# Discovery of Energy Storage Molecular Materials using Quantum Chemistry-guided Multi-objective Bayesian Optimization

| | |
|---|---|
| Journal: | *Chemistry of Materials* |
| Manuscript ID | cm-2021-020409.R3 |
| Manuscript Type: | Article |
| Date Submitted by the Author: | n/a |
| Complete List of Authors: | Agarwal, Garvit; Argonne National Laboratory, Materials Science Division<br>Doan, Hieu; Argonne National Laboratory, Materials Science Division<br>Robertson, Lily; Argonne National Laboratory, Chemical Sciences and Engineering Division<br>Zhang, Lu; Argonne National Laboratory, Electrochemical Energy Storage<br>Assary, Rajeev; Argonne National Laboratory, Materials Science Division |
| | |

# Discovery of Energy Storage Molecular Materials using Quantum Chemistry-guided Multi-objective Bayesian Optimization

Garvit Agarwal[1,2,†], Hieu A. Doan[1,2,†], Lily A. Robertson[1,3], Lu Zhang[1,3] and Rajeev S. Assary[1,2,*]

[1] Joint Center for Energy Storage Research (JCESR)

[2]Materials Science Division, Argonne National Laboratory, Lemont, IL 60439, USA

[3]Chemical Sciences and Engineering Division, Argonne National Laboratory, Lemont, IL 60439, USA

[†]GA and HAD contributed equally to this work.

[*]**Corresponding Author**

**Rajeev S. Assary –** *Materials Science Division, Argonne National Laboratory, Lemont, IL, USA, 60439*; *Phone: 630-252-3536; Email: assary@anl.gov*

## Abstract

Redox flow batteries (RFBs) are a promising technology for stationary energy storage applications due to their flexible design, scalability, and low cost. In RFBs, energy is carried in flowable redox-active materials (redoxmers) which are stored externally and pumped to the cell during operation. Further improvements in energy density of RFBs necessitates redoxmers design with wider redox potential window and higher solubility. Additionally, designing redoxmers with fluorescence enabled self-reporting functionality allows monitoring of the state-of-health of RFBs. To accelerate the discovery of redoxmers with desired properties, state-of-the-art machine learning (ML) methods such as Multi-objective Bayesian Optimization (MBO) is useful. Here, we first employed density functional theory calculations to generate a database of reduction potentials, solvation free energies, and absorption wavelengths for 1400 redoxmer molecules based on a 2,1,3-benzothiadiazole (BzNSN) core structure. From the computed properties, we identified 22 Pareto-optimal molecules that compromise all desired properties. We further utilized this data to develop and benchmark a MBO approach to identify candidates quickly and efficiently with multiple targeted properties. With MBO, optimal candidates from the 1400 molecule dataset can be identified at least 15 times more efficiently compared to brute force or random selection approach. Importantly, we utilized this approach for discovering promising redoxmers from unseen database of 1 million BzNSN based molecules, where we discovered 16 new Pareto-optimal molecules with significant improvements in properties over the initial 1400 molecules. We anticipate that this active learning technique is general and can be utilized for the discovery of any class of functional materials that satisfies multiple desired property criteria.

## Introduction

The rising energy demand requires integration of intermittent renewable energy sources, such as solar and wind, with the electric grid to maintain a carbon neutral economy [1,2]. This requires the development of long duration stationary energy storage technologies. Redox flow batteries, utilizing either aqueous or nonaqueous electrolytes, are considered a promising strategy to address this challenge [3,4,5]. Nonaqueous redox flow batteries (NRFBs) offer several advantages over their aqueous counterparts, such as wider electrochemical stability window and lower cost, thanks to the use of organic solvents and cost-effective redox-active molecules (redoxmers) [6,7]. However, the development of next-generation NRFBs with high energy density and superior cycling performance is dependent on the discovery of new and improved redoxmers, which poses a major challenge for material design.

Several properties need to be considered when designing redoxmer molecules for high performance NRFBs. For example, the 2,1,3-benzothiadiazole (BzNSN) is a well-studied anolyte redoxmer (which undergoes reduction) in NRFBs that features low redox potential, small molecular weight, high stability of the charged radical anion, and outstanding electrochemical cycling performance [8,9,10]. Recently, a BzNSN derivative, $CH_3$-Ac$BzC_6$, was also engineered with a π-extended acetamide group[11], which enabled an orthogonal molecular property, namely fluorescence, to monitor the crossover of the active species and assess the state-of-health of the battery [11]. Hence, one promising strategy for designing anolyte redoxmers is to perform molecular engineering of the BzNSN scaffold using a wide range of functional groups to achieve the desired properties (*e.g.*, use electron-withdrawing/-donating groups to tune the redox potential). However, high-throughput experimental synthesis and characterization is often a significant bottleneck and high-fidelity computational methods provide a cost-efficient alternative to enable the design and discovery.

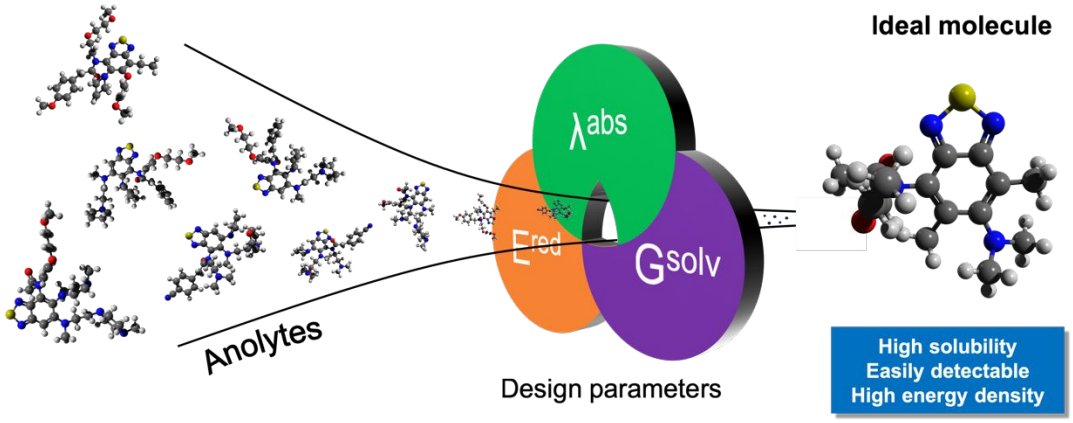
To accelerate materials discovery, high-fidelity density functional theory (DFT) calculations have been used to screen large molecular libraries and guide experiments toward the most optimal candidate molecules [12,13,14,15]. The computed properties, such as redox potentials of the organic molecules, are observed to be in good agreement with experimentally measured values using cyclic voltammetry [16,17,18]. Thus, a combination of molecular engineering and high-throughput DFT calculations are routinely used to identify redoxmer candidates with desired redox potentials for application in NRFBs [16,19,20]. For example, Pelzer *et al.* have performed high-throughput DFT calculations to screen molecules with desired reduction and oxidation potentials from a library of 4178 molecules [21]. Similarly, DFT calculations have been used to develop linear regression models to predict solubilities of the organic molecules using the computed solvation free energies and the dipole moments of the molecules [22]. While such high-throughput materials screening approaches are attractive, the brute-force computational screening methods become intractable as the size of the search space grows beyond few thousand candidates.

The recent emergence of data-driven and machine learning (ML) techniques has accelerated the screening of large search spaces for identification of molecules/materials with desired properties. Several surrogate ML models have been developed using experimental or DFT computed data to accurately and rapidly predict material properties such as band gap [23,24,25], lattice thermal conductivity [26], dielectric constant [27], refractive index [28], thermodynamic stability [29], melting temperature [30], and defect formation energies [31,32]. The prediction accuracy of the ML models typically depends on the diversity and the quantity of data used for training the models. The generation of large quantities of high-fidelity data is computationally expensive and time-consuming, which is one of the major bottlenecks in the development of

generalizable and robust ML models based on supervised learning. Thus, the surrogate ML models perform well within the domain of applicability and often fail for unseen datapoints which are outside the training domain [33,34]. The *a priori* identification of the domain of applicability of such ML models is a challenging task. To circumvent this problem, multiple active learning strategies have been proposed [35,36,37,38,39].

Active learning (AL) is a class of ML strategies in which a learning algorithm interactively queries an information source (or database) to label new data points with the desired outputs. The goal of AL is to strike a good balance between exploitation and exploration to minimize the number of computations or experimental measurements needed to optimize the property of a given material system or achieve desired accuracy of the trained ML model. For example, AL has been successfully used to guide the experiments to accelerate the discovery of new shape-memory alloys with low thermal hysteresis from a search space of 800k compositions [40] and to tune the measured electro-strain of the Pb-free piezoelectric materials [41]. Bassman *et al.* used an AL model based on Bayesian Optimization (BO) to discover layered materials with optimal band gaps [42]. Similarly, Kim *et al.* used AL to search for polymers with high glass transition temperatures using few DFT evaluations [43]. Recently, we used BO framework to identify redoxmers with optimum oxidation potentials for application in NRFBs from a large search space of 112k molecules by performing only 100 DFT calculations [44]. While a specific property of redoxmers such as redox potential, solubility, or even fluorescence may be efficiently identified via single-objective AL, it is always more desirable, albeit challenging, to search for redoxmers with multiple optimized properties. Indeed, while most of the applications of AL for materials design have been limited to optimization of a single objective/property of the materials, there are only a few examples where AL has been successfully used for simultaneous optimization of two or more material properties [45,46,47]. In particular, Janet *et al.* used multi-objective Bayesian optimization to accelerate the search for candidates with optimal combination of two properties (*i.e.*, redox potential and solubility) from a search space of 2.8 million transition metal complexes for application in RFBs [45]. Gopakumar *et al.* demonstrated superior performance of the AL strategy for simultaneous optimization of two properties compared to random search across multiple datasets [46]. Recently, Jablonka *et al.* developed a novel multi-objective AL algorithm which simultaneously optimizes 3 properties (*i.e.*, adsorption free energy, repulsion free energy of dimers and radius of gyration) of the polymer beads and efficiently identify Pareto-optimal candidates from a large search space of polymeric materials [47].

