a similar structure with learning scope (see Supporting Information Section 2), giving us a confidence that OEM candidates would have similar redox window with seeding quinone molecules. A final key observation is that, SPARKLE can predict molecules outside of the training set and recommend promising nonquinone OEMs. For example, **26** and **27** are benzidines, commonly used in dyes, a class of compounds not yet considered OEMs.

https://doi.org/10.1021/jacs.4c11663
J. Am. Chem. Soc. 2024, 146, 31230−31239

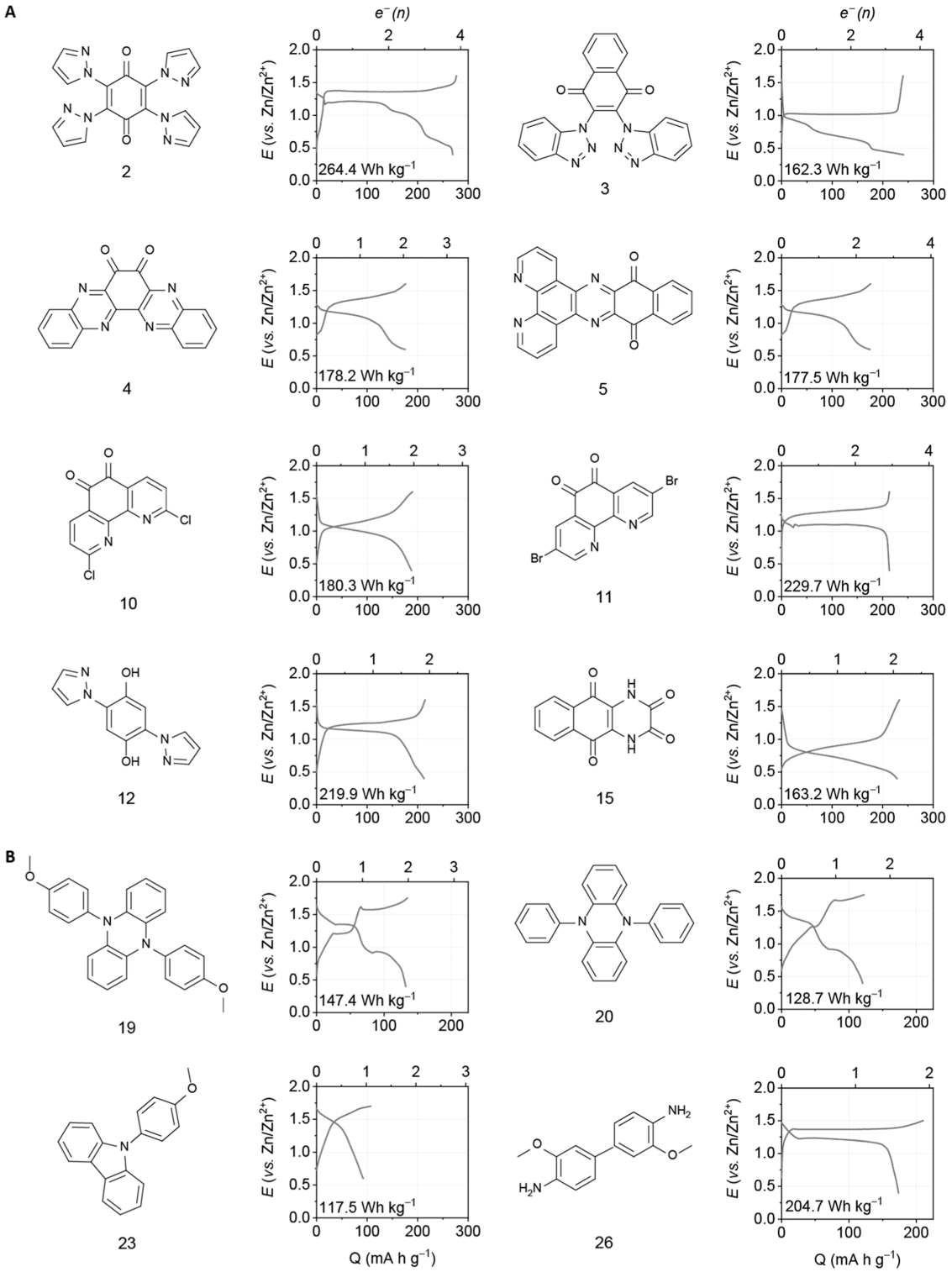


Figure 4. Selected OEMs identified by SPARKLE and corresponding galvanostatic charge−discharge (GCD) profiles at 0.5 C charge−discharge rate. (A) GCD profiles of eight representative *n*-type OEMs (**2**−**5**, **10**, **11**, **12**, and **15**). (B) GCD profiles of four representative p-type OEMs (**19**, **20**, **23**, and **26**). GCD for other successful OEMs, including **1**, **7**, **8**, **14**, and **27**, and other unsuccessful candidates were plotted in Figures S13 and S14.

