CEGAN: Crystal Edge Graph Attention Network for multiscale classification of materials environment

Suvo Banik\textsuperscript{1,2}, Deb das Dhabal\textsuperscript{3}, Henry Chan\textsuperscript{1}, Sukriti Manna\textsuperscript{1,2}, Mathew Cherukara\textsuperscript{4}, Valeria Molinero\textsuperscript{3}, Subramanian KRS Sankaranarayanan\textsuperscript{1,2,*}

\textsuperscript{1} Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, United States.
\textsuperscript{2} Department of Mechanical and Industrial Engineering, University of Illinois, Chicago, Illinois 60607, The United States.
\textsuperscript{3} Department of Chemistry University of Utah Salt Lake City, UT 84112, United States.
\textsuperscript{4} Advanced Photon Source, Argonne National Laboratory, Lemont, Illinois 60439, United States.

* Corresponding author: skr ssank@uic.edu

Abstract:

Machine learning models and applications in materials design and discovery typically involve the use of feature representations or “descriptors” followed by a learning algorithm that maps it to a user-desired property of interest. Most popular mathematical formulation-based descriptors are not unique across the atomic environments or suffer from transferability issues across different application domains and/or material classes. In this work, we introduce the Crystal Edge Graph Attention Network (CEGAN) workflow that uses graph attention-based architecture to learn unique feature representations and perform classification of materials across multiple scales (from atomic to mesoscale) and diverse classes ranging from metals, oxides, non-metals and even hierarchical materials such as zeolites. We first demonstrate a case study where the classification is based on a global, structure-level representation such as space group and structural dimensionality (e.g., bulk, clusters, 2D, etc.). Using representative materials such as polycrystals and zeolites, we next demonstrate the transferability of our network in successfully performing local atom level classification tasks, such as grain boundary identification and other heterointerfaces. Finally, we demonstrate classification in (thermal) noisy dynamical environments using a representative example of crystal nucleation and growth of a zeolite polymorph from an amorphous synthesis mixture. Overall, our approach is agnostic to the material type and allows for multiscale classification of features ranging from atomic-scale crystal structures to heterointerfaces to microscale grain boundaries.

Keywords: Multiscale material classification, Machine Learning, Graph Attention Network, Crystal Graphs
Introduction:

Characterization of materials with unique properties\(^1\) is at the core of data-driven material design and discovery\(^6\). A relatively small fraction of materials has been characterized either experimentally or with computational methods, compared to their anticipated potential diversity across a vast chemical space. Given the surge in the development of materials databases\(^8\) in recent years, there is an urgent need for automated tools to analyze large amounts of structural data. In this regard, distinguishing the unique characteristics across different classes of materials with varying dimensionality can provide key insights into learnable aspects which are crucial for state-of-the-art ML tools to be successfully implemented in the design and discovery of new materials with unique properties. To achieve such distinction, ML models typically involve the use of fingerprints or descriptors\(^11\) that allow a learning algorithm to map the fingerprint to a user-desired property of interest. A descriptor that maps the crystal features in a vector space should always be (1) invariant to basis symmetries such as rotation, reflection, translation, and permutation of atoms\(^12\), (2) unique to the system applied identified, but sensitive towards variation in properties, and (3) simple and robust. Additionally, these features play a crucial role in a plethora of applications such as quantitative structure-property relationship (QSPR)\(^1,11,17\), development of interatomic potentials\(^13,21\), and prediction of atomistic configurations based on targeted properties\(^25\), surface phenomena\(^30\).

A feature representation is constructed primarily in two ways (i) using a predefined mathematical formulation, or (ii) learning the representation by combining fundamental low-level features and correlating them to the relevant task being performed using ML methods. A plethora of mathematical formulation-based descriptors\(^12,15,18,31\) such as radial distribution functions (RDF), angular distribution function (ADF), common neighbor analysis (CNA)\(^31\), adaptive CNA\(^31\), Centro symmetry parameter (CSP)\(^31\), Voronoi analysis\(^31\), Steinhardt order parameter (SP)\(^32\), bond angle analysis (BAA)\(^33\), and neighbor distance analysis (NDA)\(^31\) are widely used for featurization. A majority of these are very simple and of a very local nature, \textit{i.e.}, mostly capable of differentiating ordered and disordered structures. Improving upon these, a set of features can be developed using pairwise feature matrices and their transformations\(^12,18,20,34,35\). These features may be as simple as pairwise distances e.g., Weyl matrices\(^35\), Z-matrices\(^36\), or pairwise electrostatic interactions between atoms (Coulomb matrix\(^18\), sine matrix\(^19\)). A more comprehensive representation of these matrices is permutation histograms\(^15\) MBTR (Many-Body Tensor Representation)\(^12,20\), BOB (Bag of Bonds). The advantages of these methods are that the pairwise features are translationally invariant, and these matrices present a unique representation of the system. However, a major setback of these matrix representations is that they are not invariant to changes in atom ordering. A very popular approach involves the use of smooth overlap of atomic positions (SOAP)\(^14\) descriptors constructed by expanding the atomic neighbor density \(\rho\) on the basis of spherical harmonics and then expanding it on a radial basis to obtain the rotational invariant power spectrum\(^14\). SOAP provides a very robust representation of the local environment.
in a smooth and continuous manner which makes it very suitable for mapping potential energy surfaces. Nevertheless, most of the existing approaches for structure characterization that involve global comparison between two structures are either based on a simple aggregation-based method (average kernel) that causes loss in resolution or are computationally expensive such as, for example, the best match kernel method\textsuperscript{14}. Another setback is that the number of descriptors increases quadratically\textsuperscript{6} with the increase in the number of chemical species, precluding their applicability to multicomponent systems.

In this context, graph neural networks (GNNs)\textsuperscript{37–39} have been widely used in node level as well as graph level classification tasks with remarkable success. Furthermore, recent developments in the area of graph attention networks\textsuperscript{40} make the task more accurate by learning the interaction between node-level features. These networks tend to learn flexible representations by combining very fundamental low-level features (interatomic distances, bond angles, etc.), and yet produce a graph-based input that very accurately maps to the target-specific tasks. This caters to the limitations of static descriptors bound by predefined mathematical formulations.