In this work, we have developed an AL framework based on Multi-objective Bayesian Optimization (MBO) to accelerate the search of desired redox-active molecules for application in high energy density NRFBs. As shown in Scheme 1, our goal is to identify anolyte molecules quickly and efficiently with three simultaneously optimized properties including reduction potential ($E^{red}$), solvation free energy ($G^{solv}$), and absorption wavelength ($\lambda^{abs}$). Although the stability of anolyte molecules is also an important criterion for designing long-duration redox flow batteries, it is a function of multiple factors (redox potential window, solvation structure, reactivity at the electrode/electrolyte interface, etc.) and is therefore beyond the scope of this work. The MBO is first benchmarked against a DFT-evaluated dataset of 1400 BzNSN molecules. Then, the model is applied to an unknown dataset of 1 million molecules to determine optimal candidates using only 100 DFT evaluations.




Scheme 1: Multi-objective design of anolyte candidate molecules for non-aqueous redox flow batteries. $E^{red}$, $G^{solv}$, and $\lambda^{abs}$ stand for computed reduction potential, solvation free energy, and absorption wavelength, respectively.

## Results and Discussion

### 1. Structure Enumeration and DFT calculations of BzNSN molecules:

A molecule dataset is generated by engineering the substituent ($R_1$-$R_5$) positions in the scaffold of acetamide-substituted BzNSN molecule with different functional groups as shown in Figure 1(a). The use of Simplified Molecular Input Line Entry System (SMILES) allows for fast and robust text-based enumeration, resulting in a dataset of c.a. 1400 BzNSN molecules. This dataset consists of 7 chemical elements, H, Br, C, N, O, S, and F. The size of molecules ranges from 20 to 36 heavy atoms (non-H atoms). Before DFT evaluations, the initial 3D molecular structures are automatically generated from SMILES representations and subsequently optimized using MMF94 forcefield as implemented in RDKit cheminformatics package [48]. Then, DFT calculations are performed to compute the reduction potentials ($E^{red}$), solvation free energies ($G^{solv}$) and absorption wavelengths ($\lambda^{abs}$) of the molecules (see Computational Details section). The distributions of the computed $E^{red}$, $G^{solv}$ and $\lambda^{abs}$ of the 1400 BzNSN molecule dataset are shown in Figure 1 (b), 1(c), and 1(d), respectively. The computed properties are observed to vary over a wide range of values, indicating a diverse dataset of molecular properties. Shown in Figure 1(b), the computed reduction potential ($E^{red}$) varies in the range of ~1.5 to ~3.0 V vs. $Li/Li^+$. Shown in Figure 1(c), the computed solvation free energy ($G^{solv}$) varies in range of -1.2 eV to -0.2 eV. Also, shown in Figure 1(d), the computed absorption wavelength varies from 300 to 500 nm. The statistics (minimum, maximum, mean, standard deviation) of the computed properties ($E^{red}$, $G^{solv}$ and $\lambda^{abs}$) are summarized in Table 1.

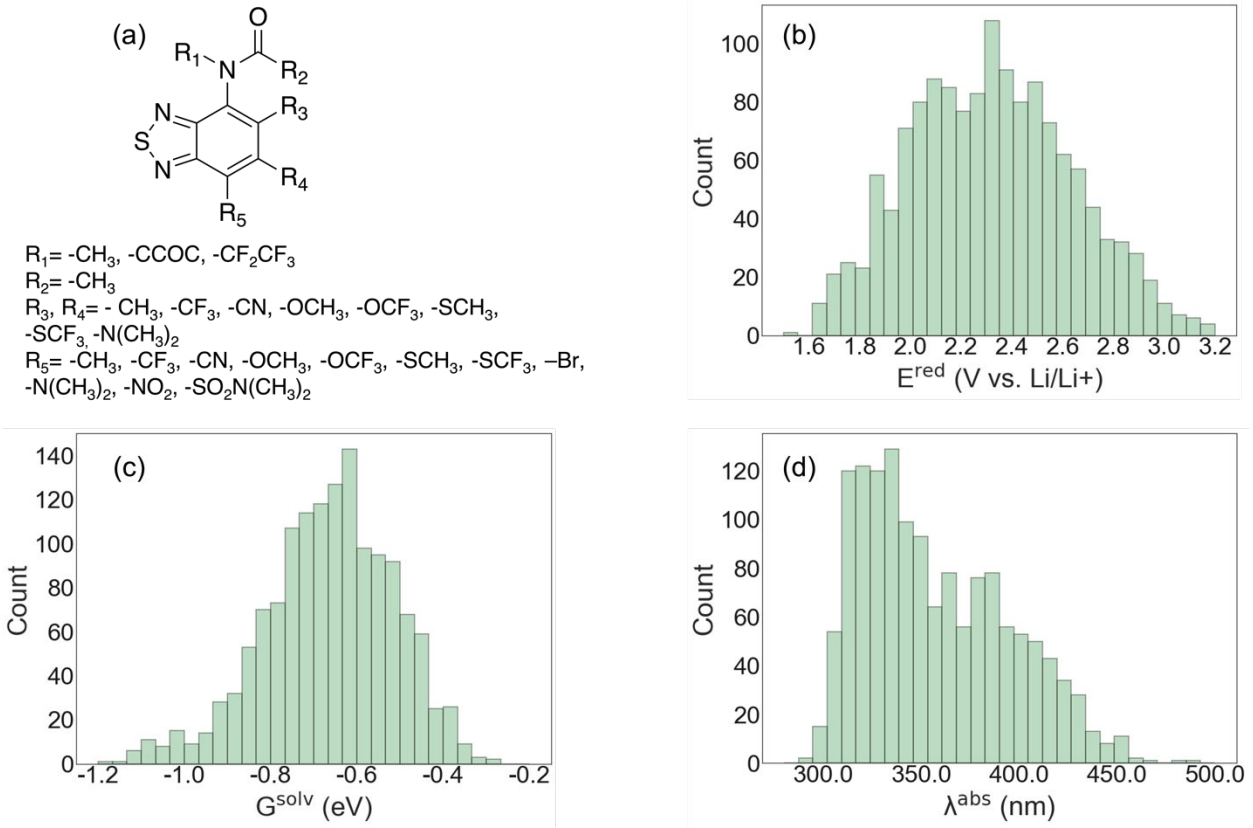



Figure 1: (a) Scaffold with "R" groups ($R_1$-$R_5$) used for enumerating the dataset of 1400 BzNSN molecules. Distributions of the DFT (wb97xd/6-31+G(d,p) using SMD solvation model using acetonitrile) computed (b) reduction potential ($E^{red}$), (c) solvation free energy ($G^{solv}$) and (d) absorption wavelength ($\lambda^{abs}$) of the 1400 molecules.

Among the computed properties, lower $E^{red}$ is crucial to expand the electrochemical window of the active species and thus helps to improve the operating voltage of a NRFB. More negative $G^{solv}$ (from DFT) is a reasonable indicator of higher solubility of the BzNSN anolytes in acetonitrile solvent, which can result in increased concentration of the active species in solution. We note that the actual solubility measurements using computations are not possible due to lack of sublimation energies of the molecular materials. In a recent work, Robertson *et al.* have identified that BzNSN molecules with $\lambda^{abs}$ values ranging from 350-400 nm exhibit fluorescent activity [11]. Our DFT calculations of a subset of 1400 BzNSNs also confirms that molecules with $\lambda^{abs}$ in this range also possess emission wavelength ($\lambda^{em}$) in the visible range (Figure S1). Since the computational evaluation of $\lambda^{em}$ is significantly more expensive than that of $\lambda^{abs}$, we used the latter as an approximate screening indicator of fluorescent activity. Particularly, to search for new molecules that are easily detectable via fluorescent activity in the electrolyte solutions, a target value of 375 nm is considered desirable for $\lambda^{abs}$. The ideal material candidate must therefore be designed by minimizing the values of $E^{red}$, $G^{solv}$, and $|\lambda^{abs}-375|$ concurrently. Unlike single property optimization, the simultaneous optimization of multiple properties requires identification of a Pareto front, in which the Pareto-optimal datapoints represent the best trade-off among properties. Here, we have used an AL strategy based on MBO to accelerate the search of the Pareto-optimal datapoints while utilizing minimal number of expensive DFT calculations.

Table 1: List of the minimum, maximum, mean, and standard deviation values of the DFT-computed reduction potential ($E^{red}$), solvation free energy ($G^{solv}$) and absorption wavelength ($\lambda^{abs}$) of the 1400 BzNSN molecule dataset

| Property | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|
| $E^{red}$ (V vs $Li/Li^+$) | 1.541 | 3.378 | 2.32 | 0.320 |
| $G^{solv}$ (eV) | -1.190 | -0.293 | -0.664 | 0.151 |
| $\lambda^{abs}$ (nm) | 293.76 | 487.49 | 357.429 | 37.453 |

**2. Multi-objective Bayesian Optimization:**

Multi-objective Bayesian Optimization (MBO) is a class of multi-objective optimization in which a surrogate model, *e.g.*, Gaussian Process Regression (GPR, See Computational Details section), is used to search for the optimal candidates based on an improvement metrics such as expected improvement (EI, see Computational Details section). By defining how the EI of a multi-objective function may be evaluated, several methods have been proposed in the recent literature. In one approach by Knowles, called ParEGO, N objectives are aggregated into a single objective via parametrized scalarizing weight factors and a single-objective EI is then used for the resulting single-objective function [49]. Further modifications of ParEGO for improved computational speed and efficiency were also proposed by Liu *et al.* [50] and Zhang *et al.* [51]. Recently, Häse and colleagues introduced a new lexicographical approach to combine multiple objectives into one, in which the contribution from each objective is based on its hierarchical order of importance[52]. In the case where objective aggregation is not preferred, either multiple-objective EI or multiple single-objective EIs must be evaluated with respect to a Pareto front. For example, Keane derives a 2-objective EI equation that computes the probability of augmenting the current Pareto front so that a new dominating solution can be determined [53]. The magnitude of improvement is then calculated with respect to the closest point on the current Pareto front. In another approach, the improvement metrics is defined by the S-metric or hypervolume increment to the Pareto front and can be computed using EI in the hypervolume formulation [54,55]. In general, multi-objective EI equations are mathematically complex and may not be easily expanded to large number of objectives ($N_{objective}$>2). A possible approach to circumvent such difficulty is to employ multiple single-objective EI equations. As shown by Jeong and Obayashi, for every candidate, an EI will be computed for each objective, and the resulting set of EIs is treated as fitness values for selecting the optimal candidate [56]. Beyond MBO, it is important to mention that Jablonka and co-workers recently proposed a novel multi-objective active learning algorithm that optimizes toward the Pareto front by directly using the GPR-predicted mean and uncertainty values [47]. In this work, we employ a similar approach to Jeong and Obayashi's method, in which the EIs for different objectives are computed independently followed by the construction of a Pareto front in the EI space. The main advantage of our approach lies in the efficient implementation that enables evaluation of multi-dimensional Pareto front in large datasets (3D Pareto front and up to 1 million datapoints in this work).