Next, the 27 discovered molecules were experimentally synthesized and evaluated as the cathode in an AZIB constructed using two-electrode coin cells with a standard aqueous zinc-based electrolyte. The first galvanostatic charge−discharge (GCD) profile at a constant current density of 0.5 C along with their chemical structure are summarized in Figure 4; results for the other five successful and five unsuccessful OEMs are reported in Supporting Information (Figures S15 and S16). The OEMs predicted by SPARKLE show excellent cycling performance. For example, compound **12** demonstrated a high operating voltage of 1.15 V, an extraordinary energy density of 220 W h kg^{-1}, and high Coulombic and

https://doi.org/10.1021/jacs.4c11663
J. Am. Chem. Soc. 2024, 146, 31230−31239

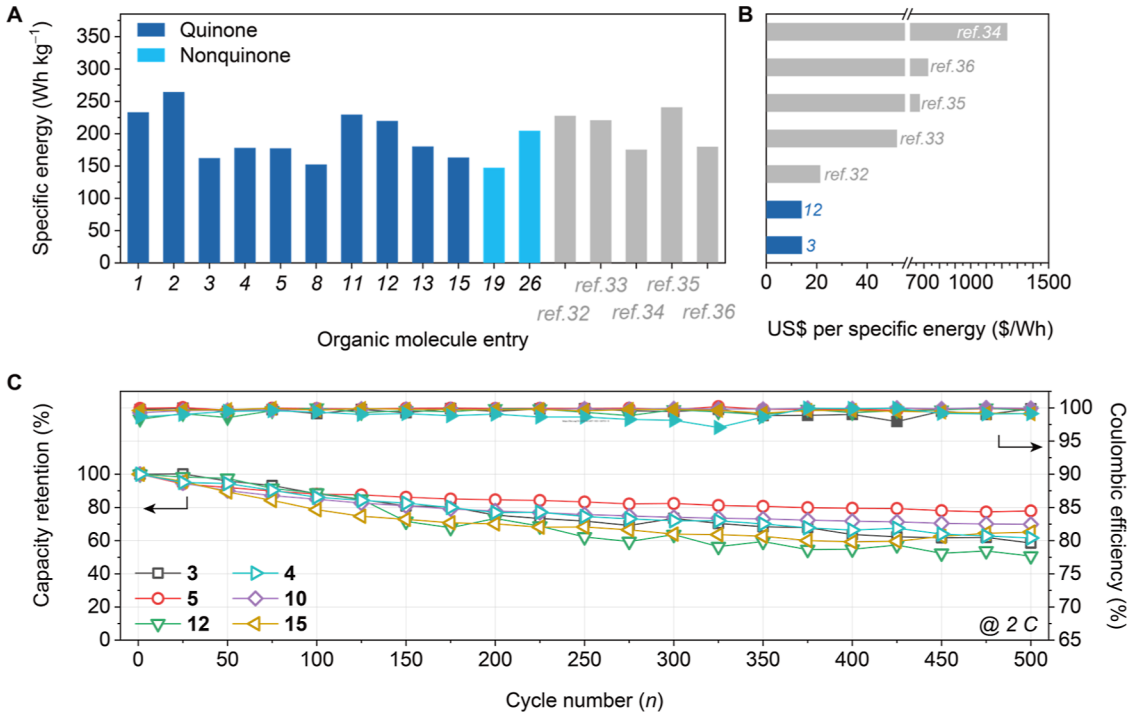


Figure 5. Comparison of OEMs identified by SPARKLE and common reported OEMs. (A) Comparison of the specific energy of OEMs identified by SPARKLE compared to existing OEMs reported in literature. (B) Comparison of the estimated cost per specific energy of selected OEMs. The cost estimation is based on the price of synthesis reagents, solvents, and reaction yields. Details can be found in Tables S3 and S4. (C). Long-term cycling stability of new OEMs identified by SPARKLE at 2 C.

round-trip efficiencies of 99.5% and 82%, respectively. It is also interesting to note that our approach found new p-type organic moieties (**19**, **20**, **26**, and **27**), providing further evidence of the ability of SPARKLE to generalize beyond the quinone-only (n-type) training set.

