Characterising the atomic structure of mono-metallic nanoparticles from x-ray scattering data using conditional generative models

Andy S. Anker*
andy@chem.ku.dk
Department of Chemistry,
University of Copenhagen
Copenhagen, Denmark

Emil T. S. Kjær*
etsk@chem.ku.dk
Department of Chemistry,
University of Copenhagen
Copenhagen, Denmark

Erik B. Dam
erikdam@di.ku.dk
Department of Computer Science,
University of Copenhagen
Copenhagen, Denmark

Simon J. L. Billinge
Department of Applied Physics and Applied Mathematics, Columbia University
Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory
New York, NY, USA

Kirsten M. Ø. Jensen
kirsten@chem.ku.dk
Department of Chemistry,
University of Copenhagen
Copenhagen, Denmark

Raghavendra Selvan
raghav@di.ku.dk
Department of Computer Science,
University of Copenhagen
Copenhagen, Denmark

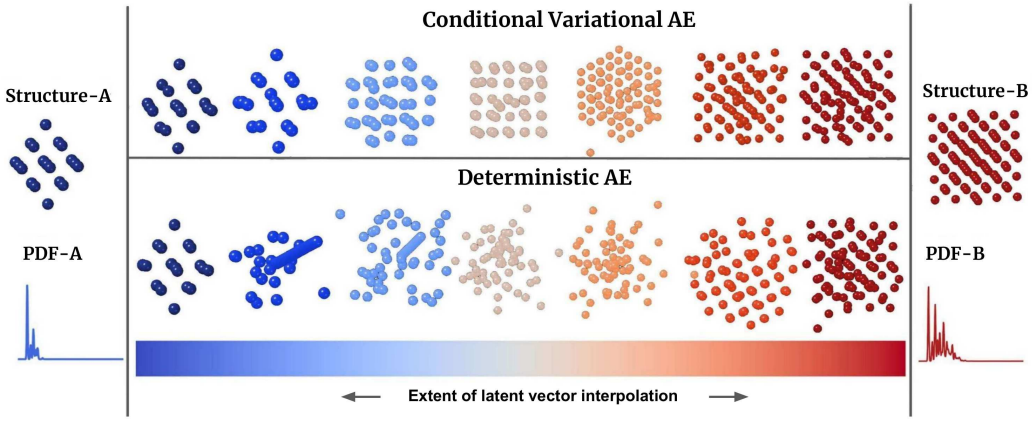

Figure 1: Structures generated by decoding different extents of interpolation of the latent variables obtained for PDF-A and PDF-B. The generated structures start from Structure-A and progressively evolve towards Structure-B. This work uses a Conditional Variational Autoencoder (CVAE) and we compare it with a Deterministic Autoencoder (DAE).

ABSTRACT

The development of new nanomaterials for energy technologies is dependent on understanding the intricate relation between material properties and atomic structure. It is, therefore, crucial to be able to routinely characterise the atomic structure in nanomaterials, and a promising method for this task is Pair Distribution Function (PDF) analysis. The PDF can be obtained through Fourier transformation of x-ray total scattering data, and represents a histogram of all interatomic distances in the sample. Going from the distance information in the PDF to a chemical structure is an unassigned distance geometry problem (uDGP), and solving this is often the bottleneck in nanostructure analysis. In this work, we propose to use a Conditional Variational Autoencoder (CVAE) to automatically solve the uDGP to obtain valid chemical structures from PDFs. We use a simple model system of hypothetical mono-metallic nanoparticles containing up to 100 atoms in the face centered cubic (FCC) structure as a proof of concept. The model is trained to predict the assigned distance matrix (aDM) from a simulated PDF of the structure as the conditional input. We introduce a novel representation of structures by projecting them inside a unit sphere and adding additional anchor points or *satellites* to help in the reconstruction of the chemical structure. The performance of the CVAE model is compared to a Deterministic Autoencoder (DAE) showing that both models are able to solve the uDGP reasonably well. We further show that the CVAE learns a structured and meaningful latent embedding space which can be used to predict new chemical structures.

*Both authors contributed equally to this research.

Presented in 16th International Workshop on Mining and Learning with Graphs, Aug 24, 2020, San Diego,CA

CCS CONCEPTS

• **Computing methodologies** → **Learning latent representations**; • **Applied computing** → **Chemistry**.

KEYWORDS

generative modeling, mono-metallic nanoparticles, CVAE, Pair Distribution Function

1 INTRODUCTION

The development of nanoscience over the last decades has given completely new possibilities for material engineering and development. [30] Compared to their bulk counterparts, nanomaterials can, for example, show improved properties in energy technologies such as catalysis, solar cells, and batteries. [38] In particular, many new properties arise in 'ultrasmall' nanoparticles, where the dimensions are smaller than 5 nm. The very large surface-to-volume ratio in such materials can lead to drastic changes in the material performance in e.g. catalysis, and a fundamental change in atomic structure may also take place when going to the nanoscale. [23, 27]

In order to engineer nanomaterials with targeted properties, the link between atomic structures and properties must be understood, and it is crucial to be able to routinely characterise the atomic structure in nanomaterials. [12, 39] However, such materials challenge many of the conventional methods for structure characterisation. Traditionally, material structure is characterised through x-ray and neutron diffraction techniques, applying crystallographic methods. [11] These diffraction techniques rely on the presence of long-range, periodic atomic order in the samples. Nanomaterials do not possess this long range-atomic order, and crystallographic methods cannot directly be applied for characterisation of many nanomaterials. [5, 6]

A way of overcoming this 'nanostructure problem' is through the use of the Pair Distribution Function (PDF). [18] The PDF of a sample is obtained through Fourier transformation of *total scattering* data, which can be collected using e.g. large scale synchrotron facilities. The PDF represents a histogram of all interatomic distances in the sample, and is a plot of intensity versus r, i.e. interatomic distance. Peak positions correspond to atom-atom distances while the intensity of a peak depends on the number of pairs having that interatomic distance and the type of atoms in the pair. A simulated PDF from a gold nanoparticle with the face centered cubic (FCC) atomic structure can be seen in Figure 2. The 9 black cubes mark the smallest repeated units, i.e. the FCC unit cells. Each peak in the PDF corresponds to an interatomic distance in the particle.

