The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

Designing building blocks of covalent organic frameworks through on-the-fly batch-based Bayesian optimization

Cite as: J. Chem. Phys. 161, 074102 (2024); doi: 10.1063/5.0223540
Submitted: 15 June 2024 • Accepted: 30 July 2024 •
Published Online: 15 August 2024

View Online Export Citation CrossMark

Yuxuan Yao[1,2] and Harald Oberhofer[2,a)]

AFFILIATIONS

[1] Department of Chemistry, TUM School of Natural Sciences, Technical University Munich, Lichtenbergstr. 4, 85748 Garching b. München, Germany
[2] Chair for Theoretical Physics VII and Bavarian Center for Battery Technology, University of Bayreuth, Universitätsstr. 30, D-95447 Bayreuth, Germany

[a)] **Author to whom correspondence should be addressed:** harald.oberhofer@uni-bayreuth.de

ABSTRACT

In this work, we use a Bayesian optimization (BO) algorithm to sample the space of covalent organic framework (COF) components aimed at the design of COFs with a high hole conductivity. COFs are crystalline, often porous coordination polymers, where organic molecular units—called building blocks (BBs)—are connected by covalent bonds. Even though we limit ourselves here to a space of three-fold symmetric BBs forming two-dimensional COF sheets, their design space is still much too large to be sampled by traditional means through evaluating the properties of each element in this space from first principles. In order to ensure valid BBs, we use a molecular generation algorithm that, by construction, leads to rigid three-fold symmetric molecules. The BO approach then trains two distinct surrogate models for two conductivity properties, level alignment vs a reference electrode and reorganization free energy, which are combined in a fitness function as the objective that evaluates BBs' conductivities. These continuously improving surrogates allow the prediction of a material's properties at a low computational cost. It thus allows us to select promising candidates which, together with candidates that are very different from the molecules already sampled, form the updated training sets of the surrogate models. In the course of 20 such training steps, we find a number of promising candidates, some being only variations on already known motifs and others being completely novel. Finally, we subject the six best such candidates to a computational reverse synthesis analysis to gauge their real-world synthesizability.

© 2024 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1063/5.0223540

10 February 2025 08:37:19

I. INTRODUCTION

Covalent organic frameworks (COFs) are a class of crystalline porous polymers formed by organic molecular building units, called building blocks (BBs), connected through covalent bonds.[1,2] Since their first inception,[3] there has been a growing number of COF applications, such as electrochemical energy storage[4–6] and opto-electronic devices.[7] There, highly conductive COFs allow for rapid charge transfer between catalysts in catalysis and in charge/discharge processes in energy storage devices.[8–11] On top of that, the versatility of the COF platform might allow their use as an active component in photovoltaic applications.[12] In all of these applications, one of the major criteria for the applicability of a material is the degree of charge carrier mobility they exhibit.[13]

While, in principle, COFs can be synthesized as one, two, or three dimensional materials, it is the class of two dimensional (possibly layered) sheets that has received the most attention in the literature.[14] In order to form an area-filling two-dimensional COF sheet, the building blocks need to exhibit at least a three-fold rotational symmetry. Compared to the space of all possible organic molecules,[15] a fixed molecular symmetry obviously limits the choice of possible BB candidates, yet still leaves a huge number of molecules to be considered. In turn, this gives rise to a large variety of chemical and physical properties of the resulting frameworks, explaining

© Author(s) 2024

the versatility of this material class. Unfortunately, such a variety also hampers targeted efforts to design COFs with specific properties through the sheer number of molecules in that design space. To exhaustively sample these massive spaces, modern approaches tend to apply surrogate machine learning models of the properties to be optimized. Thereby, efforts are mostly focused on materials' stabilities and structural properties,[16] or factors contributing to the transport of charge carriers.[17,18] Regarding COFs, machine learning (ML) approaches are mostly focused on the optimization of stability, thermal properties, or band edges.[19–23] Active machine learning (AML) and Bayesian optimization (BO) are successfully applied to the prediction, design, and engineering of porous materials.[24–26] Following our earlier work on molecular organic semiconductors,[27] we here modify an AML algorithm to efficiently sample the design space of building blocks for potential COF semiconductors.[28–31] Specifically, we focus on two properties associated with charge transport,[13] the reorganization energy—correlated with the barrier toward charge transport—and the electronic level alignment—measuring potential barriers to charge injection and extraction. In order to reduce search complexity somewhat, in this work we focus on three-fold rotationally symmetric BBs able to form hexagonal (honeycomb) networks. First, we focus on developing a morphing algorithm that generates rotationally symmetric molecules. We separate symmetric BBs into sub-units that we call bricks, which are then subjected to molecular morphing steps and replicated three times to form the full BBs. To efficiently sample this ever-growing brick space, we formulate a batch-based full-exploitation BO scheme based on the ground procedures of our original AML approach.[27] The surrogate models employed in the BO scheme are based on Gaussian progress regression (GPR),[32] which we demonstrated to have adequate predictive capabilities for the two properties we concentrate on.[27] After 20 search steps, we find in total 1736 promising BBs out of 2160 candidates that are subjected to computationally intensive electronic structure calculations.

II. METHODS

Sampling a molecular design space with BO search essentially consists of two distinct steps. First, the generation step, where new molecules can automatically be spawned from old ones, e.g., through alchemical modification or "morphing" of their constituents. Second, the evaluation step, where the properties of the newly created molecules are predicted by surrogate models or computed from first principles. As mentioned earlier, we here focus on COF building blocks. These not only need a well-defined three-fold rotational symmetry but should also be stiff enough to preserve the pores' shapes when the area-filling network is synthesized. Compared to our original approach,[27] we thus modify the morphing algorithm to fulfill these requirements. In the following, we outline our new morphing algorithm, the conductivity properties we employed in the BO search and, finally, the BO search strategy and its hyperparameters.

A. Morphing algorithm

Similarly to the original approach, we generate three-fold rotationally symmetric BBs by repeat application of simple molecular morphing operations to the smallest aromatic BB, a single benzene molecule. These operations, 61 in total and listed in Table S1 of the supplementary material, are split into two kinds, namely two backbone operations—ensuring the stiffness of the resulting BB—and 59 side group operations, including 1 annelation reaction, 1 substitution reaction, and 57 side group additions, extracted from molecules found in the Cambridge Structural Database.[33,34] Backbone operations are the addition of aromatic units to the three fold symmetric stiff scaffold, while side group operations are annelation, substitution, and side group addition reactions, all consisting of the most common organic elements C, H, N, and O. To generate the three-fold rotationally symmetric BBs, we first produce molecules, named bricks, without any specifically defined symmetry and attach them to a central benzene. Two further bricks are attached to the benzene's meta positions—in the case of single atom-capped bricks—or ortho-fused to the central benzene—in the case of ring-capped bricks. In each case, the two additional bricks are rotated by $\pm 120°$, respectively. Similar to our earlier efforts,[27] bricks themselves are the products of mutation reactions encoded as "Reaction SMARTS"[35,36] that represent pathways from a reactant molecule. Repeat application of these mutations allows us to successively enlarge the design space of potential COF bricks.