SPARKLE was able to find multiple OEMs that match or improve upon state-of-the-art OEMs in at least one objective. Figure 5 provides an overview of key performance metrics for a subset of the 17 successful OEM candidates identified by SPARKLE. We see in Figure 5A that, for example, molecule **2** has a higher specific energy than five published state-of-the-art OEMs (see Table S1 for their chemical details).[53−57] Note that we omitted OEMs that show good capacity but require a special membrane to prevent material dissolution to ensure a fair comparison.[58] More importantly, some of the high performing OEMs identified by SPARKLE are significantly less costly than the state-of-the-art (Figure 5B). Specifically, candidates **3** and **12** achieve the lowest cost per energy stored, which was estimated from the starting materials and solvents needed to synthesize the compound. We note that not all of the OEMs discovered by SPARKLE are highly cost-efficient. This outcome is not unexpected since the SASA score does not consider the commercial availability of the reagents (see Supporting Information Section 7 and Tables S2 and S3 for a detailed breakdown of the cost estimates). Nonetheless, SPARKLE demonstrates a significantly higher success rate compared to other heuristic selection methods. This further underscores its capability to effectively screen for new OEMs that are both high-performing and synthetically viable.

Figure 5C shows the results of long-term stability tests. Compounds **3**−**5**, **10**, **12**, and **15** show suitable cycling stability for realistic AZIB devices. Moreover, compound **5**, the most stable OEM that we found, could operate over 1500 cycles with a final capacity retention of 70.4% (see Figure S17). It is interesting to note that although our model cannot directly predict device stability since the decomposition pathways vary significantly as a function of molecular structure (Figure S18), its ability to eliminate bad candidates is very helpful in practice. By significantly reducing the feasible design space, one can focus experimental effort on candidates more likely to succeed, which is critical due to the large time and cost investment needed for long-term cycling studies.

To quantify the practical impacts of SPARKLE, we compared its performance to that achieved by traditional human intuition-based OEM selection in our laboratory over the past six years. Prior to this work, we had tested 72 molecules as OEM cathodes in AZIB under the same conditions. Human intuition-based selection achieved a low success rate of 20.8% (15 out of 72; Figures S19 and S20 and Table S4) where "success" is defined as achieving a GCD cycle that satisfies standard performance benchmarks, including a Coulombic efficiency greater than 95%, specific capacity greater than 150 mA h g^{-1} for n-types and 100 mA h g^{-1} for p-types, and a voltage hysteresis less than 0.5 V. Although we cannot eliminate the involvement of human intuition in SPARKLE, the ML-assisted discovery process was able to increase the success percentage by more than a factor of 3 (62.9% corresponding to 17 out of 27; Figure S21). This substantially improved probability of success was achieved despite searching over a much larger design space than was previously impossible by humans alone. While there is some discrepancy between prediction and measurements in specific energy (Figure S22), the consideration of enlarged design spaces is important, as it enabled us to discover a new, promising class of OEMs, e.g., benzidines, in an entirely zero-shot fashion.

https://doi.org/10.1021/jacs.4c11663
J. Am. Chem. Soc. 2024, 146, 31230−31239

CONCLUSIONS

In this work, we developed an efficient multiobjective discovery approach, SPARKLE, which integrates advances in computational chemistry, molecular generation, and explainable machine learning to address the time and cost bottlenecks inherent in traditional Edisonian trial-and-error search methods. At its core, we note that SPARKLE is a machine-learning model, and the insights it provides may not be as granular as those obtained through traditional trial-and-error methods. However, the primary aim of our manuscript is to showcase an accelerated discovery process. We created a publicly available web interface that allows users to enter a molecule and compute predictions of their viability as electrode materials in aqueous batteries: https://oem-webapp.streamlit.app/.

