# Machine learning classification of local environments in molecular crystals

Daisuke Kuroshima,[*,†] Michael Kilgour,[*,†] Mark E. Tuckerman,[*,†,‡,¶,§] and Jutta Rogal[*,†,‖]

†*Department of Chemistry, New York University (NYU), New York, New York 10003, USA.*

‡*Courant Institute of Mathematical Sciences, New York University, New York, New York 10012, USA.*

¶*NYU-ECNU Center for Computational Chemistry at NYU Shanghai, 3663 Zhongshan Rd. North, Shanghai 200062, China.*

§*Simons Center for Computational Physical Chemistry at New York University, New York, New York 10003, USA.*

‖*Fachbereich Physik, Freie Universität Berlin, 14195 Berlin, Germany.*

E-mail: daisuke.kuroshima@nyu.edu; michael.kilgour@nyu.edu; mark.tuckerman@nyu.edu; jutta.rogal@nyu.edu

**Abstract**

Identifying local structural motifs and packing patterns of molecular solids is a challenging task for both simulation and experiment. We demonstrate two novel approaches to characterize local environments in different polymorphs of molecular crystals using learning models that employ either flexibly learned or handcrafted molecular representations. In the first case, we follow our earlier work on graph learning in molecular

crystals, deploying an atomistic graph convolutional network, combined with molecule-wise aggregation, to enable per-molecule environmental classification. For the second model, we develop a new set of descriptors based on symmetry functions combined with a point-vector representation of the molecules, encoding information about the positions as well as relative orientations of the molecule. We demonstrate very high classification accuracy for both approaches on urea and nicotinamide crystal polymorphs, and practical applications to the analysis of dynamical trajectory data for nanocrystals and solid-solid interfaces. Both architectures are applicable to a wide range of molecules and diverse topologies, providing an essential step in the exploration of complex condensed matter phenomena.

# 1 Introduction

Elucidation of the microscopic structure of molecular materials is key to predicting and engineering their properties. Despite significant advances in experimental techniques, following structural transformations in condensed-phase systems with atomistic resolution remains a challenge due to the time- and length-scales involved. Computational approaches, such as molecular dynamics (MD) simulations, have become an invaluable tool to provide such microscopic insight, but characterizing the structural features of a molecular system from the simulation data is, in general, nontrivial. However, following the dynamical evolution of local structural environments is essential when studying polymorphic transitions, especially regarding the complex atomistic processes that govern nucleation and growth.

A number of descriptors have been developed over the years to capture local or global structural features, including Steinhardt order parameters,[1,2] common neighbor analysis,[3–5] entropy based fingerprints,[6] smooth overlap of atomic positions,[7] and atom-centred symmetry functions[8] (see also[9–16] for further overviews and examples). More recently, machine learning has been utilized to classify local environments with both supervised and unsupervised approaches.[17–28] These machine learning models for local structure classification fall

into two broad categories: models that use handcrafted structural features or descriptors together with a simple classification model, and models that use only very general information, such as atom types and distances, and letting the model learn the structural representation and intermolecular correlations simultaneously. The former approach is attractive in its ostensible simplicity but relies on the development of high-quality descriptors, while the latter requires a more complex model architecture but less intuition about the system and is more generally applicable. Here, graph neural-network (GNN) approaches are attractive in their generality, allowing one to use a single flexible model for most systems. GNNs have also been used to describe condensed-phase systems, wherein the relevant features are learned in a 'ground up' fashion from basic atomistic information.[26,27,29–36]

The structure characterization methods discussed above have primarily been established in the context of atomistic condensed-matter systems. In molecular systems, additional challenges arise since not only the positions of the molecules but also their relative orientation as well as conformational changes need to be accounted for. One idea is to include this information via a point-vector representation of the molecules where, for example, the center of mass denotes the molecule position, and vectors denote the absolute orientation of a given molecule, which can then be combined into suitable descriptors.[37,38]

In this work, we advance the state of the art of machine learning classification of local environments to capture the complex structural features in molecular solids. We present two parallel approaches, one based on handcrafted descriptors and the other on learned feature embeddings. The handcrafted descriptors extend our previous work on atomistic systems[19] to molecular symmetry functions (SF) by combining the SFs with a point-vector representation of the molecules. For the learned embeddings, we utilize our recently introduced molecular crystal graph model MXtalNet[39] and augment the architecture with a classification task. The trained models are able to distinguish different local environments in various polymorphs of complex molecular solids with high accuracy. Furthermore, both approaches are applicable to a wide range of systems, including clusters and interfaces, and can pro-

vide time-resolved information regarding melting transitions or solid-solid transformations. The potential of our classification models is exemplified for urea and nicotinamide but the methods are easily extended to arbitrary molecules. The approaches presented introduce an essential and valuable component in the analysis and interpretation of simulation data for molecular solids.

## 2 Model architectures and training

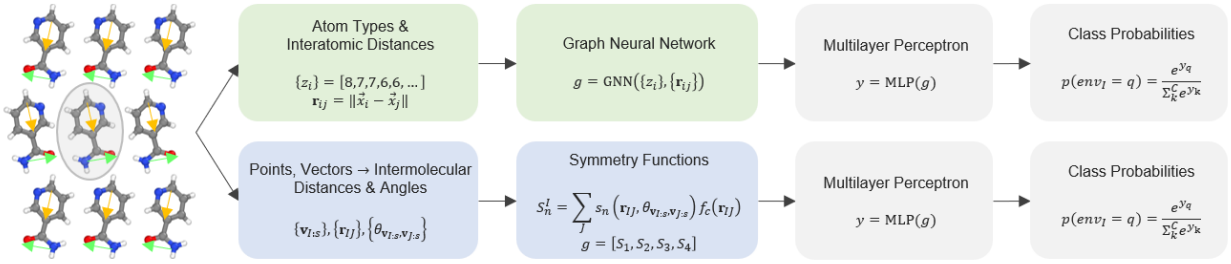


Figure 1: The workflow of the GNN and SF classifiers on top and bottom, respectively, including molecule representation, local embedding, and classification. The GNN learns the features $g$ used in the classification task, while for the SF classifier the features $g$ are given by the handcrafted molecular SFs.

The general idea of our two model architectures is schematically illustrated in Fig. 1. The classification is performed for each molecule to characterize its local structural environment. An appropriate model should be invariant to permutations of atoms of the same types, as well as global translations, rotations, and inversions of the atomic coordinates, focusing only on the structural correlations which define the respective polymorphs. For the learned feature embedding, the positions and atom types of a given molecule and its neighbors comprise the input to a graph neural network (GNN) coupled with a multilayer perceptron (MLP) to perform classification on the final embedding. For the handcrafted features, the atomic positions are used to construct a point-vector representation for each molecule which is then employed to compute a set of molecular symmetry functions as input to the classification MLP. Details of the model architectures and training protocols are given in the following.

## 2.1 Molecular crystal graph network

For the molecular GNN, we used a relatively straightforward graph neural network, similar in geometric complexity to SchNet,[40] taking interatomic distances and atom types as inputs. This GNN encodes pairwise interatomic distances to edge embeddings, atom types to node embeddings, and performs graph convolutions via the TransformerConv operator[41] implemented in the Torch Geometric package.[42]

The GNN parses a single sample in the following way, starting with embedding of the input nodes atom types $z_i$,

$$\mathbf{f}_i^0 = \mathrm{EMB}(z_i) \quad , \tag{1}$$

with EMB as a learnable discrete embedding function, followed by a fully-connected layer. The edge embedding is

$$\mathbf{e}_{ij} = \mathrm{Bessel}(|\mathbf{r}_{ij}|) \quad , \tag{2}$$

where $|\mathbf{r}_{ij}|$ is the distance between nodes $i$ and $j$, and Bessel is the radial embedding function from DimeNet[43] with cutoff $r_c = 6$ Å and 32 radial Bessel basis functions. A fully connected layer is defined as $\mathrm{FC}(\mathbf{x}) = \mathbf{W} \cdot \mathbf{x} + \mathbf{b}$, with $\mathbf{W}$ and $\mathbf{b}$ as learnable parameters. Messages are passed between nodes, conditioned on node and edge embeddings via Eqs. (3) for edge→message and (4) for node→message, over $N$ graph convolutions, with GC the graph convolution operation,

$$\mathbf{E}_{ij}^t = \mathbf{W}_{e\to m}^t \cdot \mathbf{e}_{ij} \quad , \tag{3}$$

$$\mathbf{F}_i^t = \mathbf{W}_{n\to m}^t \cdot \mathbf{f}_i^t \quad , \tag{4}$$

$$\mathbf{f}_i^{t+1} = \mathbf{W}_{m\to n}^t \cdot (\mathrm{GC}(\mathbf{F}_i^t, \mathbf{F}_j^t, \mathbf{E}_{ij}^t)) \quad . \tag{5}$$

After each graph convolution, the node embeddings are passed through a fully-connected layer with residual connection,

$$\mathbf{f}_i^t = \mathbf{f}_i^t + \sigma\left(D\left(\mathcal{N}(\mathrm{FC}_{n\to n}^t(\mathbf{f}_i^t))\right)\right) \quad , \tag{6}$$

with $\sigma$ being the activation function (here GeLU[44]), $D(x)$, a dropout function, and $\mathcal{N}(x)$, the graph layer norm operation. The final node features, corresponding to information about each atom and its local environment, are aggregated into a single embedding vector representing the entire molecule, and input to a two-layer activated fully-connected network with layer normalization and dropout, followed by a reshaping to the number of possible classes. Though there are currently many powerful graph aggregators, we find max aggregation, that is, selecting the maximum value from each feature channel, $k$, across the final atomic node embeddings in each molecule, is simple and efficient for learning the desired functions, with

$$\mathbf{g} = \mathrm{MAX}_k(\{\mathbf{f}^N\}) \tag{7}$$

and

$$\mathbf{y} = \mathrm{MLP}(\mathbf{g}) \quad , \tag{8}$$

with MLP a multilayer perceptron. The class probabilities for a molecule $I$ being in a particular environment $q$ are computed via the softmax activation function

$$p(\mathrm{env}_I = q) = \frac{\exp(y_q)}{\sum_k^C \exp(y_k)} \quad , \tag{9}$$

with $C$ the number of possible environments.