In crystal systems, GNNs can be made to operate on atom-based graphs to create node-level embedding through convolutions on neighboring nodes and edges\textsuperscript{41–48}. More layers of convolutions tend to capture higher-level information. A widely used framework for crystal systems is the Crystal Graph Convolutional Neural Network (CGCNN)\textsuperscript{47,48}. Xie and Grossman have shown that CGCNN can directly learn material properties from the connectivity of atoms in a crystal, thus enabling an interpretable representation of crystalline materials\textsuperscript{49}. Graph attention-based architecture\textsuperscript{44} has also recently been implemented for the structure to property mapping in atomic systems. Traditional CGCNN architecture tends to map structure to property by using a diverse set of atom-level features (e.g., group number, period number, atomic number, electronic structure, etc.), and crystal graphs with simple edge feature such as pairwise interatomic distances.

While it is true that predefined mathematical formulation-based descriptors are useful when there isn’t sufficient data to learn from, they largely suffer from transferability issues due to lack of flexibility. On the other hand, current existing graph-based CGCNN architectures do not incorporate orientational features\textsuperscript{41,45} that are very relevant for classification tasks in a multitude of atomic environments. Moreover, these features tend to play a more significant role in classification tasks than features belonging to different atomic species. Although there have been recent applications\textsuperscript{41,45} that include orientational features in their network architecture, they are more complex in nature and mostly focus on property prediction. To elucidate the issues involving transferability and applicability, we present two distinct classification scenarios (Fig.1). To start with, we classify the liquid and glassy-amorphous phases of a representative material such as silicon. Both liquid and amorphous phases are disordered with no symmetry whatsoever and only differ in density and coordination number. Fig.1(a) displays the variability in the coordination environment of the
Fig. 1 Classification of Si liquid and amorphous (glass) phases. (a) Typical coordination number distribution of liquid and amorphous structures used for this study. (b) tSNE plot of SOAP representation of the crystalline and amorphous phases in the validation dataset. (c) tSNE plot for Bond order parameters ($Q_2$, $Q_4$, $Q_6$ + CN(Coordination Number)) feature representation. (d) tSNE plot of embeddings of the validation dataset obtained training a classic CGCNN model. (e) Shows the classification of hexagonal and cubic stacked atoms in a coarse-grained ABABCB stacked ice with Bond order feature and CHILL+37. (f) by using SOAP (g) by using Order parameters ($Q_2$, $Q_4$, $Q_6$) + CN(Coordination Number) and (h) from the trained CGCNN model.

liquids and amorphous phases used in this study. We use a dataset containing 2000 structures with 50% liquid and 50% amorphous (See supplementary information S.1 for the details of the data generation methods). With a train-to-validation split of 80%:20%, a traditional CGCNN model is trained. From the t-SNE (t-distributed stochastic neighbor embedding) of feature representation of the validation dataset, with SOAP (cutoff 6Å) (See supplementary Fig.S1 (c) (d) for different cutoffs) (Fig.1(b)) there is no distinct separation of the phases in features space indicating the inability of SOAP to distinguish the individual phases. On the other hand, simple Bond order bases features ($Q_2$, $Q_4$, $Q_6$) + CN(Coordination Number) (cutoff 6Å) (Fig.1(c) and a trained CGCNN (Fig.1(d)) can clearly characterize the phases with decent separation in feature space. The second task involves the identification of particles belonging to local motifs (hexagonal or cubic) in a stacking disordered (ABABCB), coarse-grained ice. The correlated bond order-based CHILL+30 is used as a benchmark for labeling the data (Fig.1 (e)). Similar to the earlier case, we
employ order parameters, SOAP, and CGCNN for this classification. The training data of CGCNN comprises a pure cubic, hexagonal, and stacking fault (ABCBCB) ice structure. The results in Fig.1 (f) (g) (h) indicate that while SOAP is able to classify local motifs, CGCNN or the order parameters-based features fails to do so. Which is converse to the fact that SOAP wasn’t able to characterize to structure belonging to liquid of amorphous class while its two counterparts could. This clearly demonstrates a transferability issue in present characterization techniques across various problems at different scales. Despite traditional GNNs (such as CGCNN) showing exceptional promise in learning flexible feature representation at a graph level (global) their performance in local environments is not as good as their global counterpart and remains mostly unexplored.

Clearly, there is a need for a method that is not only transferrable but adaptable to variabilities in the material environment while providing accurate characterization at different scales. To the best of our knowledge, most efforts on crystal graph neural networks have been restricted to mapping structures to properties and a few property-based prediction tasks. There is still an immense untapped potential of GNNs in classification at both structure (global) and atomic(local) levels. In this work, we introduce a graph attention-based workflow that operates on edge graphs, convoluting on edges and bond angle features and passing messages in between (Fig.2), to learn feature representation of material environments. A great advantage of attentional-based architectures is that they can learn the importance of feature vectors (i.e., bonds, angles) in a given neighborhood of each atom and put emphasis on ones unique to the task being performed. This helps in increasing the performance by ignoring redundant and unnecessary information.

We demonstrate the efficacy of our workflow in classification tasks at both the atom-level (local) and structure-level (global) using a wide range of representative examples from materials applications. For global level classification, we perform two tasks. The first is classifying a diverse range of materials based on their space groups, and the second is classifying them based on their dimensionality (cluster, bulk, 2D, etc.). We base the local atom level classification on structural motifs (FCC, BCC, HCP, diamond cubic), and demonstrate its use on a classic problem of grain boundary identification and grain size distribution. Finally, to validate the efficacy of our workflow in environments with thermal variations or noise, we deploy our classification workflow to facilitate the study of nucleation and growth of a zeolite, a complex porous crystal, in molecular dynamics simulations of synthesis.
Fig. 2 Basic architecture of CEGAN workflow. Crystals are converted into atom graphs and edge graphs are obtained from atom graphs [(a), (b)]. (c) Shows alternate message passing and hierarchical interaction between edge and angle convolutions. Finally, the atom feature, convolved edge, and angle features are concatenated to produce the final representation. (d) shows a tSNE plot of the feature vector of liquid and amorphous structures as predicted by CEGAN workflow on an identical validation dataset as Fig. 1. (e) Shows identification of particles belonging to a hexagonal and cubic motif in an ABABCB stacked ice by a trained CEGAN workflow.

