



Inverse design of nanoporous crystalline reticular materials with deep generative models

Zhenpeng Yao^{1,2}, Benjamín Sánchez-Lengeling¹, N. Scott Bobbitt³, Benjamin J. Bucior³, Sai Govind Hari Kumar², Sean P. Collins⁴, Thomas Burns⁴, Tom K. Woo⁴, Omar K. Farha^{3,5}, Randall Q. Snurr³ and Alán Aspuru-Guzik^{1,2,6,7}

Reticular frameworks are crystalline porous materials that form via the self-assembly of molecular building blocks in different topologies, with many having desirable properties for gas storage, separation, catalysis, biomedical applications and so on. The notable variety of building blocks makes reticular chemistry both promising and challenging for prospective materials design. Here we propose an automated nanoporous materials discovery platform powered by a supramolecular variational autoencoder for the generative design of reticular materials. We demonstrate the automated design process with a class of metal-organic framework (MOF) structures and the goal of separating carbon dioxide from natural gas or flue gas. Our model shows high fidelity in capturing MOF structural features. We show that the autoencoder has a promising optimization capability when jointly trained with multiple top adsorbent candidates identified for superior gas separation. MOFs discovered here are strongly competitive against some of the best-performing MOFs/zeolites ever reported.

Reticular frameworks (which include metal-organic frameworks (MOFs) and covalent organic frameworks) are crystalline porous materials, many of which feature high internal surface area and high stability. They are formed via the self-assembly of molecular building blocks (that is, nodes and linkers) in different topologies. The notable variety of the possible building blocks and the diverse ways they can be assembled endow reticular frameworks with exceptional geometrical and chemical tunability¹. Since the first MOF², thousands of reticular frameworks have been made towards various applications with remarkable advances achieved in fields such as gas storage³, molecular separation^{4,5}, catalysis⁶, sensing⁷, electrochemical energy storage⁸ and drug delivery⁹. Aiming at a particular application, novel reticular frameworks can be designed in a trial-and-test manner through selecting plausible building blocks that assemble in a desired topology¹⁰. Given the vastness of chemical space for small molecules¹¹ that can potentially be used as linkers, reticular frameworks show a near-infinite combinatorial design space. The boundless design space substantially expands the scope of useful materials for prospective applications, yet its enormity also complicates its systematic exploration. Therefore, the search for new materials becomes a constrained global optimization problem in the high-dimension space.

One powerful approach developed to assist the discovery of reticular frameworks is high-throughput computational¹² and experimental¹³ screening. High-throughput screening proceeds via generating/synthesizing and evaluating all the frameworks (building block combinations) from a selected library. The high-throughput computational methodology has enabled the examination of a design space on the order of 10^3 – 10^5 . One main drawback of this approach is the low coverage and restriction of the search space according to the combinatorics of the building blocks. In addition

to high-throughput screening, heuristic optimization approaches include genetic algorithms and evolutionary strategies. Given a score metric and a set of candidates, these methods can transform/evolve/mutate the candidates based on their scored performance, eventually leading to higher scoring structures. This approach allows the search of larger spaces and has been successful at identifying top-performing MOFs in recent studies^{14,15}. The downside of this approach is that it requires specifying prior rules on how to transform the frameworks, which then creates a preceding constraint of the types of framework that can be explored.

Another promising approach for optimizing frameworks lies with machine learning algorithms that are able to learn from data and improve their performance automatically through experience. Among them, predictive algorithms (that is, discriminative models), those that given a datapoint x aim to predict a property y , have been used to aid or even replace physical simulations under certain circumstances. Discriminative models have been widely applied to accelerate the high-throughput screening process of reticular frameworks for properties such as storage¹⁶, mechanical stability¹⁷, synthesizability¹⁸ and so on. Another class of algorithms that do not necessarily deal with predicting a property y but modelling the data itself are generative models. For example, a Bernoulli probability distribution can be used as a model to generate a coin flip. With more complicated data distributions, we can use deep generative models such as variational autoencoders¹⁹ and generative adversarial networks²⁰. In these cases, the mapping between probability distributions and data is learned via a deep neural network; and this map can be further enhanced with additional information (physical properties) to condition or bias the generative process. By conditioning the generative process on a property of interest, the models can be employed to generate preferentially molecules with a given

¹Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. ²Department of Chemistry, University of Toronto, Toronto, Ontario, Canada. ³Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. ⁴Department of Chemistry and Biomolecular Science, University of Ottawa, Ottawa, Ontario, Canada. ⁵Department of Chemistry, Northwestern University, Evanston, IL, USA. ⁶Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada. ⁷Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario, Canada. ✉e-mail: yaozhenpeng@gmail.com; snurr@northwestern.edu; alan@aspuru.com

property. This property-to-structure approach is called inverse design²¹. Generative models can be used as a key component to realize the automated ‘closed loop’ design of materials towards targeted performances. These models have been successfully applied to a variety of molecular^{22,23} and material design applications^{24,25}. In the context of reticular frameworks, we can design and generate a framework by sampling a random vector and mapping it back to a learned data distribution. Other challenges relevant to closing the loop are the planned synthesis of a reticular framework given a set of materials and the potential automatic robotic realization of this procedure.

The primary goal in the reticular framework design presented in this work is the guided optimization of crystal structures according to a targeted functionality. In a simplified manner, a reticular framework could be seen as a large collection of regularly bonded particles (atoms) in three-dimensional space. Optimization with this representation corresponds to optimizing the number of particles, the identities of these particles and their positions. The realization of this optimization is then quite challenging due to the high and variable dimensionality, large particle number, and mix of discrete and continuous variables. Therefore, finding an efficient representation becomes the essential step for machine learning-based reticular framework optimization. One way to attack the problem is reduction approximation through exploiting symmetries and hierarchical structures of the systems. An ideal representation would encode the degrees of freedom, physical symmetries and constraints of a system and be amenable to gradient-based optimization techniques. The representation should also be decodable such that a framework can be reconstructed or decoded back. With proper representation, deep generative models show great promise for reticular framework optimization because of their potential capability to map these frameworks into a continuous vector representation. Variational autoencoders (VAEs), in particular—which can learn an invertible mapping, encode a material to a vector (that is, its latent vector) and decode it back to a framework—are a compelling solution. Optimization of materials can be ultimately made in the latent vector space within the VAE framework, which then lays the ground for the design of reticular frameworks with desired properties.

In this Article, we build an automated nanoporous materials discovery platform for the property-orientated generative design of reticular frameworks, empowered by a supramolecular variational autoencoder. We develop a semantically constrained graph-based canonical code for the efficient representation of reticular frameworks (RFcode). With MOF structures from the computation-ready, experimental MOF 2019 all-solvent removed (CoRE MOF 2019-ASR) database²⁶ as inputs and clean-energy applications (that is, CO₂/N₂ and CO₂/CH₄ separations) as the exemplified targets, we demonstrate the automated design process using a discovery platform for novel MOF structures with remarkably improved performance. By examining the latent space of our model, we illustrate that our representation captures structural features while also organized around properties. We demonstrate its capabilities for automatic targeted generation by proposing top candidates for gas-separation adsorbent materials. We believe that the MOFs discovered here are strongly competitive against some of the best-performing MOFs/zeolites ever reported in the literature. We make our trained models, results and code available as open source to aid reproducibility and adoption to broader applications (for example, covalent organic frameworks, metal-organic polyhedra, hydrogen-bonded organic frameworks and coordinational polymers).

Reticular framework representation and identification

All crystalline materials can be seen as a collection of particles with different identities arranged periodically in three-dimensional space. Given the identities and positions of the atoms, in principle, any property can be computed for the framework from the

Schrödinger equation. However, in practice, this may be difficult due to computational complexity and cost, which lead to trade-offs generally made in the form of approximations. Another approach is to estimate material properties using models such as linear models or neural networks that learn transformations on their input representations. Ideally, the representation would contain the same symmetries that the Schrödinger equation presents: translational, rotational and permutational invariance with respect to its atomic identities²¹. Meanwhile, representations and models are coupled such that different types of input will lead to distinct choices of preferred models (for example, images and convolutional networks). Materials representation currently is an open research problem, while for non-periodic chemical systems (for example, molecules), several representations have been proven successful, such as fingerprints²⁷, SMILES (simplified molecular input line entry system)²², SELFIES (self-referencing embedded strings)²⁸ and graphs²³. Defining a representation for periodic crystalline materials is more challenging because of the necessity to deal with the extra-dimensional connections at the border of unit cells. Particularly for reticular frameworks, their generally larger cell sizes (10²–10⁴ atoms²⁹ versus common crystalline materials with 10¹–10² atoms³⁰) bring further difficulties in representing them efficiently. Methods such as the smooth overlap of atomic positions³¹, Voronoi tessellation³², diffraction images³³ and multi-perspective fingerprints³⁴ have been suggested for crystalline materials classification, property prediction and so on. Some of the most promising representations under development are graph-based^{35,36} algorithms, where atoms are encoded as vertices and atom pairs (that is, bonds) as edges. They can be effective without encoding positional coordinates explicitly. However, applying this representation to typical reticular frameworks results in graphs with 10²–10⁴ vertices and 3–5 edges per vertex²⁹, leading to a space with billions of potential configurations. Barely any effective optimizations can be done in a space of this size using the graph models, and thus reductions are called for. Tiling, net and graph theories^{37–39} can be used to aid the reduction by replacing atom-based vertices with motif-based vertices and bond-based edges with polyatomic-branch-based edges that connect these motifs.

