

## ARTICLE OPEN



# Rapid design of top-performing metal-organic frameworks with qualitative representations of building blocks

Yigitcan Comlek<sup>1</sup>, Thang Duc Pham<sup>2</sup>, Randall Q. Snurr<sup>2</sup> and Wei Chen<sup>1</sup>✉

Data-driven materials design often encounters challenges where systems possess qualitative (categorical) information. Specifically, representing Metal-organic frameworks (MOFs) through different building blocks poses a challenge for designers to incorporate qualitative information into design optimization, and leads to a combinatorial challenge, with large number of MOFs that could be explored. In this work, we integrated Latent Variable Gaussian Process (LVGP) and Multi-Objective Batch-Bayesian Optimization (MOBBO) to identify top-performing MOFs adaptively, autonomously, and efficiently. We showcased that our method (i) requires no specific physical descriptors and only uses building blocks that construct the MOFs for global optimization through qualitative representations, (ii) is application and property independent, and (iii) provides an interpretable model of building blocks with physical justification. By searching only ~1% of the design space, LVGP-MOBBO identified all MOFs on the Pareto front and 97% of the 50 top-performing designs for the CO<sub>2</sub> working capacity and CO<sub>2</sub>/N<sub>2</sub> selectivity properties.

npj Computational Materials (2023)9:170; <https://doi.org/10.1038/s41524-023-01125-1>

## INTRODUCTION

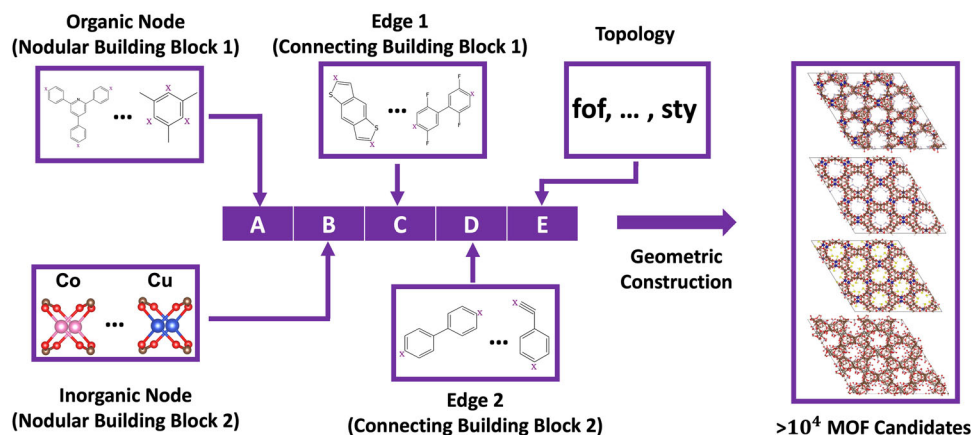
With recent advances in machine learning (ML), material system design and development has undergone rapid acceleration<sup>1,2</sup>. However, one of the major challenges in applying ML to material system design lies in finding the appropriate design representations<sup>3</sup>. Most material design applications take advantage of quantitative (or numerical) design variables to represent material systems. In most cases, these quantitative descriptors (features) require either expert knowledge or data analysis to find the most appropriate ones. On the other hand, although most qualitative (or categorical) variables (e.g., chemical elements, chemical compositions) are more accessible than quantitative variables, it is challenging to directly include qualitative variables as a part of the design variables in automated materials design. Metal-organic frameworks (MOFs) are an example of such materials systems. MOFs are a class of porous crystalline materials that have been used extensively for gas storage<sup>4,5</sup>, gas separation<sup>6–9</sup>, and catalysis<sup>10–12</sup>. Because of their highly tunable nature, MOFs have been looked at as a potential solution for different applications such as carbon dioxide (CO<sub>2</sub>) capture and separation<sup>13,14</sup>. Using a vector notation in which each element corresponds to a qualitative design variable such as topology, node, and edge, MOFs can be represented with the sole usage of qualitative variables as shown in Fig. 1. However, the versatility and different possible combinations of the MOF building blocks lead to millions of candidates. To demonstrate a simple example, consider a MOF system with a topology that requires 2 nodes and 3 edges for construction. Selecting only 20 different building blocks for each node and edge leads to a combinatorial design space of more than 10<sup>6</sup> MOF candidates. Due to the high experimental cost, both in time and resources, computational approaches have been increasingly used to replace experimental exploration<sup>3</sup>.

While high-throughput screening approaches<sup>15–20</sup> and ML techniques<sup>21–23</sup> have been utilized to computationally search or design for top-performing MOF structures in different applications, existing approaches usually rely on large data sets and high-

dimensional physical descriptors to represent the material design space. These processes can be both time consuming and property specific, meaning that the ML models and descriptors are often not transferable to different design objectives. Finally, many ML models are viewed as ‘black boxes’ that are not easily interpretable for understanding how and why the model performs the way it is<sup>24–26</sup>. Therefore, a new and a generic computational approach that (i) employs a simple but descriptor-free (featureless) design representation, (ii) requires substantially smaller amount of data, and (iii) is easily interpretable would be highly useful for the design of MOFs.

Bayesian optimization (BO) has been shown to be effective for identifying the optimum candidates for materials systems with large design spaces and local optimums in different applications such as drug discovery, additive manufacturing, and genetics<sup>27,28</sup>. BO has also been used to identify high-performance MOFs<sup>29,30</sup>. However, previous works on MOFs require expert knowledge for the choice of appropriate physical descriptors (e.g., gravimetric surface area, largest included sphere diameter) as inputs for surrogate model training. Gaussian Process (GP) is a popular surrogate model choice for BO as it provides both predictions and uncertainty quantification, which are the two main components of the acquisition function for choosing samples when applying BO. However, GP models fall short when there are qualitative design variables. This bottleneck has been recently bypassed by the Latent Variable Gaussian Process (LVGP)<sup>31</sup> approach, which can incorporate qualitative variables into GPs by implicitly mapping each qualitative variable into a quantitative space through low-dimensional latent variables. Specifically, as the influence of any qualitative variable on any quantitative response must be always due to some underlying, possibly high-dimensional, quantitative physical variables, the latent variable approach provides physics-based dimension reduction. Therefore, the latent variables and their locations in the latent space could provide physically meaningful information on how the qualitative variables influence the responses. In the context of latent space learning, the term

<sup>1</sup>Department of Mechanical Engineering, Northwestern University, Evanston, IL, USA. <sup>2</sup>Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. ✉email: [weichen@northwestern.edu](mailto:weichen@northwestern.edu)



**Fig. 1** The qualitative representation and construction of metal-organic framework materials. Each MOF can be represented with a “vector” where each element (letter) in the vector corresponds to a choice of building block or topology. With the combination of different building blocks and topologies explored in this work, there are more than  $10^4$  hypothetical MOFs to be explored.

“physically-meaningful” can be associated to explaining the cause-effect relationships between the design variables (inputs) and properties (output) through supervised learning, which is different from many existing latent space learnings through unsupervised learning methods. LVGP still possesses the qualities of a GP model in terms of providing fast surrogate modeling, capturing nonlinear responses, providing predictions, and quantifying uncertainties. Thus, LVGP bridges the gap for incorporating qualitative information into engineering design applications and has been already employed in data-driven materials design research<sup>32,33</sup>. Although LVGP and BO have been applied to materials design and development, its application has been limited to qualitative design variables with small number levels, i.e., the design options per variable.

Here we present the Latent Variable Gaussian Process Multi-Objective Batch Bayesian Optimization (LVGP-MOBBO) framework to perform rapid design of superior MOFs directly from the building blocks that construct the material. Specifically, we are interested in identifying the Pareto front for a multi-objective optimization and top-performing MOFs without any human intervention. We are particularly interested in examining the performance of the approach under both small and large numbers of levels for qualitative variables. We take advantage of the readily available qualitative building block information that is used to construct the MOFs and build an interpretable LVGP surrogate model that cooperates with MOBBO to adaptively lead towards promising MOF candidates for CO<sub>2</sub> capture and separation. With the integration of batch BO, this work shows that descriptor-free LVGP can also be effectively extended to applications with substantial number of levels.

## RESULTS

### Design spaces

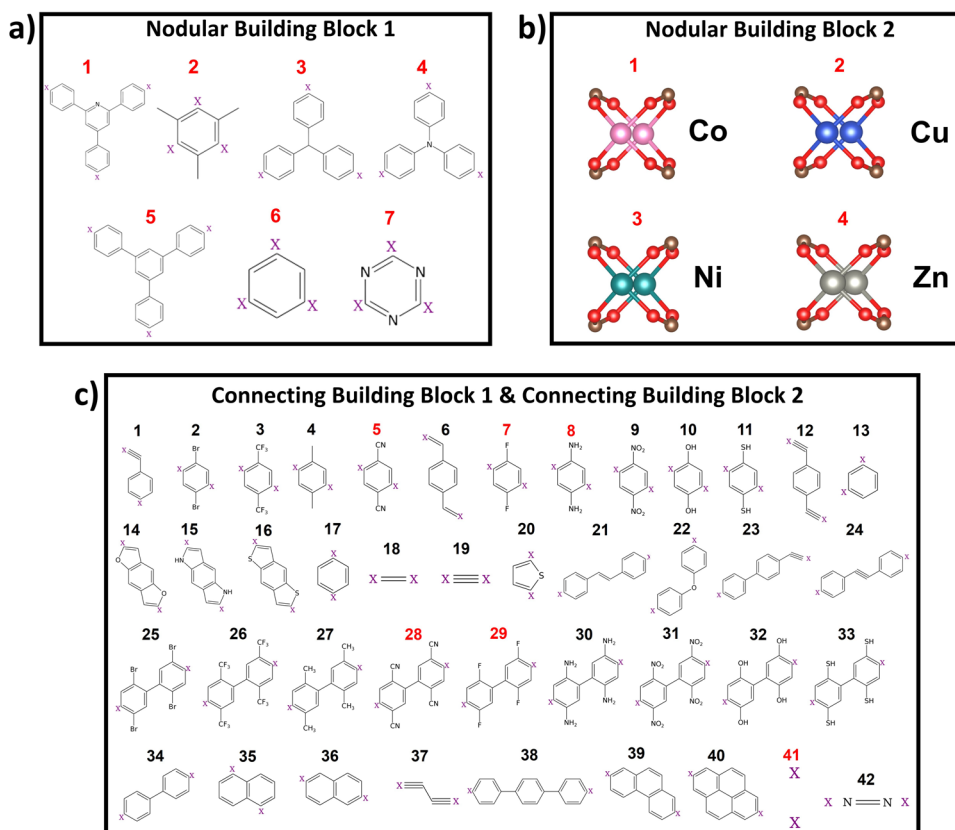
To show the effectiveness of LVGP-MOBBO, we demonstrated our framework on a design space using the **fof** topology, which consists of 4 types of building blocks (BB). We used 7 organic nodes (Nodular BB1) and 4 inorganic nodes (Nodular BB2). There are also two types of edges in the **fof** topology, and we used 41 edges for Connecting BB1 and 42 edges for Connecting BB2. All the building block choices are displayed in Fig. 2. For the use of MOFs in carbon capture, two of the most important metrics are the CO<sub>2</sub> working capacity and the CO<sub>2</sub>/N<sub>2</sub> selectivity. Since we focused on method development, we calculated these properties for all MOFs in our design space in advance to aid in testing different variations of the search methods. The first design space,

which we denoted as the Reduced Design Space (RDS) for validation purpose, consists of 1001 MOF designs that were specifically selected by choosing certain building blocks highlighted in red in Fig. 2 to demonstrate the interpretability of LVGP and the effectiveness of finding optimal MOF designs when combined with BO. The chosen building blocks have known similarities and differences in chemistry and structures, which are reflected in the distances between the building blocks in the latent space of LVGP. Further details of RDS are provided under the Performance section. The second design space, which we denoted as Entire Design Space (EDS) contains 47,740 MOF design candidates that were constructed by combining all available building blocks (7, 4, 41, 42) for the organic node, inorganic node, and the two edges (Fig. 2). Our framework, LVGP-MOBBO, was demonstrated on this design space to show the effectiveness of LVGP when a large number of building blocks (levels) are present.

