CHEMISTRY

Accelerating computational discovery of porous solids through improved navigation of energy-structure-function maps

Edward O. Pyzer-Knapp[1]*, Linjiang Chen[2], Graeme M. Day[3], Andrew I. Cooper[2]

While energy-structure-function (ESF) maps are a powerful new tool for in silico materials design, the cost of acquiring an ESF map for many properties is too high for routine integration into high-throughput virtual screening workflows. Here, we propose the next evolution of the ESF map. This uses parallel Bayesian optimization to selectively acquire energy and property data, generating the same levels of insight at a fraction of the computational cost. We use this approach to obtain a two orders of magnitude speedup on an ESF study that focused on the discovery of molecular crystals for methane capture, saving more than 500,000 central processing unit hours from the original protocol. By accelerating the acquisition of insight from ESF maps, we pave the way for the use of these maps in automated ultrahigh-throughput screening pipelines by greatly reducing the opportunity risk associated with the choice of system to calculate.

INTRODUCTION

In principle, the combination of machine learning and virtual computational screening is a powerful method for the discovery of new functional organic materials (*1*, *2*). Computational techniques show great promise for the calculation of both the thermodynamic stability and the associated functional properties of candidate materials, but it is difficult in practice to exploit these methods across a broad range of problems. A central challenge is the prohibitive computational expense of accurately calculating energies and properties for every candidate material that is to be screened, and machine learning may provide notable benefit here.

One of the most challenging cases is the a priori design of functional molecular organic crystals with desirable materials properties. Unlike their framework-based counterparts, such as zeolites and Metal Organic Frameworks (MOFs) (*3*–*5*), molecular crystals rarely obey simple geometric principles that can be exploited for rational design. Even very small changes to molecular structure can have marked effects on crystal packing and, hence, the resultant solid-state properties. Molecular crystal packing is often dictated by weak, competing intermolecular interactions: Hence, the a priori design of materials with predetermined, desirable properties requires a more subtle approach than for materials where structure (and hence function) can be "built-in" through the use of intuitive bonding rules, such as adherence to known framework topologies or other geometric bonding principles.

Energy-structure-function maps

Energy-structure-function (ESF) maps are a combination of crystal structure prediction (CSP) with per-structure property calculation, which has been shown to be a powerful tool for the virtual screening of candidate organic molecules for desirable properties such as natural gas storage capacity (*6*) and charge carrier mobility (*7*). In an ESF map, candidate crystal structures are generated using CSP methodologies, which are then screened virtually for a desired property. The resulting pairing of lattice energy and function is then used as an indicative tool for the propensity of the molecule to express the desired properties. This information can be used to guide an experimental campaign, which has been used to validate this ESF map approach (*6*, *8*). However, while this strategy can be effective, generation of the ESF map can be computationally intensive. For example, for methane storage predictions (*6*), it took around 800,000 central processing unit (CPU) hours to compute an ESF map for only one of the molecules in the study (T2E), and this computational cost was distributed roughly equally between the CSP and property calculations. The cost of computing ESF maps grows as the property of interest becomes more computationally expensive and also when the ESF maps contain larger numbers of candidate structures; this is particularly problematic for porous materials, where the energy range that includes all observable crystal structures is extended by solvent templating. Multiple components (e.g., cocrystals) and multiple stable molecular conformers also increase the dimensionality of the energy landscape markedly (*8*, *9*).

Bayesian optimization

Bayesian optimization (*10*) is a technique for evaluating a so-called black box function; that is to say, a function for which there is not access to the analytical, closed form

$$\min_{x \in X} f(x)$$

Bayesian optimization has become popular recently in the machine learning community for the efficient tuning of the hyperparameters of deep learning models (*11*), but given its strengths as a global optimizer and its powerful theoretical guarantees (*12*), it has also started to find applications in a more diverse set of domains (*13*–*16*). The core application area of Bayesian optimization is when each sample of the function, f, is expensive to acquire in financial cost, acquisition time, or both. This makes this approach highly attractive for our goal of more efficiently navigating large ESF maps.

Bayesian optimization has two fundamental principles. First, it promotes the use of a surrogate function, $\hat{f}$, to represent the true

[1]IBM Research Europe, Hartree Centre, Sci-Tech Daresbury, Warrington, UK. [2]Leverhulme Research Centre for Functional Materials Design, Department of Chemistry and Materials Innovation Factory, University of Liverpool, Liverpool, UK. [3]School of Chemistry, University of Southampton, Southampton, UK.
*Corresponding author. Email: epyzerk3@uk.ibm.com

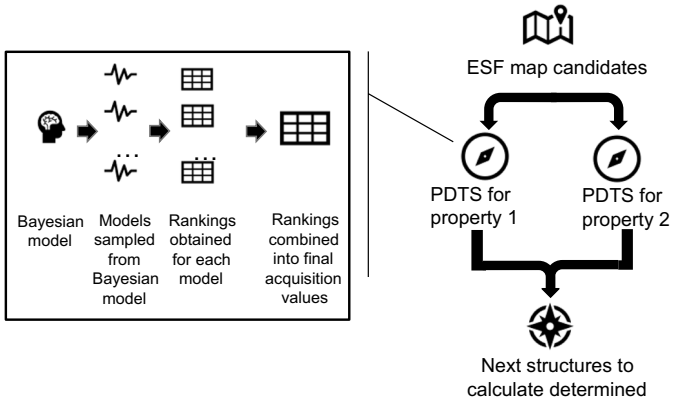

Fig. 1. Graphical illustration of the Bayesian optimization framework used in this study.

(unknown) function, f, that is being optimized. Since each data point is likely to be expensive to acquire, it is important that this surrogate function has robust and well-defined uncertainties associated with its evaluation. In this study, this model is a Gaussian process (*17*), although other models have been used (*18*, *19*).