The overall scheme for our MBO workflow is shown in Figure 2. Similar to our recently described single-objective BO scheme for identifying molecules with a desired oxidation potential [44], the preprocessing step (green) consists of feature generation for the entire candidate library followed by random selection of 10 BzNSN molecules. The molecular properties of interest, $E^{red}$, $G^{solv}$, and $\lambda^{abs}$, are computed using DFT simulations and used as initial dataset for training of the GPR models (red). Then, GPR-predicted properties and uncertainties of the remaining candidates

are used to calculate three sets of EI values, where each set corresponds to a property. The candidate(s) with the Pareto-optimal EI combination will be chosen for the next round of DFT simulations, and the cycle repeats. Perhaps, the main differences between single-objective and MBO lie in the training of different GPR models for different properties and the use of EI and Pareto front evaluation to determine subsequent training data. These specific components, *i.e.*, feature generation from SMILES strings, GPR model training, and candidate selection from multi-dimensional EI will be discussed next.

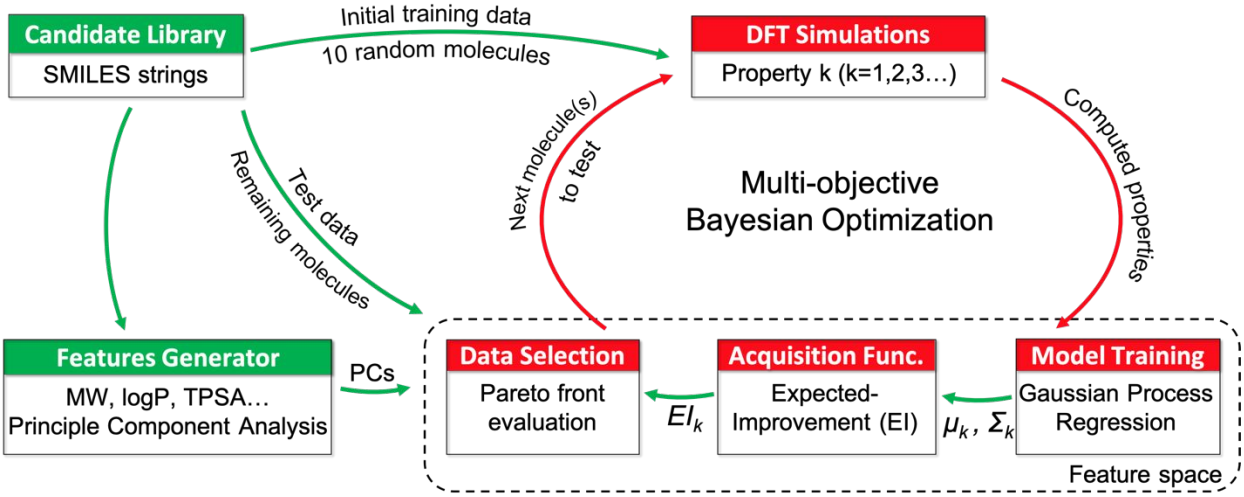



Figure 2: Active learning workflow for molecule discovery from Candidate Library (BzNSN molecules) via Multi-objective Bayesian Optimization (MBO). Here, the candidates are represented using SMILES strings. PCs, μ, Σ are principal components, GPR-predicted mean, and GPR-predicted uncertainty, respectively. The details of DFT simulation, acquisition functions and Gaussian Process Regression is given in Computational Details section.

**2.1 Feature generation:**

To build robust ML models for property prediction, it is critical to design features/fingerprints to numerically represent each molecule. It is also important to recognize that the generated features should uniquely represent the molecule and should be easy to compute for any new molecule. Here, we use the RDKit cheminformatics package [48] to generate such a set of 125 features using various physical and chemical properties of the molecules (see Table S1 of the supplementary information). Based on Pearson correlation analysis, we identified that no single feature among the original 125 features can reliably capture the trend of the computed $E^{red}$,$G^{solv}$, or $\lambda^{abs}$ (Figure S2, S3 and S4). After feature normalization, we perform principal component analysis (PCA) to reduce the dimensionality of the feature vector. From PCA, a total of 22 principal components (PCs) are found to be sufficient to account for 100% variance in the data (see Figure S5 of the supplementary information). The graphical illustration of the chemical space of 1400 BzNSN dataset is also shown in Figure S6. Thus, feature vectors consisting of 22 PCs are used as inputs for property predictions. We note that the feature generation protocol is consistent with our recent study [44].

**2.2 Gaussian process regression (GPR) models:**

Gaussian process regression (GPR) is used to train separate predictive model for each of the 3 properties in the computed dataset. The details of the GPR model are provided in the Computational Details section. The performance of the trained GPR models is evaluated using coefficient of determination ($R^2$) and root mean square error (RMSE) as the error metrics. The

entire data of 1400 BzNSN molecules are split into training and test sets. To determine the optimal training/test ratio, we examine learning curves that are generated by systematically increasing the size of the training set from 10% to 90% of the total dataset. The corresponding remaining data are used as test sets to evaluate the performance of the model. To generate statistically meaningful results, 100 evaluations are performed for each training set size and the reported test RMSE values are calculated as the average of those 100 runs.

The learning curves for the three properties ($E^{red}$, $G^{solv}$, and $\lambda^{abs}$) depicting the variation of average test RMSE as a function of training set size are shown in Figure 3 (a)-(c). The error bars denote 1 standard deviation (1σ) of the average RMSE values for the 100 runs. The plots include the learning curves for the model trained using all the original 125 features as well as reduced feature vector of 22 PCs. The GPR models trained using reduced feature vector of 22 PCs result in similar performance as compared to the models trained using all 125 features. This shows that PCA is an effective method for reducing the dimensionality of the feature vector without compromising the accuracy of the final model. It is evident from the plot that the average test RMSE decreases with increase in the training set size for all the 3 properties. The test RMSE reaches convergence at c.a. 70% training set size resulting in average test RMSE of 77 mV for reduction potential, 57 meV for solvation free energy and 15 nm for absorption wavelength. The parity plots (*i.e.*, GPR predicted property vs DFT computed property) comparing the performance of the final GPR models trained with 70% training data using 22 PCs as the feature vector are shown in Figure 3(d)-(f). The error bars in each parity plot represent the GPR uncertainty. The high $R^2$ coefficients of 0.94, 0.87 and 0.82 on the test set for reduction potential, solvation free energy and absorption wavelength, respectively, indicate good accuracy of the trained GPR models.

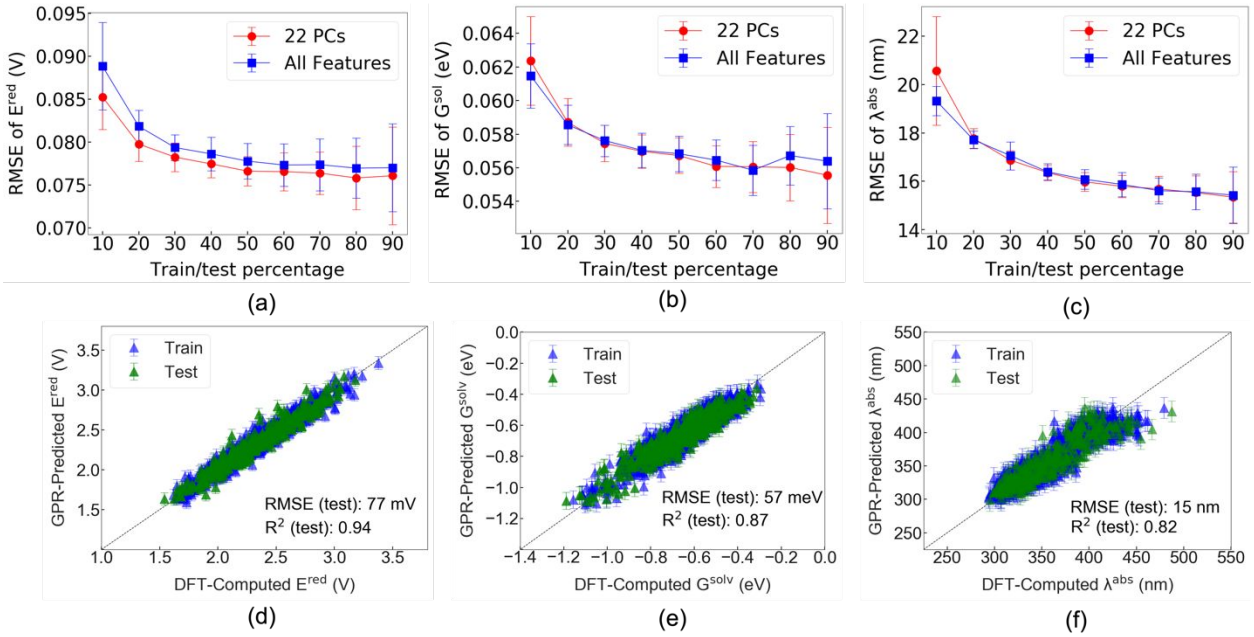



Figure 3: Learning curves of the GPR models of (a) reduction potential ($E^{red}$), (b) solvation free energy ($G^{solv}$), and (c) absorption wavelength ($\lambda^{abs}$) showing the root mean square error (RMSE) of the test set as a function of training data. The red and blue lines denote the learning curve for the GPR models trained using 22 principal components (PCs) and all 125 features generated using RDKit, respectively. The error bars denote the 1 standard deviation (1σ) of the average RMSE values for 100 runs. Parity plots showing the performance of final GPR models (using 22 PCs) of (d) reduction

potential ($E^{red}$), (e) solvation free energy ($G^{solv}$), and (f) absorption wavelength ($\lambda^{abs}$) trained using train/test ratio of 70/30%. The error bars denote GPR uncertainties.