The two main goals of our design optimization are to identify (i) the Pareto front of MOF designs between the CO<sub>2</sub> working capacity and CO<sub>2</sub>/N<sub>2</sub> selectivity, and (ii) the top-performing MOFs. The Pareto front represents the set of MOF designs that possess properties that are superior to the rest of the design space but cannot be improved without sacrificing the other properties of interest<sup>34</sup>. Furthermore, the top-performing MOFs represent the MOF designs that are closest to the Utopian MOF design in the Euclidian space. The Utopian MOF corresponds to a hypothetical MOF design that possesses the maximum available property of all objectives, which is often not achievable and therefore considered to be “Utopian”. In many multi-objective optimization applications, identifying solutions near the Utopian point is used to evaluate the performance of algorithms in terms of optimizing both objectives at the same. As a result, the Utopian MOF is used as a reference point to identify MOF designs, which we denote as “top-performing” designs, that have high values of both properties.

### Framework

We would like to explore a given design space with as few resources as possible. Thus, we implemented the LVGP-MOBBO framework to perform descriptor-free MOF design optimization with only qualitative representations of building blocks. Our proposed LVGP-MOBBO design exploration framework, consists of 5 major parts: Initial Design of Experiments, Property Evaluation, Latent Variable Gaussian Process, Multi-Objective Batch Bayesian Optimization, and Design Solution (Fig. 3).



**Fig. 2** The design space of fof topology used in the study. The fof topology consists of four building block options (a Nodular Building Block 1, b Nodular Building Block 2, c Connecting Building Block 1 and Connecting Building Block 2). Building blocks highlighted in red are selected for the Reduced Design Space (RDS). For the Entire Design Space (EDS) all of the building blocks shown in the figure are used.

**Initial Design of Experiments (DOE).** For a large design space optimization, the initial selection of design candidates plays a key role. Ideally, they should span the design space as much as possible, for which we employed the optimal sliced Latin hypercube sampling (OSLHS)<sup>35</sup>. The generated MOF designs were then passed into the next task for property evaluation. The detailed generation of the DOE is explained in the *Methods* sections.

**Property Evaluation.** Hypothetical MOFs were created using the ToBaCCo 3.0 package<sup>36</sup>, and the geometry optimization was carried out using the LAMMPS code<sup>37</sup> with the UFF4MOF force fields<sup>38</sup>. Grand canonical Monte Carlo (GCMC) simulations were performed using the RASPA package<sup>39</sup> to evaluate the CO<sub>2</sub> working capacity and CO<sub>2</sub>/N<sub>2</sub> selectivity properties. Further details of the property evaluation can be found in the *Methods* section.

**Latent Variable Gaussian Process (LVGP).** Using the available MOF designs and their associated properties from the GCMC simulations, one LVGP model for each property was trained. Next, the properties of unexplored MOFs in the design space were predicted along with their quantified uncertainty in predictions, which are utilized by the MOBBO. The details of the LVGP modeling are provided in the *Methods* section.

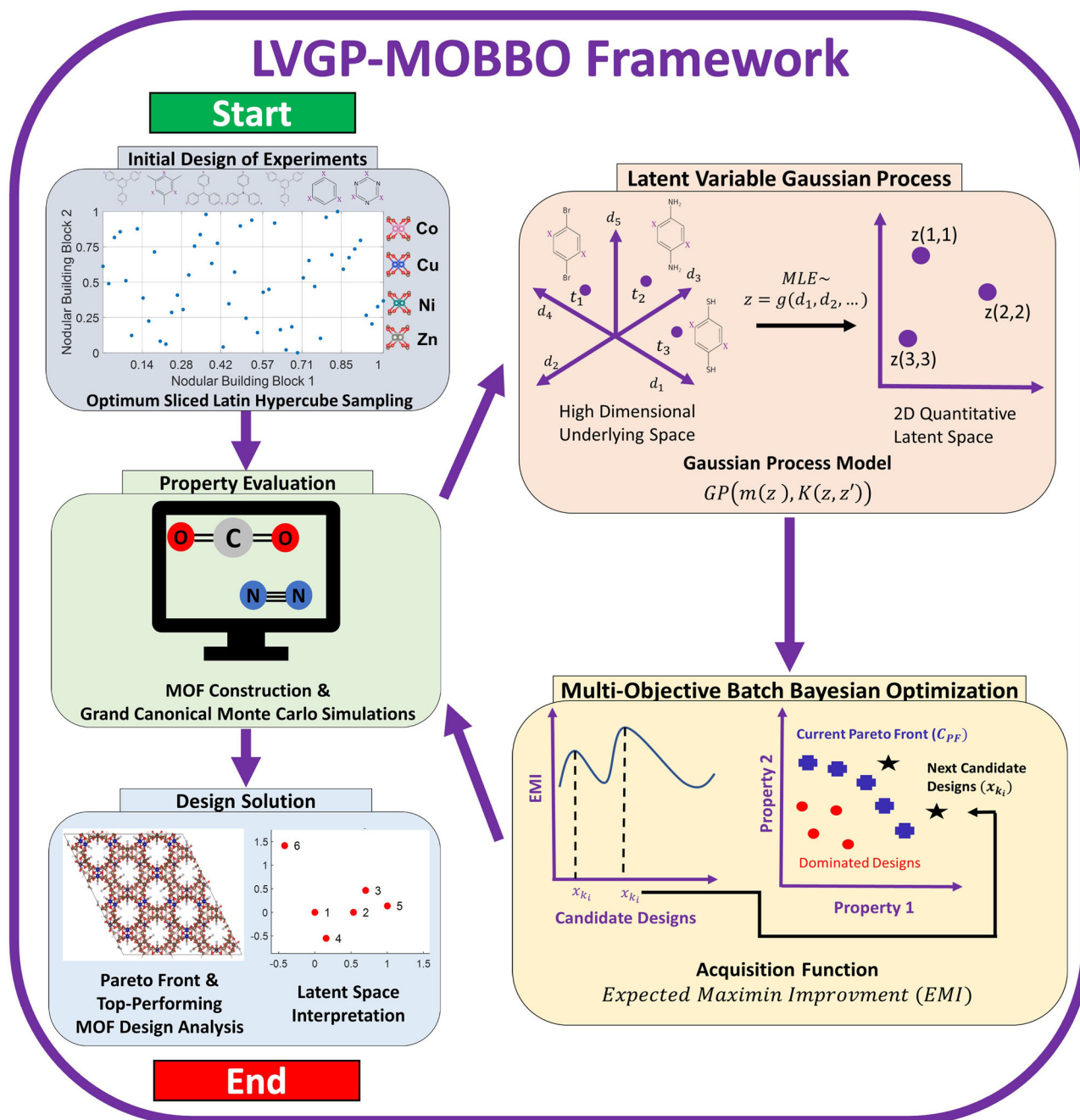
**Multi-Objective Batch Bayesian Optimization (MOBBO).** Utilizing both the predictions and the uncertainty estimates on the remaining candidates in the design space from the LVGP model, the MOBBO selects a batch of MOFs that has the highest Expected Maximin Improvement (EMI) values. The EMI is formulated in a way that both objectives have equal importance. A batch of  $B$  number of MOF designs with the highest EMI values is selected

and passed on to the Property Evaluation task once again. The framework then continues in this cycle until the stopping criterion (e.g., number of MOFs identified) is reached. Further details and formulation of the MOBBO are provided under the *Methods* Section.

**Design Solution.** Once the optimization stopping criterion is reached, the identified design candidates are analyzed further to distinguish the Pareto front and the top-performing MOF designs. Finally, the latent space of each building block is visualized to make inferences about their influence on each property of interest.

### Validation using a reduced design space (RDS)

Before applying our proposed methodology to a large design space, we validated the effectiveness of LVGP and BO on MOFs by implementing the optimization campaign on a relatively small design space. This space contains Connecting BBs that were handpicked to show the novelty of the methodology by validating the latent variables obtained at the end of the optimization campaign and assessing the efficiency of the methodology for designing MOFs that possess superior properties. All the Nodular BBs (7 and 4 levels for Nodular BB1 & BB2) and 6 building blocks from the Connecting BB1 & BB2 were selected for RDS. Specifically, we selected Connecting BBs labeled as {5, 7, 8, 28, 29, 41} (Fig. 2). The specifically selected Connecting BBs contain known differences and similarities in both structure and chemistry. The blocks {5, 7, 8} have similar molecular structures with different functional groups (-CN, -F, -NH<sub>2</sub>). Blocks {28} and {29} are extended structures of blocks {5} and {7}, respectively. Finally, block {41} is an empty building block, which facilitates a direct connection between the



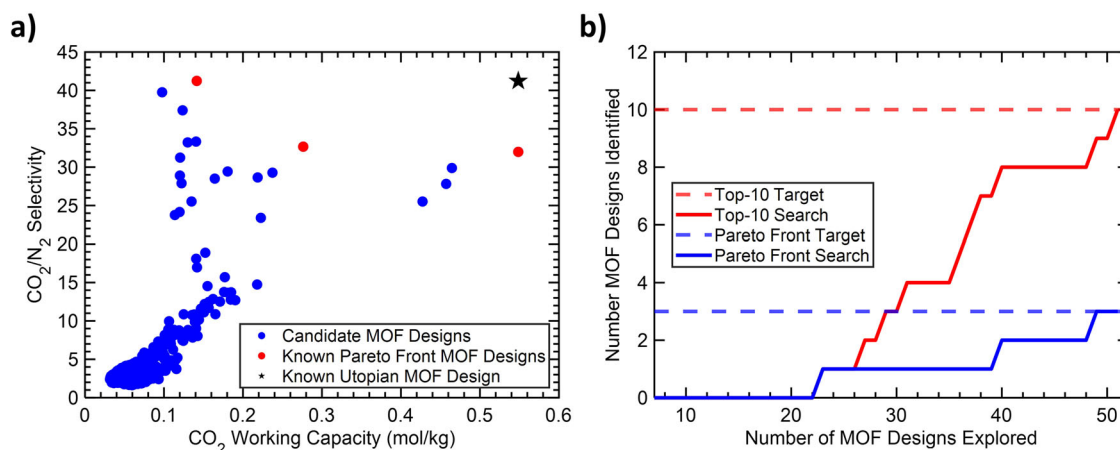
**Fig. 3** The Latent Variable Gaussian Process-Multi Objective Batch Bayesian Optimization (LVGP-MOBBO) framework. The initial set of materials, also known as DOE, is generated by optimal sliced Latin hypercube sampling. Property Evaluation includes MOF construction and prediction of their adsorption properties using Grand canonical Monte Carlo simulations. The LVGP builds the surrogate model that captures the relationship between the design and property space. MOBBO makes the next batch of MOF designs for property optimization. Design Solution analyzes the MOF designs and the latent spaces. The details of each box are explained thoroughly in the *Methods* section.