**Results**

**Edge graph representation:**

Edge graphs are higher-order representations of atomic graphs with edges as nodes and bond angles as connections between a pair of edges Fig.2(b). We start from a crystal structure, creating its atom-graph (atom as nodes, bonds as edges) based on a fixed number of nearest neighbors. The edge graph is extracted from the atom-graph afterward (Fig.2 (a) (b)). The edge features ($e_{ij}$) are obtained by expanding the pairwise distance on Gaussian basis functions while the bond angle features ($\theta_{ijk}$) are obtained by expanding the cosines of the bond angles on a gaussian basis as well.

**Hierarchical message passing:**
One main feature of the proposed CEGAN architecture is the hierarchical interaction between edge and angle layers Fig. 2(c) (See methods section). The edge layer always gets updated first. This follows the hierarchy that the bond angles are constructed from a pair of edges and any change at the edge level should get updated first before passing the information onto the corresponding angle. Which gives $n-1$ angle convolution operations for $n$ edge convolutions where “$n$” is an integer.

**CEGAN Workflow for multiscale classification:**

The architecture of the CEGAN workflow used to perform multiscale classification of materials is shown in Fig.2(c). The edge-graph representation of the structures is passed to the Hierarchical message passing block for the convolution operations. The output of the convolved feature vectors from the edge and angle convolution layers are then passed to the aggregation block via dense layers (Linear transformation), where feature representations of each of the structures are generated for the prediction task. The choice of the number of edges and angle convolution layers to be employed depends on the scale at which the classification tasks are being performed. For local-level tasks, it is preferable to have fewer convolutions while its global application requires more. In this work, we select an optimal number of convolutions that results in the best performance of our model for each of the tasks being performed (Table.1). Similar to the choice of the number of convolutions, the number of neighbors considered for the graph constructions also affect the model performance. The number of neighbors used for each of the tasks is reported in Table.1. It is to be noted that for all classification tasks performed in this work, we keep the input dimension of edge and angle feature vectors to be 80. We maintain uniformity of samples belonging to each class in both training and testing data while the splitting of any individual class is done randomly at a given ratio.

| Classification task                  | Edge Convolution | Angle Convolution | Nearest neighbors |
|--------------------------------------|------------------|-------------------|------------------|
| Amorphous and liquid                 | 2                | 1                 | 12               |
| Stacking disordered Ice              | 1                | 0                 | 16               |
| Space groups                         | 2                | 1                 | 12               |
| Dimensionality classification        | 2                | 1                 | 12               |
| Grain size distribution (FCC)        | 1                | 0                 | 12               |
| Grain size distribution (BCC)        | 1                | 0                 | 14               |
| Grain size distribution (Diamond)    | 1                | 0                 | 16               |
| Grain size distribution (HCP)        | 1                | 0                 | 12               |
| Dynamical classification with noise  | 1                | 0                 | 12               |
Classification of liquid and amorphous silicon and stacking-disordered ice:

We start by employing our CEGAN workflow for the classification tasks as discussed in Fig.1. (a) Classification of liquid and amorphous phases (Silicon) (b) Characterization of local motifs (Hexagonal or Cubic) in stacking disordered ice (ABABCB). CEGAN is trained on the same training data as CGCNN (Fig.1 (d), (h)). The validation data is also kept identical. From Fig.2 (d) the t-SNE plot of the feature vectors of liquid and amorphous structures as predicted by CEGAN, it is evident that CEGAN has been able to distinguish amorphous and liquid phases of silicon conspicuously. Fig.2 (e) also depicts the ability of CEGAN to precisely classify local cubic and hexagonal motifs in stacking fault structures where CGCNN performed miserably. CEGAN can overcome the challenge of transferability for both applications ranging from global to local levels while its counterparts, such as the traditional CGCNN, and descriptors such as SOAP fail to do so (Fig.1).

Characterization of crystal structures based on their space groups:

The space group of a crystalline system directly correlates to its structural motif, albeit at a global level. We demonstrate that the CEGAN framework can classify several different material classes based on their space groups. For this classification task, we use the same dataset as in. The space group of each crystal is calculated using a Pymatgen package. The dataset contains a total of 10517 crystal structures with 7 crystal classes belonging to 8 different space groups. For the elemental system, the classes are body-centered tetragonal (bct, 139 and 141), rhombohedral (rh, 166), hexagonal (hex, 194), simple cubic (sc, 221), fcc (225), diamond (dia, 227), and bcc (229) respectively. (See Supplementary Fig. S2).

We start with the dataset having a train-to-validation ratio of 90%:10% and train CEGAN, CGCNN, and a SOAP_ML workflow on this dataset. It is worth noting that our goal is to map SOAP feature vectors (cutoff 6Å) directly to the space group. So, instead of passing the SOAP features through consecutive dense layers (Linear transformation) with nonlinear activations, we have only one dense layer that directly maps it to the target space (SOAP_ML workflow) conforming to the specification used in CEGAN after the aggregation block (Fig.2 (c)). The accuracy on the validation dataset is shown in Fig.3 (d). The CEGAN workflow achieved an accuracy of ~100% on the validation (Fig. 3(d)) set. The confusion matrix of CEGAN (Fig.3(a)) also demonstrates a perfect identification (No off-diagonal entries) of each class belonging to different space groups (See supplementary Fig. S2 (b)-(i) for more details). The CGCNN on the other hand achieves an accuracy of ~83% on the validation dataset with major confusion (Fig.3 (b)) between the hexagonal (194) and fcc (225) structure. This is very evident from the fact that FCC and HCP are close-packed with a 74% atomic packing factor, and 12 nearest neighbors for both, which results in an identical graphical representation of the structures unless the orientational order of the particles is
Fig. 3 Global classification of crystals structures based on space groups. (a), (b), and (c) show the confusion matrix for CEGAN, CGCNN, SOAP ML, and workflow respectively. (d) Shows accuracy of prediction on the validation set for the 3 different architectures used. (e), (f), and (g) show the t-SNE plot of the embeddings in feature space as learned by the CEACN, CGNN, and SOAP ML, respectively.

