A Universal Foundation Model for Transfer Learning in Molecular Crystals

Minggao Feng[†1], Chengxi Zhao[†1], Graeme M. Day[2], Xenophon Evangelopoulos[1,3*], Andrew I. Cooper[1,3*]

[1]Materials Innovation Factory and Department of Chemistry, University of Liverpool, Liverpool, UK.
[2]School of Chemistry and Chemical Engineering, University of Southampton, Southampton, UK.
[3]Leverhulme Research Centre for Functional Materials Design, Liverpool, UK.
[†]These authors contributed equally to this work.

*Corresponding author(s). E-mail(s): evangx@liverpool.ac.uk; aicooper@liverpool.ac.uk;

Abstract

The physical and chemical properties of molecular crystals are a combined function of molecular structure and the molecular crystal packing. Specific crystal packings can enable applications such as pharmaceuticals, organic electronics, and porous materials for gas storage. However, to design such materials, we need to predict both crystal structure and the resulting physical properties, and this is expensive using traditional computational methods. Machine-learned interatomic potential methods offer major accelerations here, but molecular crystal structure prediction remains challenging due to the weak intermolecular interactions that dictate crystal packing. Moreover, machine-learned interatomic potentials do not accelerate the prediction of all physical properties for molecular crystals. Here we present Molecular Crystal Representation from Transformers (MCRT), a transformer-based model for molecular crystal property prediction that is pre-trained on 706,126 experimental crystal structures extracted from the Cambridge Structural Database (CSD). MCRT employs four different pre-training tasks to extract both local and global representations from the crystals using multi-modal features to encode crystal structure and geometry. MCRT has the potential to serve as a universal foundation model for predicting a range of properties for molecular crystals, achieving state-of-the-art results even when fine-tuned on small-scale datasets. We demonstrate MCRT's practical utility in both crystal property prediction and crystal structure prediction. We also show that model predictions can be interpreted by using attention scores.

Keywords: transfer learning, molecular crystals, transformer model, property prediction.

1 Introduction

Molecular crystals have diverse applications including pharmaceuticals [1], organic electronics [2], optical materials [3], and materials for gas storage and separation [4–6]. In all cases, the properties of molecular crystals depend on the crystal packing. For example, pharmaceutical molecules can have widely different solubilities depending on the crystalline form, and in organic electronics, charge transport is critically dependent on crystal packing. However, molecular crystal packing is notoriously difficult to predict because it is dictated by a range of weak, subtle intermolecular interactions, such as van der Waals forces, aromatic pi-stacking, and hydrogen bonds [7]. This is a major hurdle for digital material design because if we cannot predict crystal structure then we cannot, by definition, predict the functional properties of the crystal. To address this challenge, crystal structure prediction (CSP) methods have been created to identify molecular crystals with specific target functionalities. For example, the energy–structure–function (ESF) maps have guided the synthesis of various functional molecular crystals [6–10]. However, despite these successes, calculating the physical properties for each structure on an ESF map, or even a sub-set of low-energy structures, can be computationally demanding. This problem is two-fold: the prediction of lattice energy, or crystal stability, is itself computationally expensive, and the functional property calculations are usually even more expensive. To tackle this, there has been a surge of interest in machine learning (ML) techniques for the rapid prediction of materials properties and the elucidation of structure-property relationship [11, 12] at a fraction of the cost of first-principles methods, such as density functional theory (DFT).

Learning accurate representations is a crucial aspect of machine learning theory that also extends to learning molecular representations. Different types of materials pose different challenges when learning accurate latent representations. For example, in solid-state systems it is essential to capture features such as long-range interactions and periodicity in property prediction tasks [13]. This is especially challenging in organic molecular crystals due to the intermolecular interactions [7], which are typically weaker than for ionic inorganic materials. Hence, many of the inherently local graph-based deep architectures fail to capture global-driven properties, while traditional ML models that use hand-crafted descriptors have proved more successful in capturing spatial information in some cases [14, 15]. Hand-crafted descriptors such as smooth overlap of atomic positions (SOAP) [16] and atom-centered symmetry functions (ACSFs) [17] have proven effective in predicting properties like lattice energy [12], while geometric descriptors such as accessible surface area and pore diameters have been used to predict the methane deliverable capacity of molecular crystals [18] as well as other global-driven properties of porous materials [19]. More recently, persistent homology [20, 21] has been shown to successfully encode global molecular geometric features into machine-learned representations [22]. Nevertheless, calculating descriptors such as SOAP can be cumbersome in terms of memory footprint for larger organic systems, whereas geometric descriptors tend to overly compress the geometric information of the crystals, failing to adequately encode the fine detail of the molecular geometry. Another fundamental drawback of deep learning models is the need for re-training and hyperparameter re-optimization for each specific problem and property, adding further time and computational cost. A further challenge is the availability of training data: ideally, we need methods that can be fine-tuned on small scale datasets, because for chemistry problems, data is often scarce and expensive.

Transfer learning allows a model trained on one task to be adapted to a different task, significantly reducing the need for extensive retraining. Universality is a key aspect of a pre-trained model to allow it to capture simultaneously molecular features of varying modalities, as well as local and global interactions. Recently, pre-trained models using transformers [23] have been designed for metal–organic frameworks (MOFs) and showed exceptional performance across a range of different tasks [24–27]. Transformers enable multi-modal input integration combined with self-attention layers that can process data sequences in parallel, allowing for much more efficient training routines. Also, the attention scores (AS) within the self-attention layers can be used to analyse feature importance and thus offer an interpretability tool to gain insights on the prediction process itself, unlike other black-box learning systems. A leading example is BERT [28], a pre-trained language transformer model that shows state-of-the-art results across various downstream tasks after being trained on large-scale data. More recently, vision transformers architectures (ViTs) [29] have paved the way for the integration of multi-modal inputs towards more universal models [30–32] and inspired a number of recent works in materials science [24–27, 33–39].

There are two key challenges when designing the pre-training framework of a universal transformer model, namely the choice of multi-modal input features and the design of pre-training tasks. The choice of appropriate input features is crucial to enrich the representation capacity of the pre-trained model so it is applicable across a wide-range of tasks, while the pre-training tasks should be designed carefully to efficiently but accurately capture both local and global interactions across the training set. The design of a pre-training framework is challenging for organic molecular crystals that are defined by a range of inter-

and intra-molecular interactions of widely varying strength and directionality, combined with geometric information about symmetry and molecular packing [40].

Here we introduce a foundation model focused on molecular crystal structures that can be used as a universal tool for a wide range of prediction tasks for materials applications that would otherwise require time-consuming calculations. We present a novel transformer-based pre-training framework—Molecular Crystal Representation from Transformers (MCRT)—that has been pre-trained on a dataset of 706,126 experimentally-determined structures sourced from the Cambridge Structural Database (CSD) [41]. MCRT accommodates multi-modal inputs that encode both local and global features in conjunction with a set of carefully designed pre-training tasks that help capture universal representations for predicting a wide range of different crystal properties, achieving state-of-the-art performance. We tested MCRT's performance on a range of prediction tasks on crystalline properties such as lattice energy, methane deliverable capacity (as relevant for natural gas-powered vehicles), diffusivity, and charge mobility (relevant in organic electronics), demonstrating that the model can be applied to both porous and non-porous organic solids. We further explored different ablations of the proposed model, as well as its learning capacity limits under data scarcity conditions. Importantly, MCRT's attention-based architecture allows us to gain a more intuitive understanding of the structure-property relationships in molecular crystals through cumulative attention scores [42] from across the different layers of MCRT.

2 Results and discussion

Overview of pre-training framework

The overall framework of MCRT is illustrated in Fig. 1a. It comprises a transformer encoder module that is used to build a pre-trained model that then acts as foundation for fine-tuning on a range of downstream prediction tasks. The pre-trained model was built with universality in mind, and designed to distill critical features of molecular crystals without the need for labeled data and, subsequently, to extrapolate desirable physical properties across various applications after fine-tuning. This transformer is therefore designed as a multi-modal architecture that processes two distinct input modalities, which encode both local and global information: atom-based graph embeddings and persistence image embeddings.

- **Atom-based graph embeddings.** These are embeddings taken from the penultimate layer of an ALIGNN architecture [43], which performs message passing on both the interatomic bond graph and its line graph corresponding to bond angles, thus integrating bond length and angle information to provide a more enriched representation of the local environments in a crystal structure. To further enhance the positional information of each atom and to support an efficient training process, we added relative positional embeddings to the atomic features (Fig. S2), which were integrated with the atomic features before being fed into the transformer encoder during each training epoch. The positional embeddings were derived by randomly perturbing the structure, a process that enables the model to better capture the relative positions of atoms in the system, while also mitigating permutational invariance problems.
- **Persistence image embeddings.** These embeddings are generated from persistent homology images [44] and encode global structural information about each crystal structure. This complements the local information provided by the atom-based graph embeddings. Persistent homology has shown potential recently in capturing the topological features of porous materials, demonstrating improved performance in adsorption prediction [45]. More broadly, persistence images in crystal structures encode topological changes as these occur when spheres centered on the atoms increase their radii. These topological changes can include the development of channels (1D persistence image) and voids (2D persistence image) within the structure and can therefore have a critical effect in the global representation of the structure in downstream tasks where geometric information is crucial. As exemplified in Fig. S1, the existence of channels and voids is encoded in diagrams that are subsequently transformed into images that further capture the spatial distribution of topological features [44]. Here we segment persistence images of molecular crystals into patches of fixed size and feed them into the transformer encoder as an additional modality. Further details can be found in Section 4.

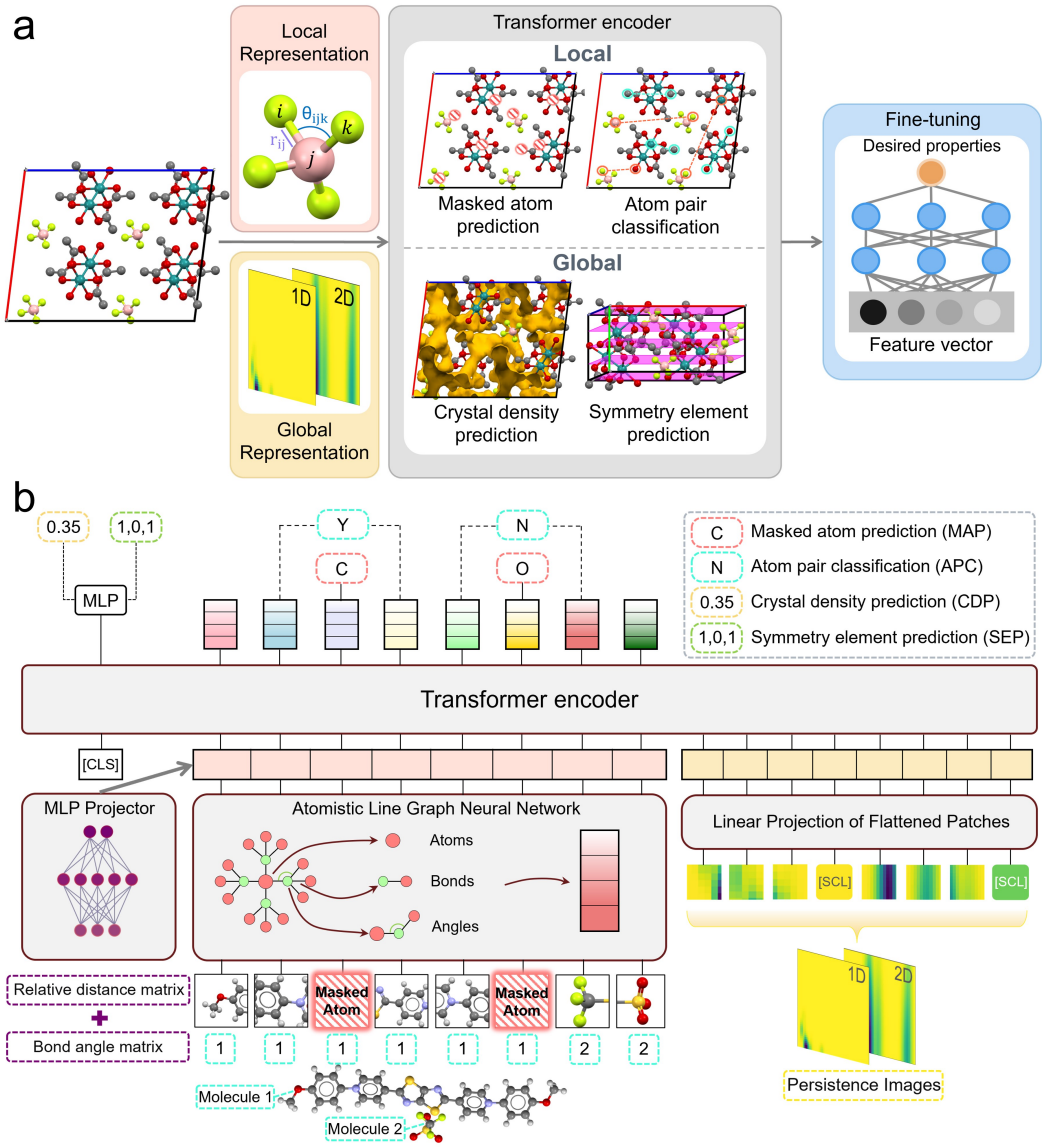

Fig. 1 Schematic overview of MCRT framework. a, Molecular crystals are represented using local and global features, which are then fed into the model. During the pre-training phase, the model undergoes pre-training on four tasks: masked atom prediction (MAP), atom pair classification (APC), crystal density prediction (CDP), and symmetry element prediction (SEP). In the fine-tuning phase, the model is initialised with parameters from the pre-trained model and a simple prediction head is added to train for the desired properties of molecular crystals. **b**, Architecture for pre-training MCRT. Before being fed into the model, 15% of the atoms are randomly masked, and the model is tasked with predicting the types of the masked atoms based on the final atomic representations. Each atom is pre-assigned a molecular label, indicating which molecule within the $P1$ unit cell it belongs to, thus providing labels for atom pairs in the subsequent APC task. Meanwhile, SEP and CDP tasks, as global pre-training tasks, leverage the output of the [CLS] token representing the entire crystal for their predictions.

The transformer encoder architecture that we used in our framework was inspired by BERT [28], which is based on a bidirectional training strategy employing masked language modeling (MLM) and next sentence prediction (NSP) objectives. A [CLS] token is used to predict desired properties by training a multi-layer perception (MLP) head on it. The subsequent tokens are embeddings of atoms and persistence images are segmented into patches, separated by a [SEP] token. At the end of each image, two [SCL] tokens are added to indicate the maximum persistence value and maximum birth value of the persistence images. This method ensures a more balanced distribution of data across pixels, instead of scaling each image to a universal maximum size, which could lead to the concentration of most of the information within just a few pixels. It also enhances the model's robustness, preventing failures when processing larger-scale images in future applications. The full pre-training architecture is illustrated in Fig. 1b.

Pre-training results

To capture universal latent representations of molecular crystals we designed four pre-training tasks performed on a dataset of 706,126 experimental structures sourced from the Cambridge Structural Database (CSD) [41], namely a masked atom prediction task (MAP), an atom pair classification task (APC), a crystal density prediction task (CDP) and a symmetry element prediction task (SEP). The MAP and APC tasks capture local chemical information, while the CDP and SEP tasks capture global structure information.