### 2.3 Candidate selection using Pareto-optimal Expected Improvements (EI)

In single objective BO, the current optimal (*e.g.*, minimum, or maximum) value in the training set is used as the reference for computing the improvement metrics associated with candidates in the test data set. With EI formulation, the candidate with maximum EI value yields the maximum potential to improve over the existing reference and is selected for evaluation. In the case of MBO, however, the Pareto front in the training set dictates the choice of the reference value(s). All members of a Pareto front are considered equivalent and an improvement over any one of them will warrant a new Pareto front. Hence, the choice of which Pareto optimal point should be the reference value is a matter of strategic preference. In this work, as shown Figure 4, we choose the data point with maximum, non-infinite, crowding distance (see Computational Details) on the Pareto front of the training data set as the reference for EI evaluations (solid red circle in Figure 4(a)).

As the Pareto-optimal reference is identified, the improvement region over this point in the property space may be explored as indicated in Figure 4(b). Such reference point and GPR-predicted mean and uncertainty values of the test set are then used to compute EI values for every property. Thus, we obtain a multi-dimensional improvement metrics with the same number of dimensions as the property space (here, it is 3 dimensions). Therefore, finding a candidate with the optimal improvement over the reference point in all dimensions of the property space is now equivalent to picking one with the maximum multi-dimensional EI value. Then, the task is to identify Pareto-optimal points in the EI space. Since the number of possible Pareto-optimal candidates increase exponentially with respect to the number of EI dimensions and the size of data set, they should be strategically selected for evaluation for optimal computational efficiency. In single-point selection approach shown in Figure 4(c), we choose only the Pareto-optimal candidate with the maximum crowding distance for DFT evaluations. Although this approach increases the size of the DFT-evaluated/training data set slowly, it provides consistent improvement of the GPR models with limited computational resources. When concurrent DFT-evaluations of multiple Pareto-optimal candidates are feasible, a set number of candidates may be collected via Latin Hypercube Sampling (LHS) [57,58,59]. As shown in Figure 4(d), in the LHS approach, the selection of data points on the EI Pareto front is evenly spaced in all dimensions and hence diversified.

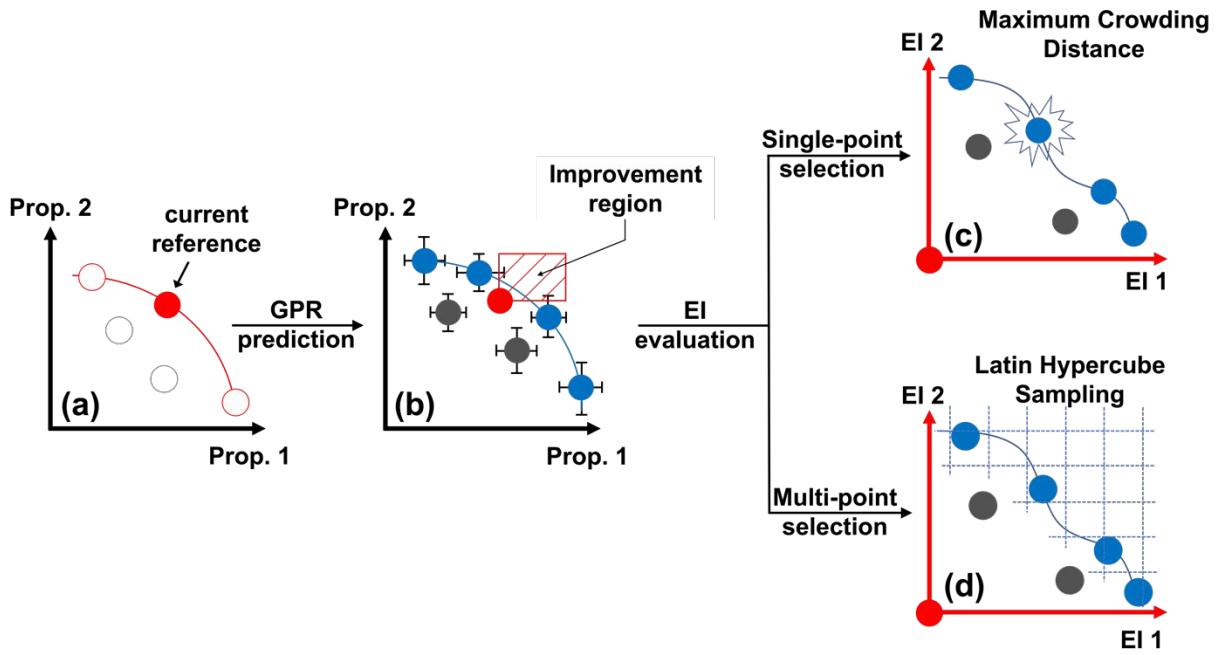



Figure 4: Candidate selection scheme via Expected Improvement (EI) and Pareto front evaluations of example property 1 and 2. (a) Selecting the Pareto-optimal reference in the training set for EI calculations. (b) Using Gaussian Process Regression to predict properties of the test set. The error bars representing the uncertainties of GPR predictions are not to scale. Due to uncertainty, the blue line is an approximation of the pareto front. (c) and (d) Computing EI values of the test set and suggesting the next candidate(s) for labeling. Details of GPR, EI and crowding distances are given in Computational Details section.

**3. Multi-objective BO performance on the 1400 BzNSN molecule dataset:**

As mentioned earlier, our multi-objective goal is to minimize the reduction potentials ($E^{red}$) and solvation free energies ($G^{solv}$) while targeting the desired absorption wavelength ($\lambda^{abs}$) of 375 nm. For the 1400 BzNSN dataset, there are 22 datapoints/molecules that form the true Pareto front, or the optimal solution set as shown by solid stars in Figure 5(a). The 2D chemical structures and the computed properties ($E^{red}$, $G^{solv}$, and $\lambda^{abs}$) of these Pareto-optimal molecules are summarized in Table 2 (IDs: 1-22). To evaluate the performance of our MBO approach, we perform 100 iterations (equivalent to 100 molecule properties evaluations) on the 1400 BzNSN dataset. After every iteration, the next molecule is suggested using the single-point selection method based on crowding distance as described in Figure 4(c). We also repeat the procedure 100 times, each using a different initial training set of 10 randomly selected data points, to obtain meaningful statistics.

A well-known metric for evaluating the performance of MBO methods is the hypervolume indicator, which is the volume enclosed by connecting the Pareto-optimal points to a chosen reference point in the multi-objective space [47, 60]. For computing the hypervolumes, we use a reference point consisting of 3 components, each of which is slightly larger than the extremum (maximum) in each property dimension, *i.e.*, $E^{red}_{ref} > E^{red}_{max}$ (3.38 V), $G^{solv}_{ref} > G^{solv}_{max}$ (-0.29 eV), $\lambda^{solv}_{ref} > \lambda^{solv}_{ref}$ (487 nm)[61]. As different solution sets/Pareto fronts yield different hypervolume values, they can be directly used for comparison. Typically, a higher hypervolume indicates a better set of Pareto optimal points, and it follows that the true Pareto front corresponds to the maximum hypervolume. For the 1400 BzNSN dataset, we can calculate the maximum hypervolume as the true Pareto front has been identified. Using maximum hypervolume as the target, the performance

of our MBO can be benchmarked against random selection. Figure 5(b) shows the hypervolume percentage (with respect to the maximum hypervolume) obtained by MBO and random selection as a function of the number of evaluations. Two observations can be made from Figure 5(b). First, based on shaded areas, which indicate 1 standard deviation around the mean, MBO provides more stable solution sets compared to random selection. Second, the Pareto-optimal molecules suggested by MBO reach the quality of the true Pareto front at a significantly faster pace than their randomly selected counterparts. Specifically, to achieve 99% of the maximum hypervolume, MBO only requires 74 molecule evaluations whereas random selection needs to investigate a total of 1126 molecules on average. Therefore, our MBO approach provides at least a 15-fold improvement in efficiency over random selection. The distribution of the number of Pareto-optimal molecules found in each of the 100 MBO runs is shown in Figure S7 of the supplementary information. Similarly, the statistics of the number of successful runs and the number evaluations required to find each of the 22 Pareto-optimal molecules are shown in Figure S8 and Table S2 of the supplementary information.

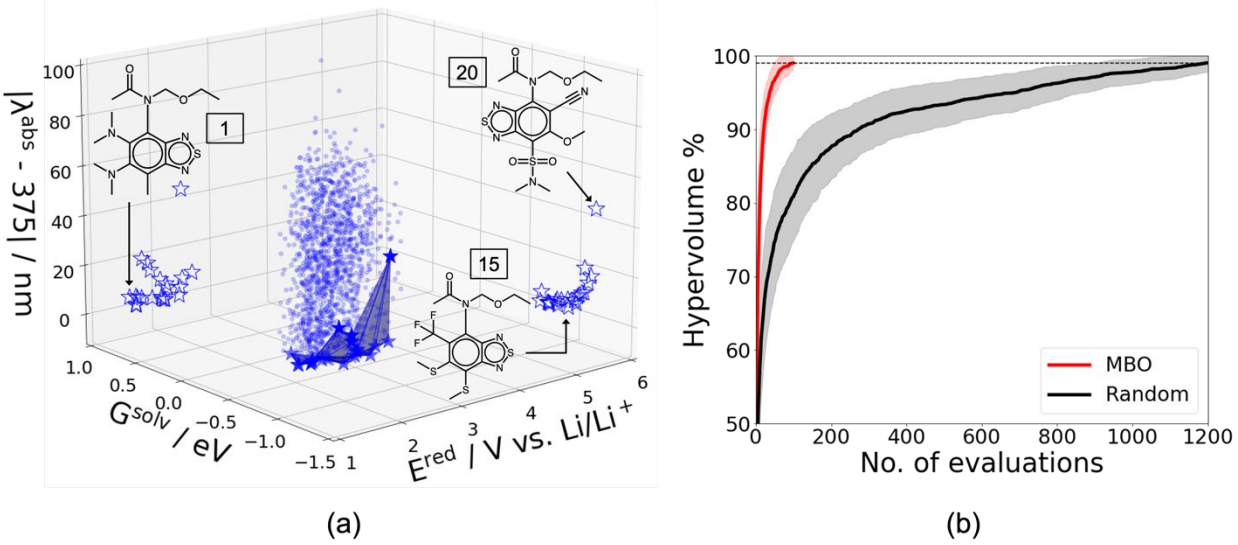
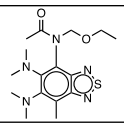
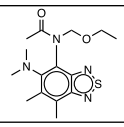
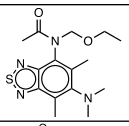
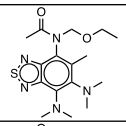
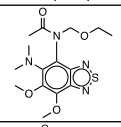
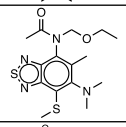
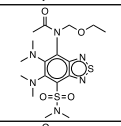
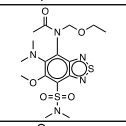
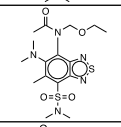
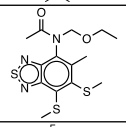
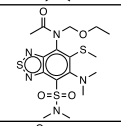
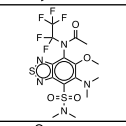
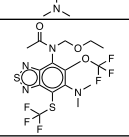
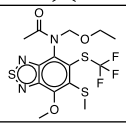
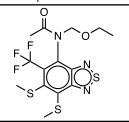
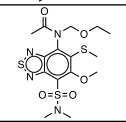
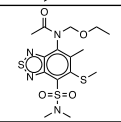
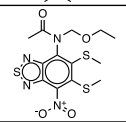
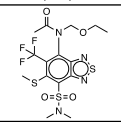
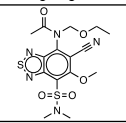
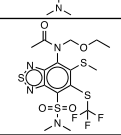
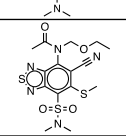