The PDF thus contains information on the atoms present in a structure and their relative interatomic distances, as is also the case for methods like solid-state Nuclear Magnetic Resonance (NMR) and Extended X-ray Absorption Fine Spectroscopy (EXAFS). In a PDF, these relative interatomic distances are not linked to their corresponding atomic identity. Going from this distance information to a structure can be described as an unassigned distance geometry problem (uDGP), compared to the assigned distance geometry problem (aDGP), where the assignment of atoms is known. [3, 4, 17] Obtaining physical structures from the unassigned distance matrix (uDM) is a combinatorial problem whereas it is straightforward to reconstruct the structure once the atoms are assigned to obtain the assigned distance matrix (aDM). This makes it challenging to go

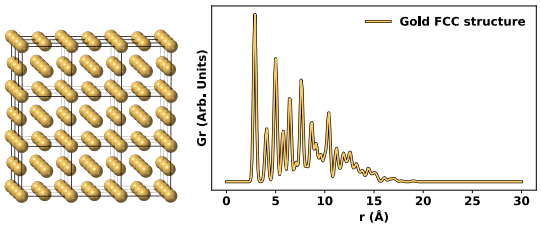

Figure 2: Illustration of an FCC gold particle and its corresponding PDF. The 9 black cubes show the smallest repeated units, i.e. the FCC unit cells.

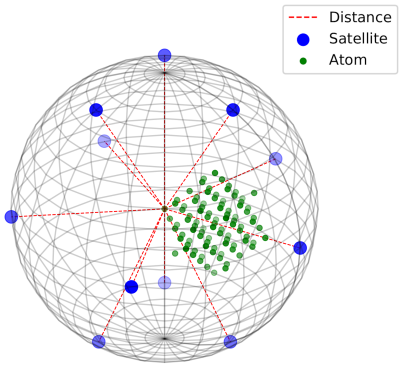

Figure 3: Illustration of the structure of a mono-metallic nanoparticle (green) represented using the unit sphere along with eleven *satellites* (blue).

directly from PDF to structure, and analysing the PDF to extract a structure model is often a bottleneck in material characterisation.

In this paper, we propose to use a deep latent variable model for structure characterisation from simulated PDF data. We address the task of going from the uDGP to the aDGP using a Conditional Variational Autoencoder (CVAE). [29, 42] As a proof of concept we use a CVAE based approach to predict atomic structures from PDFs of mono-metallic nanoparticles of sizes up to 100 atoms. We focus on a simple, hypothetical system, where the nanoparticles all are made to take the FCC structure (Figure 2), but are made of different elements, and are of different sizes. We demonstrate that a CVAE model followed by an trilateration algorithm can be used to go from a PDF to the particle structure in a matter of seconds (Figure 5).

We use a novel graph representation of structures by projecting them inside a unit sphere such that one of the atoms is at the origin of the sphere, and the farthest atom is on the sphere surface. This allows us to capture relative distances between atoms as node features across different structures in a consistent manner. Furthermore, to improve the reconstruction of the structures, several *satellites* or anchor points are introduced on the surface of the unit sphere inspired by trilateration techniques used in satellite based geo-localisation. [40] A sample structure in this representation is shown in Figure 3.

The CVAE receives the input structure (aDM) and the PDF (uDM) as input. We treat the PDF as the conditional input to the CVAE which is trained to predict both the aDM and the distances between

each atom and the satellites. The final location of atoms are determined by trilateration. We also study the influence of the number of satellites on the quality of reconstructed structures. We compare the ability of the CVAE with a Deterministic Autoencoder (DAE) which also receives the PDF as the conditional input and predicts the aDM. Additionally, we explore the latent embedding space learnt by the two models by interpolating between latent vector points to predict novel structural motifs as depicted in Figure 1.

2 BACKGROUND

2.1 Related work

In PDF analysis, the uDGP is currently addressed through modelling. A PDF can be calculated from a structure model, and the model is refined until a good agreement between the experimental and calculated PDF is obtained. Multiple programs [9, 26, 34] provide a framework for assigning the distances through refinement of one or more structure models to the data. Minimizing the difference between model and data during refinement can be a complex problem, and for nanomaterials [6, 23] in particular, identifying good starting models for the atomic structure is difficult. To overcome this bottleneck, both model-based and data-driven approaches are currently being developed. For model-based classification of PDF data, cluster-mining [2] and structure-mining [43] approaches have been proposed to automate refinements of large numbers of structure models to the data. These methods can help scientists expand their view of possible structure solutions, as thousands of structures can be tested. However, model-based algorithms are computationally heavy and are limited to already known starting structures that can be mined from databases.

Recently, machine learning (ML) methods have started surfacing in the field of structure characterisation with PDF. A Convolutional Neural Network (CNN) has been used to determine the space group of a structure from an experimental PDF. [32] A few papers have also shown the use of dimensionality reduction tools such as Principal Component Analysis (PCA) and Non-negative Matrix Factorization (NMF) on PDF data. Such methods are useful in particular when dealing with PDFs from multiphase systems, as the contributions from e.g. different chemical species can be resolved. [10, 14, 20] Both the space-group classification and the dimensionality reduction methods do not directly characterise atomic structure, but help constrain the number of structural possibilities for further analysis.

Variational Autoencoders (VAE) [29] have shown great promise as a method for generating new and/or targeted organic molecules. These data-driven models use the SMILES [7, 22, 41] string representation of graphs to generate structures. VAEs have been used to generate new and valid chemical structures for a diverse class of molecules. [22, 31] CVAEs can be trained to generate structures that respect targeted properties by conditioning the model on additional data. This allows organic molecules to be generated rapidly without having to explore all structural possibilities through theoretical computations. [7, 41] The contributions mentioned here are only a fraction of the work done applying ML to chemistry. [8, 15, 16, 19, 21] Within inorganic chemistry, CNN have shown promise in characterising microstructures from scattering data [8]. Unsupervised ML methods have also been used to discover structural relationships and link them to microscopic properties on experimentally characterised structures from chemical databases. [36] However, work focusing on applying ML methods to characterise atomic structures from x-ray data is still very sparse.