We found one or two graph convolutions gave similar performance, though more convolutions result in a larger volume for what the model considers as a 'local environment'. The number of convolutions depends on the user's desired sensitivity to longer-range structural correlations, but in the current examples, more than two convolutions resulted in training instability and overall poor convergence. For other hyperparameters, optimal performance was obtained with a relatively deep embedding (256 for node and graph embeddings, 128 for messages), aggressively regularized with layer norm and dropout of 0.25 in graph convolutions, nodewise fully-connected layers, and the embedding-to-output network. With

these settings, the model converged via the Adam optimizer to a the test minimum very quickly, usually within a few tens of epochs. Smaller models could certainly be explored, although we generally found convergence properties to be poorer in that regime. For further details of model construction, see the supplementary information (SI) and our accompanying codebase.[45]

## 2.2 Molecular symmetry functions

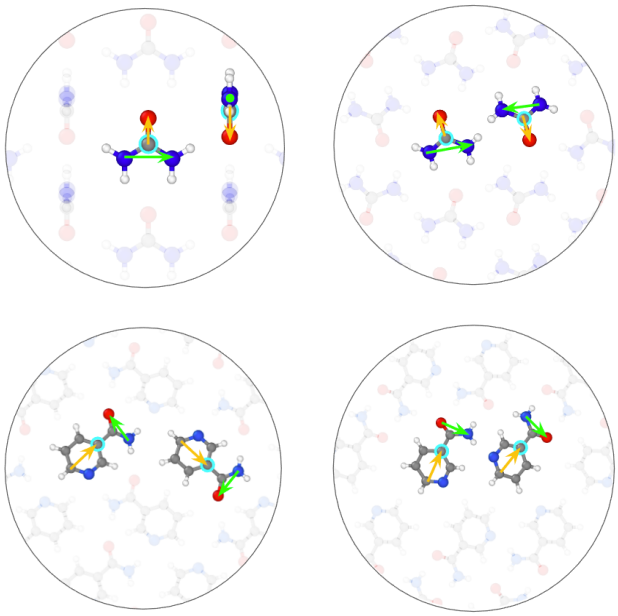

Figure 2: Point-vector representation for urea (top panels) and nicotinamide (bottom panels) in two different polymorphs, respectively. The turquoise circles indicate the positions of the molecules $\mathbf{r}_I$, and the green and orange vectors, $\mathbf{v}_{I;1}$ and $\mathbf{v}_{I;2}$, characterize their relative orientations.

Our second model derives a set of descriptors for each molecule based on the Behler-Parrinello symmetry functions[8] in combination with a point-vector representation[37,38] of the molecules. The point-vector representations for urea and nicotinamide are illustrated in Fig. 2, where the position $\mathbf{r}_I$ of molecule $I$ is represented by a selected atom (indicated by a turquoise circle in Fig. 2). Vectors $\mathbf{v}_{I;s}$ are defined between two selected atoms in the molecule, such that they can capture relative orientations of the molecules (indicated in orange, $\mathbf{v}_{I;1}$, and green, $\mathbf{v}_{I;2}$, in Fig. 2). We utilize four different types of molecular symmetry

functions $S^I$. Two are akin to radial symmetry functions for atomistic systems but using the molecule positions $\mathbf{r}_I$,

$$S_1^I(\mathbf{r}) = \sum_J e^{-\eta\ (|\mathbf{r}_{IJ}|-R_s)^2} f_c(\mathbf{r}_{IJ}) \quad , \tag{10}$$

and

$$S_2^I(\mathbf{r}) = \sum_J \cos(\kappa|\mathbf{r}_{IJ}|) f_c(\mathbf{r}_{IJ}) \quad , \tag{11}$$

where the sum runs over all other molecules, $\mathbf{r}_{IJ} = \mathbf{r}_J - \mathbf{r}_I$, $f_c$ is a cutoff function (see SI for details), and $\eta$, $R_s$, and $\kappa$ are tunable parameters. The other two types of molecular symmetry functions use the molecule vectors to characterize the relative orientation of molecule $I$ with respect to its neighbours $J$,

$$S_3^I(\mathbf{r}, \mathbf{v}_{;s}) = \sum_J \exp\left(-\eta\ (\cos\theta_{\mathbf{v}_{I;s}\mathbf{v}_{J;s}} - \cos\theta_S)^2\right) f_c(\mathbf{r}_{IJ}) \quad , \tag{12}$$

and

$$S_4^I(\mathbf{r}, \mathbf{v}_{;s}) = \sum_J \cos\left(\kappa \cos\theta_{\mathbf{v}_{I;s}\mathbf{v}_{J;s}}\right) f_c(\mathbf{r}_{IJ}), \tag{13}$$

where $\theta_{\mathbf{v}_{I;s}\mathbf{v}_{J;s}}$ is the angle between vectors $\mathbf{v}_{;s}$ on molecules $I$ and $J$, and $\cos\theta_S$ is another tunable parameter. The total number of molecular symmetry functions is 24 for both urea and nicotinamide. Details regarding the selected molecular symmetry functions and corresponding values of the tunable parameters are given in the SI.

To perform classification of molecule environments, the molecular symmetry function descriptors are input to a rather small MLP with two hidden layers, 25 nodes each, and the softmax activation in Eq. (9) for the output layer. A larger MLP with more hidden layers and nodes would provide greater flexibility but due to the simplicity of the classification task, a small network was sufficient for our applications, making both the training and evaluation rather fast. For further implementation details, see the SF classifier codebase.[46]

## 2.3 Training the models

Training data were generated by molecular dynamics (MD) simulations of all crystal polymorphs and the melt for urea and nicotinamide. Simulations were performed using the LAMMPS MD package[47] with a general Amber force field (GAFF).[48] We briefly summarize here the protocol for training the classification models. Further details regarding the MD simulations and training are given in the SI.

The graph classifier was trained on a mix of trajectory snapshots of periodic bulk cells approximately $20 \times 20 \times 20$ Å$^3$ and gas phase spherical clusters with a diameter of $\sim 30$ Å to give the effect of a 'surface'. Molecules are identified as being on the surface if their local coordination number, $\mathrm{CN}_I$, is smaller than 20, with $\mathrm{CN}_I = \sum_J \theta(-(|\mathbf{r}_{IJ}| - R_C))$, where $\theta$ is the Heaviside function and $R_c$ the molecule radius plus the graph convolution cutoff. The symmetry function classifier was trained on periodic bulk samples only.

We train the classification models on stable, low-temperature snapshots of the known polymorphs of each molecular crystal, as well as a higher temperature melt state. We test the models' generalization performance on configurations from higher temperature MD simulations, with adaptation to thermal noise standing as a proxy for overall generalization. The specific temperatures for each of the studied systems are discussed together with the results below.

The graph classifier was trained until the test loss began to increase, and the model checkpoint at test loss minimum was used for evaluation. Repeated retraining over several random seeds found variation in test loss minimum of only a few percent between runs. We used a combined cross entropy loss including both the loss for the local polymorph classification for each molecule and the molecule topology, that is 'surface' vs 'bulk'.

The symmetry function classifier was trained until the training loss converged which, generally, resulted in very small test losses.

# 3 Classification of local environments

## 3.1 Bulk polymorphs of urea and nicotinamide

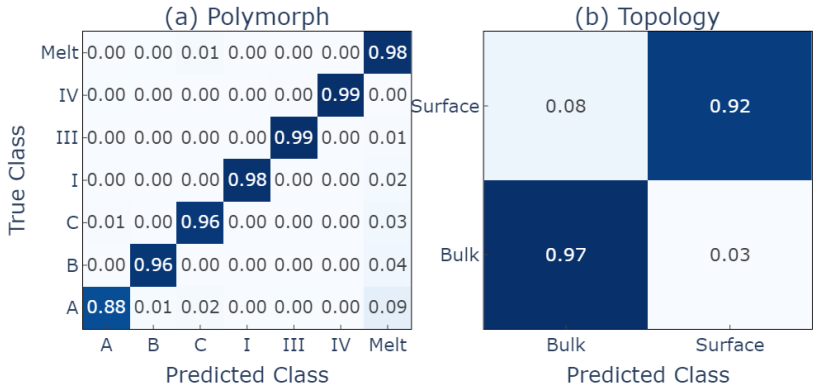


Figure 3: Confusion matrix for the graph classifier on (a) the polymorphs and (b) topologies of urea at 200K for crystals and 350K for melt. Micro F1 scores=0.969, 0.960.