Figure 5: (a) Distribution of the DFT computed reduction potentials ($E^{red}$), solvation free energies ($G^{solv}$), and absorption wavelengths ($\lambda^{abs}$) of 1400 BzNSN molecule dataset (solid circles). The 22 Pareto-optimal datapoints are marked by solid stars, while their projected counterparts are shown as empty stars. The 2D molecular structures of the extrema w.r.t to individual properties are shown for reference. (b) Performance comparison between MBO and random selection using the hypervolume percentage w.r.t the maximum hypervolume. The solid lines and shaded area represent the means and 1 standard deviations ($1\sigma$), respectively. The black dashed line indicates 99% of the maximum hypervolume.

Table 2: List of 2D structures and computed reduction potentials ($E^{red}$), solvation free energies ($G^{solv}$), absorption wavelength ($\lambda^{abs}$) properties of the Pareto-optimal molecules found in the 1400 (1-22) BzNSN datasets.

| ID | Structure | $E^{red}$ (V vs $Li/Li^+$) | $G^{solv}$ (eV) | $\lambda^{abs}$ (nm) | ID | Structure | $E^{red}$ (V vs $Li/Li^+$) | $G^{solv}$ (eV) | $\lambda^{abs}$ (nm) |
|---|---|---|---|---|---|---|---|---|---|
| 1 |  | 1.54 | -0.94 | 379.17 | 2 |  | 1.64 | -0.68 | 371.15 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3 |  | 1.65 | -0.74 | 375.37 | 4 |  | 1.65 | -0.75 | 374.71 |
| 5 |  | 1.67 | -0.87 | 371.91 | 6 |  | 1.74 | -0.83 | 377.73 |
| 7 |  | 1.75 | -1.09 | 393.97 | 8 |  | 1.85 | -1.16 | 391.44 |
| 9 |  | 1.89 | -1.13 | 388.31 | 10 |  | 1.92 | -0.82 | 373.18 |
| 11 |  | 2.05 | -1.09 | 383.72 | 12 |  | 2.07 | -0.94 | 376.13 |
| 13 |  | 2.09 | -0.66 | 374.93 | 14 |  | 2.12 | -0.75 | 374.97 |
| 15 |  | 2.13 | -0.76 | 374.98 | 16 |  | 2.15 | -1.11 | 352.01 |
| 17 |  | 2.17 | -0.99 | 370.65 | 18 |  | 2.37 | -0.76 | 357.87 |
| 19 |  | 2.39 | -1.02 | 371.41 | 20 |  | 2.43 | -1.19 | 331.07 |
| 21 |  | 2.43 | -1.04 | 367.76 | 22 |  | 2.59 | -1.07 | 366.43 |

**4. Application of MBO on an unseen molecule dataset of 1 million BzNSNs:**

To demonstrate the robustness and generalizability of the MBO approach, we applied it to a significantly larger and new molecular dataset. This new dataset was generated by expanding the molecular candidate library (similar to Figure 1(b)) with more diverse substituents ('R' positions) in the molecular scaffold. The complete list of functional groups used for generation of this large molecular dataset is provided in Table S3 of the supplementary information. We limit the maximum number of heavy atoms (non-H atoms) in the molecules to 40 to maintain reasonable computational cost for DFT calculations. The final dataset consists of 1 million BzNSN molecules. Our goal is to discover new Pareto-optimal molecules that potentially outperform the existing 22 molecules (Table 2, IDs 1-22) with minimum number of DFT calculations. For this dataset, we employ the multi-point selection strategy described in Figure 4(d), in which 10 molecules are suggested for DFT evaluations in every MBO cycle, and a total of 10 cycles or 100 molecule property evaluations (DFT) are performed. Due to the higher diversity and complexity of BzNSN molecules in the new dataset, the evaluation of the 3 properties for each of the 100 new molecules

using DFT is considerably more expensive. Thus, the number of MBO cycles were limited to 10 to strike a balance between the computational cost and the discovery of new molecules with improved properties. The SMILES strings, computed reduction potentials, solvation free energies and absorption wavelengths of the molecules are given in Table S4.

In Figure 6(a), 100 MBO-evaluated molecules from the 1 million BzNSN dataset (red) are plotted together with the initial dataset of 1400 BzNSNs (blue). For ease of visualization and analysis, the 3 properties are also projected on individual dimension of $|\lambda^{abs} - 375|$, $G^{solv}$, and $E^{red}$ in Figure 6 (b), (c), and (d), respectively. To accurately determine if any of the 100 molecules suggested by MBO yields property improvement over the 1400 BzNSN dataset, we combine them together and re-evaluate the Pareto-optimal molecules. If a new Pareto front is identified, it is an indication that molecules with more desirable properties have been discovered. Figure 6 (a-d) shows the new Pareto front (red enclosed area) together with the existing one (blue enclosed area) for comparison. Specifically, the new Pareto front consists of 19 molecules, 16 of which (red stars) belong to the 100-molecule set suggested by our MBO from the 1 million BzNSN dataset. The 2D structures and properties of those 16 molecules are tabulated in Table S5 (BzNSN 23-38, see Supplementary Information). Among the new Pareto-optimal molecules, two new extrema are identified for $E^{red}$ and $G^{solv}$ in BzNSN 34 and 37, respectively, and both provide significant improvement over the previous desired limits (1.54 V → 1.30 V, -1.19 → -1.69 eV). For the third objective in which $\lambda^{abs}$ is targeted toward 375 nm, no improvement is found since the existing extremum is already near the desired value (BzNSN 15 with $\lambda^{abs}$=374.98). Importantly, we identify four MBO-suggested molecules from the 1 million BzNSN dataset, as shown in Figure 7 (BzNSN ID 26, 29, 33, and 35), that possess both lower $E^{red}$ and $G^{solv}$ values compared to the entire 1400 BzNSN dataset, while maintaining the $\lambda^{abs}$ in the desirable range of 350-400 nm ($|\lambda^{abs}-375|<=25$). These results indicate the high efficiency and robustness of our MBO approach for identifying redoxmers of multiple design criteria. However, it is important to discuss the current practicality of our MBO-suggested solutions, especially in the case of molecular design. As seen in Figure 7 and Table S5 in SI, many of the suggested Pareto-optimal BzNSN molecules are complex and therefore difficult to synthesize. To circumvent this problem and create a more seamless interaction with experiments, future implementation of MBO will consider synthesizability [62, 63] as an additional and necessary criterion. Although our MBO algorithm has been optimized to evaluate over $10^6$ datapoints with N>=2 objectives efficiently, one bottleneck remains is the speed of property evaluations via molecular simulations. Hence, the overall efficiency of our method also depends on the complexity of the materials and their properties of interest.

