A Machine-Learning-Based Approach for Solving Atomic Structures of Nanomaterials Combining Pair Distribution Functions with Density Functional Theory

Magnus Kløve, Sanna Sommer, Bo B. Iversen,* Bjørk Hammer,* and Wilke Dononelli*

Determination of crystal structures of nanocrystalline or amorphous compounds is a great challenge in solid-state chemistry and physics. Pair distribution function (PDF) analysis of X-ray or neutron total scattering data has proven to be a key element in tackling this challenge. However, in most cases, a reliable structural motif is needed as a starting configuration for structure refinements. Here, an algorithm that is able to determine the crystal structure of an unknown compound by means of an on-the-fly trained machine learning model, which combines density functional theory calculations with comparison of calculated and measured PDFs for global optimization in an artificial landscape, is presented. Due to the nature of this landscape, even metastable configurations and stacking disorders can be identified.

1. Introduction

Characterization of functional materials at the atomic level plays a crucial role in the advancement of modern society, as the properties of functional materials are intimately related to their atomic structure.[1–3]

In this regard, X-ray and neutron scattering techniques are indispensable tools, since they are direct probes of the atomic arrangement in solids, and the atomic structure of unknown compounds can be routinely elucidated from 3D single-crystal diffraction (SCD) data.[4] For a powdered sample, however, the 3D structural information condenses onto a 1D diffractogram causing significant peak overlap and loss of directional information. Still, phase retrieval procedures such as the Patterson method and Direct Methods have been applied for solving atomic structures from powder X-ray diffraction (PXRD) experiments analogous to structure solution from SCD data.[5–8] Global optimization techniques such as genetic algorithms,[9,10] Monte-Carlo optimization,[11–14] and simulated annealing[11,15] attempt direct-space atomic structure solution from agreement with experimental PXRD data utilizing raw computational power. Hybrid methods such as the works of Wolverton and co-workers on FPASS build upon the same idea of global optimization using a genetic algorithm, however, with the use of density functional theory (DFT) energies to identify low energy structural motifs during the search followed by an evaluation of the computed PXRD.[16,17] In all cases, however, structural solution from PXRD typically requires a homogenous and well-crystalline sample together with high-quality data.

Real, functional materials are not necessarily well-crystalline, but are likely to be nanostructured, and so the traditional crystallographic approaches to structure solution fall short. For these materials, the concept of an infinitely repeating unit cell breaks down. This results in peak broadening and thus even more severe peak overlap in the 1D diffractograms, yet again reducing the amount of unique information available.[18,19]

The pair distribution function (PDF) analysis of X-ray or neutron total scattering (TS) data has proven a valuable tool for studying nanomaterials.[20–23] Still, most PDF analyses are guided by a priori structural information and are carried out as real-space counterparts to Rietveld refinements. Recent efforts by the Billinge group on the cloud platform, PDFitc.org, have provided the PDF community with real-space matching software for identifying unknown phases in PDF patterns through a structural database search.[24,25] While such software can aid real-space analyses if the structure has been previously determined, determining the atomic structures of truly unknown
nanocrystalline phases from PDF data remains yet a pending task for the community in the absence of efficient and versatile procedures.

Large-box global optimization like reverse Monte-Carlo (RMC) modeling of PDFs has long been applied to study the atomic structure of amorphous materials, where no unique structural solution exists, but where the statistical weight of the ensemble can provide structural insight. However, the RMC algorithm is inherently inefficient due to its iterative and unguided, random-guessing approach, often leading to search trapping in local minima. Also for large-box optimizations of amorphous materials, hybrid approaches such as the AIFEAR modeling by Drabold et al. have been applied to overcome the potential challenges for traditional RMC. Here, RMC is combined with ab initio calculations in a sequential manner of 1000 RMC moves followed by 10 conjugate-gradient steps using ab initio calculations and iterating this scheme until convergence.

The “Liga” method originally showcased to solve the structure of single-element, non-periodic clusters like $G_{60}$ from experimental PDFs has also been extended to solve atomic structures of multiple-element crystalline samples from experimental PDFs. While successful in most cases, the algorithm has only been tested on well-crystalline powder samples, where traditional reciprocal space analysis could be more intuitively applied.