The CCGCNN not having these attributes in its graphical representation significantly impacts its performance. The performance of SOAP_ML workflow is poor indicating that SOAP in its current mathematical state, however, does contain all the information but is not flexible enough to be directly mappable to the target space group. The degree of characterization can also be visualized in the t-SNE plot of the feature space representation on the validation dataset (Fig. 3 (e-g)). There is a clear distinction in the representation of each class for CEGAN, while CGCNN and SOAP feature vectors display a lack of resolution in the representation of each class in the feature space.
Classification of polymorphs across various structural dimensionalities:

Next, we demonstrate the ability of CEGAN to perform classification on material polymorphs across various dimensionalities, from clusters (0D) to sheets (2D) to bulk (3D). Carbon is known to have a diverse range of allotropes across these dimensionalities, making it an excellent candidate for validating the performance of our network for dimensionality classification. We start with a dataset of 511 bulk structures collected from the Samara Carbon Allotrope Database (SACADA)\(^5\), Monolayer C polymorphs\(^6\), and Graphite with varying interlayer distances, a collection of different Graphite allotrope and 2D polymorphs Carbon sampled using CASTING framework\(^7\) and LCBOP potential\(^8\) making a total of 612 2D structures and 704 C nanoclusters\(^9,\)\(^10\) resulting in a total dataset of 1827 configurations (See supplementary Fig. S3). We divide our dataset into 80% training and 20% validation.

Fig. 4 (a) shows the confusion matrix for the dimensionality classification. CEGAN workflow can classify the structures with \(\sim 99\%\) accuracy. Fig. 4 (b) shows the t-SNE plot of the embeddings of the validation set data. A clear distinction between phases can also be observed in the feature space which displays the capability of CEGAN to characterize polymorphs of different dimensions. It is worth mentioning that dimensionality is a defining material parameter, depending on which material can exhibit dramatically different properties\(^61\). Identification of materials based on their dimensionality is a crucial aspect of new material design and prediction\(^62\). While 3D crystalline objects are well documented among the experimentally known crystals, the same is not true for low dimensional structures such as 2D or 0D. For example, in a few cases where isolated 2D carbon layers tend to form porous bulk-like polymorphs which makes it difficult to categorize and distinguish them from typical layered structures.
Fig. 5 Grain size distribution of polycrystals of (a) Aluminum (FCC), (b) Tungsten (BCC), (c) Silicon (Diamond), (d) Magnesium (FCC) computed using CEGAN + DBSCAN clustering and CNA (Common neighbor analysis) + DBSCAN clustering.

Grain boundary identification:
Characterization of local motifs in full 3D samples of polycrystalline materials and accurately identifying grains and boundaries is a nontrivial task and has a plethora of applications in material science problems. Although there are many methods used for grain characterization\(^{31,33}\), there is no gold standard for identifying the grain size distribution in polycrystalline materials as the predictions widely vary with the methodology used. We use CNA\(^{31}\) (Common Neighbor Analysis) as a benchmark and use it to generate labels for the training and validation data. 4 representative polycrystal classes (i) Face-Centered Cubic (FCC-Al), Body-Centered cubic (BCC-W), Diamond (Si), and Hexagonal Closed Pack (HCP-Mg) with 40 grains are used for the prediction task. For the prediction of each of the aforementioned classes, we generate 10 polycrystalline training samples (See supplementary Fig. S4) with atomsk\(^{63}\) package. The overall characterization is carried out with a two-step approach. First, we (a) label the atoms locally in terms of their crystalline motifs (Eg, FCC, BBC, etc) and,
then, we (b) apply an unsupervised learning DBSCAN\textsuperscript{64,65} clustering to identify the size of the grains in the polycrystal samples. The grain size distribution and the number of particles belonging to crystalline motifs as predicted by CEGAN and CNA have been compared in Fig.5. It is to be noted that the ordinary CNA cannot distinguish diamond structure. Hence, we use modified CNA\textsuperscript{66} for the creation of the labels of the Si (Diamond structures). The number of nearest neighbors used for the construction of the graphs for each of the classifications is reported in Table. 1. This conforms to the number of neighbors that CNA uses\textsuperscript{31} for the prediction tasks.

The predictions of CEGAN (Fig.5 (a-d)) are almost identical as compared to CNA both in terms of the grain size distribution and the number of particles belonging to crystalline motifs of the gains. This clearly demonstrates the ability of the CEGAN in learning the different local motifs and distinguishing them from disorder atoms. The predictions of CEGAN on the local level classification tasks are largely dependent on the selection of the number of convolutional layers in the model as well as the number of neighbors used for the local neighborhood of the edge graphs. Adding more convolution layers will cause compression of too much information at a single node. This may result in a loss of resolution, which in turn will deteriorate the CEGAN performance. In Fig.6 (a), as we increase the number of convolutional layers for fixed 12 neighbors of graph construction, the performance severely declines at 4 edge-convolutional (+3 angle convolutions) layers. We have also studied the effect of the number of neighbors in the construction of the graph shown in Fig.6 (b). It seems that with an increase in the number of neighbors, CEGAN tends to slightly underpredict grain sizes. This is very intuitive given the fact that more number neighbors tend to include information from the stacking faults in the atoms of the grains resulting in an

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig6.png}
\caption{(a) Effect number of convolutional layer on the prediction of grain size distribution (b) Effect of number nearest neighbors for graph construction for grain boundary prediction.}
\end{figure}
underprediction of grains sizes. It is also worth mentioning that, unlike CNA, CEGAN is very flexible in learning environments with local noises such (as thermal) which is essential for practical applications.