- **Masked atom prediction (MAP).** The goal of the MAP task is to predict the type of randomly selected masked atoms, which gives the model a deeper understanding of the various local chemical environments of atoms. Similar to the masked word prediction task in the BERT model, 15% of the atoms were masked before being inputted into the model. Of these, 80% were replaced with a [MASK] token, 10% were replaced with another random atom, and the remaining 10% were left unchanged (Fig. 2a). This approach avoids replacing all selected atoms with [MASK] tokens as these do not appear in downstream tasks and helps mitigate the mismatch between pre-training and fine-tuning. The accuracy of the MAP task on the pre-training dataset was 99.9%.
- **Atom pair classification (APC).** In the APC task, the model attempts to distinguish whether a pair of atoms comes from the same molecule. This task is designed to help the model distinguish the different molecules within a crystal cell and to gain a deeper understanding of the crystal structure, as intermolecular and intramolecular interactions are highly diverse—more so than for ionic, inorganic crystals. Specifically, for each crystal, a certain number of atom pairs are randomly selected for this process ensuring that half of these pairs come from the same molecule, and the other half from different molecules to balance bias. To ensure that the order of atoms does not affect the prediction results, the representation vectors of the two atoms are concatenated in both forward and reverse order, passed through the same prediction head, and the outputs are summed to obtain the final prediction result, as shown in Fig. 2b. Crystal structures were represented as graphs where disconnected sub-graphs were considered as isolated molecules. If the number of atom pairs is too large, then the training process will be slowed down by the sampling process. Conversely, if the number of pairs is too small, then the training accuracy will improve very slowly. It was observed empirically that using 200 atom pairs per crystal gives a fair balance between training speed and accuracy. The accuracy of APC task on the pre-training dataset was 99.9%.
- **Crystal density prediction (CDP).** Crystal density is linked to the packing density of molecules and serves as a cheap and easy-to-obtain proxy label during pre-training. Due to the significant impact of molecular packing density on the porosity of molecular crystals, CDP is a particularly important task for applications that depend on crystal voids, such as adsorption, although it might also be expected to have correlations with other solid-state properties, such as charge mobility. Methane deliverable capacity is one example of an adsorption property task, and Fig. S4 illustrates the strong correlation between methane capacity and crystal density for a hydrogen-bonded framework (HOF) forming molecule, T2 [7]. Similar broad correlations would be expect for other gases and other materials. For the prediction of crystal density, the [CLS] token output by the model was passed through a one dense layer head. The mean absolute error (MAE) of CDP on the pre-training dataset was 0.032 g/cm^3. For reference, 99.6% of crystals in the pre-training set have a physical density of >1 g/cm^3 (average density = 1.508 g/cm^3), so this is a relatively small percentage error.
- **Symmetry element prediction (SEP).** The space group of a crystal can be considered as a blueprint that provides important information about its global structure. However, a direct prediction of space group can be quite challenging due to the strongly imbalanced distribution of space groups among crystals, where >80% of molecular crystals occupy just 5 of the 230 existing space groups [15]. This class imbalance, in conjunction with the complex symmetry information contained in underrepresented space groups, often hinders the learning of meaningful space group representations. Instead of using space groups explicitly, we focused on the less imbalanced task of predicting the total symmetry elements that define each space group, which effectively encodes the same foundational information contained in space groups. There are six types of symmetry elements [46]: inversion center, mirror plane, rotation axis, screw axis, rotoinversion axis, glide plane. Considering the case of no symmetry elements (P1 space group), the output of this task then becomes a 7-dimensional multi-hot vector (note that a structure can correspond to multiple different symmetry elements). For a successful prediction, all elements of the prediction must match the label exactly. Pre-training the SEP task on the pre-training dataset results in a prediction accuracy of 98.5%.

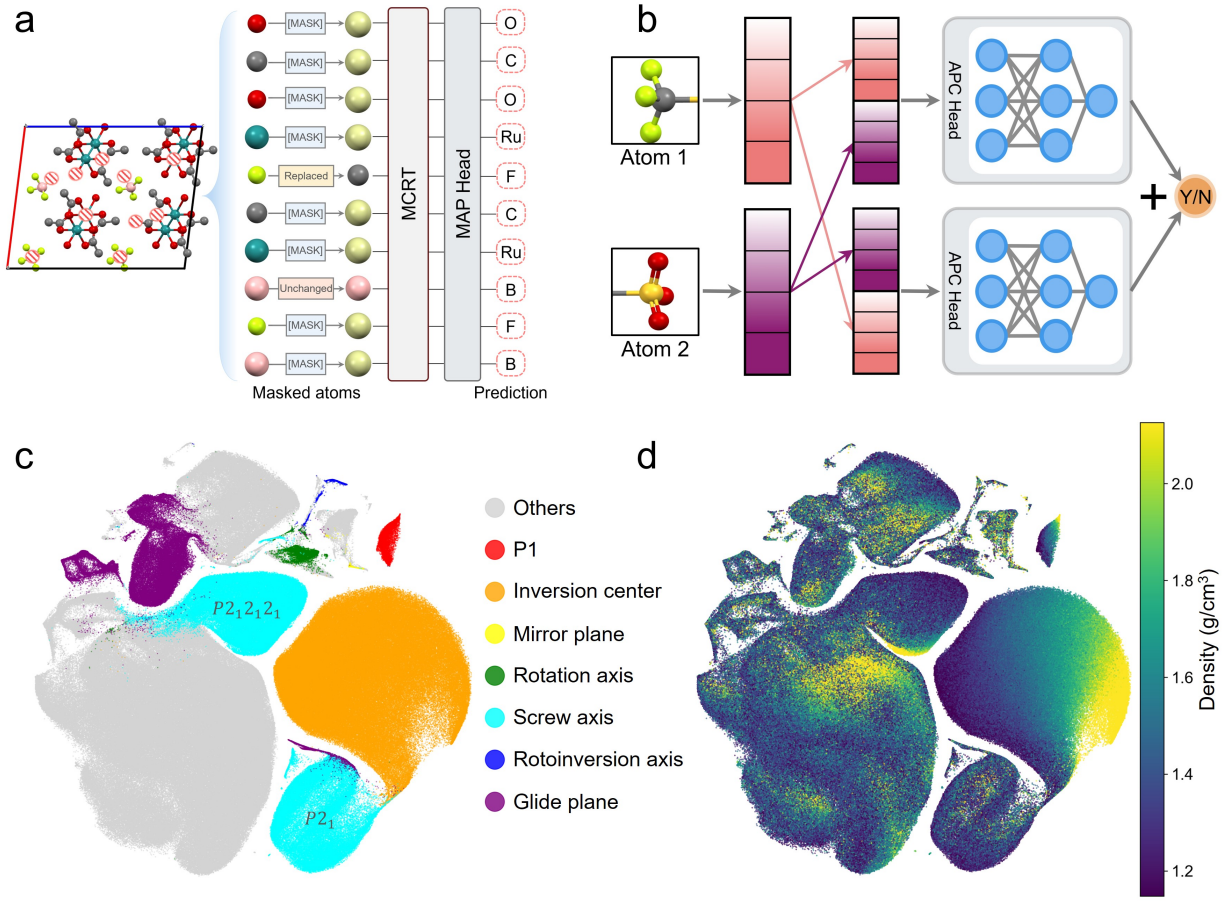

Fig. 2 Summary of pre-training tasks. a, Scheme for masked atom prediction (MAP) task. Among the masked atoms, 80% are replaced with the [MASK] token, 10% are replaced with a random atom, and 10% remain unchanged. **b**, Scheme for atom pair classification (APC) prediction head. Representations of a pair of atoms are concatenated in two orders to eliminate the impact of atom sequence, ensuring more stable predictions. **c**, The t-SNE embeddings of the [CLS] tokens of 706,126 experimental molecular crystals obtained from the pre-trained model, with crystals containing only one type of symmetry element being coloured. **d**, The t-SNE embeddings of the [CLS] tokens of 706,126 experimental molecular crystals obtained from the pre-trained model, with colour indicating density, and the top and bottom 5% of densities truncated for better visualisation.

To validate the representation learning capacity of our proposed pre-traning framework here, we visualised the learned representations by MCRT ([CLS] tokens) of all 706,126 crystals in 2D using t-SNE [47]. Since a crystal can contain multiple types of symmetry elements, we selected crystals containing only one type of symmetry element and highlighted them with a different colour. The low dimensional map of Fig. 2c validates that crystals with similar symmetry elements cluster together. It is noteworthy that certain symmetry elements manifest as multiple clusters in the t-SNE embeddings, which can be attributed to the nuanced differentiation of space groups. For example, screw axes are present in several space groups, as depicted in Fig. S5. Specifically, the two prominent clusters correspond to space groups 4 ($P2_1$) and 19 ($P2_12_12_1$). Although the model was not explicitly trained to recognize space group information, it automatically learned and differentiated between various space groups during the pre-training phase. This capability highlights the model's inherent ability to capture and classify structural features of molecular crystals and to comprehend underlying crystallographic principles. Additionally, Fig. 2d shows a re-labeling on the same map using crystal density values, where it can be seen that crystal density exhibits a gradient distribution within most of the symmetry element clusters, indicating that the embedding vectors cluster according to similar densities. Taken together, these results suggest that the pre-trained model has been successfully trained to capture the key features of molecular crystals.

Fine-tuning results

Next we demonstrated the utility of our pre-training framework through a series of fine-tuning experiments on a diverse range of crystal property prediction tasks. These include lattice energy prediction, methane

deliverable capacity prediction (298 K, pressure cycle of 65–5.8 bar), methane diffusivity prediction (298 K at infinite dilution) and charge mobility prediction, a task related to organic semiconductors. We also tested MCRT's predictive performance in Δ-E tasks; that is, the lattice energy difference between DFT and force field accuracy calculations, to assess its extrapolation to higher accuracy levels of energy prediction.

The datasets used in these tasks encompassed a wide variety of molecules. For lattice energy prediction, the dataset included a set of 10 structurally related molecules with small changes in hydrogen bonding functionality, derived from earlier crystal structure prediction (CSP) studies [6], providing a relevant test for fine-tuning a model to study a closely related family of molecules. By contrast, the Δ-E task involved a deliberately diverse set of 1018 organic molecules, designed to develop a generalised ML model capable of improving lattice energy predictions across broad and chemically diverse areas of molecular space [15]. The charge mobility predictions focused on 7 pentacene and azapentacene molecules [10], which are key compounds in organic semiconductor research, while the methane capacity and diffusivity tasks were based on CSP structures for the HOF-forming molecule T2 [7]. All datasets (except for Δ-E) were randomly split with a train-validation-test ration of 80%:10%:10%. The Δ-E dataset was split according to the original paper [15]. A detailed description of the datasets and the methods used for their generation can be found in Section 4.

Table 1 reports the mean absolute error (MAE) results for fine-tuning MCRT and its variants, compared against state-of-the-art baseline models. SOAP-based random forest (RF) [48] and kernel ridge regression (KRR) [49], graph-based CGCNN [13] and ALIGNN [43], and pre-trained crystal twins (CT) [40] were selected as baseline models due to their universality and competitive performance in predicting materials' properties [12, 45, 50]. For a detailed description of these methods and their featurisations, as well as the MCRT variants used in benchmarking, see Section 4.

Table 1 Mean absolute error (MAE) results for the fine-tuned MCRT models and baseline models for a wide range of properties.

Property (Size, Unit)	RF	KRR	CGCNN	ALIGNN	CT	MCRTp[5]	MCRTi[6]	MCRTa[7]	MCRT
LE_all (70k, kJ/mol)[1]	7.79	6.90	5.95	2.68	4.85	3.31	2.63	2.59	**2.34**
LE_T2 (8k, kJ/mol)[1]	7.79	8.44	6.13	3.45	5.21	3.84	3.20	3.27	**2.96**
LE_T2A (1k, kJ/mol)[1]	3.34	3.96	3.20	3.27	3.19	2.98	2.60	2.57	**2.15**
MC (5k, v STP/v)[2]	11.60	11.75	15.87	12.17	14.53	9.88	10.81	9.91	**8.82**
MD (5k, 10^{-8} cm^2/s)[2]	0.75	0.68	1.08	0.79	0.92	0.50	0.57	0.48	**0.42**
CM (1k, cm^2/Vs)[3]	0.59	0.60	0.70	0.62	0.62	0.62	0.54	0.60	**0.52**
Δ-E (11k, kJ/mol)[4]	2.82	3.13	2.33	2.90	2.25	1.82	1.62	1.70	**1.57**

[1]LE_all (70k), LE_T2 (8k) and LE_T2A (1k) denote lattice energy with 73,779, 8,293 and 1,367 data points, respectively. [2]MC (5k) and MD (5k) denote methane capacity and methane diffusivity with 5,687 data points, respectively. [3]CM (1k) denotes charge mobility with 1,130 data points. [4]Δ-E (11k) denotes the difference in lattice energy between DFT and force field calculations, comprising 11,458 data points. [5]MCRTp is MCRT without pre-training. [6]MCRTi is MCRT without persistence image part. [7]MCRTa is MCRT using absolute positional embedding. The best results for each property are highlighted in bold.

For lattice energy prediction, three datasets of different sizes were used to validate the model's transferability capabilities under limited data availability scenarios. We note here that LE_all includes the CSP landscapes of 10 molecules, all with hydrogen bonding functionality (see Fig. 5a, below), comprising 73,779 structures in total with their associated lattice energies. LE_T2 represents the CSP landscape of the T2 molecule with 8,293 structures and energies, while LE_T2A corresponds to the CSP landscape of the T2A molecule with 1,367 structures and energies. Both LE_T2 and LE_T2A are subsets of LE_all.

The fine-tuned MCRT model outperformed all other models across all tasks, demonstrating both superior predictive capability and universality. ALIGNN exhibited better performance compared to other baseline models when predicting LE_all (70k) and LE_T2 (8k), but its performance on LE_T2A (1k) does not stand out against other models. We hypothesise that this could be due to ALIGNN encoding angular information, and thus making the model more complex than other baseline models and more prone to overfitting with insufficient training samples. By contrast, MCRT, with proper pre-training, still demonstrates relatively good predictive performance even with this small dataset of 1,367 structures and energies. Graph-based models outperform SOAP descriptor-based models in predicting lattice energy, a property strongly related to the local chemical environment of atoms. Deep models on the other hand perform poorly in predicting methane capacity (MC) and diffusivity (MD) which are properties related to global structural features. A similar observation was also confirmed by studies using the MOFTransformer model [24]. This phenomenon is further validated by the poorer performance of the purely graph-based MCRTi model. When the persistence

image component is added, the MCRT model's performance improves significantly, further emphasising the importance of global geometric features in adsorption and diffusion predictions. Regarding the performance of the non-pre-trained MCRTp, it maintains a competitive performance, but is noticeably inferior to the pre-trained MCRT. The ablation model MCRTa used absolute positional embeddings as opposed to relative ones. Despite undergoing the same pre-training stage, its prediction accuracy across various tasks was consistently inferior to that of MCRT, indicating that absolute positional embedding, which does not satisfy translational and rotational invariance, indeed increases training difficulty.

Fig. 3 reports test-set results in R^2 for MCRT and baseline models along with performance correlation plots of MCRT across all downstream prediction tasks. From the radar plot in Fig. 3a, MCRT scores the highest R^2 across all prediction tasks, especially in the data-scarce charge mobility (CM) dataset, where it significantly outperforms other models. Although baseline models may perform well on specific tasks, they struggle to balance performance across diverse tasks. For instance, ALIGNN excels in lattice energy prediction but shows only average performance in other areas. This further highlights the universality of MCRT.

Beyond prediction errors, experimental materials researchers are interested in ranking the best-performing structures to prioritize experimental targets. For lattice energy predictions, structures with lower lattice energies are more likely to be synthesizable. Figs. 3b, c, and d, demonstrate that MCRT successfully predicts most of the lowest-energy structures for the molecules considered, highlighting the model's practical utility.

For methane capacity (MC), a property with high computational screening cost, MCRT successfully predicts all the top ten best-performing structures, highlighting its potential for robustly accelerating high-throughput computational screening procedures using crystal structure prediction. Here this may be a better measure than MAE since we are mainly interested in the best-performing crystals. Additionally, for methane diffusivity (MD), which is challenging property to capture and predict using machine-learned interatomic potentials, MCRT still delivered the best predictive performance and successfully identified nine out of the top ten best-performing structures. When it comes to charge mobility (CM), a property with sparse data due to high computational cost, all models struggle somewhat with its prediction, but MCRT still identified seven out of the top ten structures with highest charge mobility. Finally, regarding the Δ-E task, MCRT captures more accurately the relative energy relationships across various structures. As shown in Fig. 3h, the corrected energies by MCRT (turquoise points) align more closely with DFT-calculated results than those calculated by force fields (grey points), importantly in the low-energy regime, again successfully predicting all of the top ten structures.