All such cheminformatics-related tasks were carried out using RDKit,[37] if not stated otherwise. The detailed morphing algorithm is illustrated in Fig. 1 and outlined below. In both cases, we use phenyl addition and hydroxyl addition as examples for backbone and side group operations, respectively.

1. Starting from a benzene molecule, two kinds of operations can be performed. Using the two operations mentioned earlier produces a biphenyl molecule and a phenol molecule as our new bricks. Note that ring addition backbone operations such as the phenyl addition only occur at backbone sites, defined as the atom farthest from the core benzene. This way, we ensure that the ring addition only happens along the axis of symmetry. In addition, the side group operations occur at sites except backbone sites to preserve the backbone sites for further necessary connections.
2. The such generated bricks are then connected to a benzene core. Thereby, the connection between core benzene and bricks can be a single bond or a triple bond; both of them are rigid bonds. In addition, the connection can also be two rings fused together to form a final BBs central triphenylene.
3. Rotate the brick around the core benzene once by $120°$ and once by $-120°$ to finally create a three-fold rotationally symmetric molecule.

Repeating these steps for all morphing operations, we obtain a generation of potential COF BBs. Subsequent generations are produced by repeating steps (1)–(3) starting from the bricks of previous generations. Given that some combinations of operations may lead to duplicates, we prune the ensemble of such generated BB candidates, called the molecular space, after steps 1 and 2. Note that, at this stage, generated molecules are still only 2D graphs and need further processing to yield 3D structures. We first relax them with very low accuracy density functional tight-binding (DFTB) optimization calculations to obtain structural xyz-coordinates rapidly, and then coordinates are transformed to SMILES by Open Babel.[38] After this structural check, only 15%–25% BBs survive as coordinates with physically plausible structural integrity. Subsequently,

10 February 2025 08:37:19

© Author(s) 2024

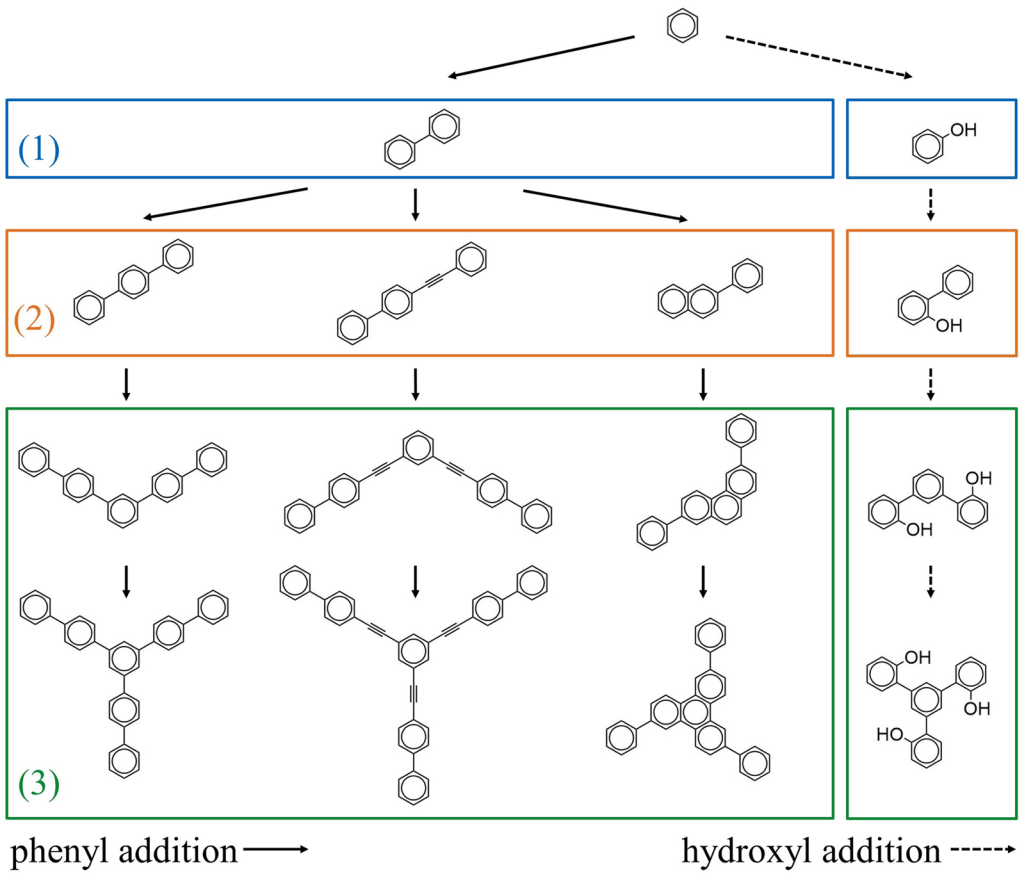

FIG. 1. Morphing algorithm of backbone (left) and side group (right) reactions starting from a simple benzene unit. Three connection ways are shown, taking two backbone and one annelation reaction as examples in step (2). Here, phenyl(solid lines) and hydroxyl(dash lines) additions were taken as examples to represent backbone and side group operations, respectively.

each candidate in the generated BBs space is labeled with ML model predicted properties. This way, we could generate an *in principle* infinite molecular space with any-fold rotational symmetry simply by changing the rotation step. In this work, specifically, we do not perform a search in a fully enumerated candidate pool but rather sample a space restricted only by symmetry and stability considerations, discussed earlier.

B. Conductivity properties

For reasons of computational efficiency and practicality, van der Waals (vdW) corrected density functional tight-binding (DFTB)[39] is used for the calculation of properties throughout the present work. Yet, recognizing the limited accuracy of such an approach, especially with regards to electronic level alignments, we aligned all DFTB results to vdW-corrected hybrid DFT[40,41] levels of accuracy using a linear scaling scheme.[27] DFTB calculations of properties were conducted with the xTB (v6.2.3) code based on GFN1-xTB theory.[42,43] Initial 3D coordinates of the BBs were created from 2D molecular graphs using the computationally cheaper dispersion corrected (DC)-self consistent charge (SCC) DFTB method implemented in the DFTB+ package.[44] These rough DFTB relaxations also guarantee the physical plausibility of generated structures, considering that structures with overlapping atoms or other unreasonable configurations would simply lead to the immediate abortion of these relaxations. All candidates surviving this step were relaxed further by the xTB internal approximate normal coordinate rational function optimizer (ANCopt) at the default geometry convergence criterion. Afterward, properties related to the conductivity of a material formed by these BBs are calculated on the resulting xTB relaxed geometries.