**Dynamical classification of structures with thermal noise:**

Zeolites are ordered microporous silicates or aluminosilicate\textsuperscript{57,68} materials widely used as solid catalysts in the chemical industry. The knowledge about the mechanistic pathways of the formation of zeolite is still limited and is the key to realizing new zeolites for catalysis and separations. The stochastic nature of nucleation processes and the small, nanoscopic size of critical nuclei within the heterogeneous reaction mixture, make the detection of the birth of a new phase challenging in experimental hydrothermal synthesis. Molecular simulations have the right spatial resolution. However, in the synthesis mixture, the zeolite crystallites and the surrounding amorphous matrix have very similar local and medium-range orders\textsuperscript{69}. Fig. 7(a-b) show that, indeed, the zeolite and the network former silica in the amorphous phase have very similar radial and $Q^n$ (number of silica neighbors) distribution. Moreover, unlike the case of ice, where the unit cell consists of 1-2 atoms, the unit cell of zeolites typically has ~100 silica nodes. Even though each silicon has a coordination number of 4, the environment of each silicon node is diverse in the zeolite. They make it very difficult to identify the particles of the zeolite nucleus inside an amorphous matrix.

Traditional approaches such as the Bond-orientational order parameter, $q_6$, could be used to detect the nucleation process of zeolites but the requirement of the large cutoff distance makes it inefficient to detect the very small nucleus\textsuperscript{70–72}. moreover, the bond-order parameter approach is specific to a particular zeolite polymorph. Identification of crystal based on mobility criteria is also not zeolite specific, but it assumes that there is a considerable mobility difference between the new crystal phase from the mother phase\textsuperscript{73}. This approach does not work if the new phase crystallizes from a glassy state. These necessitate the development new classification technique that distinguishes the zeolite nucleus from the amorphous phases during the formation of zeolites.
Fig. 7 Performance of CEGAN for dynamical classification during a zeolite nucleation and growth process. (a-b) The radial distribution function between silica nodes ($g_{si-si}$) and the number of silica neighbors ($Q_n$) is very similar between the amorphous and zeolite phase. (c) The fraction of nucleating Z1 particles and its growth predicted by CEGAN. The snapshots of the simulation box corresponding to points i)-iv) are shown on the right-hand side. Waters are removed for clarity. Silica nodes of the amorphous phase are shown in purple, whereas the crystalline silica detected by CEGAN is shown with green sticks.

We use the CEGAN framework to probe the evolution of the zeolite nucleus and growth in the simulation mentioned above. To train our network we use a total of 400 structures consisting of 50% pure crystalline zeolites at different temperatures, noisy zeolite crystals (added Gaussian noise to the atomic positions) as well as 50% amorphous structures at different temperatures (See supplementary S.2 for details on data generation methods). We use 12 NN (nearest neighbors) for the graph construction and although the effects of 4 and 8 nearest neighbors on the construction of the graph are also explored (See supplementary Fig.S5 (c)). Fig. 7 (c) shows the zeolite fraction in the simulation trajectory as a function of time for the case of 12NN. A sharp change in the fraction of zeolite starting at time 16.5 ns suggest the formation of stable nuclei of zeolite Z1 that grow into a full slab at time > 25 ns. The same is evident from the snapshots presented at different instances during the crystallization (panel i-iv in Fig. 7(b)). This case study clearly illustrates that the proposed CEGAN workflow is not only capable of performing accurate classification in static local environments but also equally effective in heterogeneous simulation.
environments with considerable thermal noise. This is remarkable because our CEGAN workflow can identify a crystal nucleus smaller than its unit-cell size.

**Discussion**

Characterization of materials at different scales and domains of application is a must for any data-driven material science application. In this work, we develop graph attention-based CEGAN workflow which is not only transferrable across scales but adaptable to variabilities in the material environment and provides accurate characterization at the same time. We demonstrate the efficacy of our workflow by taking on challenging and yet relevant classification problems in material science. Unlike other Graph bases architecture (CGCNN) or mathematical formulation-based descriptors (SOAP, Order parameters), CEGAN is not only able to classify disordered (liquid and amorphous phases) at a global level but equally accurate in classifying local motifs in sticking disordered structures displaying transferability in the application domain. It is equally effective in performing global levels classification tasks such as space group classification and characterization of structures based on their dimensionality. These play crucial roles in a plethora of material science problems. CEGAN performs the classifications with near-perfect accuracy while its counterparts (SOAP, CGCNN, etc.) fail to do so. We then move to a local atomic level where CEGAN is employed to perform characterization of grain boundaries in polycrystalline materials and their grain distribution in terms of size. Our method managed to obtain identical accuracy in terms of grain size distribution to the traditional CNA method. Lastly, we perform a very challenging task of identifying of onset of nucleation and growth of zeolite polymorphs in synthesis solutions with strong thermal fluctuations. CEGAN not only manages to detect the onset of nucleation but captures the growth process very accurately. This showcases the applicability of CEGAN in problems involving variability in the environments. Overall, our approach is agnostic to the problem and allows the classification of features at different scales with equal efficacy.

**Methods:**

**Angle convolution:**

In the angle convolutional layer uses bond angle \((\theta_{ijk})\) cosines expanded on a gaussian basis as the initial input. The idea is that each angle learns and collects the messages from its adjacent edges through the convolutions. We use a simple graph attention-based architecture and convolutional operation is performed according to
\[ \theta_{ij}^{l+1} = \text{softplus} \left( \theta_{ij}^l + \alpha_{ijkl} \ast \left( W_{ijkl}^f (\theta_{ijkl} \oplus e_{ij}^l \oplus e_{jk}^l) + b_{ijkl}^f \right) \right), \quad (1) \]

where \( e_{ij}^l, e_{jk}^l \) are edge features from previous edge convolution layers and \( \alpha_{ijkl} \) is the attention coefficient calculated using:

\[ \alpha_{ijkl} = \text{softmax} \left( (W_{ijkl}^\text{att} (\theta_{ijkl} \oplus e_{ij}^l \oplus e_{jk}^l) + b_{ijkl}^\text{att}) \right), \quad (2) \]

where \( W_{ijkl}^f, W_{ijkl}^\text{att} \) and \( b_{ijkl}^f, b_{ijkl}^\text{att} \) are feature and attention weights and biases, respectively. We use softmax activation as a normalizer for calculating the attention coefficient and the final output of the angle convolution is passed through a softplus activation to obtain the final representation. Batch normalization is applied after the aggregation operation.