We initially trained and applied our classification models to two different systems, urea and nicotinamide. Urea is a relatively small and rigid molecule, which is also significantly polymorphic, having six distinct crystal structures with unique intermolecular packing character[49–52] (see the SI for a visualization of the respective polymorphs). The models were trained on $T = 100$ K crystal samples and $T = 350$ K melts, and evaluation metrics were computed on samples at 200 K for the crystal polymorphs and 350 K for the melt. At low temperatures, the graph classifier achieves perfect accuracy for both polymorphs and local topologies. This means that the GNN learns an embedding where the different molecule environments are clearly separated without overlap. This is expected as the graph model is rather expressive and in all the thousands of individual molecular environments, the local structure seen by the model should be quite similar within each polymorph. The graph model also generalizes well to higher temperature samples at $T = 200$ K, as evidenced by the confusion matrix shown in Fig. 3, meaning that larger thermal fluctuations can be captured within the trained model. Only urea form A shows a slighter larger classification error, with about 9% of the samples being identified as 'melt', which might be due to the lower stability

of form A. The symmetry function classifier also demonstrated excellent performance on urea, achieving comparable or better performance at polymorph classification ($F1 \gtrsim 0.98$) to the GNN model in training and evaluation while being lightweight and fast to run at inference. The corresponding confusion matrix can be found in the SI.

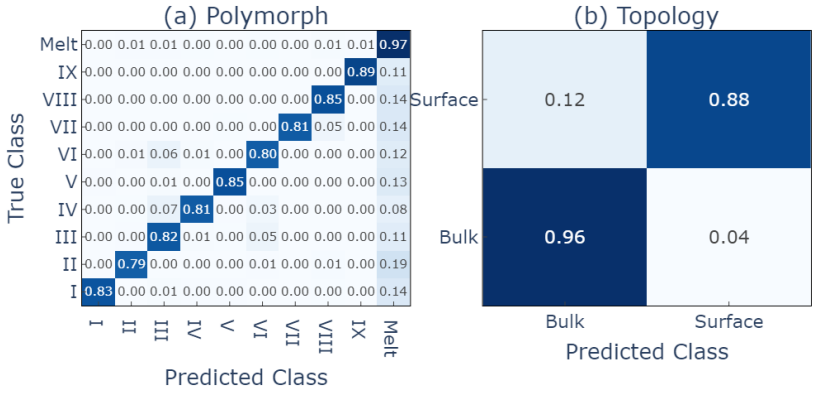


Figure 4: Confusion matrix for the graph classifier on (a) the polymorphs and (b) topologies of nicotinamide at 350 K. Micro F1 scores=0.875, 0.922.

As a second example, we chose nicotinamide as a more challenging molecule. Nicotinamide is larger than urea and more flexible with internal degrees of freedom that allow for polymorphs consisting of different conformers of the molecule. Nine polymorphs of nicotinamide have been experimentally crystallized[53,54](see the SI for a visualization of the respective polymorphs). Despite this significant added complexity in the molecular system, the performance of our classification models is again very good. As with urea, the training samples, both crystal polymorphs at 100 K and melts at 350 K, are essentially perfectly learned, and the model generalizes well to the high temperature test samples at 350 K. The corresponding confusion matrix for the GNN classifier is shown in Fig. 4. The F1 score for nicotinamide at high temperatures is slightly worse than for urea, 0.875 compared to 0.969, which reflects the increased flexibility in the thermal fluctuations at this even higher temperature. This is, however, not a fundamental limit of the model, as, when retrained across the full range of temperatures, the accuracy again approaches 100%.

We see that the generality and high capacity of the GNN model allow it to classify

each polymorph and local topology, without the need for model customization of any kind. Likewise, the symmetry function classifier performs excellently on the nicotinamide polymorphs (see the SI for the corresponding confusion matrices). This indicates that the chosen set of molecular symmetry function provides suitable descriptors to capture the additional complexity and flexibility in nicotinamide crystal polymorphs and melt.

One interesting point is that the GNN classifier exhibits a somewhat lower performance on the nicotinamide high temperature samples compared to the SF classifier, when both are trained on low temperature crystals and high temperature melts only. From the confusion matrix in Fig. 4 it becomes clear that the accuracy loss of the graph classifier is primarily due to over-prediction of the melt state. For a model trained only at 100 K and evaluated at 350 K, this should perhaps not be surprising. The larger thermal fluctuations in inter- and intramolecular degrees of freedom increase the general similarity of bulk crystals to the melt, and they are interpreted as such by the model. That we do not see this effect as strongly in the SF classifier results indicates that the handcrafted descriptors are quite robust to fluctuations, yet sensitive enough to achieve high classification accuracy.

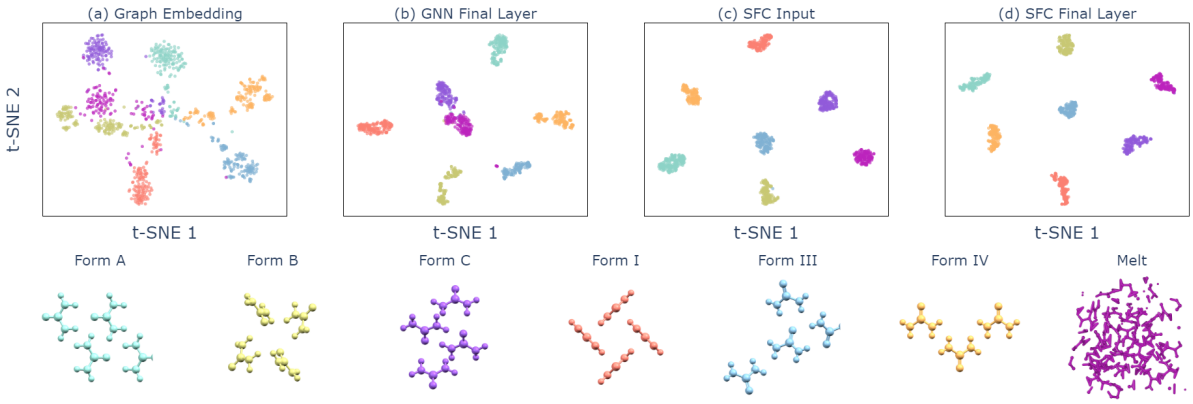


Figure 5: The t-distributed stochastic neighbor embedding (t-SNE) of urea samples from (a) the 256-dimensional graph embedding (output of (7)), (b) 256-dimensional final layer activation, (c) 24 symmetry functions, and (d) 25-dimensional SFC final layer activation; samples are taken from three different temperatures of 100 K, 200 K, and 350 K.

To get a better understanding of the learned and handcrafted features in our molec-

ular graph and symmetry function classifiers, respectively, we compare the corresponding embedding spaces. In Fig. 5, the embedding spaces of the representations and final layer activations for urea are visualized using the t-distributed stochastic neighbor embedding (t-SNE).[55] Fig.5(a) shows that the molecular representation learned by the GNN already separates the different polymorphs of urea reasonably well. The quality of the handcrafted symmetry functions is obvious when examining the t-SNE of the symmetry function inputs in Fig. 5(c), which cluster essentially perfectly before applying any learned transformations. Figs.5(b) and (d) show the t-SNE of the final layer activations for the GNN and symmetry function classifier, respectively. The class separation is excellent, as expected from the very high classification accuracy observed for both models.

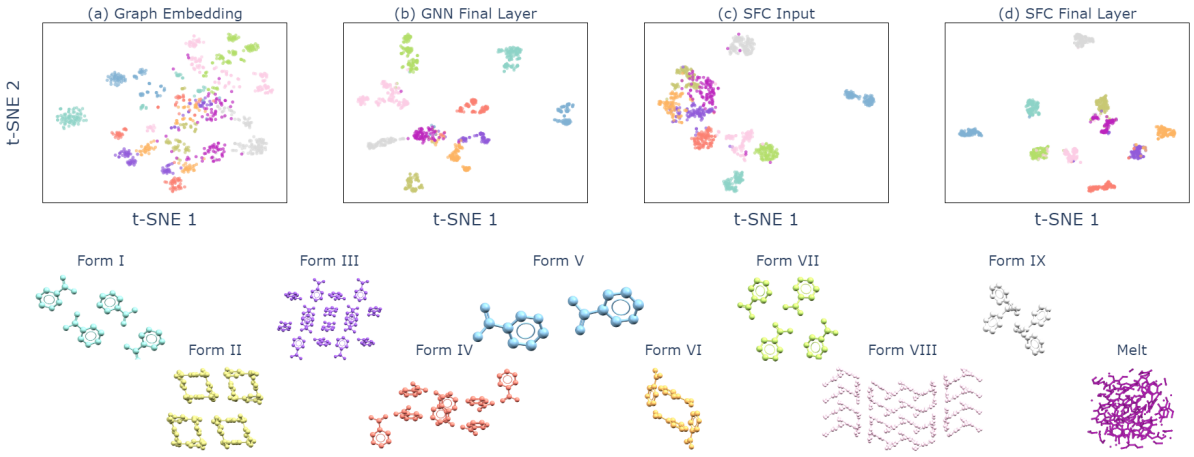


Figure 6: t-SNE of nicotinamide samples from (a) the graph embedding (output of 7), (b) final layer activation, (c) symmetry functions, and (d) SFC final layer activation, at temperatures of 100 K and 350 K. Embedding dimensionality is the same as in Figure 5.

The t-SNE visualization of the embedding spaces for nicotinamide are shown in Fig. 6. Both the learned and handcrafted embedding spaces in Figs. 6(a) and (c) show imperfect classwise separation between the various polymorphs in nicotinamide. This again underscores the increased challenge in characterizing structural environments in more complex and flexible systems. In particular, samples from the melt seem to cover a wide range and are less clustered in the embedding spaces. We also see greater separation of samples from

the same crystalline polymorphs in Figs. 6(a)-(b), including bifurcation of some classes, corresponding to the different sampled temperatures and topologies. The overlap between the melt and crystal embeddings visible in Figs. 6(a)-(b) is also consistent with the GNN classifier confusing some crystalline polymorphs mainly with the melt, as seen in Fig. 4. Nevertheless, the final learned representations in Figs. 6(b) and (d) show again a very good separation between the different polymorph classes, even for the high-temperature samples.