2.2 X-ray total scattering and Pair Distribution Function (PDF) for structural analysis

A PDF can be obtained by Fourier transforming experimental x-ray total scattering data. In an x-ray scattering experiment, one measures the angle-dependent interference pattern that arises when monochromatic x-rays scatter off a sample, as this scattering pattern contains information on the sample structure. X-ray total scattering experiments are best done at large scale synchrotron facilities where high energy x-rays are available, as scattering data must be collected to high values of the *scattering vector* Q:

$$Q = 4\pi sin(\theta)/\lambda \tag{1}$$

Here, λ is the radiation wavelength, and θ is the scattering angle. For PDF analysis, high quality scattering data are generally needed in the range ca. $0.5 - 25$ Å^{-1}.

The measured x-ray total scattering data is corrected and normalized to obtain the structure function $S(Q)$. This function is then Fourier transformed over the available Q-range yielding the PDF, $G(r)$: [18]

$$G(r) = \frac{2}{\pi} \int_{Q_{min}}^{Q_{max}} Q[S(Q) - 1] \sin(Q \cdot r) dQ, \tag{2}$$

The Fourier transform of the scattering data in Q thus yields a function in r, which represents a histogram of interatomic distances or the uDM. [13, 24, 25]

In the current work, all scattering data and PDFs have been simulated from structure models. This has been done using the Diffpy-CMI software. [26] Details are given in Section 4.3 and simulation parameters are shown in Appendix C, Table 3.

3 METHODS

We are interested in reconstructing structures of mono-metallic nanoparticles, $\mathbf{s} \in \mathcal{S}$, from their corresponding PDFs, $\mathbf{x} \in \mathcal{X}$. Each structure, $\mathbf{s} \in \mathbb{R}^{N \times F}$, comprises N nodes with F node attributes. The node attributes are the atomic number, interatomic distances and distance to the satellites. The corresponding PDFs are represented as the sequence $\mathbf{x} \in \mathbb{R}^D$. The data used in this work are detailed in Section 4.3.

Reconstructing structures from PDFs or the task of going from uDGP to aDGP can be formulated as the learning task: $f(\cdot) : \mathcal{X} \rightarrow \mathcal{S}$. Seen from a density point of view, we are interested in estimating $p(\mathbf{s}|\mathbf{x})$. In this work, we take up a latent variable approach to capture the dependencies between $\mathbf{s}$ and $\mathbf{x}$. This is done by introducing the latent variable $\mathbf{z} \in \mathbb{R}^H$, where H is the latent (hidden) dimension. Specifically, we use the CVAE framework to derive a model and to access a meaningful latent space that can be used for conditional generative sampling. [29, 42]

The CVAE extends the VAE framework to incorporate conditioning inputs to the encoder-decoder structure. The CVAE objective minimizes the Kullback-Leibler (KL) divergence between the true

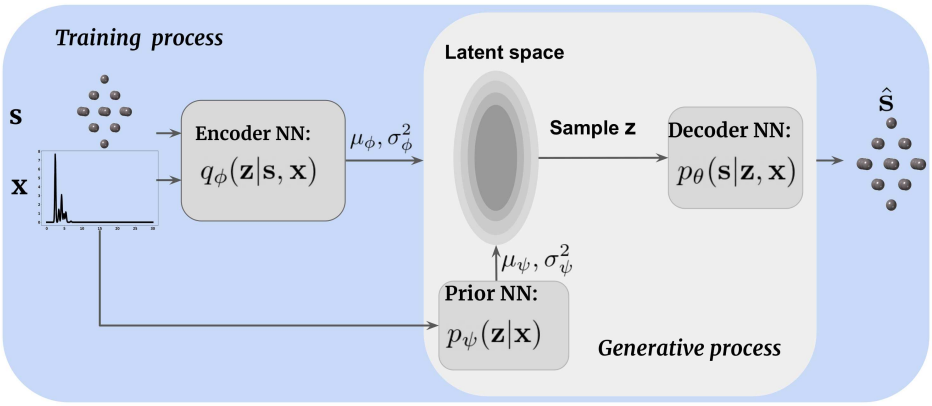

Figure 4: CVAE utilized to predict the aDM used to reconstruct the mono-metallic nanoparticle structure with conditional input provided in the form of the PDF. The encoder gets the structure, s, and its PDF, x, as input. The PDF is also input to the prior network. The decoder network predicts the aDM, ŝ, from a latent sample, z, which is used to reconstruct the structure.

posterior distribution $p(\mathbf{z}|\mathbf{s},\mathbf{x})$ and its variational approximation $q_\phi(\mathbf{z}|\mathbf{s},\mathbf{x})$ resulting in an objective of the form: [42]

$$\mathcal{L}_{\text{CVAE}} = \mathcal{L}_{rec} + \mathcal{L}_{reg} \tag{3}$$

$$= -\mathbb{E}_{q_\phi}\big[\log p_\theta(\mathbf{s}|\mathbf{z},\mathbf{x})\big] + \text{KL}\big[q_\phi(\mathbf{z}|\mathbf{s},\mathbf{x})||p_\psi(\mathbf{z}|\mathbf{x})\big] \tag{4}$$

This objective is derived in Appendix A.

The first term in the objective in Eq. (4) can be interpreted as the reconstruction loss, $\mathcal{L}_{rec}$. More formally, it is the expected conditional log-likelihood under the approximate posterior density predicted by the encoder. The expectation is with respect to the variational density, $q_\phi(\mathbf{z}|\mathbf{s},\mathbf{x}) = \mathcal{N}(\mathbf{z};\boldsymbol{\mu}_\phi,\boldsymbol{\sigma}^2_\phi)$, which is constrained to be a Gaussian with mean $\boldsymbol{\mu}_\phi$ and variance $\boldsymbol{\sigma}^2_\phi$. The mean and variance of the posterior density are predicted by the encoder neural network parameterised by ϕ. The encoder maps the input structure, $\mathbf{s}$, and the corresponding PDF, $\mathbf{x}$, to a latent random variable $\mathbf{z} \in \mathbb{R}^H$. The term inside the expectation is the conditional log likelihood, $p_\theta(\mathbf{s}|\mathbf{z},\mathbf{x})$, predicted by the decoder neural network parameterised by θ. The decoder predicts the aDM of interest, $\hat{\mathbf{s}}$, from the latent representation, $\mathbf{z}$, and the conditioning input, $\mathbf{x}$. The reconstruction loss in practice is computed as the mean-squared error between the input aDM $\mathbf{s}$ and the predicted aDM $\hat{\mathbf{s}}$, i.e.,

$$\mathcal{L}_{rec} = ||\mathbf{s} - \hat{\mathbf{s}}||^2. \tag{5}$$

The second term in Eq. (4), $\mathcal{L}_{reg}$, acts as the regularisation term forcing the posterior density to match the conditional prior density $p_\psi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z};\boldsymbol{\mu}_\psi,\boldsymbol{\sigma}^2_\psi)$ which is also constrained to be Gaussian with mean $\boldsymbol{\mu}_\psi$ and variance $\boldsymbol{\sigma}^2_\psi$. The prior network is parameterised by ψ.