Inspired by these reduction theories, we construct our representation of the reticular frameworks (that is, RFcode) using their unique, decomposed nets as a tuple: edges|vertices|topologies. Edges are molecular fragments with two connection points, vertices are multi-connected metal or organic nodes, and topologies define how these components are connected to form a specific reticular framework. Note that in the RFcode and throughout this paper, topologies are indicated by a three-letter code in bold font. Within the RFcode, we consider the edges as semantically constrained graphs²⁸, while vertices and topologies are categorical variables from known frameworks considering their relatively limited variety. In addition, we consider metal and organic vertices separately in the RFcode. The advantages of the RFcode are: (1) efficiency, as there is no redundant information, edges and vertices are only described once in the RFcode; (2) uniqueness, each representation encodes a unique framework; and (3) invertibility, as all components can be readily translated back and forth. Moreover, for each component of the RFcode, generative models have been effectively developed, and therefore a model that takes the full RFcode is realizable. To illustrate this method, we use MOF-117⁴⁰ as an example, and its representation is shown in Fig. 1a.

The RFcode representations of reticular frameworks can be determined automatically using a previously developed identification algorithm supplemented with framework deconstruction⁴¹ and reconstruction tools⁴². As a demonstration of our method, we decomposed all MOF structures from the CoRE 2019-ASR MOF database into their building blocks and identified all their RFcodes. Meanwhile, collections of edges, vertices (metal and organic) and

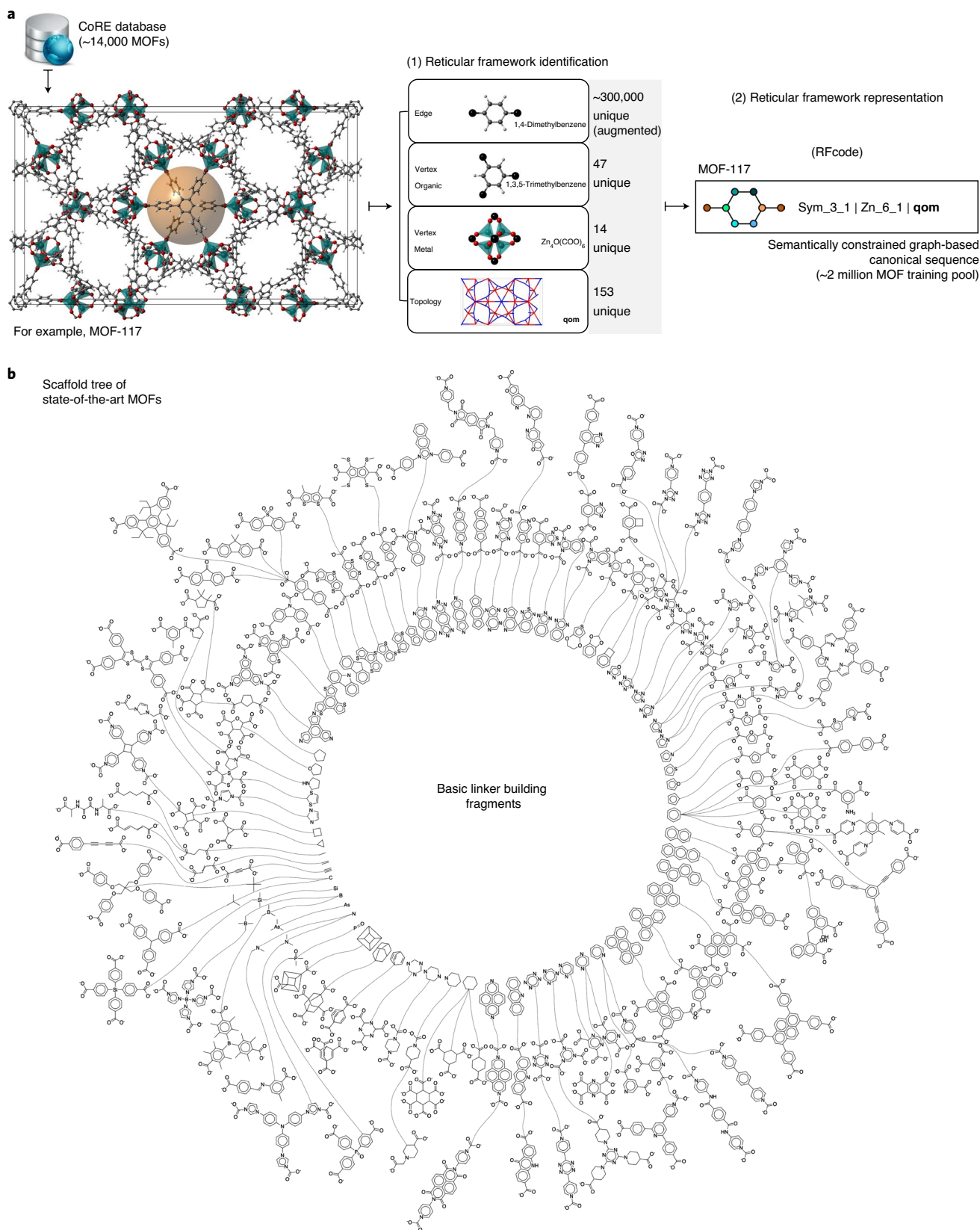


Fig. 1 | Reticular framework identification and representation, exemplified with MOF structures from the CoRE MOF database²⁶. **a**, Reticular frameworks (for example, MOF-117)⁴⁰ are: (1) decomposed to their building blocks (edges, organic/metal vertices) and topology using a previously developed identification method⁴¹, which are then recognized and labelled; and (2) the labels are further converted to semantically constrained graph-based canonical sequences, namely the Rfcode (edge | organic vertex | metal vertex | topology). **b**, A fragmentation analysis was conducted on the linkers of all MOF structures from the CoRE MOF database. Here we illustrate all the basic building fragments of state-of-the-art MOFs with high occurrence rate (the inner circle) and linkers derived from them (the second and third circles) in a scaffold tree plot.

topologies were also built. To have a sense of the chemical variety of all linkers in the CoRE database, we conducted a fragmentation analysis of them using molBLOCKS⁴³, and the derivations of different linkers are illustrated using a scaffold tree plot shown in Fig. 1b. Note that here and in the traditional MOF terminology, an organic 'linker' may be a single edge (connecting two metal vertices) or may contain an organic vertex and several edges.

Reticular framework (MOF) library generation

While there are no established rules on the sufficient size of training datasets for deep generative models, empirically these models start to be useful when input datasets are on the order of 10^6 . With the correct architecture, at this scale, the model can begin to generate new data that are likely to come from the empirical data distribution. Considering that there are only around 14,000 MOFs in the CoRE MOF database, a training data augmentation is necessary. Starting with all the MOF edges obtained from CoRE MOF identification (372 edges), we did random functionalizations (Supplementary Fig. 2) with selected common functional groups of known MOF structures (Supplementary Table 1). An augmented edges dataset of ~300,000 was generated. Vertex and topology datasets are constructed during the identification of the CoRE database as mentioned in the previous section by selecting vertices and topologies that are compatible with the current reticular framework reconstructor^{42,44}. Therefore, all these datasets are subject to further expansions in the future with improvements of the reconstructor. We then used this augmented edge dataset with the vertex dataset (metal: 14, Supplementary Fig. 3; organic: 47, Supplementary Fig. 4) and topology dataset (153, Supplementary Fig. 5), resulting in an augmented dataset with around two million MOF structures. An underlying assumption in our dataset is that the current vertex and topology pools represent plausible and realizable structures for reticular frameworks. Our search space does not include new vertices and topologies.