As in our earlier work, these properties are the highest occupied molecular orbital (HOMO) energy $\varepsilon_{\mathrm{HOMO}}$ and the reorganization energy.[13,34] The former measures the barrier to charge injection from a standard gold electrode[45–47] as $\varepsilon_{\mathrm{align}} = |\varepsilon_{\mathrm{HOMO}} - \Phi_{\mathrm{Au}}|$,[48] the alignment vs the electrode's Fermi level, here taken as $\Phi_{\mathrm{Au}} = -5.1$ eV.[49] In an ideal semiconductor, the energy level of the hole matches the electrode's Fermi level. In addition, note that

10 February 2025 08:37:19

© Author(s) 2024

hereby we assume our resulting COF to be of p-type semiconductor character.[50] The intra-molecular (hole) reorganization energy λ_h, which is computed via the popular four-point scheme,[51] measures the energy cost of accommodating a new charge state after the carrier has moved to another site.[52] In a solid, λ_h is connected to electron–phonon coupling[13] and, therefore, should be minimized for efficient transport. To evaluate the candidates' performance and optimize the selection of BB candidates, these two properties $\varepsilon_{\mathrm{HOMO}}$ and λ_h are combined in the scalar fitness function,

$$F = -\left\| \begin{pmatrix} \lambda_h \\ \varepsilon_{align} \end{pmatrix} \cdot \mathbf{w} \right\|_2 . \tag{1}$$

Here, a weight vector $\mathbf{w} = (1.0, 0.7)^T$ is employed to balance the scales between two properties. Comparing to known organic semiconducting materials, the values of $\mathbf{w}$ are chosen such that if λ_h falls into a favorable range, the value of 0.7 scaling ε_{align} will yield an Ohmic alignment with the electrode of $|\varepsilon_{align} < 0.3\ \mathrm{eV}|$. In an earlier study,[27] we found this scalar fitness function to yield values of $F \geq -0.2$ for promising candidates. Consequently, we use this threshold to probe the discovery success of our AML search strategy.[53] Ideal candidates are expected to maximize the value of F.

C. BO search strategy and hyperparameter optimization

In this work, batch-based BO is utilized to create ever-improving surrogate models of the conductivity properties based on the GPR algorithm.[32] Note that we employ two distinct GPR models, one for each property, to account for the fact that the prediction of $\varepsilon_{\mathrm{align}}$ might be more accurate than that of λ or vice versa. The GPR kernel function based on the MinMax kernel[54] closely related to the Tanimoto kernel[37] is illustrated in Sec. II of the supplementary material. In both models, the molecular BB candidates are represented by Morgan fingerprints[55] extracted up to a radius of five bonds around each atom. The input dataset into the models $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$, $\mathbf{x_N}$ is the Morgan Fingerprints vector. In our previous work, during each learning step, in order to balance the exploitation and exploration of the surrogate model, candidates were elected based on the upper confidence bound (UCB) acquisition function[56] F_{acq},

$$F_{\mathrm{acq}} = F + \kappa \cdot \sigma, \tag{2}$$

where F is the fitness function of candidates defined earlier and σ denotes the prediction uncertainty obtained from the standard deviation of the GP.[57] The free parameter κ thus balances exploitation and exploration,[58] where larger κ implies that higher uncertainty is preferred and more distinct candidates are chosen.[56,59] In this work, we aim to optimally explore the unlimited COF BB molecular space in a limited number of learning steps and optimize the hyperparameter κ accordingly. In our earlier work,[27] $\kappa > 0$ was found to be optimal, resulting in an active machine learning scheme. As discussed in more detail below, here we find $\kappa = 0$ to be more favorable. Our search thus reduced to a Bayesian optimization approach focused on exploitation.

As a quantitative benchmark to measure the performance of the BO strategy, we employ the discovery success S(N),[27] which is the ratio of well-performing molecules with $F \geq -0.2$ divided by the full population of molecules queried with DFT-corrected xTB, cf. Sec. II B. The surrogate model is initialized from a set of $N_{\mathrm{initial}} = 160$ molecules, which are BBs formed from 62 bricks, generated by applying 61 morphing operations and the original benzene, and used to predict the fitness of each candidate in the current generation. As the workflow depicted in Fig. 2 shows, at each learning step, we choose N_{mutation} bricks from the current population to yield a BB space. Therefore, elements of the search space are predicted by pre-trained GPR models of the two properties. Of these, N_{batch} molecules with a predicted fitness of $F \geq -0.2$ are queried and labeled by fist-principles calculated properties as added training data.

We further realize a reinforcement learning strategy[27,60] with tree search in the brick space extension. At every learning step, $N_{\mathrm{mutation\ batch}}$ bricks are selected to undergo morphing reactions and generate a new layer of brick space. From this freshly generated layer, $N_{\mathrm{mutation\ batch}}$ bricks are selected thereafter to produce second layer brick space. This operation could be repeated as many times as needed to generate a multi-layer tree-like molecular space. The number of layers, namely the search depth, is noted as d.

In the previous work,[27] molecules in the tree-like search space were labeled by ML prediction of two properties, and N_{mutation} of them was sampled by maximizing the acquisition function through roulette-wheel selection. Here, however, mutations are performed on bricks—cf. step (1) in Fig. 1—and, therefore, not directly on the BBs, which are subject to our surrogate model evaluation. This mismatch between the heuristically growing brick space and ML models trained by BBs requires a different sampling criterion instead of the acquisition function. Instead, we label them with similarity scores, which are measured by a Tanimoto distance matrix[61,62] between a specific molecule and training set, each evaluated from their molecular fingerprints. We compare three sampling approaches, that are (1) dissimilarity sampling, i.e., bricks with high dissimilarities are assigned higher probabilities to be selected; (2) similarity sampling, i.e., bricks with low dissimilarities are assigned higher probabilities to be selected; and (3) random sampling. As Fig. 3 shows, the performance of dissimilarity sampling is most robust. Random sampling is most powerful at the early steps but further falls down without considering the heuristics between ever-growing molecules at different steps. In these three sampling approaches, the brick selection is wrapped into a BO search algorithm using the ML surrogate models to select promising BBs to be labeled by DFT-corrected DFTB calculations. As a comparison, though, we also tested an approach solely based on the dissimilarity measure, cf. pink diamonds in Fig. 3, which we find to yield very poor results. Instead, dissimilarity sampling of bricks together with the evaluation of BBs by surrogate ML models ensures a robust algorithm.

Therefore, in the following, we focus on dissimilarity sampling (green area in Fig. 2). In detail, $N_{\mathrm{mutation\ batch}}$ bricks with the highest dissimilarities are chosen to generate the brick space. From this space, N_{mutation} bricks with the highest dissimilarities are chosen to form the three-fold rotational symmetric BBs following the steps (2)–(4) in Fig. 1.