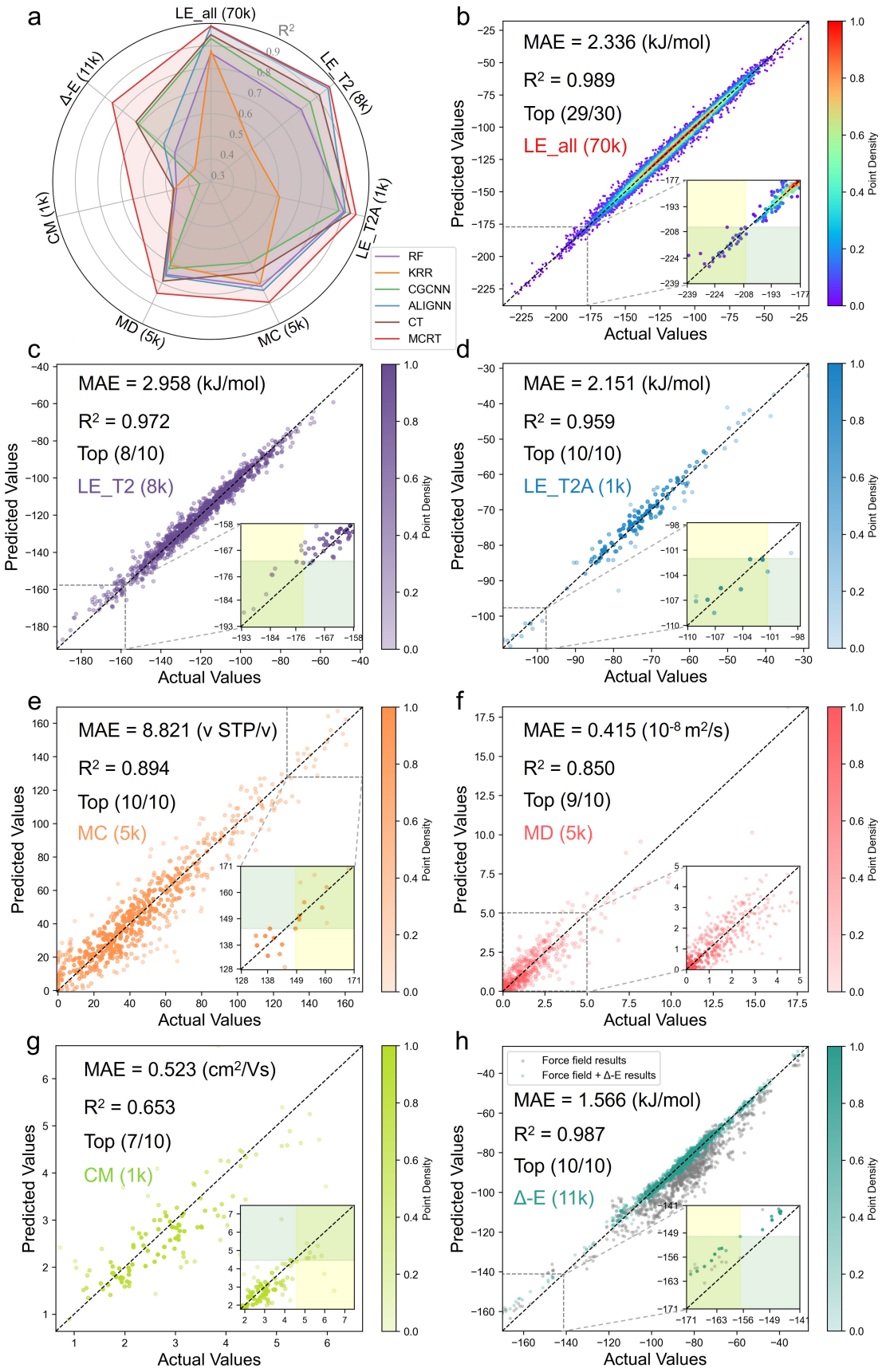

Fig. 3 MCRT outperforms other baseline models for all downstream prediction tasks, identifying the top few structures of interest. a, The coefficient of determination R^2 on test sets of MCRT and baseline models. The prediction results on test sets of fine-tuned MCRT for **b**, LE_all (70k), **c**, LE_T2 (8k) **d**, LE_T2A (1k), **e**, MC (5k), **f**, MD (5k), **g**, CM (1k), **h**, Δ-E (11k). Inset sub-figures are illustrations of areas of interest, the yellow regions represent the areas where the top n actual values are located, while the green regions indicate the areas where the top n predicted values are located. The points in the intersections represent the top n points that were successfully predicted. For MD (5k), the sub-figure is intended to provide a clearer visualisation of the densely populated region.

Interpretability

Feature importance analysis is an inherent capability of transformer models that helps us here to better understand the relationship between molecular crystal structures and their properties. Cumulative attention scores of the [CLS] token were calculated to measure the model's assigned attention to input features according to their importance. An attention rollout strategy [42] was employed by recursively multiplying attention weights across layers, providing more focused patterns for interpreting which input tokens contribute the most to the model's output. Higher attention scores indicate greater importance for the model's prediction. Fig. 4 provides an intuitive visualisation of the explainability scores for an experimentally synthesized structure, T2-γ [7], for methane capacity (left column) and lattice energy (right column), illustrating explainable feature importance for both modalities.

When predicting methane capacity, the model shows little attention toward the atomic graph modality (Fig. 4a). However, the model places significantly higher attention on persistence images when predicting methane capacity than when predicting lattice energy (Fig. 4c / 4g, and Fig. 4d / 4h), as highlighted by the salmon-coloured areas in the persistence images. This further validates that global geometric features are more important for adsorption predictions, aligning with findings from previous studies [21, 51]. In particular, when predicting methane capacity, the model places particularly high attention to the 1D persistence image, which encodes information about the pores in the crystal. The [SCL] token indicating the largest persistence value is identified as the model's most significant feature, receiving an attention score far exceeding those of other image patches. The persistence value represents the radius of the largest sphere that can pass through the topological object, effectively corresponding to the pore radius. Subsequently we mapped the topological objects within the patches near the largest persistence back to the representative cycles in the original T2-γ structure. The objects near the largest persistence value correspond to the large pores in T2-γ as shown in Fig. 4b. The largest persistence value thus contains essential information about the pore size, which is crucial for predicting adsorption properties [52].

For lattice energy predictions, the model exhibits marked attention to the hydrogen-bonding benzidimazole groups of the T2 molecule (Fig. 4e). To evaluate the accuracy of the chemical insights provided by these attention scores, an electron density difference (EDD) analysis was performed on T2-γ, as shown in Fig. 4f. The yellow isosurfaces represent regions with increased electron density, while the blue isosurfaces indicate regions with decreased electron density. Normally, a larger magnitude of electron density shift indicates a stronger intermolecular interactions. Notably, expanding to a supercell, we observe strong attention in regions corresponding to the areas of intense intermolecular interactions (Fig. S7d). A similar phenomenon can also be observed in other experimental T2 polymorphs, as shown in Fig. S8 and Fig. S9. This suggests that the model's attention aligns with key regions of strong intermolecular interactions, which play a crucial role in stabilising the crystal structure. Given that subtle changes in these strong interactions can lead to significant differences in lattice energy [7], the fine-tuned model appears to have effectively captured the critical chemical features necessary for accurate lattice energy predictions.

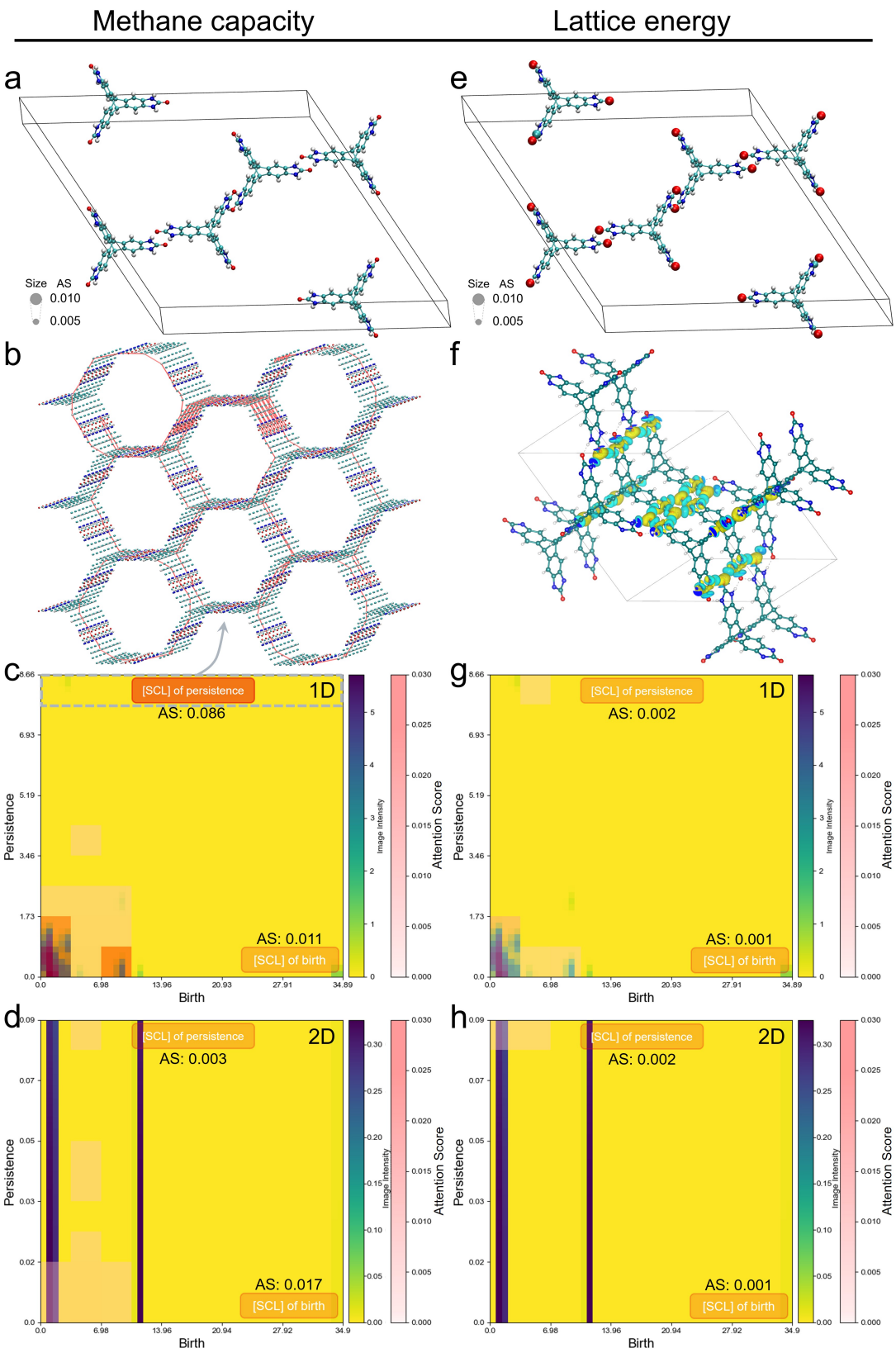

Fig. 4 Attention scores for atom-based embeddings and persistence image embeddings for a porous framework, T2-γ. a, Unit cell, **c**, 1D and **d**, 2D persistence images of T2-γ with attention scores for CH_4 capacity. **b**, The point cloud of atoms in T2-γ, with red lines representing the representative cycles corresponding to the topological objects in the persistence image. **e**, Unit cell, **g**, 1D and **h**, 2D persistence images of T2-γ with attention scores for lattice energy. **f**, Electron density difference plot of T2-γ highlighting the region of intermolecular interactions (yellow isosurfaces = increased electron density; blue isosurfaces = decreased electron density). The strong intermolecular interactions are found for the atoms that were attended to in **e**, also for other experimental T2 polymorphs (Fig. S8 and Fig. S9). In the unit cells, the atomic size is proportional to normalized attention scores, with scores less than 0.005 being clipped to avoid extremely small atoms (colour code: C, cyan; H, white; N, blue; O, red). In the persistence images, the 10 patches with the highest attention scores are visualized with a salmon-coloured overlay, where stronger intensities represent higher attention scores.

Few-shot learning

Data can be the limiting resource both in practical synthetic chemistry and in computational materials studies due to the high costs of synthetic or computational methods. Here we explored our model's extrapolation capabilities on extremely small datasets. Specifically, we tested MCRT's predictive performance in zero- and few-shot learning [53] scenarios for predicting lattice energies of T2 structures, using datasets containing analogues of T2 to assess its generalization capability within a related molecular family.

We first formed a test set using all T2-based structures contained in LE_all (a dataset comprising CSP landscapes of T2 and analogues of T2). The remaining structures were then randomly divided into training and validation sets in a 90%:10% ratio to fine-tune MCRT. Subsequently, from the T2 structures, we sequentially separated 100, 200, 300, ..., up to 1,000 structures to further fine-tune the model obtained from the previous step. The remaining structures were used as a test set to assess the robustness of MCRT predictions. To provide a clear performance comparison, we conducted the same experiment using ALIGNN, the most competitive baseline model for lattice energy prediction (Table 1). Fig. 5a illustrates our training setup and Fig. 5b presents the MAE, R^2 and top-10 prediction performances for MCRT and ALIGNN. In the zero-shot scenario, MCRT exhibits significantly higher prediction accuracy compared to ALIGNN, indicating that after learning from similar structures, MCRT can generalize more effectively to related but unseen structures. Furthermore, with just an extra step of further fine-tuning on 100 structures, MCRT scores a low prediction MAE (4.34 kJ/mol). By contrast, even after being trained on 10 times more data, ALIGNN still fails to achieve MCRT's extrapolation capacity, with a prediction MAE of 6.40 kJ/mol. This was further echoed by a series of energy-density landscape reconstruction tasks on the different fine-tuned MCRT models. As shown in Fig. S14, MCRT accurately reproduces the relative positions of the four experimental structures within the energy-density landscape even with zero-shot learning, whereas ALIGNN fails to effectively capture these relative positions, particularly misplacing T2-δ by conflating it with numerous other structures, as illustrated in Fig. S15. This further underscores MCRT's practicality for computationally costly scenarios such as crystal structure prediction.

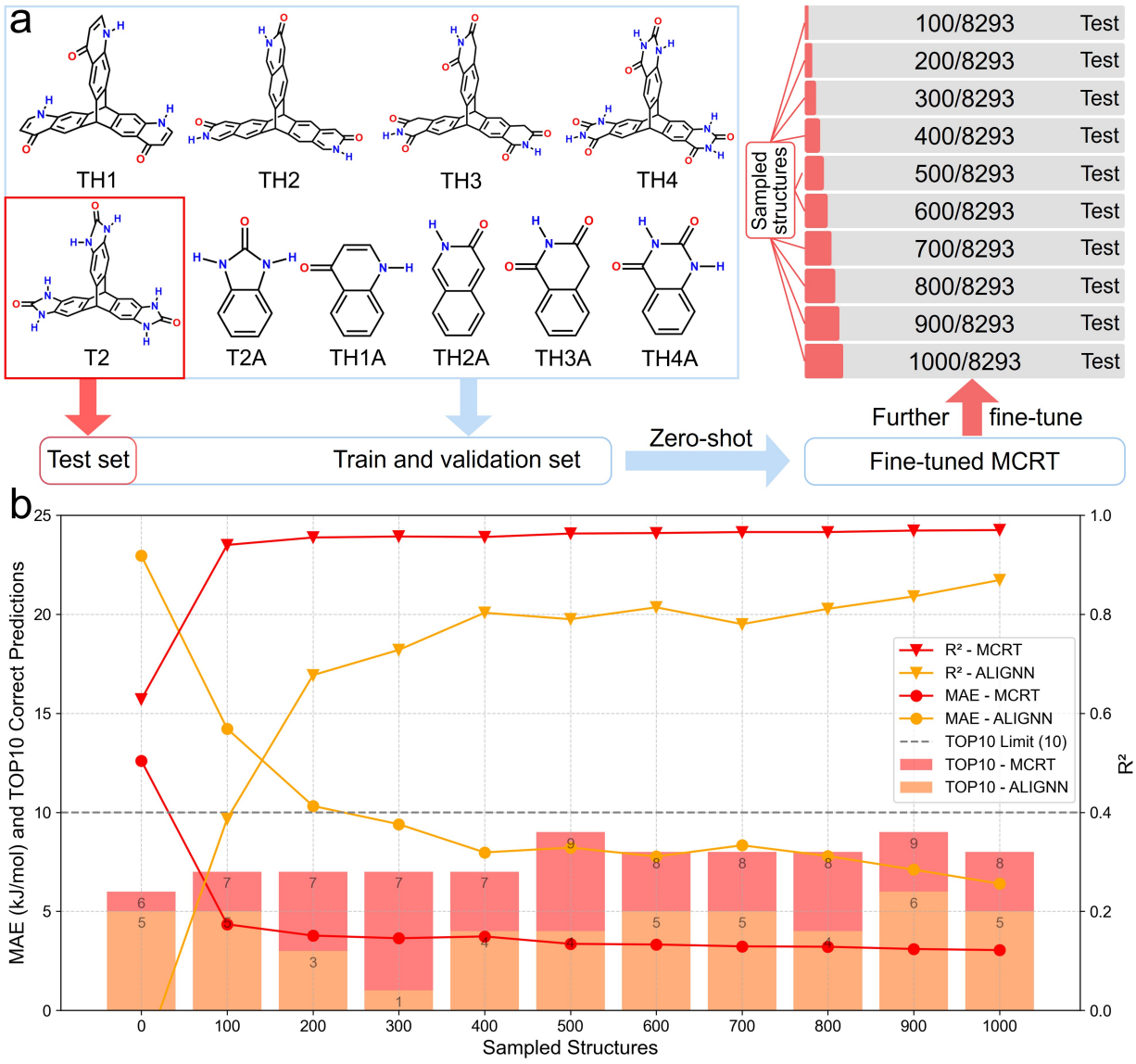

Fig. 5 Few-shot learning of lattice energies for 10 related hydrogen-bonding molecules. a, Chemical structures of T2 and 9 other related hydrogen-bonding molecules, including 4 triptycene framework candidates (TH1–TH4) and 5 small, monoaromatic molecules with the same representative hydrogen bonding functionalities. In the few-shot learning experiment, the T2 structures in the LE_all dataset were extracted and used as the test set. The remaining structures of other 9 molecules were used for training and validation of MCRT, resulting in a fine-tuned MCRT model, yielding the zero-shot prediction results (sampled structures = 0). Subsequently, a small number of T2 structures (100–1,000) was randomly sampled as training and validation sets to further fine-tune this model. The remaining T2 structures and associated lattice energies were used as the test set to evaluate the few-shot performance of MCRT. **b**, The prediction results of the few-shot learning experiments of MCRT and ALIGNN. Sampled structures refers to the number of T2 structures and associated lattice energies extracted as the training and validation set during few-shot learning scenarios. TOP10 represents the number of correctly predicted structures among the 10 lowest-energy structures.