Besides generating new structures, we are interested in making our model aware of properties of interests. Doing so with deep neural networks requires having a large dataset of reticular frameworks (RFcodes) as well as properties, preferably experimental. However, such a dataset is currently lacking; therefore, we resorted to computational simulations on around 45,000 randomly selected MOF structures. The randomness allows coverage of multiple types of framework, and the quantity is to keep the computational cost at a reasonable level. We considered properties as follows: four textural properties (pore-limiting diameter (PLD), largest cavity diameter (LCD), density and accessible gravimetric surface area (AGSA)), three properties related to natural gas separation (CO_2 uptake, CH_4 uptake and CO_2/CH_4 selectivity, all at 5 bar and 300 K for a 10/90 mole fraction mixture of CO_2/CH_4) and three properties related to flue gas separation (CO_2 uptake, N_2 uptake and CO_2/N_2 selectivity, all at 1 bar and 313 K for a 15/85 mole fraction mixture of CO_2/N_2). Textural properties were calculated geometrically, and gas uptake properties were calculated using grand canonical Monte Carlo simulations. Gas-separation selectivities, which are entirely dependent on the uptake values of the mixed-gas phases from the mixed-gas simulations, were then derived numerically. We use pore blocking to prevent insertions into cavities that are inaccessible to the adsorbate molecules due to narrow windows. Therefore, some reticular frameworks may have extremely small or even zero uptakes of the larger radius molecules such as CH_4 and N_2 . As a result, these frameworks are predicted to have enormous or even infinite (∞) selectivity of CO_2 against CH_4 and N_2 . In reality, the observed selectivities may not be perfect (infinite) because the frameworks may not be completely rigid and large adsorbate molecules may not be totally blocked. The gas adsorption simulation of flexible reticular frameworks is still an open question that goes beyond the scope of this study. As a result, these infinite selectivity numbers should be

seen as a sign of gas-separation performance that is predicted to be outstanding compared with frameworks with lower selectivity values, rather than truly infinite selectivity. Further details are described in Supplementary Note 1. The distributions of the textural properties for these 45,000 MOFs are shown in Supplementary Fig. 6 and the distributions of gas uptake properties for these 45,000 MOFs are shown in Supplementary Fig. 7.

Supramolecular variational autoencoder

For our deep generative model, we utilize a VAE⁴⁵. A VAE is trained to process and reconstruct non-labelled data in an unsupervised manner. In its simplest form, a VAE is composed of two components: an encoder and a decoder. For a given datapoint \mathbf{x} , the encoder compresses the information to a vector \mathbf{z} , and the decoder decompresses the data into a reconstructed sample $\tilde{\mathbf{x}}$. To learn these transformations, neural networks are used as computational and optimizable building blocks for the encoder and decoder. The encoder and decoder are then optimized according to a loss, which is a low reconstruction error ($\|\mathbf{x} - \tilde{\mathbf{x}}\|$). To generalize to new data points, a VAE imposes a prior over the structure of the vector space \mathbf{z} , and this lower-dimensional space, namely the latent space, is in our case normally distributed. To enforce this constraint, an additional term is introduced in the loss function, the Kullback–Leibler (KL) divergence of the variational approximation²². This term can also be interpreted as a regularization term. It measures how our latent space resembles a normal Gaussian distribution. A cyclical annealing scheduler, which has been proven to be effective in boosting the training performance and mitigating KL vanishing⁴⁶, was also adopted.

Considering that our reticular framework representation, namely the RFcode, is a multiple component input, we build our supramolecular variational autoencoder (SmVAE) with several corresponding components that are in charge of encoding and decoding each part of the RFcode. When properly trained, this model allows us to map the frameworks with discrete representations (RFcodes) into continuous vectors (\mathbf{z}) and then back. As the latent space is a vector space, continuous optimization and search algorithms will be used to find local minima or maxima. By decoding, we can sample and reconstruct new frameworks. To posit information relating structure to physical properties in our latent space, the SmVAE has a property prediction component and is jointly trained for property prediction and framework generation. As the size of our property dataset is much smaller than our structural dataset, we train this component in a semi-supervised fashion. In the joint training, SmVAE was fed with 45,000 MOFs with the property data (textural and gas uptake properties) and another ~2 million MOFs without property data. Predictive network parameters are only optimized when labelled data are observed during training⁴⁷. When the model is correctly trained, we can identify principal axes that align with increasing and decreasing values of physical properties. This feature improves the optimization capabilities of our model. Gas-separation selectivities are then derived using the corresponding uptake values of the gas phases. Taking all the components into account, we propose a multi-component loss function L_{total} as follows:

$$\begin{aligned} L_{\text{total}} &= L_{\text{edge}} + L_{\text{vertex}} + L_{\text{topo}} + L_{\text{property}} + L_{\text{KL}} \\ &= L_{\text{RFcodeRecon.}} + L_{\text{Semi-superProp.}} + L_{\text{VAEConstraint}} \end{aligned} \quad (1)$$

After the realization of property prediction, we ultimately add one property-guided optimization component to the SmVAE for automated reticular framework inverse design. A Gaussian process (GP) model is built and trained with labelled frameworks from the jointly trained latent space from the SmVAE to predict the targeted properties. The entire structure of our SmVAE with all components is illustrated in Fig. 2. GP models are known to be

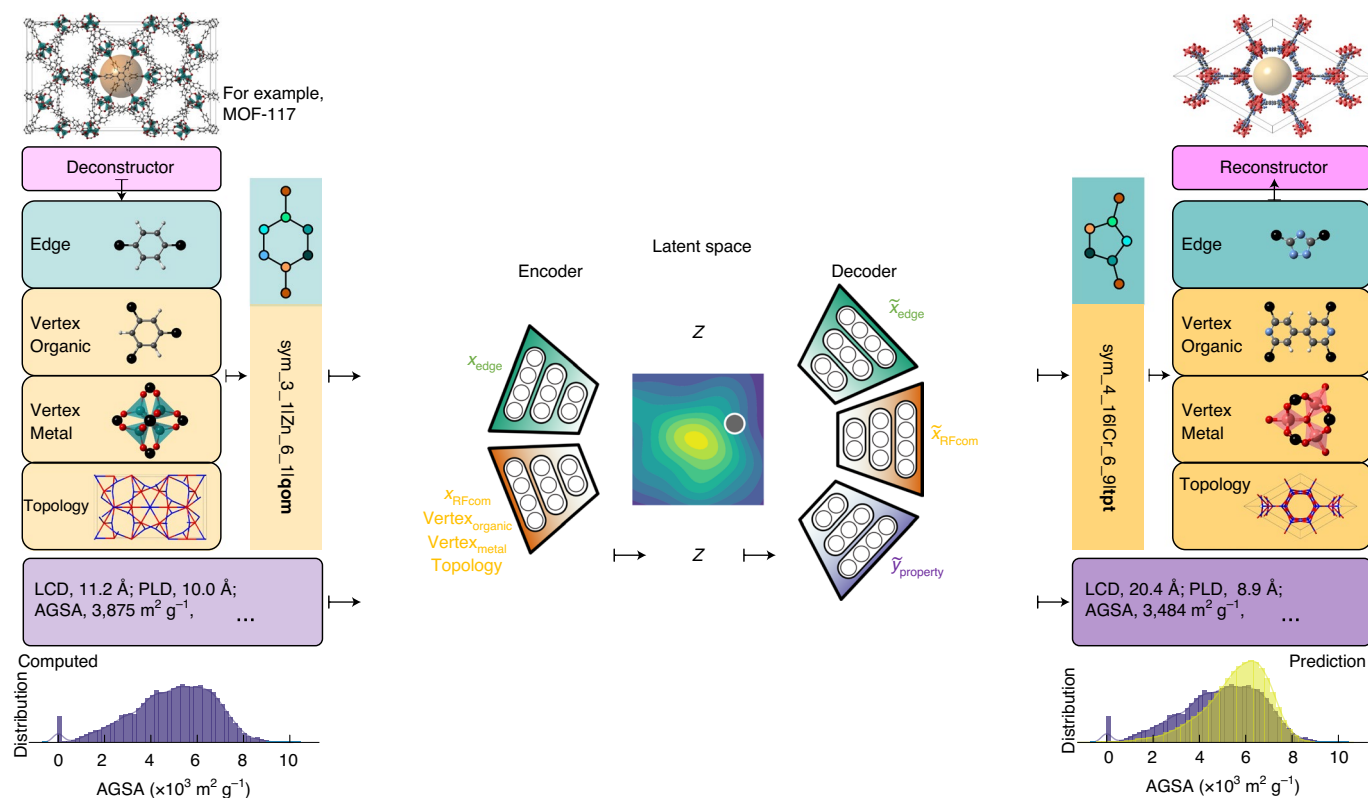


Fig. 2 | Schematic of the automated reticular framework discovery platform empowered by the SmVAE. The SmVAE is a multi-component variational autoencoder with modules that are in charge of encoding and decoding each part of the Rfcode ($x_{\text{edge}} \rightarrow \tilde{x}_{\text{edge}}, x_{\text{Rfcom}} \rightarrow \tilde{x}_{\text{Rfcom}}$). Reticular frameworks are mapped with discrete RfCodes, transferred into continuous vectors (z) and then transferred back. To have the latent space organized around properties of interest, we add an extra component to the model that uses labelled data (y). This process is realized with the additional model that learns to predict properties ($\tilde{y}_{\text{property}}$) from the latent space. Rfcom: components of Rfcode except edge. Topology: **qom**, **tpt**.

effective in prediction with even a limited amount of training data⁴⁸. The detailed SmVAE architecture and hyperparameter tuning process are described in Supplementary Note 2.