The BB optimization itself is then, as already mentioned earlier, performed by a BO search algorithm, in contrast to our earlier work. The reason for this is that we found an acquisition function with $\kappa = 0$ to be most efficient after five-step optimization, as depicted in

10 February 2025 08:37:19

© Author(s) 2024

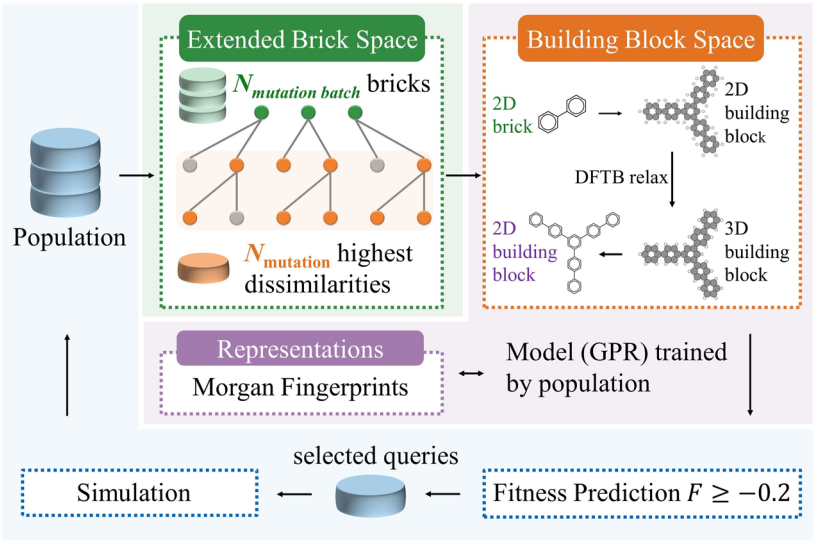

FIG. 2. The workflow of BO including brick space extension (green area), machine learning representation and training (orange–purple area), and selection of molecules based on the model prediction queried into the full population (blue area).

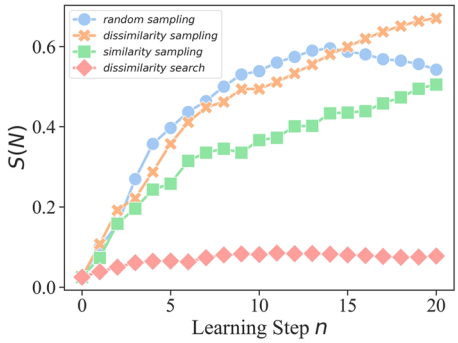

FIG. 3. Three sampling metrics of bricks are compared based on the performances that are evaluated by success ratio $S(N)$ at every learning step n. The performance of the full-exploitation dissimilarity search strategy is also depicted.

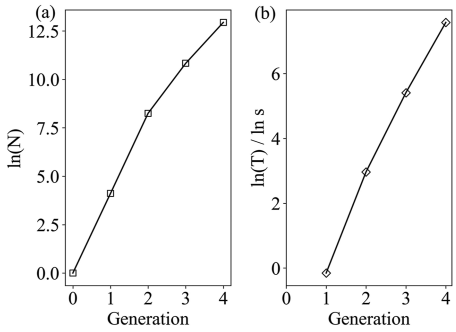

FIG. 4. Natural logarithm of (a) the number of molecular bricks generated in each generation N and (b) the cost time T in units of seconds plotted with respect to generation steps.

Fig. S2 of the supplementary material. $\kappa = 0$ indicates that this is a full-exploitation BO without enforced exploration since the structural dissimilarity is already taken into account in the mutation process.

Candidates with $F \geq -0.2$ are queried, and the queried number is noted as N_{batch}. At each learning step, the full population of DFT-corrected DFTB-valuated molecules is thus $N_{\text{pop}} = N_{\text{initial}} + \sum^{n} N_{\text{batch}}$. These molecules are used to retrain the surrogate model. The complete BO search workflow is illustrated in Fig. 2.

The hyperparameters of the BO process are d, N_{batch}, N_{mutation}, and $N_{\text{mutation batch}}$. The metric of our hyperparameter search is again $S(N)$. The hyperparameters are in the range of $d = 1, 2, 3$, $N_{\text{batch}} = 100, 200, 500$, $N_{\text{mutation}} = 1200, 800, 500$, and $N_{\text{mutation batch}} = 10, 20, 50$, as illustrated in Fig. S3. In general, different hyperparameters strongly affect the calculation and search efficiency. For example, the $N_{\text{mutation batch}}$ is taken as 20 molecules to achieve the highest efficiency. As Fig. S3(d) shows, a too small $N_{\text{mutation batch}}$ would lead to a small candidate pool, with potentially many invalid structures after rough relaxation, decreasing the feasibility of the sampling. In accordance with our optimization results, we set $(d, N_{\text{batch}}, N_{\text{mutation}}, N_{\text{mutation batch}})$ to (1, 100, 800, 20) to optimize the procedure's performance. Based on Figs. S2 and S3, five-step optimization for parameters (including κ) can provide a stable assessment of search efficiency. In the earlier steps, $S(N)$ is occasionally non-monotonic, and the comparison of distinct searches with different parameters is not able to be carried out. Apart from that, five-step calculation expense is justifiable. Note that we further

10 February 2025 08:37:19

© Author(s) 2024

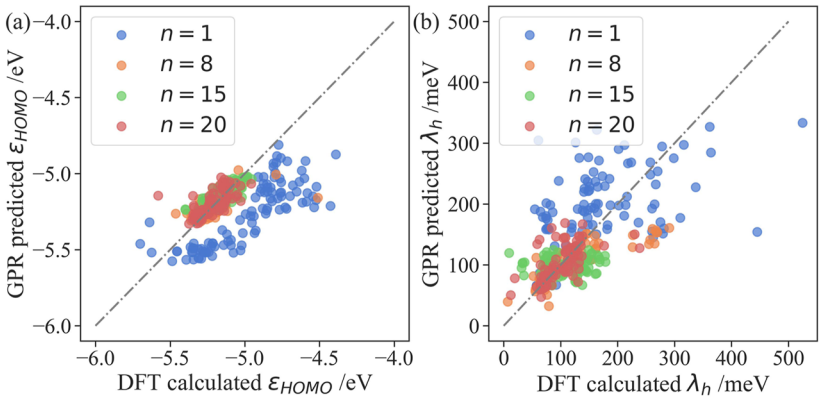

FIG. 5. At learning step n = 1, 8, 15, 20, the distributions of queried N_{batch} molecules that are not contained in the train set of GPR models are depicted. Their (a) DFT-corrected xTB calculated ε_{HOMO}, GPR predicted ε_{HOMO}, and (b) DFT-corrected xTB calculated λ_h, GPR predicted λ_h.