## 3.2 Analyzing molecular simulations

Being able to reliably characterize local environments in unknown structures will be particularly useful when analyzing and interpreting trajectory data from molecular simulations. In the following, we discuss two examples: the stability of gas phase nanocrystals at different temperatures and the migration of an interface during a solid-solid transformation in a molecular crystal.

### 3.2.1 Dynamical structure characterization of molecular clusters

The GNN classifier trained on the bulk polymorphs of nicotinamide is used to identify the local environments of nicotinamide molecules in small nanocrystals. We set up spherical clusters of nicotinamide form I with a diameter of 34 Å containing 148 molecules. Molecular dynamics simulations for the clusters in vacuum are run at $T = 100$ K and 350 K (further simulation details are given in the SI). In Fig. 7, the structural evolution of the nicotinamide nanoclusters at these two temperatures is shown, obtained using the graph classifier. Since the classifier provides information for each molecule individually, we can separate our analysis for molecules that are in the core region of the clusters, Fig. 7(a) and (c), and at their surfaces, Fig. 7(b) and (d). At 100 K, the nanocluster clearly keeps its crystalline structure over the entire simulation time. While the majority of molecules in the core region are identified as nicotinamide form I, molecules at the surface are partially classified as melt or others, which is expected since the structural environment at the surface is significantly different from the

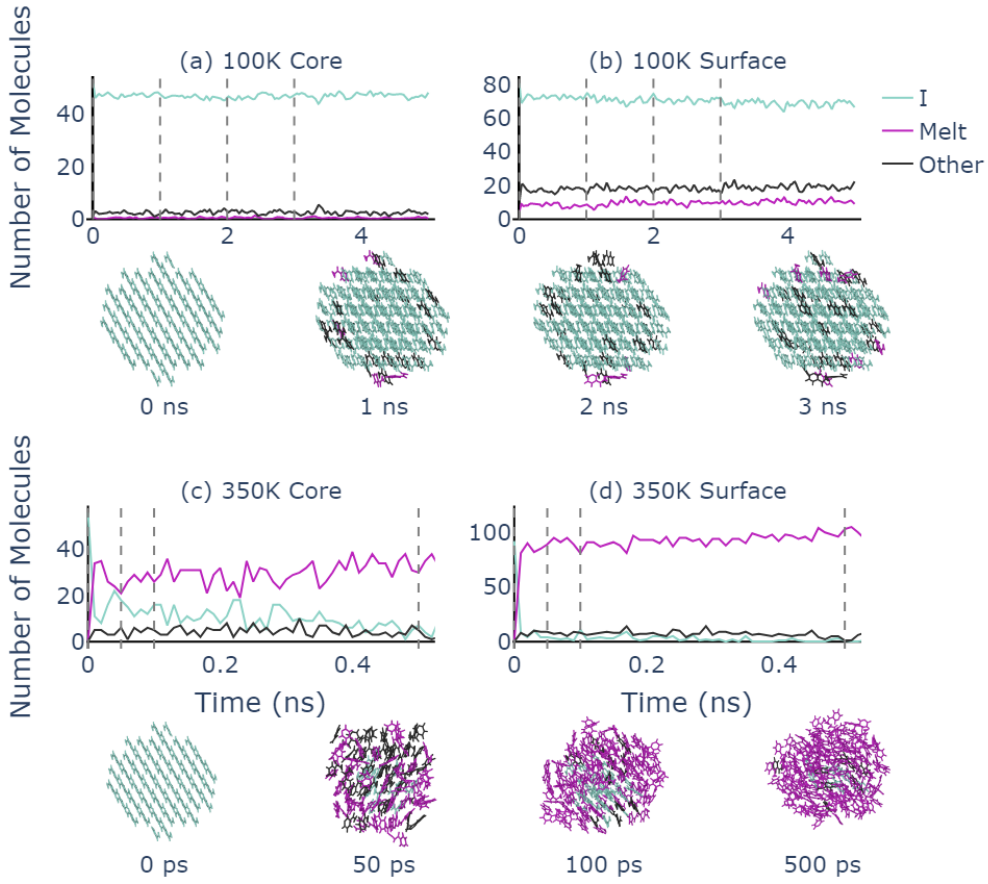


Figure 7: Time evolution of the number of molecules classified as form I, melt, or other, (a)-(b) at 100 K and (c)-(d) at 350 K. The analysis is shown separately for high-coordination 'core' molecules in (a) and (c) and low-coordination 'surface' molecules in (b) and (d). Vertical dashed lines identify the time points for the cluster snapshots, with molecules coloured according to their most probable form. Snapshots were visualized using OVITO[56]

bulk. At 350 K, the crystalline cluster quickly melts starting from the surface. Within a few picoseconds, molecules at the surface are identified as liquid with a handful labeled as others. The core region melts a little more slowly with a few molecules initially remaining as form I and others. After approximately 500 ps, the cluster appears to be completely melted with only a small number of core molecules identified as others.

Despite not having been trained on clusters in vacuum or mixtures of polymorphs, the performance of our graph classifier in the analysis of the simulation data is sensible and very informative, allowing to evaluate the structural stability and the onset of melting as a function of temperature.

### 3.2.2 Time evolution of solid-solid phase boundaries

Pushing our analysis tools even further, we apply our classification models to track the position of the interface between two different polymorphs of urea during a solid-solid transformation. A semi-coherent interface between form I and IV of urea is set up by pairing both phases along the [001] direction. The $xy$-dimensions parallel to the interface are fixed resulting in 1.7% compression in $x$ and 1.4% strain in $y$ of urea I and 2.8% compression in $x$ and 0.8% strain in $y$ of urea IV, respectively. Periodic boundary conditions are applied in all dimensions, keeping molecules at one of the interfaces fixed and simulations are run in the NVAP$_z$ ensemble at $T = 100$ K (further simulation details are given in the SI).

In Fig. 8, the analysis of the structural transformation using the graph classifier is shown. Initially, the system is mainly composed of urea form I (green molecules) in the top half of the simulation cell with some form IV at the bottom. Molecules at the interface between the two polymorphs are primarily identified as 'others' due to deviations of their local environments from the pure bulk polymorphs. Since within the chosen setup form I is rather unfavourable, transformation to form IV rapidly takes place over a few hundred femtoseconds, which is indicated by the continued increase of molecules identified as form IV and decrease of form I in the top graph of Fig. 8.

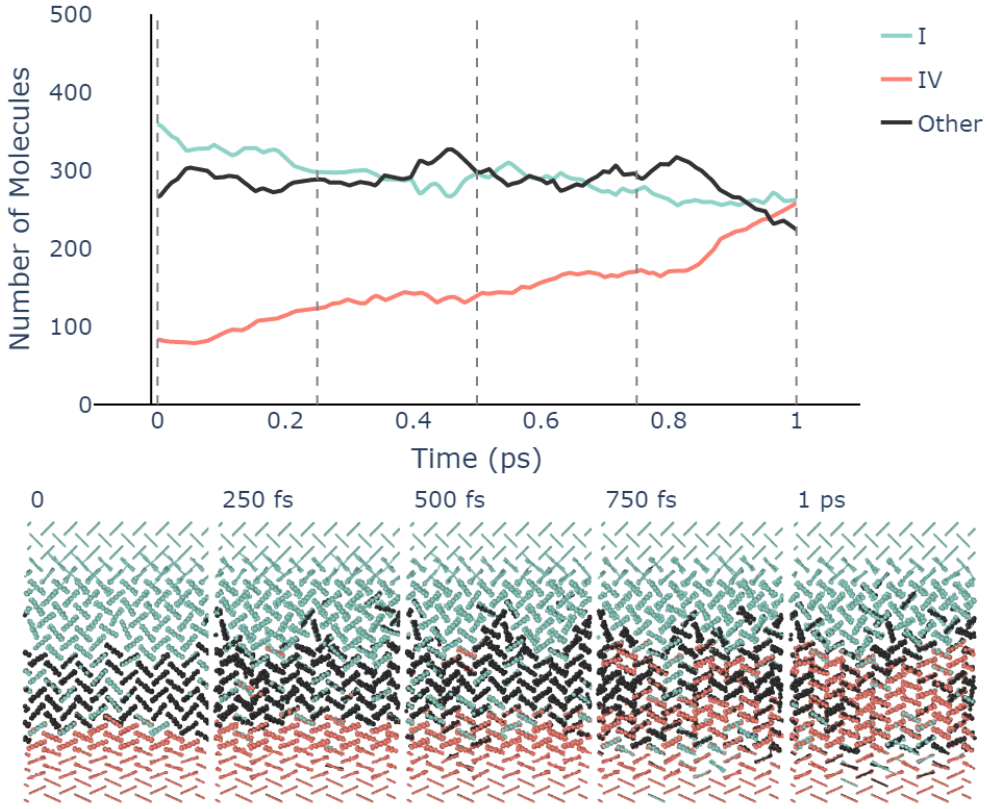


Figure 8: Time series of the molecule-wise composition of a system with a moving interface between form I and IV of urea. In the top graph, only molecules in the central region of the simulation cell, highlighted in bold in the snapshots below, are included. Vertical dashed lines correspond to the time points from which the snapshots were sampled, with molecules coloured according to their assigned polymorph. Snapshots were visualized using OVITO.[56]

Here again, the utility of accurate local environment classification is clearly evidenced, as subtle changes in local spacing and orientations of molecules can be seen to correspond to the transformation between distinct polymorphs, in this case form I and IV of urea. Interestingly, we also see that the conversion from form I to IV is not perfect, as some defects are left in the wake of the phase boundary as it moves upward through the sample.