**Edge convolution:**

We follow a similar attention type mechanism for the edge convolutional layer. The convolutional function is represented as

\[ e_{ij}^{l+1} = \text{softplus} \left( e_{ij}^l + \sum_{k \in N} \text{softplus} \left( \alpha_{ijk} \ast \left( W_{ijk}^f (\theta_{ijkl} \oplus e_{ij}^l \oplus e_{jk}^l) + b_{ijkl}^f \right) \right) \right), \quad (3) \]

where \( W_{ijk}^f \) and \( b_{ijkl}^f \) are the weights and biases for the feature matrix and \( \theta_{ijkl}^l \) is the angle features from the previous angle convolutional stage. \( \alpha_{ijk} \), the attention coefficient computed using an equation analogous to Eq.2, with different weights and biases. We apply a nonlinear softplus activation function before and after the aggregation over the neighborhood; the additional nonlinearity helps the features to adapt to the target task.

**Feature aggregation and concatenation:**

The aggregation block (Fig. 2(c)) consists of 3 stages. First, the edge and angle features are aggregated as

\[ e_{il}^{l+1} = \sum_{j \in N} \text{softplus} (e_{ij}^l) \quad (4) \]

\[ \theta_{il}^{l+1} = \sum_{j \in N} \text{softplus} (\sum_{k \in N} \text{softplus} (\theta_{ijkl}^l)) \quad (5) \]
The final feature representation is given as concatenation $Z_i = e^{l+1} \oplus \theta^{l+1}$. To provide more resolution to the aggregated feature, we take a linear transformation before the aggregation stage. The pooling of the features follow the concatenation operation. It should be noted that the pooling on the features is applied only if a global level classification task is being performed. For local classification tasks, no pooling is applied to the features. Batch normalization is applied after the aggregation operation. We also apply dropouts’ (0.5 rates) before subsequent transformation after the convolutional layer. This helps in reducing overfitting. We use cross-entropy loss as the loss metric.\textsuperscript{75}

**Training the model:**

The network is trained on 1 GPU-accelerated to compute node on the NERSC computing cluster with 20-core Intel Xeon Gold 6148 (‘Skylake’) @ 2.40 GHz and 1 NVIDIA Tesla V100 (‘Volta’) GPU. The feature vector for the Angle convolution and edge convolutions are kept being 80. The hidden feature for the dense layer following the edge and angle convolution layers is 256. Upon aggregation, the overall dimension of the feature vector is 512.

**Code & Data Availability:**

The dataset used for space group classification is available from https://www.nomad-coe.eu.\textsuperscript{52} The Carbon bulk structures used in this work are available in https://www.sacada.info/. The CEGAN code along with the trained models and datasets are available in https://github.com/sbanik2/CEGAN.

**Acknowledgment:**

The authors acknowledge support by the U.S. Department of Energy through BES award DE-SC0021201. Use of the Center for Nanoscale Materials, an Office of Science user facility, was supported by the U. S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. This research also used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U. S. Department of Energy under contract DE-AC02-06CH11357. This research used resources of the National Energy Research Scientific Computing Center; a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We gratefully acknowledge the computing resources provided on Fusion and Blues; high performance computing clusters operated by the Laboratory Computing Resource Center (LCRC) at Argonne National Laboratory.
Author contributions

S.B. and S.K.R.S. conceived the project. S.B. developed the CEGAN workflow for multiscale classification with feedback from H.C. S.B evaluated the performance of workflow on different classification tasks and analyzed the results. D.D and V.M. generated the dataset and contributed to the wiring, and analysis of the dynamical classification task presented in the manuscript. All the authors contributed to data analysis and provided input on training data. SB wrote the manuscript with guidance from S.S and V.M. S.M, H.C, and D.D provided input on the writing and improvement of the manuscript. M.C. assisted with the computing resources and provided feedback on the workflow. S.K.R.S. supervised and directed the overall project.

References:

1. Zhang, H.-T. et al. Reconfigurable perovskite nickelate electronics for artificial intelligence. Science (1979) 375, 533–539 (2022).
2. Dwivedi, N. et al. Unusual High Hardness and Load-Dependent Mechanical Characteristics of Hydrogenated Carbon–Nitrogen Hybrid Films. ACS Applied Materials & Interfaces 14, 20220–20229 (2022).
3. Mansouri Tehrani, A. et al. Machine learning directed search for ultraincompressible, superhard materials. J Am Chem Soc 140, 9844–9853 (2018).
4. Schaibley, J. R. et al. Valleytronics in 2D materials. Nature Reviews Materials 1, 1–15 (2016).
5. Gogotsi, Y. & Anasori, B. The rise of MXenes. ACS nano vol. 13 8491–8494 Preprint at (2019).
6. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials 5, 1–36 (2019).
7. Liu, Y., Zhao, T., Ju, W. & Shi, S. Materials discovery and design using machine learning. Journal of Materiomics 3, 159–177 (2017).
8. Curtarolo, S. et al. AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations. Computational Materials Science 58, 227–235 (2012).
9. Jain, A. et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Mater 1, 011002 (2013).
10. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). Jom 65, 1501–1509 (2013).
11. Isayev, O. et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat Commun* **8**, 1–12 (2017).
12. Huo, H. & Rupp, M. Unified representation of molecules and crystals for machine learning. *arXiv preprint arXiv:1704.06439* (2017).
13. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J Chem Phys* **134**, 074106 (2011).
14. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics* **18**, 13754–13769 (2016).
15. Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chemical Reviews* **121**, 9759–9815 (2021).
16. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2**, 1–7 (2016).
17. 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).
18. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys Rev Lett* **108**, 058301 (2012).
19. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry* **115**, 1094–1101 (2015).
20. Hansen, K. et al. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J Phys Chem Lett* **6**, 2326–2331 (2015).
21. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys Rev Lett* **104**, 136403 (2010).
22. Seko, A., Takahashi, A. & Tanaka, I. Sparse representation for a potential energy surface. *Physical Review B* **90**, 024101 (2014).
23. Jiang, B. & Guo, H. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. *J Chem Phys* **139**, 054112 (2013).
24. Deringer, V. L. & Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Physical Review B* **95**, 094203 (2017).
25. Revard, B. C., Tipton, W. W. & Hennig, R. G. Genetic algorithm for structure and phase prediction. Preprint at (2018).
26. Oganov, A. R., Pickard, C. J., Zhu, Q. & Needs, R. J. Structure prediction drives materials discovery. *Nature Reviews Materials* **4**, 331–348 (2019).
27. Banik, S. et al. Learning with Delayed Rewards—A Case Study on Inverse Defect Design in 2D Materials. *ACS Applied Materials & Interfaces* **13**, 36455–36464 (2021).
28. Loeffler, T. D., Banik, S., Patra, T. K., Sternberg, M. & Sankaranarayanan, S. K. R. S. Reinforcement learning in discrete action space applied to inverse defect design. *Journal of Physics Communications* **5**, 031001 (2021).
29. Ryan, K., Lengyel, J. & Shatruk, M. Crystal structure prediction via deep learning. *J Am Chem Soc* **140**, 10158–10168 (2018).
30. Jäger, M. O. J., Morooka, E. v, Federici Canova, F., Himanen, L. & Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials* **4**, 1–8 (2018).
31. Stukowski, A. Structure identification methods for atomistic simulations of crystalline materials. *Modelling and Simulation in Materials Science and Engineering* **20**, 045021 (2012).
32. Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. Bond-orientational order in liquids and glasses. *Physical Review B* **28**, 784 (1983).
33. Ackland, G. J. & Jones, A. P. Applications of local crystal structure measures in experiment and simulation. *Physical Review B* **73**, 054104 (2006).
34. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Physical Review B* **87**, 184115 (2013).
35. Weyl, H. *The classical groups: their invariants and representations*. (Princeton university press, 1946).
36. Jensen, F. *Introduction to computational chemistry*. (John wiley & sons, 2017).
37. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE Trans Neural Netw* **20**, 61–80 (2008).
38. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).
39. Zhou, J. et al. Graph neural networks: A review of methods and applications. *AI Open* **1**, 57–81 (2020).
40. Velickovic, P. et al. Graph attention networks. *stat* **1050**, 20 (2017).
41. Klicpera, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123* (2020).
42. Duvenaud, D. K. et al. Convolutional networks on graphs for learning molecular fingerprints. *Adv Neural Inf Process Syst* **28**, (2015).
43. Fung, V., Zhang, J., Juarez, E. & Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *npj Computational Materials* **7**, 1–8 (2021).
44. Schmidt, J., Pettersson, L., Verdozzi, C., Botti, S. & Marques, M. A. L. Crystal graph attention networks for the prediction of stable materials. *Science Advances* **7**, eabi7948 (2021).