A Gaussian process is a nonparametric machine learning model, which can be described by a Normal distribution $\mathcal{N}$ with mean function, μ, and a kernel function, $K(x, x')$

$$p(\boldsymbol{f} \mid X) = \mathcal{N}(\mu, K(x, x'))$$

where $p(\boldsymbol{f} \mid X)$ is the probability of $\boldsymbol{f}$ given X, and $\boldsymbol{f}$ is the vector of function values $[f(x_1), f(x_2)\ldots f(x_N)]$ evaluated at input points x_1, $x_2\ldots x_N$. There are many potential choices for the kernel function $K(x, x')$ and, for this study, we used a Matérn kernel (*20*)

$$C_{\frac{3}{2}} = \sigma^2\left(1 + \frac{\sqrt{3}}{l}\right)\exp\left(-\frac{\sqrt{3}}{l}\right)$$

where the length scale l is determined on a per-feature basis using the automatic relevance determination (*21*) protocol and σ^2 is the signal SD. We also introduce a white noise kernel, whose scale is determined as a hyperparameter of the overall Gaussian process and tuned to maximize the log-likelihood of the model with respect to the data.

The second major principle of Bayesian optimization is to balance exploration (the acquisition of new knowledge) and exploitation (the reliance on existing knowledge) when deciding which data points to acquire (*22*). This takes advantage of the existence of the uncertainties associated with the evaluations of the surrogate function, $\hat{f}$, and is controlled through a construct known as the acquisition function. There are a number of potential acquisition functions, with the most popular being expected improvement (EI) (*23*), which aims to maximize the EI to the optimization of collecting a data point. While EI is seemingly a serial methodology, there have been strategies implemented recently that generalize to the parallel setting (*18*, *24*–*27*). Typically, these do not scale well with the number of dimensions and those which do require sparsity and incoherence properties of the feature space that are not present in this problem (*26*). Thompson sampling (*28*) solves this problem by approximating the predictive distribution as follows

$$p(y_j \mid x_j, \mathcal{D}_{\mathfrak{T}}) = \int p(y_j \mid x_j, \theta)\, p(\theta, \mathcal{D}_{\mathfrak{T}})\, d\theta$$

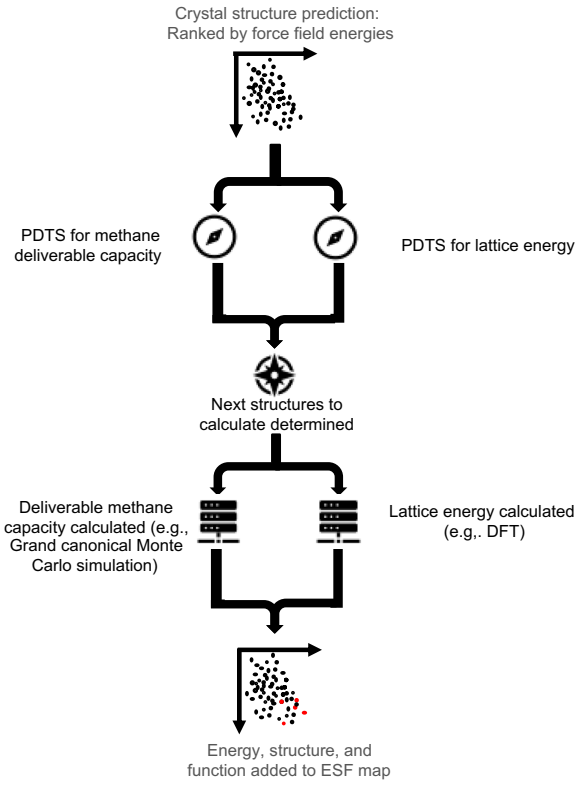

Fig. 2. Flowchart representing the use of MO-PDTS for accelerating ESF map construction. Note that, in some cases, a sufficiently accurate value for lattice energy is calculated at the initial generation stage, and, in these cases, calculation of lattice energy is not necessary, but a second, optional calculation at a higher level may also be used.

where $p(\theta)$ represents the prior distribution given a set of data $\mathcal{D}_{\mathfrak{T}}$, thus approximating the posterior distribution using Monte Carlo, based on a single sample from $p(\theta, \mathcal{D}_{\mathfrak{T}})$. This method thus scales significantly better with the scale and dimensionality of the problem.

The use of Thompson sampling for parallel Bayesian optimization requires an adaptation of this methodology known as parallel and distributed Thompson sampling (PDTS) (*18*), which is described visually in the inset of Fig. 1 and extensively in pseudo-code in the electrospray ionization. PDTS extends the Thompson sampling framework to a parallel case, exploiting the fact that PDTS with batch size S is the same as running sequential Thompson sampling S times without updating the current posterior. This allows the parallel and distributed calculation of the acquisition function, ensuring that this method is highly scalable with increasing batch size. This is particularly important in the case described here since it allows for evaluations to be distributed over a cluster computer system or even over a completely distributed system, such as IBM's World Community Grid (*29*), which harnesses the power of volunteer compute by harvesting "idle" cycles from volunteer devices such as laptops, small computational systems, or even mobile devices.

In this study, we further extend the use of PDTS to the multiobjective case (MO-PDTS) without harming the scalability and thus the parallel performance. To achieve this, we assign a separate PDTS sampler to each objective, the acquisition functions of which

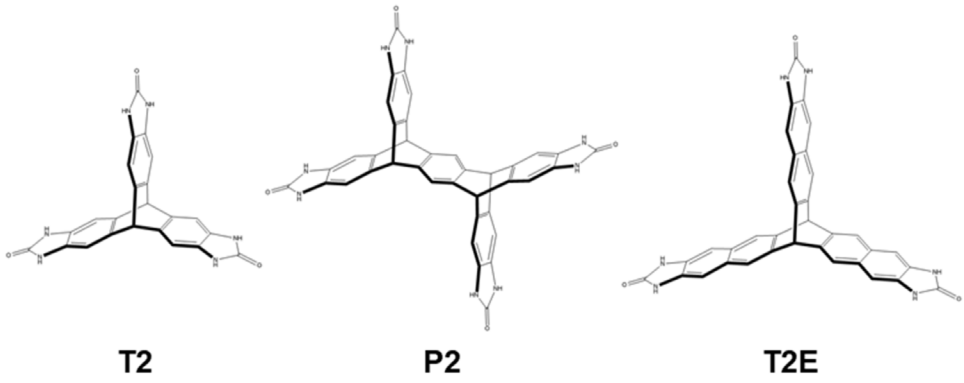

Fig. 3. Chemical structures of the three molecules in this study.