# 4 Conclusions

We have introduced two machine learning based approaches for the classification of local structural environments in molecular solids. Both the GNN classifier with learned feature embeddings and the SF classifier with handcrafted descriptors identify molecular environments in various bulk polymorphs with high accuracy. While the performance of the two machine learning models is comparable for the studied systems, there are differences in their practical application.

The GNN model can be used for most molecular systems 'out of the box' with minimal customization but may require hyperparameter tuning to achieve good generalization. Due to its flexibility and expressive power, with the model presented here containing 356k parameters, the GNN classifier is somewhat sensitive to overfitting the training data. Again, one could train a smaller GNN model, at the empirically observed cost of slower convergence to inferior evaluation minima. Still, the model evaluates relatively quickly, at 35 training iterations, each comprising some hundreds of molecules, per second on V100 GPU compute and $\sim$1 per second on a single CPU. During evaluation, the performance bottleneck is more often the conversion from MD trajectory output files into the appropriate data format for the GNN model than the model forward pass itself, with 500 trajectory frames of 20 $\text{Å}^3$ bulk systems taking usually only several minutes to analyse. For sufficiently complex problems, a GNN classifier could in the future be upgraded with more sophisticated geometric features, convolutional methods or global aggregators, to efficiently capture longer-range intra- and

inter-molecular dependencies within a particular system. Today, such architectural improvements are relatively well understood and adoptable 'off the shelf'.

The performance of the SF classifier strongly depends on the handcrafted input features. The molecular symmetry functions proposed here do provide the flexibility to capture complex environments in molecular solids but need to be carefully chosen for each new system. This includes both the point-vector representation of the respective molecule as well as the tunable parameters of the symmetry function. For larger and more flexible molecules, it might be necessary to expand the molecular symmetry functions to explicitly account for conformational changes, for example, by introducing symmetry functions that depend on different vectors in the same molecule. It is, however, desirable to keep the number of molecular SFs small since calculating the input descriptors is the main computational cost when evaluating the SF classifier.

Both models are trivially parallelizable as they only require the information for a given molecule and its environment. They are also applicable to multi-component systems, such as co-crystals, or can be used to identify defects, such as impurities, vacancies, surfaces, or interfaces. The main challenge in these more complex scenarios is the preparation of labeled training data for the supervised learning task.

The two classification models presented in this study provide a general approach for the analysis and interpretation of simulation data in molecular solids. This will be particularly useful for the study of structural transformations, including nucleation and growth. Additionally, information about the local environment can also be used to construct collective variables used in enhanced sampling of structural transformations, as we have shown previously for atomistic systems.[19,20] We expect that the characterization of local structural motifs using classification models will become an essential tool in the simulation of molecular solids, as these models are easy to train and extremely versatile.

## Acknowledgement

The authors would like to thank Leslie Vogt-Maranto for fruitful discussions. The work of MK was supported by a Natural Science and Engineering Research Council of Canada (NSERC) post-doctoral fellowship. JR acknowledges financial support from the Deutsche Forschungsgemeinschaft (DFG) through the Heisenberg Programme project 428315600. JR and MET acknowledge funding from grants from the National Science Foundation, DMR-2118890, and MET from CHE-1955381. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

## Supporting Information Available

The supplemental information contains details regarding the descriptors, architecture, and training of the machine learning models as well as regarding the molecular dynamics simulations to create training and test data and molecular simulation examples.

## References

(1) Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B* **1983**, *28*, 784, DOI: 10.1103/PhysRevB.28.784.

(2) Lechner, W.; Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *J. Chem. Phys.* **2008**, *129*, 114707, DOI: 10.1063/1.2977970.

(3) Honeycutt, J. Dana.; Andersen, H. C. Molecular Dynamics Study of Melting and Freezing of Small Lennard-Jones Clusters. *J. Phys. Chem.* **1987**, *91*, 4950–4963, DOI: 10.1021/j100303a014.

(4) Faken, D.; Jónsson, H. Systematic analysis of local atomic structure combined with 3D

computer graphics. *Comput. Mater. Sci.* **1994**, *2*, 279–286, DOI: 10.1016/0927-0256(94)90109-0.

(5) Stukowski, A. Structure identification methods for atomistic simulations of crystalline materials. *Model. Simul. Mater. Sci. Eng.* **2012**, *20*, 045021, DOI: 10.1088/0965-0393/20/4/045021.

(6) Piaggi, P. M.; Parrinello, M. Entropy based fingerprint for local crystalline order. *J. Chem. Phys.* **2017**, *147*, 114112, DOI: 10.1063/1.4998408.

(7) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115, DOI: 10.1103/PhysRevB.87.184115.

(8) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401, DOI: 10.1103/PhysRevLett.98.146401.

(9) Neha,; Tiwari, V.; Mondal, S.; Kumari, N.; Karmakar, T. Collective Variables for Crystallization Simulations – from Early Developments to Recent Advances. *ACS Omega* **2022**, *8*, 127–146, DOI: 10.1021/acsomega.2c06310.

(10) Tanaka, H.; Tong, H.; Shi, R.; Russo, J. Revealing key structural features hidden in liquids and glasses. *Nat. Rev. Phys.* **2019**, *1*, 333–348, DOI: 10.1038/s42254-019-0053-3.

(11) Tong, H.; Xu, N. Order parameter for structural heterogeneity in disordered solids. *Phys. Rev. E* **2014**, *90*, 010401, DOI: 10.1103/PhysRevE.90.010401.

(12) Yang, X.; Liu, R.; Yang, M.; Wang, W.-H.; Chen, K. Structures of local rearrangements in soft colloidal glasses. *Phys. Rev. Lett.* **2016**, *116*, 238003, DOI: 10.1103/PhysRevLett.116.238003.

(13) Reinhardt, A.; Doye, J. P.; Noya, E. G.; Vega, C. Local order parameters for use in

driving homogeneous ice nucleation with all-atom models of water. *J. Chem. Phys.* **2012**, *137*, 194504, DOI: 10.1063/1.4766362.

(14) Eslami, H.; Khanjari, N.; Müller-Plathe, F. A local order parameter-based method for simulation of free energy barriers in crystal nucleation. *J. Chem. Theory Comput.* **2017**, *13*, 1307–1316, DOI: 10.1021/acs.jctc.6b01034.

(15) Piaggi, P. M.; Valsson, O.; Parrinello, M. Enhancing entropy and enthalpy fluctuations to drive crystallization in atomistic simulations. *Phys. Rev. Lett.* **2017**, *119*, 015701, DOI: 10.1103/PhysRevLett.119.015701.

(16) Song, H.; Vogt-Maranto, L.; Wiscons, R.; Matzger, A. J.; Tuckerman, M. E. Generating Cocrystal Polymorphs with Information Entropy Driven by Molecular Dynamics-Based Enhanced Sampling. *J. Phys. Chem. Lett.* **2020**, *11*, 9751–9758, DOI: 10.1021/acs.jpclett.0c02647.

(17) Geiger, P.; Dellago, C. Neural networks for local structure detection in polymorphic systems. *J. Chem. Phys.* **2013**, *139*, 164105, DOI: 10.1063/1.4825111.

(18) Cubuk, E. D.; Schoenholz, S. S.; Rieser, J. M.; Malone, B. D.; Rottler, J.; Durian, D. J.; Kaxiras, E.; Liu, A. J. Identifying structural flow defects in disordered solids using machine-learning methods. *Phys. Rev. Lett.* **2015**, *114*, 108001, DOI: 10.1103/PhysRevLett.114.108001.

(19) Rogal, J.; Schneider, E.; Tuckerman, M. E. Neural-network-based path collective variables for enhanced sampling of phase transformations. *Phys. Rev. Lett.* **2019**, *123*, 245701, DOI: 10.1103/PhysRevLett.123.245701.

(20) Rogal, J.; Tuckerman, M. E. *Multiscale Dynamics Simulations: Nano and Nano-bio Systems in Complex Environments*; Royal Society of Chemistry, 2021; Chapter 11, pp 312–348, DOI: 10.1039/9781839164668-00312.

(21) DeFever, R. S.; Targonski, C.; Hall, S. W.; Smith, M. C.; Sarupria, S. A Generalized Deep Learning Approach for Local Structure Identification in Molecular Simulations. *Chem. Sci.* **2019**, *10*, 7503–7515, DOI: 10.1039/C9SC02097G.

(22) Emanuele Boattini, F. S., Michel Ram; Filion, L. Neural-network-based order parameters for classification of binary hard-sphere crystal structures. *Mol. Phys.* **2018**, *116*, 3066–3075, DOI: 10.1080/00268976.2018.1483537.

(23) Boattini, E.; Marín-Aguilar, S.; Mitra, S.; Foffi, G.; Smallenburg, F.; Filion, L. Autonomously revealing hidden local structures in supercooled liquids. *Nat. Commun.* **2020**, *11*, 5479, DOI: 10.1038/s41467-020-19286-8.

(24) Scheiber, H.; Patey, G. Binary salt structure classification with convolutional neural networks: Application to crystal nucleation and melting point calculations. *J. Chem. Phys.* **2022**, *157*, 204108, DOI: 10.1063/5.0122274.