45. Choudhary, K. & DeCost, B. Atomistic Line Graph Neural Network for improved materials property predictions. *npj Computational Materials* **7**, 1–8 (2021).

46. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials* **31**, 3564–3572 (2019).

47. Park, C. W. & Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials* **4**, 063801 (2020).

48. 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).

49. Xie, T. & Grossman, J. C. Hierarchical visualization of materials space with graph convolutional neural networks. *J Chem Phys* **149**, 174111 (2018).

50. Nguyen, A. H. & Molinero, V. Identification of clathrate hydrates, hexagonal ice, cubic ice, and liquid water in simulations: The CHILL+ algorithm. *The Journal of Physical Chemistry B* **119**, 9369–9376 (2015).

51. Schmidt, J., Pettersson, L., Verdozzi, C., Botti, S. & Marques, M. A. L. Crystal graph attention networks for the prediction of stable materials. *Science Advances* **7**, eabi7948 (2021).

52. Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat Commun* **9**, 1–10 (2018).

53. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68**, 314–319 (2013).

54. Leitherer, A., Ziletti, A. & Ghiringhelli, L. M. Robust recognition and exploratory analysis of crystal structures via Bayesian deep learning. *Nat Commun* **12**, 1–13 (2021).

55. Hoffmann, R., Kabanov, A. A., Golov, A. A. & Proserpio, D. M. Homo citans and carbon allotropes: For an ethics of citation. *Angewandte Chemie International Edition* **55**, 10962–10976 (2016).

56. Sharma, B. R., Manjanath, A. & Singh, A. K. pentahexoctite: A new two-dimensional allotrope of carbon. *Scientific Reports* **4**, 1–6 (2014).

57. Zhang, H.-T. *et al.* Reconfigurable perovskite nickelate electronics for artificial intelligence. *Science (1979)* **375**, 533–539 (2022).

58. Los, J. H. & Fasolino, A. Intrinsic long-range bond-order potential for carbon: Performance in Monte Carlo simulations of graphitization. *Physical Review B* **68**, 024107 (2003).

59. Manna, S. *et al.* A Database of Low-Energy Atomically Precise Nanoclusters. (2021). DOI: 10.26434/chemrxiv-2021-0fq3q.
60. Manna, S. et al. Learning in continuous action space for developing high dimensional potential energy models. *Nat Commun* **13**, 1–10 (2022).
61. Novoselov, K. S. et al. Two-dimensional atomic crystals. *Proceedings of the National Academy of Sciences* **102**, 10451–10453 (2005).
62. Oganov, A. R., Pickard, C. J., Zhu, Q. & Needs, R. J. Structure prediction drives materials discovery. *Nature Reviews Materials* **4**, 331–348 (2019).
63. Hirel, P. Atomsk: A tool for manipulating and converting atomic data files. *Computer Physics Communications* **197**, 212–219 (2015).
64. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)* **42**, 1–21 (2017).
65. Chan, H., Cherukara, M., Loeffler, T. D., Narayanan, B. & Sankaranarayanan, S. K. R. S. Machine learning enabled autonomous microstructural characterization in 3D samples. *npj Computational Materials* **6**, 1–9 (2020).
66. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool. *Model Simul Mat Sci Eng* **18**, 015012 (2009).
67. Chen, Y., Manna, S., Ciobanu, C. v & Reimanis, I. E. Thermal regimes of Li-ion conductivity in β-eucryptite. *Journal of the American Ceramic Society* **101**, 347–355 (2018).
68. Chen, Y. et al. Pressure-induced phase transformation in β-eucryptite: An X-ray diffraction and density functional theory study. *Scripta Materialia* **122**, 64–67 (2016).
69. Bertolazzo, A. A., Dhabal, D. & Molinero, V. Polymorph Selection in Zeolite Synthesis Occurs after Nucleation. *The Journal of Physical Chemistry Letters* **13**, 977–981 (2022).
70. Bertolazzo, A. A., Dhabal, D., Lopes, L. J. S., Walker, S. K. & Molinero, V. Unstable and Metastable Mesophases Can Assist in the Nucleation of Porous Crystals. *The Journal of Physical Chemistry C* **126**, 3776–3786 (2022).
71. Kumar, A. & Molinero, V. Two-step to one-step nucleation of a zeolite through a metastable gyroid mesophase. *The Journal of Physical Chemistry Letters* **9**, 5692–5697 (2018).
72. Kumar, A., Nguyen, A. H., Okumu, R., Shepherd, T. D. & Molinero, V. Could mesophases play a role in the nucleation and polymorph selection of zeolites? *J Am Chem Soc* **140**, 16071–16086 (2018).
73. Dhabal, D., Bertolazzo, A. A. & Molinero, V. Coarse-Grained Model for the Hydrothermal Synthesis of Zeolites. *The Journal of Physical Chemistry C* **125**, 26857–26868 (2021).
74. Velickovic, P. et al. Graph attention networks. *stat* **1050**, 20 (2017).
Supplementary Information