3 Conclusions

We present a new universal transformer model, MCRT, together with a pre-training framework for predicting a wide range of physical properties for molecular crystals. We have designed a multi-modal architecture that has the capacity to comprehensively learn both local and global representations of molecular crystals. This ensures universal transferability for MCRT across different tasks and structures, at least for the tasks attempted here. We tested MCRT's predictive performance by fine-tuning it on various diverse properties such as lattice energy, methane capacity and diffusivity, as well as charge mobility. Our proposed model both outperformed current state-of-the-art models and showed strong generalisability performance in limited data availability scenarios. This highlights the practical utility in robustly accelerating materials discovery — for

example, by rapidly estimating crystal structure prediction energy landscapes based on landscapes calculated for related molecules. Our MCRT model was also provided insights into structure-property relationships for molecular crystals through its permeable, interpretable architecture design. While such interpretability should not be equated to causal, physical understanding, it is striking that the attention scores in the MCRT model correlate so strongly with the key intermolecular interactions in lattice-energy prediction tasks, at least for molecules such as T2 that feature dominant hydrogen-bonding patterns (Fig. 4e,f; Figs. S8-S10).

MCRT performs well both in predicting properties across a broad range and in highlighting the top few 'best-performing' cystals (Figs. 3b-h). From a practical perspective, both of these tasks are important since there is very often a trade-off between different properties for real applications. To give just one example, to design materials for methane storage, we need to predict structures that have low lattice energies—that is, materials that will be formed in experiments—while having methane capacities that are high enough, rather than simply identifying the crystals that absorb the most methane. Both properties, lattice energy and methane capacity, are expensive to calculate. There are other important physical properties relevant to methane storage materials, not investgated here, such as mechanical stability and thermal conductivity, which would add further computational cost to a digital materials screening programme. As such, the development of universal, inexpensive prediction tools is a key priority in computational materials design. We believe that MCRT can serve as a foundational infrastructure for the molecular crystal research community, aiding us in the accelerated exploration of the vast space of molecular crystals.

4 Methods

Pre-training datasets

To ensure high-quality crystal structures for the pre-training dataset of MCRT, we selected 706,126 molecular structures from the Cambridge Structural Database (CSD) database [41], pre-filtered to satisfy the following criteria: (i) only structures with fully determined three-dimensional coordinates were included to ensure comprehensive spatial information; (ii) only structures with an R factor of 0.1 or less were included to ensure high-quality refinement and accuracy of the crystal structures; (ii) structures exhibiting any form of disorder were excluded to avoid complications in subsequent analysis and to maintain data consistency; (iv) only structures without reported errors were included; (v) we excluded polymeric structures, such as metal-organic frameworks, focusing solely on discrete molecular crystals; (vi) only single crystal structures were considered, ensuring higher precision in the determination of atomic positions. This robust pre-filtering process was crucial to ensure the robustness and reliability of our subsequent analysis and training.

Materials analysis

For the manipulation and labeling of the collected pre-training set, we used the Python Materials Genomics (pymatgen) library [54]. In particular, for the APC task, the crystal structures were represented as graphs, where disconnected sub-graphs were considered as isolated molecules, for the SEP task the space groups of the crystals were first identified and subsequently were mapped to their corresponding symmetry elements. For the remaining tasks the label generation was straightforward using pymatgen. Before being inputted into the model, the crystals were converted into the $P1$ space group to ensure the feasibility of subsequent SEP and APC tasks during pre-training phase. For the persistence image generation, we used MoleculeTDA [21] to compute persistence images with a resolution of 50×50 and a spread of 0.15, consistent with previous studies [55]. For the t-SNE embedding, we used a perplexity parameter of 50, due to the large size of the pre-training dataset.

Fine-tuning data collection

We fine-tuned MCRT on diverse properties of different molecular crystals to validate the generality of MCRT's predictive capabilities. The details of the datasets are as follows:

- **Lattice energy.** Lattice energy calculations were performed with an anisotropic atom–atom potential using DMACRYS [56]. Electrostatic interactions were modelled using an atomic multipole description of the molecular charge distribution (up to hexadecapole on all atoms) from the B3LYP/6-311G(d,p)-calculated charge density using a distributed multipole analysis [57]. Atom–atom repulsion and dispersion interactions were modelled using a revised Williams intermolecular potential [58], which has been benchmarked against accurate, experimentally determined lattice energies for a range of molecular crystals [59]. We specifically generated three fine-tuning datasets of different size to test MCRT's predictive capacity on limited data availability scenarios, namely *LE_all*, a dataset with 73,779 structures composed of CSP

landscapes on all the molecules listed in Fig. 5a, *LE_T2*, the CSP landspace of T2 with 8,293 structures and *LE_T2A*, the CSP landspace of T2A with 1,367 structures.

- **CH_4 deliverable capacity.** A dataset of 5,687 T2-based structures with calculated CH_4 deliverable capacity (298 K, 65–5.8 bar) was directly retrieved from the previous work [7].
- **CH_4 diffusivity.** A dataset of 5,687 T2-based structures with calculated CH_4 diffusion coefficients at infinite dilution using the MD simulations. The simulations were conducted at 298 K with a time step of 1 fs for a total of 5 million cycles, with 1,000 cycles used for the initialization and 10,000 cycles for equilibration. DREIDING force field was used with the Lorentz–Berthelot mixing rule and a cut-off distance of 13 Å. The CH_4 molecule was modeled as a single atom. Prior to the simulations, 30 CH_4 molecules were randomly introduced into the pores of crystals. The mean square displacement (MSD) of gas molecules during 1–5 ns is used to calculate the diffusion coefficient through Einstein's relation [57]. All these simulations were carried out at NVT ensemble using RASPA2 package [60].
- **Charge mobility.** The charge carrier mobility values in this dataset were obtained from previous work [10] and were calculated using the Marcus theory of charge transport. The dataset is based on crystal structure prediction (CSP) studies, with the studied molecules including pentacene and azapentacenes. The charge carrier mobility calculations were restricted to crystal structures within a 7 kJ/mol energy range of the global minimum on the energy-density landscapes, capturing low-energy polymorphs most likely to be observed experimentally.
- **Δ-E.** The training target, Δ-E, represents the lattice energy difference between DFT (B86bPBE+XDM) and force field (FIT+DMA) accuracy. Following the approach in the original paper [15], which includes 1,000 CSP landscapes, we split the dataset by selecting 10 crystal structures from each of around 900 landscapes for training and validation, while about 100 landscapes, with 10 structures each, were reserved as a test set. An exclusion of duplicate structures was applied that led to a final dataset of 11,458 structures.

Training details

For pre-training, we randomly split the 706,126 molecular crystal dataset with a train-validation ration of 90%:10%. The model was trained for 50 epochs with a batch size of 512. The AdamW optimizer with a learning rate of 10^{-4} and weight decay of 10^{-2} was used [41]. The learning rate was warmed up during the first 5% of the total epoch and was then linearly decayed to zero for the remaining epochs. For the SEP task we assigned higher weights to the elements with fewer occurrences due to the great variance in the frequency of the occurrence of different symmetry elements. The individual weights are calculated as follows:

$$w(x_i) = \frac{1}{ln(\epsilon + x_i)} \tag{1}$$

where $w(x_i)$ is the weight of element i, x_i is the frequency of element i and ϵ ($\epsilon \geq 1$) is a parameter to adjust the weight distribution which was set to 1.1 to avoid extremely large weights. The resolution of the persistence image during training was set to 50×50 with a patch size of 5×5 in accordance to previous studies [21, 55].

For fine-tuning, all datasets (except for Δ-E) were randomly split with a train-validation-test ration of 80%:10%:10%. The Δ-E dataset was split according to the original paper [15]. By initializing a single dense layer to the [CLS] token, all model weights are fine-tuned to predict desired properties for 50 epochs with a batch size of 32. All other settings are the same as in the pre-training step.

Baselines and ablations

We test the prediction performance of MCRT against a wide range of baselines and state-of-the-art methods. These include

- *Random forest (RF):* A robust ensemble learning algorithm that aggregates the predictions of multiple decision trees, typically leading to enhanced generalisation performance by reducing overfitting and variance in predictive modeling tasks [48].
- *Kernel Ridge Regression (KRR):* KRR integrates ridge regression with the kernel trick, enabling it to perform nonlinear regression in high-dimensional feature spaces while controlling for model complexity through regularisation [49].
- *Crystal Graph Convolutional Neural Network (CGCNN):* CGCNN represents crystalline materials as graphs, where atoms serve as nodes and bonds as edges, and learns material properties by applying convolutional operations over the graphs [13].

- *Atomistic Line Graph Neural Network (ALIGNN):* ALIGNN enhances conventional graph neural networks by incorporating bond angle information from line graphs, thereby improving the model's capability to predict complex material properties with higher accuracy [43].
- *Crystal Twins (CT):* CT is a self-supervised pre-trained model for crystalline material property prediction, using twin CGCNNs to learn robust representations from large unlabeled datasets, which are then fine-tuned for specific tasks [40].

For descriptor-based models, the Smooth Overlap of Atomic Positions (SOAP) descriptor was used due to its universality [16]. The parameters for the SOAP descriptor were set as follows: a cutoff for the local region of 4.0 Å, 6 radial basis functions, and a maximum degree of spherical harmonics of 6. RF and KRR implemented in scikit-learn [61] were adopted, and the hyperparameters were tuned using grid search. For RF, the number of trees was searched from 10 to 1000. For KRR, the regularization strength ω was searched from 0.001 to 100. For Graph Neural Networks (GNNs), CGCNN was trained with the following hyperparameters: 32 batch size, 100 epochs, 5 message passing layers, 1 hidden layer after pooling, 64 hidden atom features in message passing layers. ALIGNN was trained with the following hyperparameters: 32 batch size, 100 epochs, 4 message passing layers, 1 hidden layer after pooling, 256 hidden atom features in message passing layers, in line with the original paper [43]. For the crystal twins pre-trained model (CT), the same fine-tuning hyperparameters as the original paper were used (128 batch size, 200 epochs, 3 message passing layers) [40].

Additionally, the following variants of MCRT were included in the fine-tuning comparisons to assess the importance of the different learning components of the proposed framework

- *MCRTp:* The complete architecture of MCRT used directly for prediction without pre-training.
- *MCRTi:* The architecture of MCRT without persistence image modality input module.
- *MCRTa:* The architecture of MCRT using absolute positional embeddings instead of relative ones.

Both MCRTi and MCRTa underwent the same pre-training process as MCRT. For the latter, the absolute positional input features are processed as in [27], that is by employing BERT's native positional embedding module to embed each of the three-dimensional atomic coordinates separately, and eventually summing them up.

The electron density difference analysis

Periodic DFT calculations, including the electron density calculation, were carried out within the plane-wave pseudopotential formalism, using the Vienna ab initio simulation package (VASP) code version 5.4.4 [62]. Projector augmented-wave (PAW) method was applied to describe the electron-ion interactions [63]. Generalized gradient approximation (GGA) with the Perdew–Burke–Ernzerhof (PBE) exchange-correlation functional was adopted to treat electron interaction energy [64]. Grimme's semi-empirical DFT-D3 scheme with Becke–Johnson damping functions was used here to give a better description of interactions [65–67]. A kinetic-energy cut-off of 600 eV was used to define the plane-wave basis set. The electronic Brillouin zone was integrated with the smallest allowed spacing between k-points (KSPACING) being 0.4 Å^{-1}, and the generated grid was centered at the Γ-point. The convergence threshold for self-consistency was set to 10^{-6} eV during total energy and force calculations.

The electron density difference (EDD) plots were generated by subtracting the electron densities of each isolated molecule from the electron density of the entire crystal:

$$\Delta\rho = \rho_{\text{crystal}} - \sum_{i=1}^{N} \rho_{\text{molecule}_i} \tag{2}$$

where ρ_{crystal} is the electron density of the crystal, and ρ_{molecule_i} represents the electron density of the i-th isolated molecule in the crystal.

5 Data availability

Source data and datasets used in this work, including CIF files of molecular crystals screened from the CSD, are available via Figshare at https://doi.org/10.6084/m9.figshare.27844302. Additionally, we provide pre-trained MCRT model and fine-tuned versions for all datasets, accessible via Figshare at https://doi.org/10.6084/m9.figshare.27822705.

6 Code availability

The MCRT library is available at https://github.com/fmggggg/MCRT. For ease of use, pre-defined Apptainer images are available on Figshare at https://doi.org/10.6084/m9.figshare.26390275. To ensure reproducibility, all results in this paper are obtained from version 1.0.2 of the MCRT library, which is available at https://pypi.org/project/MCRT-tools/1.0.2.

7 Author Contributions

M.F, C.Z and X.E conceived the project, and X.E and A.I.C supervised the project. M.F and C.Z performed the computational experiments and X.E, G.M.D and A.I.C analysed the output of the experiments. All authors contributed to writing the manuscript.

Acknowledgements

This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 856405). X.E. and A.I.C. acknowledge financial support from the Leverhulme Trust via the Leverhulme Research Centre for Functional Materials Design. C.Z. acknowledges the financial support from the China Scholarship Council (No. 202106745008). A.I.C. thanks the Royal Society for a Research Professorship (RSRP\S2\232003). The authors would also like to thank Dmitriy Morozov, Maciej Haranczyk, Christopher R. Taylor, Jay Johal and Tao Liu for useful discussions.

References

[1] Qiao, N. *et al.* Pharmaceutical cocrystals: An overview. *Int. J. Pharm.* **419**, 1–11 (2011).

[2] Punzi, A. *et al.* Croconaines as molecular materials for organic electronics: Synthesis, solid state structure and use in transistor devices. *J. Mater. Chem. C* **4**, 3138–3142 (2016).

[3] Feiler, T. *et al.* Tuning the mechanical flexibility of organic molecular crystals by polymorphism for flexible optical waveguides. *CrystEngComm* **23**, 5815–5825 (2021).

[4] Li, W., Zhang, J., Guo, H. & Gahungu, G. Adsorption of gases in microporous organic molecular crystal, a multiscale computational investigation. *J. Phys. Chem. C* **115**, 4935–4942 (2011).

[5] Chen, T.-H., Kaveevivitchai, W., J. Jacobson, A. & Š. Miljanić, O. Adsorption of fluorinated anesthetics within the pores of a molecular crystal. *Chem. Commun.* **51**, 14096–14098 (2015).

[6] Zhao, C. *et al.* Digital navigation of energy–structure–function maps for hydrogen-bonded porous molecular crystals. *Nat. Commun.* **12**, 817 (2021).

[7] Pulido, A. *et al.* Functional materials discovery using energy–structure–function maps. *Nature* **543**, 657–664 (2017).

[8] O'Shaughnessy, M. *et al.* Porous isoreticular non-metal organic frameworks. *Nature* **630**, 102–108 (2024).

[9] Aitchison, C. M. *et al.* Photocatalytic proton reduction by a computationally identified, molecular hydrogen-bonded framework. *J. Mater. Chem. A* **8**, 7158–7170 (2020).

[10] Campbell, J. E., Yang, J. & Day, G. M. Predicted energy–structure–function maps for the evaluation of small molecule organic semiconductors. *J. Mater. Chem. C* **5**, 7574–7584 (2017).