In this letter, we present an efficient machine learning (ML)-enhanced global optimization algorithm for solving atomic structures from experimental PDFs of materials on several length scales including nanomaterials obtained with only short temporal resolution. By combining the global optimization method GOFEFEE by Bisbo and Hammer with the DiffPy-CMI software for PDF modeling, the algorithm performs a global optimization in an on-the-fly, ML-trained surrogate R-factor landscape, thus solving the atomic structure from the PDF. Combining DFT calculations with PDF modeling the ML model can be trained in a combined artificial R-factor/DFT landscape to automatically determine the structure of even metastable phases correctly. In consecutive global multiphase refinements, the combined artificial R-factor/DFT landscape can be used to determine additional stacking motifs, i.e., to identify stacking faults for layered compounds.

2. Combining R-Factor Landscape and DFT Energy for Global Optimization

In the original work of Bisbo and Hammer, the GOFEFEE algorithm was used to optimize structures to their global minimum in the potential energy surface (PES) calculated using DFT. For a detailed description of the GOFEFEE method, readers are referred to the original work by Bisbo and Hammer. In this work, we want to present an alternative approach to a pure DFT search, where the global optimization is additionally guided based on agreement between a calculated and an experimentally measured PDF through the R-factor in Equation (1). A ML potential is trained on-the-fly on this comparison function and performs local optimizations of new candidate structures in this artificial (R-factor) landscape in combination with the DFT PES (see Equation (2)), eventually optimizing the structures to their global minimum after several iterations of the algorithm.

In contrast to the previously mentioned hybrid schemes, the cost function used in this combined R-factor/DFT landscape optimizes simultaneously towards structures showing high thermodynamic stability and good agreement to the experimental PDF and not in a sequential manner. This ensures the ability to find structures that might not be minimum structures in one of the two landscapes, but rather in the other.

$$R_w = \sqrt{\frac{\sum w_i (G_{i,\text{obs}} - G_{i,\text{calc}})^2}{\sum w_i G_{i,\text{obs}}^2}}$$

Equation (1) is defined as the deviation between the calculated PDF, $G_{i,\text{calc}}$, and the observed PDF, $G_{i,\text{obs}}$, at each data point $i$ with the weights, $w_i$, defined as $1/\sigma_i^2$, with $\sigma_i$ being the uncertainty of $G_{i,\text{obs}}$. In the underlying work, these R-factors are determined using the DiffPy-CMI software.

The fitness $F$ of new candidate structures within the GOFEFEE framework is calculated by combining the R-factor with prefactor $\alpha$, with the total energy of the system $E_{\text{DFT}}$ (in units of eV) calculated at DFT level of theory with prefactor $(1 - \alpha)$ as shown in Equation (2).

$$F = \alpha R_w + (1 - \alpha) E_{\text{DFT}}$$

Equation (2) allows different mixing factors for $R_w$ and the total DFT energy to be used as the fitness function in the global optimization run. The two extrema correspond to setting either $\alpha = 0$ or $\alpha = 1$. In the first case, the global optimization is only based on the DFT energy, in the latter the search is only guided based on agreement with the PDF through $R_w$. The choice of the parameter $\alpha$ is discussed in Section 3.1.

3. Results and Discussion

3.1. Finding the Crystal Structure of 6 nm Anatase Nanoparticles

To demonstrate the ability of the algorithm to find the crystal structure of an unknown compound and to analyze the influence of the prefactor $\alpha$, we globally optimize the atomic structure of 6 nm Anatase (TiO$_2$) nanoparticles from a PDF obtained with only 3 s of X-ray exposure. Each run is initialized by creating four different random structures by placing 4 Ti and 8 O atoms at random non-overlapping positions in the simulation cell. Each of the four initial candidates is then locally optimized for 10 BFGS steps. After the initialization and local optimization, 1500 GOFEFEE cycles are used to guarantee convergence of the algorithm. Details on the algorithm can be found in the Experimental and Computational Details and for details on GOFEFEE in general in the original work by Bisbo and Hammer.

