DeepStruc: Towards structure solution from pair distribution function data using deep generative models

Emil T. S. Kjær[†1], Andy S. Anker[†1], Marcus N. Weng[1], Simon J. L. Billinge[2,3], Raghavendra Selvan*[4,5], Kirsten M. Ø. Jensen*[1]*

†Both authors contributed equally to this work.

*Correspondence to sb2896@columbia.edu, (SJLB), raghav@di.ku.dk (RS), kirsten@chem.ku.dk (KMØJ)

1: Department of Chemistry and Nano-Science Center, University of Copenhagen, 2100 Copenhagen Ø, Denmark

2: Department of Applied Physics and Applied Mathematics Science, Columbia University, New York, NY 10027, USA

3: Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, NY 11973, USA

4: Department of Computer Science, University of Copenhagen, 2100 Copenhagen Ø, Denmark

5: Department of Neuroscience, University of Copenhagen, 2200, Copenhagen N

Abstract

Structure solution of nanostructured materials that have limited long-range remains a bottleneck in materials development. We present a deep learning algorithm, DeepStruc, that can solve a simple nanoparticle structure directly from a Pair Distribution Function obtained from total scattering data by using a conditional variational autoencoder (CVAE). We first apply DeepStruc to PDFs from seven different structure types of monometallic

nanoparticles, and show that structures can be solved from both simulated and experimental PDFs, including PDFs from nanoparticles that are not present in the training distribution. We also apply DeepStruc to a system of *hcp*, *fcc* and stacking faulted nanoparticles, where DeepStruc recognizes stacking faulted nanoparticles as an interpolation between *hcp* and *fcc* nanoparticles and is able to solve stacking faulted structures from PDFs. Our findings suggests that DeepStruc is a step towards a general approach for structure solution of nanomaterials.

Introduction

Crystallographic methods, such as single crystal and powder diffraction, have been foundational in the development of functional materials over the past century. They yield atomic-scale structural models for crystalline materials and allow establishing the links between material structure and properties that are at the heart of materials development.[1,2] However, other approaches for structure determination are needed for nanostructured materials that have limited long-range order, and total scattering methods such as atomic pair distribution function (PDF) analysis have become increasingly important tools.[3-7] Currently, PDF analysis is mainly done by fitting a known starting model to an experimental PDF, a process known as structure refinement. Recent developments in automated modelling[8-10] have made it possible to extend the searched structural space, but identifying a model or solving a structure *de novo* from a PDF is still an enormous challenge. So far, only highly symmetrical nanostructures such as the C_{60} buckyball have been solved *ab initio* from a PDF.[11-15] Determining the structure of less symmetrical nanostructures is limited by the lost information caused by PDF peak overlap, which challenges the use of PDF for structure solution of more complicated nanomaterials.

An approach to handle the challenges due to the information barrier in PDFs is to employ supervised machine learning (ML) methods that can learn from well-known PDF-structure pairs. In this work, we use deep generative models (DGMs). DGMs are a class of ML models that can estimate the underlying data distribution from a reasonably small set of training examples.[16] A well-known use case of DGMs is in the generation of synthetic

'deep-fake' images[17,18] based on large datasets of real images. We here train our DGM to identify new structure models by training on known chemical structures. The DGM learns the relation between PDF and atomic structure, which enables it to solve a structures, based on a PDF it has not seen befeore and its learned chemical knowledge.

We apply our DGM, which we refer to as 'DeepStruc', for structural analysis of a model system of monometallic nanoparticles (MMNPs) with seven different structure types (Fig. 1a) and demonstrate the method for both simulated and experimental PDFs. DeepStruc is generative, which means that it can be used to construct structures that are not in the training set, i.e., solve a structure from a PDF. We demonstrate this capability on a dataset of face-centered cubic (*fcc*), hexagonal closed packed (*hcp*) and stacking faulted structures, where DeepStruc can recognize the stacking faulted structures as an interpolation between *fcc* and *hcp* and construct new structural models based on a PDF.

Results

Training DeepStruc to determine the structure of MMNPs from PDF data

DeepStruc, illustrated in Fig. 1a and discussed below, is a conditional variational autoencoder (CVAE). Autoencoders are a class of deep learning (DL) methods where high-dimensional inputs, such as chemical structures,[19,20] are reduced in dimensionality. The transformation into 2 or 3 dimensional vectors is achieved using an information bottleneck by an encoder neural network (NN),[19,21,22] and the resulting lower-dimensional, compressed feature space is known as the latent space. A decoder NN can reconstruct the input from these low-dimensional representations. When the latent space is regularized (smoothed) using normal distributions instead of discrete points we obtain a variational autoencoder (VAE). The VAE can be made to be dependent (conditioned) on additional information by the prior NN resulting in a CVAE.[22]

We here use MMNP structures (Fig. 1b) as input, and condition them on their simulated PDFs (Fig. 1c). The MMNP structures span seven different structure types computed using a variety of metals to emulate the variability in bond lengths in real metallic nanoparticle samples. The structure types are simple cubic (*sc*), body-centered cubic (*bcc*), face-centered cubic (*fcc*), hexagonal closed packed (*hcp*), decahedral, icosahedral, and octahedral, and all structure types have been constructed in sizes from 5 to 200 atoms. We used 3743 MMNP structures, which were split into training- (60 %), validation- (20 %) and testing-sets (20 %). A histogram of the distribution of the seven structure types are provided in section A in the Supplementary Information. During the training process (blue + green region Fig. 1a), DeepStruc learns to map the conditioning PDFs to their structures in the latent space. After the training process is complete, DeepStruc can be used on data that have not been part of the training set, which is referred to as 'inference'. Further details about the DeepStruc network can be found in the Method section.

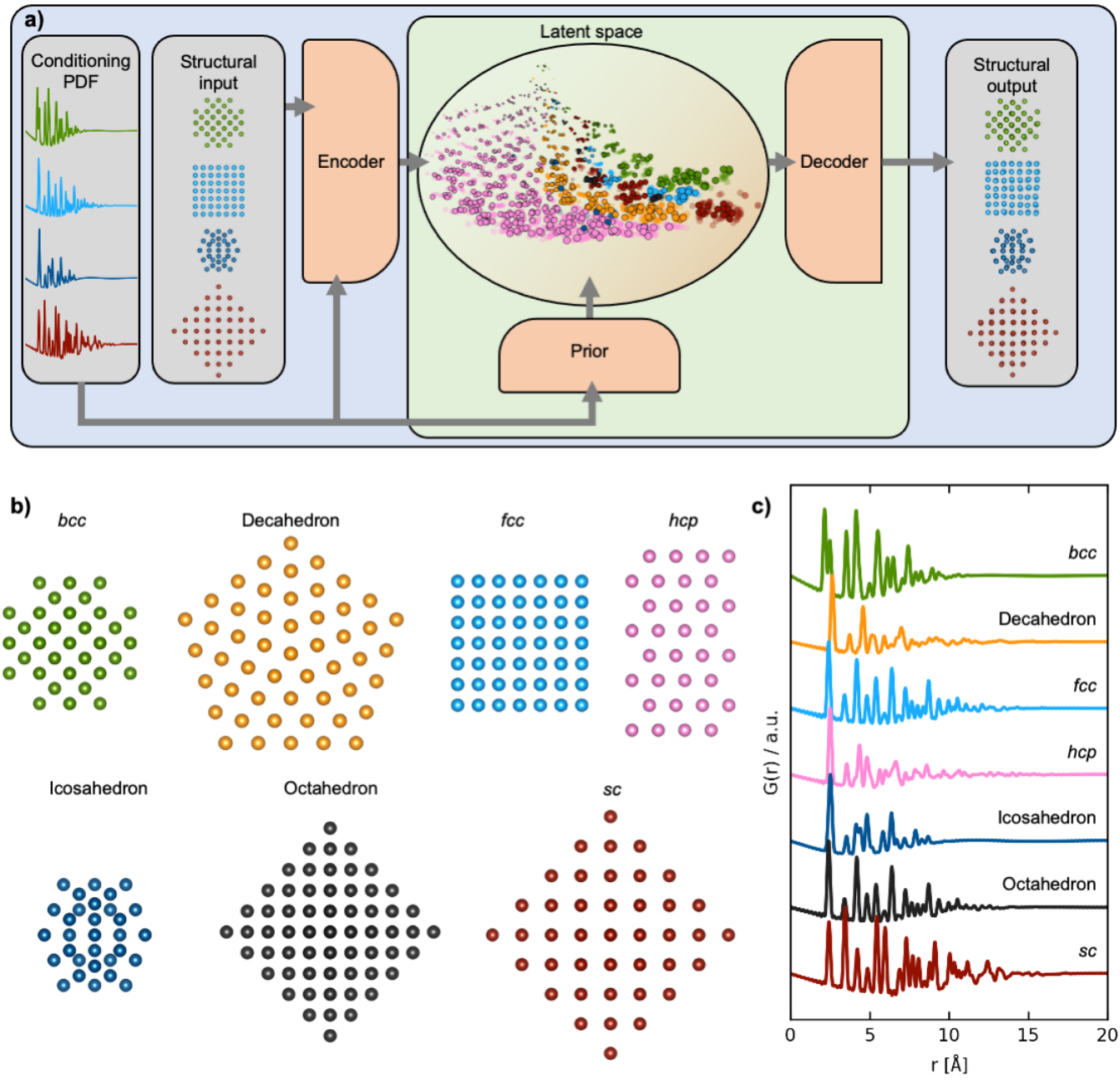

Fig. 1 | Training DeepStruc to determine the structure of MMNPs from PDFs. a) DeepStruc predicts the xyz-coordinates of the MMNP structure with conditional input provided in the form of a PDF. The encoder uses the structure and its PDF as input while the prior only takes the PDF as input. To obtain the structural output a latent space embedding is given as input to the decoder which produces the corresponding MMNP xyz-coordinates. During training of DeepStruc both the blue and green regions are used, while only the green region

is used for structure prediction during the inference process. b) Examples of the seven different structure types which are used as input to DeepStruc together with their c) simulated PDFs used as conditioning in DeepStruc. Each structure type has been included in the training set with varying sizes of 5 to 200 atoms and with varying lattice constants. The 3743 structures were split into training- (60 %), validation- (20 %), and testing sets (20 %).

Mapping of structures in a latent space

We first evaluate DeepStruc's ability to map the MMNP structures in a low-dimensional latent space by investigating structural trends and clustering. Fig. 2 shows a visualization of the two-dimensional latent space with selected MMNP reconstructions indicated. The colour of the points indicates the structure type, and the relative point size indicates the size of the MMNP cluster. We observe that DeepStruc learns to map the chemical structures in the latent space by size and symmetry. It maps the cubic structure types (*sc*, *bcc,* and *fcc*) together, and it learns that the octahedral MMNPs are closely related to the *fcc* structure type. Interestingly, DeepStruc also allocates the decahedral structures to be in between the *fcc* and *hcp* structures. This can be rationalized by considering that decahedral structures are constructed from five tetrahedrally shaped *fcc* crystals which are separated by {111} twin boundaries that resemble stacking faults.[9,23,24] The twin boundaries will resemble stacking faulted regions of *fcc* justifying that they exist in the latent space between *fcc* and *hcp*.

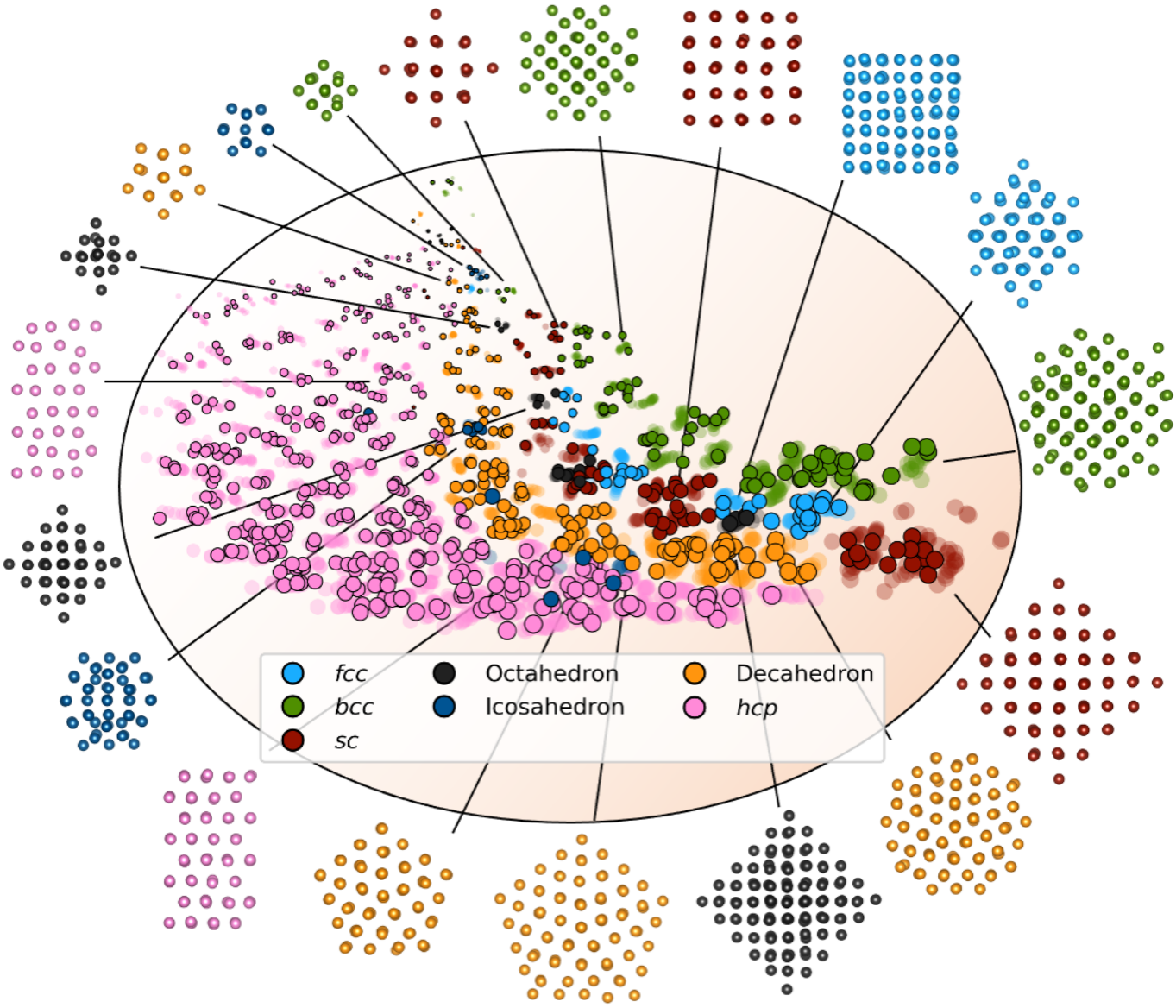

Fig. 2 | The two-dimensional latent space with structure reconstructions. The points in the latent space correspond to a structure and its simulated PDF. Data points from the test set are shown in solid colour and outlined. The points from the training and validation sets are shown as semi-transparent. The size of the points relates to the size of the embedded MMNP, and the orange background indicates the general size increase throughout the latent space. The colour of each point resemblances its structure type, *fcc* (light blue), octahedral (dark grey), decahedral (orange), *bcc* (green), icosahedral (dark blue), *hcp* (pink), and *sc* (red). Note that the test set structures shown here are the predicted structures from DeepStruc obtained during inference.

DeepStruc for structure determination from PDF

We now move on to identify structures directly from a PDF. The results of using DeepStruc on seven simulated PDFs of MMNPs not used in the training process are illustrated in Fig 3. Here, we show the structure that the input PDF was calculated from (left), the reconstructed structure (right), and its agreement with the input PDF after structure refinement (middle, discussed below). In all seven cases, the structures are correctly reconstructed from the PDF input. Before structure refinement, the mean absolute error (MAE) of the atom positions is 0.128 ± 0.073 Å as described in section B in the Supplementary Information. However, the MAE is artificially high due to a common aberration by DeepStruc, where it predicts the right geometric atomic arrangement, but isotropically contracted or expanded compared to the original structure. After refining the structure to the PDF[25] by fitting a contraction/expansion factor, a scale factor and an isotropic atomic displacement parameter (ADP), as described in section B in the Supplementary Information, the MAE of the atom positions is reduced to 0.093 ± 0.058 Å. The inference is thus robust against moderate changes in lattice parameter between a provided PDF and the structures that DeepStruc were trained on. The reconstructed structures exhibit some artificial positional atomic disorder that broadens the PDF peaks. The fitted ADP values (section B in the Supplementary Information) are thus lower than the ADP values of the conditioning PDFs.

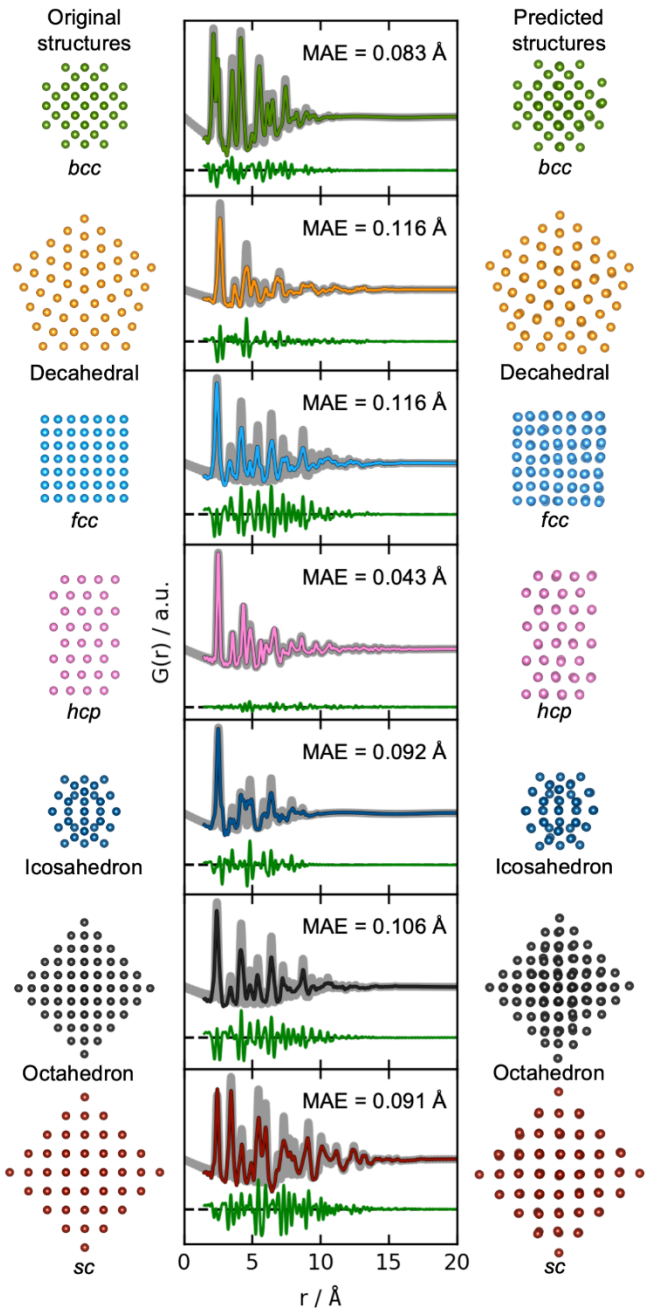

Fig. 3 | Structure determination from PDFs. Simulated PDFs (grey) from the original structures of the seven different structure types (left) are used during inference for structure prediction (right). The middle column shows the fitted PDFs of the predicted structures to the simulated PDFs of the original structures. Only the scale-factor, contraction/expansion-factor, and ADP are refined, see section B in the Supplementary Information.

Having established that DeepStruc works for structures highly resembling those in the training set, we now consider more challenging cases and explore the capabilities of DeepStruc on data which is far from the training distribution. As described above, the largest structures in the training set contained only 200 atoms. We now evaluate it on a test set of simulated MMNPs with 5 to 1000 atoms, i.e., containing much larger particles. The latent space obtained from this new test set is plotted using diamond markers in Fig. 4, where the latent space from the training process is shown with semi-transparent markers. We observe that the trends in the training area are comparable for the training set and the test set of larger MMNPs. Notably, the trends of both the size and the structure types continue beyond the training area to structures containing about 400 atoms. Beyond 400 atoms, all structure types collapse onto a line, however a size estimate of the structure can still be obtained from DeepStruc. Of course, DeepStruc could be retrained on a larger training set if reconstructions are desired on clusters larger than 200 atoms. However, this test shows that DeepStruc can extrapolate significantly in the latent space. It can thereby give useful information about PDFs from structures not represented in the training set and is generative in a meaningful way. This can be compared to, for example, a tree-based ML-classifier, which is limited to a predefined structural database and cannot extrapolate. The capability of DeepStruc to extrapolate arises from each structure in the latent space being predicted as a normal distribution instead of a discrete point. We have previously demonstrated that VAEs can do a better job interpolating in the latent space compared to deterministic AEs.[19]

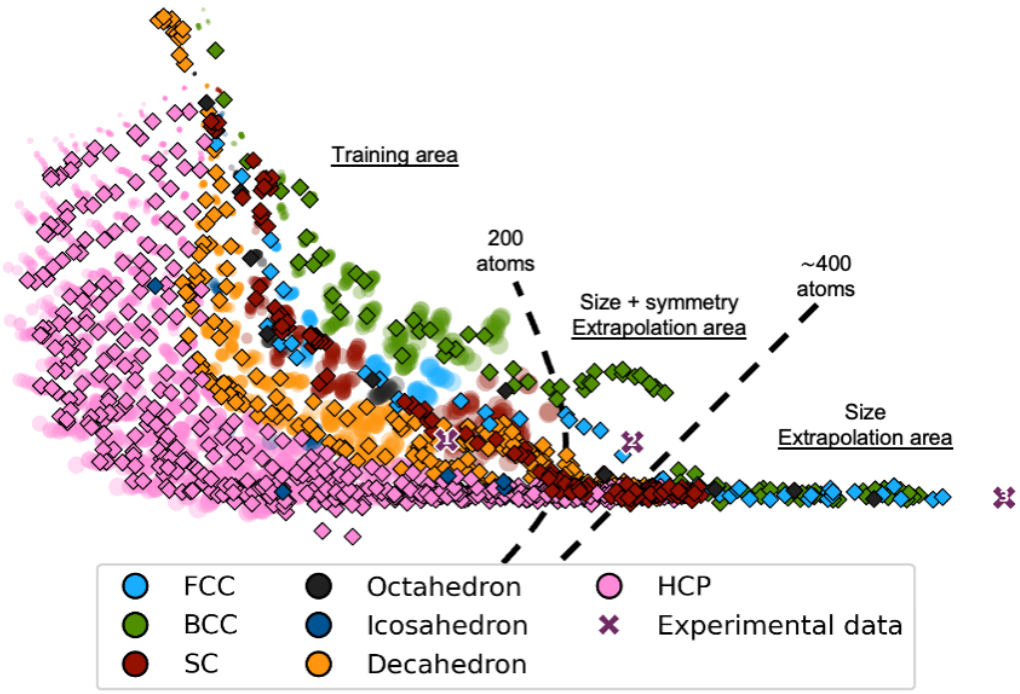

Fig. 4 | DeepStruc applied on PDFs of structures up to 1000 atoms. Each point is coloured after its structure type, i.e. *fcc* (light blue), octahedral (dark grey), decahedral (orange), *bcc* (green), icosahedral (dark blue), *hcp* (pink), and *sc* (red). Each point in the latent space corresponds to a structure based on its simulated PDF. Test PDFs from structures up to 1000 atoms are plotted as diamond markers on top of the training and validation data which are made semi-transparent. Note that the training set latent space is identical to that plotted in Fig. 2. DeepStruc has only been trained on structures up to 200 atoms. Three experimental PDFs (shown in section C in the Supplementary Information) obtained from differently sized *fcc* nanocrystals estimated to contain 203 (cross marker 1), 371 (cross marker 2), and 1368 (cross marker 3) atoms are illustrated as purple cross markers in the latent space.

In practice, DeepStruc must be able to yield valid reconstructed structures from experimental data that contain noise and other aberrations. We therefore use DeepStruc to infer structures from previously published experimental PDFs from MMNPs. Fig. 5a shows the latent space with the predicted location of structures from

three experimental PDFs. Here, the location in the latent space is represented as distributions rather than as discrete points, and multiple structures are sampled from each distribution and compared to the experimental PDF to select the best candidate. The mean of the experimental PDF distributions is represented as a black diamond with three ellipsoids indicating different confidence intervals with σ: 3, 5 and 7, where σ is the standard deviation of the normal distribution.

The first experimental dataset that we evaluate was published by Jensen et al.,[26] who identified a decahedral structure as the core motif of Au_{144}(*p*-MBA)$_{60}$ nanoparticles. DeepStruc locates the Au_{144}(*p*-MBA)$_{60}$ PDF (Fig. 5b) in a decahedral region (orange distributions in Fig. 5a) in the latent space. Given the generative capabilities of DeepStruc, in theory, we can sample an unlimited number of structures for a given PDF. As described in section D of the Supplementary Information, we here sampled up to 1000 structures from the three normal distributions (σ: 3, 5, and 7), and compared their fit to the experimental PDF. Fig. 5b shows the fit of the best structural prediction, which was among the structures sampled from the σ: 3 distributions. DeepStruc predicts a decahedral structure, which agrees well with the literature.[26] Other structures sampled from the three distributions are shown in Section E of the Supplementary Information, where we also compared the DeepStruc analysis to two baseline methods, a brute-force structure-mining method, and a tree-based ML classifier.

The second dataset that we evaluate, published by Quinson et al.,[27] are from 1.8 nm Pt nanoparticles with the *fcc* structure (described further in Section C in the Supplementary Information). This size corresponds to ca. 203 atoms, i.e. the number of atoms in the particle goes slightly beyond the *fcc* structures in the training set that contain only 165 atoms.[27] The location of the predicted mean is again shown as a black diamond in Fig. 5a, enclosed by three blue ellipsoids illustrating different magnitudes of standard deviation. The mean of the predicted structure is placed near the largest *sc* structures. If DeepStruc only favoured symmetry it would be placed directly on the *fcc* structures. Interestingly, DeepStruc does not purely favour size either, as it does not position the PDF near the largest structures which are *hcp* structures of 200 atoms. Instead, we observe that

DeepStruc takes both symmetry and size into account by placing the mean predicted structure adjacent to the largest *sc* structures containing 185 atoms. To identify the structure from the experimental PDF, we again sample 1000 structures from the σ: 3, 5 and 7 distributions. When fitting these sampled structures to the dataset, we obtain the best fit from an *fcc* structure of 146 atoms that is visualized in Fig. 5c and which agrees with the baseline models (section E in the Supplementary Information). DeepStruc thus identifies an *fcc* structure even though the size of the MMNP is outside the training set distribution.

We also attempted to input PDFs from even larger *fcc* nanoparticles, estimated to have diameters of 2.2 and 3.4 nm, corresponding to 371 and 1368 atoms, respectively (section C in the Supplementary Information).[27] Their positions in the latent space are shown in Fig. 4 along with the 1.8 nm *fcc* nanoparticles using cross markers labelled 1, 2, and 3 for increasing size. We observe that they follow the trend of the simulated *fcc* structures discussed above: while it is possible to estimate both size and symmetry for the 2.2 nm particles through extrapolation, DeepStruc can only estimate size for the 3.4 nm particle. Overall, the ability of DeepStruc to predict on experimental data for structures beyond those in the training set is promising for structure solution from PDF.

While DeepStruc only has been trained on simple MMNPs, we finally evaluate it on a PDF from $Au_{144}(PET)_{60}$ nanoparticles, consisting of an icosahedral core of 54 atoms surrounded by a rhombicosidodecahedron shell of 60 atoms (Fig. 5d and e).[26,28] We show the predicted mean position of the structure with a black diamond enclosed by pink ellipsoids. DeepStruc positions the PDF in the *hcp* region of the latent space, and when sampling 1000 structures from the distribution with σ: 7, the best fitting structures is an *hcp* structure with 40 atoms for the $Au_{144}(PET)_{60}$ nanoparticle (Fig. 5d). Similar structures are found when sampling from the σ: 3 and σ: 5 distributions. However, the PDF fit reveals that the reconstructed structure does not capture all peaks in the experimental PDF. When considering further the latent space, icosahedral structures are strongly

underrepresented in our dataset (section A in the Supplementary Information) which results in an inconsistency when placing icosahedral structures in the latent space. DeepStruc is thus challenged when solving the icosahedral core structure of the nanoparticle. However, we observe that one of the test icosahedral structures is placed near the experimental PDF in latent space within the σ: 5 distribution. Therefore, we again try to sample 1000 structures by moving the mean of the σ: 3 distribution to the nearest cluster of icosahedral structures in the latent space, which are located right outside the σ: 7 distribution. The best fitting structure (Fig. 5e) captures all main peaks of the experimental PDF. Strategies for sampling of underrepresented structures is discussed further in section D in the Supplementary Information.

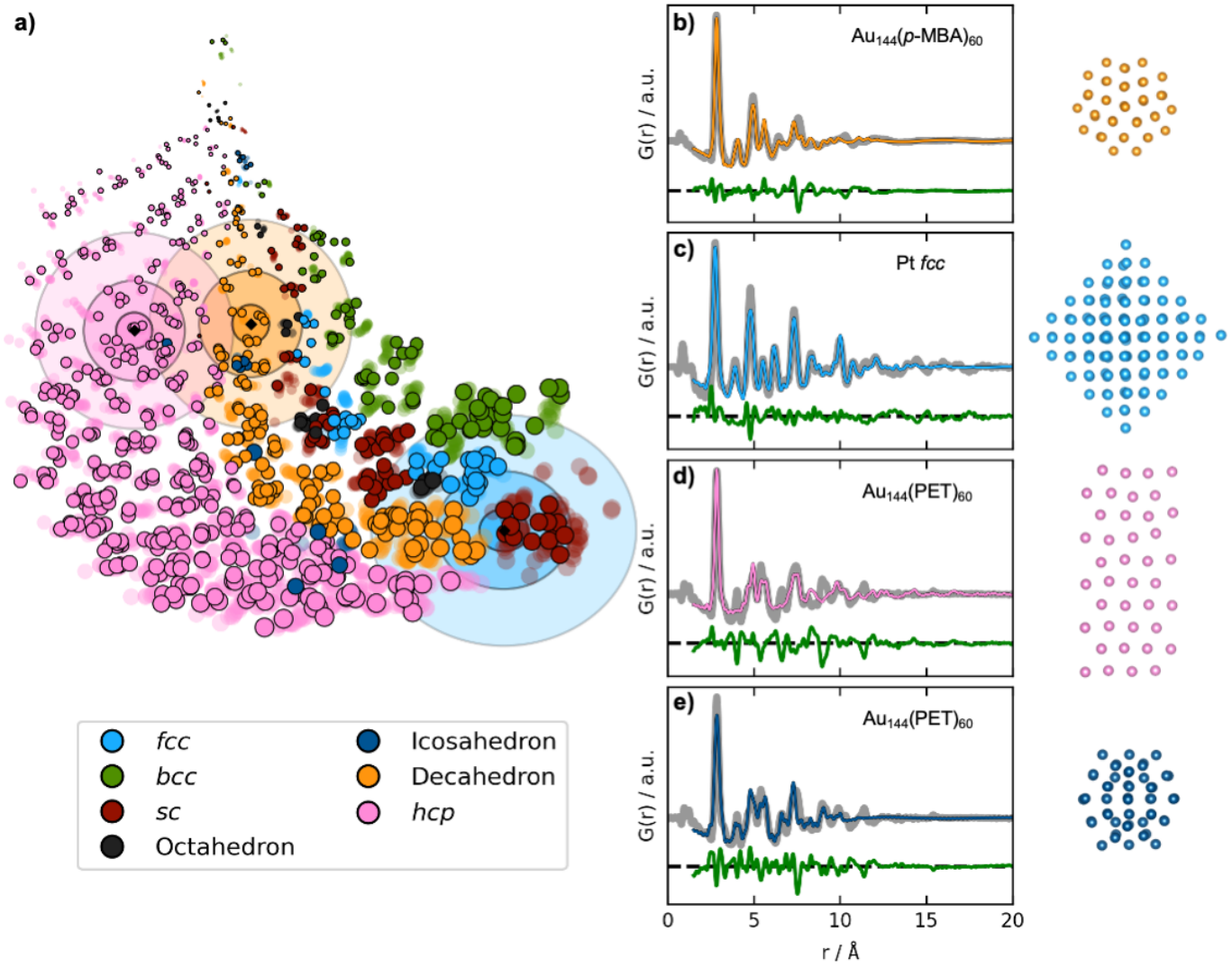

Fig. 5 | Fitting experimental PDFs with structures obtained by DeepStruc. a) The DeepStruc latent space showing predicted latent space positions for structures from three experimental PDFs. The predicted means are shown as diamond markers, which are enclosed by three rings, indicating the sampling regions for σ: 3, 5, and 7. b) PDF fit of the reconstructed structure from the Au_{144}(*p*-MBA)$_{60}$ PDF[26] c) PDF fit of the reconstructed structure from the 1.8 nm Pt nanoparticle PDF from Quinson et al.[27], d) PDF fit of the reconstructed structure from the $Au_{144}(PET)_{60}$ PDF[26] using a *hcp* structure. e) PDF fit of the reconstructed structure from the $Au_{144}(PET)_{60}$ PDF[26] using an icosahedral structure. Note that the test set structures shown here are the predicted structures from DeepStruc obtained during inference on experimental PDFs.

Structure determination from PDF: fcc, hcp, and stacking faulted nanoparticles

To obtain a deeper understanding of the latent space's behaviour, we investigate a simpler dataset only containing *fcc*, *hcp,* and stacking faulted structures. *Fcc* and *hcp* structures are distinguished by the stacking sequence of closed packed layers in their structures: while *fcc* structures can be described by ABCABC stacking, *hcp* structures has ABABAB stacking. Structures with other sequences are stacking faulted structures. We hypothesize that stacking faulted structures can be considered an 'interpolation' in the discrete space between the *fcc* and *hcp* structure type.[29]

Examples of reconstructed *fcc* (blue), *hcp* (pink), and different stacking faulted structures (purple) and their position in the new latent space are illustrated in Fig. 6a. The MMNPs cluster in size, whilst we also observe that *fcc* and *hcp* structures separate in the latent space. It is evident that the stacking faulted structures are located in between the *fcc* and *hcp* structures in the latent space as hypothesized. It is chemically reasonable that they are positioned in this exact order based on their similarity to *fcc* and *hcp*. For example, the structure with ABCABA layers, shown in Fig. 6 with a purple star is structurally close *fcc.* We see that it is also located closer to the *fcc* structures in the latent space. On the other hand, the structure with ABCBCB layers (marked as a purple diamond in Fig. 6) can be considered structurally more closely related to *hcp* than *fcc*. DeepStruc places this structure adjacent to *hcp* structures of the same size in the latent space. DeepStruc can thus insert stacking faulted structures between *fcc* and *hcp* into the latent space in a chemically meaningful way.

Fig. 6b illustrates the fits of the reconstructed structures to the PDF data. The difference curves indicate that the predicted and true structures are very close to being identical, which is supported by the MAE (section G in the Supplementary Information). While disorder causes a broadening of the peaks, the disorder in the generated structures is minor and structures with distinct difference between the layers and in the correct sequence can be reconstructed to a satisfying degree. This is a promising result, showing that a CVAE can be used as a tool to

determine the structure of stacking faulted nanoparticles from PDFs,[30,31] which is a topic of significant current interest.[32-36]

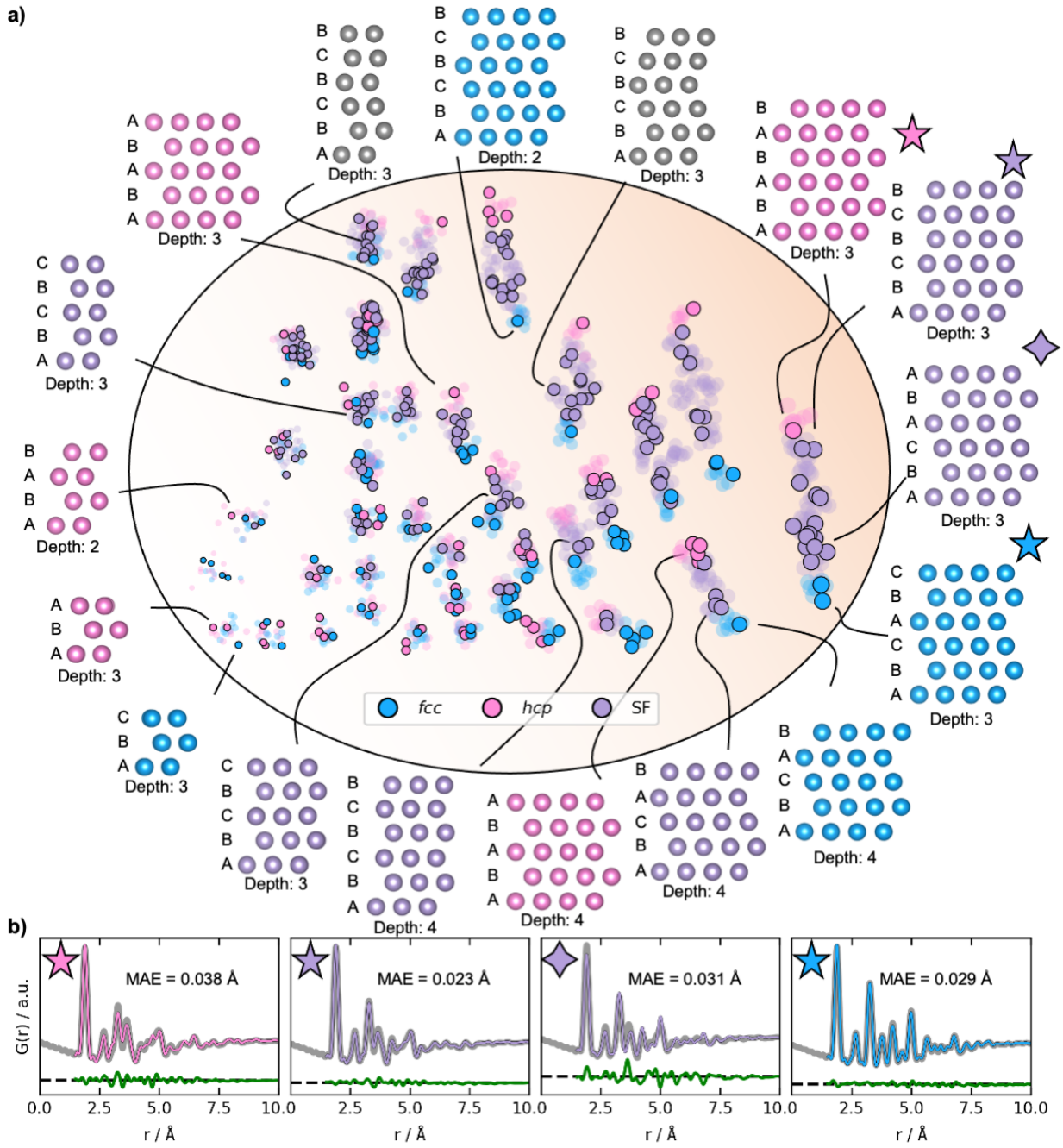

Fig. 6 | Latent space and reconstructions of stacking faulted nanoparticles. a) The latent space and reconstructed structures shown with their stacking sequence. The structures are shown in two dimensions, and

the size (number of atoms) in the third dimension is given as 'depth'. The semi-transparent dots in the latent space represent the training and validation data, and the solid dots represent the test data. *Fcc* structures are plotted in blue, *hcp* in pink, and the stacking faulted structures in purple. The marker size represents the size of the structures. B) Fits from reconstructed structures from the test PDF from a *fcc* (ABCABC stacking), a *hcp* (ABABAB stacking), and two stacking faulted structures. The original conditioning PDFs are shown in grey, while the PDFs of the generated structures are coloured according to their structure type. The difference curves are shown in green. The latent space is two-dimensional, hence allowing it to be directly visualized. Note that the test set structures shown here are the predicted structures obtained from DeepStruc during inference.

Discussion

We have shown the potential of using a DGM for structure determination from simulated and experimental PDFs. Our CVAE algorithm DeepStruc provides valuable information through its latent space, as the MMNP structures cluster based on symmetry and size in agreement with their structural chemistry. Using experimental data, the Au_{144}(*p*-MBA)$_{60}$ nanoparticle was determined to be decahedral, Pt nanoparticles were determined to be *fcc* and the $Au_{144}(PET)_{60}$ was determined to have an icosahedral core structure, all in agreement with previous literature. Our approach is only restricted by the distribution of the structural training set. When DeepStruc is trained on *fcc*, *hcp,* and stacking faulted structures, it will locate the stacking faulted structures in between the *fcc* and *hcp* structures. This suggests a strategy for training DeepStruc models on different chemical systems that also 'interpolate' from one to another when this can be identified. DeepStruc does not yet provide a completely general structure solution approach but gives critical insight into how DGMs can interact with structural and diffraction information to yield candidate structures and ultimately structure solutions.

We plan to implement DeepStruc as part of PDF-in-the-cloud (PDFitc.org),[37] where the training data can gradually be expanded over time. Combining the PDF conditioning with data from complimentary techniques

could prove important for structure determination of more complex systems. Such studies would both enable structure determination from a combined modelling perspective, but it would also reveal fundamental aspects of the information content of the different datasets for solving structure problems.

References

1 David, W. I. F. & Shankland, K. Structure determination from powder diffraction data. *Acta Crystallogr. A* **64**, 52-64 (2008).

2 Cheetham, A. K. & Goodwin, A. L. Crystallography with powders. *Nat. Mater.* **13**, 760-762 (2014).

3 Billinge, S. J. L. & Kanatzidis, M. G. Beyond crystallography: the study of disorder, nanocrystallinity and crystallographically challenged materials with pair distribution functions. *Chem. Commun.*, 749-760 (2004).

4 Young, C. A. & Goodwin, A. L. Applications of pair distribution function methods to contemporary problems in materials chemistry. *J. Mater. Chem.* **21**, 6464-6476 (2011).

5 Christiansen, T. L., Cooper, S. R. & Jensen, K. M. Ø. There's no place like real-space: elucidating size-dependent atomic structure of nanomaterials using pair distribution function analysis. *Nanoscale Adv.* **2**, 2234-2254 (2020).

6 Zhu, H., Huang, Y., Ren, J., Zhang, B., Ke, Y., Jen, A. K.-Y., Zhang, Q., Wang, X.-L. & Liu, Q. Bridging Structural Inhomogeneity to Functionality: Pair Distribution Function Methods for Functional Materials Development. *Adv. Sci.* **8**, 2003534 (2021).

7 Billinge, S. J. L. & Levin, I. The problem with determining atomic structure at the nanoscale. *Science* **316**, 561-565 (2007).

8 Yang, L., Juhas, P., Terban, M. W., Tucker, M. G. & Billinge, S. J. L. Structure-mining: screening structure models by automated fitting to the atomic pair distribution function over large numbers of models. *Acta Crystallogr. A* **76**, 395-409 (2020).

9 Banerjee, S., Liu, C.-H., Jensen, K. M. O., Juhas, P., Lee, J. D., Tofanelli, M., Ackerson, C. J., Murray, C. B. & Billinge, S. J. L. Cluster-mining: an approach for determining core structures of metallic nanoparticles from atomic pair distribution function data. *Acta Crystallogr. A* **76**, 24-31 (2020).

10 Christiansen, T. L., Kjær, E. T. S., Kovyakh, A., Röderen, M. L., Høj, M., Vosch, T. & Jensen, K. M. Ø. Structure analysis of supported disordered molybdenum oxides using pair distribution function analysis and automated cluster modelling. *J. Appl. Crystallogr.* **53**, 148-158 (2020).

11 Juhás, P., Cherba, D. M., Duxbury, P. M., Punch, W. F. & Billinge, S. J. L. Ab initio determination of solid-state nanostructure. *Nature* **440**, 655-658 (2006).

12 Juhás, P., Granlund, L., Duxbury, P. M., Punch, W. F. & Billinge, S. J. The Liga algorithm for ab initio determination of nanostructure. *Acta Crystallogr. A* **64**, 631-640 (2008).

13 Juhas, P., Granlund, L., Gujarathi, S. R., Duxbury, P. M. & Billinge, S. J. Crystal structure solution from experimentally determined atomic pair distribution functions. *J. Appl. Crystallogr.* **43**, 623-629 (2010).

14 Cliffe, M. J., Dove, M. T., Drabold, D. & Goodwin, A. L. Structure determination of disordered materials from diffraction data. *Phys. Rev. Lett.* **104**, 125501 (2010).

15 Cliffe, M. J. & Goodwin, A. L. Nanostructure determination from the pair distribution function: a parametric study of the INVERT approach. *J. Phys.: Condens. Matter* **25**, 454218 (2013).

16 Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., WooPark, C., Choudhary, A., Agrawal, A., Billinge, S. J. L., Holm, E., Ong, S. P. & Wolverton, C. Recent Advances and Applications of Deep Learning Methods in Materials Science. *arXiv*, 2110.14820 (2021).

17 Razavi, A., Van den Oord, A. & Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *Adv. Neural Inf. Process. Syst.* **32** (2019).

18 Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. & Aila, T. Analyzing and improving the image quality of stylegan. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110-8119 (2020).

19 Anker, A. S., Kjær, E. T. S., Dam, E. B., Billinge, S. J. L., Jensen, K. M. Ø. & Selvan, R. Characterising the Atomic Structure of Mono-Metallic Nanoparticles from X-Ray Scattering Data Using Conditional Generative Models. *Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG)* (2020).
20 Lim, J., Ryu, S., Kim, J. W. & Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J. Cheminformatics* **10**, 1-9 (2018).
21 Samarakoon, A. M., Barros, K., Li, Y. W., Eisenbach, M., Zhang, Q., Ye, F., Sharma, V., Dun, Z. L., Zhou, H., Grigera, S. A., Batista, C. D. & Tennant, D. A. Machine-learning-assisted insight into spin ice $Dy_2Ti_2O_7$. *Nat. Commun.* **11**, 892 (2020).
22 Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P. & Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **4**, 268-276 (2018).
23 Marks, L. D. Surface structure and energetics of multiply twinned particles. *Philos. Mag. A* **49**, 81-93 (1984).
24 Banerjee, S., Liu, C.-H., Lee, J. D., Kovyakh, A., Grasmik, V., Prymak, O., Koenigsmann, C., Liu, H., Wang, L., Abeykoon, A. M. M., Wong, S. S., Epple, M., Murray, C. B. & Billinge, S. J. L. Improved Models for Metallic Nanoparticle Cores from Atomic Pair Distribution Function (PDF) Analysis. *J. Phys. Chem. C* **122**, 29498-29506 (2018).
25 Juhas, P., Farrow, C. L., Yang, X., Knox, K. R. & Billinge, S. J. L. Complex modeling: a strategy and software program for combining multiple information sources to solve ill posed structure and nanostructure inverse problems. *Acta Crystallogr. A* **71**, 562-568 (2015).
26 Jensen, K. M. Ø., Juhas, P., Tofanelli, M. A., Heinecke, C. L., Vaughan, G., Ackerson, C. J. & Billinge, S. J. L. Polymorphism in magic-sized $Au_{144}(SR)_{60}$ clusters. *Nat. Commun.* **7** (2016).
27 Quinson, J., Kacenauskaite, L., Christiansen, T. L., Vosch, T., Arenz, M. & Jensen, K. M. Ø. Spatially Localized Synthesis and Structural Characterization of Platinum Nanocrystals Obtained Using UV Light. *ACS Omega* **3**, 10351-10356 (2018).
28 Yan, N., Xia, N., Liao, L., Zhu, M., Jin, F., Jin, R. & Wu, Z. Unraveling the long-pursued Au_{144} structure by x-ray crystallography. *Sci. Adv.* **4**, eaat7259 (2018).
29 Bertolotti, F., Moscheni, D., Migliori, A., Zacchini, S., Cervellino, A., Guagliardi, A. & Masciocchi, N. A total scattering Debye function analysis study of faulted Pt nanocrystals embedded in a porous matrix. *Acta Crystallogr. A* **72**, 632-644 (2016).
30 Masadeh, A. S., Bozin, E. S., Farrow, C. L., Paglia, G., Juhas, P., Billinge, S. J. L., Karkamkar, A. & Kanatzidis, M. G. Quantitative size-dependent structure and strain determination of CdSe nanoparticles using atomic pair distribution function analysis. *Phys. Rev. B* **76** (2007).
31 Yang, X., Masadeh, A. S., McBride, J. R., Božin, E. S., Rosenthal, S. J. & Billinge, S. J. L. Confirmation of disordered structure of ultrasmall CdSe nanoparticles from X-ray atomic pair distribution function analysis. *Phys. Chem. Chem. Phys.* **15**, 8480-8486 (2013).
32 Cenker, J., Sivakumar, S., Xie, K., Miller, A., Thijssen, P., Liu, Z., Dismukes, A., Fonseca, J., Anderson, E., Zhu, X., Roy, X., Xiao, D., Chu, J.-H., Cao, T. & Xu, X. Reversible strain-induced magnetic phase transition in a van der Waals magnet. *Nat. Nanotechnol.* (2022).
33 Rong, X., Liu, J., Hu, E., Liu, Y., Wang, Y., Wu, J., Yu, X., Page, K., Hu, Y.-S., Yang, W., Li, H., Yang, X.-Q., Chen, L. & Huang, X. Structure-Induced Reversible Anionic Redox Activity in Na Layered Oxide Cathode. *Joule* **2**, 125-140 (2018).

34 Charles, D. S., Feygenson, M., Page, K., Neuefeind, J., Xu, W. & Teng, X. Structural water engaged disordered vanadium oxide nanosheets for high capacity aqueous potassium-ion storage. *Nat. Commun.* **8**, 15520 (2017).
35 Gao, P., Metz, P., Hey, T., Gong, Y., Liu, D., Edwards, D. D., Howe, J. Y., Huang, R. & Misture, S. T. The critical role of point defects in improving the specific capacitance of δ-MnO_2 nanosheets. *Nat. Commun.* **8**, 14559 (2017).
36 Metz, P. C., Koch, R. & Misture, S. T. Differential evolution and Markov chain Monte Carlo analyses of layer disorder in nanosheet ensembles using total scattering. *J. Appl. Crystallogr.* **51**, 1437-1444 (2018).
37 Yang, L., Culbertson, E. A., Thomas, N. K., Vuong, H. T., Kjaer, E. T. S., Jensen, K. M. O., Tucker, M. G. & Billinge, S. J. L. A cloud platform for atomic pair distribution function analysis: PDFitc. *Acta Crystallogr. A* **77**, 2-6 (2021).
38 Egami, T. & Billinge, S. J. L. *Underneath the Bragg Peaks, Pergamon* (2012).
39 Hjorth Larsen, A., Jørgen Mortensen, J., Blomqvist, J., Castelli, I. E., Christensen, R., Dułak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., Hermes, E. D., Jennings, P. C., Bjerre Jensen, P., Kermode, J., Kitchin, J. R., Leonhard Kolsbjerg, E., Kubal, J., Kaasbjerg, K., Lysgaard, S., Bergmann Maronsson, J., Maxson, T., Olsen, T., Pastewka, L., Peterson, A., Rostgaard, C., Schiøtz, J., Schütt, O., Strange, M., Thygesen, K. S., Vegge, T., Vilhelmsen, L., Walter, M., Zeng, Z. & Jacobsen, K. W. The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* **29**, 273002 (2017).
40 Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: going beyond euclidean data. *IEEE Signal Process. Mag.* **34**, 18-42 (2017).
41 Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* **20**, 61-80 (2008).
42 Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
43 Sohn, K., Lee, H. & Yan, X. Learning structured output representation using deep conditional generative models. *Adv. Neural Inf. Process. Syst.* **28**, 3483-3491 (2015).
44 Duxbury, P. M., Granlund, L., Gujarathi, S., Juhas, P. & Billinge, S. J. The unassigned distance geometry problem. *Discret. Appl. Math.* **204**, 117-132 (2016).
45 Shao, H., Xiao, Z., Yao, S., Sun, D., Zhang, A., Liu, S., Wang, T., Li, J. & Abdelzaher, T. ControlVAE: Tuning, Analytical Properties, and Performance Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).
46 Rydhmer, K. & Selvan, R. Dynamic beta-VAEs for quantifying biodiversity by clustering optically recorded insect signals. *arXiv preprint arXiv:2102.05526* (2021).
47 Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

Method

In the following sections, we briefly explain what a PDF is, how we obtained the simulated PDFs and their structures, and finally we elaborate on the CVAE method developed here to analyse PDFs. A more detailed description of the PDF is given elsewhere.[38]

The Pair Distribution Function (PDF)

The PDF is the Fourier transform of total scattering data, which can be obtained through x-ray, neutron, or electron scattering. In this work we focus on the usage of x-ray total scattering data. The scattering vector Q is defined as follows, where λ is the radiation wavelength, and θ is the scattering angle:

$$Q = \frac{4\pi sin(\theta)}{\lambda}$$

The measured scattering intensities are denoted I(Q), which are corrected for incoherent scattering, fluorescence, etc. and normalized such that the total scattering structure function S(Q) is obtained.

$$S(Q) = \frac{I(Q) - \langle f(Q)^2 \rangle + \langle f(Q) \rangle^2}{\langle f(Q) \rangle^2}$$

Here f is the atomic form factor. To obtain the structural real-space information, the total scattering structure function is Fourier transformed over the truncated Q-range, hence yielding the reduced PDF also known as G(r):

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

G(r) can be interpreted as a histogram of real-space interatomic distances and the information is equivalent to that of an unassigned distance matrix (uDM). Simulated PDFs are shown in Fig. 1b and all simulation parameters can be found in section H in the Supplementary Information. The PDFs used in this project are normalised to have I(G(r)) = 1 as illustrated in section I in the Supplementary Information.

Simulated and experimental data

To simulate the nanoparticles used in the training process of DeepStruc, the Python library atomic simulation environment (ASE) was used.[39] The seven different structure types: *fcc*, *bcc*, *sc*, *hcp*, icosahedral, decahedral, and octahedral were constructed with the cluster module in ASE in the same manner as described by Banerjee et al.[9] and Anker & Kjær et al.[19] All MMNPs were generated in sizes ranging from 5 to 200 atoms. Each MMNP was then populated with different atoms hence changing the lattice spacing/bond distances in the MMNP. To ensure that there were no duplicate MMNPs within the dataset, all MMNPs were decomposed into a distance list of all atom-atom distances. The distance lists are a reduced format of the xyz representation as they are rotation- and translation-invariant in Euclidean space. All the distance-lists were sorted and duplicate structures with equivalent distance lists were removed. This yielded a total of 3742 unique MMNPs, see section A in the Supplementary Information for the distribution of the seven structure types. The xyz-coordinates will be the label that DeepStruc has to reconstruct. Nanoparticles with each of the seven structure types can be seen in Fig. 1b along with their simulated PDF, Fig. 1a. All the simulation parameters used can been seen in section H in the Supplementary Information.

To further investigate the latent space behaviour of DeepStruc, a more chemically simple and intuitive dataset was made of *fcc*, *hcp,* and stacking faulted structures. *Fcc* and *hcp* can be considered layered structures that are only differentiated by the repetition of layers within the structure. *Fcc* consists of a repeated ABCABC layered structure where *hcp* is an ABABAB layered structure. A 5 layered stacking fault structure could then be described as ABCAC, as it does not satisfy either of the *fcc* or *hcp* stacking criteria, see Fig. 6. A total of 1620 stacking fault structures were generated.

Data representation

In this work, the structures from ASE are converted into a graph-based representation in order to capture the interatomic relationships, as the original representation generated with ASE are not optimal as input to

DeepStruc. Graph representations have seen increasing success in machine learning applications related to materials science as the interatomic relations in graphs are invariant to transformations of the structure such as solid translations and rotations.[40,41] Each structure in graph representation can be described as G = (**X**,**A**), where $\mathbf{X} \in \mathbb{R}^{N\times F}$ is the node feature matrix which contains F features that can describe each of the N atoms in the structure. We use F = 3 comprising only the Euclidean coordinates of the atom in a 3-dimensional space. The interatomic relationships are captured using the adjacency matrix $\mathbf{A} \in R^{N\times N}$. In our case, the entries of the adjacency matrix are the Euclidean distance between each pair of atoms, resulting in a soft adjacency matrix. However, to make the adjacency matrix sparse, when the distance between any pair of nodes is larger than the lattice constant the corresponding edge weight is set to zero. When the edge weight is zero this corresponds to absence of an edge between the pair of nodes, and in other cases the edges have a weight given by the interatomic distance. Section J in the Supplementary Information shows a decahedron consisting of seven atoms alongside the components describing it in our chosen graph representation.

The Conditional Deep Generative Model (DGM)

DGMs such as variational autoencoders (VAEs) are commonly used to synthesize novel, synthetic data by approximating the underlying data-generating processes based on the training data.[42] In this work, we are interested in generating structures based on properties such as the PDF resulting in the conditional DGM scenario. The specific formulation of the conditional DGM used in this work is the CVAE, initially proposed for computer vision tasks[43] and more recently it has also been explored for synthesizing novel drug molecules.[20] The CVAE in this work is trained to solve the unassigned distance geometry problem[44] (uDGP) as it solves the task of converting the distances within a PDF to a chemical structure. In the uDGP the problem of taking a starting point of a list of distances and reconstructing it into a structure is broken down into two discrete problems. First, is to discover the graph that connects pairs of atoms, with the edges labelled by the distances from the distance

list (the assignment problem). Second is to embed this graph into Euclidian space. An illustration of the CVAE can be seen in Fig. 1a. Here, the blue area is the training process, and the green area is the prediction/inference process. During training of the CVAE, the encoder takes pairs of structures and their corresponding PDFs as input. The encoder learns to map the structure-PDF pairs into a low-dimensional, latent Gaussian distribution, known as the encoder distribution. Each structure-PDF pair is mapped to certain regions of the latent space. When trained with large amounts of diverse data, the latent space is able to capture relationships between different structures and PDF pairs so that similar structures are closer in this latent space than very different structures. CVAEs are different from classical autoencoders in that the latent space is probabilistic, which makes it possible to sample structures from these latent encoder distributions. This is achieved during training by forcing the encoder distributions to align with a simpler prior distribution which only takes the PDF as input. The two distributions are matched by minimizing the Kullback-Leibler Divergence between the encoder and prior distributions and is interpreted as the regularization term, L_{reg}.

The prior NN gets the PDF as input and maps it to the low-dimensional prior distribution. The low-dimensional latent vector conditioned on the PDF is then input to the decoder, which is tasked to predict the xyz-coordinates of the structural input. During the training process, the mean squared error (MSE) between the xyz-coordinates of the input and output are computed to force the decoder to predict xyx-coordinates from the latent representations. The MSE is defined as the reconstruction loss, L_{rec}. The CVAE is trained by jointly optimizing these two loss components:

$$L_{CVAE} = L_{rec} + \beta \cdot L_{reg}$$

where β is a scaling factor that controls the relative influence of the regularization- and reconstruction-terms. In our training process, at initialization β is set to 0 which allows the model to focus on minimizing L_{rec}. Each time L_{rec} gets below a certain threshold β is increased. This helps keep the model from falling into a local minimum and the process is repeated until convergence has been reached. Similar strategies for annealing β in VAEs have

been attempted.[45,46] At inference (test) time, the prior NN receives the PDF as input which is then mapped to the low-dimensional latent space which during training has been trained to match the encoder distribution. A sufficiently well trained CVAE is then able to predict structures from the latent space based on the PDF input. A simplified version of the CVAE used for this work, DeepStruc, can be seen in Fig. 1a. The CVAE is presented more formally in our earlier work.[19]

Graph Conditional Variational Autoencoder (CVAE)

In this work, two types of CVAEs were utilized depending on the type of encoder. In the conventional CVAE, the encoder was based on Multi-Layered Perceptrons which operate on a tabular format of the node features, and the adjacency matrix populated with atom–atom distances. For the second type of CVAE – that we call the graph CVAE – the encoder consists of a graph neural network (GNN)[41,47] and is able to process graph structured data, taking the neighbourhood information into consideration. GNNs are generalized message passing methods that can aggregate information from the neighbourhood of a node by passing messages along the edges. These messages are learned during training and can summarize the information present at the node necessary for the downstream tasks. Further, by making the encoder deep, i.e. adding additional GNN layers, nodes can get access to information from nodes that are farther from them. For instance, in a k-layered GNN each node had access to information from nodes that are k-hops away. In our experiments, we observed that the generative capabilities of the graph CVAE was better than the conventional CVAE, part E in the Supplementary Information. Further, we were able to obtain comparable reconstruction quality from the graph CVAE with only two latent dimensions compared to using eight dimensions for the conventional CVAE. This indicates that the graph encoder is able to better compress the information present in the node and adjacency matrices. A minor technical detail in our CVAE models is that the predictions from the decoder do not exactly match the input features. That is, the

decoder does not reconstruct the full input comprising node features and adjacency matrix but only the node features. The algorithm we refer to as DeepStruc refers to the graph based CVAE.

Data availability

Code for the baseline models and DeepStruc is available at:

https://github.com/EmilSkaaning/DeepStruc

https://github.com/AndyNano/Brute-force-PDF-modelling

https://github.com/AndyNano/MetalFinder

https://github.com/AndyNano/CVAE

Acknowledgements

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). We are grateful to the Villum Foundation for financial support through a Villum Young Investigator grant (VKR00015416). Funding from the Danish Ministry of Higher Education and Science through the SMART Lighthouse is gratefully acknowledged. We acknowledge support from the Danish National Research Foundation Center for High Entropy Alloy Catalysis (DNRF 149). Work in the Billinge group was supported by the U.S. National Science Foundation through grant DMREF-1922234.

Author contributions

ETSK and ASA contributed to all aspects of the paper. MNW wrote the code associated to the tree-based classifier. SJLB, RS and KMØJ supervised the project. All authors contributed to the writing of the manuscript.

Competing interests

The authors declare no competing interests.