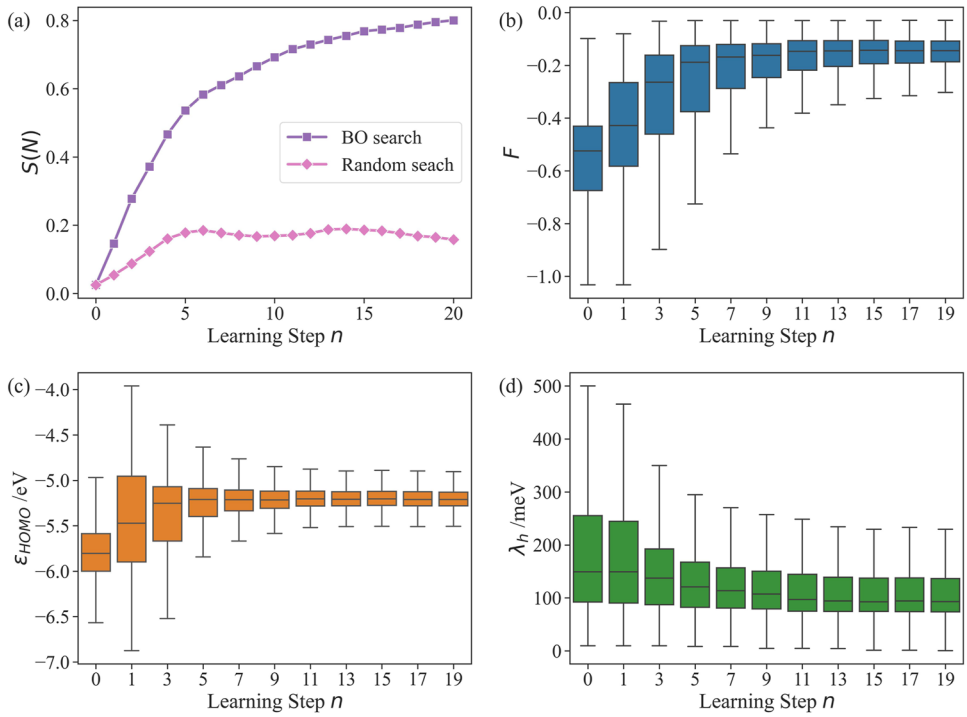

FIG. 6. Visualization of BO search results with optimal parameters. (a) Success ratio denoted by $S(N)$ of BO search and random search is plotted with respect to learning step n. Note that all parameters in these two searches are the same. Distributions of (b) fitness F, (c) ε_{HOMO}, and (d) λ_h are outlined over N_{batch} molecules at each n. When n = 0, it is the distribution of the initial population $N_{initial}$.

perform these optimization searches until 20-step in order to ensure the trends established by five-step ones are reliable. These verification runs are not counted as hyperparamter optimization costs for BO searches.

III. RESULTS

As already mentioned earlier, we initialize the BO search strategy from a very limited molecular space containing 160 BBs. These not only serve to initially train the surrogate models but also as the basis for further BB generation steps. The generation rate and estimated computational cost per generation are depicted in Fig. 4 and Table S2. The generation rate is not fully exponential since several constraints are set during the morphing process. The most important two rules are the longest allowed topological distance, i.e., the maximum number of involved bonds in a molecule, and the maximum number of allowed side groups. These two values are 11 and 2,

10 February 2025 08:37:19

© Author(s) 2024

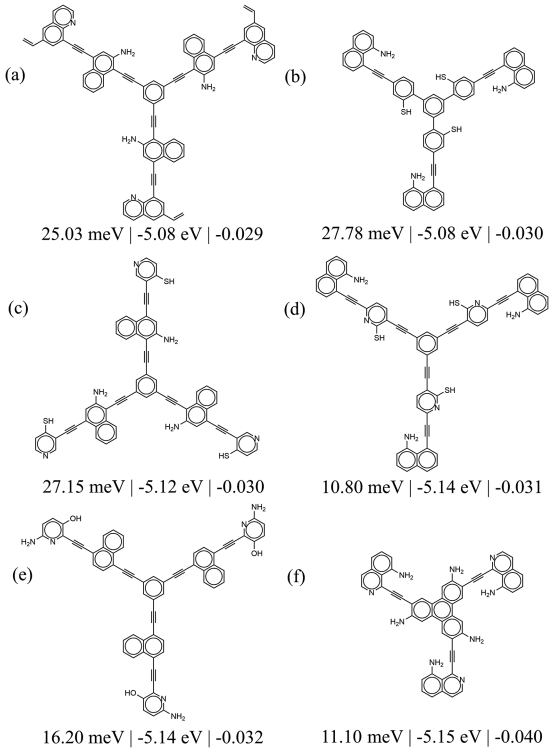

FIG. 7. Six top-performing molecular structures from the BO search space are listed; in addition, λ_h, ε_{HOMO}, and F are noted below.

respectively, to ensure the synthesizability of a final COFs structure. In principle, one generated brick could produce eight three-fold rotationally symmetric building blocks. The three considered connection types between the generated brick and the core benzene of the BB are single and triple bond connections and two rings fused together. As Fig. S1(a) illustrates, the single and triple bond connection sites are two vertices along the molecular backbone axis for bond connections. For ring fusion [Fig. S1(b)], the connection sites are four pairs of meta- and para-points in the molecule. Note that some bricks are not able to form eight BBs; for instance, the meta-position of phenol in 3-phenylphenol is occupied and further formation is prohibited. Apart from these limitations, molecular symmetries cause structural duplicates; for example, biphenyl at most produces three BBs. After the generation procedure of eight BBs, we exclude duplicates and illegal structures where, e.g., side groups would overlap, geometry optimization failed (for whichever reason), or those structures that could otherwise not be used to form a 2D area-filling network. With all duplicate or illegal structures removed, 160 molecules survive and are counted as the initial population to be fed to batch-based BO. We fully evaluate this initial space with xTB to already find four molecules reaching our fitness baseline. From this, we then perform our BO search based on ever-improving surrogate models. In Fig. 5, we sketch the distributions of N_{batch} molecules. These molecules are not contained in the train set of GPR models that are used to predict their ε_{HOMO} and λ_h before the query. Their corresponding DFT-corrected xTB calculations label them after the query.

It shows that the distributions continuously reach the identity line with learning step n going. Meanwhile, the correlations between predictions and calculations of two properties converge to the ideal points gradually. This proves that the robustness of models at deep steps could lead the search to an unknown but promising space. The performances of each model on the training set are depicted in Fig. S4. Figure 6 outlines the BO search and random search results after 20eaning steps with optimal parameters, where $(\kappa, d, N_{\text{batch}}, N_{\text{mutation}}, N_{\text{mutation batch}}) = (0, 1, 100, 800, 20)$. The parameter optimization is of great significance in our approach, and the target criterion during it is the balance between the calculation expense and the search efficiency. As Figs. S2 and S3 show, five-step optimization is able to provide reliable parameters that still perform as expected in the further steps. Figure 6(a) shows that 1736 promising BBs along with $F \geq -0.2$ are collected with a success ratio $S(2160) = 0.80$, while the distribution of fitness F is depicted in Fig. 6(b). There, similar to our earlier uses of AML, we find that the median fitness of the search population rises quickly, while the distribution itself remains broad due to our measures to improve the exploration of the molecular design space. This is illustrated in more detail in Figs. 6(c) and 6(d). Especially the level alignment gets narrower with the median value of $\varepsilon_{\text{homo}}$ approaching -5.1. Therefore, $\varepsilon_{\text{align}}$ tends toward a perfect fit of 0.0. Meanwhile, also, the median of λ_h tends toward smaller values. In principle, with each learning step, the molecules intuitively get more and more complex, leading to a higher rate of failure of our property calculations, e.g., due to more potential for steric overlap between side groups. Therefore, in later generations we need to go to greater lengths to relax the geometries to avoid such invalid molecular overlaps.