[11] Han, Y. *et al.* Machine learning accelerates quantum mechanics predictions of molecular crystals. *Phys. Rep.* **934**, 1–71 (2021).

[12] Musil, F. *et al.* Machine learning for the structure–energy–property landscapes of molecular crystals. *Chem. Sci.* **9**, 1289–1300 (2018).

[13] 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).

[14] Fung, V., Zhang, J., Juarez, E. & Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *npj Comput. Mater.* **7**, 1–8 (2021).

[15] Taylor, C., Butler, P. & Day, G. M. Predictive crystallography at scale: Mapping, validating, and learning from 1,000 crystal energy landscapes. *Faraday Discuss.* https://doi.org/10.1039/D4FD00105B (2024).

[16] Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).

[17] Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).

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

[19] Townsend, J., Micucci, C. P., Hymel, J. H., Maroulas, V. & Vogiatzis, K. D. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nat. Commun.* **11**, 3230 (2020).

[20] Lee, Y. *et al.* Quantifying similarity of pore-geometry in nanoporous materials. *Nat. Commun.* **8**, 15396 (2017).

[21] Krishnapriyan, A. S., Montoya, J., Haranczyk, M., Hummelshøj, J. & Morozov, D. Machine learning with persistent homology and chemical word embeddings improves prediction accuracy and interpretability in metal-organic frameworks. *Sci. Rep.* **11**, 8888 (2021).

[22] Boyd, P. G. *et al.* Data-driven design of metal–organic frameworks for wet flue gas CO_2 capture. *Nature* **576**, 253–256 (2019).

[23] Vaswani, A. *et al.* Attention is all you need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2017).

[24] Kang, Y., Park, H., Smit, B. & Kim, J. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nat. Mach. Intell.* **5**, 309–318 (2023).

[25] Cao, Z., Magar, R., Wang, Y. & Farimani, A. B. MOFormer: Self-supervised transformer model for metal–organic framework property prediction. *J. Am. Chem. Soc.* **145**, 2958–2967 (2023).

[26] Wang, J. *et al.* A comprehensive transformer-based approach for high-accuracy gas adsorption predictions in metal-organic frameworks. *Nat. Commun.* **15**, 1904 (2024).

[27] Cui, J. *et al.* Direct prediction of gas adsorption via spatial atom interaction learning. *Nat. Commun.* **14**, 7043 (2023).

[28] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. Preprint at https://doi.org/10.48550/arXiv.1810.04805 (2019).

[29] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2021).

[30] Ramesh, A. *et al.* Zero-shot text-to-image generation. Preprint at https://doi.org/10.48550/arXiv.2102.12092 (2021).

[31] Alayrac, J.-B. *et al.* Flamingo: A visual language model for few-shot learning. Preprint at https://doi.org/10.48550/arXiv.2204.14198 (2022).

[32] Yuan, L. *et al.* Florence: A new foundation model for computer vision. Preprint at https://doi.org/10.48550/arXiv.2111.11432 (2021).

[33] Schwaller, P. *et al.* Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).

[34] Kreutter, D., Schwaller, P. & Reymond, J.-L. Predicting enzymatic reactions with a molecular transformer. *Chem. Sci.* **12**, 8648–8659 (2021).

[35] Kuenneth, C. & Ramprasad, R. polyBERT: A chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.* **14**, 4099 (2023).

[36] Tu, Z. & Coley, C. W. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *J. Chem. Inf. Model.* **62**, 3503–3513 (2022).

[37] Jin, T., Singla, V., Hsu, H.-H. & Savoie, B. M. Large property models: A new generative machine-learning formulation for molecules. *Faraday Discuss.* (2024).

[38] Gao, W., Luo, S. & Coley, C. W. Generative artificial intelligence for navigating synthesizable chemical space. Preprint at https://doi.org/10.48550/arXiv.2410.03494 (2024).

[39] Antunes, L. M., Butler, K. T. & Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *Nat. Commun.* **15**, 10570 (2024).

[40] Magar, R., Wang, Y. & Farimani, A. B. Crystal twins: Self-supervised learning for crystalline material property prediction. *npj Comput. Mater.* **8**, 1–8 (2022).

[41] Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge structural database. *Acta Cryst. B* **72**, 171–179 (2016).

[42] Abnar, S. & Zuidema, W. Quantifying attention flow in transformers. Preprint at https://doi.org/10.48550/arXiv.2005.00928 (2020).

[43] Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **7**, 1–8 (2021).

[44] Adams, H. *et al.* Persistence images: A stable vector representation of persistent homology. *J. Mach. Learn. Res.* **18**, 1–35 (2017).

[45] Moosavi, S. M., Xu, H., Chen, L., Cooper, A. I. & Smit, B. Geometric landscapes for material discovery within energy–structure–function maps. *Chem. Sci.* **11**, 5423–5433 (2020).

[46] Glazer, M., Burns, G. & Glazer, A. N. *Space Groups for Solid State Scientists* (Elsevier, 2012).

[47] van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

[48] Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

[49] Vovk, V. in *Kernel ridge regression* (eds Schölkopf, B., Luo, Z. & Vovk, V.) *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* 105–116 (Springer, Berlin, Heidelberg, 2013).

[50] Rosen, A. S. *et al.* Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter* **4**, 1578–1597 (2021).

[51] Pardakhti, M., Moharreri, E., Wanik, D., Suib, S. L. & Srivastava, R. Machine learning using combined structural and chemical descriptors for prediction of methane adsorption performance of metal organic frameworks (MOFs). *ACS Comb. Sci.* **19**, 640–645 (2017).

[52] Bénard, P. & Chahine, R. Storage of hydrogen by physisorption on carbon and nanostructured materials. *Scr. Mater.* **56**, 803–808 (2007).

[53] Brown, T. B. *et al.* Language models are few-shot learners. Preprint at https://doi.org/10.48550/arXiv.2005.14165 (2020).

[54] Ong, S. P. *et al.* Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).

[55] Krishnapriyan, A. S., Haranczyk, M. & Morozov, D. Topological descriptors help predict guest adsorption in nanoporous materials. *J. Phys. Chem. C* **124**, 9360–9368 (2020).

[56] Price, S. L. *et al.* Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Phys. Chem. Chem. Phys.* **12**, 8478–8490 (2010).

[57] Anthony, S. GDMA: A program for performing distributed multipole analysis of wave functions calculated using the gaussian program system. (University of Cambridge, 2010).

[58] Pyzer-Knapp, E. O., Thompson, H. P. G. & Day, G. M. An optimized intermolecular force field for hydrogen-bonded organic molecular crystals using atomic multipole electrostatics. *Acta Cryst. B* **72**, 477–487 (2016).

[59] Nyman, J., Pundyke, O. S. & Day, G. M. Accurate force fields and methods for modelling organic molecular crystals at finite temperatures. *Phys. Chem. Chem. Phys.* **18**, 15828–15837 (2016).

[60] 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).

[61] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

[62] Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).

[63] Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).

[64] Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

[65] Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132**, 154104 (2010).

[66] Becke, A. D. & Johnson, E. R. A density-functional model of the dispersion interaction. *J. Chem. Phys.* **123**, 154101 (2005).

[67] Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).

Declarations

The authors declare no competing interests.

Supplementary Information

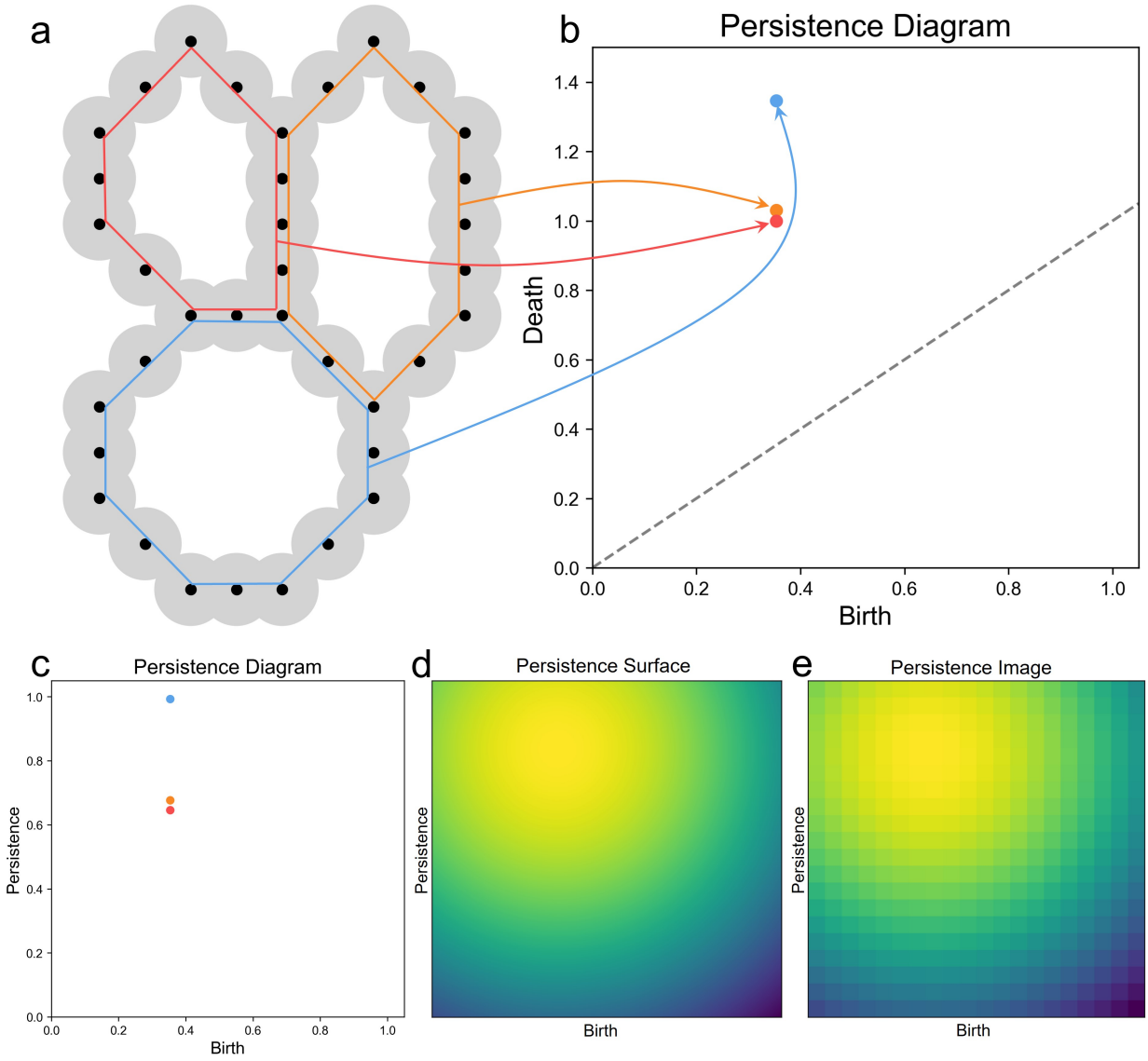

Fig. S1 Scheme outlining mapping of point cloud to persistence image. a, A point set (representing atomic centers) with balls of increasing radius around the points. **b**, 1-Dimensional persistence diagram of the point set in birth-death coordinates. Representative cycles, corresponding to the points in the diagram, are highlighted with matching colours. **c**, 1-Dimensional persistence diagram of the point set in birth-persistence coordinates. **d**, The corresponding persistence surface is defined as a weighted sum of Gaussian functions, with each Gaussian centered at a point in the PD. **e**, The corresponding persistence image is obtained by converting the surface into a finite-dimensional vector, which is achieved by discretising a relevant subdomain and then integrating the surface values over each discrete region. The set resolution of the persistence images is 50 × 50, which is subsequently partitioned into 5 × 5 patches for transformer input.

Persistence images

The process of generating persistence images from point clouds is simplified into a two-dimensional representation, as illustrated in Fig. S1. The molecular crystal is represented as a union of spheres centered on atomic positions, with radii systematically increased to track changes in the topology of their union. These changes, which correspond to the appearance or disappearance of features like loops and voids, are recorded as birth–death pairs in persistence diagrams. The birth value represents the scale at which a topological object appears, and the death value represents when it disappears, with the difference defining the persistence of the topological object. Persistence diagrams, computed using the Dionysus library

(https://github.com/mrzv/dionysus), are transformed into birth–persistence pairs to emphasize the stability and prominence of features. These pairs are then convolved with Gaussian functions, smoothed, and discretized onto a fixed-resolution grid to create persistence images. This approach translates the structural and topological information of the crystal into a standardized vector format suitable for machine learning, enabling direct comparison across structures and efficient extraction of meaningful features.

To normalize the size of molecular crystals, we fill a $(100\ \text{Å})^3$ super cell with the atoms of the crystal. The size is large enough to capture the statistics of the distribution of the topological features in every structure. The resolution of persistence image is 50×50 and the Gaussian spread is 0.15.

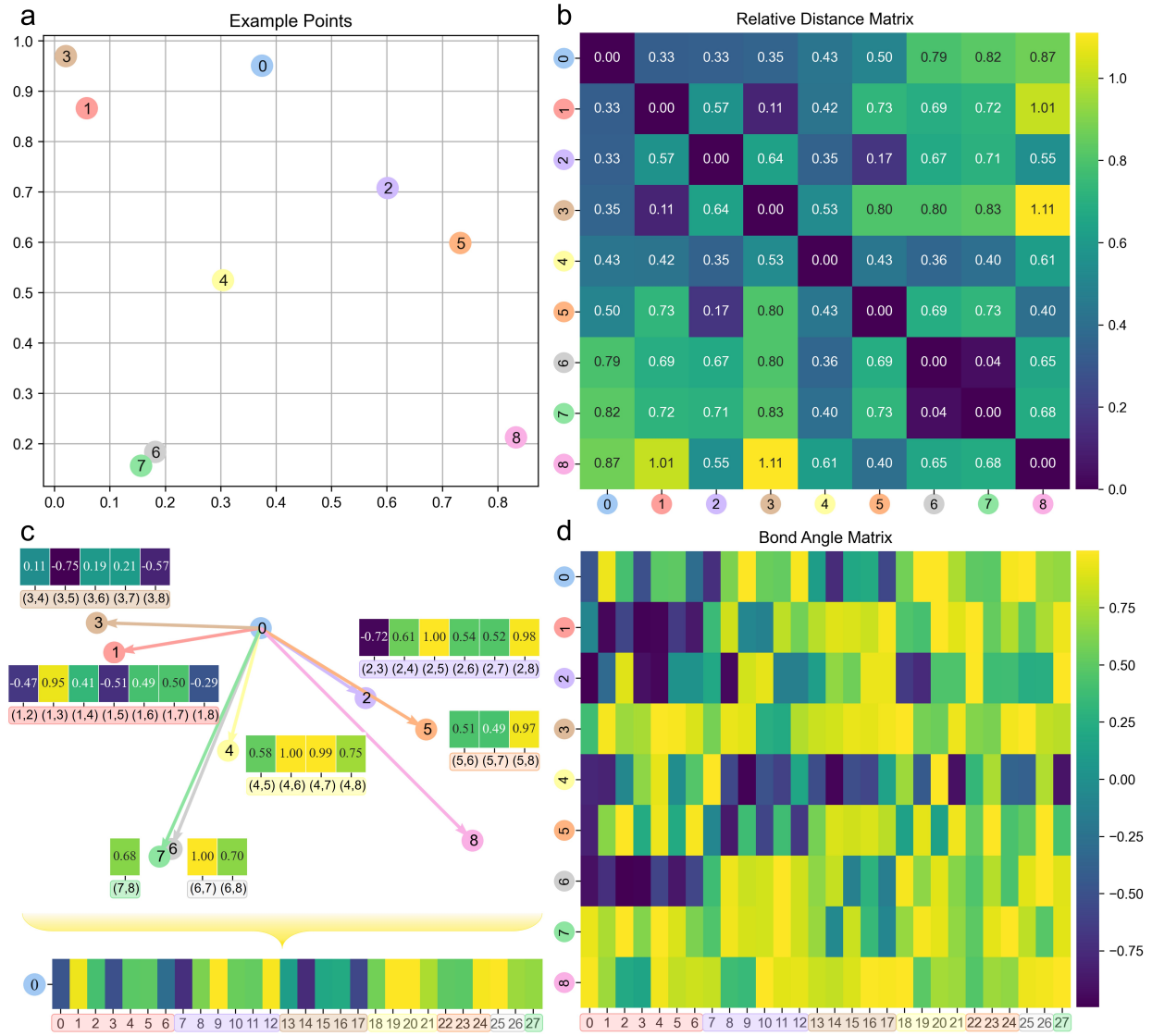

Fig. S2 Scheme of relative positional embedding. a, Example points with randomly generated coordinates. **b**, Relative distance matrix of the example points. **c**, Scheme showing the generation of bond angle vector of point 0. The order of angles is based on the distances between point 0 and neighbour points. **d**, Bond angle matrix of the example points.