Demonstration of SmVAE on MOF design and optimization

To evaluate the fidelity of the trained SmVAE and the capability of its latent space to capture MOF structure information, we estimate the kernel density of each dimension in the latent space (288 in total). As shown in Supplementary Fig. 8, all data distributions in different dimensions are normal, indicating the effectiveness of the variational regularizer as implemented in the SmVAE. Furthermore, we use MOF-117⁴⁰ as an example by feeding its Rfcode to the encoder to obtain its latent representation and sampling its neighbouring latent points at various distances. We check the decoding results of the original representation and neighbouring points. We are able to get the original MOF-117 back at the original point, and decoded MOF structures at the sampled neighbouring points demonstrate more and more variations with increasing distance as shown in Fig. 3c. The autoencoder also provides us a critical opportunity to explore the geometrical correlation between different MOF structures. We encode two well-known yet topologically distinct (topology: **ftw**, **csq**) MOF structures (that is, cubic, **ftw** NU-1104⁴⁹ and hexagonal, **csq** NU-1000⁵⁰) and perform an interpolation between their latent points in space (Fig. 3d). The intermediate frameworks along the interpolation path are then decoded, which demonstrate a clear geometrical evolution from the cubic framework to the hexagonal framework.

Discovering systems with improved properties is the essential goal of materials design. We examine the mapping of property

values to the latent representation in the jointly trained SmVAE latent space using PCA (principal component analysis) (Fig. 3a,b), and we find that the distribution of frameworks shows an explicit gradient, with high-performance MOFs located in one domain and low-performance MOFs in other domains. For comparison, another SmVAE was trained with about two million MOFs without any property as a control group. The resulting latent representation distribution shows no noticeable pattern with respect to property values (Supplementary Fig. 9), confirming the ability of the SmVAE to organize the latent space according to property values. Performance metrics such as prior and posterior scores for sampling and constructing valid MOFs, as well as the mean absolute error (MAE) on predicting MOF properties, are computed and shown in Supplementary Table 2. Our SmVAE demonstrates superb accuracy in designing MOFs and predicting their properties.

Ultimately, we optimize MOF structures in the latent space of the jointly trained autoencoder. We build a GP model, which has been proven to be lightweight and effective for smooth function prediction⁴⁸, to learn the property landscape of the latent representations. A GP model is then trained to predict the target property of the latent vector of a given MOF Rfcode. We then choose CO₂ uptake in the natural gas separation (CO₂/CH₄) as the target and demonstrate two optimization processes: (1) isoreticular MOF design, where the topology is constrained; and (2) globally optimized MOF design (Fig. 4), with maximized property frameworks identified and intermediate structures interpolated. In the isoreticular design process, we pick the MOF NU-1104⁴⁹ (CO₂ uptake of 0.65 mol kg⁻¹) as the starting point and optimized the framework with constrained **ftw** topology. Going through a series of intermediate linkers (Fig. 4b),

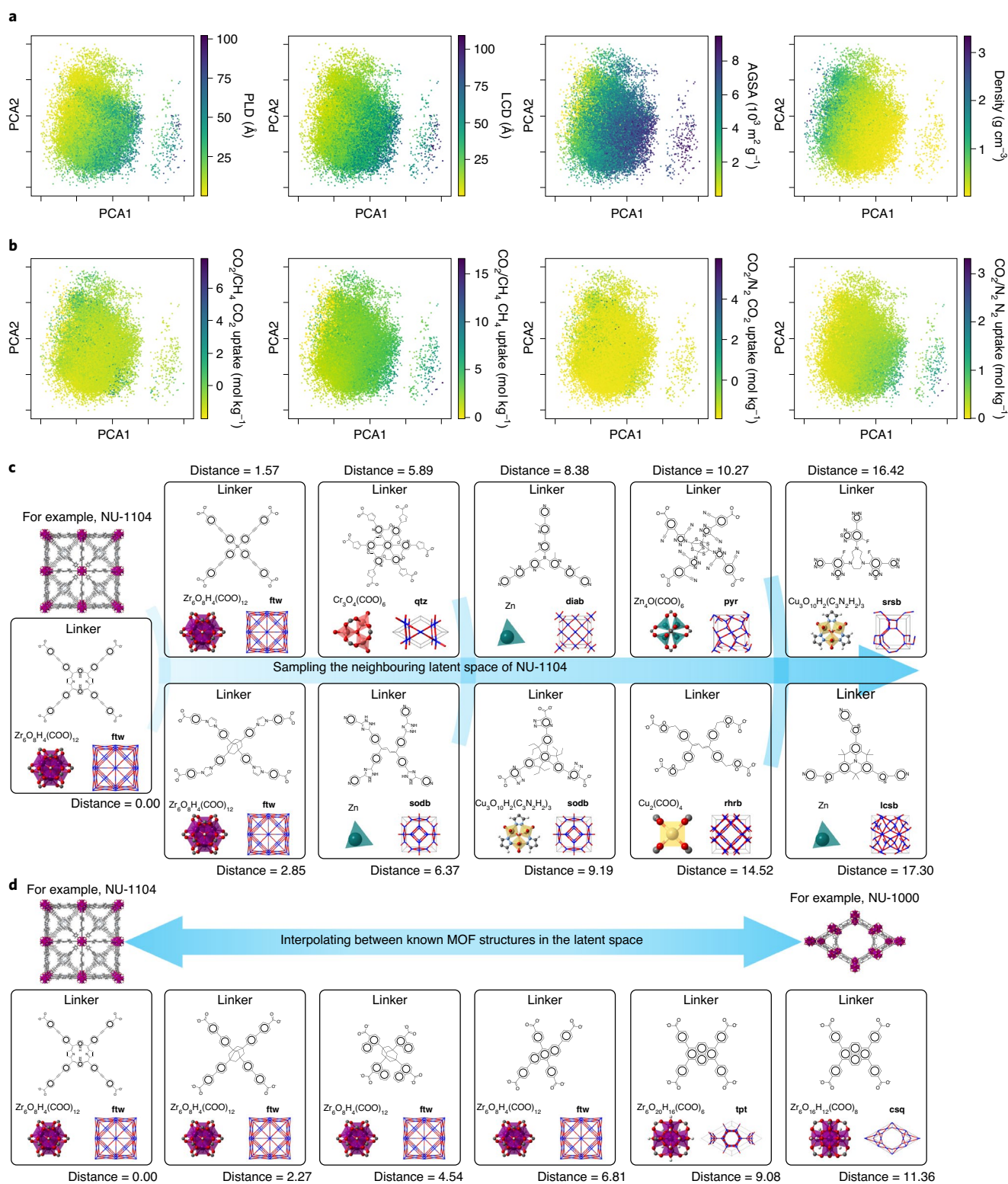


Fig. 3 | Illustration of the latent space of the jointly trained SmVAE using PCA analysis conditioned by MOF properties and exemplified sampling of the latent space. a, b, Latent space of the SmVAE after joint training exhibits notable gradients by textural (a) and gas uptake (b) property values. **c,** Starting with NU-1104⁴⁹, we sample its neighbouring latent points at various distances and check the decoding results. **d,** We interpolate the latent points of two known distinct MOF structures (for example, NU-1104 and NU-1000⁵⁰) and identify the intermediate structures. A clear structure evolution is observed from two geometrically different frameworks. Topology: **ftw**, **qtz**, **diab**, **sodb**, **pyr**, **rhrb**, **srsb**, **lcsb**, **tpt**, **csq**.

we are able to optimize the targeted CO_2 uptake to 4.33 mol kg^{-1} . In the global optimization process without topology constraint, we begin with MOF-5⁵¹ (CO_2 uptake of 2.80 mol kg^{-1} (ref. ⁵²)) and

search for MOFs with optimized uptake (Fig. 4c). At the end, we discover a **spn** (topology) MOF with a remarkably high CO_2 uptake of 7.55 mol kg^{-1} for natural gas separation.