The CVAE model used in this work is illustrated in Figure 4 and the network architecture details are described in Section 4.1. The details of the training and generative (inference) processes are described below:

Training process: The encoder gets a structure in the form of an aDM and its corresponding PDF as input to predict the latent density parameters. The prior network predicts the parameters of the prior density from the PDF. The KL divergence is computed between the posterior and the prior densities, resulting in the $\mathcal{L}_{reg}$ term. A sample from the latent posterior density is then decoded in the decoder network. The decoder's output is the aDM which is compared with the input structure to obtain the reconstruction loss $\mathcal{L}_{rec}$. The combined loss is backpropagated through the network, training the model to reconstruct the aDM from the latent space.

Generative process: During the inference time, only the trained prior and decoder networks are used. The PDF is input to the prior network and a sample from the resulting prior density is decoded by the decoder, *generating* a new aDM. The final location of atoms are determined by trilateration from the predicted distances, as described in Section 4.2.

4 EXPERIMENTS

4.1 Network Architecture

The CVAE model used in this work has three high level components that are implemented using neural networks.

The encoder, $q_\phi(\cdot)$, consists of 4 fully connected layers with rectified linear unit activation's except for the last layer. These encoder layers use $[F, 384, 256, 128, 2H]$ hidden units where F is the input feature dimension and H is the latent dimension. The PDF which is the conditional input, $\mathbf{x}$, is concatenated to the input structure, $\mathbf{s}$, by upsampling it to match the node feature dimensions. The upsampling operation is performed with 3 layers of gated 1-D transpose convolution layers. The encoder predicts the mean, $\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}^2_\phi \in \mathbb{R}^H$, of the encoder density. Latent vectors, $\mathbf{z} \in \mathbb{R}^H$, are sampled from the predicted density: $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}^2_\phi)$.

The prior, $p_\psi(\cdot)$, comprises 3 gated 1-D convolution layers with rectified linear unit except for the last layer with $[D, 48, 24, 2H]$ hidden nodes where D is the input PDF dimension. The prior network predicts the mean, $\boldsymbol{\mu}_\psi, \boldsymbol{\sigma}^2_\psi \in \mathbb{R}^H$, of the prior density.

The decoder, $p_\theta(\cdot)$, predicts $\hat{\mathbf{s}}$ from the latent variable, $\mathbf{z}$. The decoder is implemented with 5 fully connected layers with rectified linear unit except for the last layer. These decoder layers have $[H, 128, 256, 384, 512]$ hidden units and outputs $\hat{\mathbf{s}} \in \mathbb{R}^{N\times F}$ which are the relative distance between the atoms and the atoms distances to the satellites.

4.2 Post-processing

The CVAE model reduces the uDGP to aDGP. The prediction from the CVAE model, $\hat{\mathbf{s}}$, is a combination of the aDM and atom-satellite distances. From this prediction, one can reconstruct the structures using trilateration. We start with the first atom in the distance list as the origin and add subsequent atoms with respect to the previous ones based on the relative distances between atoms and the satellites, which were introduced for the purpose of trilateration. The reconstruction can be seen as solving the aDGP but with help from the atom-satellite distances, thereby minimizing the uncertainty on the placed atoms. If the predicted distances are accurate there is a unique solution. As the predicted distances are not exact, we pose the trilateration task as an optimization problem and use an L-BFGS-B optimizer to solve for the reconstructed structure. [44]

4.3 Data and model hyperparameters

Data: The nanoparticle structures were simulated using the Atomic Simulation Environment (ASE). [1] All nanoparticles were made to take the FCC structure, but multiple particles were obtained by varying the type of atoms, exposed surfaces, layers of atoms and lattice constant. This simulation procedure yielded 3137 unique nanoparticle structures consisting of fewer than 100 atoms. From each of the 3137 structures, a PDF was simulated using Diffpy-CMI. [26]. Details of the simulation procedure and the parameters used which reflect typical values for synchrotron experiments are described in Appendix C, Table 3.

Baseline model: There are no established baseline models to compare with for the tasks considered in this work. We compare the CVAE with a DAE model which has exactly the same architecture as the CVAE model except for the stochasticity. The DAE has been chosen as a baseline as it can perform the regression task of going from uDM to aDM and it also learns a (deterministic) latent representation of the input. The latent variable, $\mathbf{z}$, in the DAE model is treated as a deterministic variable. And instead of the KL divergence based regularisation inherent to CVAE, L2 regularisation is introduced between the latent vectors predicted by the encoder network and the prior networks during training.

Performance metric: Comparing generative models for the task under consideration is not straightforward as there are no established metrics or measures to quantify the validity of these nanoparticle structures. One surrogate measure that could signify the quality of structures can be derived by obtaining the PDFs of the generated structures, $\hat{\mathbf{x}}$, and comparing them to the input PDFs. Note that going from structure to PDF is easier than the converse task. The difference in PDFs is quantified using a Mean-Squared Error (MSE) term, $R_p = ||\mathbf{x} - \hat{\mathbf{x}}||_2$ and the Pearson correlation coefficient.

Hyperparameters: The CVAE and DAE models were trained with 2400, validated on 600 and tested on 137 structures. The models were implemented in PyTorch. [37] Training was performed with a batch size of 20. Adam optimizer with a learning rate of $5e^{-5}$ was used to optimize the loss in Eq. (4) for all the experiments. [28] The models were assumed to have converged if there was no improvement in validation loss for 50 epochs. The latent space dimension was set to 2 and the number of satellites 11. All the parameters were tuned based on experiments on the training set performance.