The application of SPARKLE toward the discovery of OEMs holds significant potential for developing sustainable, safe, and low-cost battery technology. Using small amounts of property data calculated by DFT, SPARKLE learns accurate and interpretable models for specific energy and solvation energy, key properties that impact battery performance. By combining these models with a novel synthesizability metric, SPARKLE identified over 5000 unique OEM candidates suitable for aqueous zinc-ion battery applications. Analyzing the Pareto frontier for these performance metrics, we discovered 27 candidates predicted to optimally balance reward (specific energy), risk (solvation energy), and cost (synthesizability). Experimental synthesis and testing of these 27 candidates in coin-cell batteries revealed a 62.9% success rate in exceeding benchmark performance metrics, a more than 3-fold increase compared to the 20.8% success rate from human intuition-guided OEM selection alone. In addition to increased efficiency, SPARKLE identified OEMs that improve upon the existing state-of-the-art for each property, including an OEM with a specific energy greater than 250 W h kg^{-1}. Two candidates achieved the lowest estimated synthetic cost per specific energy to date.

Another key observation is the ability of SPARKLE-discovered models to generalize beyond the training set, which is not possible with existing black-box machine-learning methods. SPARKLE discovered several nonquinones, including four new p-type OEMs. Two of these p-type OEMs are benzidines, representing a previously unreported OEM moiety that warrants further study. Thus, this work not only provides specific recommendations for promising OEMs but also demonstrates the immense potential for extrapolation. Finally, we note that the algorithms developed in SPARKLE are broadly applicable, and we expect the framework to be useful in many end-use applications beyond OEMs.

ASSOCIATED CONTENT

Data Availability Statement

The code used to generate the results in this paper can be found on GitHub at this link: https://github.com/PaulsonLab/SPARKLE. DOI: 10.5281/zenodo.13947711. We also created a publicly available web interface that allows users to enter a molecule and compute predictions of its specific energy as the cathode in an aqueous zinc-ion battery, aqueous solvation energy, and symmetry-adapted synthetic accessibility score. This tool is available at the following link: https://oem-webapp.streamlit.app/.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.4c11663.

Experimental procedures, synthetic details, figures S1−S56, tables S1−S4 and references (PDF)

Complete lists of 5000 OEM information (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Joel A. Paulson − *Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio 43210, United States*; orcid.org/0000-0002-1518-7985; Email: paulson.82@osu.edu

Shiyu Zhang − *Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States*; orcid.org/0000-0002-2536-4324; Email: zhang.8941@osu.edu

Authors

Jaehyun Park − *Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States*

Farshud Sorourifar − *Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio 43210, United States*

Madhav R. Muthyala − *Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio 43210, United States*

Abigail M. Houser − *Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States*

Madison Tuttle − *Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.4c11663

Author Contributions

[§]J.P. and F.S. authors contributed equally.

Notes

The authors declare the following competing financial interest(s): We have filed a patent on the application of the reported machine learning algorithm in discovering new organic electrode materials.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Graduate Research Fellowship under Grant no. DGE-1343012, the National Science Foundation under CBET-2124604, the Sustainability Institute at Ohio State University, and the Center for Emergent Materials, NSF MRSEC, under award number DMR-2011876. We acknowledge the Nagib group (OSU) for helping validate the synthetic accessibility model.

REFERENCES

(1) Liang, Y.; Jing, Y.; Gheytani, S.; Lee, K. Y.; Liu, P.; Facchetti, A.; Yao, Y. Universal Quinone Electrodes for Long Cycle Life Aqueous Rechargeable Batteries. *Nat. Mater.* **2017**, *16* (8), 841−848.

(2) Cong, G.; Wang, W.; Lai, N. C.; Liang, Z.; Lu, Y. C. A High-Rate and Long-Life Organic−Oxygen Battery. *Nat. Mater.* **2019**, *18* (4), 390−396.