Nodular BB1 and BB2 and it is known to result in superior properties for some gas separation applications when included in the design of MOFs. As a result, we first aimed to validate the influence of different choices of building blocks on the properties through analyzing and explaining the spatial relationships between latent variables obtained from the LVGP models. Next, we aimed to demonstrate the effectiveness of LVGP for identifying optimal building blocks when used with BO.

The RDS contains 1001 MOF design candidates and three Pareto front MOFs designs. The property space with the known Pareto front and the Utopian designs is shown in Fig. 4a. In addition to

demonstrating the effectiveness of LVGP for MOFs, the design optimization goal of this study was to identify both the Pareto and other top-performing designs. To account for all the possible Nodular BB1s, 7 MOFs were chosen for the initial DOE using OSLHS. Each of the 7 MOFs corresponds to one level of the Nodular BB1. Furthermore, due to the small number of candidates, we chose to add one MOF design,  $B = 1$ , during each iteration of BO. Batch BO was implemented on a larger design space, as discussed later in the paper.

The LVGP-BO design optimization campaign ran until the stopping criteria of identifying both the Pareto front and the 10



**Fig. 4** The LVGP-BO results for the Reduced Design Space (RDS) exploration. **a** The property space of the available MOF candidates. The known Utopian and Pareto front MOF designs are highlighted with black and red points, respectively. **b** Design optimization history for 10 top-performing and Pareto front MOF designs. The blue color represents the Pareto front search, and the red color represents the 10 top-performing design search.

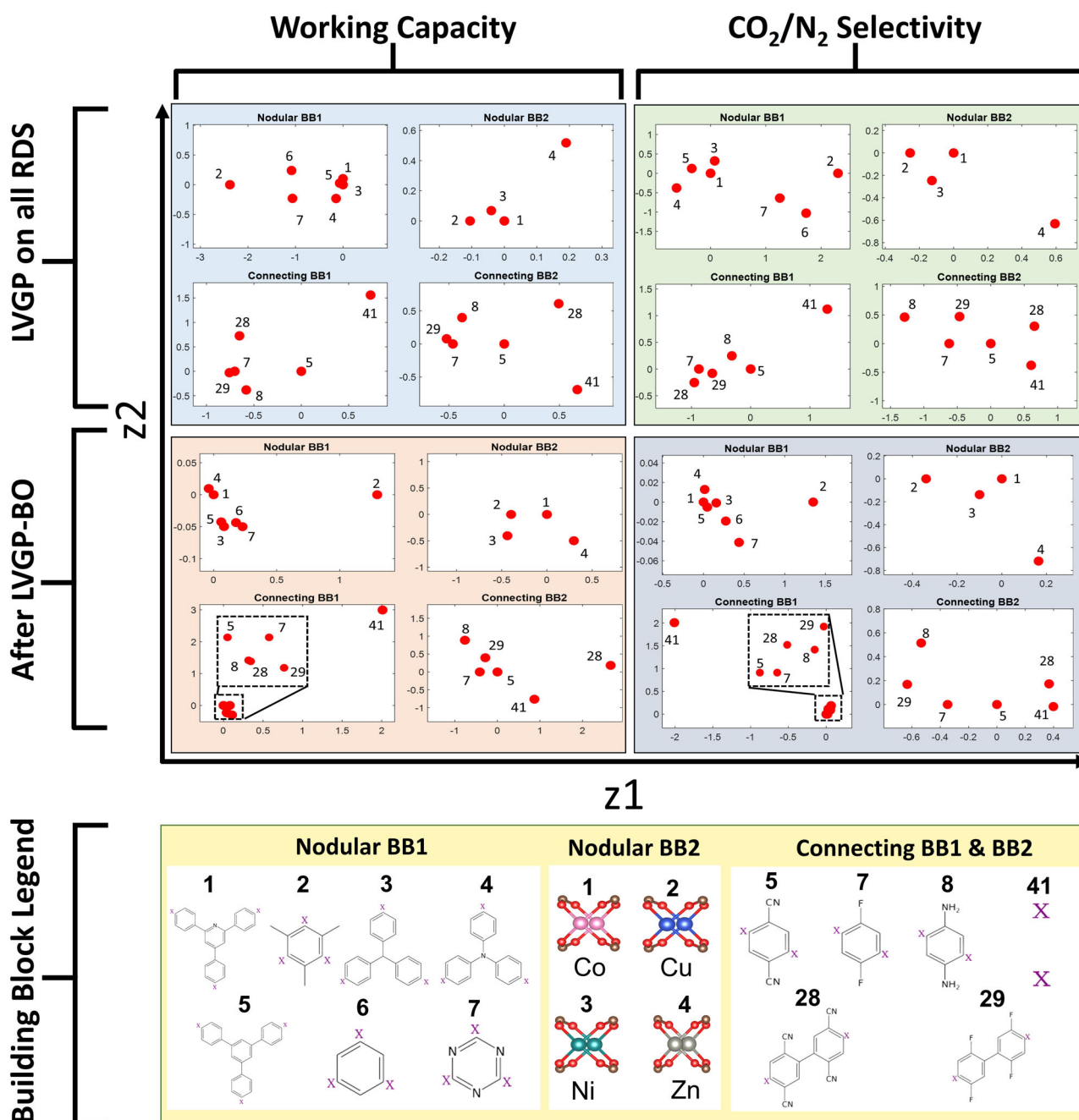
top-performing MOFs were satisfied. Starting with 7 initial MOFs in the DOE, this stopping criterion led to 44 iterations, which in return shows that a total of 51 (7 + 44) MOFs (5.1% of the design space) are explored as the next design candidates. Specifically, the LVGP-BO framework found all three Pareto front MOF designs in 42 iterations, which corresponds to 4.9% of the design space. The optimization history of identified MOF designs can be seen in Fig. 4b. The fast design exploration of the LVGP-BO demonstrated its capability of finding top-performing MOFs.

To verify the interpretability of LVGP models, we examined the latent spaces obtained from training the LVGP surrogate model for both properties after the 44-iteration optimization campaign. In addition, to validate the correctness of the latent variables obtained from the optimization campaign, we also trained the LVGP models on the entire RDS for both properties. We then compared the latent spaces obtained from these two instances. The comparison of the latent spaces for both objectives is shown in Fig. 5. The four large colored boxes represent the latent space obtained for each BB after training the LVGP, the two columns represent different properties, and the two rows represent the different training instances. By comparing boxes in each column, we observed that the latent space representations obtained at the end of the design optimization show differences with the latent spaces obtained from the LVGP model trained on the entire design space. This was an expected result since LVGP-BO is optimization driven. However, independent of orientation and the scale of  $z_1$  and  $z_2$  (the 2D latent space axis), the relative distances between latent variables, which reflect the relationships between design choices (building blocks) and their influence on the properties, are preserved after LVGP-BO. For example, for Connecting BB1, level {41} is far from the other levels for both properties in both training instances in Fig. 5. This similarity shows that even though the LVGP used in the design optimization framework was trained on a very small portion of the entire design space that is biased towards promising building block candidates, it can capture the underlying latent variables and the relationships between building blocks very well. This can be very advantageous for designers to understand and extract true meanings from the design decisions that our framework makes.

The next question then becomes, what do these latent variables represent? As we previously mentioned, since the influence of every qualitative variable on the quantitative response of interest must be due to some physical quantitative variables, the low dimensional latent variables could provide physically meaningful information regarding the cause-effect relationships between

qualitative design variables (inputs) and properties (outputs). Specifically, the spatial relations (distances) between different qualitative design choices in the latent space can show similarity and differences regarding the influence of these properties on the response. Similarly, spatial relationships between latent variables can also imply the dimensionality of underlying physical descriptors. Figure 6 shows the importance of the input space, in terms of the textural characteristics, on the property space. For both the RDS and the EDS that will be demonstrated later, we found that most top-performing MOFs for  $\text{CO}_2/\text{N}_2$  separation often have small pores, characterized by low values of the largest cavity diameter (LCD) and small gravimetric surface area (GSA). MOFs with smaller pores could result in stronger van der Waals interaction and thus favor  $\text{CO}_2$  adsorption over  $\text{N}_2$  adsorption. Knowing the importance of the input space on the latent space, we further investigated how different building blocks affect the pore size, and ultimately the latent space (Fig. 7). For the latent plot of Nodular BB1 (as shown in Fig. 5), we found that the distance among the blocks {1, 3, 4, 5} and {6, 7} are small, and block {2} is always far from the rest of the variables. Building blocks {2, 6, 7} are smaller blocks than {1, 3, 4, 5}, resulting in MOFs with smaller LCD (Fig. 7a). This could explain why blocks {1, 3, 4, 5} are always closer in the latent space than {2, 6, 7}. Moreover, block {2} is bulkier, with a branching  $-\text{CH}_3$  group, than block {6, 7}, resulting in MOFs with slightly smaller pores, and thus far away from the other building blocks. In the Connecting BB1 latent variable plots, we observed that the 5 blocks {5, 7, 8, 28, 29} formed a cluster and are located far away from the block {41}. Because {41} is an “empty” building block (Fig. 5), using block {41} resulted in MOFs with significantly smaller pores than other building blocks (Fig. 7b), and thus different in the property space.

For Nodular BB2 and Connecting BB2, we found that the building blocks lead to minimal differences in the pore sizes (Supplementary Fig. 1), and thus LCD could not be used to explain the latent space. For Nodular BB2, the building blocks have the same shape and differ only in their metal elements {1: Co, 2: Cu, 3: Ni and 4: Zn}. A potential explanation for the latent space is the difference in Lennard-Jones parameters (Supplementary Table 1), in which Zn has an  $\epsilon$  value of about one order of magnitude larger than the other elements, suggesting a stronger van der Waals interaction for Zn, which could favor  $\text{CO}_2$  adsorption over  $\text{N}_2$ . As a result, block {4} (or Zn) is far apart from the other designs. Although the chemical identities of the building blocks in Connecting BB2 are the same as in Connecting BB1, Connecting BB2 has small effect on the pore size, and thus the property space.