Relative postional embedding

The relative positional embedding, in contrast to absolute positional embedding that uses 3D coordinates, employs a relative distance matrix and a bond angle matrix to generate positional embedding. These two matrices are individually processed through separate multi-layer perceptrons (MLPs) to match the dimensionality of the atomic representation and are then directly added to the atomic representation. The relative distance matrix provides information about the distances between each atom and all other atoms, while the arrangement of bond angles in the bond angle matrix is based on the order of atomic distances. This

enables the model to deduce the relative positions of atoms from these matrices. It is important to note that the use of relative positional embedding achieves translational and rotational invariance, capabilities not possible with absolute positional embedding.

The calculation of the relative distance matrix is straightforward, but the computation of the bond angle matrix requires careful design. As shown in Fig. S2c and d, we calculate and organize the angles formed between a central atom and its surrounding atoms based on their distances. Starting from the closest neighbors, we sort the surrounding atoms by their distance to the central atom. For each neighbor, we progressively calculate the angles between its vector and the vectors of other, more distant neighbors. This approach results in a structured matrix of angles, ordered by distance, which captures the spatial relationships around the central atom.

In crystal or molecular structures, multiple atoms may be equidistant from a central atom, especially in symmetric structures. These equidistant atoms are grouped to capture local geometric features and improve computational efficiency, while also ensuring that angle representations are unique and unaffected by the order of atoms within the same distance group. To accurately capture the angular relationships within each group, we calculate a squared mean of the angles between atoms in the group as the internal average angle of this group, introducing a nonlinearity that enhances the distinction between different angular arrangements compared to a simple average. We then calculate the central vector by summing the unit vectors of each atom in the group, ensuring that the resulting vector maintains rotational equivariance. In subsequent angle calculations, when two neighboring atoms belong to the same group, the internal average angle of that group is used. When only one neighboring atom belongs to a group, the central vector of that group is used to calculate the angle with the other atom. This approach reduces computational complexity and preserves the overall directional information, while also ensuring rotational invariance and uniqueness in the encoding. To capture all the possible bond angles, the number of neighboring atoms is set to 8.

To demonstrate the practical utility of using relative positional embeddings in the architecture of MCRT, we compare its pre-training performance against a variant without relative positional embeddings (MCRT-wope). Fig. S3 highlights the importance of relative positional embeddings in better capturing both local and global structural information essential for the APC and SEP tasks.

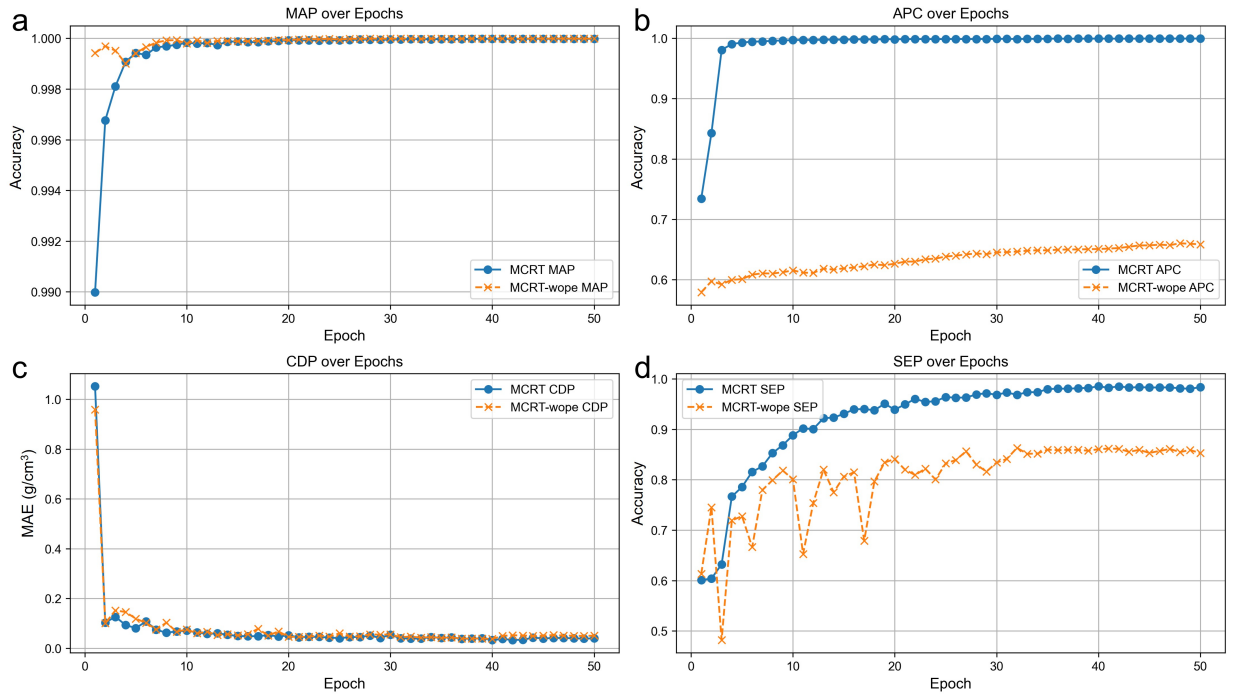

Fig. S3 Pre-training curve of MCRT and MCRT-wope. MCRT-wope denotes MCRT without positional embeddings and evidently fails to capture structural information for the APC and SEP pre-training tasks.

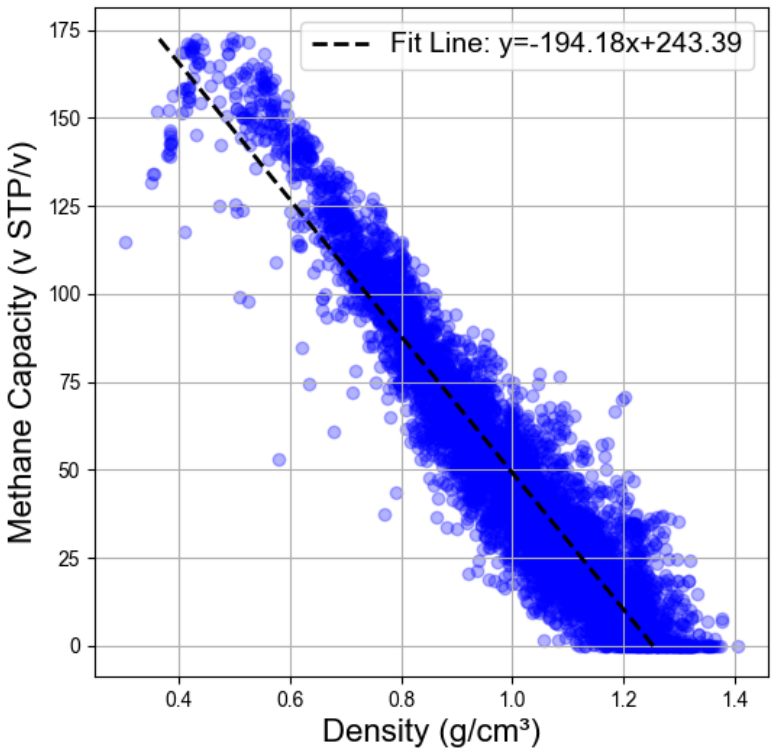

Fig. S4 Correlation plot of methane deliverable capacity vs. density for T2 structures (*Nature* **543**, 657 (2017)). A similar broad correlation between density, which is inversely proportional to pore volume, and gas storage capacity might be excepted for other gases and other frameworks.

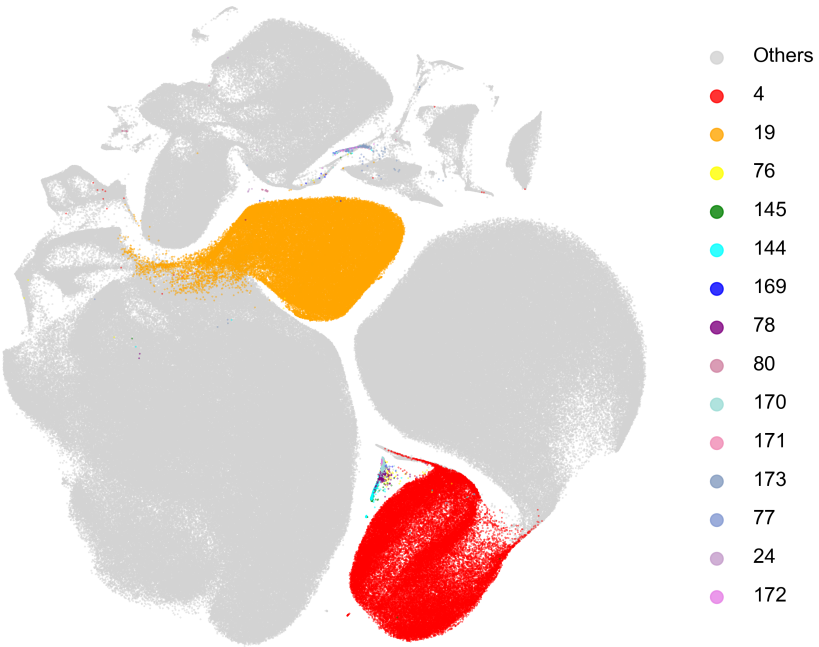

Fig. S5 The t-SNE embeddings of the [CLS] tokens of 706,126 experimental molecular crystals. Obtained from the pre-trained model, with crystals containing only screw axis being coloured by space groups.

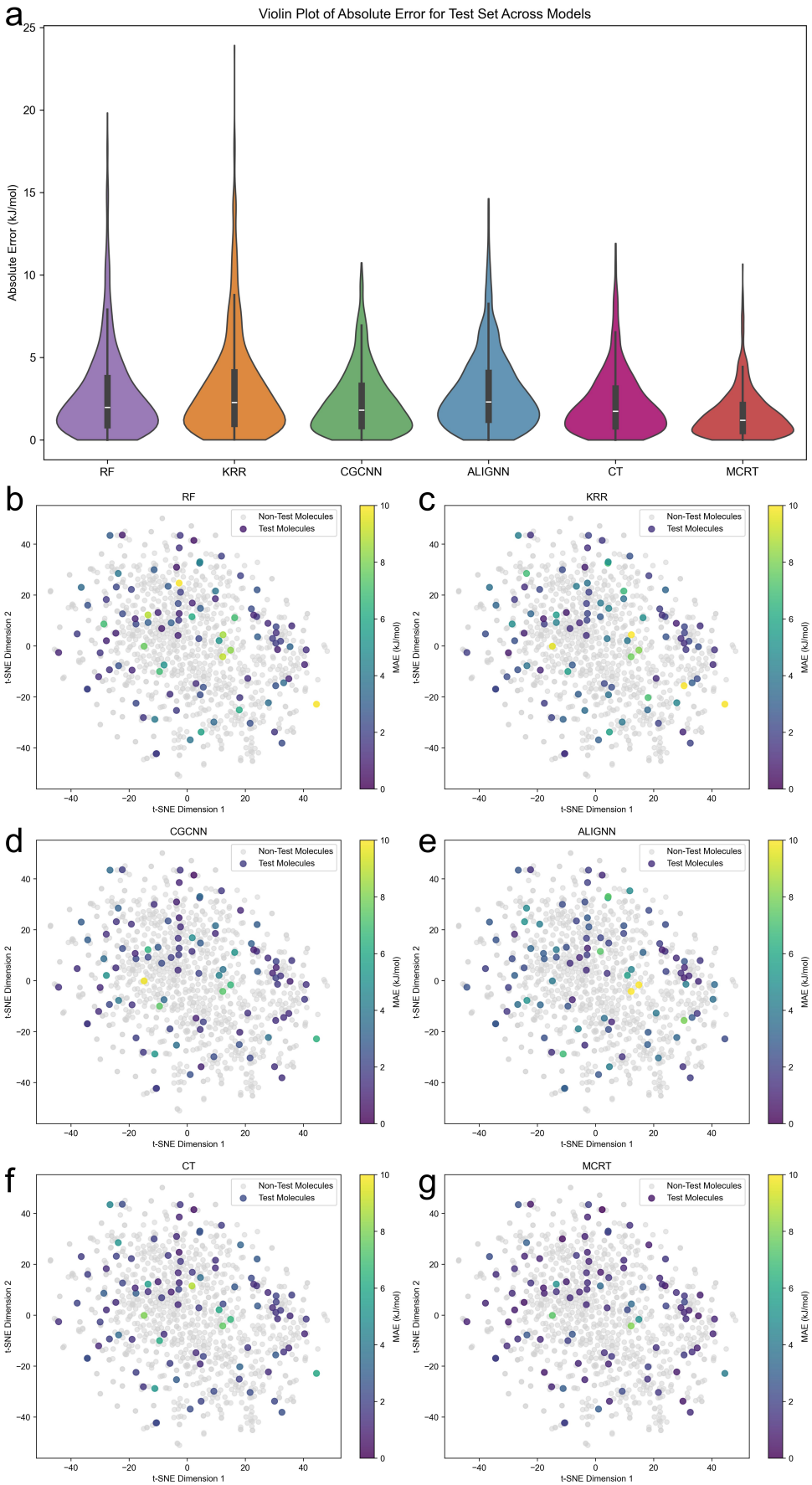

Fig. S6 Test-set results for MCRT and baseline models for the Δ-E task. a, The violin plot showing the absolute errors on the test set for the Δ-E task across MCRT and baseline models. **b**–**g**, The t-SNE embedding of extended connectivity fingerprints (ECFPs) for molecules in the Δ-E task, with colors representing the MAE of each molecule in the test set. MAEs larger than 10 kJ/mol were capped.

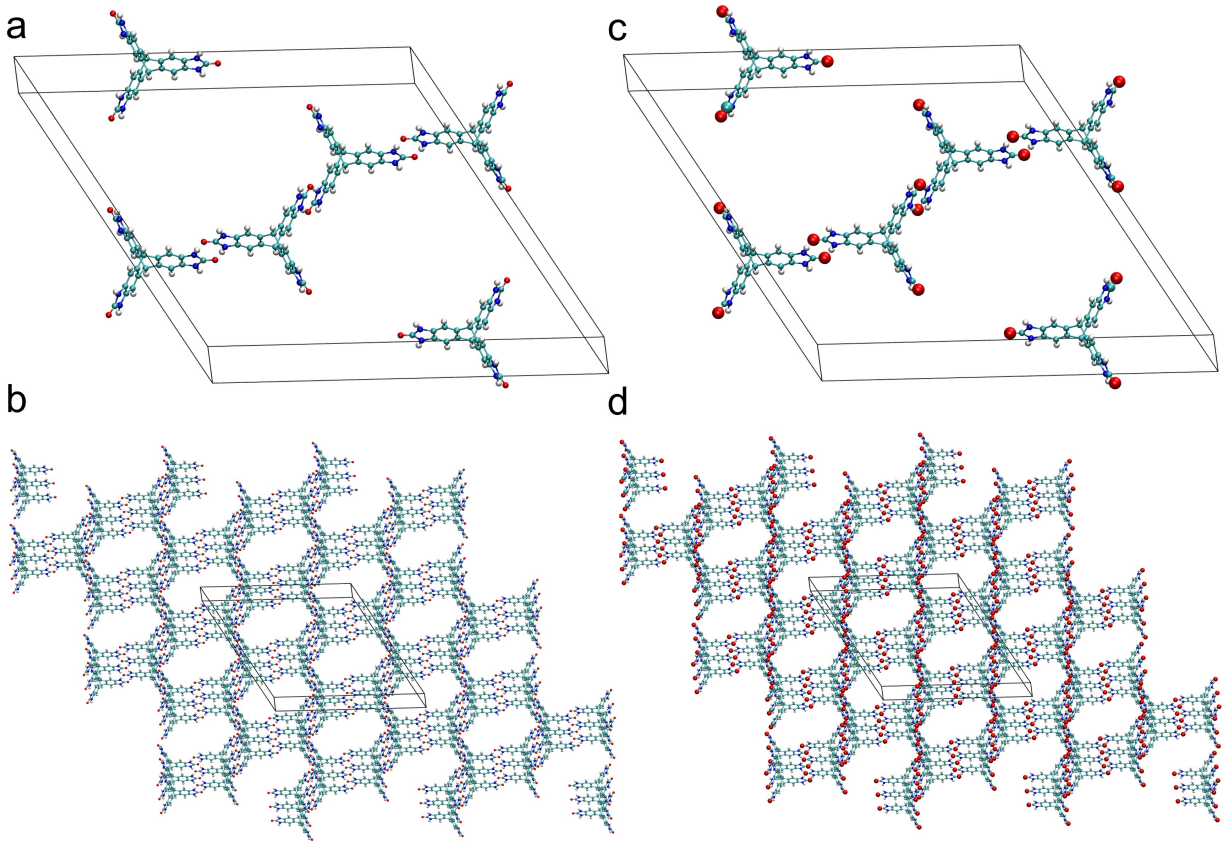

Fig. S7 Attention scores for atom-based embeddings in T2-γ. a, Unit cell and **b**, Supercell of T2-γ with attention scores from fine-tuned MCRT that predicts CH_4 capacity. **c**, Unit cell and **d**, Supercell of T2-γ with attention scores from fine-tuned MCRT that predicts lattice energy. The atomic size is proportional to normalized attention scores, with scores less than 0.5 being clipped to avoid extremely small atoms (colour code: C, cyan; H, white; N, blue; O, red).