Table 1: Mean-squared error E_p and average Pearson correlation between the input PDFs and the PDFs reconstructed from the generated structures for the the DAE and CVAE models. Significant differences ($p < 0.001$) based on two-tailed paired sample t-tests are shown in bold.

Model	R_p	Pearson
DAE	**0.4548 ± 0.30**	**0.7482 ± 0.17**
CVAE	0.4823 ± 0.34	0.7321 ± 0.19

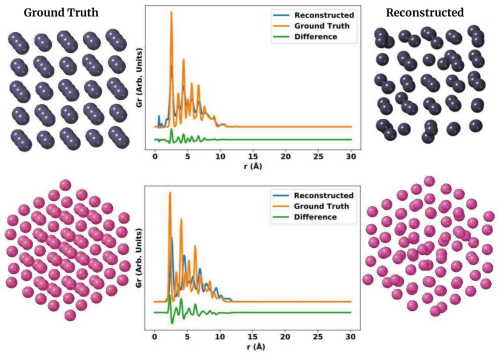

Figure 5: Visualisation of the PDF of the ground truth and the reconstructed PDF of the reconstructed mono-metallic nanoparticle for two typical cases from the validation set. Both cases show reasonable reconstructions, but the top example has broader peaks than the ground truth due to disorder in the structure. For the bottom example, the PDF is shifted due to incorrect distance between the atoms. The difference curve is scaled by 0.2 for better visualisation.

5 RESULTS AND DISCUSSION

The two models were trained on the same training/validation/test splits and the performance metrics are reported in Table 1 where we observe that the DAE model is better at the reported metrics. This is expected as there are no stochastic elements in the DAE model. However, an important feature of the CVAE model is the access to a *meaningful* latent space which can be used to explore and obtain new structural motifs. This is clearly demonstrated in Figure 1. The CVAE model can not only reasonably reconstruct structure A and structure B from their PDFs, but it can also generate reasonable mono-metallic nanoparticle structures from the interpolated latent vectors. As we move from left to right, we are traversing between the latent points for the two structures and we see a clear evolution of structures. This is not the case for the DAE model, which produces invalid structures when decoding from the interpolated latent vectors.

One common feature of both DAE and CVAE models is that they can generate structures comprising the right amount of atoms, as illustrated on the test set for the CVAE model in Figure 8. When interpolating in latent space, the number of atoms take discrete steps.

This behaviour is probably due to the discrete nature of the training set. A more diverse training set can be achieved by adding additional structure geometries or facets of the nanoparticles.

In Figure 5, two structures obtained from PDFs with the CVAE model are visualised along with the ground truth structures. Apart from the reconstructed structures, the PDFs reconstructed from the generated structure are shown. The reconstructed structure in the top row has some disorder which is also reflected in the reconstructed PDF. The reconstructed PDF in the top row has overlapping and broader peaks when compared to the ground truth. The structure in the bottom row has longer distances between the atoms than the ground truth, which results in a slight offset in the reconstructed PDF. Currently, we do not explicitly optimize for the PDFs to be aligned. These inconsistencies in the PDFs can also be incorporated into the training process by introducing an additional loss term that is dependent on the PDFs which we expect could improve the quality of generated structures.

The satellites were introduced to help with the trilateration. The effect of using $[4, 11, 100]$ satellites is shown in Figure 6. It can be seen that the reconstructed structures improve with the number of satellites. There is a small improvement in the reconstructed structures when using 100 satellites compared to 11 but it comes at a large increase in computation time as shown in Table 2, where the computer time is calculated as how long time it takes to solve the trilateration problem. This might be a factor to consider if this method is to be used to get instantaneous results. A reasonable trade-off between performance and computation time was found to be with 11 satellites.

6 CONCLUSION AND FUTURE WORK

Deep learning based methods have proven to be extremely powerful at embedding complex data into low dimensions, while still learning complex and useful features. In this work, we have demonstrated a proof of concept on embedding structures of mono-metallic nanoparticles while conditioning each structure on their corresponding simulated PDF. Both DAE and CVAE models show that valid structures can be obtained from the PDFs. Furthermore, the models show that they are not constrained to reconstruct a fixed number of atoms. These methods could allow for fast analysis of PDF data, which can prove extremely potent e.g. for real-time data analysis during synchrotron experiments.

Furthermore, we have shown that the latent space learnt by the CVAE model is highly structured (Figure 7). This latent space can be used to obtain novel structural motifs as illustrated in Figure 1 where interpolated latent vectors resulted in valid structures when compared to the DAE model.

By normalizing the structures to reside inside a unit sphere we show how trilateration can be used to reconstruct a chemical structure from an aDM. We investigated a range of fixed satellite points for structure reconstruction showing that using 11 satellites was a reasonable trade-off between structure reconstruction accuracy and computation time.

This work has shown great promise which calls for future investigations. The data used for this model only consist of a small number of mono-metallic structures, all with the FCC structure. For the model to be useful for chemists the chemical-space must be massively expanded, for instance by including structure types other than FCC. In addition, the model could benefit from having a PDF difference regularization term implemented in its loss function.

Table 2: Average reconstruction time with varying number of satellites on a standard laptop with 8 CPU cores. The mean and deviation are calculated based on 50 reconstructions.

# of satellites	Time (s)
4	8.6 ± 5.4
11	9.6 ± 5.9
100	23.5 ± 12.0

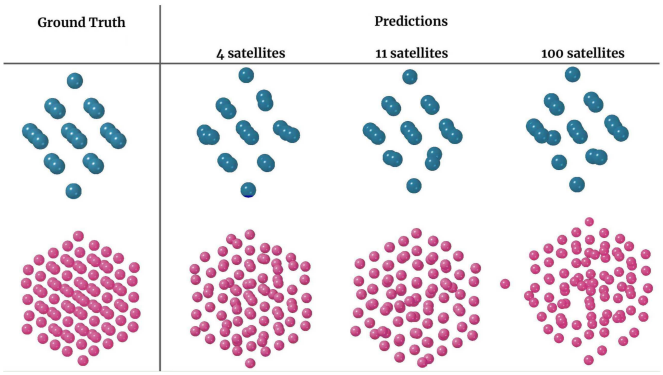

Figure 6: Influence of the number of satellites used for trilateration, on the quality of reconstructed structures for two types of mono-metallic nanoparticles with: 19 atoms (top row) and 87 atoms (bottom row).