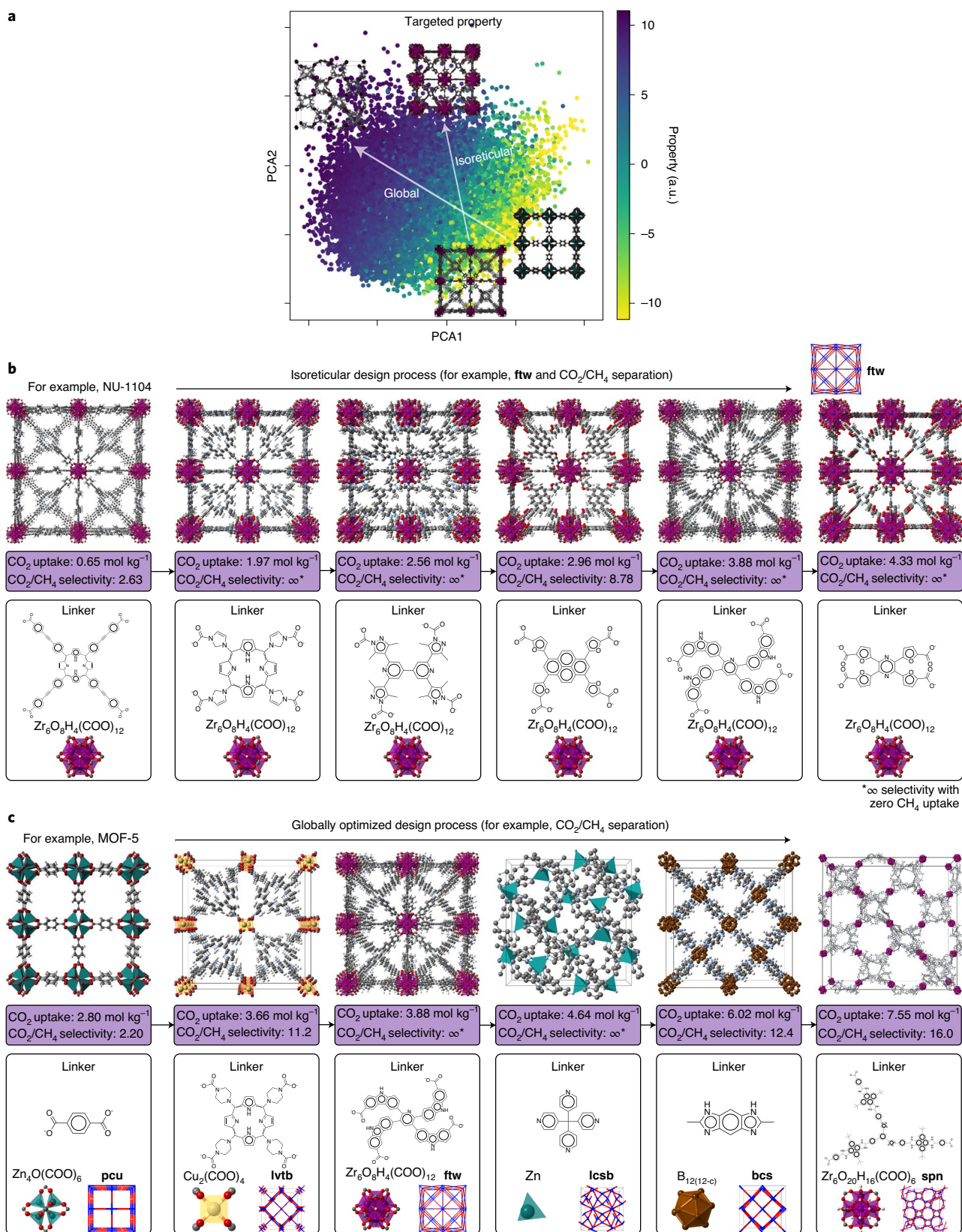


Fig. 4 | Reticular framework design and optimization using the SmVAE with natural gas separation (CO₂ uptake) as the exemplified target.

a–c. Two optimized design processes of isorecticular (**b**) and global (**c**) with the optimization paths for a particular target in the latent space schematically shown (**a**). Through the isorecticular design process (starting with MOF NU-1104⁴⁹) constrained to the **ftw** topology, a MOF with notable CO₂ uptake of 4.33 mol kg⁻¹ and infinite selectivity (zero CH₄ uptake) is discovered (5 bar, 313 K, 10/90 CO₂/CH₄). The global optimization design process without topology constraint (starting with MOF-5⁵¹) leads to a MOF with the remarkably high CO₂ uptake of 7.55 mol kg⁻¹ and high selectivity of 16.0 (5 bar, 313 K, 10/90 CO₂/CH₄). Topology: **ftw**, **pcu**, **lvtb**, **lcsb**, **bcs**, **spn**.

Table 1 | Top GMOF candidates targeted at gas separations (natural gas, CO₂/CH₄; flue gas, CO₂/N₂) sorted with increasing synthesizability SCScore (increasing synthesis complexity)

	GMOF-1	GMOF-2	GMOF-3	GMOF-4	GMOF-5	GMOF-6	GMOF-7	GMOF-8	GMOF-9
Topology	lcsb	ftw	ftw	tpt	spn	spn	spn	spn	spn
Linker									
(SCScore)	1.6	3.2	3.5	3.8	4.8	4.9	4.9	4.9	5.0
Metal node	Zn	Zr ₆ O ₈ H ₄ (COO) ₁₂	Zr ₆ O ₈ H ₄ (COO) ₁₂	Cr ₃ O ₄ (COO) ₆	Zr ₆ O ₂₀ H ₁₆ (COO) ₆	Zr ₆ O ₂₀ H ₁₆ (COO) ₆	Zr ₆ O ₂₀ H ₁₆ (COO) ₆	Zr ₆ O ₂₀ H ₁₆ (COO) ₆	Zr ₆ O ₂₀ H ₁₆ (COO) ₆
CO ₂ capacity (mol kg ⁻¹)	CO ₂ /CH ₄ separation (10:90, 5 bar, 300 K)								
Selectivity	4.64	4.33	4.22	3.97	4.80	4.51	4.34	7.21	7.55
CO ₂ capacity (mol kg ⁻¹)	CO ₂ /N ₂ separation (15:85, 1 bar, 313 K)								
Selectivity	∞	∞	∞	3058.0	10.8	10.0	11.4	17.5	16.0
LCD (Å)	3.06	2.09	2.47	2.51	1.29	1.26	1.45	2.61	2.80
PLD (Å)	∞	∞	48.5	3157.2	12.3	11.7	16.3	27.1	24.6
AGSA (m ² g ⁻¹)	5.79	8.73	9.49	5.56	71.16	70.86	62.61	74.06	81.08
PLD (Å)	3.59	3.46	3.71	3.40	58.63	60.83	55.99	68.92	61.29
AGSA (m ² g ⁻¹)	1337	1184	1470	429	5261	5423	5076	5025	5233

Top MOF candidates proposed for gas separation

Aiming at CO₂ loading in natural gas (5 bar, 300 K, 10/90 CO₂/CH₄) and flue gas separation (1 bar, 313 K, 15/85 CO₂/N₂), we repeat the globally optimized design process and select the top candidates for further validations. When we rank all the generatively designed MOFs (GMOFs), we consider their gas-separation properties as well as the MOF synthesizability to make suggestions for further experimental measurements. To estimate the latter, we calculate the synthetic complexity score (SCScore)³³ of the organic linkers used in the GMOFs. Complete linkers of all MOF candidates are assembled using the appropriate edge and organic vertex, as shown in Supplementary Fig. 10. The ranking procedure for the designed MOFs proceeds as follows: (1) sort them by their CO₂ uptakes in CO₂/CH₄ separation and then select the top nine high-capacity candidate systems; (2) sort them by their synthesizability (reversed SCScore order since higher means more challenging to synthesize). The top candidates with superior performance are shown in Table 1 sorted according to decreasing SCScore. We are able to identify multiple MOFs with enhanced gas-separation properties, including GMOF-9, which shows the highest 7.55 mol kg⁻¹ CO₂ uptake and reasonably large selectivity of 16.0 for CO₂/CH₄ separation. All candidate MOF structures are stable through relaxation, and corresponding properties predicted have been reconfirmed with grand canonical Monte Carlo (GCMC) simulations (Supplementary Fig. 11).

By examining the corresponding porosities, we identify two types of promising MOF with distinct gas-separation mechanisms:

(1) Size exclusion frameworks (GMOF-1, -2, -3 and -4) with small PLD (3.40–3.71 Å) that fall between CO₂ (3.3 Å) and CH₄ (3.8 Å) or N₂ (3.64 Å), therefore effectively permitting CO₂ to diffuse into the MOF while excluding CH₄ or N₂. The MOFs we have identified here (GMOF-1, -2, -3 and -4) have very small PLDs that

do not permit the adsorption of N₂ or CH₄, resulting in a theoretically infinite selectivity. In our GCMC simulations, the MOF atoms are held fixed at their crystallographic positions; however, in reality, some of these MOFs might exhibit a degree of flexibility. This flexibility might permit some adsorption of N₂ or CH₄, which would then bring the selectivities down to high yet finite values. For CO₂/CH₄ separation, they exhibit remarkable CO₂ uptakes (4.64, 4.33, 4.22 and 3.97 mol kg⁻¹, respectively). They are also strong CO₂/N₂ separation candidates with high CO₂ uptakes (3.06, 2.09, 2.47 and 2.51 mol kg⁻¹, respectively) and high selectivities.

(2) Thermodynamic separation frameworks (GMOF-5, -6, -7, -8 and -9), which show large pores (LCD, 62.61–81.08 Å; PLD, 55.99–68.92 Å), compared with the size of the targeted molecules. These MOFs all have high AGSAs (>5,000 m² g⁻¹), offering many binding sites for CO₂, which results in high capacity. They exhibit strong selective CO₂ adsorption as a result of the stronger van der Waals interactions between CO₂ and the frameworks versus CH₄ and N₂. For CO₂/CH₄ separation, we observe notably high CO₂ uptakes (4.80, 4.51, 4.34, 7.21 and 7.55 mol kg⁻¹, respectively) at reasonably high CO₂/CH₄ selectivities (10.0–17.5). They are also competent flue gas-separation materials with reasonable CO₂ uptakes (1.29, 1.26, 1.45, 2.61 and 2.80 mol kg⁻¹, respectively) and good CO₂/N₂ selectivities (11.7–27.1).