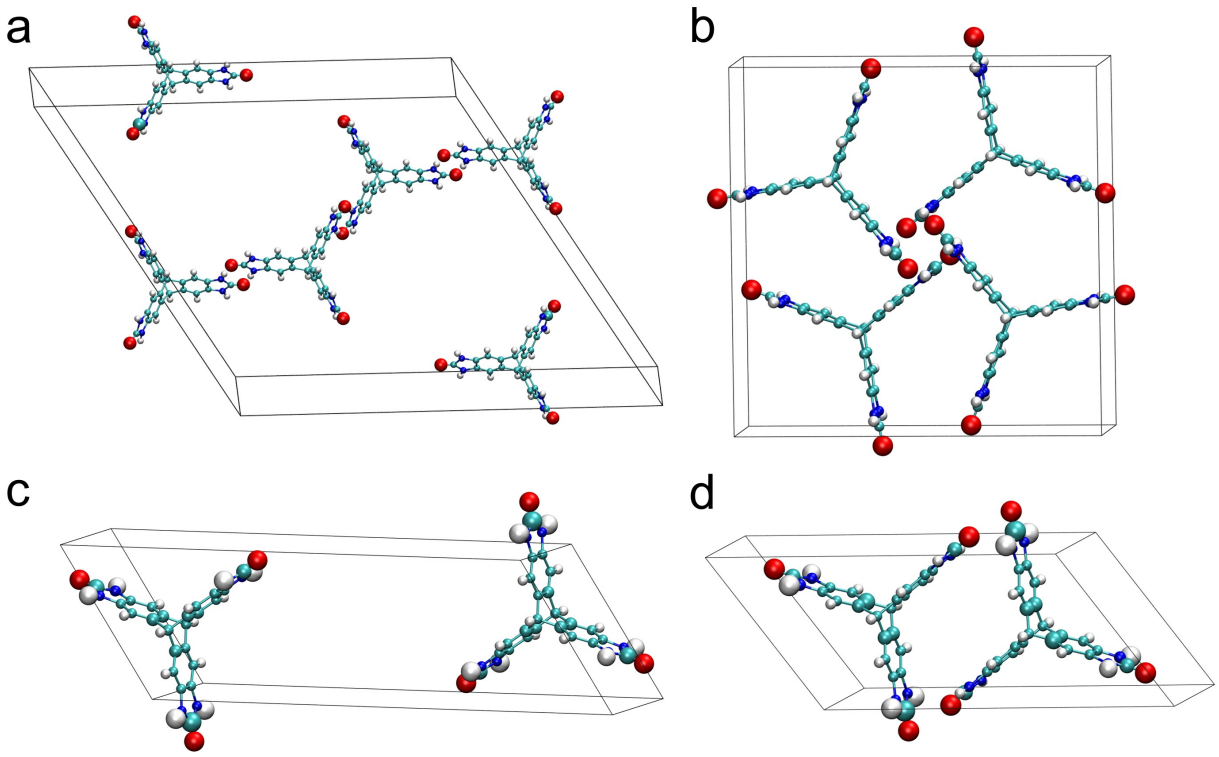

Fig. S8 Attention scores for atom-based embeddings in unit cells of experimental T2 structures. Unit cell of **a**, T2-γ **b**, T2-α **c**, T2-β and **d**, T2-δ with attention scores from fine-tuned MCRT that predicts lattice energy. The atomic size is proportional to normalized attention scores, with scores less than 0.5 being clipped to avoid extremely small atoms (colour code: C, cyan; H, white; N, blue; O, red). For all four of these T2 polymorphs, the atoms involved in hydrogen bonding are the most attended to by the MRCT model. This agrees with experiment, since all four of these polymorphs are dictated by hydrogen bonded 1-D tapes that run parallel to the 1-D pores (*Nature* **543**, 657 (2017)).

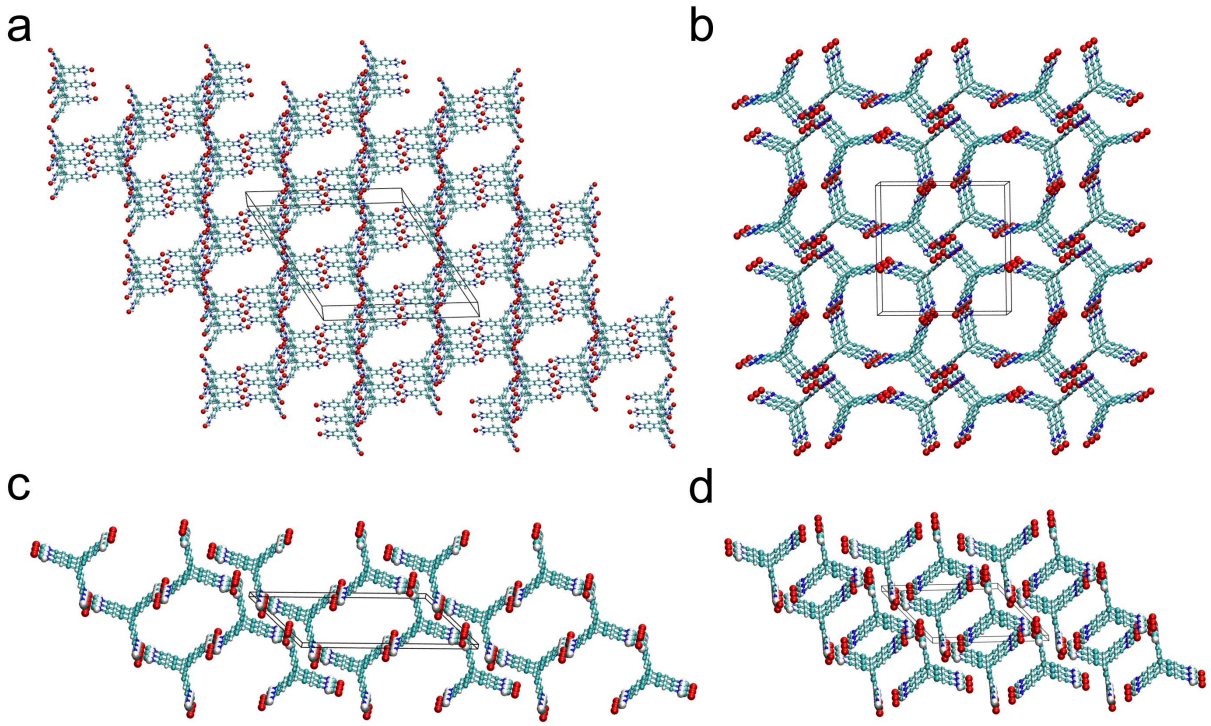

Fig. S9 Attention scores for atom-based embeddings in supercells of experimental T2 structures. Supercell of **a**, T2-γ **b**, T2-α **c**, T2-β and **d**, T2-δ with attention scores from fine-tuned MCRT that predicts lattice energy. The atomic size is proportional to normalized attention scores, with scores less than 0.5 being clipped to avoid extremely small atoms (colour code: C, cyan; H, white; N, blue; O, red).

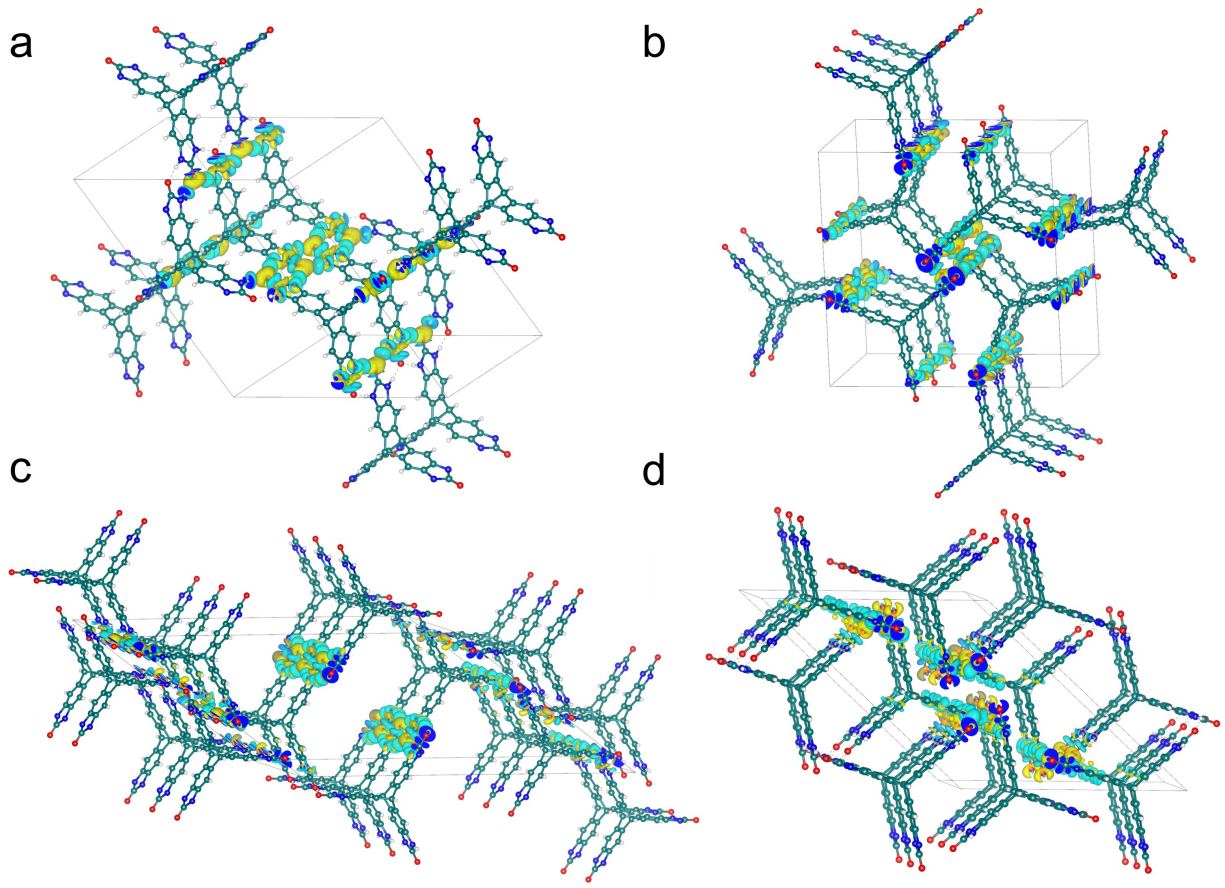

Fig. S10 Electron density difference (EDD) plots. a, T2-γ, **b**, T2-α, **c**, T2-β and **d**, T2-δ highlighting the region of intermolecular interactions. The yellow isosurfaces represent regions with increased electron density, while the blue isosurfaces indicate regions with decreased electron density. These areas of strong intermolecular interaction correspond to the atoms attended to in Fig. S8 and Fig. S9.

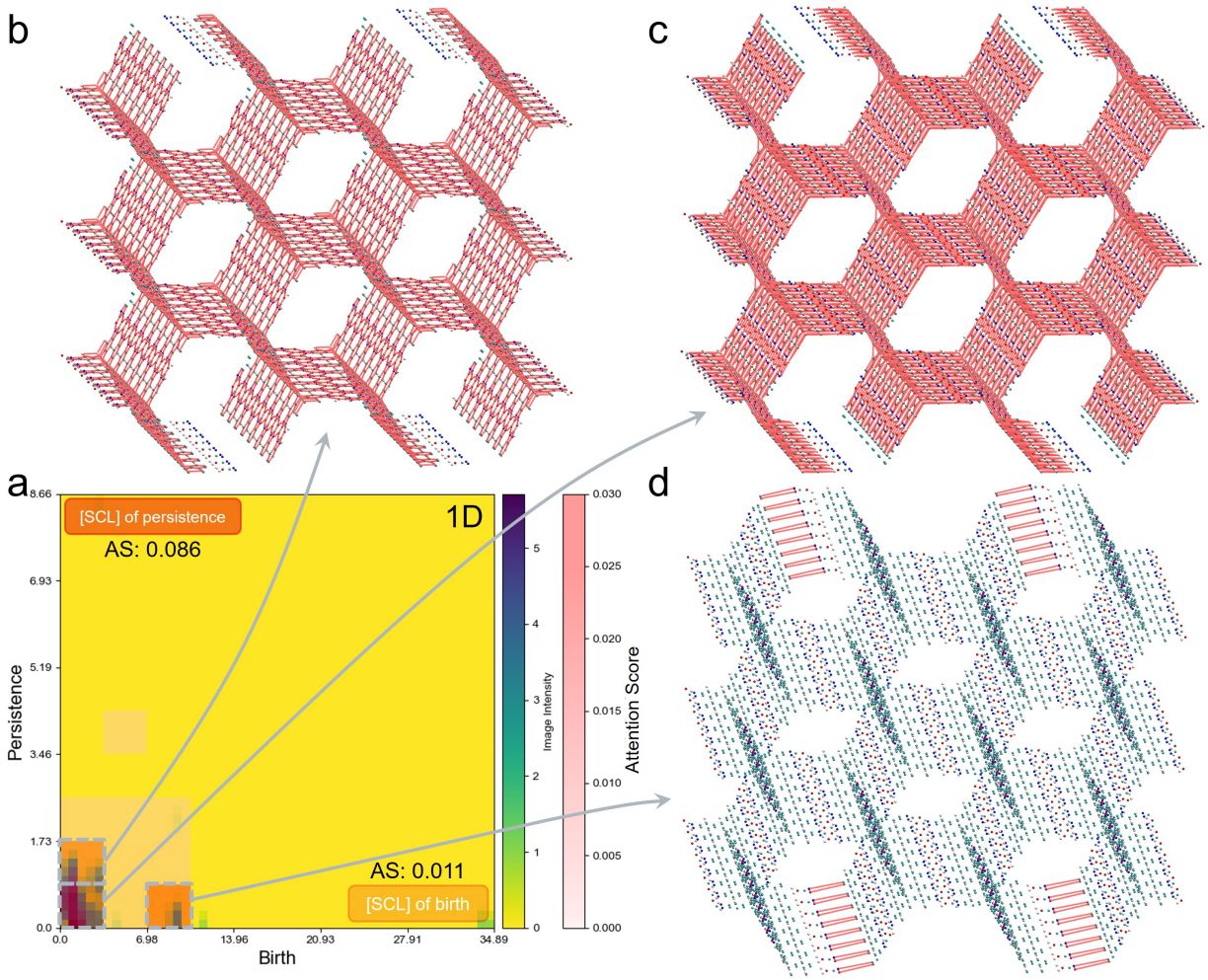

Fig. S11 Representative cycles of three patches with relatively high attention scores in the 1D persistence image of T2-γ. a, 1D persistence images of T2-γ with attention scores from fine-tuned MCRT that predicts CH_4 capacity. **b**, **c** and **d**, The representative cycles of topological objects within the patches with large attention scores.