Currently the model only predicts the positions of the atoms for mono-metallic nanoparticles as the composition is assumed to be known. To expand the use cases of the model, the atom labels need also be predicted. These are directions we look forward to pursuing.

In conclusion, the results reported in this work have been encouraging and the proposed model can be used as a generative model for characterisation of mono-metallic FCC nanoparticles.

Acknowledgments. This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, (Grant agreement no. 804066).

Work in the Billinge group was supported by the US National Science Foundation through grant DMR-1922234.

REFERENCES

[1] Jakob Blomqvist Ivano E. Castelli Rune Christensen Marcin Dułak Jesper Friis Michael N. Groves Bjørk Hammer Cory Hargus Eric D. Hermes Paul C. Jennings Peter Bjerre Jensen James Kermode John R. Kitchin Esben Leonhard Kolsbjerg Joseph Kubal Kristen Kaasbjerg Steen Lysgaard Jón Bergmann Maronsson Tristan Maxson Thomas Olsen Lars Pastewka Andrew Peterson Carsten Rostgaard Jakob Schiøtz Ole Schütt Mikkel Strange Kristian S. Thygesen Tejs Vegge Lasse Vilhelmsen Michael Walter Zhenhua Zeng Ask Hjorth Larsen, Jens Jørgen Mortensen and Karsten W Jacobsen. 2017. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* 29, 27 (2017), 273002.

[2] Soham Banerjee, Chia-Hao Liu, Kirsten M. Ø. Jensen, Pavol Juhás, Jennifer D. Lee, Marcus Tofanelli, Christopher J. Ackerson, Christopher B. Murray, and Simon J. L. Billinge. 2020. *Cluster-mining*: an approach for determining core structures of metallic nanoparticles from atomic pair distribution function data. *Acta Crystallographica Section A* 76, 1 (Jan 2020), 24–31.

[3] Simon J. L. Billinge, Phillip M Duxbury, Douglas S. Gonçalves, Carlile Lavor, and Antonio Mucherino. 2016. Assigned and unassigned distance geometry:

applications to biological molecules and nanostructures. *4OR* 14, 4 (2016), 337–376.
[4] Simon J. L. Billinge, Phillip M. Duxbury, Douglas S. Gonçalves, Carlile Lavor, and Antonio Mucherino. 2018. Recent results on assigned and unassigned distance geometry with applications to protein molecules and nanostructures. *Annals of Operations Research* 271, 1 (December 2018), 161–203.
[5] Simon J. L. Billinge and Mercouri G. Kanatzidis. 2004. Beyond crystallography: the study of disorder, nanocrystallinity and crystallographically challenged materials with pair distribution functions. *Chemical communications* 7 (2004), 749–760.
[6] Simon J. L. Billinge and Igor Levin. 2007. The Problem with Determining Atomic Structure at the Nanoscale. *Science* 316, 5824 (2007), 561–565.
[7] John Bradshaw, Brooks Paige, Matt J. Kusner, Marwin H. S. Segler, and José Miguel Hernández-Lobato. 2019. A Model to Search for Synthesizable Molecules. *CoRR* abs/1906.05221 (2019).
[8] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. 2018. Machine learning for molecular and materials science. *Nature* 559, 7715 (2018), 547–555.
[9] J. W. Liu D Bryndin E. S. Božin J. Bloch Th Proffen C L Farrow, P Juhas and Simon J. L. Billinge. 2007. PDFfit2 and PDFgui: computer programs for studying nanostructure in crystals. *Journal of Physics: Condensed Matter* 19, 33 (jul 2007), 335219.
[10] Karena W. Chapman, Saul H. Lapidus, and Peter J. Chupas. 2015. Applications of principal component analysis to pair distribution function data. *Journal of Applied Crystallography* 48, 6 (Dec 2015), 1619–1626.
[11] Anthony K. Cheetham and Andrew L. Goodwin. 2014. Crystallography with powders. *Nature materials* 13, 8 (2014), 760–762.
[12] Troels Lindahl Christiansen, Susan R. Cooper, and Kirsten M. Ø. Jensen. 2020. There's no place like real-space: elucidating size-dependent atomic structure of nanomaterials using pair distribution function analysis. *Nanoscale Adv.* (2020).
[13] Troels Lindahl Christiansen, Emil T. S. Kjær, Anton Kovyakh, Morten L. Röderen, Martin Høj, Tom Vosch, and Kirsten M. Ø. Jensen. 2020. Structure analysis of supported disordered molybdenum oxides using pair distribution function analysis and automated cluster modelling. *Journal of Applied Crystallography* 53, 1 (Feb 2020), 148–158.
[14] Jacqueline M. Cole, Xie Cheng, and Michael C. Payne. 2016. Modeling Pair Distribution Functions of Rare-Earth Phosphate Glasses Using Principal Component Analysis. *Inorganic Chemistry* 55, 21 (2016), 10870–10880.
[15] Tânia F. G. G. Cova and Alberto A. C. C. Pais. 2019. Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns. *Frontiers in Chemistry* 7 (2019), 809.
[16] David K. Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*. 2224–2232.
[17] P.M. Duxbury, L. Granlund, S.R. Gujarathi, P. Juhas, and Simon J. L. Billinge. 2016. The unassigned distance geometry problem. *Discrete Applied Mathematics* 204 (2016), 117 – 132.
[18] Takeshi Egami and Simon J. L. Billinge. 2003. *Underneath the Bragg peaks: structural analysis of complex materials*. Elsevier.
[19] Daniel C. Elton, Zois Boukouvalas, Mark D. Fuge, and Peter W. Chung. 2019. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* 4 (2019), 828–849. Issue 4.
[20] Harry S. Geddes, Helen Blade, James F. McCabe, Leslie P. Hughes, and Andrew L. Goodwin. 2019. Structural characterisation of amorphous solid dispersions via metropolis matrix factorisation of pair distribution function data. *Chem. Commun.* 55 (2019), 13346–13349. Issue 89.
[21] Garrett B. Goh, Nathan O. Hodas, and Abhinav Vishnu. 2017. Deep learning for computational chemistry. *Journal of computational chemistry* 38, 16 (2017), 1291–1307.
[22] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. 2018. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* 4, 2 (2018), 268–276.
[23] Kirsten M. Ø. Jensen, Pavol Juhas, Marcus A. Tofanelli, Christine L. Heinecke, Gavin Vaughan, Christopher J. Ackerson, and Simon J. L. Billinge. 2016. Polymorphism in magic-sized Au144(SR)60 clusters. *Nature Communications* 7 (2016).
[24] Kirsten M. Ø. Jensen, Mogens Christensen, Pavol Juhas, Christoffer Tyrsted, Espen D. Bøjesen, Nina Lock, Simon J. L. Billinge, and Bo B. Iversen. 2012. Revealing the Mechanisms behind SnO2 Nanoparticle Formation and Growth during Hydrothermal Synthesis: An In Situ Total Scattering Study. *Journal of the American Chemical Society* 134, 15 (2012), 6785–6792.
[25] Mikkel Juelsholt, Troels Lindahl Christiansen, and Kirsten M. Ø. Jensen. 2019. Mechanisms for Tungsten Oxide Nanoparticle Formation in Solvothermal Synthesis: From Polyoxometalates to Crystalline Materials. *The Journal of Physical Chemistry C* 123, 8 (2019), 5110–5119.
[26] Pavol Juhás, Christopher L. Farrow, Xiaohao Yang, Kevin R. Knox, and Simon J. L. Billinge. 2015. Complex modeling: a strategy and software program for combining multiple information sources to solve ill posed structure and nanostructure inverse problems. *Acta Crystallographica Section A* 71, 6 (Nov 2015), 562–568.
[27] Byung Hyo Kim, Michael J Hackett, Jongnam Park, and Taeghwan Hyeon. 2014. Synthesis, characterization, and application of ultrasmall nanoparticles. *Chemistry of Materials* 26, 1 (2014), 59–71.
[28] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[29] Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations*.
[30] Maksym V Kovalenko, Liberato Manna, Andreu Cabot, Zeger Hens, Dmitri V Talapin, Cherie R Kagan, Victor I Klimov, Andrey L Rogach, Peter Reiss, Delia J Milliron, et al. 2015. Prospects of nanoscience with nanocrystals.
[31] Youngchun Kwon, Jiho Yoo, Youn-Suk Choi, Won-Joon Son, Dongseon Lee, and Seokho Kang. 2019. Efficient learning of non-autoregressive graph variational autoencoders for molecular graph generation. *Journal of Cheminformatics* 11, 1 (2019), 70.
[32] Chia-Hao Liu, Yunzhe Tao, Daniel Hsu, Qiang Du, and Simon J. L. Billinge. 2019. Using a machine learning approach to determine the space group of a structure from the atomic pair distribution function. *Acta Crystallographica Section A* 75, 4 (2019), 633–643.
[33] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.
[34] Martin T. Dove Andrew L. Goodwin Matthew G. Tucker, David A. Keen and Qun Hui. 2007. RMCProfile: reverse Monte Carlo for polycrystalline materials. *Journal of Physics: Condensed Matter* 19, 33 (jul 2007), 335218.
[35] Łukasz Mentel. [n.d.]. *mendeleev – A Python resource for properties of chemical elements, ions and isotopes*.
[36] Thomas C. Nicholas, Andrew L. Goodwin, and Volker L. Deringer. 2020. Understanding the Geometric Diversity of Inorganic and Hybrid Frameworks through Structural Coarse-Graining. *arXiv preprint arXiv:2005.09939* (2020).
[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. 8024–8035.
[38] Ekaterina Pomerantseva, Francesco Bonaccorso, Xinliang Feng, Yi Cui, and Yury Gogotsi. 2019. Energy storage: The future enabled by nanomaterials. *Science* 366, 6468 (2019).
[39] Emil Roduner. 2006. Size matters: why nanomaterials are different. *Chem. Soc. Rev.* 35 (2006), 583–592. Issue 7.
[40] P. E. Schmid and J. J. Lynn. 1975. Results of the 3 November 1974 Applications Technology Satellite-6 (ATS-6) trilateration test. *NASA Technical Reports Server* (1975).
[41] Martin Simonovsky and Nikos Komodakis. 2018. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *CoRR* abs/1802.03480 (2018).
[42] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*. 3483–3491.
[43] Long Yang, Pavol Juhás, Maxwell W. Terban, Matthew G. Tucker, and Simon J. L. Billinge. 2020. Structure-mining: screening structure models by automated fitting to the atomic pair distribution function over large numbers of models. *Acta Crystallographica Section A* 76, 3 (May 2020), 395–409.
[44] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Trans. Math. Softw.* 23, 4 (1997), 550–560.

A DERIVING THE CVAE OBJECTIVE

We start with the motivation of approximating the latent posterior distribution with a variational density that minimizing the *reverse* KL divergence,

$$\begin{aligned}\mathrm{KL}\Big[q(\mathbf{z}|\mathbf{s},\mathbf{x})||p(\mathbf{z}|\mathbf{s},\mathbf{x})\Big] &= \mathbb{E}_{q(\mathbf{z}|\mathbf{s},\mathbf{x})}\Big[\log\frac{q(\mathbf{z}|\mathbf{s},\mathbf{x})}{p(\mathbf{z}|\mathbf{s},\mathbf{x})}\Big] \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{s},\mathbf{x})}\Big[\log\Big(\frac{q(\mathbf{z}|\mathbf{s},\mathbf{x})}{p(\mathbf{s}|\mathbf{z},\mathbf{x})}\frac{p(\mathbf{s}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x})}\Big)\Big] \quad (6)\end{aligned}$$

The last equality is due to Bayes' Rule. Some algebraic manipulations yield,

$$\begin{aligned}\mathrm{KL}\Big[q(\mathbf{z}|\mathbf{s},\mathbf{x})||p(\mathbf{z}|\mathbf{s},\mathbf{x})\Big] &= \mathbb{E}_{q(\mathbf{z}|\mathbf{s},\mathbf{x})}\Big[\log\frac{q(\mathbf{z}|\mathbf{s},\mathbf{x})}{p(\mathbf{z}|\mathbf{x})}\Big] \quad (7)\\ -\mathbb{E}_{q(\mathbf{z}|\mathbf{s},\mathbf{x})}\Big[\log p(\mathbf{s}|\mathbf{z},\mathbf{x})\Big] &+ \log p(\mathbf{s}|\mathbf{x})\end{aligned}$$