are then combined in a single step, determining the final acquisition function for the overall optimization process. Under a Gaussian process prior, this combination is equivalent to optimizing a single objective consisting of a weighted combination of objective values, with one significant advantage. Since the acquisition values are distinct from the models used to predict them in our MO-PDTS setting, each sampler can be built from a completely different set of descriptors. Under the reasonable assumption that a model built from specifically chosen descriptors is more likely to have strong predictive ability than one built from a general set of descriptors, the ability to separate the predictors affords the user a framework that is significantly more transferable across a range of property types. The ability to fully distribute this calculation is maintained because the domain over which this optimization is performed is a discrete set of structures.

RESULTS AND DISCUSSION

We investigated the extent of MO-PDTS acceleration on three ESF maps that were calculated to evaluate the potential of three molecular materials for methane storage and delivery. We demonstrate how this new navigation workflow (Fig. 2) would have reduced the necessary computation and resulting time to insight for three systems, i.e., T2, T2E, and P2 (Fig. 3), recently predicted to have stable crystal structures with desirable methane deliverable capacities. These molecules were originally chosen because they represent a set of awkwardly shaped molecules. Hence, they have the potential to form porous structures with high methane capacities, but intuitive packing arguments alone cannot provide sufficient insight to make a priori arguments about the relative potential of these three molecules to perform well in this application. Even if we could predict crystal packing intuitively, the methane deliverable capacity does not scale in a simple way with crystal density; hence, both lattice energy and function must be computed. ESF maps for this application are very expensive because of the large energy range of viable predicted crystal structures, taking into account the effects of solvent stabilization coupled with the high cost of the methane adsorption calculations. In the study of T2, P2, and T2E (Fig. 3), crystal structures in the range up to 100 kJ/mol above the global minimum were considered, as compared to a more usual energy range of 10 or 15 kJ/mol for crystal structure landscapes for nonporous packings. Since the number of structures on the landscape increases rapidly as we move away from the global minimum, this 7- to 10-fold increased energy range leads to a much larger concomitant increase in the number of crystal structures that must be considered.

Table 1. Average performance achieved over 10 replicates for the three systems studied. Mean encounter time is the mean sample number at which the minimum is found, and mean epochs required is the sampling epoch in which this sample fell.

Structure	Mean encounter time	Mean epochs required
T2E	39.0	4
T2	14.3	2
P2	34.0	4

The ESF maps for methane deliverable capacity for T2, P2, and T2E contained ~5400, ~9800, and ~ 30,000 structures, respectively. To ensure reproducibility and to display the robustness of our approach, we tested the intelligent navigation workflow for each system with 10 replicate experiments, each of which was seeded with different initial structures chosen from the landscape. Using these replicate experiments, we were able to use the bootstrap methodology to calculate confidence intervals for the convergence of each of the three systems with respect to ideal behavior. All of the samplers converged on an ideal solution before 100 samples, or 10 epochs, have been completed. Since the executions are completed in parallel, when we calculate the first encounter time, we must only base this evaluation on the epoch in which the global minimum was found; that is, there is no advantage to being found halfway through a batch.

Table 1 shows the distribution of performance over the 10 repeats with the best performance being achieved by the T2 system, which shows a mean first encounter time of 14.3 samples, or within two completed epochs. Both P2 and T2E have a mean first encounter time of around four epochs. This can be rationalized by considering the full ESF maps for these systems: There are more that have both a high methane deliverable capacity and a low lattice energy for T2 than for other systems, facilitating the discovery of high performing systems. T2 also exhibits superior performance in the magnitude of our normalized objectives, with a score of circa 1.6, as compared to P2's score of circa 1.5 indicating that there is a more favorable trade-off between low energy and high methane capacity structures.

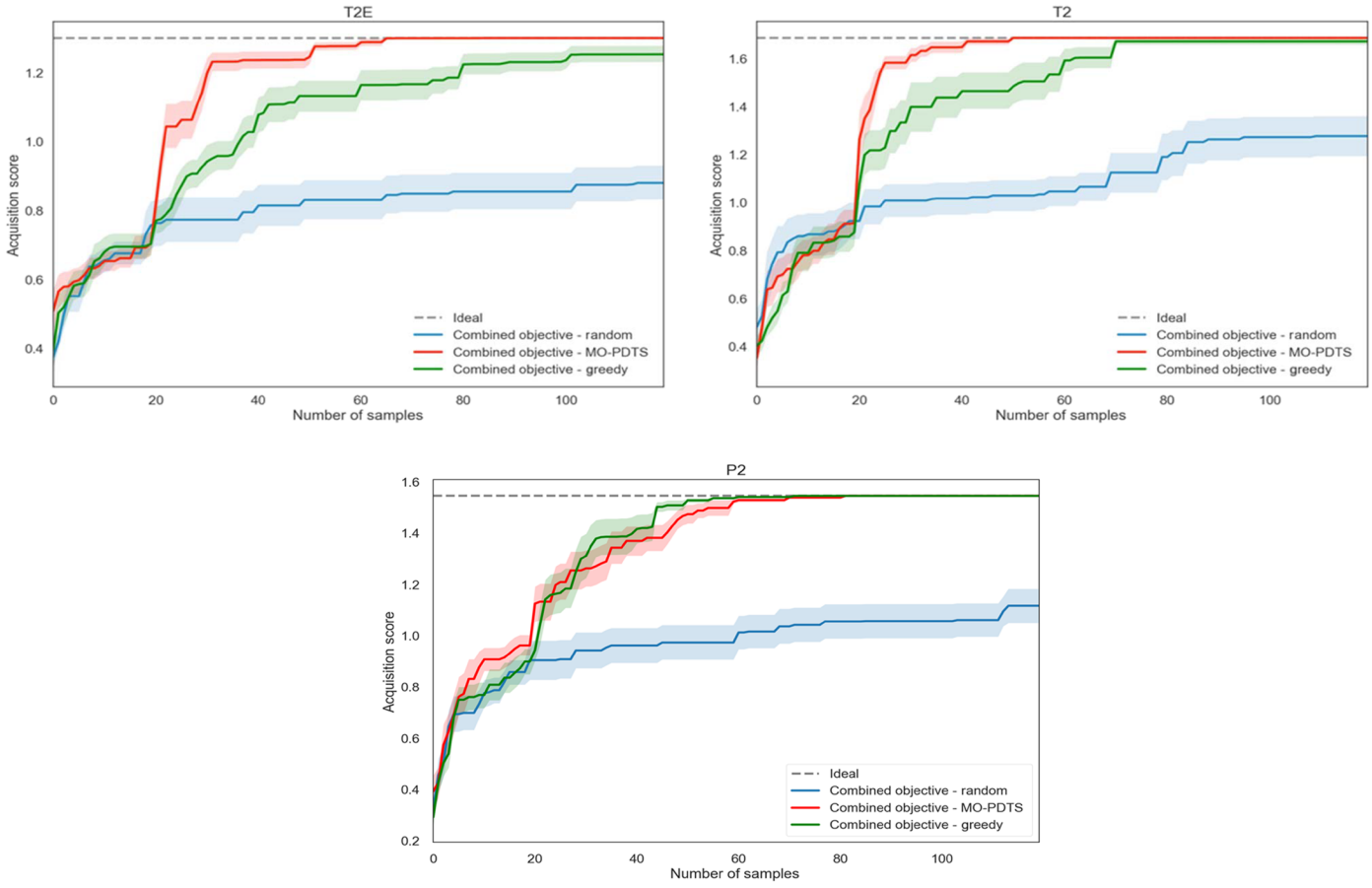

Fig. 4. Performance of the MO-PDTS sampler for the three systems studied. Confidence intervals are generated using the bootstrap methodology from 10 replicate experiments seeded with different candidate structures.

Table 2. Computational savings as a fraction of the potential ESF map for the systems T2, T2E, and P2.

Structure	Number of structures in ESF map	% of ESF sampled	CPU hours to generate full ESF map	Computational saving* (CPU hours)
T2E	29,848	0.14	392,213	391,427
T2	5403	0.32	74,469	73,945
P2	9817	0.39	96,369	95,583

*Computational saving is based on averaging the cost of each calculation over the entire set.

Comparison to greedy sampling

An alternative approach to the reduction in computational cost for the exploration of large ESF maps, or other compound libraries, is to use a greedy sampling method. For this class of search algorithm, a model is built from existing data and used to predict values for data that have not yet been acquired. At each epoch of sampling, the candidate that has the largest predicted value is selected—or the smallest value, for minimization purposes—and added to the training set, from which the model is then refitted. Most traditional Qualitative Structure Property Relationship (QSPR) methods use this methodology, either implicitly or explicitly, for accelerated materials discovery.

As shown in Fig. 4, the MO-PDTS sampler locates the ideal solution in all cases and outperforms the baseline random sampler significantly. For T2E and T2 systems, there is a clear advantage over the greedy sampler, indicating that these are systems where there are competing local maxima and demonstrating the advantage of the more sophisticated MO-PDTS method. In the case of P2, the performances are similar, indicating that there is a single clear structure-property relationship, which can be exploited by the greedy sampler.

The dangers of a greedy sampler are illustrated in the case of T2E (Fig. 5). The greedy sampler identifies a reasonably well-performing structure-property relationship and concentrates its sampling in this area. Unfortunately, this structure-property relationship does not indicate the existence of a second "peak" of activity with a higher value. The balance of exploration and

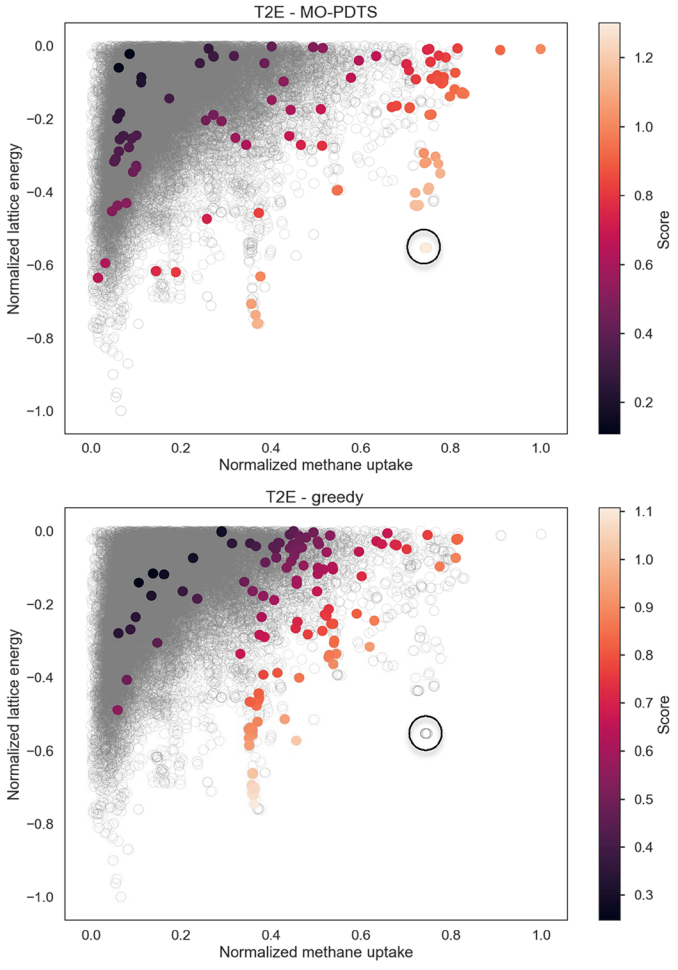

Fig. 5. Comparison of the MO-PDTS and greedy sampling strategies. Candidate structures are colored by their combined energy-structure score, and no color indicates that the structure was not sampled. It can be seen, for the T2E case, that the greedy sampler gets stuck in a local maxima but that PDTS is able to locate the global maximum (circled in black)

exploitation in MO-PDTS avoids this situation and samples in a more intelligent and robust manner. The performance curves in Fig. 4 indicate that there is little "cost" to adopting this more sophisticated strategy over a more traditional, greedy approach when the structure-property landscape is simple but significant benefits when it is not.

Computational savings

We have seen that the proposed intelligent navigation approach to ESF maps yields considerable computational savings. The exact details are shown in Table 2. In all cases, we see greater than two orders of magnitude improvement in the "time to insight," which results in hundreds of thousands of saved CPU hours.

In total, 544,955 hours were saved using this technique over the entire campaign; for context, this saving is similar in magnitude to a small grant on a supercomputer system. For many functional properties, this high level of computational acceleration could transform ESF maps from a proof-of-concept demonstration to an important, routine practical tool for in silico high-throughput screening, particularly for physical properties that are expensive to compute. We assessed these savings based solely on the savings in property evaluations, but where higher-level energies are required for lattice energy rankings, for example, by density functional theory (DFT), the savings would be even greater. Even when we only consider the property calculation savings, these benchmark figures suggest that this technique could allow a user to screen orders of magnitude more candidates for the same computational expense. As with all accelerations of this kind, there is not the same completeness guarantee that is possible by calculating the entire ESF map. However, we believe that this is more than compensated by the huge increase in throughput and the ability to evaluate a much broader range of candidate molecular structures. In many cases, the use of this technique may be the difference between an ESF map for a particular property being calculated and being deemed too expensive. This represents a significant practical advance in the ESF methodology, allowing us to tackle new functional properties that have hitherto been deemed impossible because of their high computational cost.

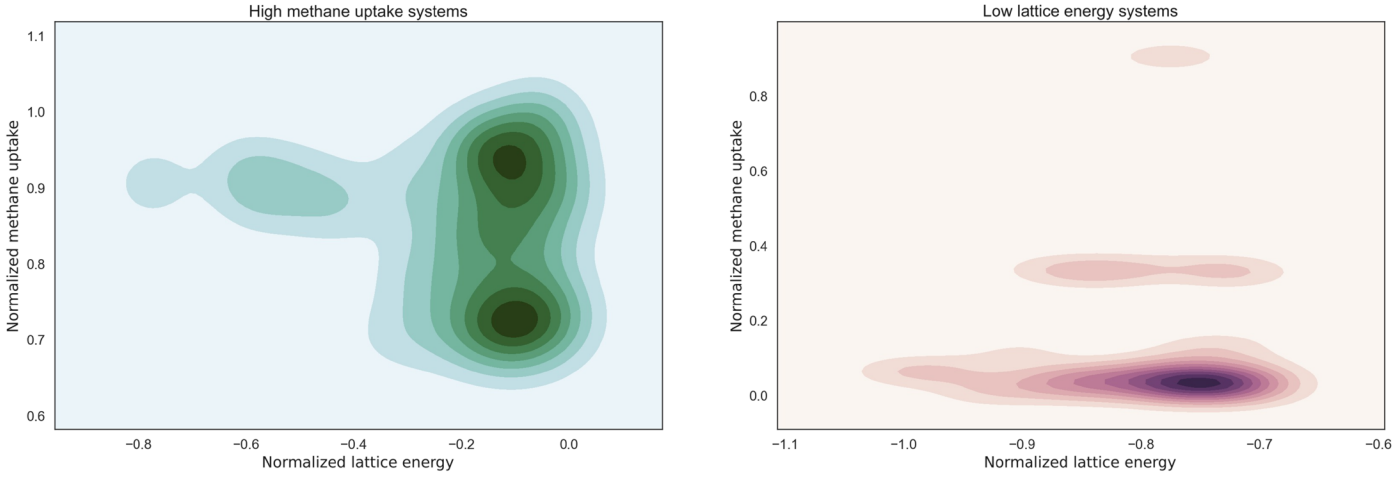

Fig. 6. Estimated density plot for the normalized lattice energy and the normalized methane deliverable capacity for the T2 system, with values normalized over the entire dataset between 0 and 1 for methane capacity (−1 and 0 for energy). Color hue is used to indicate contour planes of increasing density. The left plot focuses on crystal packings with high methane deliverable capacity; the right plot highlights systems with low lattice energies. Most of the low energy systems have poor methane deliverable capacity, and the largest high methane capacity systems are relatively high in lattice energy; that is, these two properties are, broadly speaking, orthogonal.

In conclusion, we present an important evolution in the ESF mapping technique for the a priori prediction of materials properties: a smart navigator for ESF maps based on MO-PDTS. The scalability of this method adds negligible overhead to the computation of the ESF map; by selectively sampling the map and only requiring the use of expensive function calculations for a fraction of the structures, we are able to make significant computational savings. For the three structures here studied, we were able to save more than half a million CPU hours. This has two key advantages. First, we significantly reduce the opportunity risk for the selection of systems for ESF map calculation; that is, did I choose the right molecule to spend this resource on? Second, through the reduction of computational requirements, we extend the power of the ESF map approach both to researchers who are not able to access the necessary computational resources and also to expensive property calculations for large, complex ESF maps that are simply intractable today.

MATERIALS AND METHODS

MO-PDTS optimization details

We posed the problem as a multiobjective optimization over both energy and methane deliverable capacity, thus searching for the ESF maps for low-energy, highly porous, crystalline forms. For the purposes of this study, we are testing the methodology as if we do not have the final energies that we require, mimicking the case where higher-level energy calculations are required than were used in the structure search itself. We use the calculated force field energies as a proxy for these higher-level energies. We did not consider the expense for these energy calculations when calculating savings, and so this study represents a lower bound on the potential for this method.