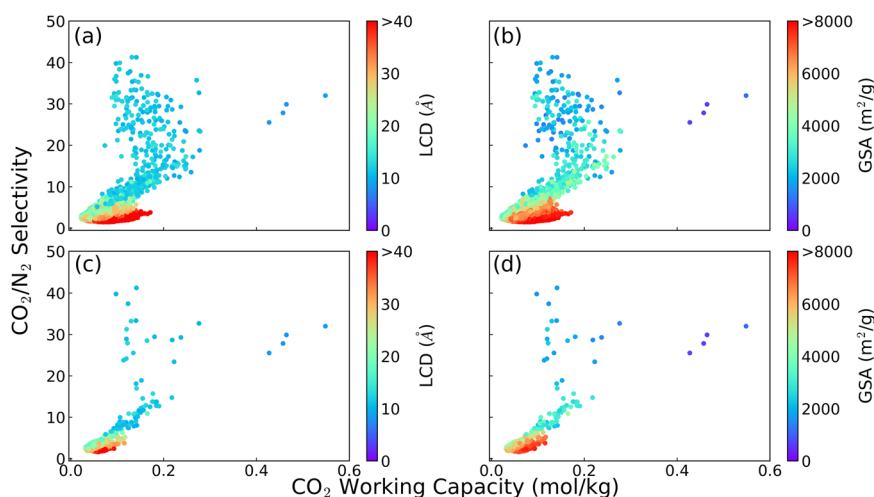
**Fig. 5** The latent variables obtained from the Reduced Design Space (RDS) study. Each colored box shows the 4 building block design variables and red dots show their respective latent variables. The numbers represent the design choice for the specific building block and the legends for the numbers are found in the yellow box on the bottom of the figure. The axes  $z_1$  and  $z_2$  represent the 2D latent space obtained from the LVGP model. The 1<sup>st</sup> row represents the latent variables obtained by training LVGP on the all RDS and the 2<sup>nd</sup> row represents the latent variables obtained after 44 iterations of LVGP-BO on RDS. The 1<sup>st</sup> and 2<sup>nd</sup> columns represent the  $\text{CO}_2$  working capacity and  $\text{CO}_2/\text{N}_2$  selectivity properties, respectively. Finally, the dashed boxes show the zoomed in images of clustered latent variables.

As a result, the points are evenly spread out in the latent space of Connecting BB2. Both the optimization performance and the physically interpretable model obtained from this design study demonstrate the effectiveness of LVGP and BO for further design applications.

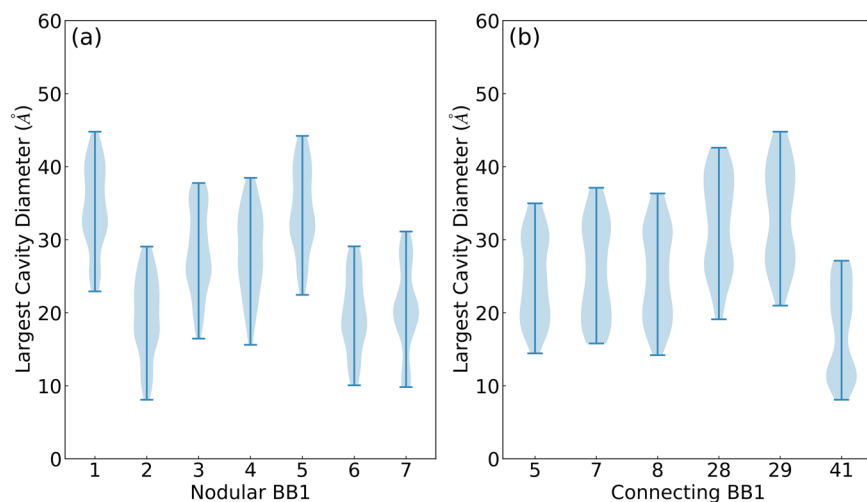
#### Entire design space (EDS)

After confirming the effectiveness of our methodology on the RDS, we applied our framework to the entire design space (EDS) that contains 47,740 MOF candidates with **fof** topology through

combination of 7, 4, 41 and 42 building blocks for Nodular BB1, Nodular BB2, Connecting BB1, and Connecting BB2, respectively. The MOF candidates and their respective properties can be seen in Fig. 8a. The design space contains 7 Pareto front MOF designs of interest. Incorporating our knowledge from previous LVGP implementations<sup>32,40</sup> and considering the large number of available building blocks in the design space, we decided to match each edge (Connecting BB2, 42 options) with each metal node (Nodular BB2, 4 options), resulting in a total of 168 MOFs to be selected for the initial DOE. To create this DOE, we used OSLS



**Fig. 6 Structure – property relationship of the Entire Design Space (EDS) and Reduced Design Space (RDS) datasets.** The CO<sub>2</sub>/N<sub>2</sub> selectivity versus the CO<sub>2</sub> working capacity for the EDS (a, b) and for the RDS (c, d), colored by the largest cavity diameter (LCD) (a, c) and the gravimetric surface area (GSA) (b, d).

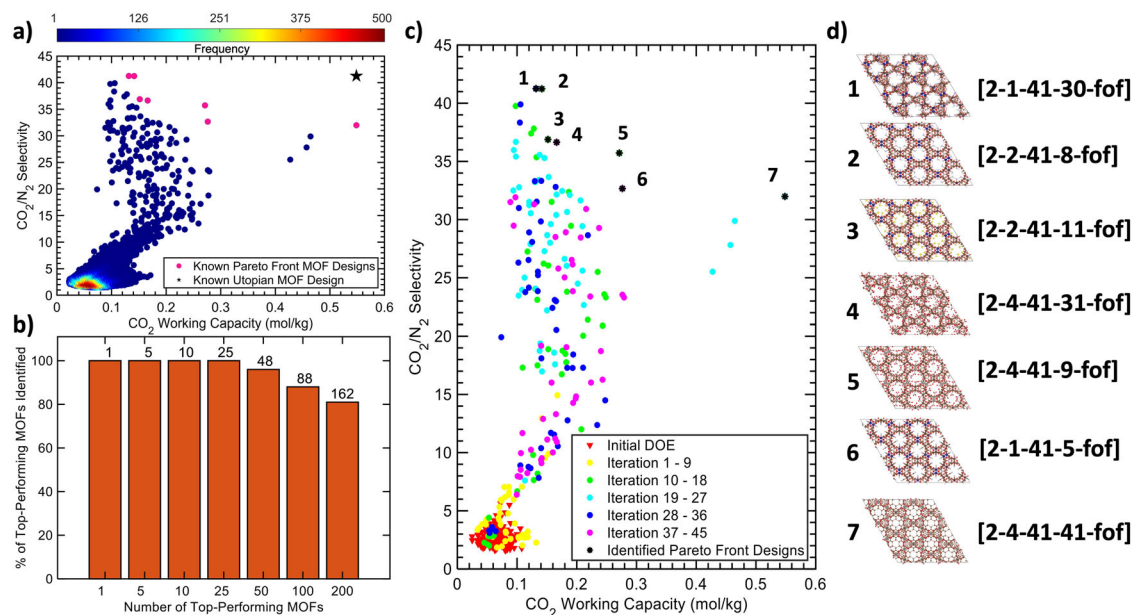


**Fig. 7 The distribution of the largest cavity diameters of 1001 MOFs in the Reduced Design Space (RDS) for different building blocks.** Largest Cavity Diameter (LCD) distribution for (a) Nodular BB1 and (b) Connecting BB1 on RDS.

once again. After starting the LVGP-MOBBO framework with 168 MOFs, we proceeded by adding batches of  $B = 5$  new MOFs with highest Expected-Maximin Improvement (EMI) values until the design space search campaign reached the stopping criterion. The design optimization campaign stopped when the mean EMI values of the 5 MOFs that are selected for property evaluation in each iteration is less than a constant,  $\delta$ , taken as  $\delta = 10^{-5}$  in our study. We implemented this stopping criterion to account for the fact that in practice, it is not plausible to know if we recovered the entire Pareto front set without exploring all MOF designs in the Pareto front. As a result, the implemented stopping criterion tells us that our metamodels, LVGP, are confident that there can be no further improvements made to the MOF design optimization if more MOF designs are added into the framework.

With the aforementioned stopping criterion, the LVGP-MOBBO design optimization campaign continued for 66 iterations, identifying 498 MOF designs in total, including the initial 168 MOFs. Our results show that by scanning only 1.04% of the entire design space, LVGP-MOBBO identified all MOF designs that lie on the Pareto front. Specifically, as seen in Fig. 8c, all the Pareto front designs are identified within 45 iterations, which corresponds to

exploration of only 0.82% of the entire design space. This shows that our methodology is very effective and efficient. Although the initial DOE covers the MOF input design space as evenly as possible, the MOFs in the DOE are not distributed evenly in the property space (Fig. 8c). Figure 8d shows the images of the seven MOFs that lie in the Pareto front (Fig. 8c). The five-dimensional vector representation shows the design choices selected for the Nodular Building Block 1, Nodular Building Block 2, Connecting Building Block 1, Nodular Connecting Block 2, Topology respectively. The choices of building blocks can be found in Fig. 2. For some machine learning and optimization methods, this can be problematic, as we show later with the Random Forest approach. However, our methodology was swift in guiding the design decisions towards MOFs with high properties. Figure 8b shows the result of exploring the different number of top-performing MOF designs that are closest to the Utopian MOF design. The LVGP-MOBBO found all of the 25 top-performing MOFs. Furthermore, our methodology identified more than 97%, 87%, 80% of the 50, 100, 200 top-performing MOF designs, respectively. Finally, out of all 330 MOFs explored, 206 MOF designs (63.3%) belong to the 330 top-performing MOFs. The high efficiency in identifying a



**Fig. 8 Performance of the LVGP-MOBBO on the Entire Design Space (EDS).** **a** The property space of MOF design candidates along with Pareto front and Utopian MOF designs. **b** Percentage of top-performing MOFs identified after 66 iterations. Numbers on top of bars indicate the amount of identified top-performing MOF designs. **c** The initial DOE and the identified MOF designs after different numbers of iterations. **d** The building block representations of Pareto front MOFs and their crystal structures. Each MOF is represented as a vector [A-B-C-D-E], where each letter represents Nodular BB1, Nodular BB2, Connecting BB1, Connecting BB2, Topology respectively.