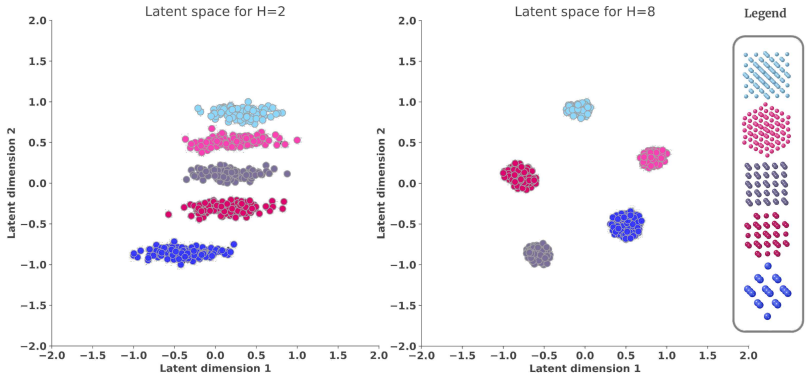

Figure 7: Comparison of embedded latent space for latent dimensions $H = 2$ and $H = 8$. 600 PDFs are converted to the latent variables through the prior network. The points are colored with respect to the number of atoms in the structure in the same color as the ground truth structures shown in the legend.

Table 3: Simulation parameters for the PDF data.

Q (Å^{-1})		r (Å)		Vibrations (Å^2)	
Qmin	0	rmin	0	ADP	1
Qmax	30	rmax	30	delta2	2
Qdamp	0.04	rstep	0.1	–	–

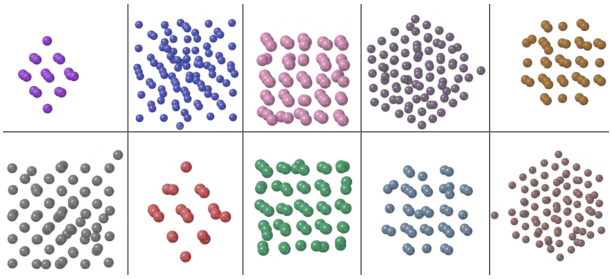

Figure 8: Ten sample reconstructions from the test set predicted by the CVAE model with latent space dimension $H = 2$ and 11 satellites.

Note the first term is $\mathrm{KL}\Big[q(\mathbf{z}|\mathbf{s},\mathbf{x})||p(\mathbf{z}|\mathbf{x})\Big]$ and the first two terms form a lower bound on the conditional likelihood. Thus,

$$\mathrm{KL}\Big[q(\mathbf{z}|\mathbf{s},\mathbf{x})||p(\mathbf{z}|\mathbf{s},\mathbf{x})\Big] = -\mathcal{L}b\Big[p(\mathbf{s}|\mathbf{x})\Big] + \log p(\mathbf{s}|\mathbf{x}) \tag{8}$$

where

$$\mathcal{L}b\Big[p(\mathbf{s}|\mathbf{x})\Big] = \mathbb{E}_{q(\mathbf{z}|\mathbf{s},\mathbf{x})}\Big[\log p(\mathbf{s}|\mathbf{z},\mathbf{x})\Big] - \mathrm{KL}\Big[q(\mathbf{z}|\mathbf{s},\mathbf{x})||p(\mathbf{z}|\mathbf{x})\Big] \tag{9}$$

The bound $\mathcal{L}b\big(p(\mathbf{s}|\mathbf{x})\big)$ is equal to the conditional likelihood when the KL divergence between the variational density and the true posterior density is zero. This is essentially the reason why we can maximize a surrogate objective such as the bound on the conditional likelihood to indirectly minimize the KL divergence we set to in Eq. (6).

The bound in Eq. (9) is a maximization objective. We choose to minimize the negative of this bound in the CVAE objective. In (4), each of the densities are parameterised by neural networks reflected in the subscripts $\{\phi, \theta, \psi\}$, resulting in the final objective:

$$\mathcal{L}_{\mathrm{CVAE}} = -\mathbb{E}_{q_\phi}\Big[\log p_\theta(\mathbf{s}|\mathbf{z},\mathbf{x})\Big] + \mathrm{KL}\Big[q_\phi(\mathbf{z}|\mathbf{s},\mathbf{x})||p_\psi(\mathbf{z}|\mathbf{x})\Big].$$

B LATENT SPACE COMPARISON

In order to test the effect of the latent space dimension, two CVAE models using latent space dimensions of 2 and 8 were trained. Figure 7 shows the latent space of the trained models with $H = [2, 8]$. For both latent space dimensions, 600 PDFs were input to the trained prior network to obtain 600 latent variables. The 8-dimensional space was reduced to 2 dimensions by the use of t-SNE dimensionality reduction method. [33] The points were colored according to the number of atoms in the structure, to visually emphasize how the latent space clusters the structures. The reconstructions from the model using latent space dimension 8 is slightly better than of a model with latent space dimension 2. Figure 7 shows that the latent space is equally compact with latent space dimension 2 and 8, but t-SNE makes it challenging to directly compare the compactness of the two latent spaces. Also, it is possible that the latent space with dimension 8 captures some chemical details that is not present in the 2-dimensional space, however, we have not been able to identify it. In the end, we choose to use a smaller latent space dimension of 2 which could be beneficial when interpolating in the latent space as demonstrated in Figure 1. However, if the goal is simply to reconstruct structures from the PDF a larger latent space or the DAE model can be utilized.

C SIMULATION PARAMETERS FOR PDF

Mono-metallic nanoparticle structures for each of the atoms in the d-block of the periodic table (period 4-6) were generated. All nanoparticles were built to have the FCC structure. The size and shape of each nanoparticle were controlled by parameters used in the ASE framework: [1] The number of atoms is given by the number of layers, as counted from the center of the particles. The shape is defined by specifying Miller indices for the exposed facets of the particles. Each structure was generated with layers ranging from 2 to 8 and with exposed (100), (110) and (111) facets on the surface. For each configuration of metal, layers and surfaces, 10 structures were obtained using lattice constants equally spaced in the range of 99% to 101% of the ideal lattice constant based on the atomic metallic radius's which were obtained from the mendeleev python package. [35] This process yielded 3137 unique structures consisting of fewer than 100 atoms.