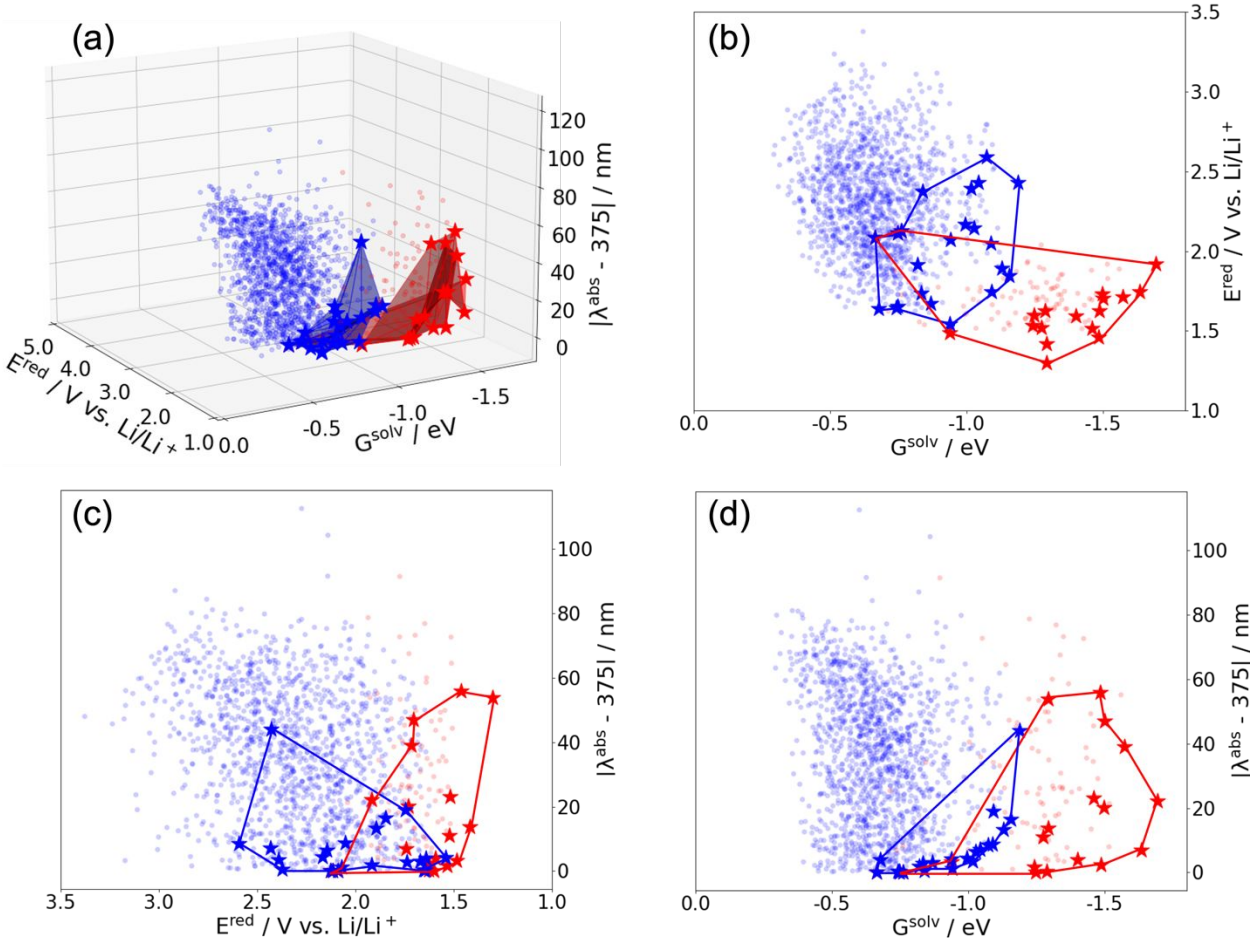



Figure 6: (a) Pareto front improvement over the known 1400 BzNSN dataset via multi-objective BO (MBO) application on the new 1 million BzNSN molecular dataset. Projection of datapoints on 2D surface of $G^{solv}$ and $E^{red}$ (b), $E^{red}$ and $|\lambda^{abs}$-375| (c), and $G^{solv}$ and $|\lambda^{abs}$-375| (d). All molecules in the 1400 BzNSN dataset and 100 MBO-suggested molecules from the 1 million BzNSN dataset are shown in blue and red circles, respectively. The Pareto-optimal molecules of the 1400 BzNSN dataset are shown as blue stars while the red stars are the updated Pareto front when 100 MBO-suggested molecules are added to the 1400 BzNSN dataset.

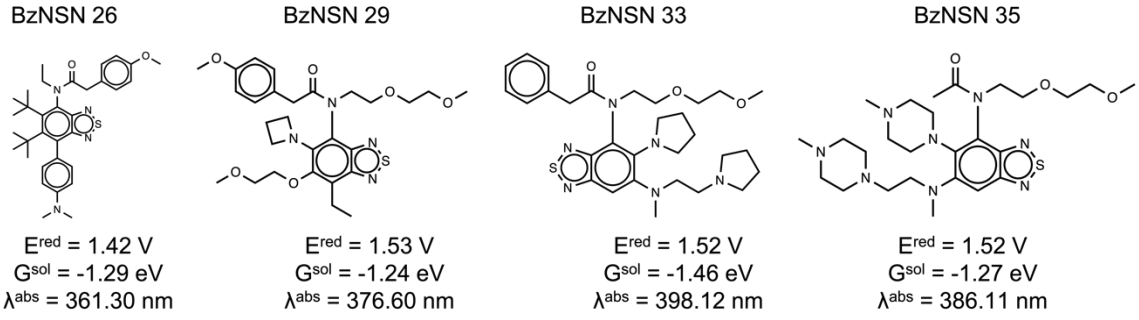



Figure 7: 2-D chemical structures and DFT computed properties of the 4 BzNSN molecules identified by MBO from the 1 million BzNSN molecular dataset which possess lower reduction potential ($E^{red}$) and solvation free energies ($G^{sol}$) compared to the entire 1400 BzNSN dataset with absorption wavelength ($\lambda^{abs}$) in the range of 350-400 nm. Sixteen new Pareto-optimal molecules (see Table S5 in SI for complete list of molecules) were identified within 10 MBO cycles from the expanded dataset of 1 million BzNSN molecules.

## Conclusion

Discovery of new and improved organic materials are essential for developing redox flow battery technology, and atomistic simulations and AI-based approaches can provide design guidelines to accelerate materials development. For non-aqueous redox flow battery (NRFB) technology, molecules must satisfy multiple properties such as redox windows, solubility, stability, rheology, conductivity, and other self-reporting/repairing properties. In this work, high-throughput DFT calculations were first carried out to evaluate the reduction potentials, solvation free energies, and absorption wavelengths of 1400 BzNSN molecules to search for the ideal anolyte redoxmers for NRFBs. Twenty-two Pareto-optimal BzNSN molecules that best compromise all three properties were identified and suggested for experimental validation. To accelerate the discovery of the Pareto-optimal candidates while keeping the computational cost minimal, an active learning model based on Multi-objective Bayesian Optimization (MBO) was developed and benchmarked on the 1400 BzNSN molecule dataset. The results indicate at least a 15-fold efficiency improvement over random selection in searching the optimal molecules. Finally, when applied to a new molecular dataset consisting of one million BzNSNs, our MBO quickly identified 16 new Pareto-optimal molecules with significant property improvement over the 1400 BzNSN dataset. Our robust and flexible multi-objective active learning approach provides an accelerated discovery framework for multi-criteria functional materials.

## Computational Details

*Density Functional Theory (DFT) Simulations:*

All DFT calculations were carried out using Gaussian 16 software [64] at wb97xd/6-31+G(d,p) [65,66] level of theory. The geometries of the molecules in the neutral and reduced charge states were optimized and frequency calculations were performed to compute the Gibbs free energies at 298 K. The SMD continuum model [67] with acetonitrile as the solvent medium was used to compute the solvation free energies. The reduction potentials ($E^{red}$, $Li/Li^+$) of the molecules were calculated using the change in Gibbs free energy in solution medium at 298 K upon addition of 1 $e^-$ to the neutral molecule ($\Delta G^{red} = G^{reduced} - G^{neutral}$) as given by equation 1.

$$E^{red} = \frac{-\Delta G^{red}}{nF} - 1.24\ V \quad (1)$$

Here, F is the Faraday constant (eV) and n is the number of electrons added to the neutral molecule (n = 1). The constant value of 1.24 V is subtracted to convert the change in Gibbs free energy to the reduction potential ($Li/Li^+$ reference electrode). More details about the calculation of redox potential can be found elsewhere [68,69,70,71].

The solvation free energies of the neutral molecules were calculated as the difference in the Gibbs free energies of the molecules in acetonitrile solvent medium ($G^{MeCN}$) and in gas-phase ($G^{gas}$) using equation 2.

$$G^{solv} = (G^{MeCN} - G^{gas}) \quad (2)$$

The absorption wavelengths of the molecules were calculated by performing single-point calculation at the optimized geometry of the neutral molecule using time-dependent DFT (TD-DFT) [72,73,74,75] as implemented in Gaussian 16.

*Machine Learning*

*Gaussian Process Regression (GPR) Models:*

The GPR models [76] with Matérn kernel were trained using Scikit-learn package [77]. Based on our benchmark of GPR predictions on $E^{red}$, the parameter ν that controls the smoothness of the approximated function was chosen to be 1.5 (Figure S9). The covariance function between the two molecules with feature vectors x, x' is given by equation 3

$$k(x, x') = \left(1 + \frac{\sqrt{3}\|x - x'\|}{\sigma_l}\right)\exp\left(-\frac{\sqrt{3}\|x - x'\|}{\sigma_l}\right) + \sigma_n^2 \quad (3)$$

Here, $\sigma_l$ and $\sigma_n$ are the length scale and the expected noise level in the dataset, respectively. Each parameter was determined using maximum likelihood estimate during model training.

*Expected Improvement (EI):*

The expected improvement (EI) acquisition function was independently calculated for each property as given by equation 4 [42,44]

$$EI(x) = \begin{cases} (\mu(x) - f(x^+))\Phi(Z) + \sigma(x)\phi(Z) & \sigma(x) > 0 \\ 0 & \sigma(x) = 0 \end{cases} \quad (4)$$

$$Z = \frac{\mu(x) - f(x^+) - \epsilon}{\sigma(x)} \quad (5)$$

Here, $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are the GPR predicted mean and uncertainties, $f(x^+)$ is the best property value in the current training set and $x^+$ is the feature vector of the material with best property value, $\Phi(Z)$ and $\varphi(Z)$ are the cumulative and probability density functions, respectively. The parameter $\epsilon$ in Equation 5 determines the amount exploration during optimization, and we used a constant value of 0.01 as it yields the optimal balance between exploration and exploitation in our dataset (Figure S10).

*Crowding distance assignment:*

The crowding distance assignment was originally developed for the NSGA-II algorithm and used to estimate the density of data points surrounding a particular point in the multi-objective solution space [78]. The crowding distance estimates the cuboid perimeter around a data point using its nearest neighbors as vertices. Hence, data points with larger crowding distances are more isolated. The crowding distance for each Pareto-optimal point i, is $d^i_{crowding}$, which was computed using the following pseudo code:

- Initialize $d^i_{crowding} = 0$ (i = 1, 2, …n Pareto-optimal points)
- for each objective/property m:
  - sort i based on its value in m ($y^i_m$)
  - if i is an extremum:
    - $d^i_{crowding}$ = infinity
  - else:
    - $d^i_{crowding}$ += $\frac{y^{i+1}_m - y^{i-1}_m}{y^{max}_m - y^{min}_m}$

## Associated Content

**Supporting Information:** Principal component analysis, analysis of 100 multi-objective Bayesian optimization (MBO) runs on the 1400 BzNSN dataset, SMILES representation and DFT computed properties of the 1400 BzNSN dataset, SMILES representation and DFT computed properties of the 100 MBO suggested molecules from the 1 million BzNSN dataset.