(25) Bapst, V.; Keck, T.; Grabska-Barwińska, A.; Donner, C.; Cubuk, E. D.; Schoenholz, S. S.; Obika, A.; Nelson, A. W.; Back, T.; Hassabis, D., et al. Unveiling the predictive power of static structure in glassy systems. *Nat. Phys.* **2020**, *16*, 448–454, DOI: 10.1038/s41567-020-0842-8.

(26) Banik, S.; Dhabal, D.; Chan, H.; Manna, S.; Cherukara, M.; Molinero, V.; Sankaranarayanan, S. K. CEGANN: Crystal Edge Graph Attention Neural Network for multiscale classification of materials environment. *npj Comput. Mater.* **2023**, *9*, 23, DOI: 10.1038/s41524-023-00975-z.

(27) Ishiai, S.; Endo, K.; Yasuoka, K. Graph neural networks classify molecular geometry and design novel order parameters of crystal and liquid. *J. Chem. Phys.* **2023**, *159*, 064103, DOI: 10.1063/5.0156203.

(28) Ishiai, S.; Yasuda, I.; Endo, K.; Yasuoka, K. Graph-Neural-Network-Based Unsupervised Learning of the Temporal Similarity of Structural Features Observed in Molec-

ular Dynamics Simulations. *J. Chem. Theory Comput.* **2024**, *20*, 819–831, DOI: 10.1021/acs.jctc.3c00995.

(29) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301, DOI: 10.1103/PhysRevLett.120.145301.

(30) Park, C. W.; Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* **2020**, *4*, 063801, DOI: 10.1103/PhysRevMaterials.4.063801.

(31) Kim, Q.; Ko, J.-H.; Kim, S.; Jhe, W. GCIceNet: a graph convolutional network for accurate classification of water phases. *Phys. Chem. Chem. Phys.* **2020**, *22*, 26340–26350, DOI: 10.1039/D0CP03456H.

(32) Beyerle, E. R.; Zou, Z.; Tiwary, P. Recent advances in describing and driving crystal nucleation usingmachine learning and artificial intelligence. *Curr. Opin. Solid State Mater. Sci.* **2023**, *27*, 101093, DOI: 10.1016/j.cossms.2023.101093.

(33) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **2019**, *31*, 3564–3572, DOI: 10.1021/acs.chemmater.9b01294.

(34) Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI open* **2020**, *1*, 57–81, DOI: 10.1016/j.aiopen.2021.01.001.

(35) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S. Y. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4–24, DOI: 10.1109/TNNLS.2020.2978386.

(36) Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *npj Comput. Mater.* **2021**, *7*, 84, DOI: 10.1038/s41524-021-00554-0.

(37) Santiso, E. E.; Trout, B. L. A general set of order parameters for molecular crystals. *J. Chem. Phys.* **2011**, *134*, 064109, DOI: 10.1063/1.3548889.

(38) Shah, M.; Santiso, E. E.; Trout, B. L. Computer simulations of homogeneous nucleation of benzene from the melt. *J. Phys. Chem. B* **2011**, *115*, 10400–10412, DOI: 10.1021/jp203550t.

(39) Kilgour, M.; Rogal, J.; Tuckerman, M. Geometric Deep Learning for Molecular Crystal Structure Prediction. *J. Chem. Theory Comput.* **2023**, *19*, 4743–4756, DOI: 10.1021/acs.jctc.3c00031.

(40) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet–a deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722, DOI: 10.1063/1.5019779.

(41) Shi, Y.; Huang, Z.; Feng, S.; Zhong, H.; Wang, W.; Sun, Y. Masked label prediction: Unified message passing model for semi-supervised classification. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. 2021; pp 1548–1554, DOI: 10.24963/ijcai.2021/214.

(42) Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. ICLR Workshop on Representation Learning on Graphs and Manifolds. 2019; DOI: 10.48550/arXiv.1903.02428.

(43) Gasteiger, J.; Groß, J.; Günnemann, S. Directional message passing for molecular graphs. International Conference on Learning Representations. 2020; DOI: 10.48550/arXiv.2003.03123.

(44) Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs), arXiv:1606.08415. **2023**, DOI: 10.48550/arXiv.1606.08415.

(45) Kilgour, M. MXtalTools. https://github.com/InfluenceFunctional/MXtalTools/tree/mol_classifier, 2024.

(46) Kuroshima, D.; Rogal, J. MolStrucClassifier. https://github.com/rogalj/MolStrucClassifier, 2024.

(47) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.* **2022**, *271*, 108171, DOI: 10.1016/j.cpc.2021.108171.

(48) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174, DOI: 10.1002/jcc.20035.

(49) Swaminathan, S.; Craven, B.; McMullan, R. The crystal structure and molecular thermal motion of urea at 12, 60 and 123 K from neutron diffraction. *Acta Crystallogr. B.* **1984**, *40*, 300–306, DOI: 10.1107/S0108768184002135.

(50) Olejniczak, A.; Ostrowska, K.; Katrusiak, A. H-bond breaking in high-pressure urea. *J. Phys. Chem. C* **2009**, *113*, 15761–15767, DOI: 10.1021/jp904942c.

(51) Giberti, F.; Salvalaglio, M.; Mazzotti, M.; Parrinello, M. Insight into the nucleation of urea crystals from the melt. *Chem. Eng. Sci.* **2015**, *121*, 51–59, DOI: https://doi.org/10.1016/j.ces.2014.08.032.

(52) Shang, C.; Zhang, X.-J.; Liu, Z.-P. Crystal phase transition of urea: What governs

the reaction kinetics in molecular crystal phase transitions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32125–32131, DOI: 10.1039/C7CP07060H.

(53) Li, X.; Ou, X.; Wang, B.; Rong, H.; Wang, B.; Chang, C.; Shi, B.; Yu, L.; Lu, M. Rich polymorphism in nicotinamide revealed by melt crystallization and crystal structure prediction. *Commun. Chem.* **2020**, *3*, 152, DOI: 10.1038/s42004-020-00401-1.

(54) Fellah, N.; Zhang, C. J.; Chen, C.; Hu, C. T.; Kahr, B.; Ward, M. D.; Shtukenberg, A. G. Highly polymorphous nicotinamide and isonicotinamide: Solution versus melt crystallization. *Cryst. Growth Des.* **2021**, *21*, 4713–4724, DOI: 10.1021/acs.cgd.1c00547.

(55) Hinton, G. E.; Roweis, S. Stochastic neighbor embedding. Advances in Neural Information Processing Systems. 2002; p 857–864.

(56) Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool. *Model. Simul. Mater. Sci. Eng.* **2009**, *18*, 015012, DOI: 10.1088/0965-0393/18/1/015012.

# Supplementary Information
# Machine learning classification of local environments in molecular crystals

Daisuke Kuroshima,[1] Michael Kilgour,[1] Mark E. Tuckerman,[1,2,3,4] and Jutta Rogal[1,5]

[1]*Department of Chemistry, New York University (NYU), New York, New York 10003, USA.*
[2]*Courant Institute of Mathematical Sciences, New York University, New York, New York 10012, USA.*
[3]*NYU-ECNU Center for Computational Chemistry at NYU Shanghai, 3663 Zhongshan Rd. North, Shanghai 200062, China.*
[4]*Simons Center for Computational Physical Chemistry at New York University, New York, New York 10003, USA.*
[5]*Fachbereich Physik, Freie Universität Berlin, 14195 Berlin, Germany.*

## I. CRYSTAL STRUCTURES OF UREA AND NICOTINAMIDE POLYMORPHS

Figs. 1 and 2 visualize the different polymorphs of urea and nicotinamide, respectively. These structures have been visualized using Ovito [1].

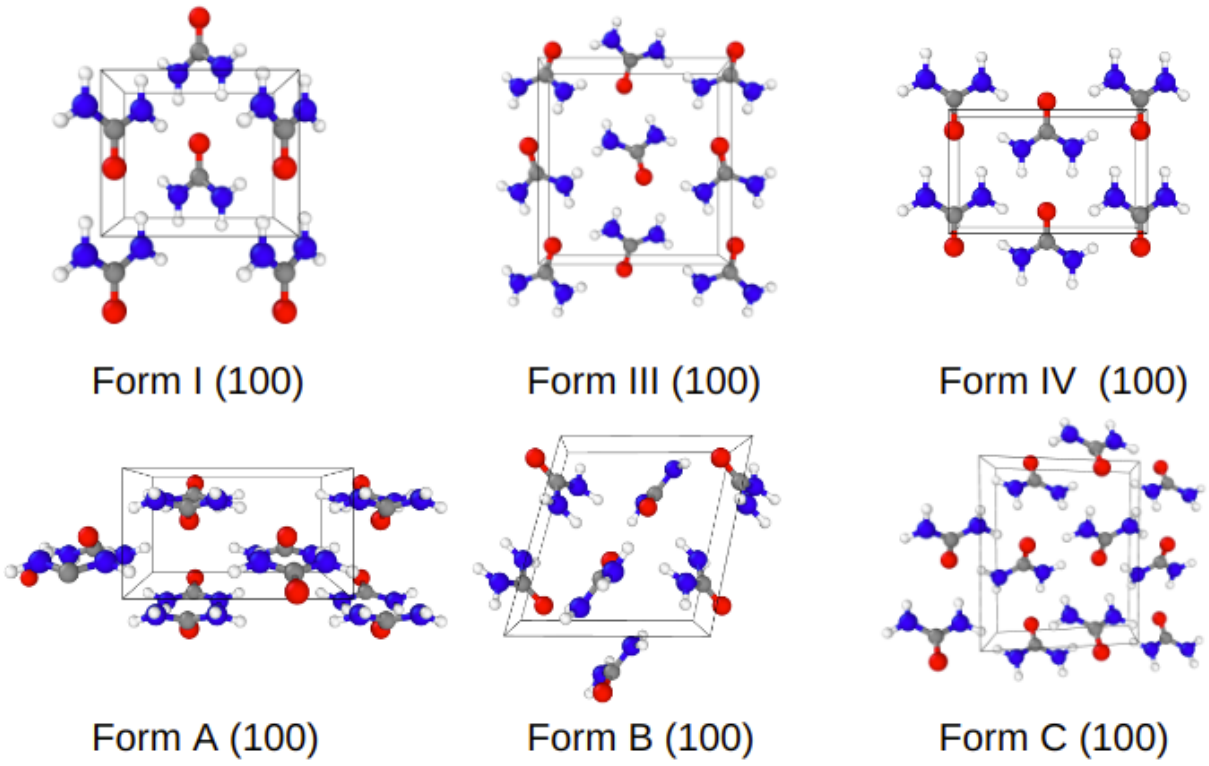


FIG. 1: Crystal structures of six urea polymorphs used in this study viewed along the [100] direction, including the experimentally crystallized forms I, III, and IV, as well as computationally proposed forms A, B, and C. [2–5]