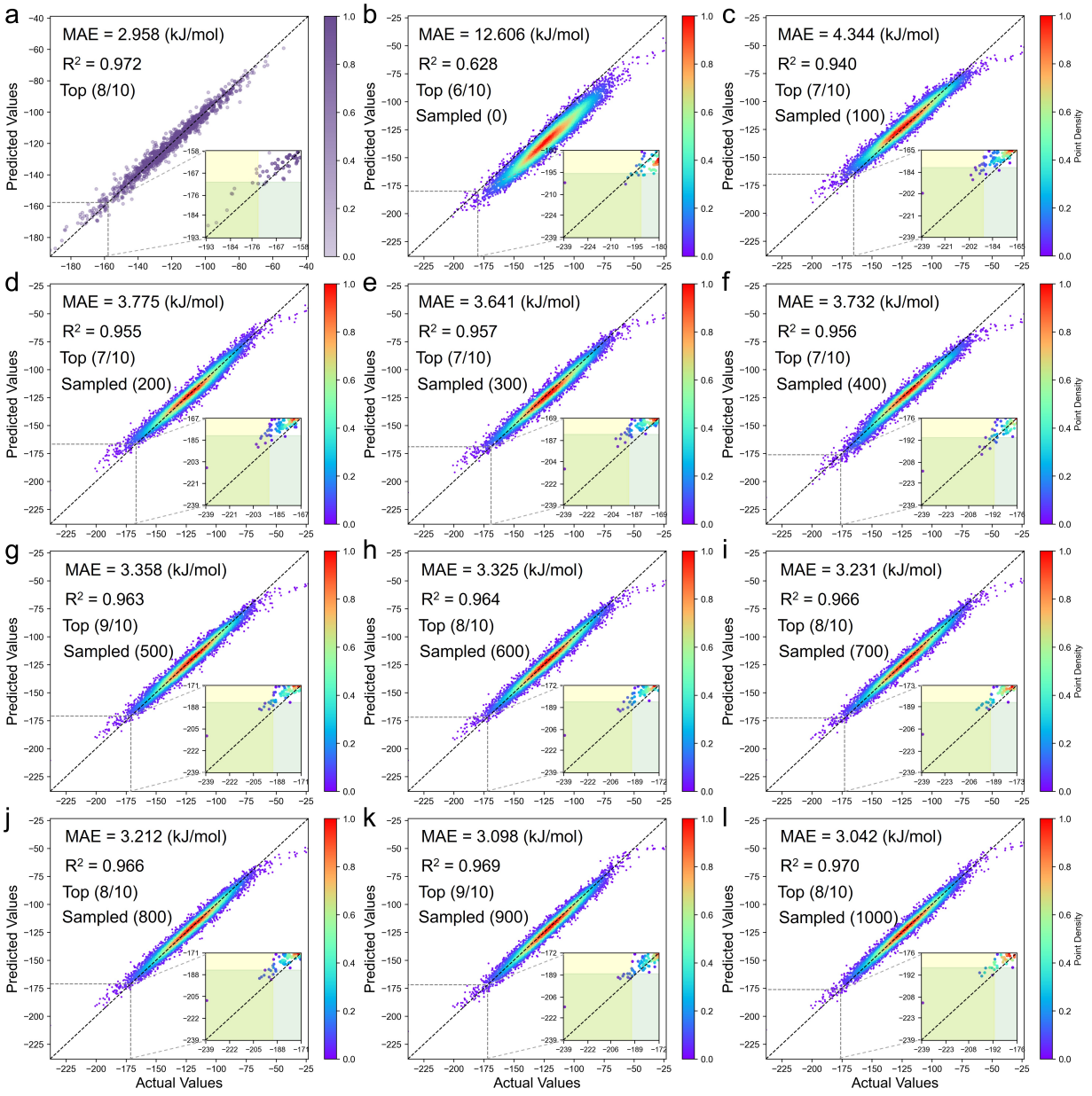

Fig. S12 Prediction results on test set of fine-tuned MCRT. a, LE_T2 (8k). **b-l**, The few-shot prediction results on test sets of fine-tuned MCRT for T2 with 0 (**b**), 100 (**c**), 200 (**d**), 300 (**e**), 400 (**f**), 500 (**g**), 600 (**h**), 700 (**i**), 800 (**j**), 900 (**k**), 1000 (**l**) samples.

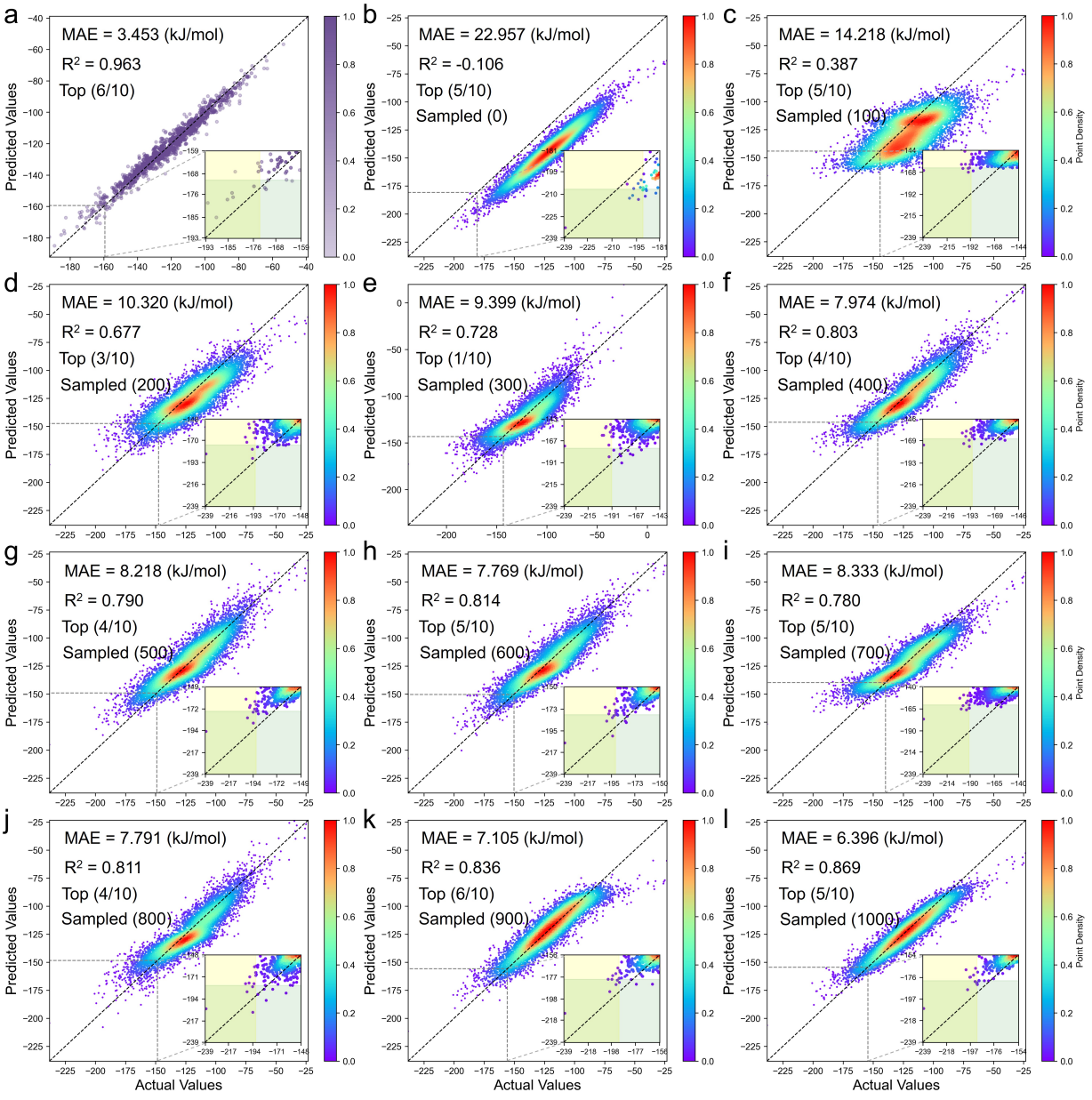

Fig. S13 Prediction results on test set of ALIGNN. a, LE_T2 (8k). **b-l**, The few-shot prediction results on test sets of fine-tuned ALIGNN for T2 with 0 (**b**), 100 (**c**), 200 (**d**), 300 (**e**), 400 (**f**), 500 (**g**), 600 (**h**), 700 (**i**), 800 (**j**), 900 (**k**), 1000 (**l**) samples.

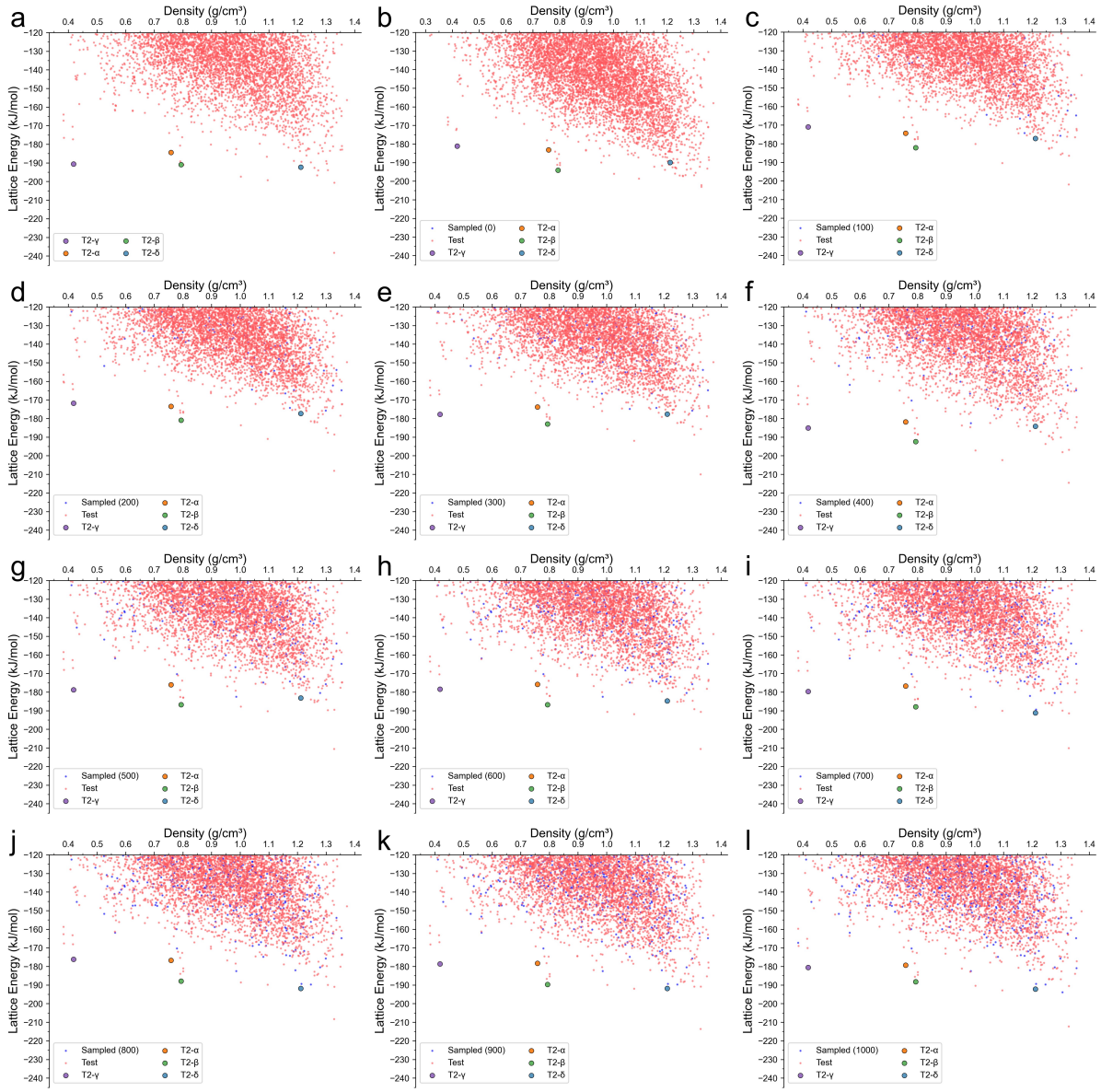

Fig. S14 MCRT-derived CSP energy–density plots. **a**, CSP energy–density plots for T2. **b-l**, Predicted CSP energy–density plots using MCRT for T2 with 0 (**b**), 100 (**c**), 200 (**d**), 300 (**e**), 400 (**f**), 500 (**g**), 600 (**h**), 700 (**i**), 800 (**j**), 900 (**k**), 1000 (**l**) samples. Only structures with a lattice energy of less than -120 kJ/mol are plotted. Even after zero-shot learning, **b**, the four key experimentally-observed porous structures for T2 are identified as being on the leading edge of the energy-density landscape, and 'spikes' emerge that correspond to families of related 1-D porous crystals, as observed in CSP plots generated with forcefield or DFT enery calculations (*Nature* **543**, 657 (2017); *Nature* **630**, 102 (2024)).

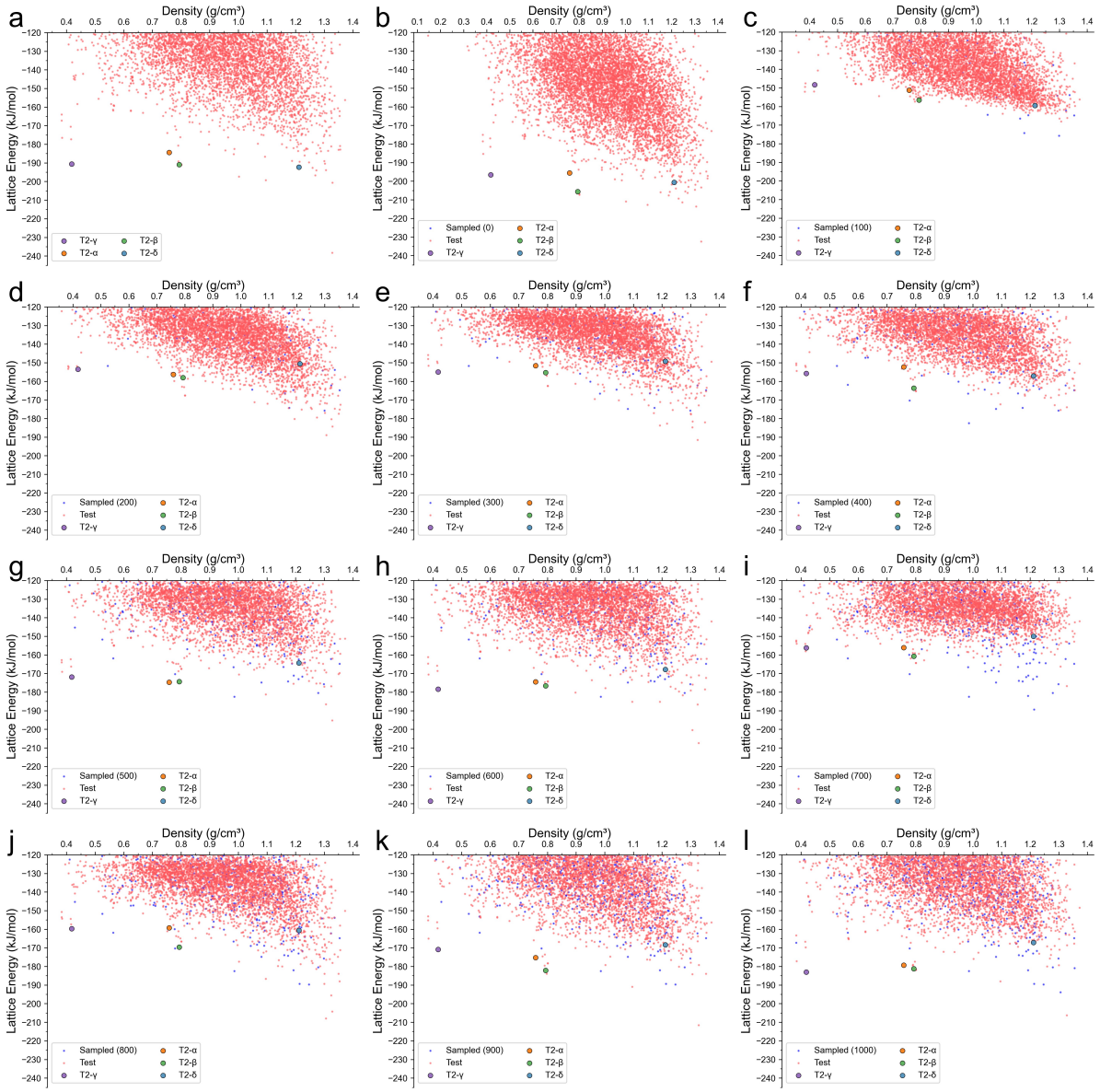

Fig. S15 ALIGNN-derived CSP energy–density plots. a, CSP energy–density plots for T2. **b-l**, Predicted CSP energy–density plots using ALIGNN for T2 with 0 (**b**), 100 (**c**), 200 (**d**), 300 (**e**), 400 (**f**), 500 (**g**), 600 (**h**), 700 (**i**), 800 (**j**), 900 (**k**), 1000 (**l**) samples. Only structures with a lattice energy of less than -120 kJ/mol are plotted.