## Author Information

**Corresponding Author**

**Rajeev S. Assary –** *Materials Science Division, Argonne National Laboratory, Lemont, IL, USA, 60439*; orcid.org/0000-0002-9571-3307; *Phone: 630-252-3536; Email:* assary@anl.gov

**Author Contributions**

GA, HAD, and RSA conceived the idea and directed the research. GA and HAD performed DFT simulations and developed machine learning models. LZ and LAR provided details regarding molecular dataset and redoxmer chemistry. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

**Notes**

The authors declare no competing financial interest. The multi-objective Bayesian optimization code used in this work was developed as a part of our Chemistry Oriented Bayesian Optimization Library (COBOL), which can be found on GitHub at https://github.com/MolecularMaterials/COBOL/tree/main/case-study/multi-objective-BzNSN

**Authors**

**Garvit Agarwal** - *Materials Science Division, Argonne National Laboratory, Lemont, IL, USA, 60439*; orcid.org/0000-0002-1896-1024, Email: garvit.iitr@gmail.com

**Hieu A. Doan** - *Materials Science Division, Argonne National Laboratory, Lemont, IL, USA, 60439*; orcid.org/0000-0003-1460-9004, Email: hadoan@anl.gov

**Lily A. Robertson** - *Chemical Sciences & Engineering Division, Argonne National Laboratory, Lemont, IL, USA, 60439*; orcid.org/0000-0002-8784-0568, Email: robertla@anl.gov

**Lu Zhang** - *Chemical Sciences & Engineering Division, Argonne National Laboratory, Lemont, IL, USA, 60439*; orcid.org/0000-0003-0367-0862; Email: luzhang@anl.gov

**Acknowledgement**

This work was supported as part of the Joint Center for Energy Storage Research, an Energy Innovation Hub funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences. We gratefully acknowledge the computing resources provided on "Bebop", a 1024-node computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory. We also acknowledge the computational resources from Center for Nanoscale Materials, an Office of Science user facility, which was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

## References

1. Trahey, L.; Brushett, F. R.; Balsara, N. P.; Ceder, G.; Cheng, L.; Chiang, Y.-M.; Hahn, N. T.; Ingram, B. J.; Minteer, S. D.; Moore, J. S.; Mueller, K. T.; Nazar, L. F.; Persson, K. A.; Siegel, D. J.;

Xu, K.; Zavadil, K. R.; Srinivasan, V.; Crabtree, G. W., Energy storage emerging: A perspective from the Joint Center for Energy Storage Research. *Proceedings of the National Academy of Sciences* **2020,** 117, (23), 12550.

2. Gür, T. M., Review of electrical energy storage technologies, materials and systems: challenges and prospects for large-scale grid storage. *Energy & Environmental Science* **2018,** 11, (10), 2696-2767.

3. DeBruler, C.; Hu, B.; Moss, J.; Luo, J.; Liu, T. L., A Sulfonate-Functionalized Viologen Enabling Neutral Cation Exchange, Aqueous Organic Redox Flow Batteries toward Renewable Energy Storage. *ACS Energy Letters* **2018,** 3, (3), 663-668.

4. Chai, J.; Lashgari, A.; Cao, Z.; Williams, C. K.; Wang, X.; Dong, J.; Jiang, J. J., PEGylation-Enabled Extended Cyclability of a Non-aqueous Redox Flow Battery. *ACS Applied Materials & Interfaces* **2020,** 12, (13), 15262-15270.

5. Li, M.; Case, J.; Minteer, S. D., Bipolar Redox-Active Molecules in Non-Aqueous Organic Redox Flow Batteries: Status and Challenges. *ChemElectroChem* **2021,** 8, (7), 1215-1232.

6. Gong, K.; Fang, Q.; Gu, S.; Li, S. F. Y.; Yan, Y., Nonaqueous redox-flow batteries: organic solvents, supporting electrolytes, and redox pairs. *Energy & Environmental Science* **2015,** 8, (12), 3515-3530.

7. Li, M.; Rhodes, Z.; Cabrera-Pardo, J. R.; Minteer, S. D., Recent advancements in rational design of non-aqueous organic redox flow batteries. *Sustainable Energy & Fuels* **2020,** 4, (9), 4370-4389.

8. Duan, W.; Huang, J.; Kowalski, J. A.; Shkrob, I. A.; Vijayakumar, M.; Walter, E.; Pan, B.; Yang, Z.; Milshtein, J. D.; Li, B.; Liao, C.; Zhang, Z.; Wang, W.; Liu, J.; Moore, J. S.; Brushett, F. R.; Zhang, L.; Wei, X., "Wine-Dark Sea" in an Organic Flow Battery: Storing Negative Charge in 2,1,3-Benzothiadiazole Radicals Leads to Improved Cyclability. *ACS Energy Letters* **2017,** 2, (5), 1156-1161.

9. Huang, J.; Duan, W.; Zhang, J.; Shkrob, I. A.; Assary, R. S.; Pan, B.; Liao, C.; Zhang, Z.; Wei, X.; Zhang, L., Substituted thiadiazoles as energy-rich anolytes for nonaqueous redox flow cells. *Journal of Materials Chemistry A* **2018,** 6, (15), 6251-6254.

10. Zhang, J.; Huang, J.; Robertson, L. A.; Shkrob, I. A.; Zhang, L., Comparing calendar and cycle life stability of redox active organic molecules for nonaqueous redox flow batteries. *Journal of Power Sources* **2018,** 397, 214-222.

11. Robertson, L. A.; Shkrob, I. A.; Agarwal, G.; Zhao, Y.; Yu, Z.; Assary, R. S.; Cheng, L.; Moore, J. S.; Zhang, L., Fluorescence-Enabled Self-Reporting for Redox Flow Batteries. *ACS Energy Letters* **2020,** 5, (9), 3062-3068.

12. Cai, Y.; Xie, W.; Teng, Y. T.; Harikesh, P. C.; Ghosh, B.; Huck, P.; Persson, K. A.; Mathews, N.; Mhaisalkar, S. G.; Sherburne, M., High-throughput computational study of halide double perovskite inorganic compounds. *Chemistry of Materials* **2019,** 31, (15), 5392-5401.

13. Woods-Robinson, R.; Broberg, D.; Faghaninia, A.; Jain, A.; Dwaraknath, S. S.; Persson, K. A., Assessing high-throughput descriptors for prediction of transparent conductors. *Chemistry of Materials* **2018,** 30, (22), 8375-8389.

14. Li, S.; Xia, Y.; Amachraa, M.; Hung, N. T.; Wang, Z.; Ong, S. P.; Xie, R.-J., Data-driven discovery of full-visible-spectrum phosphor. *Chemistry of Materials* **2019,** 31, (16), 6286-6294.

15. Cheng, L.; Assary, R. S.; Qu, X.; Jain, A.; Ong, S. P.; Rajput, N. N.; Persson, K.; Curtiss, L. A., Accelerating electrolyte discovery for energy storage with high-throughput screening. *The journal of physical chemistry letters* **2015,** 6, (2), 283-291.
16. de la Cruz, C.; Molina, A.; Patil, N.; Ventosa, E.; Marcilla, R.; Mavrandonakis, A., New insights into phenazine-based organic redox flow batteries by using high-throughput DFT modelling. *Sustainable Energy & Fuels* **2020,** 4, (11), 5513-5521.
17. Davis, A. P.; Fry, A. J., Experimental and computed absolute redox potentials of polycyclic aromatic hydrocarbons are highly linearly correlated over a wide range of structures and potentials. *The Journal of Physical Chemistry A* **2010,** 114, (46), 12299-12304.
18. Han, Y.-K.; Jung, J.; Cho, J.-J.; Kim, H.-J., Determination of the oxidation potentials of organic benzene derivatives: theory and experiment. *Chemical physics letters* **2003,** 368, (5-6), 601-608.
19. Bachman, J. E.; Curtiss, L. A.; Assary, R. S., Investigation of the Redox Chemistry of Anthraquinone Derivatives Using Density Functional Theory. *The Journal of Physical Chemistry A* **2014,** 118, (38), 8852-8860.
20. Assary, R. S.; Brushett, F. R.; Curtiss, L. A., Reduction potential predictions of some aromatic nitrogen-containing molecules. *RSC Advances* **2014,** 4, (101), 57442-57451.
21. Pelzer, K. M.; Cheng, L.; Curtiss, L. A., Effects of functional groups in redox-active organic molecules: A high-throughput screening approach. *The Journal of Physical Chemistry C* **2017,** 121, (1), 237-245.
22. Kucharyson, J.; Cheng, L.; Tung, S.; Curtiss, L.; Thompson, L., Predicting the potentials, solubilities and stabilities of metal-acetylacetonates for non-aqueous redox flow batteries using density functional theory calculations. *Journal of Materials Chemistry A* **2017,** 5, (26), 13700-13709.
23. Patra, A.; Batra, R.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Ramprasad, R., A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. *Computational Materials Science* **2020,** 172, 109286.
24. Pilania, G.; Gubernatis, J. E.; Lookman, T., Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Computational Materials Science* **2017,** 129, 156-163.
25. Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I., Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Physical Review B* **2016,** 93, (11), 115104.
26. Chen, L.; Tran, H.; Batra, R.; Kim, C.; Ramprasad, R., Machine learning models for the lattice thermal conductivity prediction of inorganic materials. *Computational Materials Science* **2019,** 170, 109155.
27. Chen, L.; Kim, C.; Batra, R.; Lightstone, J. P.; Wu, C.; Li, Z.; Deshmukh, A. A.; Wang, Y.; Tran, H. D.; Vashishta, P.; Sotzing, G. A.; Cao, Y.; Ramprasad, R., Frequency-dependent dielectric constant prediction of polymers using machine learning. *npj Computational Materials* **2020,** 6, (1), 61.
28. Lightstone, J. P.; Chen, L.; Kim, C.; Batra, R.; Ramprasad, R., Refractive index prediction models for polymers using machine learning. *Journal of Applied Physics* **2020,** 127, (21), 215105.
29. Talapatra, A.; Uberuaga, B. P.; Stanek, C. R.; Pilania, G., A Machine Learning Approach for the Prediction of Formability and Thermodynamic Stability of Single and Double Perovskite Oxides. *Chemistry of Materials* **2021,** 33, (3), 845-858.