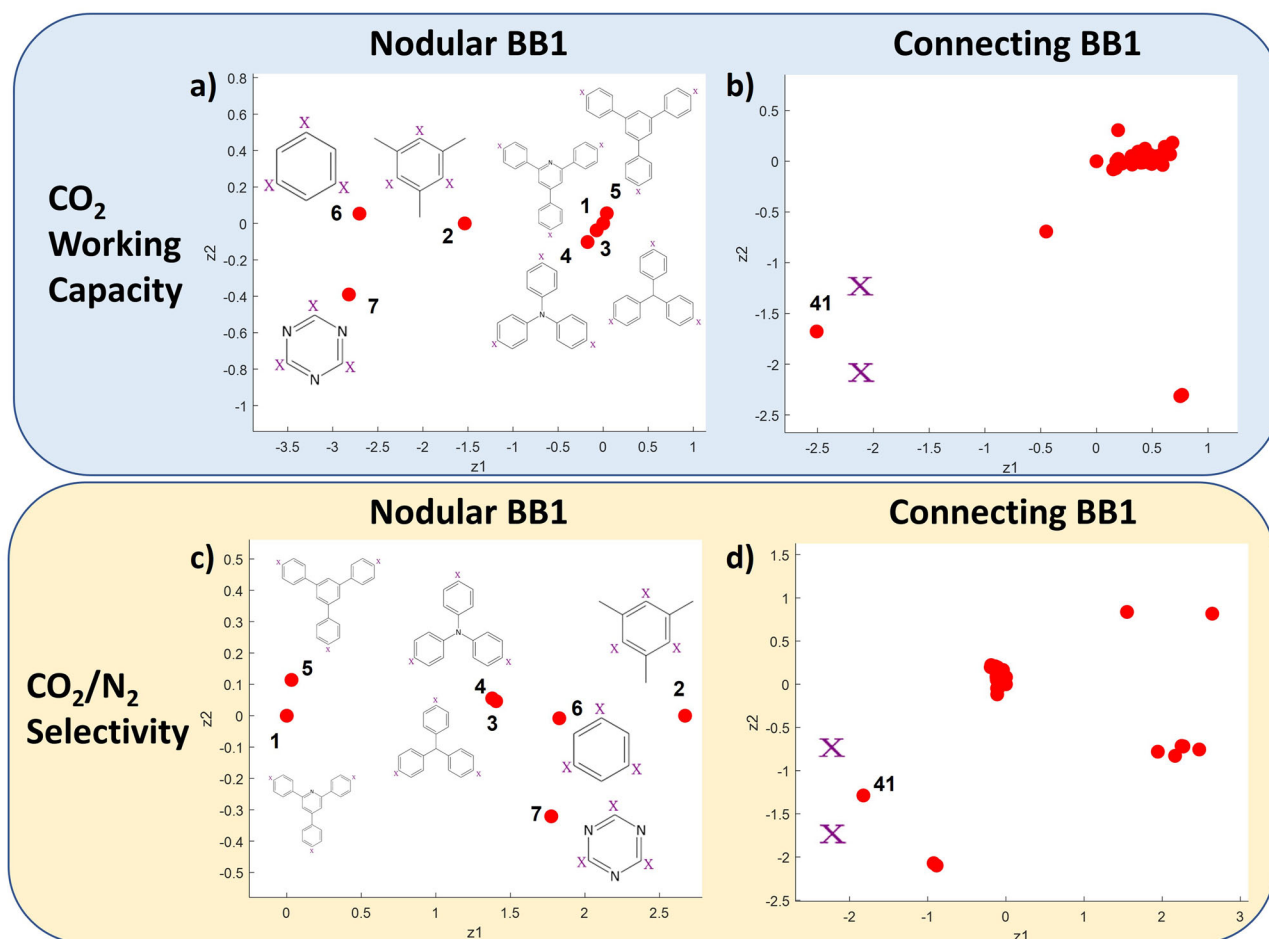
large number of solutions is advantageous due to two potential main reasons. First, it is possible that not every proposed MOF can be synthesized in the laboratory. Second, there are other criteria that must be addressed in practice beyond the  $\text{CO}_2$  working capacity and  $\text{CO}_2/\text{N}_2$  selectivity, such as cost and stability. Thus, it is useful to have alternative promising candidates at hand, so that a practical solution can be found.

By looking further into the histogram of selected building blocks at the end of the optimization campaign (Supplementary Fig. 2), we observed a bias towards particular building blocks. Specifically, for Nodular BB2 and Connecting BB1, the blocks {2} and {41} are favored because all the Pareto front MOFs possess these building blocks. Therefore, LVGP-MOBBO can identify the promising building blocks effectively and choose them as the next MOF designs at the very early steps of the optimization campaign. Due to very low amounts of training data used for design optimization (starting with 0.35% and ending with only ~1% of the design space), the overall predictive capability of LVGP models is somewhat limited for the complete design spectrum. The parity plots and the mean absolute error (MAE), described by the mean of absolute difference between the predicted and the true property values of the remaining MOFs (47,242 MOFs) in the EDS that are unseen to the LVGP model, are shown in Supplementary Fig. 3. The relatively high MAE scores obtained from the LVGP models were expected because this framework is design (objective) oriented. For Bayesian optimization to perform well, the LVGP does not have to be accurate for all design candidates<sup>41</sup>. This is evident in our result because even though LVGP is not an accurate model for global predictive capability, the model is good enough to identify where in the design space to look for the optimal solution. The predictions and prediction uncertainties quantified by the LVGP model are satisfactory in the neighborhood of the optimum design candidates, which led to the promising Pareto solutions observed. The low accuracy in regions of the property space far from the optimum does not have a large effect on the overall performance in identifying top-performing materials.

The interpretability of the LVGP approach can be demonstrated using the results for the entire design space. At the end of the 66 iterations, we observed that the latent spaces of the Nodular BB1 (Fig. 9a, c) and BB2 (Supplementary Fig. 3a, c) converged to a final state. This means that after each iteration, the latent spaces obtained for these design variables did not change. On the other hand, we observed non-convergent latent spaces for Connecting BB1 & BB2 that contain 41 and 42 different design choices, respectively. This is because the LVGP model is trained with a very small percentage of the design space (~1%). The optimization campaign still works well although the latent spaces are not stable. Specifically, block {41} is always separate from the rest in the latent space plots, meaning that its superior effect on the properties is identified clearly. Furthermore, the non-converging behavior is observed for the blocks that have minimal effect on the performance properties. Therefore, our framework can identify the specific building blocks that are superior with a physical justification using the physics-aware LVGP approach.

The latent variable plots of the Nodular BB1 (Fig. 9a, c) and Connecting BB1 (Fig. 9b, d) can also be explained using the MOF textural properties. For Nodular BB1, blocks {1, 3, 4, 5} form a cluster in the latent space of the  $\text{CO}_2$  working capacity, while blocks {2, 6, 7} are spread out (Fig. 9a). A similar trend was observed in the RDS (Fig. 5), which we ascribed to the size of building blocks that determine the LCD of the MOFs. However, the latent space for the  $\text{CO}_2/\text{N}_2$  selectivity changed slightly compared to the RDS. Specifically, blocks {3, 4} are away from blocks {1, 5} and become closer to block {6}, while the positions of the other building blocks remain similar. For Connecting BB1 (Fig. 9b, d), block {41} is distant from the other building blocks, which was also observed in the small dataset. Although some building blocks are also further apart from the clusters, their locations change from one iteration to another. The latent variables for the Nodular BB2 and the Connecting BB2 can be found in Supplementary Fig. 3. The Nodular BB2 plots can be explained by a similar reasoning as for the RDS plots, whereas the Connecting BB2 is non-convergent due to low training percentage of the LVGP.



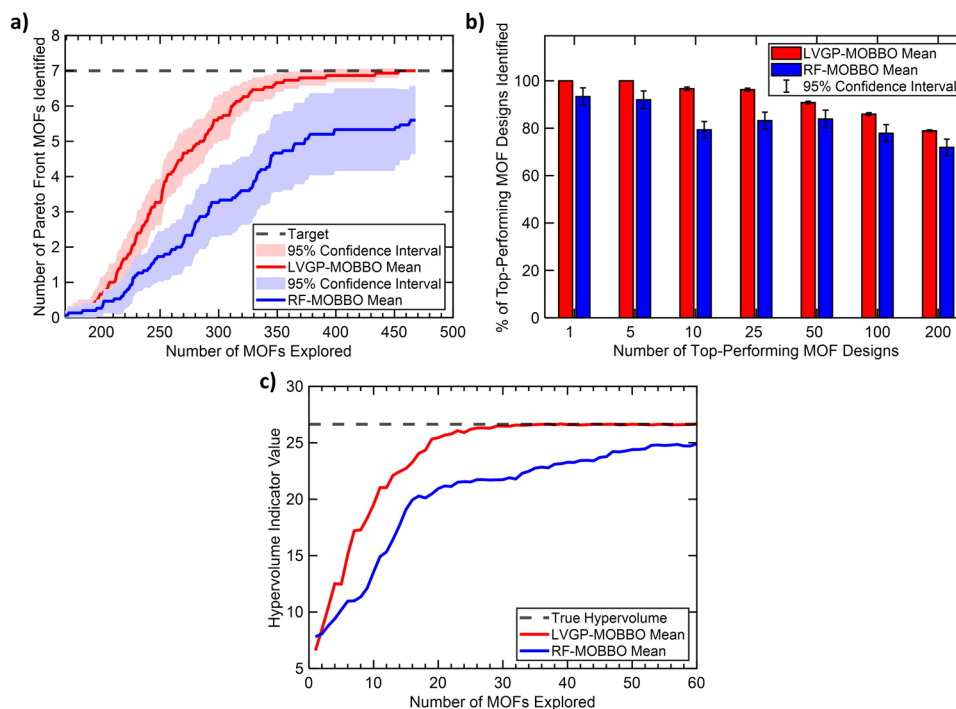


**Fig. 9** Latent variable plots after the LVGP-MOBBO campaign on the Entire Design Space (EDS). **a** Latent spaces of the Nodular BB1 for **(a)** the  $\text{CO}_2$  working capacity and **(c)** the  $\text{CO}_2/\text{N}_2$  selectivity. Latent spaces of Connecting BB1 for **(b)** the  $\text{CO}_2$  working capacity and **(d)** the  $\text{CO}_2/\text{N}_2$  selectivity.