CEGAN: Crystal Edge Graph Attention Network for Multiscale Classification of Materials Environment

Suvo Banik\textsuperscript{1,2}, Debdas Dhabal\textsuperscript{3}, Henry Chan\textsuperscript{1}, Sukriti Manna\textsuperscript{1,2}, Mathew Cherukara\textsuperscript{4}, Valeria Molinero\textsuperscript{3}, Subramanian KRS Sankaranarayanan\textsuperscript{1,2}

\textsuperscript{1} Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, United States.
\textsuperscript{2} Department of Mechanical and Industrial Engineering, University of Illinois, Chicago, Illinois 60607, The United States.
\textsuperscript{3} Department of Chemistry University of Utah Salt Lake City, UT 84112, United States.
\textsuperscript{4} Advanced Photon Source, Argonne National Laboratory, Lemont, Illinois 60439, United States.

S.1. Generation of amorphous and liquid silicon structures:

To generate the liquid and amorphous (glassy) structure, pure cubic diamond silicon is melted at 6000 K using an NPT ensemble with pressure ranges 0.1-4 GPa, which gives the liquid dataset. The atomistic simulations are carried out using a Stillinger-Weber\textsuperscript{1} interatomic potential with a LAMMPS\textsuperscript{2} package. To create glassy amorphous structures, we quench the liquid structures at temperatures within 50-350 K (below the glass transition of silicon) in an NVT ensemble. The typical amorphous structure has a sharp radial distribution function (rdf) peak at \(~2.4\ \AA\) and a secondary peak at \(~4\ \AA\) as compared to the liquid, where the distribution is flat beyond the first peak (Fig. S1 (a)). Despite the difference in the rdfs, they both take similar volume per atom (Fig. S1 (b)).
Fig.S1. Classification of liquid and amorphous Si structures modeled using Stillinger Weber (SW) potential. (a) Comparison of radial distribution functions between liquid and amorphous structures. (b) Volume per atom of liquid and amorphous silicon. PCA plot of SOAP representation of the crystalline, liquid and amorphous phases for a cutoff of (c) 3 Å and (d) 6 Å on the overall dataset.
Fig.S2 (a) The data points in each class in the dataset used for space group classification. (b-i) The ROC-AUC curves as obtained on the validation datasets results from the CEGAN workflow. (j-k) Typical training and validation accuracy during the training of the CEGAN for the classification task performed.
Fig. S3 Performance of CEGAN in dimensionality classification. (a) Shows the data points in each class in the dataset used for dimensionality classification. (b) Size distribution of 2d polymorphs (c) Size distribution of clusters (d) Size distribution of bulk structures used in this study. (e) Cross entropy loss during training on the training set (I did not understand) (f) Accuracy score on the validation dataset during training.
Fig. S4 (a-d) percentage crystalline and stacking fault particles in training data for the Al (FCC), W (BCC), Si (diamond), and Mg (HCP) system. To me there is no need to write “(a) shows or (b) shows” in the figure caption. That is written when you explain about a figure in main text.

S.2. Generation of data for probing nucleation and growth in zeolites:

The prediction data used for the identification of Z1 zeolite in this study has the structure of the FIR-30 metal-organic framework. We select a molecular dynamics simulation trajectory at 670 K that depicts the nucleation and growth process simulated using a coarse-grained model. The overall training data contains pure zeolite Z1 and the amorphous silica phase. We use both energy minimized zeolite structure and equilibrated framework at 670 K-720 K (I am not sure whether we equilibrated Z1 at 720 K? it should be...
only 670 K, I think) for the training purpose. Similar is the case for the amorphous phase too, except that is equilibrated at (Amorphous is equilibrated at 720 K only, I don’t know if you used any other data which are equilibrated at 607 K). We observe a significant fluctuation in the high-temperature zeolite structure even though the microscopic structural environments are the same between the low and high-temperature states (Fig. S5a-b). To account for the noises in the nucleated crystal, we consider pure zeolites with Gaussian noise added to their atomic positions. Overall, the training data contains a diverse set of crystal and amorphous phases that helps the network to learn beyond thermal fluctuations and identify critical features of nucleation and growth.

Fig.S5 (a) The rdfs of the pure energy minimized zeolite crystal and a zeolite at a temperature of 670 K used for the prediction. (b) The coordination number distribution, $Q^n$ of the same. (c) The fraction of nucleating ZI particles and its growth predicted by CEGAN for a different number of neighbors.
References:

1. Stillinger FH, Weber TA. Computer simulation of local order in condensed phases of silicon. *Physical review B*. 1985;31(8):5262.

2. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. *J Comput Phys*. 1995;117(1):1-19.

3. Wang F, Fu HR, Zhang J. Homochiral metal–organic framework with intrinsic chiral topology and helical channels. *Crystal Growth & Design*. 2015;15(4):1568-1571.

4. Dhabal D, Bertolazzo AA, Molinero V. Coarse-Grained Model for the Hydrothermal Synthesis of Zeolites. *The Journal of Physical Chemistry C*. 2021;125(48):26857-26868.