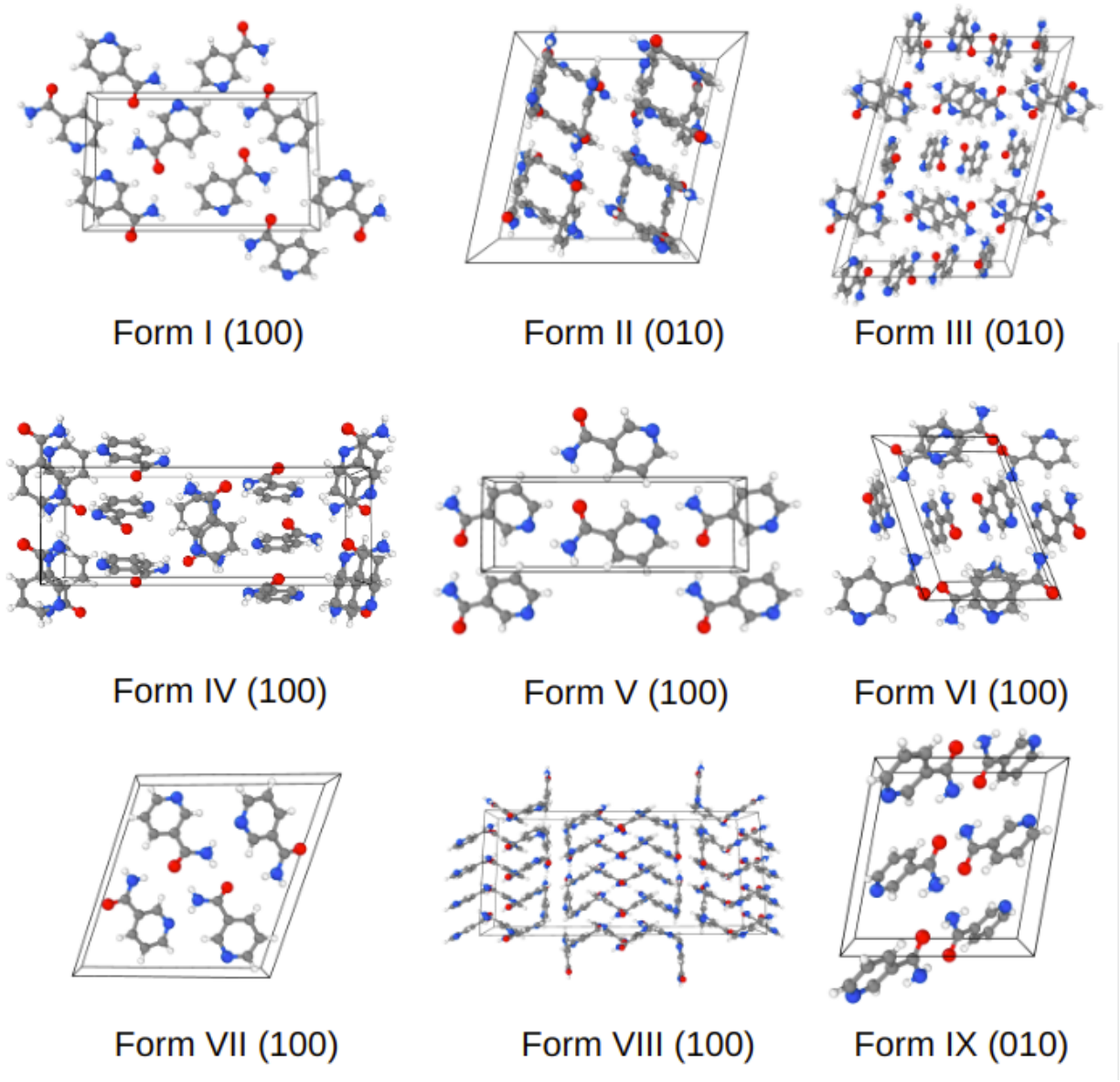


FIG. 2: Crystal structure of nine nicotinamide polymorphs used in this study. Form I, IV, V, VII, and VIII are viewed along the [100] direction, and Form II, III, VI, and IX along [010].

## II. MOLECULAR SYMMETRY FUNCTIONS AND TRAINING

The cutoff function used in the molecular symmetry functions has the following form [6]

$$f_c(\mathbf{r}_{IJ}) = \begin{cases} 1 & \text{if } |\mathbf{r}_{IJ}| < r_{\min} \\ \frac{1}{2}\left(\cos\left[\frac{(|\mathbf{r}_{IJ}|-r_{min})}{r_c-r_{min}}\pi\right]+1\right) & \text{if } r_{\min} < |\mathbf{r}_{IJ}| \leq r_c \\ 0 & \text{if } |\mathbf{r}_{IJ}| > r_c \end{cases} \quad (1)$$

where $|\mathbf{r}_{IJ}|$ is the distance between molecule $I$ and $J$. The cutoff radii are set to $r_{\min} = 9.8$ Å and $r_c = 10.0$ Å for urea and $r_{\min} = 6.8$ Å and $r_c = 7.0$ Å for nicotinamide. A set of input function was carefully selected by computing the distributions of symmetry function values for a series of the tunable parameters $R_s$, $\cos\theta_S$, $\eta$, and $\kappa$. The overlap of distributions for different polymorphs were compared and parameters resulting in small overlaps were selected. In total, 24 molecular symmetry functions were selected for both urea and nicotinamide. The corresponding values for the parameters are given in Tab. I for urea and Tab. II for nicotinamide.

TABLE I: Parameters of the molecular symmetry functions used for urea.

| symmetry function | $R_s$ | $\cos\theta_S$ | $\eta$ | $\kappa$ | vector |
|---|---|---|---|---|---|
| $S_1^I$ | | | | | |
| 1 | 6.16 | - | 2.44 | - | - |
| 2 | 6.28 | - | 2.68 | - | - |
| 3 | 6.76 | - | 1.00 | - | - |
| 4 | 6.88 | - | 1.00 | - | - |
| $S_2^I$ | | | | | |
| 5 | - | - | - | 2.50 | - |
| 6 | - | - | - | 4.54 | - |
| 7 | - | - | - | 4.90 | - |
| 8 | - | - | - | 6.22 | - |
| $S_3^I$ | | | | | |
| 9 | - | 0.368 | 1.00 | - | C-O |
| 10 | - | 0.08 | 1.00 | - | C-O |
| 11 | - | 0.36 | 1.12 | - | C-O |
| 12 | - | 0.28 | 6.76 | - | C-O |
| 13 | - | -0.64 | 3.28 | - | N-N |
| 14 | - | -0.36 | 3.28 | - | N-N |
| 15 | - | 0.88 | 3.28 | - | N-N |
| 16 | - | 1.00 | 3.28 | - | N-N |
| $S_4^I$ | | | | | |
| 17 | - | - | - | 2.50 | C-O |
| 18 | - | - | - | 3.58 | C-O |
| 19 | - | - | - | 4.78 | C-O |
| 20 | - | - | - | 8.26 | C-O |
| 21 | - | - | - | 2.50 | N-N |
| 22 | - | - | - | 8.12 | N-N |
| 23 | - | - | - | 8.24 | N-N |
| 24 | - | - | - | 8.36 | N-N |

The parameters for symmetry functions were adjusted by comparing the histograms of symmetry functions with different parameters. The overlap of the histogram was calculated for each polymorph, and eight parameters each from the point, first point-vector, and second point-vector were selected as descriptors. These symmetry functions were applied to trajectory of each bulk system, and the resulting calculations from each molecule at each snapshot were stored for use in the classification NN.

To train the classification NN with these sets of descriptors, 5,000 and 10,000 training samples were used for urea and nicotinamide, respectively.