### Comparison with random forest and robustness of LVGP-MOBBO

We compared our LVGP approach with another ML approach, Random Forest (RF), which is also used for optimization problems with qualitative variables. To conduct a fair study between two ML models, the same descriptor-free (featureless) MOF representation is used for the inputs to the RF model. Further details of MOF representation are explained in the *MOF Representation and Database Construction* subsection under the *Methods* section. Both approaches employed the same MOBBO method defined previously. To conduct the study, we ran both frameworks 15 times on the EDS for 60 MOBBO iterations with 15 different initial DOEs. The Pareto front and top-performing MOF design exploration performance of the study can be seen in Fig. 10. We observed that the LVGP can identify all the Pareto front MOF designs whereas the RF approach fails to do so in most cases (Fig. 10a). The small confidence interval in the performance shows that the LVGP approach is robust and reliable in identifying the Pareto front candidates. On the other hand, the confidence interval of RF is large since some of the RF-MOBBO instances fail to identify any Pareto Front MOF designs. This is because RF-MOBBO is stuck in local optimum designs since the algorithm cannot predict beyond the training data, which usually contained initial DOEs with low properties. In contrast, LVGP was able to expand beyond the low property region towards the high property region by its capabilities of extrapolating beyond training data. Moreover, the Bayesian prediction of uncertainty provided by the LVGP

compared to frequentist prediction of RF, leads to better and more effective design space exploration<sup>41</sup>. The LVGP is able to extrapolate uncertainty well, whereas RF fails to do so, as it provides a fixed uncertainty prediction for MOF designs outside of training data due to unavailable tree splitting. More importantly, the LVGP approach makes the correct design decisions at a faster rate compared to the RF approach, which is crucial if the cost of conducting simulations or experiments is very high. Similarly, for all number of top-performing MOF identification categories, the LVGP approach resulted in a better and more robust performance (Fig. 10b). As we already knew both the Pareto front and the top-performing designs beforehand, the two metrics we used could be considered as greedy metrics. In practice, since these designs are typically unknown, a well-known metric used for multi-objective optimization comparison between different algorithms is the hypervolume indicator<sup>42</sup>. Hypervolume indicator provides a scalar metric of how much hypervolume the Pareto front designs obtained by the algorithm dominate the reference point, where the reference point is usually chosen as the nadir point (known lowest values of objectives). Typically, larger values represent better Pareto front designs. Figure 10c shows the mean hypervolume indicator values obtained at each iteration of both methods in 15 different runs. We can easily observe that (i) the explored MOF designs by the LVGP approach span a much larger hypervolume at a much faster rate, and (ii) the true hypervolume can be achieved for all 15 different initializations compared to the RF approach.



**Fig. 10** Comparative study with Random Forest and LVGP-MOBBO. **a** Pareto front and **b** top-performing MOF designs identified by the two methods after 15 different runs. **c** Hypervolume indicator comparison between two methods after 15 different runs.

The advantages provided by the LVGP approach are not limited to design optimization. The physical justification provided by the latent variables make the LVGP an interpretable and a favorable ML model for MOF design. The latent variables obtained at the end of the optimization campaign enabled us to gain physical insights behind the design decisions. Although RF is an explainable ML model, model agnostic methodologies are required to draw conclusions or understand its performance. Finally, since both models aim to perform design exploration in the most efficient way possible, the models are trained on the fly. As a result, we do not have the luxury to perform hyperparameter tuning, as it could require additional property evaluations, which contradicts our goal with efficient materials design optimization. All hyperparameters in LVGP modeling, including both the hyperparameters for regular GP modeling and the locations of different levels of a categorical variable in the latent variable space, are identified through maximum likelihood maximization with the available explored data. On the other hand, RF models could require external hyperparameter tuning through additional property evaluations, such as identifying the number of trees or depth of the trees in the model. This demonstrates another significant advantage of LVGP over RF or other models that require hyperparameter tuning since LVGP does hyperparameter tuning within itself. Thus, together with the better performance and accuracy results, the interpretability and efficiency of LVGP makes our approach more desirable and meaningful for materials design applications.

## DISCUSSION

Due to their versatile and tunable nature, MOFs have very large design spaces, and it is impossible to simulate or perform experiments for every MOF to find the promising candidates for an application of interest. Although numerous ML and high-throughput screening approaches exist, they require either large databases or property-specific descriptors. To tackle these challenges, we presented the LVGP-MOBBO framework to design

superior MOFs by only employing qualitative representations of building blocks. The framework presented here provides three main advantages compared to current similar efforts: (i) the framework requires no specific descriptors and only uses the MOF building blocks to perform the adaptive design space search, (ii) the framework is application independent, meaning that it can be applied to any property without the need to select important descriptors for the application of interest, and (iii) the physically justifiable latent variable approach provides interpretability on how each building block influences the resulting performance properties. We demonstrated our framework on a design space with 47,740 MOF candidates. The LVGP-MOBBO successfully identified all Pareto front designs and more than 97% of the 50 top-performing MOF candidate designs by scanning only ~1% of the design space. Compared to Random Forest, LVGP has better performance and robustness, and provides interpretability regarding the design through physically justifiable latent variables. Finally, although we demonstrated our framework on a MOF design space with adsorption properties, LVGP-MOBBO can be applied to any property that requires time consuming simulations such as quantum mechanical calculations.

A key challenge in the presented framework lies in the high number of building blocks. When a large number of blocks are present, although the design optimization campaign works efficiently to identify the top building blocks and MOF designs, the LVGP model struggles to converge to a final latent space due to high number of parameter (latent variables) estimations during model fitting. We expect that by incorporating prior knowledge, when available, into the framework such as assigning prior known distributions to latent variables, the latent variable realizations can be more accurate<sup>43</sup>. We can also incorporate additional fingerprints, i.e., physical descriptors<sup>44–47</sup>, that can further differentiate the MOF candidates from each other to build more accurate LVGP models, which in return can further improve design optimization. For enhancing the original LVGP method, we developed, in our recent work, an approach that combines both categorical variables and physical descriptors to address the many-level challenge

when using LVGP<sup>48</sup>. However, choosing the right fingerprints that can be uniquely mapped to MOF designs is often application specific and requires additional work. Therefore, due to the formulation of the LVGP modeling, the main goal of this paper was to search for MOF candidate designs that are known a priori. As a result, predicting the properties of MOFs with unseen building blocks through LVGP is a promising area of research. Finally, an interesting application of our framework would involve performing materials design and development through autonomous experimentation studies. As there is no human intervention in LVGP-MOBBO, and the experimental inputs can be both qualitative and quantitative, we envision that the method we presented here can help researchers guide their experiments efficiently.

## METHODS

### MOF representation and database construction

To validate our implementation of LVGP-MOBBO, we created a design space using the **fof** topology, 7 organic nodes (Nodular BB1), 4 inorganic nodes (Nodular BB1), 41 edges (Connecting BB1), and 42 edges (Connecting BB2). The **fof** topology is a derived net of **nbo** topology, in which the tetratopic linker in **nbo** is decomposed into two organic nodes and two edges in **fof**. Further details on **fof** topology were discussed in literature<sup>49</sup>. From the combinations of the building blocks, we created 48,216 hypothetical MOFs. We eliminated 104 MOFs that had poor initial geometries and missing bonds. We performed geometry optimization on the remaining 48,112 MOFs, in which we found and eliminated 372 MOFs that collapsed after the geometry optimization. Therefore, 47,740 MOFs were considered for this study. Among the 47,740 MOFs, at least 8 MOFs were experimentally synthesized (Supplementary Table 2).

Each MOFs with **fof** topology consisted of an inorganic node, an organic node, and two edge blocks. Thus, to represent each MOF we use a 5-element 'vector' representation with integer encoding  $[A - B - C - D - E]$ , where each letter represents Nodular BB1, Nodular BB2, Connecting BB1, Connecting BB2, and Topology, respectively. For each letter, an integer value is assigned to represent a specific choice of building block and the choices of building blocks can be seen in Fig. 2. In this study, we kept the topology as a fixed variable to keep the design space at a reasonable size for comprehensive validation of the method. The letter "E" can be represented with "**fof**" or take the value of 1. A visualization of this representation is shown in Fig. 1.

For the initial set of materials, also known as design of experiments (DOE), that initialize the optimization framework, first an optimal Latin hypercube sample with specified number of samples and variables was created. Then, for each qualitative variable, the design space was sliced into  $p_i$  sections, where  $p_i$  represents the number of unique options for each qualitative variable. Each DOE design is assigned to a qualitative variable that falls under the sliced section. This approach enables us to select initial MOF designs that cover the design space as evenly as possible. An example DOE with two qualitative variables (Nodular BB1 & Nodular BB2) that each have 7 and 4 levels is shown in Fig. 3 under the *Initial Design of Experiments* box.

### MOF construction and geometry optimization

MOFs were created using the topologically based crystal constructor (ToBaCCo 3.0)<sup>36</sup> software. Geometry optimization was carried out to optimize the unit cell parameters and atomic position using LAMMPS<sup>37</sup> with the UFF4MOF<sup>38</sup> force field. For each structure, the geometry optimization was performed in a cycle that consisted of two steps, as recommended by Anderson et al.<sup>36</sup>. The unit cell parameters and atomic positions were first relaxed using a conjugate gradient (CG) algorithm, followed by

atomic position relaxation using the FIRE algorithm (we chose a timestep of 0.1 fs). Each minimization converged only when the change in energy from the previous step to the current step divided by the current energy magnitude was less than  $10^{-8}$  and the forces on atoms were less than  $10^{-8}$  kcal mol<sup>-1</sup> Å. The cycle stopped when the change in energy between the previous cycle and the current cycle was less than  $10^{-8}$  kcal mol<sup>-1</sup>.

### GCMC simulations

Grand canonical Monte Carlo (GCMC) simulation was carried out using the RASPA package<sup>39</sup>. Each simulation consisted of 10,000 equilibration cycles and 10,000 production cycles. The Monte Carlo moves used were translation, rotation, insertion, deletion, and random reinsertion. Lennard-Jones (LJ) and Coulombic interactions were used to calculate the energies between non-bonded atoms. LJ parameters between different atom types were computed using the Lorentz-Berthelot mixing rules. CO<sub>2</sub> and N<sub>2</sub> were modeled as three-site rigid molecules with charges on each site, using the LJ parameters and partial charges from the TraPPE force field<sup>50</sup>. LJ parameters for the framework atoms were from the Universal Force Field (UFF)<sup>51</sup>. Previous studies have shown that using UFF parameters for framework atoms can adequately demonstrate the interaction between MOFs and various adsorbates<sup>52–57</sup>. The framework atom partial charges were calculated using the PACMOF (Partial Atomic Charges in Metal-Organic Frameworks) software<sup>58</sup>. For each MOF, we carried out two GCMC simulations; the first was at the adsorption condition of 1 bar, 313 K, and a bulk molar composition of CO<sub>2</sub>: N<sub>2</sub> = 0.15 : 0.85, and the second was at the desorption condition of 0.1 bar, 313 K, and a bulk molar composition of CO<sub>2</sub> : N<sub>2</sub> = 0.9 : 0.1.