30. Gu, T.; Lu, W.; Bao, X.; Chen, N., Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors. *Solid State Sciences* **2006,** 8, (2), 129-136.
31. Sharma, V.; Kumar, P.; Dev, P.; Pilania, G., Machine learning substitutional defect formation energies in ABO3 perovskites. *Journal of Applied Physics* **2020,** 128, (3), 034902.
32. Mannodi-Kanakkithodi, A.; Toriyama, M. Y.; Sen, F. G.; Davis, M. J.; Klie, R. F.; Chan, M. K. Y., Machine-learned impurity level prediction for semiconductors: the example of Cd-based chalcogenides. *npj Computational Materials* **2020,** 6, (1), 39.
33. Pilania, G., Machine learning in materials science: From explainable predictions to autonomous design. *Computational Materials Science* **2021,** 193, 110360.
34. Sutton, C.; Boley, M.; Ghiringhelli, L. M.; Rupp, M.; Vreeken, J.; Scheffler, M., Identifying domains of applicability of machine learning models for materials science. *Nature Communications* **2020,** 11, (1), 4428.
35. Lookman, T.; Balachandran, P. V.; Xue, D.; Yuan, R., Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials* **2019,** 5, (1), 21.
36. Sivaraman, G.; Krishnamoorthy, A. N.; Baur, M.; Holm, C.; Stan, M.; Csányi, G.; Benmore, C.; Vázquez-Mayagoitia, Á., Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Computational Materials* **2020,** 6, (1), 104.
37. Sivaraman, G.; Guo, J.; Ward, L.; Hoyt, N.; Williamson, M.; Foster, I.; Benmore, C.; Jackson, N., Automated Development of Molten Salt Machine Learning Potentials: Application to LiCl. *The Journal of Physical Chemistry Letters* **2021**, 4278-4285.
38. Wen, C.; Zhang, Y.; Wang, C.; Xue, D.; Bai, Y.; Antonov, S.; Dai, L.; Lookman, T.; Su, Y., Machine learning assisted design of high entropy alloys with desired property. *Acta Materialia* **2019,** 170, 109-117.
39. Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E., Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics* **2018,** 148, (24), 241733.
40. Xue, D.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T., Accelerated search for materials with targeted properties by adaptive design. *Nature Communications* **2016,** 7, (1), 11241.
41. Xue, D.; Balachandran, P. V.; Yuan, R.; Hu, T.; Qian, X.; Dougherty, E. R.; Lookman, T., Accelerated search for BaTiO<sub>3</sub>-based piezoelectrics with vertical morphotropic phase boundary using Bayesian learning. *Proceedings of the National Academy of Sciences* **2016,** 113, (47), 13301-13306.
42. Bassman, L.; Rajak, P.; Kalia, R. K.; Nakano, A.; Sha, F.; Sun, J.; Singh, D. J.; Aykol, M.; Huck, P.; Persson, K.; Vashishta, P., Active learning for accelerated design of layered materials. *npj Computational Materials* **2018,** 4, (1), 74.
43. Kim, C.; Chandrasekaran, A.; Jha, A.; Ramprasad, R., Active-learning and materials design: the example of high glass transition temperature polymers. *MRS Communications* **2019,** 9, (3), 860-866.
44. Doan, H. A.; Agarwal, G.; Qian, H.; Counihan, M. J.; Rodríguez-López, J.; Moore, J. S.; Assary, R. S., Quantum Chemistry-Informed Active Learning to Accelerate the Design and Discovery of Sustainable Energy Storage Materials. *Chemistry of Materials* **2020,** 32, (15), 6338-6346.

45. Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J., Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Central Science* **2020,** 6, (4), 513-524.
46. Gopakumar, A. M.; Balachandran, P. V.; Xue, D.; Gubernatis, J. E.; Lookman, T., Multi-objective Optimization for Materials Discovery via Adaptive Design. *Scientific Reports* **2018,** 8, (1), 3738.
47. Jablonka, K. M.; Jothiappan, G. M.; Wang, S.; Smit, B.; Yoo, B., Bias free multiobjective active learning for materials design and discovery. *Nature Communications* **2021,** 12, (1), 2312.
48. Landrum, G., Rdkit: Open-source cheminformatics software. *GitHub and SourceForge* **2016,** 10, 3592822.
49. Knowles, J., ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation* **2006,** 10, (1), 50-66.
50. Liu, W.; Zhang, Q.; Tsang, E.; Liu, C.; Virginas, B. In *On the Performance of Metamodel Assisted MOEA/D*, Berlin, Heidelberg, 2007; Springer Berlin Heidelberg: Berlin, Heidelberg, 2007; pp 547-557.
51. Zhang, Q.; Liu, W.; Tsang, E.; Virginas, B., Expensive Multiobjective Optimization by MOEA/D With Gaussian Process Model. *IEEE Transactions on Evolutionary Computation* **2010,** 14, (3), 456-474.
52. Häse, F.; Roch, L. M.; Aspuru-Guzik, A., Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chemical Science* **2018,** 9, (39), 7642-7655.
53. Keane, A. J., Statistical Improvement Criteria for Use in Multiobjective Design Optimization. *AIAA Journal* **2006,** 44, (4), 879-891.
54. Ponweiser, W.; Wagner, T.; Biermann, D.; Vincze, M. In *Multiobjective Optimization on a Limited Budget of Evaluations Using Model-Assisted $\mathcal{S}$-Metric Selection*, Berlin, Heidelberg, 2008; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 784-794.
55. Emmerich, M. T.; Giannakoglou, K. C.; Naujoks, B., Single-and multiobjective evolutionary optimization assisted by Gaussian random field metamodels. *IEEE Transactions on Evolutionary Computation* **2006,** 10, (4), 421-439.
56. Jeong, S.; Obayashi, S. In *Efficient global optimization (EGO) for multi-objective problem and data mining*, 2005 IEEE Congress on Evolutionary Computation, 2-5 Sept. 2005, 2005; 2005; pp 2138-2145 Vol. 3.
57. Viana, F. A. C., A Tutorial on Latin Hypercube Design of Experiments. *Quality and Reliability Engineering International* **2016,** 32, (5), 1975-1985.
58. Helton, J. C.; Davis, F. J., Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety* **2003,** 81, (1), 23-69.
59. Paulson, N. H.; Libera, J. A.; Stan, M., Flame spray pyrolysis optimization via statistics and machine learning. *Materials & Design* **2020,** 196, 108972.
60. Zitzler, E.; Thiele, L. In *Multiobjective optimization using evolutionary algorithms — A comparative case study*, Berlin, Heidelberg, 1998; Springer Berlin Heidelberg: Berlin, Heidelberg, 1998; pp 292-301.
61. Sambo, F.; Borrotti, M.; Mylona, K., A coordinate-exchange two-phase local search algorithm for the D- and I-optimal designs of split-plot experiments. *Computational Statistics & Data Analysis* **2014,** 71, 1193-1207.

62. Ertl, P.; Schuffenhauer, A., Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* **2009,** 1, (1), 8.

63. Gao, W.; Coley, C. W., The Synthesizability of Molecules Proposed by Generative Models. *Journal of Chemical Information and Modeling* **2020,** 60, (12), 5714-5723.

64. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.

65. Chai, J.-D.; Head-Gordon, M., Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Physical Chemistry Chemical Physics* **2008,** 10, (44), 6615-6620.

66. Rassolov, V. A.; Ratner, M. A.; Pople, J. A.; Redfern, P. C.; Curtiss, L. A., 6-31G* basis set for third-row atoms. *Journal of Computational Chemistry* **2001,** 22, (9), 976-984.

67. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B* **2009,** 113, (18), 6378-6396.

68. Agarwal, G.; Doan, H. A.; Assary, R. S., Molecular Structure and Electron Affinity of Metal-Solvent Complexes: Insights from Density Functional Theory Simulations. *Journal of The Electrochemical Society* **2020,** 167, (10), 100545.

69. Borodin, O.; Behl, W.; Jow, T. R., Oxidative Stability and Initial Decomposition Reactions of Carbonate, Sulfone, and Alkyl Phosphate-Based Electrolytes. *The Journal of Physical Chemistry C* **2013,** 117, (17), 8661-8682.

70. Kelly, C. P.; Cramer, C. J.; Truhlar, D. G., Single-Ion Solvation Free Energies and the Normal Hydrogen Electrode Potential in Methanol, Acetonitrile, and Dimethyl Sulfoxide. *The Journal of Physical Chemistry B* **2007,** 111, (2), 408-422.

71. Vollmer, J. M.; Curtiss, L. A.; Vissers, D. R.; Amine, K., Reduction Mechanisms of Ethylene, Propylene, and Vinylethylene Carbonates: A Quantum Chemical Study. *Journal of The Electrochemical Society* **2004,** 151, (1), A178-A183.

72. Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J., An efficient implementation of time-dependent density-functional theory for the calculation of excitation energies of large molecules. *The Journal of Chemical Physics* **1998,** 109, (19), 8218-8224.

73. Casida, M. E.; Jamorski, C.; Casida, K. C.; Salahub, D. R., Molecular excitation energies to high-lying bound states from time-dependent density-functional response theory: Characterization and correction of the time-dependent local density approximation ionization threshold. *The Journal of Chemical Physics* **1998,** 108, (11), 4439-4449.

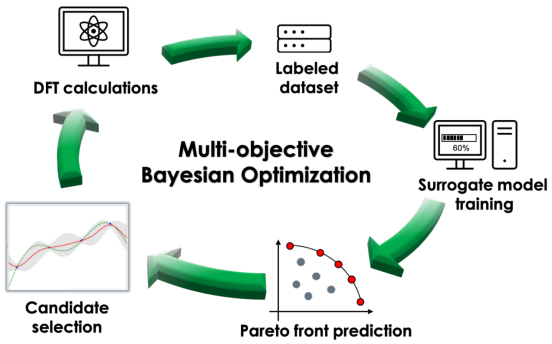
74. Bauernschmitt, R.; Ahlrichs, R., Treatment of electronic excitations within the adiabatic approximation of time dependent density functional theory. *Chemical Physics Letters* **1996,** 256, (4), 454-464.

75. Furche, F.; Ahlrichs, R., Adiabatic time-dependent density functional methods for excited state properties. *The Journal of Chemical Physics* **2002,** 117, (16), 7433-7447.

76. Rasmussen, C. E.; Williams, C., Gaussian processes for machine learning the MIT press. *Cambridge, MA* **2006**.

77. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011,** 12, 2825-2830.

78. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T., A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **2002,** 6, (2), 182-197.


DFT calculations
Labeled dataset
Multi-objective Bayesian Optimization
Surrogate model training
Candidate selection
Pareto front prediction