TABLE II: Parameters of the molecular symmetry functions used for nicotinamide.

| symmetry function | $R_s$ | $\cos\theta_S$ | $\eta$ | $\kappa$ | vector |
|---|---|---|---|---|---|
| $S_1^I$ | | | | | |
| 1 | 3.75 | - | 1.26 | - | - |
| 2 | 5.25 | - | 0.01 | - | - |
| 3 | 4.9 | - | 0.1 | - | - |
| 4 | 5.9 | - | 0.016 | - | - |
| $S_2^I$ | | | | | |
| 5 | - | - | - | 1.06 | - |
| 6 | - | - | - | 0.51 | - |
| 7 | - | - | - | 1.03 | - |
| 8 | - | - | - | 2.41 | - |
| $S_3^I$ | | | | | |
| 9 | - | 0.01 | 0.66 | - | C-C |
| 10 | - | 1.66 | 0.01 | - | C-C |
| 11 | - | 2.9 | 0.9 | - | C-C |
| 12 | - | 1.69 | 4.0 | - | C-C |
| 13 | - | 0.56 | 3.0 | - | O-N |
| 14 | - | 0.66 | 0.01 | - | O-N |
| 15 | - | 1.18 | 17.2 | - | O-N |
| 16 | - | 1.15 | 2.2 | - | O-N |
| $S_4^I$ | | | | | |
| 17 | - | - | - | 0.13 | C-C |
| 18 | - | - | - | 0.33 | C-C |
| 19 | - | - | - | 5.05 | C-C |
| 20 | - | - | - | 3.2 | C-C |
| 21 | - | - | - | 1.05 | O-N |
| 22 | - | - | - | 0.57 | O-N |
| 23 | - | - | - | 2.36 | O-N |
| 24 | - | - | - | 13.22 | O-N |

## III. GRAPH MODEL HYPERPARAMETERS AND TRAINING

The graph neural network classifier was constructed with one convolutional layer, a nodewise fully-connected layer, followed by two fully-connected layers after graph aggregation. The graph convolution cutoff was 6 Å. The feature depth was 256 throughout, except for during message passing where it was bottlenecked down to 128. Regularization was added with a dropout probability of 0.5 on all fully-connected layers, graphwise layernorm on the graph nodes, and standard layernorm on the graph embedding. We used the Adam optimizer [7] with a constant learning rate of $10^{-4}$, and a batch size of 5, synthesized via gradient accumulation over 5 MD snapshots.

The train and test datasets were comprised of 1050 and 250 MD snapshots, respectively, sampled at randomly spaced time intervals, containing on average 370 molecules each, adding up to approximately 390k total molecular environments. Convergence studies showed similar convergence on as little as 10% of this data, which is unsurprising, since at low temperature, most local molecular environments for a given polymorph should be very similar.

## IV. DATASET PREPARATION

Bulk periodic molecular dynamics trajectories of the known polymorphs of urea and nicotinamide were undertaken under the following conditions. Simulations were undertaken using the LAMMPS [8] molecular dynamics program. Simulations were run for 1 ns with a time step $\Delta t = 1$ fs in the NPT ensemble using a Nosé-Hoover thermostat and barostat implemented in LAMMPS [9–12].

In this work, we employed the AMBER force field for urea and nicotinamide which relies on second generation of Generalized Amber Force Field (gaff2) [13]. Partial charges for urea were taken from OPLS [14] and for nicotinamide using RESP-charges from PBE calculations [15].

Simulation box sizes were set as the minimum number of unit cell replicas in each cell direction to achieve at least the desired box length, where box lengths of 20 Å and 40 Å were used. The 20 Å samples were used in the GNN model for training on periodic bulk structures. The 40 Å boxes were used to carve out spheres with a 30 Å diameter to create molecular environments on a surface. Surface molecules were identified as having intermolecular coordination numbers less than 20, with that value identified via coordination number histograms within several test clusters, and visually confirmed by inspection of the clusters themselves. Initial configurations of nicotinamide gas phase clusters used were generated in the same way and placed in large periodic boxes to simulate vacuum.

Trajectories were run at temperatures of 100 K and 200 K for urea crystal polymorphs, and 350 K for melts, and at 100 K and 350 K for nicotinamide crystal polymorphs and 350 K for melts. These temperatures were chosen to ensure that sample structures were clearly melted or crystalline for each molecule.

Melt structures were prepared starting from a stable crystal. After relaxing the system, we gradually increased the temperature from 350 K to 2,000 K over a duration of 10 picoseconds to melt the system. Subsequently, we reduced the temperature of the system back to 350 K on the same timescale. A simulation was then run for 1 nanosecond, and the resulting data was used to characterize the molten structure.

The interface structure was prepared using form I and IV urea structures. To minimize the mismatch within the system, we oriented both structures along the [001] plane, resulting in 1.7% compression in $x$ and 1.4% strain in $y$ of urea I and 2.8% compression in $x$ and 0.8% strain in $y$ of urea IV, respectively. . To avoid having two moving interfaces, we fixed one of the interface of form I and IV in the $z$-dimensions, then proceeded to relax the system using MD simulation.

The collected data were randomly divided into testing and training sets. Various sizes of training data, ranging from N = 100 to N = 50,000 unique molecular environments, were used.

## V. SYMMETRY FUNCTION CLASSIFIER ACCURACY

Figs. 3-4 show the evaluation accuracy of the SF classifier on high temperature samples of urea and nicotinamide, respectively. The overall accuracy is nearly perfect in both cases. Note that the SF classifier was only trained on bulk samples, therefore surface vs. bulk classification performance is ommited in this analysis.

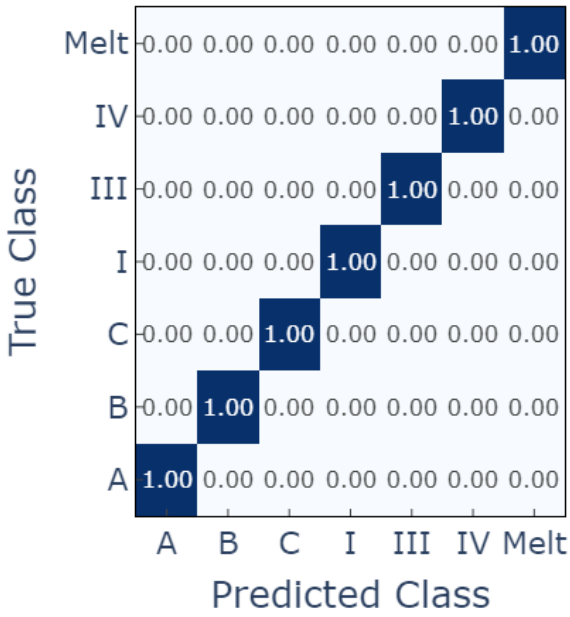


FIG. 3: Confusion matrix for the symmetry function classifier on the polymorphs of urea at 200 K for crystals and 350 K for melt. Micro F1 score=1.0.

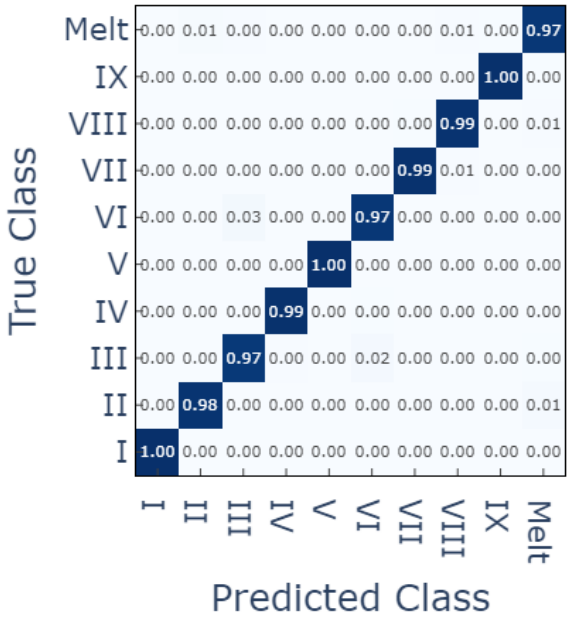


FIG. 4: Confusion matrix for the symmetry function classifier on the polymorphs of nicotinamide at 350 K. Micro F1 score=0.986.

[1] A. Stukowski, Model. Simul. Mater. Sci. Eng. **18**, 015012 (2009).
[2] S. Swaminathan, B. Craven, and R. McMullan, Acta Crystallogr. B. **40**, 300 (1984).
[3] A. Olejniczak, K. Ostrowska, and A. Katrusiak, J. Phys. Chem. C **113**, 15761 (2009).
[4] F. Giberti, M. Salvalaglio, M. Mazzotti, and M. Parrinello, Chem. Eng. Sci. **121**, 51 (2015).
[5] C. Shang, X.-J. Zhang, and Z.-P. Liu, Phys. Chem. Chem. Phys. **19**, 32125 (2017).
[6] M. Chen, M. A. Cuendet, and M. E. Tuckerman, J. Phys. Chem. B **137**, 024102 (2012).
[7] D. P. Kingma and J. Ba, CoRR **abs/1412.6980** (2014).
[8] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, et al., Comp. Phys. Comm. **271**, 108171 (2022).
[9] W. Shinoda, M. Shiga, and M. Mikami, Phys. Rev. B **69**, 134103 (2004).
[10] G. J. Martyna, D. J. Tobias, and M. L. Klein, J. Chem. Phys. **101**, 4177 (1994).
[11] M. Parrinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981), ISSN 0021-8979.
[12] M. E. Tuckerman, J. Alejandre, R. López-Rendón, A. L. Jochim, and G. J. Martyna, J. Phys. A: Math. Gen. **39**, 5629 (2006).
[13] X. He, V. H. Man, W. Yang, T.-S. Lee, and J. Wang, J. Chem. Phys. **153** (2020).
[14] E. M. Duffy, D. L. Severance, and W. L. Jorgensen, Isr. J. Chem. **33**, 323 (1993).
[15] J. Wang, P. Cieplak, and P. A. Kollman, J. Comput. Chem. **21**, 1049 (2000).