To demonstrate the modular nature of this approach, the two considered properties were modeled with different features—22 geometrically defined features for porosity, and the National Institute of Standards and Technology JARVIS (*30*) descriptor set for the lattice energy.

Topological analysis of the pore space within a crystal structure was performed using the void analysis tool zeo++ (*31*). The outputs from this analysis included the pore dimensionality [zero dimension (0D), 1D, 2D, or 3D], pore diameters, surface areas, and pore volumes. A probe radius of 1.70 Å was used in all calculations. A total of 22 pore descriptors were calculated for each of the predicted crystal structures. These 22 descriptors are simple extensions to four basic pore descriptors: crystal density, largest pore diameter, total surface area, and total pore volume. First, the total surface area and the total pore volume were decomposed into accessible and nonaccessible contributions. Second, to capture the heterogeneity of the pore geometry within a structure, we derived several descriptors based on the surface areas and pore volumes of individual channels and pockets. Last, the total surface area was also decomposed into elemental contributions. A description of each descriptor is as follows:

1) Crystal density (in grams per cubic centimeter);

2 to 4) Pore diameters (in angstrom): the largest included sphere (D_i), the largest free sphere (D_f), and the largest included sphere along the free sphere path (D_{if});

5 to 8) Accessible surface area (in square meters per gram), nonaccessible surface area (in square meters per gram), accessible volume (in cubic centimeters per gram), nonaccessible volume (in cubic centimeters per gram);

9 to 12) Absolute (in cubic centimeters per gram) and fraction (–) of probe-occupied accessible volume, absolute (in cubic centimeters per gram), and fraction (–) of probe-occupied nonaccessible volume;

13 to 16) Elemental surface areas (in square meters per gram), i.e., total (accessible + nonaccessible) surface area decomposed into individual contributions from the H, C, N, and O atoms;

17 to 22) Variants based on the surface areas and pore volumes of individual channels [accessible (acc)] and pockets [nonaccessible (nacc)] to capture, to some extent, the heterogeneity of the pore space within a crystal structure:

Average of the accessible surface areas divided by the corresponding accessible volumes for all individual channels

$$A_{\mathrm{acc}} = \frac{1}{n}\sum_{n=1}^{n}\frac{S_{n,\mathrm{acc}}}{V_{n,\mathrm{acc}}}$$

Median of the accessible surface areas divided by the corresponding accessible volumes for all individual channels

$$M_{\mathrm{acc}} = \mathrm{Median}\left(\frac{S_{1,\mathrm{acc}}}{V_{1,\mathrm{acc}}}, \frac{S_{2,\mathrm{acc}}}{V_{2,\mathrm{acc}}} \cdots \frac{S_{n,\mathrm{acc}}}{V_{n,\mathrm{acc}}}\right)$$

Variance of the accessible surface areas divided by the corresponding accessible volumes for all individual channels

$$\sigma^2{}_{\mathrm{acc}} = \frac{1}{n}\sum_{n=1}^{n}\left(\frac{S_{n,\mathrm{acc}}}{V_{n,\mathrm{acc}}} - A_{\mathrm{acc}}\right)^2$$

Average of the nonaccessible surface areas divided by the corresponding nonaccessible volumes for all individual pockets

$$A_{\mathrm{nacc}} = \frac{1}{n}\sum_{n=1}^{n}\frac{S_{n,\mathrm{nacc}}}{V_{n,\mathrm{nacc}}}$$

Median of the nonaccessible surface areas divided by the corresponding nonaccessible volumes for all individual pockets

$$M_{\mathrm{nacc}} = \mathrm{Median}\left(\frac{S_{1,\mathrm{nacc}}}{V_{1,\mathrm{nacc}}}, \frac{S_{2,\mathrm{nacc}}}{V_{2,\mathrm{nacc}}} \cdots \frac{S_{n,\mathrm{nacc}}}{V_{n,\mathrm{nacc}}}\right)$$

Variance of the nonaccessible surface areas divided by the corresponding nonaccessible volumes for all individual pockets

$$\sigma^2{}_{\mathrm{nacc}} = \frac{1}{n}\sum_{n=1}^{n}\left(\frac{S_{n,\mathrm{nacc}}}{V_{n,\mathrm{nacc}}} - A_{\mathrm{nacc}}\right)^2$$

where n is the number of channels or pockets, $S_{n,\mathrm{acc}}$ and $V_{n,\mathrm{acc}}$ are the accessible surface area and accessible volume for the nth channel, respectively; and $S_{n,\mathrm{nacc}}$ and $V_{n,\mathrm{nacc}}$ are the nonaccessible surface area and nonaccessible volume for the nth pocket, respectively.

Since JARVIS is a very high-dimensional set of features with significant information redundancy, we use a principle component analysis to reduce the number of features while retaining 99% of the

Table 3. Dimensionality of JARVIS descriptors, once reduced using principal components analysis to retain 99% of original variance.

System	Number of dimensions
T2	45
P2	59
T2E	28

variance. This resulted in the feature dimensions for the systems shown in Table 3.

To quantify the acceleration achieved, we compare our results here to the calculation of full ESF maps, previously reported by some of the authors (*6*); that is, we accurately computed both lattice energies and methane deliverable capacities for all structures on the three associated ESF maps. Since ESF maps are used as indicators of the potential for a molecule to behave in a desirable way, we based our metric of success on the first encounter time for the global minimum on the ESF landscape; that is to say, the structure that has the best combination of low energy and high methane deliverable capacity. For this study, we weighted the contribution to this score from the energy term and the property term equally

$$S = aE_i + bP_i$$

Where a and b are weighting coefficients to energy and property, respectively, and, in this study, are equal and normalized to remove units and ensure that the scales of the two properties are comparable. We note that, for a more conservative approach, it is possible to weight the energy term more highly, that is, to increase the likelihood that the identified structure is thermodynamically accessible in the laboratory.

Figure 6 shows that, in general, structures with high deliverable methane capacity have a high lattice energy. Thus, we expect that the number of structures, which have both desirable methane deliverable capacity and low lattice energy to be small, is further emphasizing the need for an efficient, accelerated approach and also the importance of the multiobjective nature of our search strategy.

The MO-PDTS was seeded with an initialization strategy based on k-means inspired by the generation of inducing points for sparse Gaussian processes. In this methodology, k-centroids were determined over input descriptor (feature) space using the k-means algorithm. The structures that minimized the distance to these centroids were chosen to initialize the search; that is, we selected the nearest structure to each of the k-centroids. Under a uniform distribution, this is equivalent to a Latin hypercube due to the spherical repulsion of k-means. However, under a nonuniform distribution, we believe that this initialization captures the underlying data structure better, leading to increased model stability throughout the optimization process. MO-PDTS was then run for 10 epochs, at each of which 10 structures were selected and properties were calculated. To account for the difference in magnitudes of the two objectives, the values for each were scaled for each objective based on the 20 selected structures from which the search was seeded.

Simulation details

For each ESF map, candidate crystal structures were generated using a quasi-random sampling procedure, as implemented in the Global Lattice Energy Explorer software (*32*). Molecules were first sketched in ChemDraw, followed by an initial molecular geometry optimization with the COMPASS force field (*33*), as implemented in the Materials Studio software package (*34*). Force field–optimized molecular geometries were further refined by reoptimization using DFT with the M06-2X exchange-correlation functional and 6-311G** basis set. Molecular DFT calculations were performed with the Gaussian09 software (*35*). These molecular geometries were held rigid throughout crystal structure generation and lattice energy minimization.

Lattice energy calculations were performed with an anisotropic atom-atom potential using DMACRYS (*36*). Electrostatic interactions were modeled using an atomic multipole description of the molecular charge distribution (up to hexadecapole on all atoms) from the B3LYP/6-31G**-calculated charge density using a distributed multipole analysis (*37*). Atom-atom repulsion and dispersion interactions were modeled using a revised Williams intermolecular potential (*38*).

Methane adsorption was predicted for each structure at a temperature of 298 K and pressures of 5.8 and 65 bar; methane deliverable capacity was calculated as the difference in methane uptake at 65 and 5.8 bar (assuming gas storage at 65 bar and gas delivery at 5.8 bar). All of the adsorption predictions were performed using grand canonical Monte Carlo simulations involving a 50,000-cycle equilibration period and a 50,000-cycle production run, using the RASPA code (*39*). The adsorbent-adsorbate and adsorbate-adsorbate intermolecular interactions were modeled using Lennard-Jones (LJ) potentials, with a cutoff radius of 12.0 Å (beyond which a simple truncation was applied). Methane (CH_4) was described by the TraPPE united-atom force field (*40*), in which CH_4 is considered a single entity, i.e., the carbon atom and its bonded hydrogen atoms are grouped together to form one interaction site. The LJ parameters for the adsorbent structures were assigned on the basis of the DREIDING force field (*41*). The Lorentz-Berthelot combining rules were used to calculate the LJ cross-parameters.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/33/eabi4763/DC1

REFERENCES AND NOTES

1. E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, A. Aspuru-Guzik, What is high-throughput virtual screening? A perspective from organic materials discovery. *Annu. Rev. Mater. Res.* **45**, 195–216 (2015).
2. C. Suh, C. Fare, J. A. Warren, E. O. Pyzer-Knapp, Evolving the materials genome: How machine learning is fueling the next generation of materials discovery. *Annu. Rev. Mater. Res.* **50**, 1–25 (2020).
3. O. K. Farha, A. Ö. Yazaydın, I. Eryazici, C. D. Malliakas, B. G. Hauser, M. G. Kanatzidis, S. T. Nguyen, R. Q. Snurr, J. T. Hupp, De novo synthesis of a metal–organic framework material featuring ultrahigh surface area and gas storage capacities. *Nat. Chem.* **2**, 944–948 (2010).
4. C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp, R. Q. Snurr, Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* **4**, 83–89 (2012).
5. C. M. Simon, J. Kim, D. A. Gomez-Gualdron, J. S. Camp, Y. G. Chung, R. L. Martin, R. Mercado, M. W. Deem, D. Gunter, M. Haranczyk, D. S. Sholl, R. Q. Snurr, B. Smit, The materials genome in action: Identifying the performance limits for methane storage. *Energy Environ. Sci.* **8**, 1190–1199 (2015).
6. A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, A. Stephenson, C. M. Kane, R. Clowes, T. Hasell, A. I. Cooper, G. M. Day, Functional materials discovery using energy–structure–function maps. *Nature* **543**, 657–664 (2017).
7. J. Yang, S. De, J. E. Campbell, S. Li, M. Ceriotti, G. M. Day, Large-scale computational screening of molecular organic semiconductors using crystal structure prediction. *Chem. Mater.* **30**, 4361–4371 (2018).
8. P. Cui, D. P. McMahon, P. R. Spackman, B. M. Alston, M. A. Little, G. M. Day, A. I. Cooper, Mining predicted crystal structure landscapes with high throughput crystallisation: Old molecules, new insights. *Chem. Sci.* **10**, 9988–9997 (2019).
9. D. P. McMahon, A. Stephenson, S. Y. Chong, M. A. Little, J. T. Jones, A. I. Cooper, G. M. Day, Computational modelling of solvent effects in a prolific solvatomorphic porous organic cage. *Faraday Discuss.* **211**, 383–399 (2018).
10. E. Brochu, V. M. Cora, N. de Freitas, A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599 [cs.LG] (12 December 2010).
11. J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms. arXiv:1206.2944 [stat.ML] (29 August 2012).

12. A. D. Bull, Convergence rates of efficient global optimization algorithms. *J. Mach. Learn. Res.* **12**, 2879–2904 (2011).
13. L. Schultz, V. Sokolov, Bayesian optimization for transportation simulators. *Proc. Comput. Sci.* **130**, 973–978 (2018).
14. J. Park, K. H. Law, in *Smart Sensor Phenomena, Technology, Networks, and Systems Integration 2015* (International Society for Optics and Photonics, 2015), vol. 9436, pp. 943608.
15. E. O. Pyzer-Knapp, Bayesian optimization for accelerated drug discovery. *IBM J. Res. Dev.* **62**, 2:1–2:7 (2018).
16. J. L. McDonagh, A. Shkurti, D. J. Bray, R. L. Anderson, E. O. Pyzer-Knapp, Utilizing machine learning for efficient parameterization of coarse grained molecular force fields. *J. Chem. Inf. Model.* **59**, 4278–4288 (2019).
17. C. E. Rasmussen, *Gaussian Processes for Machine Learning* (MIT Press, 2006).
18. J. M. Hernández-Lobato, J. Requeima, E. O. Pyzer-Knapp, A. Aspuru-Guzik, in *Proceedings of the 34th International Conference on Machine Learning* (JMLR. org, 2017), vol. 70, pp. 1470–1479.
19. J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Ali, R. P. Adams, Scalable Bayesian optimization using deep neural networks. arXiv:1502.05700 [stat.ML] (13 July 2015).
20. B. Matérn, *Spatial Variation. Medd. fr. St. Skogsf. Inst. 49 (5). Also Appeared as Number 36 of Lecture Notes in Statistics* (Springer-Verlag, 1960).
21. R. M. Neal, *Bayesian Learning for Neural Networks* (Springer, 1996).
22. D. Jasrasaria, E. O. Pyzer-Knapp, in *Science and Information Conference* (Springer, 2018), pp. 1–15.
23. J. Mockus, in *System Modeling and Optimization* (Lecture Notes in Control and Information Sciences, Springer, 1982), pp. 473–481.
24. J. González, Z. Dai, P. Hennig, N. D. Lawrence, Batch Bayesian optimization via local penalization. arXiv:1505.08052 [stat.ML] (15 October 2015).
25. A. Shah, Z. Ghahramani, Parallel predictive entropy search for batch global optimization of expensive objective functions. arXiv:1511.07130 [cs.LG] (23 November 2015).
26. M. Groves, E. O. Pyzer-Knapp, Efficient and scalable batch bayesian optimization using K-Means. arXiv:1806.01159 [stat.ML] (19 September 2018).
27. J. Wang, S. C. Clark, E. Liu, P. I. Frazier, Parallel Bayesian global optimization of expensive functions. arXiv:1602.05149 [stat.ML] (5 May 2019).
28. W. R. Thompson, On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**, 285–294 (1933).
29. J. Hindo, E. Pyzer-Knapp, in *Abstracts of Papers of the American Chemical Society* (American Chemical Society, 2016), vol. 252.
30. K. Choudhary, B. DeCost, F. Tavazza, Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Phys. Rev. Mater.* **2**, 083801 (2018).
31. T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza, M. Haranczyk, Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous Mesoporous Mater.* **149**, 134–141 (2012).
32. D. H. Case, J. E. Campbell, P. J. Bygrave, G. M. Day, Convergence properties of crystal structure prediction by quasi-random sampling. *J. Chem. Theory Comput.* **12**, 910–924 (2016).
33. H. Sun, COMPASS: An ab initio force-field optimized for condensed-phase applicationsoverview with details on alkane and benzene compounds. *J. Phys. Chem. B* **102**, 7338–7364 (1998).
34. Accelrys Software Inc., *Materials Studio v6.10* (Accelrys Software Inc., 2012).
35. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, *Gaussian~09 Revision E.01*.
36. S. L. Price, M. Leslie, G. W. A. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis, G. M. Day, Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Phys. Chem. Chem. Phys.* **12**, 8478–8490 (2010).
37. A. J. Stone, M. Alderton, Distributed multipole analysis. *Mol. Phys.* **56**, 1047–1064 (1985).
38. E. O. Pyzer-Knapp, H. P. G. Thompson, G. M. Day, An optimized intermolecular force field for hydrogen-bonded organic molecular crystals using atomic multipole electrostatics. *Acta Crystallogr. B* **72**, 477–487 (2016).
39. D. Dubbeldam, S. Calero, D. E. Ellis, R. Q. Snurr, RASPA: Molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol. Simul.* **42**, 81–101 (2016).
40. M. G. Martin, J. I. Siepmann, Transferable potentials for phase equilibria. 1. United-atom description ofn-alkanes. *J. Phys. Chem. B* **102**, 2569–2577 (1998).
41. S. L. Mayo, B. D. Olafson, W. A. Goddard, DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909 (1990).

Acknowledgments

Funding: E.O.P.-K. acknowledges the support from the STFC Hartree Centre's Innovation Return on Research Programme, funded by the Department for Business, Energy and Industrial Strategy. A.I.C. and L.C. acknowledge the Leverhulme Trust for supporting the Leverhulme Research Centre for functional materials design. G.M.D. thanks the European Research Council for funding under the European Union's Seventh Framework Programme (FP/2007-2013) through grant agreement number 307358 (ERC-stG-2012-ANGLE). **Author contributions:** The initial idea was developed by E.O.P.-K. and AIC, and its implementation was discussed with L.C. The descriptors were provided by L.C. and G.M.D. E.O.P.-K. managed the project. All authors participated in the data analysis and writing and reading of the paper. **Competing interests:** The authors declare that they have no competing interest. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The datasets used in this study, relating to the P2, T2, and T2E systems, are publicly available and can be accessed at the following link: http://dx.doi.org/10.5258/SOTON/404749. Additional data related to this paper may be requested from the authors.

Submitted 12 March 2021
Accepted 30 June 2021
Published 13 August 2021
10.1126/sciadv.abi4763

Citation: E. O. Pyzer-Knapp, L. Chen, G. M. Day, A. I. Cooper, Accelerating computational discovery of porous solids through improved navigation of energy-structure-function maps. *Sci. Adv.* **7**, eabi4763 (2021).