Performance comparison on gas separations between MOFs is practically difficult as the measurements are often conducted at different experimental conditions (for example, temperature, pressure and gas phase composition). However, we believe that the MOFs discovered here are strongly competitive against some of the best-performing MOFs/zeolites ever reported in the literature (Table 2). Our top candidates show high-performance for natural gas separation (that is, GMOF-8, 7.21 mol kg⁻¹; GMOF-9, 7.55 mol kg⁻¹) at a condition of 5 bar, 300 K with a low CO₂:CH₄

Table 2 | Gas-separation performance of well-known MOFs and zeolites

	SIFSIX-2-Cu-i	SIFSIX-3-Zn	Mg-MOF-74	UTSA-16	13X	CuBTC	ZIF-8	IRMOF-1
CO ₂ /CH ₄ separation								
	(50:50, 1 bar, 298 K) ^{9a}	(50:50, 1 bar, 298 K) ^{9a}	(50:50, 5 bar, 313 K) ⁵⁴	(50:50, 2 bar, 296 K) ⁵	(50:50, 5 bar, 313 K) ^{54a}	(25:75, 5 bar, 303 K) ⁷¹	(10:90, 5 bar, 293 K) ^{72a}	(10:90, 5 bar, 298 K) ^{73,74a}
CO ₂ capacity (mol kg ⁻¹)	4.16	2.46	8.0	4.25	4.4	3.6	0.48	0.79
Selectivity	33	231	105.1	29.8	36	7.2	3.95	3.86
CO ₂ /N ₂ separation								
	(10:90, 1 bar, 298 K) ^{9a}	(10:90, 1 bar, 298 K) ^{9a}	(15:75, 0.9 bar, 313 K) ⁵⁶	(15:85, 1 bar, 296 K) ⁵	(16:84, 1.1 bar, 288 K) ⁵⁷	(15:85, 1 bar, 296 K) ^{73,74a}	(15:85, 1 bar, 296 K) ^{72a}	(15:85, 1 bar, 296 K) ^{73,74a}
CO ₂ capacity (mol kg ⁻¹)	1.59	2.27	4.43	2.37	3.0	0.85	0.15	0.26
Selectivity	140	1818	175	314.7	20	24	11.7	11.1

^aIdeal adsorbed solution theory^{75,76}.

ratio of 1:9 while the notable Mg-MOF-74 and zeolite 13X show CO₂ comparable or even lower capacities (8.0 mol kg⁻¹, 4.4 mol kg⁻¹)⁵⁴ at 5 bar, 313 K, with higher CO₂:CH₄ ratio of 1:1 (ref. ⁵⁴). SIFSIX-2-Cu-i, SIFSIX-3-Zn and UTSA-16 exhibit capacities of 4.16 mol kg⁻¹ (ref. ⁵), 2.46 mol kg⁻¹ (ref. ⁵) and 4.25 mol kg⁻¹ (ref. ⁵⁵), respectively, at lower-pressure conditions (1–2 bar, ~300 K) with a CO₂:CH₄ ratio of 1:1. Their selectivities against CH₄ (29.8–231)^{5,54,55} are all lower than our top selectivity candidates (that is, GMOF-1, -2, -3 and -4: 3,058–∞ with zero CH₄ uptake). For flue gas separation at similar conditions as this study (1 bar, 313 K, 15/85 CO₂/N₂), our top candidate GMOF-1 exhibits CO₂ uptake of 3.06 mol kg⁻¹ with extremely high selectivity (∞ with zero N₂ uptake), which is only lower than the capacity of Mg-MOF-74: 4.43 mol kg⁻¹ with a selectivity of 175 (0.9 bar, 313 K, 15/75 CO₂/N₂)⁵⁶, while higher than SIFSIX-2-Cu-i (1.59 mol kg⁻¹ at 140 selectivity)⁵, SIFSIX-3-Zn (2.27 mol kg⁻¹ at 1,818 selectivity)⁵, UTSA-16 (2.37 mol kg⁻¹ at 314.7 selectivity)⁵⁵, and 13X (3.0 mol kg⁻¹ at 20 selectivity)⁵⁷. Furthermore, our top candidates show potentially strong chemical and hydrothermal stabilities, with the exclusive usage of well-known stable metal nodes such as Zr₆O₈/Zr₆O₂₀, Cr₃O₄ and Zn₄ (ref. ⁵⁸). This is particularly important for carbon capture applications in a harsh flue gas environment⁵⁹.

Conclusions

We developed an automated nanoporous materials discovery platform using a supramolecular variational autoencoder for the generation of reticular frameworks with optimized properties. We have demonstrated the automated design process with MOF structures starting from the computation-ready, experimental (CoRE) MOF database²⁶ and generating new proposed structures with improved properties for CO₂ separations. Our model exhibits high fidelity in capturing structural features and reconstructing MOF structures. The autoencoder shows great prediction and optimization capability when jointly trained with multiple top candidates identified for superior gas separation and confirmed via atomistic Monte Carlo simulations. We use this platform to design novel MOFs with improved capacity and good selectivity for CO₂/N₂ and CO₂/CH₄ separations, which are important clean-energy-relevant applications. The top-performing MOF has a CO₂ capacity of 7.55 mol kg⁻¹ and a selectivity over CH₄ of 16. This platform can be applied to a broad range of materials (for example, covalent organic frameworks, metal–organic polyhedra, hydrogen-bonded organic frameworks and coordinational polymers) and lays the groundwork for the design of reticular frameworks for a variety of applications.

Methods

Reticular framework textural and gas-separation property calculations. We performed computational simulations on around 45,000 randomly selected MOF structures from the augmented MOF set of two million. We calculate the textural

properties of the MOF crystals including PLD, LCD and AGSA using Zeo++⁶⁰ with high-accuracy settings (-ha flag), and a hard sphere with a diameter of 3.31 Å (the Lennard–Jones σ parameter of nitrogen in the TraPPE model⁶¹). We optimize the geometry of the MOF structure in the Forcite module of Materials Studio⁶² using the universal force field (UFF)⁶³ through a two-step process. In the first phase, the cell shape and size are held constant while the atom positions are moved, and then in the second phase, the cell shape is also allowed to change. The distribution of the sampled AGSAs are shown in Fig. 2, and the distributions of the remaining textural properties are shown in Supplementary Fig. 6. We select gas separations as the targeted applications (removal of CO₂ from natural gas and flue gas) and calculated MOF properties for CO₂/CH₄ and CO₂/N₂ separation. Partial charges on the framework atoms are computed using the SQE-MEPO method of Collins and Woo⁶⁴, which is an empirical charge model that has been fit to reproduce density functional theory (DFT) derived electrostatic potential fitted⁶⁵ charges in MOFs and yields accurate results for CO₂ adsorption. These charges are fed into GCMC simulations, which are performed using our in-house multipurpose simulation code RASPA⁶⁶. We use the Lennard–Jones parameters from the UFF⁶³ for the MOF framework atoms, and the MOF structures are held fixed during the simulations. We use the TraPPE models for CO₂ (ref. ⁶⁷), N₂ (ref. ⁶⁷) and CH₄ (ref. ⁶⁸). Van der Waals interactions beyond 12.8 Å are neglected, and tail corrections are not used. We use a sufficient number of unit cells so that the simulation box exceeds 25.6 Å in all dimensions. Coulomb interactions are computed using the Ewald summation method. For the GCMC simulations, we use 4,000 initialization cycles and 4,000 production cycles. We use pore blocking to prevent insertions into cavities that are inaccessible to the adsorbate molecules. Therefore, some reticular frameworks may have extremely small or even zero uptakes of the larger radius molecules compared with CO₂, like CH₄ and N₂. As a result, these frameworks will have derived enormous or infinite selectivity of CO₂ against CH₄ and N₂. The pore-blocking spheres are computed in Zeo++⁶¹. The Monte Carlo moves are translation, regrow, swap (insert/delete) and identity change with a relative probability of 1, 1, 1 and 2, respectively. We compute the CO₂ and N₂ uptake from a 15/85 mole fraction mixture of CO₂/N₂ at 1 bar and 313 K, and we compute the CO₂ and CH₄ uptake from a 10/90 mole fraction mixture of CO₂/CH₄ at 5 bar and 300 K. To confirm the accuracy of our simulation methodology, we computed isotherms at 298 K for CO₂ adsorption in IRMOF-1 and IRMOF-3 and compared them with the experimental counterparts. Agreement between simulation and experiment is achieved, as shown in Supplementary Fig. 1.

Autoencoder architecture and hyperparameter tuning. The multi-component SmVAE consists of an edge encoder/decoder, a reticular framework information encoder/decoder and a property predictor. The edge encoder and decoder are paired recurrent neural networks. Edge molecular SMILES are converted to the semantically constrained graph-based strings (SELFIES)³⁸, and the strings are then encoded and decoded in a sequence-to-sequence manner. There is no length limit set for the SMILES and SELFIES strings. In the edge encoder, one gated recurrent unit⁶⁸ layer with a hidden dimension of 768 is implemented. In the edge decoder, another gated recurrent unit hidden layer of 704 is used. Reticular framework information, including the vertices (organic and metal) and topology types, are one-hot encoded as categories. The information is fed into the reticular framework encoder and decoder containing two and one fully connected hidden layers, respectively. For property prediction (that is, textural and gas uptake properties), three layers of networks are used to predict properties from the latent space. All the encoded edges, framework information and properties during joint training are passed to a comprehensive latent space with a dimension of 288. Considering that we only have properties for part of our training framework set (~45,000 of 2 million), a masked function is used to colour only the latent points with properties determined and realize the semi-supervised learning.