(3) Jezowski, P.; Crosnier, O.; Deunf, E.; Poizot, P.; Béguin, F.; Brousse, T. Safe and Recyclable Lithium-Ion Capacitors Using Sacrificial Organic Lithium Salt. *Nat. Mater.* **2018**, *17* (2), 167−173.
(4) Armand, M.; Grugeon, S.; Vezin, H.; Laruelle, S.; Ribière, P.; Poizot, P.; Tarascon, J. M. Conjugated Dicarboxylate Anodes for Li-Ion Batteries. *Nat. Mater.* **2009**, *8* (2), 120−125.
(5) Wang, J.; Lakraychi, A. E.; Liu, X.; Sieuw, L.; Morari, C.; Poizot, P.; Vlad, A. Conjugated Sulfonamides as a Class of Organic Lithium-Ion Positive Electrodes. *Nat. Mater.* **2021**, *20* (5), 665−673.
(6) Lin, K.; Chen, Q.; Gerhardt, M. R.; Tong, L.; Kim, S. B.; Eisenach, L.; Valle, A. W.; Hardee, D.; Gordon, R. G.; Aziz, M. J.; Marshak, M. P. Alkaline Quinone Flow Battery. *Science* **2015**, *349* (6255), 1529−1532.
(7) Liang, Y.; Yao, Y. Designing Modern Aqueous Batteries. *Nat. Rev. Mater.* **2023**, *8* (2), 109−122.
(8) Lee, M.; Hong, J.; Lopez, J.; Sun, Y.; Feng, D.; Lim, K.; Chueh, W. C.; Toney, M. F.; Cui, Y.; Bao, Z. High-Performance Sodium-Organic Battery by Realizing Four-Sodium Storage in Disodium Rhodizonate. *Nat. Energy* **2017**, *2* (11), 861−868.
(9) Kim, D. J.; Yoo, D. J.; Otley, M. T.; Prokofjevs, A.; Pezzato, C.; Owczarek, M.; Lee, S. J.; Choi, J. W.; Stoddart, J. F. Rechargeable Aluminium Organic Batteries. *Nat. Energy* **2019**, *4* (1), 51−59.
(10) Lu, Y.; Chen, J. Prospects of Organic Electrode Materials for Practical Lithium Batteries. *Nat. Rev. Chem* **2020**, *4* (3), 127−142.
(11) Peng, J.; Schwalbe-Koda, D.; Akkiraju, K.; Xie, T.; Giordano, L.; Yu, Y.; Eom, C. J.; Lunger, J. R.; Zheng, D. J.; Rao, R. R.; Muy, S.; Grossman, J. C.; Reuter, K.; Gómez-Bombarelli, R.; Shao-Horn, Y. Human- and Machine-Centred Designs of Molecules and Materials for Sustainability and Decarbonization. *Nat. Rev. Mater.* **2022**, *7* (12), 991−1009.
(12) Pineda Flores, S. D.; Martin-Noble, G. C.; Phillips, R. L.; Schrier, J. Bio-Inspired Electroactive Organic Molecules for Aqueous Redox Flow Batteries. 1. Thiophenoquinones. *J. Phys. Chem. C* **2015**, *119* (38), 21800−21809.
(13) Allam, O.; Cho, B. W.; Kim, K. C.; Jang, S. S. Application of DFT-Based Machine Learning for Developing Molecular Electrode Materials in Li-Ion Batteries. *RSC Adv.* **2018**, *8* (69), 39414−39420.
(14) Xu, S.; Liang, J.; Yu, Y.; Liu, R.; Xu, Y.; Zhu, X.; Zhao, Y. Machine Learning-Assisted Discovery of High-Voltage Organic Materials for Rechargeable Batteries. *J. Phys. Chem. C* **2021**, *125* (39), 21352−21358.
(15) Kristensen, S. B.; van Mourik, T.; Pedersen, T. B.; Sørensen, J. L.; Muff, J. Simulation of Electrochemical Properties of Naturally Occurring Quinones. *Sci. Rep.* **2020**, *10*, 13571.
(16) Robinson, S. G.; Yan, Y.; Hendriks, K. H.; Sanford, M. S.; Sigman, M. S. Developing a Predictive Solubility Model for Monomeric and Oligomeric Cyclopropenium-Based Flow Battery Catholytes. *J. Am. Chem. Soc.* **2019**, *141* (26), 10171−10176.
(17) Er, S.; Suh, C.; Marshak, M. P.; Aspuru-Guzik, A. Computational Design of Molecules for an All-Quinone Redox Flow Battery. *Chem. Sci.* **2015**, *6* (2), 885−893.
(18) Tuttle, M. R.; Brackman, E. M.; Sorourifar, F.; Paulson, J.; Zhang, S. Predicting the Solubility of Organic Energy Storage Materials Based on Functional Group Identity and Substitution Pattern. *J. Phys. Chem. Lett.* **2023**, *14* (5), 1318−1325.
(19) Sevov, C. S.; Hickey, D. P.; Cook, M. E.; Robinson, S. G.; Barnett, S.; Minteer, S. D.; Sigman, M. S.; Sanford, M. S. Physical Organic Approach to Persistent, Cyclable, Low-Potential Electrolytes for Flow Battery Applications. *J. Am. Chem. Soc.* **2017**, *139* (8), 2924−2927.
(20) Lee, B.; Yoo, J.; Kang, K. Predicting the Chemical Reactivity of Organic Materials Using a Machine-Learning Approach. *Chem. Sci.* **2020**, *11* (30), 7813−7822.
(21) Sowndarya, S. V. S.; St. John, P. C.; Paton, R. S. A Quantitative Metric for Organic Radical Stability and Persistence Using Thermodynamic and Kinetic Features. *Chem. Sci.* **2021**, *12* (39), 13158−13166.
(22) Assary, R. S.; Zhang, L.; Huang, J.; Curtiss, L. A. Molecular Level Understanding of the Factors Affecting the Stability of Dimethoxy Benzene Catholyte Candidates from First-Principles Investigations. *J. Phys. Chem. C* **2016**, *120* (27), 14531−14538.
(23) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268−276.
(24) Koscher, B. A.; Canty, R. B.; McDonald, M. A.; Greenman, K. P.; McGill, C. J.; Bilodeau, C. L.; Jin, W.; Wu, H.; Vermeire, F. H.; Jin, B.; Hart, T.; Kulesza, T.; Li, S. C.; Jaakkola, T. S.; Barzilay, R.; Gómez-Bombarelli, R.; Green, W. H.; Jensen, K. F. Autonomous, Multiproperty-Driven Molecular Discovery: From Predictions to Measurements and Back. *Science* **2023**, *382* (6677), No. eadi1407.
(25) Bradford, G.; Lopez, J.; Ruza, J.; Stolberg, M. A.; Osterude, R.; Johnson, J. A.; Gomez-Bombarelli, R.; Shao-Horn, Y. Chemistry-Informed Machine Learning for Polymer Electrolyte Discovery. *ACS Cent. Sci.* **2023**, *9*, 206−216.
(26) Walters, W. P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc. Chem. Res.* **2021**, *54* (2), 263−270.
(27) Agarwal, G.; Doan, H. A.; Robertson, L. A.; Zhang, L.; Assary, R. S. Discovery of Energy Storage Molecular Materials Using Quantum Chemistry-Guided Multiobjective Bayesian Optimization. *Chem. Mater.* **2021**, *33* (20), 8133−8144.
(28) Lombardo, T.; Duquesnoy, M.; El-Bouysidy, H.; Årén, F.; Gallo-Bueno, A.; Jørgensen, P. B.; Bhowmik, A.; Demortière, A.; Ayerbe, E.; Alcaide, F.; Reynaud, M.; Carrasco, J.; Grimaud, A.; Zhang, C.; Vegge, T.; Johansson, P.; Franco, A. A. Artificial Intelligence Applied to Battery Research: Hype or Reality? *Chem. Rev.* **2022**, *122* (12), 10899−10969.
(29) Szymanski, N. J.; Bartel, C. J. Computationally Guided Synthesis of Battery Materials. *ACS Energy Lett.* **2024**, *9*, 2902−2911.
(30) Vermeire, F. H.; Chung, Y.; Green, W. H. Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures. *J. Am. Chem. Soc.* **2022**, *144* (24), 10785−10797.
(31) Jin, W.; Barzilay, R.; Jaakkola, T. Chapter 11: Junction Tree Variational Autoencoder for Molecular Graph Generation. In *Artificial Intelligence in Drug Discovery*; *RSC Drug Discovery Series*; The Royal Society of Chemistry, 2020; 2021-Janua, pp 228−249..
(32) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947−1958.
(33) Chen, C.-H.; Tanaka, K.; Funatsu, K. Random Forest Model with Combined Features: A Practical Approach to Predict Liquid-Crystalline Property. *Mol. Inf.* **2019**, *38* (4), 1800095.
(34) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **2021**, *121* (16), 10073−10141.
(35) Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45* (3), 549−561.
(36) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO A Compressed-Sensing Method for Identifying the Best Low-Dimensional Descriptor in an Immensity of Offered Candidates. *Phys. Rev. Mater.* **2018**, *2*, 083802.
(37) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley, 2009; ..
(38) Long, S.; Zhou, Y.; Dai, X.; Zhou, H. Zero-Shot 3D Drug Design by Sketching and Generating. In *Advances in Neural Information Processing Systems*; NeurIPS, 2022; Vol. 35.
(39) Zhao, H.; Liu, S.; Ma, C.; Xu, H. *A Unified Graph-Text Model for Instruction-Based Molecule Zero-Shot Learning GIMLET*. In *NeurIPS*, 2023, pp 1−38.
(40) Seidl, P.; Renz, P.; Dyubankova, N.; Neves, P.; Verhoeven, J.; Wegner, J. K.; Segler, M.; Hochreiter, S.; Klambauer, G. Improving Few- and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks. *J. Chem. Inf. Model.* **2022**, *62*, 2111−2120.