We used the CO<sub>2</sub> working capacity ( $\Delta N_{\text{CO}_2}$ ) and the CO<sub>2</sub>/N<sub>2</sub> selectivity at adsorption ( $\alpha_{\text{CO}_2/\text{N}_2}^{\text{ads}}$ ) as the criteria to determine top-performing MOFs for CO<sub>2</sub>/N<sub>2</sub> separation. The two properties are defined as follows:

$$\Delta N_{\text{CO}_2} = N_{\text{CO}_2}^{\text{ads}} - N_{\text{CO}_2}^{\text{des}} \quad (1)$$

$$\alpha_{\text{CO}_2/\text{N}_2}^{\text{ads}} = \frac{N_{\text{CO}_2}^{\text{ads}} y_{\text{N}_2}^{\text{ads}}}{N_{\text{N}_2}^{\text{ads}} y_{\text{CO}_2}^{\text{ads}}} \quad (2)$$

Here,  $\Delta N_{\text{CO}_2}$  is the CO<sub>2</sub> working capacity,  $N_{\text{CO}_2}^{\text{ads}}$  and  $N_{\text{CO}_2}^{\text{des}}$  are the CO<sub>2</sub> adsorption loadings at the adsorption and desorption conditions,  $\alpha_{\text{CO}_2/\text{N}_2}^{\text{ads}}$  is the selectivity of CO<sub>2</sub> over N<sub>2</sub> at adsorption condition,  $N_{\text{N}_2}^{\text{ads}}$  is the N<sub>2</sub> loading at adsorption, and  $y_{\text{N}_2}^{\text{ads}}$  and  $y_{\text{CO}_2}^{\text{ads}}$  are the bulk mole fractions of N<sub>2</sub> and CO<sub>2</sub> at adsorption, respectively. While CO<sub>2</sub> working capacity reflects how effective the MOF is at both capturing and releasing CO<sub>2</sub>, the selectivity determines how selectively the MOF can separate CO<sub>2</sub> from the mixture of CO<sub>2</sub> and N<sub>2</sub>.

### Latent variable Gaussian process (LVGP)

One of the main contributions of this paper lies in the design optimization of MOFs using only the readily available qualitative representations of building blocks. On the other hand, due to the nature of the correlation functions, it is not possible to directly implement the building blocks into the Gaussian Process (GP) models as the difference between variables becomes unclear. Therefore, in this paper, we implemented the Latent Variable Gaussian Process (LVGP) to account for the qualitative variables in the GP model<sup>31</sup>. It is known that for every qualitative variable, there are underlying, possibly high-dimensional, quantitative variables that explain its effect on properties. The latent variable approach helps us to map the qualitative variables to a quantitative latent space. Consider a GP model input with  $t = [t_1^q, t_2^q, \dots, t_n^q] \in R^{q \times n}$  with  $n$  qualitative variables and  $q$  number of points, where each point here represents a unique MOF design. Each variable,  $t_i$ , has  $p_i$  unique levels (design choices)

$\{I_1(t_i), I_2(t_i), \dots, I_{p_i}(t_i)\}$  for  $i = 1 : n$ , (e.g., Cu, Co, Ni options for nodular building block ( $p_i = 3$ )). Then, each qualitative variable can be represented with a latent variable vector  $z_t(I_{p_i}) = \{z_t^1(I_{p_i}), \dots, z_t^k(I_{p_i})\}$  for  $z^i \in \mathbb{R}^k$ . The developers of the algorithm have stated that users are free to choose the dimensions of  $z^i$  but also demonstrated that  $k = 2$  is enough to represent the underlying high-dimensional quantitative space. Consequently, we chose  $k = 2$ . Thus, each level within a qualitative variable can be represented with a latent vector of  $z_t(I_{p_i}) = \{z_t^1(I_{p_i}), z_t^2(I_{p_i})\}$ , and the input to the GP model becomes  $z(t) = [z_t^1(I_{p_1}), z_t^2(I_{p_1}), \dots, z_t^1(I_{p_n}), z_t^2(I_{p_n})] \in \mathbb{R}^{q \times 2n}$ . An illustration of the latent variable representation of qualitative building blocks is shown in Fig. 3 under the *Latent Variable Gaussian Process* box. Consider a typical single response Gaussian Process model, which consists of prior constant mean  $\mu$  and  $K_Z(t)$ , describing the mean response at any given point in the input space, and a zero-mean Gaussian Process with a covariance function  $K(t, t')$ , respectively. The covariance function  $K(t, t')$  determines the relationship or the correlation between variables in the model. The covariance function can be further extended to  $K(t, t') = \sigma^2 \cdot c(t, t')$  where the  $\sigma^2$  represents the prior variance of the GP model and  $c(t, t')$  describes the correlation between each point in the model through the specified correlation function. To explain the relationship between each design candidate for this application, we have implemented the Gaussian correlation function:

$$c(t, t') = \exp \left\{ - \sum_{i=1}^p \left( \|z_t^1(I_{p_i}) - z_{t'}^1(I_{p_i})\|_2^2 + \|z_t^2(I_{p_i}) - z_{t'}^2(I_{p_i})\|_2^2 \right) \right\} \quad (3)$$

The Gaussian correlation function shown in Eq. (3) evaluates the correlation between points  $t$  and  $t'$  based on 2-norm distance. The main reason behind choosing this correlation function is because we assume that points that are close in the spatial input space should also reflect a similar behavior in the output space as well. Along with the 2D mapped latent variables  $z = (z_t^1(I), z_t^2(I))$  for level  $l$  of each qualitative variable  $t$ , the parameters,  $\mu$  and  $\sigma$  are estimated through Maximum Likelihood Estimation (MLE) of the log-likelihood function

$$l(\mu, \sigma, z) = -\frac{q}{2} \ln(\sigma) - \frac{1}{2} \ln|C(z)| - \frac{1}{2\sigma^2} (\mathbf{y} - \mu \mathbf{1})^T C(z)^{-1} (\mathbf{y} - \mu \mathbf{1}) \quad (4)$$

where  $q$  is the number of samples,  $C$  is the  $q \times q$  correlation matrix with  $C_{ij} = c(t^i, t^j)$  for  $i, j = 1, 2, 3, \dots, q$ ,  $\mathbf{1}$  is a vector of ones with dimensions of  $q \times 1$ , and  $\mathbf{y}$  is the observed response with dimensions of  $q \times 1$ . Finally, the 2D quantitative latent variables are then used to construct a GP model that provides both prediction and statistical representation of uncertainty in the design space for Bayesian optimization.

### Multi-objective batch bayesian optimization (MOBBO)

Bayesian Optimization is a well-known efficient, fast, and easy-to-implement optimization technique that has been used in numerous materials design applications. For single objective optimization, BO makes the decision on which design in the design space should be sampled next based on the choice of an acquisition function. Three well-known acquisition functions are Lower Confidence Bound<sup>59</sup>, Probability of Improvement<sup>60</sup>, and Expected Improvement (EI)<sup>61</sup>. With its ability to balance exploration of the design space and exploitation of the objective, EI has been a popular choice for most materials design applications. Considering the large MOF design space, we have also chosen EI as our base acquisition function. For a given candidate design  $x$ , with its predicted objective value  $y'_x$  and quantified uncertainty  $\sigma_x$  from the LVGP model, EI for single objective optimization can be

calculated using,

$$EI(x) = (y^* - y'_x) \cdot \psi \left( \frac{(y^* - y'_x)}{\sigma_x} \right) + \sigma_x \cdot \phi \left( \frac{(y^* - y'_x)}{\sigma_x} \right) \quad (5)$$

where  $y^*$  is the best observed objective so far in the optimization campaign and  $\psi, \phi$  represent the cumulative distribution function (CDF) and probability distribution function (PDF), respectively. As Eq. (5) shows, the EI function suggests a new design by not only considering the exploitation of the objective function,  $(y^* - y'_x)$ , but also the uncertainties associated with the design space,  $\sigma_x$ .

Often there are tradeoffs between objectives, meaning that one objective cannot be optimized without sacrificing the other one. This type of problem is also known as Pareto front optimization and is frequently observed in material systems<sup>34</sup>. Thus, for multi-objective optimization problems, the goal becomes discovering the Pareto front of the property space. Therefore, we have expanded the EI formulation by implementing the Expected-Maximin Improvement (EMI) acquisition function to serve as the balancer of the exploration and exploitation for multi-objective optimization. For the case of optimizing two objectives, the formulation of EMI is

$$EMI(x_k) = \min(\max(EI_1, EI_2), 0) \quad (6)$$

where  $EI_j$  corresponds to the Expected Improvement value of each objective  $j$ . The EI formula was used to compare the candidates with respect to the observed number of  $p$  Pareto front designs so far in the optimization campaign. Therefore, each  $EI_j$  is a vector of  $p \times 1$  that contains the EI values of a candidate design on the observed Pareto front designs for each objective. Lastly, we first take the maximum of EI's for both objectives to observe the dominance of the candidate on the current Pareto frontier and then select the minimum of the maximum EIs to balance the multi-objective search. As a result, the EMI is formulated in a way that both objectives have equal importance. Eq. (6) selects the single best candidate in each multi-objective BO iteration.

Due to the large number of candidate designs and the cost of training GP models, it is not ideal to train the LVGP with a single design candidate at each iteration. Therefore, to extend single candidate BO to select a batch of promising candidates, we select  $B$  candidates that possess the highest EMI values in each iteration and use them as the next design candidates. A demonstration of a single MOBBO iteration is demonstrated in Fig. 3 under the *Multi-Objective Batch Bayesian Optimization* box.

### DATA AVAILABILITY

The crystal structures and calculated properties of 47,740 MOFs in this study are deposited on Zenodo. (<https://doi.org/10.5281/zenodo.7951588>).

### CODE AVAILABILITY

The LVGP-MOBBO code used to carry out this work are described in the Methods section. The MATLAB codes used in this study for the LVGP-MOBBO framework are provided at <https://github.com/ideal-nu/MOF-LVGP-MOBBO>. For interested readers, the LVGP-code can be also accessed through the Comprehensive R Archive Network (CRAN) at <https://cran.r-project.org/package=LVGP>.

Received: 10 February 2023; Accepted: 11 September 2023;  
Published online: 21 September 2023

### REFERENCES

- Moosavi, S. M., Jablonka, K. M. & Smit, B. The role of machine learning in the understanding and design of materials. *J. Am. Chem. Soc.* **142**, 20273–20287 (2020).
- Ramprasad, R., Batra, R., Pilia, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: Recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).

3. Jablonka, K. M., Ongari, D., Moosavi, S. M. & Smit, B. Big-data science in porous materials: Materials genomics and machine learning. *Chem. Rev.* **120**, 8066–8129 (2020).
4. Suh, M. P., Park, H. J., Prasad, T. K. & Lim, D.-W. Hydrogen storage in metal–organic frameworks. *Chem. Rev.* **112**, 782–835 (2012).
5. He, Y., Zhou, W., Qian, G. & Chen, B. Methane storage in metal–organic frameworks. *Chem. Soc. Rev.* **43**, 5657–5678 (2014).
6. Li, H. et al. Recent advances in gas storage and separation using metal–organic frameworks. *Mater. Today* **21**, 108–121 (2018).
7. Shah, M., McCarthy, M. C., Sachdeva, S., Lee, A. K. & Jeong, H.-K. Current status of metal–organic framework membranes for gas separations: Promises and challenges. *Ind. Eng. Chem. Res.* **51**, 2179–2199 (2012).
8. Roohollahi, H., Zeinalzadeh, H. & Kazemian, H. Recent advances in adsorption and separation of methane and carbon dioxide greenhouse gases using metal–organic framework-based composites. *Ind. Eng. Chem. Res.* **61**, 10555–10586 (2022).
9. Li, J.-R., Kuppler, R. J. & Zhou, H.-C. Selective gas adsorption and separation in metal–organic frameworks. *Chem. Soc. Rev.* **38**, 1477–1504 (2009).
10. Wang, Q. & Astruc, D. State of the art and prospects in metal–organic framework (mof)-based and mof-derived nanocatalysis. *Chem. Rev.* **120**, 1438–1511 (2020).
11. Wei, Y.-S., Zhang, M., Zou, R. & Xu, Q. Metal–organic framework-based catalysts with single metal sites. *Chem. Rev.* **120**, 12089–12174 (2020).
12. Freund, R. et al. The current status of mof and cof applications. *Angew. Chem. Int. Ed.* **60**, 23975–24001 (2021).
13. Li, J.-R. et al. Carbon dioxide capture-related gas adsorption and separation in metal–organic frameworks. *Coord. Chem. Rev.* **255**, 1791–1823 (2011).
14. Sumida, K. et al. Carbon dioxide capture in metal–organic frameworks. *Chem. Rev.* **112**, 724–781 (2012).
15. Avci, G., Velioglu, S. & Keskin, S. High-throughput screening of mof adsorbents and membranes for H<sub>2</sub> purification and CO<sub>2</sub> capture. *ACS Appl. Mater. Interfaces* **10**, 33693–33706 (2018).
16. Altintas, C. et al. An extensive comparative analysis of two mof databases: High-throughput screening of computation-ready mofs for CH<sub>4</sub> and H<sub>2</sub> adsorption. *J. Mater. Chem. A* **7**, 9593–9608 (2019).
17. Gu, C., Yu, Z., Liu, J. & Sholl, D. S. Construction of an anion-pillared mof database and the screening of mofs suitable for Xe/Kr separation. *ACS Appl. Mater. Interfaces* **13**, 11039–11049 (2021).
18. Li, S., Chung, Y. G. & Snurr, R. Q. High-throughput screening of metal–organic frameworks for CO<sub>2</sub> capture in the presence of water. *Langmuir* **32**, 10368–10376 (2016).
19. Colón, Y. J. & Snurr, R. Q. High-throughput computational screening of metal–organic frameworks. *Chem. Soc. Rev.* **43**, 5735–5749 (2014).
20. Islamov, M. et al. High-throughput screening of hypothetical metal–organic frameworks for thermal conductivity. *npj Comput. Mater.* **9**, 11 (2023).
21. Lee, S. et al. Computational screening of trillions of metal–organic frameworks for high-performance methane storage. *ACS Appl. Mater. Interfaces* **13**, 23647–23654 (2021).
22. Park, J., Lim, Y., Lee, S. & Kim, J. Computational design of metal–organic frameworks with unprecedented high hydrogen working capacity and high synthesizability. *Chem. Mater.* **35**, 9–16 (2023).
23. Yao, Z. et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.* **3**, 76–86 (2021).
24. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
25. Wei, J. et al. Machine learning in materials science. *InfoMat* **1**, 338–358 (2019).
26. Guo, K., Yang, Z., Yu, C.-H. & Buehler, M. J. Artificial intelligence and machine learning in design of mechanical materials. *Mater. Horiz.* **8**, 1153–1172 (2021).
27. Wang, K. & Dowling, A. W. Bayesian optimization for chemical products and functional materials. *Curr. Opin. Chem. Eng.* **36**, 100728 (2022).
28. Frazier, P. I. & Wang, J. Bayesian optimization for materials design in *Information science for materials discovery and design* (eds Lookman, T., Alexander, F. J., & Rajan, K.) 45–75 (Springer International Publishing, 2016).
29. Taw, E. & Neaton, J. B. Accelerated discovery of CH<sub>4</sub> uptake capacity metal–organic frameworks using bayesian optimization. *Adv. Theor. Simul.* **5**, 2100515 (2022).
30. Deshwal, A., Simon, C. M. & Doppa, J. R. Bayesian optimization of nanoporous materials. *Mol. Syst. Des. Eng.* **6**, 1066–1086 (2021).
31. Zhang, Y., Tao, S., Chen, W. & Apley, D. W. A latent variable approach to gaussian process modeling with qualitative and quantitative factors. *Technometrics* **62**, 291–302 (2020).
32. Zhang, Y., Apley, D. W. & Chen, W. Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Sci. Rep.* **10**, 4924 (2020).
33. Iyer, A. et al. Data centric nanocomposites design via mixed-variable bayesian optimization. *Mol. Syst. Des. Eng.* **5**, 1376–1390 (2020).
34. Censor, Y. Pareto optimality in multiobjective problems. *Appl. Math. Optim.* **4**, 41–59 (1977).
35. Ba, S., Myers, W. R. & Breneman, W. A. Optimal sliced latin hypercube designs. *Technometrics* **57**, 479–487 (2015).
36. Anderson, R. & Gómez-Gualdrón, D. A. Increasing topological diversity during computational “synthesis” of porous crystals: How and why. *CrystEngComm* **21**, 1653–1665 (2019).
37. Thompson, A. P. et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271**, 108171 (2022).
38. Coupury, D. E., Addicoat, M. A. & Heine, T. Extension of the universal force field for metal–organic frameworks. *J. Chem. Theory Comput.* **12**, 5215–5225 (2016).
39. Dubbeldam, D., Calero, S., Ellis, D. E. & Snurr, R. Q. Raspa: Molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol. Simul.* **42**, 81–101 (2016).
40. Wang, Y., Iyer, A., Chen, W. & Rondinelli, J. M. Featureless adaptive optimization accelerates functional electronic materials design. *Appl. Phys. Rev.* **7**, 041403 (2020).
41. Zhang, H., Chen, W., Iyer, A., Apley, D. W. & Chen, W. Uncertainty-aware mixed-variable machine learning for materials design. *Sci. Rep.* **12**, 19760 (2022).
42. Jablonka, K. M., Jothiappan, G. M., Wang, S., Smit, B. & Yoo, B. Bias free multi-objective active learning for materials design and discovery. *Nat. Commun.* **12**, 2312 (2021).
43. Yerramilli, S., Iyer, A., Chen, W. & Apley, D. W. Fully bayesian inference for latent variable gaussian process models. *arXiv preprint at Preprint at arxiv.org/abs/2211.02218* (2022).
44. Shi, K. et al. Two-dimensional energy histograms as features for machine learning to predict adsorption in diverse nanoporous materials. *J. Chem. Theory Comput.* **19**, 4568–4583 (2023).
45. Kang, Y., Park, H., Smit, B. & Kim, J. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nat. Mach. Intell.* **5**, 309–318 (2023).
46. Boyd, P. G., Lee, Y. & Smit, B. Computational development of the nanoporous materials genome. *Nat. Rev. Mater.* **2**, 17037 (2017).
47. Nandy, A. et al. Mofsimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks. *Sci. Data* **9**, 74 (2022).
48. Iyer, A., Yerramilli, S., Rondinelli, J. M., Apley, D. W. & Chen, W. Descriptor aided bayesian optimization for many-level qualitative variables with materials design applications. *J. Mech. Des.* **145**, 031701 (2022).
49. Bucior, B. J. et al. Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. *Cryst. Growth Des.* **19**, 6682–6697 (2019).
50. Potoff, J. J. & Siepmann, J. I. Vapor–liquid equilibria of mixtures containing alkanes, carbon dioxide, and nitrogen. *AIChE J.* **47**, 1676–1682 (2001).
51. Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A. III & Skiff, W. M. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).
52. Gong, W. et al. Creating optimal pockets in a clathrocholate-based metal–organic framework for gas adsorption and separation: Experimental and computational studies. *J. Am. Chem. Soc.* **144**, 3737–3745 (2022).
53. García-Holley, P. et al. Benchmark study of hydrogen storage in metal–organic frameworks under temperature and pressure swing conditions. *ACS Energy Lett.* **3**, 748–754 (2018).
54. Wilmer, C. E. et al. Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* **4**, 83–89 (2012).
55. Anderson, R., Schweitzer, B., Wu, T., Carreon, M. A. & Gómez-Gualdrón, D. A. Molecular simulation insights on Xe/Kr separation in a set of nanoporous crystalline membranes. *ACS Appl. Mater. Interfaces* **10**, 582–592 (2018).
56. Chen, Z. et al. Balancing volumetric and gravimetric uptake in highly porous materials for clean energy. *Science* **368**, 297–303 (2020).
57. Polat, H. M., Zeeshan, M., Uzun, A. & Keskin, S. Unlocking CO<sub>2</sub> separation performance of ionic liquid/cubtc composites: Combining experiments with molecular simulations. *Chem. Eng. J.* **373**, 1179–1189 (2019).
58. Kancharlapalli, S., Gopalan, A., Haranczyk, M. & Snurr, R. Q. Fast and accurate machine learning strategy for calculating partial atomic charges in metal–organic frameworks. *J. Chem. Theory Comput.* **17**, 3052–3064 (2021).
59. Zheng, J., Li, Z., Gao, L. & Jiang, G. A parameterized lower confidence bounding scheme for adaptive metamodel-based design optimization. *Eng. Comput.* **33**, 2165–2184 (2016).
60. Couckuyt, I., Deschrijver, D. & Dhaene, T. Fast calculation of multiobjective probability of improvement and expected improvement criteria for pareto optimization. *J. Glob. Optim.* **60**, 575–594 (2014).
61. Jones, D. R. A taxonomy of global optimization methods based on response surfaces. *J. Glob. Optim.* **21**, 345–383 (2001).

## ACKNOWLEDGEMENTS

W.C. acknowledges support from the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0001209, and National Science Foundation (NSF) Grant 1729743. R.Q.S. acknowledges support from the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences under Award DE-FG02-17ER16362. T.D.P. acknowledges computing support from Quest high performance computing facility at Northwestern University. T.D.P. acknowledges computing support from the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC award BES-ERCAP0020094.

## AUTHOR CONTRIBUTIONS

Y.C. and W.C. conceived and designed the project. Y.C. led the collaboration, created the LVGP-MOBBO framework, performed the optimization, and analyzed the results. T.D.P. carried out the GCMC simulations and analyzed the results. Y.C. and T.D.P. wrote the manuscript. R.Q.S. and W.C. supervised the work. All authors (Y.C., T.D.P., R.Q.S., and W.C.) reviewed and edited the manuscript.

## COMPETING INTERESTS

R.Q.S. has a financial interest in the start-up company NuMat Technologies, which is seeking to commercialize metal–organic frameworks. The remaining authors declare no competing financial interests, and all authors declare no competing non-financial interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41524-023-01125-1>.

**Correspondence** and requests for materials should be addressed to Wei Chen.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023