Using the R-factor alone ($\alpha = 1$), the global optimization yields a best-fit structure where one oxygen atom is displaced with respect to the perfectly ordered crystalline structure (cf. Figure 1A,B). As shown in Figure 1C,D, the difference in the corresponding $R_w$ values is only very small (0.025). Nevertheless, such a small difference is enough to force the algorithm to focus on structures that show such a structural defect.
To analyze this effect in a statistical fashion, ten different instances of the same global optimization were started for four different values of $\alpha$, respectively. The results are shown in Figure 2B. GOFEE was able to determine the correct atomic arrangement, i.e., the crystal structure as shown in Figure 2A of non-defected TiO$_2$ from the PDF ($\alpha = 1$) with a success rate of 60% (cf. Figure 2B red line). The criterion for success is defined as finding a structure in the pool of candidates that measured by comparison of the sum of all bond lengths agrees with a corresponding non-defected crystal structure as shown in Figure 2A. We found a similar result for CeO$_2$ as shown in the Supporting Information, although due to the higher symmetry of the CeO$_2$ crystal structure, the success rate was found to be 80%. In contrast to the global optimization of the Anatase structure based on the R-factor ($\alpha = 1$), a global optimization based entirely on the DFT energy ($\alpha = 0$) gives a 100% success rate already after a few (54) iterations, however, without considering the agreement with experimental data (the $R_w$ value) during the search. Using values of $\alpha = 0.5$ and $\alpha = 0.8$ lead to success rates in between the ones of the two extreme cases. Note: a value of $\alpha = 0.5$ does not necessarily lead to an equal weighting of the DFT energy and the $R_w$ values.

$R_w$ ranges from 0 to 1, whereas the total energy of a system $E_{\text{DFT}}$ might differ by orders of magnitude for depending on chemical composition and atomic arrangement.

Now, the question arises why the global optimization based on the DFT energy is much more efficient than the same search based on the R-factor. The answer to this question is illustrated in Figure 3, where the landscapes for placing the...
shaded Ti atom at every position in the (100)-plane are shown for four combinations of DFT/PDF weighting. The pure DFT PES shows five maxima (blue) and just three distinct favorable minima (red). In comparison, the corresponding pure R-factor landscape shows multiple different minima. Exploration of all these minima takes much more effort than exploring the few local minima present in the DFT PES. Combining the DFT energy with the R-factor yields intermediate landscapes with fewer or at least more pronounced local minima (Figure 3) and consequently leads to success rates in between the ones of a pure DFT or R-factor-based search (Figure 2B). As a side effect, even for weighting the R-factor stronger than the DFT energy (Figure 2B, green line), the somewhat unphysical features of the pure R-factor-based search can be discarded and a structure similar to the perfect, non-defected Anatase crystal structure results.

The choice of \( \alpha \) influences the result of the structure search as shown in Figure 2. The parameter has to be chosen with an eye on the number of atoms per unit cell and the quality of the experimental measurement. A more detailed discussion can be found in the Supporting Information. Nevertheless, for statistical reasons, multiple instances of the same search should be started for any global optimization method anyway, including the method presented in this work. Therefore, different values of \( \alpha \) should be used for all instances, which makes the explicit choice of this parameter less important.

Besides the fact that the pure R-factor-based landscape (\( \alpha = 1 \)) is more complex, it is worth mentioning that the resulting \( R_w \) strongly depends on the quality of the measurement as well as the purity of the measured sample. If the material is amorphous or impure, it is almost impossible to represent the material using a single average structure model.

**3.2. Studying the Defected Crystal Structure of Gahnite ZnAl\(_2\)O\(_4\)**

The R-factor landscape of solid matter can become quite complex, as demonstrated for Anatase resulting in only a 60% success rate in a pure R-factor-based search for this rather simple crystal structure. In the case of Gahnite (ZnAl\(_2\)O\(_4\)), which crystallizes in the more complex spinel structure that features a large unit cell with both octahedrally and tetrahedrally coordinated metal ions, the R-factor landscape seems to be too complex for the GOFEE algorithm to determine a physical meaningful average structure only based on the R-factor even after many iterations of the global optimization algorithm (see Figure S2, Supporting Information). In the R-factor optimized structure, some local structural motifs of the metal ions that resemble the octahedral and tetrahedral coordination found in the spinel crystal structure are present. Yet, the overall structure is far from being a realistic representation of the crystal structure. Using a pure DFT PES (\( \alpha = 0 \)), the GOFEE algorithm yields the perfect, non-defected crystal structure as known from databases. Depending, for example, on the function used for energy evaluations (e.g., a pure GGA like PBE or a hybrid functional like PBE0), some distances and angles between different atoms might differ slightly, however, the overall motif still matches the spinel structure.

Regardless, both the R-factor and DFT optimized structures provide fits to the experimental PDFs of similar quality, and this example therefore clearly shows the limitations of using GOFEE or any other global optimization method weighted only by the R-factor, whenever the structure is more complex than a simple, highly symmetric binary compound.

If on the other hand, the R-factor and DFT energy are combined in the global search, a different structure was found showing an inversion of one Al atom with one Zn atom (Figure 4A). This is a well-known point defect in spinel structures\(^{[34,35]} \) and therefore not too surprising.

Since a local minimum structure in the DFT PES does not necessarily need to correspond exactly to a minimum in the R-factor landscape, for further comparison both the non-defected structure and the one with inversion have been refined using either local optimization in DFT or local optimization with respect to minimizing the \( R_w \) as explained in the Experimental Section.

The non-defected structure is 0.93 eV more stable than the structure showing an inversion. On the other hand, the calculated PDF of this defective structure shows 0.63% better agreement with the measured PDF than the calculated PDF of the non-defected structure (cf. Figure 4E). Therefore, although the non-defected structure might represent the thermodynamically most stable configuration, the defected structure found with the combination of DFT energy evaluation and R-factor calculation, leads to the best average representation of the studied Gahnite nanoparticles with respect to the measured PDF pattern.

In the above, we worked with a 1 × 1 × 1 cell and hence a high concentration of defects whenever these are present. In an attempt to study the occurrence of defects at lower concentrations, a global optimization of a 2 × 1 × 1 Gahnite unit cell was run and similar results were obtained. The search based on the DFT energy leads to a non-defected crystal structure, whereas the search in the combined DFT and R-factor landscape leads to a structure with an inversion of Zn and Al. Comparable to the single unit cell, the non-defected structure is 0.95 eV more stable. On the other hand, again, locally optimizing the resulting structures in the R-factor landscape leads to a 0.86% better fit of the PDF of the defective structure. This is slightly more pronounced than in the case of a single-unit
Increasing the cell size to a $3 \times 1 \times 1$ supercell, the non-defected structure is 0.96 eV more stable than the one showing one inversion comparable to a single unit cell and a $2 \times 1 \times 1$ supercell. On the other hand, the defected supercell yields now only a 0.72% better fit than the non-defected structure. Therefore, one can estimate that the concentration of point defects is likely lower than one Al/Zn inversion per unit cell, but bigger than 0.33 inversions per unit cell. Note, however, that in the case of the $3 \times 1 \times 1$ supercell the dimension of the periodically repeated cell almost reaches the size of the nanoparticle itself. Therefore, some additional effects might influence the decrease in the stability difference. In addition, it should be noted that we have considered a perfect stoichiometry in the search, whereas non-stoichiometric amounts of the three elements were used in the refinement in the original study by Sommer et al.\cite{35} from which the data originate. Using non-stoichiometric amounts of any of the three elements in our GOFEE search could potentially result in other structural motifs showing up such as the different point defects found in their work.

### 3.3. Identifying Stacking Disorder in SnS$_2$

As shown in the previous section, the PDF of a single model structure can be used to analyze point-defects in nanocrystalline materials. However, if more complex structural motifs or extended defects like stacking disorder are present in the material, a single-crystal structure might not encapsulate all features of the average structure. In a similar fashion, as shown before for single-phase optimization, multiple structural models can be generated with global optimization in the combined DFT and R-factor landscape.

Previously, we have studied the bonding topology as well as structural deformation under high pressure in single crystals of layered metal chalcogenide compounds and found dominant, but weak interlayer van der Waals (vdW) interactions responsible for the atomic and electronic structure.\cite{36,37} With only weak interlayer interactions and therefore only subtle energy differences between different stacking sequences, these vdW compounds provide a great example of identifying potential stacking disorder. As the last example, we therefore analyze a powder sample of SnS$_2$ to demonstrate the capabilities of the algorithm to study such extended defects.

In the first step, we use our algorithm to globally optimize the structure in a $1 \times 1 \times 2$ unit cell consisting of 2 Sn and 4 S atoms in the combined DFT and R$_w$ landscape. Two identical layers of SnS$_2$ are placed on top of each other. The resulting structure (phase 1) in Figure 5A is in agreement with the crystal structure of SnS$_2$ known from the literature. It consists of layers of edge-sharing SnS$_6$ octahedra with Sn atoms of subsequent layers placed directly on top of each other in an AAA stacking sequence. Using a higher value of $\alpha$, a second phase (cf. Figure 5A, phase 2) is found that is very similar to phase 1, but where every second layer is mirrored in a plane parallel to the $ab$ plane. This results in a stacking sequence we denote as AA*$A$, where * represents the mirroring. This second phase fits slightly better with the PDF with $R_w = 0.37$ compared to $R_w = 0.41$ for phase 1. Nevertheless, the differences in their

---

**Figure 4.** Modeled PDF and structure of Gahnite global optimized based on the R-factor and DFT energy. A) Inversion in a single unit cell. B) Inversion in a double unit cell. C) Inversion in a triple unit cell. All three models were predicted by GOFEE in the combined DFT PES and R-factor landscape. D) Computed PDF of structure A with corresponding $R_w$ value. E) Stability/likelihood comparison of Gahnite unit cells with and without inversion. The relative change with respect to the lowest energy structure (non-defected cells) or the structure with the lowest $R_w$ value (double cell with inversion) is shown. The gray atoms are aluminum, the blue atoms are zinc, and the red atoms are oxygen.
calculated PDFs are very small as shown in Figure 5B (cf. \( G(\mathbf{r})_{\text{phase}1} \) and \( G(\mathbf{r})_{\text{phase}2} \)). On the other hand, phase 2 is 0.55 eV per \( 1 \times 1 \times 2 \) unit cell less stable than phase 1.

Following this, we now proceed to search for a structural fit based on two phases. To limit the search, we assume that one of the phases is indeed the already identified phase 1. With this choice, the multi-phase search involves the optimization of the structure of the material in a second phase, which could in principle recover phase 1 or phase 2 once more. However, when conducting the search, we identify a new structure, phase 3, depicted in Figure 5A. While phase 3 in itself would not fit the PDF very well, \( R_w = 0.77 \) (cf. Figure 5B), when it occurs together with phase 1 as a multi-phase, it improves the fit considerably to \( R_w = 0.23 \) (cf. Figure 5C). The structure of phase 3 has in fact previously been reported as a polymorph of SnS\(_2\).[38] The local coordination environment of Sn and S are similar in all three phases, i.e., with layers of regular, edge-sharing octahedra. However, as for phase 2, every second layer in phase 3 is mirrored in a plane parallel to the \( ab \)-plane, but in addition, this layer is shifted within the \( ab \)-plane by a \( (2/3, 1/3, 0) \)-vector. This stacking pattern we can denote (AB*A). Energetically, this phase shows very similar stability to phase 1 being only 0.004 eV per \( 1 \times 1 \times 2 \) unit cell more stable. In fact, this very slight stability difference means that if a global optimization based only on DFT energies (\( \alpha = 0 \)) is performed, phase 3 results. Note, however, that for the DFT calculations the PBE functional has been used, and especially for calculations on vdW structures like SnS\(_2\), this functional is known to behave poorly. Surprisingly, adding dispersion corrections to the PBE functional does not change the energetic order of phases 1 and 3, but choosing another functional might do so.

These results seem to indicate either stacking disorder within SnS\(_2\) crystals or inclusions of some crystals with AB*A stacking sequence in the ensemble of measured SnS\(_2\) crystals.

### 4. Conclusion and Outlook

We have introduced a global optimization method, which utilizes methods from the field of machine learning to find the crystal structure of unknown compounds only based on the comparison of a model to a measured PDF (based on the agreement factor, \( R_w \)). This method alone works well for simple crystal structures, but for more complex materials, one needs to combine the R-factor with an energy evaluation of the system thus introducing chemical intuition in the algorithm, for example by means of DFT in order to sort out physically non-meaningful structures. In addition, the combined DFT+R-factor corrected landscape may lead to meta-stable configurations as global minima, whereas a search based on DFT would always try to predict the thermodynamically most stable configuration as the global minimum. Using this method, crystal structures of unknown compounds can be solved fully automatically based on a combination of PDF measurements and DFT calculations within only a few hours to days of computational time. In addition, the automated detection of additional phases in multicomponent materials is possible by combining DFT energies with multiphase PDF analyses in machine learning enhanced global optimization.

Although only tested with a combination of the DFT energy with the R-factor, the method is not limited to combining these two values and creating an artificial landscape for global
optimization. As alternatives, classical potentials or force fields could potentially be used in order to force the algorithms to generate physically meaningful structures. We additionally assume that instead of using the $R_w$ factor from a PDF agreement evaluation, other experimental comparisons could be used. One example could be to utilize the R-factor from a comparison of a measured and a calculated PXRD in combination with DFT, or even in a combined DFT+$R_{PDF}$+$R_{PXRD}$ landscape evaluation.

5. Experimental and Computational Details

**Experimental Methods:** The TS data on TiO$_2$ and ZnAl$_2$O$_4$ was collected at the P07 beamline at PETRA-III (DESY, Hamburg, Germany) with a photon energy of $\approx 60$ keV ($\lambda = 0.2072$ Å)$^{[39]}$ on a 2D PerkinElmer XRD 1621 area detector for data acquisition. The TS data on CeO$_2$ shown in the Supporting Information was collected at the P07 beamline at PETRA-III (DESY, Hamburg, Germany) with a photon energy of $\approx 100$ keV ($\lambda = 0.124$ Å)$^{[40]}$ on a DECTRIS PILATUS3 $\times 2$ m CdTe detector. The 2D detector images were azimuthally integrated with the Dioptas$^{[41]}$ software and calibrated using either a LaB$_6$ 660b NIST line standard for TiO$_2$ and ZnAl$_2$O$_4$ or the CeO$_2$ data itself. The TS data on SnS$_2$ was collected at the RIKEN BL44B2 beamline at SPring-8 (Japan) with a photon energy of $24.8$ keV ($\lambda = 0.500408$ Å) on an image plate detector. All data was appropriately background subtracted and subsequently Fourier transformed with PDFgetX$^{[42]}$ in the xPDFsuite$^{[43]}$ software, thus giving the PDF.

The TS data of Anatase (TiO$_2$) nanoparticles was obtained from an in situ synthesis by Søndergaard-Pedersen et al.$^{[44]}$ An industrial-grade TiO$_2$ precursor was diluted with sulfuric acid to a titanium concentration of 1.29% (M). The solution was injected into a fused silica capillary mounted in a revised version of the in situ setup described in detail elsewhere,$^{[45]}$ pressurized to 250 bar and heated to 200 °C. Data collection was carried out with a temporal resolution of 3 s. The data was background subtracted with a measurement of diluted sulfuric acid and Fourier transformed with a $Q_{\text{max}} = 16.8$ Å$^{-1}$. Instrumental dampening was accounted for by separately obtaining and refining TS data of a LaB$_6$ 660b NIST line standard in the PDFgui software.$^{[46]}$

The TS data of Gahnite (ZnAl$_2$O$_4$) nanoparticles originated from a study by Sommer et al.$^{[35]}$ and readers are referred to their paper on the preparation of the nanoparticles. The data was background subtracted with a measurement of an empty capillary and Fourier transformed with a $Q_{\text{max}} = 23.5$ Å$^{-1}$. Instrumental dampening was accounted for by separately obtaining and refining TS data of a LaB$_6$ 660b NIST line standard in the PDFgui software.$^{[46]}$

The TS data of CeO$_2$ was background subtracted with a measurement of an empty capillary and Fourier transformed with a $Q_{\text{max}} = 25.0$ Å$^{-1}$. Instrumental dampening was accounted for by performing an initial refinement of the data in the TOPAS v6 software.$^{[47]}$

The TS data of SnS$_2$ was Fourier transformed with a $Q_{\text{max}} = 15.7$ Å$^{-1}$. Instrumental dampening was accounted for by performing an initial refinement of the data in the TOPAS v7 software.$^{[47]}$

**Local Structure Optimization:** A local geometry optimization is a minimization of the target function towards the nearest minimum. Local geometry optimizations can be performed straightforwardly for DFT calculations using analytical forces as implemented in GPAW.$^{[48]}$ The exchange-correlation interaction is treated by the generalized gradient approximation using the PBE functional.

In addition, local geometry optimizations can be performed in the artificial R-factor landscape. Instead of using an analytical expression, forces are calculated numerically by displacing the atom positions $\pm 0.001$ Å into each Cartesian direction and calculating the corresponding R-factor (from Equation (1)). Then, the BFGS algorithm as implemented in ASE$^{[49]}$ is used to optimize the structure into its local minimum.

Global Structure Optimization Using a Machine-Learning-Based Surrogate Landscape: In this work, the global optimization method GOFEE is used. In contrast to the original work by Bisbo and Hammer,$^{[11,12]}$ the machine learning potential is not always exclusively trained on DFT energies, but additionally on the R-factor (as defined in Equation (1)) or a combination of the R-factor and DFT energy according to Equation (2).

The workflow of the code can be divided into five steps. First (I), a user-defined number of initial random atomic structures is created, e.g., 10 structures. All structures share a common set of lattice parameters and stoichiometry. Each of these structures are geometrically locally relaxed either using DFT forces or numerical R-factor forces. Based on this data, the initial machine learning potential can be trained. In the second step (II), $m$ new structures are created via random guess or mutation of the previous population ($m = \text{number of CPU cores}$). In the current version of the code, the mutations are either permutations of two non-identical atoms or rattling some of the atoms to new positions. In the third step (III), all new structures’ atomic positions are relaxed in the model machine learning potential. Next (IV), the fitness of each of these new structures are evaluated in the machine learning potential. Then a structure, where the machine learning potential either predicts a good fitness (exploit) or has a high uncertainty (explore), is chosen. For this structure, the correct fitness is evaluated. The fitness can either be the DFT potential energy, the R-factor as introduced in Equation (1), or a combination of both as introduced in Equation (2). In the last step (V), the chosen new structure and its corresponding fitness are added to the training data and the model potential is retrained. The algorithm is started again from step (II) until $n$ iterations have been reached (e.g., $n = 2000$), resulting in hundreds of different local minimum structures. The resulting structures including the global minimum structure are only relaxed in the ML potential during the search. Therefore, the resulting global minimum structure has to be post-processed and has to be locally optimized as described above. If the energetic stability should be compared, the final structure is relaxed using analytical DFT forces. If the fit to the experiment is important, then numerical R-factor forces are used for the final local relaxation.

**Supporting Information**

Supporting Information is available from the Wiley Online Library or from the author.

**Acknowledgements**

The authors acknowledge financial support through the APF project “Materials on Demand” within the “Humans on Mars” Initiative funded by the Federal State of Bremen and the University of Bremen and support from the Villum Fonden. In addition, this work was supported by the Danish National Research Foundation through the Center of Excellence “InterCat” (Grant Agreement No. DNRF150). The authors thank the synchrotron beamline RIKEN BL44B2 at SPring-8 for the beamtimeallocation. W.D. thanks L. Colombi Ciacchi and T. M. Gesing for fruitful discussions related to the underlying work.

**Conflict of Interest**

The authors declare no conflict of interest.

**Data Availability Statement**

The data that support the findings of this study are available from the corresponding author upon reasonable request.
Keywords
crystal structure prediction, density functional theory, global optimization, machine learning, nanomaterials, pair distribution function

Received: September 7, 2022
Revised: November 25, 2022
Published online: February 12, 2023