(41) Wang, Y.; Xia, Y.; Yan, J.; Yuan, Y.; Shen, H.-B.; Pan, X. ZeroBind: A Protein-Specific Zero-Shot Predictor with Subgraph Matching for Drug-Target Interactions. *Nat. Commun.* **2023**, *14* (1), 7861.
(42) Tabor, D. P.; Gómez-Bombarelli, R.; Tong, L.; Gordon, R. G.; Aziz, M. J.; Aspuru-Guzik, A. Mapping the Frontiers of Quinone Stability in Aqueous Media: Implications for Organic Aqueous Redox Flow Batteries. *J. Mater. Chem. A* **2019**, *7* (20), 12833−12841.
(43) Berenger, F.; Tsuda, K. Molecular Generation by Fast Assembly of (Deep)SMILES Fragments. *J. Cheminf.* **2021**, *13*, 88.
(44) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminf.* **2018**, *10* (1), 4.
(45) Sondhi, P. *Feature Construction Methods: A Survey*, 2009; Vol. *69*, pp 70−71. sifaka.cs.uiuc.edu.
(46) Iguyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157−1182.
(47) Donoho, D. L. Compressed Sensing. *IEEE Trans. Inf. Theory* **2006**, *52* (4), 1289−1306.
(48) Santosa, F.; Symes, W. W. Linear Inversion of Band-Limited Reflection Seismograms. *SIAM J. Sci. Stat. Comput.* **1986**, *7* (4), 1307−1330.
(49) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. Roy. Stat. Soc. B Stat. Methodol.* **1996**, *58* (1), 267−288.
(50) Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian Estimation of Synthetic Accessibility of Organic Compounds. *J. Cheminf.* **2020**, *12*, 35.
(51) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58* (2), 252−261.
(52) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1*, 8.
(53) Nam, K. W.; Kim, H.; Beldjoudi, Y.; Kwon, T.; Kim, D. J.; Stoddart, J. F. Redox-Active Phenanthrenequinone Triangles in Aqueous Rechargeable Zinc Batteries. *J. Am. Chem. Soc.* **2020**, *142* (5), 2541−2548.
(54) Tie, Z.; Liu, L.; Deng, S.; Zhao, D.; Niu, Z. Proton Insertion Chemistry of a Zinc−Organic Battery. *Angew. Chem., Int. Ed.* **2020**, *59* (12), 4920−4924.
(55) Gao, Y.; Li, G.; Wang, F.; Chu, J.; Yu, P.; Wang, B.; Zhan, H.; Song, Z. A High-Performance Aqueous Rechargeable Zinc Battery Based on Organic Cathode Integrating Quinone and Pyrazine. *Energy Storage Mater.* **2021**, *40*, 31−40.
(56) Guo, Z.; Ma, Y.; Dong, X.; Huang, J.; Wang, Y.; Xia, Y. An Environmentally Friendly and Flexible Aqueous Zinc Battery Using an Organic Cathode. *Angew. Chem., Int. Ed.* **2018**, *57* (36), 11737−11741.
(57) Lin, Z.; Shi, H.-Y.; Lin, L.; Yang, X.; Wu, W.; Sun, X. A High Capacity Small Molecule Quinone Cathode for Rechargeable Aqueous Zinc-Organic Batteries. *Nat. Commun.* **2021**, *12* (1), 4424.
(58) Zhao, Q.; Huang, W.; Luo, Z.; Liu, L.; Lu, Y.; Li, Y.; Li, L.; Hu, J.; Ma, H.; Chen, J. High-Capacity Aqueous Zinc Batteries Using Sustainable Quinone Electrodes. *Sci. Adv.* **2018**, *4* (3), No. eaao1761.

https://doi.org/10.1021/jacs.4c11663
J. Am. Chem. Soc. 2024, 146, 31230−31239