After running the BO discovery over 20 steps, we extract the six top-performing BBs with high F scores and display them in Fig. 7. The central motif of the top five well-performing molecules is 1,3,5-triphenylbenzol, and it is a compound of BTPB/BTPA in the synthesis of hexagonal BTP-COFs with 4 nm open pores.[63] The fused ring naphthalene motifs contribute to an increased conductivity of BBs through conjugate bonds. Naphthalene-based linkers are widely used in the synthesis of COF to connect BBs.[64–66] Here, we involve this scaffold into promising BBs directly, which leads to potentially different synthesis and characterization of COFs. Furthermore, we focus on the well performing molecule (f), which exhibits a central triphenylene motif, which is also commonly synthesized in some COFs, such as COF-5[3] and T-COF-OH.[67,68] Such motifs are widely used in diverse applications owing to their high thermal stability and strong fluorescence. The similar nitrogen substituted and nitrogen-containing functional groups attached to the triphenylene motif were also reported by Yilun *et al.*[69] in HPP-COF, which shows high-rate H+/OH− conduction. Overall, we find that the S, N, and O rich functional groups tethered to the backbone tend to increase the conductivity. Additionally, they could provide anchoring points for small molecules in the pores for chemical or photochemical applications of the COF. In order to assess the synthesize ability of our BBs, we employed the computer-assisted organic synthesis software SPAYA,[70,71] which indeed found synthetic routes for (a) and (e).

10 February 2025 08:37:19

© Author(s) 2024

Finally, looking at the collection of best performing molecules as a whole, we found that our more stringent symmetry and stability requirements compared to earlier work[27] alleviated one of the downsides of the earlier AML.

There, as a consequence of the minimization of λ,[34] the molecules tended toward ever larger and softer scaffolds, that are bricks in our current BO search. Indeed, a generation algorithm without the built-in rigidity constraints, only ensuring 3-fold symmetry of the BB candidates, would lead to quite similar results, as depicted in Fig. S5. Specifically, floppy bonds (e.g., -O-, -N-) were included in the AML morphing algorithm to ensure the diversity of the search space. BBs generated with these soft bricks are also floppy and show no potential to form a structural framework. Focusing on those BBs that could feasibly form area-filling COF sheets, we thus feel justified in enforcing rigidity already at the generation level.

IV. CONCLUSION

In this work, we adapted a successful batch-based Bayesian optimization strategy to sample the design space of building blocks for conductive covalent organic frameworks. Thereby, we rely on two proven properties, measuring charge injection barriers (ε_{HOMO}) and barriers to charge transport across COF layers (λ). Owing to the high discovery success rate while also ensuring a high candidate diversity, the BO strategy allowed us to efficiently mine BBs among the overwhelming wealth of potential COF BB molecules. By adapting the generation algorithm, we ensure that only valid and sufficiently rigid BBs are generated. After 20 BO steps, we extract the six top-performing BB candidates to find both well-known and uncommon structural motifs. Our modified generation algorithm counteracts a tendency of BO based on the used properties to maximize the size of the molecules in order to minimize λ. Instead, the constraints we put into the generation algorithm were shown to lead to very plausible BB candidates. Finally, recognizing the need to connect our results to experiments, we also subjected the six best performing BBs to a computational retrosynthesis analysis. While not perfect, the used tool already found synthetic routes for at least two of the BBs.

However, in this work, we only focused on BBs and did not target full 2D COF sheets or even layered 3D structures. Given that most conductive COFs known so far mostly show transport in the out-of-plane direction,[72] we would need to generate full 3D models of the layered COF material. Yet, our morphing algorithm and the molecular representation are based on "SMILES." This 2D encoding and decoding of molecular structures does not give us access to real 3D COF structures. Future work could remedy this by also including 3D structural data in our surrogate modeling. Nevertheless, there are hints that, due to fortuitous error cancellation, gas phase calculations can actually be used to predict $\varepsilon_{\mathrm{HOMO}}$ of experimental condensed phase results.[73] Overall, though, this study once again demonstrated how the use of machine learned surrogate models, paired with suitable constraints derived from a physical and chemical understanding of the design space, can lead to tangible results potentially guiding further development of materials.

SUPPLEMENTARY MATERIAL

Supplementary material is available online as a single PDF containing a complete list of morphing operations, an estimate of the size of the design space at each step, the GPR kernel function, two graphs depicting the results of the hyperparameter optimization, and a graph of BBs resulting from an application of our approach without the molecular stiffness constraint.

ACKNOWLEDGMENTS

H.O. acknowledges support from the German Science Foundation (DFG) through the Heisenberg scheme under Grant No. OB 425/9-1. This work was partially funded by the DFG under Germany's Excellence Strategy (Grant No. EXC 852 2089/1-390776260, Econversion) and the State of Bavaria as part of the "Solar Technologies go Hybrid" initiative.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Yuxuan Yao: Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (equal); Software (lead); Validation (equal); Visualization (lead); Writing – original draft (lead). **Harald Oberhofer**: Conceptualization (lead); Funding acquisition (lead); Methodology (equal); Project administration (lead); Resources (lead); Validation (equal); Writing – review & editing (lead).

DATA AVAILABILITY

The full code and the population used for the initial training of the surrogate models are available at https://github.com/yuxuan1023/COF_Brick_BO_Search. The dataset of 1736 potential building blocks for the construction of COF is available at https://zenodo.org/records/12698407.

REFERENCES

[1] K. Geng, T. He, R. Liu, S. Dalapati, K. T. Tan, Z. Li, S. Tao, Y. Gong, Q. Jiang, and D. Jiang, "Covalent organic frameworks: Design, synthesis, and functions," Chem. Rev. **120**, 8814–8933 (2020).

[2] J. Keupp and R. Schmid, "Topoff: Mof structure prediction using specifically optimized blueprints," Faraday Discuss. **211**, 79–101 (2018).

[3] A. P. Côté, A. I. Benin, N. W. Ockwig, M. O'Keeffe, A. J. Matzger, and O. M. Yaghi, "Porous, crystalline, covalent organic frameworks," Science **310**, 1166–1170 (2005).

[4] Y. Zhou, Z. Wang, P. Yang, X. Zu, and F. Gao, "Electronic and optical properties of two-dimensional covalent organic frameworks," J. Mater. Chem. **22**, 16964–16970 (2012).

[5] X. Zhao, P. Pachfule, and A. Thomas, "Covalent organic frameworks (COFs) for electrochemical applications," Chem. Soc. Rev. **50**, 6871–6913 (2021).

[6] D. Zhu, G. Xu, M. Barnes, Y. Li, C.-P. Tseng, Z. Zhang, J.-J. Zhang, Y. Zhu, S. Khalil, M. M. Rahman, R. Verduzco, and P. M. Ajayan, "Covalent organic frameworks for batteries," Adv. Funct. Mater. **31**, 2100505 (2021).

10 February 2025 08:37:19

© Author(s) 2024

[7]N. Keller and T. Bein, "Optoelectronic processes in covalent organic frameworks," Chem. Soc. Rev. **50**, 1813–1845 (2021).
[8]R. Wang, H. Lyu, G. S. H. Poon Ho, H. Chen, Y. Yuan, K.-T. Bang, and Y. Kim, "Highly conductive covalent–organic framework films," Small **20**, 2306634 (2024).
[9]C. S. Diercks, S. Lin, N. Kornienko, E. A. Kapustin, E. M. Nichols, C. Zhu, Y. Zhao, C. J. Chang, and O. M. Yaghi, "Reticular electronic tuning of porphyrin active sites in covalent organic frameworks for electrocatalytic carbon dioxide reduction," J. Am. Chem. Soc. **140**, 1116–1122 (2018).
[10]J. Liang, Q. Liu, A. A. Alshehri, and X. Sun, "Recent advances in nanostructured heterogeneous catalysts for n-cycle electrocatalysis," Nano Res. Energy **1**, 9120010 (2022).
[11]H. Liao, H. Wang, H. Ding, X. Meng, H. Xu, B. Wang, X. Ai, and C. Wang, "A 2D porous porphyrin-based covalent organic framework for sulfur storage in lithium–sulfur batteries," J. Mater. Chem. A **4**, 7416–7421 (2016).
[12]M. Dogru, M. Handloser, F. Auras, T. Kunz, D. Medina, A. Hartschuh, P. Knochel, and T. Bein, "A photoconductive thienothiophene-based covalent organic framework showing charge transfer towards included fullerene," Angew. Chem., Int. Ed. **52**, 2920–2924 (2013).
[13]H. Oberhofer, K. Reuter, and J. Blumberger, "Charge transport in molecular materials: An assessment of computational methods," Chem. Rev. **117**, 10319–10357 (2017).
[14]M. S. Lohse and T. Bein, "Covalent organic frameworks: Structures, synthesis, and applications," Adv. Funct. Mater. **28**, 1705553 (2018).
[15]P. G. Polishchuk, T. I. Madzhidov, and A. Varnek, "Estimation of the size of drug-like chemical space based on GDB-17 data," J. Comput. Aided Mol. Des. **27**, 675–679 (2013).
[16]K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit, "Big-data science in porous materials: Materials genomics and machine learning," Chem. Rev. **120**, 8066–8129 (2020).
[17]J. Lederer, W. Kaiser, A. Mattoni, and A. Gagliardi, "Machine learning–based charge transport computation for pentacene," Adv. Theory Simul. **2**, 1800136 (2019).
[18]P. Reiser, M. Konrad, A. Fediai, S. Léon, W. Wenzel, and P. Friederich, "Analyzing dynamical disorder for charge transport in organic semiconductors via machine learning," J. Chem. Theory Comput. **17**, 3750–3759 (2021).
[19]Y. Lan, X. Han, M. Tong, H. Huang, Q. Yang, D. Liu, X. Zhao, and C. Zhong, "Materials genomics methods for high-throughput construction of COFs and targeted synthesis," Nat. Commun. **9**, 5274 (2018).
[20]P. Yang, H. Zhang, X. Lai, K. Wang, Q. Yang, and D. Yu, "Accelerating the selection of covalent organic frameworks with automated machine learning," ACS Omega **6**, 17149–17161 (2021).
[21]S. Kumar, G. Ignacz, and G. Szekely, "Synthesis of covalent organic frameworks using sustainable solvents and machine learning," Green Chem. **23**, 8932–8939 (2021).
[22]C.-W. Wu, F. Li, Y.-J. Zeng, H. Zhao, G. Xie, W.-X. Zhou, Q. Liu, and G. Zhang, "Machine learning accelerated design of 2D covalent organic frame materials for thermoelectrics," Appl. Surf. Sci. **638**, 157947 (2023).
[23]D. Wang, H. Lv, Y. Wan, X. Wu, and J. Yang, "Band-edge prediction of 2D covalent organic frameworks from molecular precursor via machine learning," J. Phys. Chem. Lett. **14**, 6757–6764 (2023).
[24]A. Deshwal, C. M. Simon, and J. R. Doppa, "Bayesian optimization of nanoporous materials," Mol. Syst. Des. Eng. **6**, 1066–1086 (2021).
[25]Y. Xie, C. Zhang, H. Deng, B. Zheng, J.-W. Su, K. Shutt, and J. Lin, "Accelerate synthesis of metal–organic frameworks by a robotic platform and Bayesian optimization," ACS Appl. Mater. Interfaces **13**, 53485–53491 (2021).
[26]C. Chowdhury, "Bayesian optimization for efficient prediction of gas uptake in nanoporous materials," ChemPhysChem e202300850 (published online, 2024).
[27]C. Kunkel, J. T. Margraf, K. Chen, H. Oberhofer, and K. Reuter, "Active discovery of organic semiconductors," Nat. Commun. **12**, 2422 (2021).
[28]T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design," npj Comput. Mater. **5**, 21 (2019).
[29]J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," J. Chem. Phys. **148**, 241733 (2018).
[30]J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, and B. Kozinsky, "On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events," npj Comput. Mater. **6**, 20 (2020).
[31]J. P. Janet, S. Ramesh, C. Duan, and H. J. Kulik, "Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization," ACS Cent. Sci. **6**, 513–524 (2020).
[32]C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, *Adaptive Computation and Machine Learning* (MIT Press, 2006), pp. I–XVIII, 1–248.
[33]F. H. Allen, "The Cambridge structural database: A quarter of a million crystal structures and rising," Acta Crystallogr., Sect. B: Struct. Sci. **58**, 380–388 (2002).
[34]C. Kunkel, C. Schober, J. T. Margraf, K. Reuter, and H. Oberhofer, "Finding the right bricks for molecular legos: A data mining approach to organic semiconductor design," Chem. Mater. **31**, 969–978 (2019).
[35]C. A. James and D. Weininger, *Daylight Theory Manual* (Daylight Chemical Information Systems, Inc., Laguna Niguel, CA, 2008).
[36]E. S. R. Ehmki, R. Schmidt, F. Ohm, and M. Rarey, "Comparing molecular patterns using the example of smarts: Applications and filter collection analysis," J. Chem. Inf. Model. **59**, 2572–2586 (2019).
[37]G. Landrum, Rdkit: Open-source cheminformatics (2016), https://www.rdkit.org/.
[38]N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox," J. Cheminf. **3**, 33 (2011).
[39]M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties," Phys. Rev. B **58**, 7260–7268 (1998).
[40]A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior," Phys. Rev. A **38**, 3098–3100 (1988).
[41]C. Lee, W. Yang, and R. G. Parr, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density," Phys. Rev. B **37**, 785–789 (1988).
[42]S. Grimme, C. Bannwarth, and P. Shushkov, "A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all SPD-block elements (z = 1–86)," J. Chem. Theory Comput. **13**, 1989–2009 (2017).
[43]C. Bannwarth, S. Ehlert, and S. Grimme, "GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions," J. Chem. Theory Comput. **15**, 1652–1671 (2019).
[44]B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshaye, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. van der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W.-Z. Yu, and T. Frauenheim, "DFTB+, a software package for efficient approximate density functional theory based atomistic simulations," J. Chem. Phys. **152**, 124101 (2020).
[45]C. Wang, H. Dong, W. Hu, Y. Liu, and D. Zhu, "Semiconducting π-conjugated systems in field-effect transistors: A material odyssey of organic electronics," Chem. Rev. **112**, 2208–2267 (2011).
[46]J. Takeya, M. Yamagishi, Y. Tominari, R. Hirahara, Y. Nakazawa, T. Nishikawa, T. Kawase, T. Shimoda, and S. Ogawa, "Very high-mobility organic single-crystal transistors with in-crystal conduction channels," Appl. Phys. Lett. **90**, 102120 (2007).
[47]O. Jurchescu, J. Baas, and T. Palstra, "Effect of impurities on the mobility of single crystal pentacene," Appl. Phys. Lett. **84**, 3061–3063 (2004).
[48]H. Ishii, K. Sugiyama, E. Ito, and K. Seki, "Energy level alignment and interfacial electronic structures at organic/metal and organic/organic interfaces," Adv. Mater. **11**, 605–625 (1999).
[49]H. B. Michaelson, "The work function of the elements and its periodicity," J. Appl. Phys. **48**, 4729–4733 (1977).
[50]H. Xu, S. Tao, and D. Jiang, "Proton conduction in crystalline and porous covalent organic frameworks," Nat. Mater. **15**, 722–726 (2016).

10 February 2025 08:37:19

© Author(s) 2024

[51] S. F. Nelsen, S. C. Blackstock, and Y. Kim, "Estimation of inner shell Marcus terms for amino nitrogen compounds by molecular orbital calculations," J. Am. Chem. Soc. **109**, 677–682 (1987).
[52] C. Brückner, C. Walter, M. Stolte, B. Braïda, K. Meerholz, F. Würthner, and B. Engels, "Structure–property relationships for exciton and charge reorganization energies of dipolar organic semiconductors: A combined valence bond self-consistent field and time-dependent Hartree-Fock and DFT study of merocyanine dyes," J. Phys. Chem. C **119**, 17602–17611 (2015).
[53] B. Settles, Active learning literature survey (2009).
[54] L. Wilbraham, D. Smajli, I. Heath-Apostolopoulos, and M. A. Zwijnenburg, "Mapping the optoelectronic property space of small aromatic molecules," Commun. Chem. **3**, 14 (2020).
[55] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," J. Chem. Inf. Model. **50**, 742–754 (2010).
[56] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," J. Mach. Learn. Res. **3**, 397–422 (2002).
[57] L. Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi, "Graph kernels for chemical informatics," Neural Networks **18**, 1093–1110 (2005).
[58] F. Häse, L. M. Roch, C. Kreisbeck, and A. Aspuru-Guzik, "Phoenics: A Bayesian optimizer for chemistry," ACS Cent. Sci. **4**, 1134–1145 (2018).
[59] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting," IEEE Trans. Inf. Theory **58**, 3250–3265 (2012).
[60] Z. Zhou, S. Kearnes, L. Li, R. N. Zare, and P. Riley, "Optimization of molecules via deep reinforcement learning," Sci. Rep. **9**, 10752 (2019).
[61] G. Maggiora, M. Vogt, D. Stumpfe, and J. Bajorath, "Molecular similarity in medicinal chemistry," J. Med. Chem. **57**, 3186–3204 (2014).
[62] D. Bajusz, A. Rácz, and K. Héberger, "Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations?," J. Cheminf. **7**, 20 (2015).
[63] M. Dogru, A. Sonnauer, A. Gavryushin, P. Knochel, and T. Bein, "A covalent organic framework with 4 nm open pores," Chem. Commun. **47**, 1707–1709 (2011).
[64] S. Royuela, E. Martínez-Periñán, M. P. Arrieta, J. I. Martínez, M. M. Ramos, F. Zamora, E. Lorenzo, and J. L. Segura, "Oxygen reduction using a metal-free naphthalene diimide-based covalent organic framework electrocatalyst," Chem. Commun. **56**, 1267–1270 (2020).
[65] S. Jhulki, C. H. Feriante, R. Mysyk, A. M. Evans, A. Magasinski, A. S. Raman, K. Turcheniuk, S. Barlow, W. R. Dichtel, G. Yushin, and S. R. Marder, "A naphthalene diimide covalent organic framework: Comparison of cathode performance in lithium-ion batteries with amorphous cross-linked and linear analogues, and its use in aqueous lithium-ion batteries," ACS Appl. Energy Mater. **4**, 350–356 (2021).
[66] Y. Li, X. Su, W. Zheng, J.-J. Zheng, L. Guo, M. Bonn, X. Gao, H. I. Wang, and L. Chen, "Targeted synthesis of isomeric naphthalene-based 2D kagome covalent organic frameworks," Angew. Chem., Int. Ed. **62**, e202216795 (2023).
[67] S. Rager, M. Dogru, V. Werner, A. Gavryushin, M. Götz, H. Engelke, D. D. Medina, P. Knochel, and T. Bein, "Pore wall fluorescence labeling of covalent organic frameworks," CrystEngComm **19**, 4886–4891 (2017).
[68] L. Zhang, L. Yi, Z.-J. Sun, and H. Deng, "Covalent organic frameworks for optical applications," Aggregate **2**, e24 (2021).
[69] Y. Lin, H. Cui, C. Liu, R. Li, S. Wang, G. Qu, Z. Wei, Y. Yang, Y. Wang, Z. Tang, H. Li, H. Zhang, C. Zhi, and H. Lv, "A covalent organic framework as a long-life and high-rate anode suitable for both aqueous acidic and alkaline batteries," Angew. Chem., Int. Ed. Engl. **62**, e202218745 (2023).
[70] Spaya, a community-enriched algorithm for data-driven retrosynthetic planning developed by Iktos, spaya.ai (2020).
[71] M. Parrot, H. Tajmouati, V. B. R. da Silva, B. R. Atwood, R. Fourcade, Y. Gaston-Mathé, N. Do Huu, and Q. Perron, "Integrating synthetic accessibility with AI-based generative drug design," J. Cheminf. **15**, 83 (2023).
[72] S.-W. Kim, H. Jung, M. S. Okyay, H.-J. Noh, S. Chung, Y. H. Kim, J.-P. Jeon, B. M. Wong, K. Cho, J.-M. Seo, J.-W. Yoo, and J.-B. Baek, "Hexaazatriphenylene-based two-dimensional conductive covalent organic framework with anisotropic charge transfer," Angew. Chem., Int. Ed. **62**, e202310560 (2023).
[73] S. Bhandari, M. S. Cheung, E. Geva, L. Kronik, and B. D. Dunietz, "Fundamental gaps of condensed-phase organic semiconductors from single-molecule calculations using polarization-consistent optimally tuned screened range-separated hybrid functionals," J. Chem. Theory Comput. **14**, 6287–6294 (2018).

10 February 2025 08:37:19

© Author(s) 2024