During training, a cyclical annealing scheduler⁴⁶ is adopted with a period of 15 epochs, and the full training runs for 120 epochs in total. The property and

reticular framework information prediction loss annealing is initiated at the same time with the variational loss in the joint training and stops before the last ten epochs. A random optimization of 200 trials is conducted to optimize the key hyperparameters of the full model, using the prior and posterior validities as criteria. When evaluating the performance of the SmVAE, prior validity is calculated by randomly sampling 10,000 points from the trained latent space and counting the number of valid reticular frameworks decoded (simultaneously valid edge SMILES and reticular information). Posterior validity and reconstruction ratio are calculated by randomly sampling 1,000 MOF structures and feeding them into the SmVAE. Then the decoding is tried ten times, and we check how many of the decoded MOFs are valid and whether the original MOF can be reconstructed. Meanwhile, MAEs are computed for all properties to estimate the general accuracy of the SmVAE on property predictions compared with the geometric results (textural properties) from Zeo++⁶⁰ or the gas adsorption results (gas uptake properties) from the GCMC simulations (Supplementary Table 2).

The GP model for the identification of reticular frameworks with the optimized target property is trained with the latent vectors of 5,000 randomly selected MOFs and their corresponding properties. With this model, we are able to search through the whole reticular framework latent space and maximize the target property. We decided to use a GP model instead due to the following reasons. (1) Empirically, we found that GP has fewer local minima and therefore tends to converge faster using an optimization algorithm²². (2) GP models provide prediction uncertainty estimates which are useful for Bayesian optimization settings. (3) We wish to showcase that other optimization strategies (evolution strategies and so on) can be adopted with the latent vectors. The regression statistics of GP are shown in Supplementary Table 3 with MAEs of all properties are no larger in magnitude compared with the counterparts of the property prediction networks in SmVAE, as shown in Supplementary Table 2. For the top candidates newly designed, as shown in Table 1, we calculated all their textural and gas uptake properties using Zeo++⁶⁰ and GCMC simulations. Gas-separation selectivities of these candidates are then derived using the gas uptake values of corresponding phases. To further confirm the effectiveness of the GP property predictor, we made parity plots between GP predicted and GCMC computed textural and gas-separation properties (Supplementary Fig. 11) for all the simulated frameworks (45,000 structures + top candidates). Great agreements between the predicted and computed properties can be then observed.

We use the Pytorch packages⁶⁹ to build and train this model and the RDKit⁷⁰ package for cheminformatics.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data for the training of the SmVAE including the augmented two million MOF set and the tabulated textural and gas-separation property data for the randomly selected MOF structures are available at https://github.com/zhenpengyao/Supramolecular_VAE/tree/master/data.

Code availability

Code for the SmVAE is available at <https://doi.org/10.24433/CO.8185164.v1>.

Received: 7 May 2020; Accepted: 9 November 2020;

Published online: 11 January 2021

References

- Yaghi, O. M. et al. Reticular synthesis and the design of new materials. *Nature* **423**, 705–714 (2003).
- Li, H., Eddaoudi, M., Groy, T. L. & Yaghi, O. M. Establishing microporosity in open metal–organic frameworks: gas sorption isotherms for Zn(BDC) (BDC = 1,4-benzenedicarboxylate). *J. Am. Chem. Soc.* **120**, 8571–8572 (1998).
- Mason, J. A. et al. Methane storage in flexible metal–organic frameworks with intrinsic thermal management. *Nature* **527**, 357–361 (2015).
- Chen, K.-J. et al. Synergistic sorbent separation for one-step ethylene purification from a four-component mixture. *Science* **366**, 241–246 (2019).
- Nugent, P. et al. Porous materials with optimal adsorption thermodynamics and kinetics for CO₂ separation. *Nature* **495**, 80–84 (2013).
- Diercks, C. S., Liu, Y., Cordova, K. E. & Yaghi, O. M. The role of reticular chemistry in the design of CO₂ reduction catalysts. *Nat. Mater.* **17**, 301–307 (2018).
- Hu, Z., Deibert, B. J. & Li, J. Luminescent metal–organic frameworks for chemical sensing and explosive detection. *Chem. Soc. Rev.* **43**, 5815–5840 (2014).
- Sheberla, D. et al. Conductive MOF electrodes for stable supercapacitors with high areal capacitance. *Nat. Mater.* **16**, 220–224 (2017).
- Tan, L. L. et al. Stimuli-responsive metal–organic frameworks gated by pillar[5]arene supramolecular switches. *Chem. Sci.* **6**, 1640–1644 (2015).
- Li, M., Li, D., O’Keeffe, M. & Yaghi, O. M. Topological analysis of metal–organic frameworks with polytopic linkers and/or multiple building units and the minimal transitivity principle. *Chem. Rev.* **114**, 1343–1370 (2014).
- Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432**, 823 (2004).
- Wilmer, C. E. et al. Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* **4**, 83–89 (2012).
- Boyd, P. G. et al. Data-driven design of metal–organic frameworks for wet flue gas CO₂ capture. *Nature* **576**, 253–256 (2019).
- Collins, S. P., Daff, T. D., Piotrkowski, S. S. & Woo, T. K. Materials design by evolutionary optimization of functional groups in metal–organic frameworks. *Sci. Adv.* **2**, e1600954 (2016).
- Chung, Y. G. et al. In silico discovery of metal–organic frameworks for precombustion CO₂ capture using a genetic algorithm. *Sci. Adv.* **2**, e1600909 (2016).
- Fernandez, M., Boyd, P. G., Daff, T. D., Aghaji, M. Z. & Woo, T. K. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO₂ capture. *J. Phys. Chem. Lett.* **5**, 3056–3060 (2014).
- Moghadam, P. Z. et al. Structure–mechanical stability relations of metal–organic frameworks via machine learning. *Matter* **1**, 219–234 (2019).
- Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
- Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014—Conference Track Proceedings* (International Conference on Learning Representations, 2014).
- Goodfellow, I. J. et al. Generative adversarial networks. Preprint at <https://arxiv.org/abs/1406.2661> (2014).
- Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
- Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
- Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In *Proc. 35th International Conference on Machine Learning ICML 2018* Vol. 5 3632–3648 (IMLS, 2018).
- Noh, J. et al. Inverse design of solid-state materials via a continuous representation. *Matter* **1**, 1370–1384 (2019).
- Kim, B., Lee, S. & Kim, J. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* **6**, eaax9324 (2020).
- Chung, Y. G. et al. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019. *J. Chem. Eng. Data* <https://doi.org/10.1021/acs.jced.9b00835> (2019).
- Duvenaud, D. et al. Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process. Syst.* **2**, 2224–2232 (2015).
- Krenn, M., Hase, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* <https://doi.org/10.1088/2632-2153/aba947> (2020).
- Li, P. et al. Bottom-up construction of a superstructure in a porous uranium–organic crystal. *Science* **356**, 624–627 (2017).
- Jain, A. et al. Commentary: The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
- Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, (2013).
- Isayev, O. et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, 15679 (2017).
- Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **9**, 2775 (2018).
- Ryan, K., Lengyel, J. & Shatruk, M. Crystal structure prediction via deep learning. *J. Am. Chem. Soc.* **140**, 10158–10168 (2018).
- Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
- Park, C. W. & Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* **4**, 063801 (2020).
- Eon, J. G. Topological features in crystal structures: a quotient graph assisted analysis of underlying nets and their embeddings. *Acta Crystallogr. A* **72**, 268–293 (2016).
- Delgado-Friedrichs, O., Hyde, S. T., O’Keeffe, M. & Yaghi, O. M. Crystal structures as periodic graphs: the topological genome and graph databases. *Struct. Chem.* **28**, 39–44 (2017).
- O’Keeffe, M. & Yaghi, O. M. Deconstructing the crystal structures of metal–organic frameworks and related materials into their underlying nets. *Chem. Rev.* **112**, 675–702 (2012).
- Furukawa, H., Kim, J., Ockwig, N. W., O’Keeffe, M. & Yaghi, O. M. Control of vertex geometry, structure dimensionality, functionality, and pore metrics in the reticular synthesis of crystalline metal–organic frameworks and polyhedra. *J. Am. Chem. Soc.* **130**, 11650–11661 (2008).

41. 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).
42. Anderson, R. & Gómez-Gualdrón, D. A. Increasing topological diversity during computational “synthesis” of porous crystals: how and why. *CrystEngComm* **21**, 1653–1665 (2019).
43. Gherzi, D. & Singh, M. molBLOCKS: decomposing small molecule sets and uncovering enriched fragments. *Bioinformatics* **30**, 2081–2083 (2014).
44. Colón, Y. J., Gómez-Gualdrón, D. A. & Snurr, R. Q. Topologically guided, automated construction of metal–organic frameworks and their evaluation for energy-related applications. *Cryst. Growth Des.* **17**, 5801–5810 (2017).
45. Kingma, D. P. & Welling, M. An introduction to variational autoencoders. *Found. Trends Mach. Learn.* **12**, 307–392 (2019).
46. Fu, H. et al. Cyclical annealing schedule: a simple approach to mitigating. In *Proc. 2019 Conference of the North 240–250* (Association for Computational Linguistics, 2019); <https://doi.org/10.18653/v1/N19-1021>
47. Kingma, D. P., Rezende, D. J., Mohamed, S. & Welling, M. Semi-supervised learning with deep generative models. *Adv. Neural Inf. Process. Syst.* **4**, 3581–3589 (2014).
48. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, 2005); <https://doi.org/10.7551/mitpress/3206.001.0001>
49. Deria, P. et al. Ultraporous, water stable, and breathing zirconium-based metal–organic frameworks with ftw topology. *J. Am. Chem. Soc.* **137**, 13183–13190 (2015).
50. Mondloch, J. E. et al. Vapor-phase metalation by atomic layer deposition in a metal–organic framework. *J. Am. Chem. Soc.* **135**, 10294–10297 (2013).
51. Li, H., Eddaoudi, M., O’Keeffe, M. & Yaghi, O. M. Design and synthesis of an exceptionally stable and highly porous metal–organic framework. *Nature* **402**, 276–279 (1999).
52. Gu, Z. Y., Jiang, J. Q. & Yan, X. P. Fabrication of isoreticular metal–organic framework coated capillary columns for high-resolution gas chromatographic separation of persistent organic pollutants. *Anal. Chem.* **83**, 5093–5100 (2011).
53. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).
54. Herm, Z. R., Krishna, R. & Long, J. R. CO₂/CH₄, CH₄/H₂ and CO₂/CH₄/H₂ separations at high pressures using Mg₂(dobdc). *Micropor. Mesopor. Mater.* **151**, 481–487 (2012).
55. Xiang, S. et al. Microporous metal–organic framework with potential for carbon dioxide capture at ambient conditions. *Nat. Commun.* **3**, 954 (2012).
56. Mason, J. A., Sumida, K., Herm, Z. R., Krishna, R. & Long, J. R. Evaluating metal–organic frameworks for post-combustion carbon dioxide capture via temperature swing adsorption. *Energy Environ. Sci.* **4**, 3030–3040 (2011).
57. Cavenati, S., Grande, C. A. & Rodrigues, A. E. Adsorption equilibrium of methane, carbon dioxide, and nitrogen on zeolite 13X at high pressures. *J. Chem. Eng. Data* **49**, 1095–1101 (2004).
58. Howarth, A. J. et al. Chemical, thermal and mechanical stabilities of metal–organic frameworks. *Nat. Rev. Mater.* **1**, 15018 (2016).
59. Rieth, A. J., Wright, A. M. & Dincă, M. Kinetic stability of metal–organic frameworks for corrosive and coordinating gas capture. *Nat. Rev. Mater.* **4**, 708–725 (2019).
60. Willems, T. F., Rycroft, C. H., Kazi, M., Meza, J. C. & Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Micropor. Mesopor. Mater.* **149**, 134–141 (2012).
61. Bae, Y. S., Yazaydn, A. Ö. & Snurr, R. Q. Evaluation of the BET method for determining surface areas of MOFs and zeolites that contain ultra-micropores. *Langmuir* **26**, 5475–5483 (2010).
62. Biovia, D. S. Materials Studio (San Diego Dassault Systèmes, 2019).
63. Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A. & 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).
64. Collins, S. P. & Woo, T. K. Split-charge equilibration parameters for generating rapid partial atomic charges in metal–organic frameworks and porous polymer networks for high-throughput screening. *J. Phys. Chem. C* **121**, 903–910 (2017).
65. Campaña, C., Mussard, B. & Woo, T. K. Electrostatic potential derived atomic charges for periodic systems using a modified error functional. *J. Chem. Theory Comput.* **5**, 2866–2878 (2009).
66. 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).
67. Martin, M. G. & Siepmann, J. I. Transferable potentials for phase equilibria. 1. United-atom description of *n*-alkanes. *J. Phys. Chem. B* **102**, 2569–2577 (1998).
68. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. Preprint at <https://arxiv.org/abs/1412.3555> (2014).
69. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. Preprint at <https://arxiv.org/abs/1912.01703> (2019).
70. Landrum, G. *RDKit: Open-source Cheminformatics Software* (RDKit, 2006); <http://www.rdkit.org>
71. Hamon, L., Jolimaître, E. & Pirngruber, G. D. CO₂ and CH₄ separation by adsorption using Cu-BTC metal–organic framework. *Ind. Eng. Chem. Res.* **49**, 7497–7503 (2010).
72. Liu, H. et al. A hybrid absorption–adsorption method to efficiently capture carbon. *Nat. Commun.* **5**, 5147 (2014).
73. Millward, A. R. & Yaghi, O. M. Metal–organic frameworks with exceptionally high capacity for storage of carbon dioxide at room temperature. *J. Am. Chem. Soc.* **127**, 17998–17999 (2005).
74. Li, J., Li, J., Yang, J. & Li, L. Separation of CO₂/CH₄ and CH₄/N₂ mixtures using MOF-5 and Cu₂(BTC)₂. *J. Energy Chem.* **23**, 453–460 (2014).
75. Myers, A. L. & Prausnitz, J. M. Thermodynamics of mixed-gas adsorption. *AIChE J.* **11**, 121–127 (1965).
76. Simon, C. M., Smit, B. & Haranczyk, M. PyIAST: ideal adsorbed solution theory (IAST) Python package. *Comp. Phys. Commun.* **200**, 364–380 (2016).

Acknowledgements

Z.Y., N.S.B., B.J.B., S.G.H.K., O.K.F., R.Q.S. and A.A.-G. were supported as part of the Nanoporous Materials Genome Center by the US Department of Energy, Office of Science, Office of Basic Energy Sciences under award number DE-FG02-17ER16362. Funding for T.B., S.P.C. and T.K.W. were provided by NSERC. Computations were made on the supercomputer ‘beluga’ from École de technologie supérieure, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l’Économie, de la science et de l’innovation du Québec (MESI) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT). This research was supported in part through the computational resources and staff contributions provided for the Quest high-performance computing facility at Northwestern University, which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. A.A.-G. is a Canadian Institute for Advanced Research (CIFAR) Lebovic Fellow.

Author contributions

Z.Y. conceived the overall project. Z.Y., B.J.B. and R.Q.S. designed the reticular framework representation approach. N.S.B. and R.Q.S. conducted the MOF property determination calculations. Z.Y. and B.S.-L. developed the deep learning variational autoencoder. S.P.C., T.B. and T.K.W. did the charge calculations for the framework charges for property simulations. A.A.-G. led the project and provided the overall directions. All authors participated in preparing the manuscript.

Competing interests

O.K.F. and R.Q.S. have a financial interest in NuMat Technologies, a startup company that is seeking to commercialize MOFs.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s42256-020-00271-1>.

Correspondence and requests for materials should be addressed to Z.Y., R.Q.S. or A.A.-G.

Peer review information *Nature Machine Intelligence* thanks Jihan Kim, Joshua Schrier and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

© The Author(s), under exclusive licence to Springer Nature Limited 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-----|-----------|
| n/a | Confirmed |
|-----|-----------|
- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
 - A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
 - The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
 - A description of all covariates tested
 - A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
 - A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
 - For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
 - For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
 - For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
 - Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Zeo++ (<https://doi.org/10.1016/j.micromeso.2011.08.020>), Materials Studio, RASPA (<https://doi.org/10.1080/08927022.2015.1010082>), and Tobacco (https://github.com/tobacco-mofs/tobacco_3.0) were used for data collection.

Data analysis Scikit-learn library was used for data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data for the training of the SmVAE including the augmented 2 million MOF set and the tabulated textural and gas separation property data for the randomly selected MOF structures are available at https://github.com/zhenpengyao/Supramolecular_VAE/tree/master/data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<i>Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.</i>
Data exclusions	<i>Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.</i>
Replication	<i>Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.</i>
Randomization	<i>Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.</i>
Blinding	<i>Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.</i>

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	<i>Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).</i>
Research sample	<i>State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.</i>
Sampling strategy	<i>Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.</i>
Data collection	<i>Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.</i>
Timing	<i>Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.</i>
Data exclusions	<i>If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.</i>
Non-participation	<i>State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.</i>
Randomization	<i>If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.</i>

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	<i>Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.</i>
Research sample	<i>Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i>, all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National</i>

Research sample *Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*

Sampling strategy *Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*

Data collection *Describe the data collection procedure, including who recorded the data and how.*

Timing and spatial scale *Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*

Data exclusions *If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*

Reproducibility *Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*

Randomization *Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*

Blinding *Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? Yes No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involvement
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging