

ARTICLE

DOI: 10.1038/s41467-018-05169-6

OPEN

# Insightful classification of crystal structures using deep learning

Angelo Ziletti <sup>1</sup>, Devinder Kumar <sup>2,3</sup>, Matthias Scheffler<sup>1</sup> & Luca M. Ghiringhelli <sup>1</sup>

Computational methods that automatically extract knowledge from data are critical for enabling data-driven materials science. A reliable identification of lattice symmetry is a crucial first step for materials characterization and analytics. Current methods require a user-specified threshold, and are unable to detect average symmetries for defective structures. Here, we propose a machine learning-based approach to automatically classify structures by crystal symmetry. First, we represent crystals by calculating a diffraction image, then construct a deep learning neural network model for classification. Our approach is able to correctly classify a dataset comprising more than 100,000 simulated crystal structures, including heavily defective ones. The internal operations of the neural network are unraveled through attentive response maps, demonstrating that it uses the same landmarks a materials scientist would use, although never explicitly instructed to do so. Our study paves the way for crystal structure recognition of—possibly noisy and incomplete—three-dimensional structural data in big-data materials science.

<sup>1</sup>Theory Department, Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany. <sup>2</sup>University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada. <sup>3</sup>Vector Institute of AI, Toronto ON M5G 1M1, Canada. Correspondence and requests for materials should be addressed to A.Z. (email: [ziletti@fhi-berlin.mpg.de](mailto:ziletti@fhi-berlin.mpg.de))

Crystals play a crucial role in materials science. In particular, knowing chemical composition and crystal structure—the way atoms are arranged in space—is an essential ingredient for predicting properties of a material<sup>1–3</sup>. Indeed, it is well known that the crystal structure has a direct impact on materials properties<sup>4</sup>. Just to give a concrete example: in iron, carbon solubility (important for steel formation) increases nearly forty times going from body-centered cubic (bcc)  $\alpha$ -Fe (ferrite) to face-centered cubic (fcc)  $\gamma$ -Fe (austenite)<sup>5</sup>. From the computational point of view, identification of crystal symmetries allows, for example, to construct appropriate  $k$ -point grids for Brillouin zone sampling, generate paths between high-symmetry points in band structure calculations, or identify distortions for finite-displacement phonon calculations.

Given the importance of atomic arrangement in both theoretical and experimental materials science, an effective way of classifying crystals is to find the group of all transformations under which the system is invariant; in three dimensions, these are described by the concept of space groups<sup>6</sup>. Currently, to determine the space group of a given structure, one first determines the allowed symmetry operations, and then compare them with all possible space groups to obtain the correct label; this is implemented in existing symmetry packages such as FINDSYM<sup>7</sup>, Platon<sup>8</sup>, Spglib<sup>9–11</sup>, and, most recently, the self-consistent, threshold-adaptive AFLOW-SYM<sup>12</sup>. For idealized crystal structures, this procedure is exact. But in most practical applications atoms are displaced from their ideal symmetry positions due to (unavoidable) intrinsic defects or impurities or experimental noise. To address this, thresholds need to be set in order to define how loose one wants to be in classifying (namely, up to which deviations from the ideal structures are acceptable); different thresholds may lead to different classifications (see for instance Table 1). So far, this was not a big problem because individual researchers were manually finding appropriate tolerance parameters for their specific dataset.

However, our goal here is to introduce an automatic procedure to classify crystal structures starting from a set of atomic coordinates and lattice vectors; this is motivated by the advent of high-throughput materials science computations, owing to which millions of calculated data are now available to the scientific community (see the Novel Materials Discovery (NOMAD) Laboratory<sup>13</sup> and references therein). Clearly, there is no universal threshold that performs optimally (or even sub-optimally) for such a large number of calculations, nor a clear procedure to check if the chosen threshold is sound. Moreover, the aforementioned symmetry-based approach fails—regardless of the tolerance thresholds—in the presence of defects such as, for example, vacancies, interstitials, antisites, or dislocations. In fact, even removing a single atom from a structure causes the system to lose most of its symmetries, and thus one typically obtains the (low symmetry, e.g.  $P1$ ) space group compatible with the few symmetry operations preserved in the defective structure. This

label—although being technically correct—is practically always different from the label that one would consider appropriate (i.e., the most similar space group, in this case the one of the pristine structure). Robustness to defects, however, is paramount in local and global crystal structure recognition. Grain boundaries, dislocations, local inclusions, heterophase interfaces, and in general all crystallographic defects can have a large impact on macroscopic materials properties (e.g., corrosion resistance<sup>14, 15</sup>). Furthermore, atom probe tomography—arguably the most important source of local structural information for bulk systems—provides three-dimensional atomic positions with an efficiency up to 80%<sup>16</sup> and near-atomic resolution, which, on the other hand, means that at least 20% of atoms escaped detection, and the uncertainty on their positions is considerable.

Here, we propose a procedure to efficiently represent and classify potentially noisy and incomplete three-dimensional materials science structural data according to their crystal symmetry (and not to classify x-ray diffraction images, or powder x-ray diffraction data<sup>17</sup>). These three-dimensional structural data could be, for example, atomic structures from computational materials science databases, or elemental mappings from atom probe tomography experiments. Our procedure does not require any tolerance threshold, and it is very robust to defects (even at defect concentrations as high as 40%). First, we introduce a way to represent crystal structures (by means of images, i.e., two-dimensional maps of the three-dimensional crystal structures, see below), then we present a classification model based on convolutional neural networks (ConvNets), and finally we unfold the internal behavior of the classification model through visualization. An interactive online tutorial for reproducing the main results of this work is also provided<sup>18</sup>.

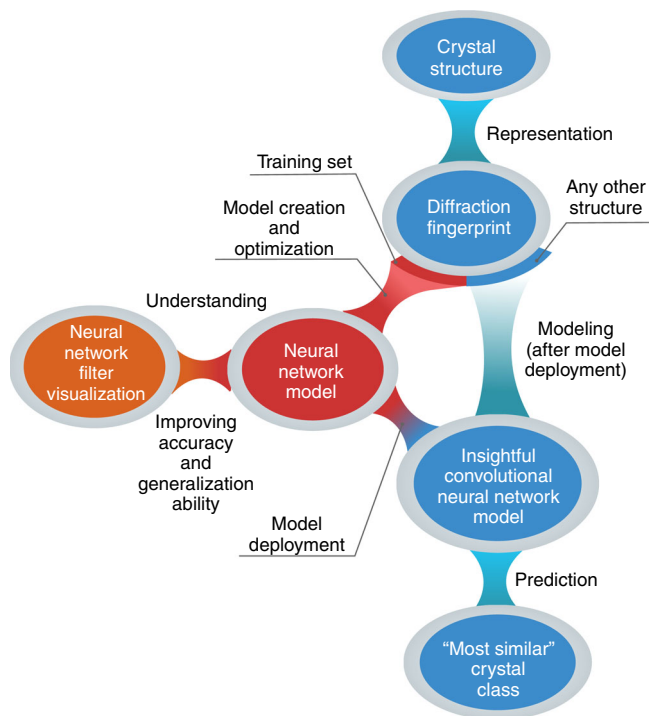
## Results

**How to represent a material.** The first necessary step to perform any machine learning and/or automatized analysis on materials science data (see Fig. 1) is to represent the material under consideration in a way that is understandable for a computer. This representation—termed “descriptor”<sup>19</sup>—should contain all the relevant information on the system needed for the desired learning task. Numerous structural descriptors have been proposed to represent physical systems, most notable examples being atom-centered symmetry functions<sup>20</sup>, Coulomb matrix<sup>21</sup>, smooth overlap of atomic positions<sup>22</sup>, deep tensor neural networks<sup>23</sup>, many-body tensor representation<sup>24</sup>, and Voronoi tessellation<sup>25, 26</sup>. However, these descriptors are either not applicable to extended systems<sup>21, 23</sup>, not size-invariant by construction<sup>24</sup>, or base their representation of infinite crystals on local neighborhoods of atoms in the material<sup>20, 22, 25–27</sup>. If on the one hand these local approaches are able to produce accurate force fields<sup>28, 29</sup>, on the other hand their strategy of essentially partitioning the crystal in patches (defined by a certain cut-off radius, generally 4–6 Å<sup>20, 28</sup>) makes it difficult to detect global

**Table 1 Accuracy in identifying the correct (most similar) crystal class in the presence of defects**

|                 | Random displacements ( $\sigma$ ) |         |         |        |        |        | Vacancies ( $\eta$ ) |        |        |        |
|-----------------|-----------------------------------|---------|---------|--------|--------|--------|----------------------|--------|--------|--------|
|                 | 0.001 Å                           | 0.002 Å | 0.005 Å | 0.01 Å | 0.02 Å | 0.06 Å | 1%                   | 2%     | 15%    | 25%    |
| Spglib (tight)  | 0.00                              | 0.00    | 0.00    | 0.00   | 0.00   | 0.00   | 0.02                 | 0.00   | 0.00   | 0.00   |
| Spglib (medium) | 73.70                             | 0.00    | 0.00    | 0.00   | 0.00   | 0.00   | 0.02                 | 0.00   | 0.00   | 0.00   |
| Spglib (loose)  | 99.99                             | 99.99   | 99.99   | 75.22  | 0.00   | 0.00   | 0.01                 | 0.00   | 0.00   | 0.00   |
| This work       | 100.00                            | 100.00  | 100.00  | 100.00 | 100.00 | 100.00 | 100.00               | 100.00 | 100.00 | 100.00 |

The defective structures are calculated randomly displacing atoms according to Gaussian distribution with standard deviation  $\sigma$  (left), or removing  $\eta\%$  of the atoms (right). For details regarding the Spglib thresholds chosen see Supplementary Note 1. The accuracy values shown in the table are in percentage



**Fig. 1** The model workflow of automatic crystal structure classification. First, every crystal structure is represented by the two-dimensional diffraction fingerprint. Then, a small subset of these structures are used as training set to generate a classification model. In particular, a convolutional neural network is used, and optimized minimizing the training set classification error. However, this is in general not enough to have a sound and generalizable model. Thus, we unfold the neural network internal operations by visualization, and ensure that the model arrives at its classification decision on physically motivated grounds. Finally, a classification model is deployed, and crystal structures can be directly and efficiently classified without any additional model optimization

structural properties, in particular where recognizing long-range order is crucial.

In the case of crystal structure recognition, however, it is essential that the descriptor captures system's symmetries in a compact way, while being size invariant in order to reflect the infinite nature of crystals. Periodicity and prevailing symmetries are evident—and more compact—in reciprocal space, and therefore we introduce an approach based on this space. For every system, we first simulate the scattering of an incident plane wave through the crystal, and then we compute the diffraction pattern in the detector plane orthogonal to that incident wave. This is schematically depicted in Fig. 2a.

The amplitude  $\Psi$ , which originates from the scattering of a plane wave with wave vector  $\mathbf{k}_0$  by  $N_a$  atoms of species  $a$  at positions  $\{\mathbf{x}_j^{(a)}\}$  in the material can be written as:

$$\Psi(\mathbf{q}) = r^{-1} \sum_a f_a^\lambda(\theta) \left[ \sum_{j=1}^{N_a} r_0 \exp(-i\mathbf{q} \cdot \mathbf{x}_j^{(a)}) \right], \quad (1)$$

where  $r_0$  is the Thomson scattering length,  $\mathbf{q} = \mathbf{k}_1 - \mathbf{k}_0$  is the scattering wave vector,  $\mathbf{x}'$  the corresponding position in the detector plane, and  $r = |\mathbf{x}'|$  (see Fig. 2a). Assuming elastic scattering, we have that  $|\mathbf{k}_0| = |\mathbf{k}_1| = 2\pi/\lambda$ , where  $\lambda$  is the wavelength of the incident radiation. The quantity  $f_a^\lambda(\theta)$  is the so-called x-ray form factor; it describes how an isolated atom of species  $a$  scatters incident radiation with wavelength  $\lambda$  and scattering angle  $\theta$ . Since x-rays are scattered by the electronic

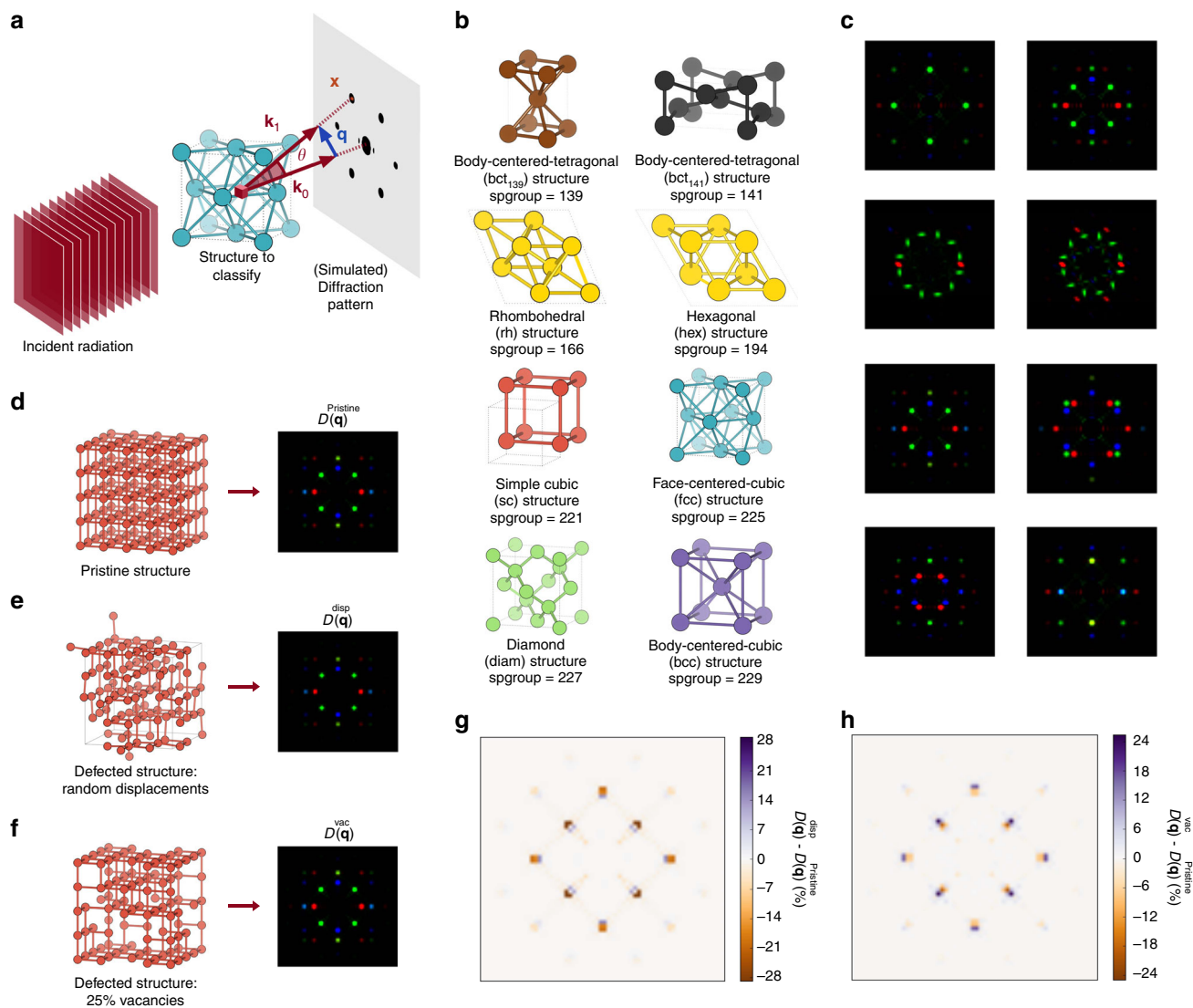
cloud of an atom, its amplitude increases with the atomic number  $Z$  of the element<sup>30</sup>. Following the successful application of scattering concepts in determining atomic structures (using for example x-rays<sup>31</sup>, electrons<sup>32</sup> or neutrons<sup>33</sup>), we propose the diffraction pattern intensity as the central quantity to describe crystal structures:

$$I(\mathbf{q}) = A \cdot \Omega(\theta) |\Psi(\mathbf{q})|^2, \quad (2)$$

where  $\Omega(\theta)$  is the solid angle covered by our (theoretical) detector, and  $A$  is a (inessential) constant determined by normalization with respect to the brightest peak (see section Methods). For each structure we first construct the standard conventional cell according to ref<sup>34</sup>. Then, we rotate the structure 45° clockwise and counterclockwise about a given crystal axis (e.g.,  $x$ ), calculate the diffraction pattern for each rotation, and superimpose the two patterns. Any other choice of rotation angle is in principle valid, provided that the diffraction patterns corresponding to different crystal classes do not accidentally become degenerate. This procedure is then repeated for all three crystal axes. The final result is represented as one RGB image for crystal structure, where each color channel shows the diffraction patterns obtained by rotating about a given axis (i.e., red (R) for  $x$ -axis, green (G) for  $y$ -axis, and blue (B) for  $z$ -axis). Each system is thus described as an image, and we term this descriptor two-dimensional diffraction fingerprint ( $D_F$ ). We point out that this procedure does not require to already know the crystal symmetry, and  $x$ ,  $y$ , and  $z$  are arbitrary, for example, determined ordering the lattice vectors by length<sup>34</sup> (or whatever the chosen criterion). For additional computational details on the descriptor  $D_F$ , please refer to the section Methods.

Despite its rather complicated functional form (see Eqs. (1) and (2)), the descriptor  $D_F$  is one image for each system being represented (data point); the eight crystal classes considered in this work (see below) and examples of their calculated two-dimensional diffraction fingerprints are shown in Fig. 2b, c, respectively. This descriptor compactly encodes detailed structural information (through Eq. (1)) and—in accordance with scattering theory—has several desirable properties for crystal structure classification, as we outline below.

It is invariant with respect to system size: changing the number of periodic replicas of the system will leave the diffraction peak locations unaffected. This allows to treat extended and finite systems on equal footing, making our procedure able to recognize global and local order, respectively. We exploit this property, and instead of using periodically repeated crystals, we calculate  $D_F$  using clusters of approximately 250 atoms. These clusters are constructed replicating the crystal unit cell (see Methods). By using finite samples, we explicitly demonstrate the local structure recognition ability of our procedure. The diffraction fingerprint is also invariant under atomic permutations: re-ordering the list of atoms in the system leads to the same  $D_F$  due to the sum over all atoms in Eq. (1). Moreover, its dimension is independent of the number of atoms and the number of chemical species in the system being represented. This is an important property because machine learning models trained using this descriptor generalize to systems of different size by construction. This is not valid for most descriptors: for example, the Coulomb matrix dimension scales as the square of atoms in the largest molecule considered<sup>21</sup>, while in symmetry functions-based approaches<sup>20</sup> the required number of functions (and thus model complexity) increases rapidly with the number of chemical species and system size. Being based on the process of diffraction, the diffraction fingerprint mainly focuses on atomic positions and crystal symmetries; the information on the atomic species—encoded in the form factor  $f_a^\lambda$  in Eq. (1)—plays a less prominent role in the

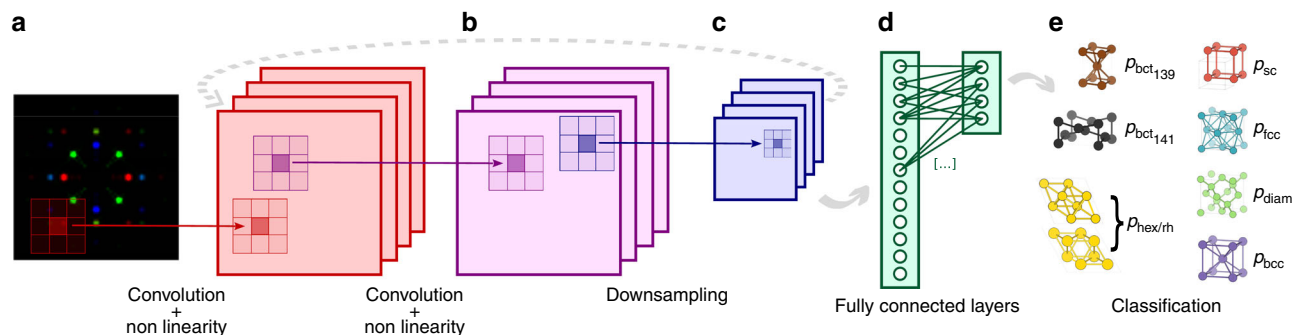


**Fig. 2** The two-dimensional diffraction fingerprint. **a** Schematic representation of the two-dimensional diffraction fingerprint calculation. An incident plane wave is scattered by the material, and the diffraction pattern on a plane perpendicular to the incident radiation is computed. **b** Prototypes of the crystal classes considered in this work. **c** Examples of two-dimensional diffraction patterns for materials belonging to each of the eight classes. The ordering is the same as **b**. Rhombohedral and hexagonal structures have the same two-dimensional diffraction fingerprint. **d–f** A pristine simple cubic structure (**d**), the same structure with 25% of vacancies (**e**), and with atoms displaced randomly according to a Gaussian distribution with standard deviation of 0.08 Å (**f**), together with their diffraction fingerprints. **g, h** Difference between the diffraction fingerprints of the defective (**e**), (**f**) and the pristine structure (**d**)

descriptor. As a result, materials with different atomic composition but similar crystal structure have similar representations. This is the ideal scenario for crystals classification: a descriptor which is similar for materials within the same class, and very different for materials belonging to different classes. Finally, the diffraction fingerprint is straightforward to compute, easily interpretable by a human (it is an image, see Fig. 2c), has a clear physical meaning (Eqs. (1) and (2)), and is very robust to defects. This last fact can be traced back to a well-known property of the Fourier transform: the field at one point in reciprocal space (the image space in our case) depends on all points in real space. In particular, from Eq. (1) we notice that the field  $\Psi$  at point  $\mathbf{q}$  is given by the sum of the scattering contributions from all the atoms in the system. If, for example, some atoms are removed, this change will be smoothed out by the sum over all atoms and spread over—in principle—all points in reciprocal space. Practically, with increasing disorder new low-intensity peaks will gradually appear in the diffraction fingerprint due to the now

imperfect destructive interference between the atoms in the crystal. Examples of pristine and highly defected structures, together with their corresponding diffraction fingerprints, are shown in Fig. 2d–f, respectively. It is evident that the diffraction fingerprint is indeed robust to defects. This property is crucial in enabling the classification model to obtain a perfect classification even in the presence of highly defective structures (see below).

A disadvantage of the two-dimensional diffraction fingerprint is that it is not unique across space groups. This is well known in crystallography: the diffraction pattern does not always determine unambiguously the space group of a crystal<sup>35,36</sup>. This is primarily because the symmetry of the diffraction pattern is not necessarily the same as the corresponding real-space crystal structure; for example, Friedel's law states that—if anomalous dispersion is neglected—a diffraction pattern is centrosymmetric, irrespective of whether or not the crystal itself has a center of symmetry. Thus, the diffraction fingerprint  $D_F$  cannot represent non-centrosymmetric structures by construction. The non-



**Fig. 3** Schematic representation of the convolutional neural network (ConvNet) used for crystals classification. **a** A learnable filter (also called kernel) is convolved across the image, and the scalar product between the filter and the input at every position is computed. This results in a two-dimensional activation map (in red) of that filter at each spatial position, which is then passed through a rectified linear unit (ReLU)<sup>66</sup>. **b** The same procedure as point a is applied to this activation map (instead of the original image), producing another activation map (in purple). **c** A downsampling operation (in blue) is performed to coarse grain the representation. Six convolutional and two downsampling (max pooling) layers are stacked sequentially (see Methods for additional details). **d** The output of the convolutional/downsampling layers sequence is passed to fully connected layers (regularized using dropout<sup>67</sup>) to complete the classification procedure. **e** The ConvNet outputs the probabilities that the input image, and therefore the corresponding material, belongs to a given class. Minimizing the classification error, the above-mentioned filters are learned—through backpropagation<sup>68</sup>—and they will activate when a similar feature (e.g., edges or curves for initial layers, and more complex motifs for deeper layers) appears in the input

uniqueness of the diffraction pattern  $I(\mathbf{q})$  across space groups also implies that crystal structures belonging to different space groups can have the same diffraction fingerprints. Nevertheless, from Fig. 2c we notice that out of the eight crystal structure prototypes considered (covering the large majority of the most thermodynamically stable structures formed in nature by elemental solids<sup>37</sup>), only the rhombohedral and hexagonal structures—whose real-space crystal structures are quite similar—have the same two-dimensional diffraction fingerprint.

**The classification model.** Having introduced a way to represent periodic systems using scattering theory, we tackle the problem of their classification in crystal classes based on symmetries. A first (and naive) approach to classify crystals—now represented by the diffraction descriptor  $D_F$ —would be to write specific programs that detect diffraction peaks in the images, and classify accordingly. Despite appearing simple at first glance, this requires numerous assumptions and heuristic criteria; one would need to define what is an actual diffraction peak and what is just noise, when two contiguous peaks are considered as one, how to quantify relative peak positions, to name but a few. In order to find such criteria and determine the associated parameters, one in principle needs to inspect all (thousands or even millions) pictures that are being classified. These rules would presumably be different across classes, require a separate—and not trivial—classification paradigm for each class, and consequently lead to a quagmire of ad hoc parameters and task-specific software. In addition, the presence of defects leads to new peaks or alters the existing ones (see Fig. 2g, h), complicating matters even further. Thus, this approach is certainly not easy to generalize to other crystal classes, and lacks a procedure to systematically improve its prediction capabilities.

However, it has been shown that all these challenges can be solved by deep learning architectures<sup>38–40</sup>. These are computational nonlinear models sequentially composed to generate representations of data with increasing level of abstraction. Hence, instead of writing a program by hand for each specific task, we collect a large amount of examples that specify the correct output (crystal class) for a given input (descriptor image  $D_F$ ), and then minimize an objective function which quantifies the difference between the predicted and the correct classification labels. Through this minimization, the weights (i.e., parameters) of the neural network are optimized to reduce such classification

error<sup>41, 42</sup>. In doing so, the network automatically learns representations (also called features) which capture discriminative elements, while discarding details not important for classification. This task—known as feature extraction—usually requires a considerable amount of heuristics and domain knowledge, but in deep learning architectures is performed with a fully automated and general-purpose procedure<sup>40</sup>. In particular, since our goal is to classify images, we use a specific type of deep learning network which has shown superior performance in image recognition: the ConvNet<sup>43–45</sup>. A schematic representation of the ConvNet used in this work is shown in Fig. 3. ConvNets are inspired by the multi-layered organization of the visual cortex<sup>46</sup>: filters are learned in a hierarchical fashion, composing low-level features (e.g., points, edges, or curves) to generate more complex motifs. In our case, such motifs encode the relative position of the peaks in the diffraction fingerprint for the crystal classes considered, as we will show below.

**The model performance.** For every calculation in the AFLOW-LIB elemental solid database<sup>47, 48</sup>, we determine its space group using a symmetry-based approach<sup>9, 10</sup> as implemented by the Spglib code. We then extract all systems belonging to centrosymmetric space groups which are represented with more than 50 configurations. This gives us systems with the following space group numbers: 139, 141, 166, 194, 221, 225, 227, and 229. For the case of elemental solids presented here, these space groups correspond to body-centered tetragonal (bct, 139 and 141), rhombohedral (rh, 166), hexagonal (hex, 194), simple cubic (sc, 221), fcc (225), diamond (diam, 227), and bcc (229) structures. This represents a rather complete dataset since it includes the crystal structures adopted by more than 80% of elemental solids under standard conditions<sup>37</sup>. It is also a challenging dataset because it contains 10,517 crystal structures comprising 83 different chemical species, cells of various size, and structures that are not necessarily in the most stable atomic arrangement for a given composition, or even at a local energy minimum. This last point in particular could potentially be a problem for the symmetry-based approach: when crystals are not in a perfect arrangement, it can fail in returning the correct labels. In fact, if atoms are slightly displaced from their expected symmetry positions, the classification could return a different space group because symmetries might be broken by this numerical noise. To avoid this, we include in the pristine dataset only systems which

are successfully recognized by the symmetry-based approach to belong to one of the eight classes above, thus ensuring that the labels are correct. We refer to the above as pristine dataset; the dataset labels are the aforementioned space groups, except for rh and hex structures, which we merge in one class (hex/rh) since they have the same diffraction fingerprint (see Fig. 2c).

We apply the workflow introduced here (and schematically shown in Fig. 1) to this dataset. For each structure, we first compute the two-dimensional diffraction fingerprint  $D_F$ ; then, we train the ConvNet on (a random) 90% of the dataset, and use the remaining 10% as test set. We obtain an accuracy of 100% on both training and test set, showing that the model is able to perfectly learn the samples and at the same time capable of correctly classifying systems which were never encountered before. The ConvNet model optimization (i.e., training) takes 80 min on a quad-core Intel(R) Core(TM) i7-3540M CPU, while one class label is predicted—for a given  $D_F$ —in approximately 70 ms on the same machine (including reading time). The power of machine learning models lies in their ability to produce accurate results for samples that were not included at training. In particular, the more dissimilar test samples are from the training samples, the more stringent is the assessment of the model generalization performance. To evaluate this, starting from the pristine dataset, we generate heavily defective structures introducing random displacements (sampled from Gaussian distributions with standard deviation  $\sigma$ ), randomly substituting atomic species (thus forming binaries and ternaries alloys), and creating vacancies. This results in a dataset of defective systems, for some of which even the trained eyes of a materials scientist might have trouble identifying the underlying crystal symmetries from their structures in real space (compare, e.g., the crystal structures in Fig. 2d, f).

As mentioned in the Introduction and explicitly shown below, symmetry-based approaches for space group determination fail in giving the correct (most similar) crystal class in the presence of defects. Thus, strictly speaking, we do not have a true label to compare with. However, since in this particular case the defective dataset is generated starting from the pristine, we do know the original crystal class for each sample. Hence, to estimate the model generalization capability, we label the defective structures with the class label of the corresponding pristine (parental) system. This is a sensible strategy given that displacing, substituting, or removing atoms at random will unlikely change the materials' crystal class. Using the ConvNet trained on the pristine dataset (and labels from the pristine structures), we then predict the labels for structures belonging to the defective dataset. A summary of our findings is presented in Table 1, which comprises results for  $10,517 \times (6 + 4) = 105,170$  defective systems; additional data are provided in Supplementary Notes 1 and 2.

When random displacements are introduced, Spglib accuracy varies considerably according to the threshold used; moreover, at  $\sigma \geq 0.02 \text{ \AA}$  Spglib is never able to identify the most similar crystal class, regardless of threshold used. Conversely, the method proposed in this work always identifies the correct class up to  $\sigma$  as high as  $0.06 \text{ \AA}$ . Similar are the results for vacancies: Spglib accuracy is  $\sim 0\%$  already at vacancies concentrations of 1%, while our procedure attains an accuracy of 100% up to 40% vacancies, and  $>97\%$  for vacancy concentrations as high as 60% (Table 1 and Supplementary Table 2). Since no defective structure was included at training, this represents a compelling evidence of both the model robustness to defects and its generalization ability.

If random changes will unlikely modify a crystal class, it is however possible to apply targeted transformations in order to change a given crystal from one class to another. In particular, starting from a bcc one can obtain an sc crystal removing all

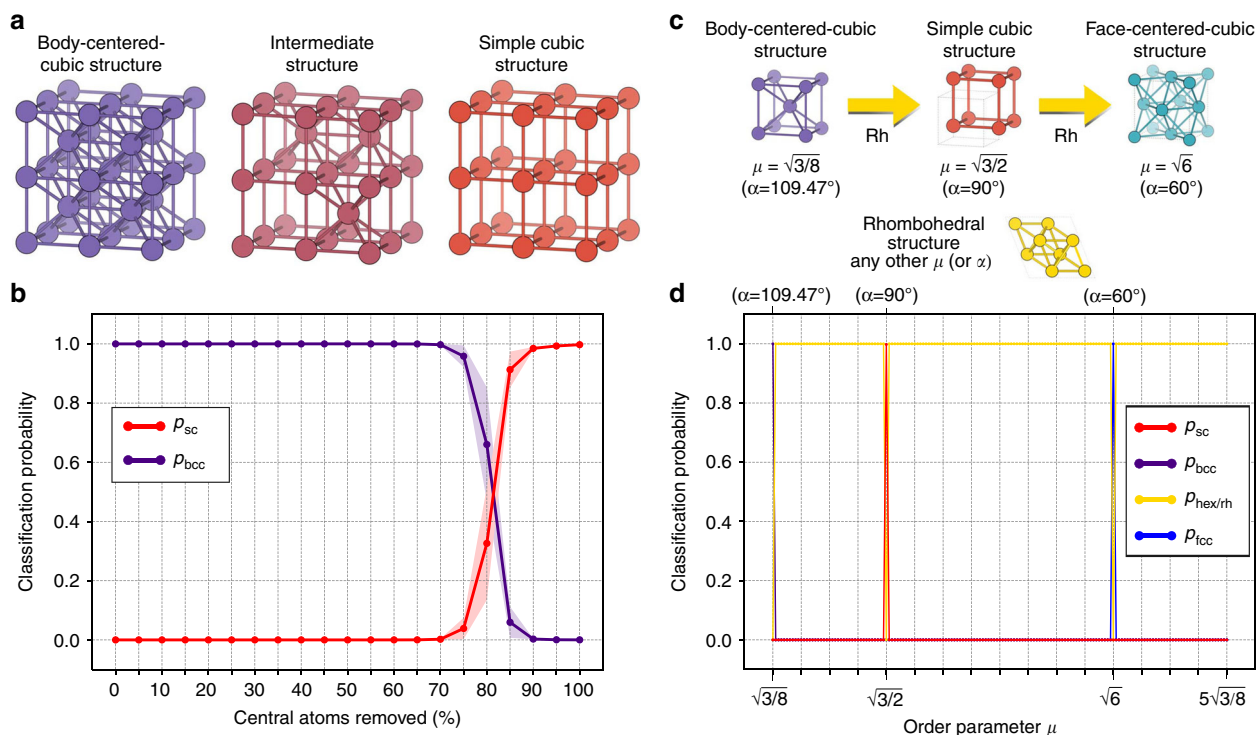
atoms at the center of the bcc unit cell (Figs. 2b and 4a). We remove different percentages of central atoms (from 0 to 100%, at 10% steps) from a subset of bcc structures in the pristine dataset; this gives us a collection of structures which are intermediate between bcc and sc by construction (see Fig. 4a center for a concrete example).

Let us now recall that the output of our approach is not only the crystal class but also the probability that a system belongs to a given class; this quantifies how certain the neural network is regarding its classification. The probability of the aforementioned structures being fcc (purple) or sc (red) according to our model are plotted in Fig. 4b as function of the percentage of central atoms removed (the shaded area indicates the standard deviation of such distributions). This percentage can be seen as a order parameter of the bcc-to-sc structural phase transition. If no atoms are removed, the structures are pure bcc, and the model indeed classifies them as bcc with probability 1, and zero standard deviation. At first, removing (central) atoms does not modify this behavior: the structures are seen by the model as defective bcc structures. However, at 75% of central atoms removed, the neural network judges that such structures are not defective bcc anymore, but are actually intermediate between bcc and sc. This is reflected in an increase of the classification probability of sc, a corresponding decrease in bcc probability, and a large increment in the standard deviation of these two distributions. When all central atoms are removed, we are left with pure sc structures, and the model classifies again with probability 1, and vanishing standard deviation: the neural network is confident that these structures belong to the sc class.

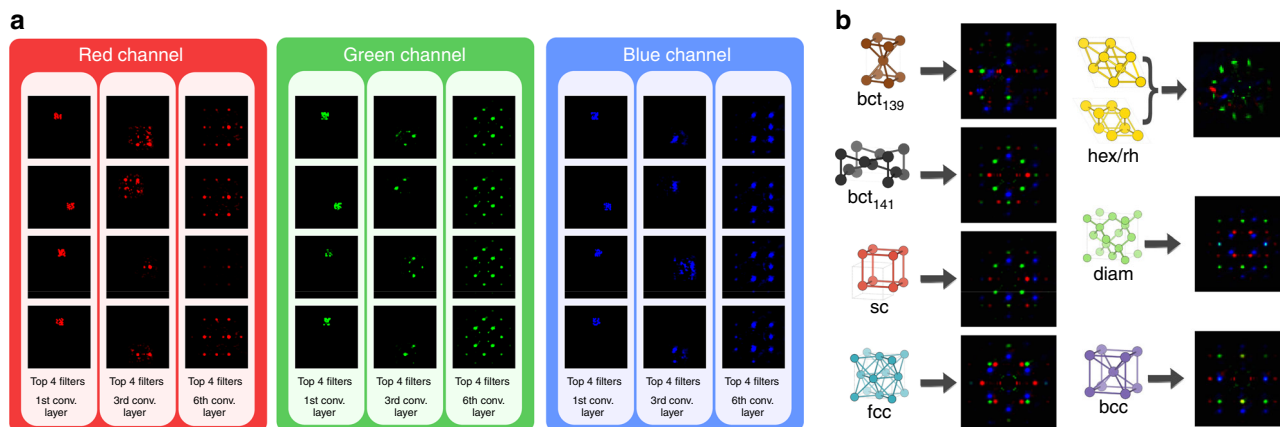
We conclude our model exploration applying the classification procedure on a structural transition path encompassing rh, bcc, sc, and fcc structures (Fig. 4c). From the AFLOW Library of Crystallographic Prototypes<sup>49</sup>, we generate rhombohedral structures belonging to space group 166 (prototype  $\beta$ -Po A\_hR1\_166\_a) with different values of  $\mu \equiv c/a$  or  $\alpha$ , where  $a$  and  $c$  are two of the lattice vectors of the conventional cell<sup>34</sup>, and  $\alpha$  is the angle formed by the primitive lattice vectors<sup>49</sup>. Particular values of  $\mu$  (or  $\alpha$ ) lead this rhombohedral prototype to reduce to bcc ( $\mu_{\text{bcc}} = \sqrt{3/8}$  or  $\alpha = 109.47^\circ$ ), sc ( $\mu_{\text{sc}} = \sqrt{3/2}$  or  $\alpha = 90^\circ$ ), or fcc ( $\mu_{\text{fcc}} = \sqrt{6}$  or  $\alpha = 60^\circ$ ) structures<sup>49</sup>. To test our model on this structural transition path, we generate crystal structures with  $\sqrt{3/8} \leq \mu \leq 5\sqrt{3/8}$ , and use the neural network trained above to classify these structures. The results are shown in Fig. 4d. Our approach is able to identify when the prototype reduces to the high-symmetry structures mentioned above (at  $\mu_{\text{bcc}}$ ,  $\mu_{\text{sc}}$ , and  $\mu_{\text{fcc}}$ ), and also correctly classify the structure as being rhombohedral for all other values of  $\mu$ . This is indeed the correct behavior: outside the high-symmetry bcc/sc/fcc the structure goes back to hex/rh precisely because that is the lower symmetry family ( $\mu$  not equal to  $\mu_{\text{bcc}}$ ,  $\mu_{\text{sc}}$ , or  $\mu_{\text{fcc}}$ ).

**Opening the black box using attentive response maps.** Our procedure based on diffraction fingerprints and ConvNet correctly classifies both pristine and defective dataset, but are we obtaining the right result for the right reason? And how does the ConvNet arrive at its final classification decision?

To answer these questions, we need to unravel the neural network internal operations; a challenging problem which has recently attracted considerable attention in the deep learning community<sup>50–55</sup>. The difficulty of this task lies in both the tendency of deep learning models to represent the information in a highly distributed manner, and the presence of non-linearities in the network's layers. This in turn leads to a lack of interpretability which hindered the widespread use of neural networks in natural sciences: linear algorithms are often preferred



**Fig. 4** Neural network predictions on structural transitions. **a, b** Body-centered-cubic (bcc) to simple cubic (sc) structural transition. **a** Examples of a bcc, an intermediate bcc/sc, and a sc structure. **b** Distributions of classification probability for the bcc (purple) and sc (red) classes as a function of the percentage of central atoms being removed (see text for more details). The shaded area corresponds to a range of one standard deviation above and below these distributions. **c, d** Structural transition: transition path including rhombohedral, body-centered cubic, simple cubic and face-centered cubic structures. The prototypes are generated using the AFLOW Library of Crystallographic Prototypes<sup>49</sup>



**Fig. 5** Visualizing the convolutional neural network (ConvNet) attentive response maps. **a** Attentive response maps from the top four most activated filters of the first, third and last convolutional layers for the simple cubic class. The brighter the pixel, the most important is that location for classification. Comparing across layers, we notice that the ConvNet filters are composed in a hierarchical fashion, increasing their complexity from one layer to another. At the third convolutional layer, the ConvNet discovers that the diffraction peaks, and their relative arrangement, are the most effective way to predict crystal classes. **b** Sum of the last convolutional layer filters for all seven crystal classes: the ConvNet learned crystal templates automatically from the data

over more sophisticated (but less interpretable) models with superior performance.

To shed light on the ConvNet classification process, we resort to visualization: using the fractionally strided convolutional technique introduced in ref.<sup>53</sup> we back-project attentive response maps (i.e., filters) in image space<sup>50, 51, 55</sup>. Such attentive response maps—shown in Fig. 5—identify the parts of the image which are the most important in the classification decision<sup>53</sup>.

The top four most activated (i.e., most important) filters from the first, third, and last convolutional layers for each of the three color channels are shown in Fig. 5a for the sc class. The complexity of the learned filters grows layer by layer, as demonstrated by the increasing number of diffraction peaks spanned by each motif. The sum of the last convolutional layer filters for each class is shown in Fig. 5b; they are class templates automatically learned from the data by the ConvNet. Comparing Figs. 2c and 5b, we see that our deep learning model is able to

**Table 2 Architecture of the convolutional neural network used in this work**

| Layer type                      | Specifications                                    |
|---------------------------------|---|
| Convolutional layer             | (Kernel: $7 \times 7$ ; 32 filters)               |
| Convolutional layer             | (Kernel: $7 \times 7$ ; 32 filters)               |
| Max pooling layer               | (Pool size: $2 \times 2$ ; stride: $2 \times 2$ ) |
| Convolutional layer             | (Kernel: $7 \times 7$ ; 16 filters)               |
| Convolutional layer             | (Kernel: $7 \times 7$ ; 16 filters)               |
| Max pooling layer               | (Pool size: $2 \times 2$ ; stride: $2 \times 2$ ) |
| Convolutional layer             | (Kernel: $7 \times 7$ ; 8 filters)                |
| Convolutional layer             | (Kernel: $7 \times 7$ ; 8 filters)                |
| Fully connected layer + dropout | (Size: 128; dropout: 25%)                         |
| Batch normalization             | (Size: 128)                                       |
| Softmax                         | (Size: 7)   |

autonomously learn, and subsequently use, the same features that a domain expert would use. This not only confirms the soundness of the classification procedure but also explains its robustness in terms of generalization.

## Discussion

We have introduced a way of representing crystal structures by means of (easily interpretable) images. Being based on reciprocal space, this descriptor—termed two-dimensional diffraction fingerprint—compactly encodes crystal symmetries, and possesses numerous attractive properties for crystal classification. In addition, it is complementary with existing real-space-based representations<sup>22</sup>, making possible to envision a combined use of these two descriptors. Starting from these diffraction fingerprints, we use a convolutional neural network to predict crystal classes. As a result, we obtain an automatic procedure for crystals classification which does not require any user-specified threshold, and achieves perfect classification even in the presence of highly defective structures. In this regard, we argue that—since materials science data are generated in a relatively controlled environment—defective datasets represent probably the most suitable test to probe the generalization ability of any data-analytics model. Given the solid physical grounds of the diffraction fingerprint representation, our deep learning model is modest in size, which translates in short training and prediction times. Finally, using recently developed visualization techniques, we uncover the learning process of the neural network. Owing to its multi-layered architecture, we demonstrate that the network is able to learn, and then use in its classification decision the same landmarks a human expert would use. Further work is needed to make the approach proposed here unique across space groups and to widen its domain of applicability to non-centrosymmetric crystals, which can exhibit technologically relevant ferroelectric, piezoelectric, or nonlinear optical effects. In accordance with the principle of reproducible research<sup>56, 57</sup>, we also provide an online tutorial<sup>18</sup> where users can interactively reproduce the main results of this work (but also produce their own) within the framework of the NOMAD Analytics Toolkit. As an outlook, our method could also be applied to the problem of local microstructure determination in atomic probe tomography experiments, with the ultimate goal of discovering structural–property relationships in real materials.

## Methods

**Two-dimensional diffraction fingerprint.** First, for each structure in the dataset (specified by a set of atomic coordinates and lattice vectors), we concatenate three random rotations around the three crystal axes to randomize the initial crystal orientation. Then, we construct the standard conventional cell according to ref.<sup>34</sup> using a customized implementation based on the Python Materials Genomics (pymatgen) package<sup>58</sup>; in particular, we use the convention for triclinic cells—

irrespective of the actual lattice type—and no symmetry refinement of the atomic position. This procedure is therefore completely independent from traditional symmetry approaches and robust against randomization of the initial crystal orientation. Finally, we replicate this standard cell in all three directions such that the resulting cluster contains a number of atoms which is as close as possible to a given target number (namely, 250). The size invariance of the diffraction peak locations guarantees that the results are independent from this choice, only the peak widths will slightly change, in accordance with the indeterminacy principle<sup>59</sup> (this was expressly checked for systems ranging from 32 to 1024 atoms). Defective structures are then generated from these supercells by removing or randomly displacing atoms. We have also tested that a random rotation followed by the conventional cell determination applied to already generated defective structures leads to the same result, since this depends on the lattice vectors only.

As mentioned in the main text, we used finite samples instead of periodically repeated crystal structures to explicitly prove the local structure recognition capabilities of the method. Each system is then isotropically scaled by its average atomic bond length (i.e., distance between nearest neighboring atoms). We also noticed that for materials formed by hydrogen or helium the diffraction fingerprint contrast is low due to the small  $f_a^h$  (Eq. (1)) of these elements; H and He are indeed notoriously difficult to detect with x-ray diffraction methods because of their small number of electrons ( $Z = 1$  and  $Z = 2$ , respectively)<sup>36</sup>. However, our main goal here is to introduce a transferable descriptor for crystal structure representation, and not to compare with experimental data. Thus, we are free to choose a different value for the atomic number in order to augment the contrast in the diffraction fingerprint. In particular, we increase the atomic number of the elements by two when calculating the diffraction fingerprint, that is, H is mapped to Li, He to Be, and so on. Moreover, given that the task is to distinguish crystals classes with an image for each system, one needs to choose a wavelength which is much smaller than the spacing between atoms, such that many beams are diffracted simultaneously (because the corresponding Ewald sphere radius is much larger than the lattice spacing)<sup>36</sup>. Therefore, we use a wavelength of  $\lambda = 5.0 \times 10^{-12}$  m for the incident plane wave (Eq. (1)), a wavelength typically used in electron diffraction experiments. Indeed, the two-dimensional diffraction fingerprint bears resemblance to experimental scattering techniques such as single-crystal or selected-area electron diffraction; from this perspective, the angle of rotation could be chosen based on specific crystal orientations<sup>60, 61</sup>.

For the (computational) detector, we use a pixel width and height of  $4.0 \times 10^{-4}$  m, and produce a  $64 \times 64$  pixel image as diffraction fingerprint. Since the direct beam does not carry any structural information, and gives raise to a very bright central diffraction spot which compromises the contrast of high-order peaks, we remove this central spot from the diffraction fingerprint setting to zero the intensity within a radius of five pixels from the image center. The two-dimensional diffraction patterns are calculated using the open-source software Condor<sup>62</sup>.

**Dataset.** Our pristine dataset consists of materials from the AFLOWLIB elemental solid database<sup>47</sup> belonging to centrosymmetric space groups which are represented with more than 50 configurations in the database. Specifically, we extract structures that have a consistent space group classification for different symmetry tolerances, as determined by the Python Materials Genomics (pymatgen)<sup>58</sup> wrapper around the Spglib<sup>11</sup> library with `SYMPREC = {10-3 Å, 10-6 Å, 10-9 Å}` for all except rh and hex structures, for which `SYMPREC = {10-3 Å, 10-6 Å}` is employed since some symmetries are missed for `SYMPREC = 10-9 Å`. This gives us crystal structures belonging to the following space groups: 139 (bcc), 141 (bcc), 166 (rh), 194 (hex), 221 (sc), 225 (fcc), 227 (diam), and 229 (fcc). From this, we apply the defective transformations described in the main text (random displacements, vacancies, and chemical substitutions) to the pristine structures; the resulting dataset is used as test set. For this defective dataset we use labels from the pristine structures because the materials' class will unlikely be changed by the transformations above. To quantify this, let us consider the transformation of bcc into sc crystals for the case of random vacancies as illustrative example. As stated in the main text, an sc structure can be obtained removing all atoms laying at the center of the bcc unit cell (see Fig. 2b). Therefore, for a structure comprising  $N$  atoms, one needs to remove exactly the  $N/2$  atoms which are at the center of the cubic unit cell (note that each corner atom is shared equally between eight adjacent cubes and therefore counts as one atom). For  $N/2$  randomly generated vacancies, the probability of removing all and only these central atoms is  $P_N = 2 \left[ \binom{N}{N/2} \right]^{-1}$  which

—for the removing sizes considered in this work—leads to negligible probabilities ( $P_{64} \approx 10^{-18}$ ,  $P_{128} \approx 10^{-38}$ ). The same holds for chemical substitutions: even in principle they could change the space group (e.g., diamond to zincblende structure), the probability of this to happen is comparable with the example above, and therefore negligible. Finally, in the case of displacements, atoms are randomly moved about their original positions, and—due to this randomness—it is not possible to obtain any long-range re-organization of the crystal, necessary to change the materials' class; moreover, for large displacements the system becomes amorphous (without long-range order).

**Neural network architecture and training procedure.** The architecture of the convolutional neural network used in this work is detailed in Table 2. Training was performed using Adam optimization<sup>63</sup> with batches of 32 images for 5 epochs with



a learning rate  $10^{-3}$ , and cross-entropy as cost function. The convolutional neural network was implemented with TensorFlow<sup>64</sup> and Keras<sup>65</sup>.

**Data availability.** Calculation data can be downloaded from the NOMAD Repository and Archive (<https://www.nomad-coe.eu/>); the uniform resource locators (URLs) are provided in the Supplementary Note 3. Additional data including spatial coordinates and diffraction fingerprint for each structure of the pristine dataset is available at the Harvard Dataverse: <https://doi.org/10.7910/DVN/ZDKBRF>. An online tutorial<sup>18</sup> to reproduce the main results presented in this work can be found in the NOMAD Analytics Toolkit.

Received: 1 September 2017 Accepted: 18 June 2018

Published online: 17 July 2018

## References

- Olson, G. B. Designing a new material world. *Science* **288**, 993–998 (2010).
- Curtarolo, S., Morgan, D., Persson, K., Rodgers, J. & Ceder, G. Predicting crystal structures with data mining of quantum calculations. *Phys. Rev. Lett.* **91**, 135503 (2003).
- Fischer, C. C., Tibbetts, K. J., Morgan, D. & Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nat. Mater.* **5**, 641–646 (2006).
- Nye, J. F. *Physical Properties of Crystals: Their Representation by Tensors and Matrices*, ser. Oxford science publications (Clarendon Press, Oxford, 1985).
- Smith, W. F. & Hashemi, J. *Foundations of Materials Science and Engineering*, ser. McGraw-Hill Series in Materials Science and Engineering (McGraw-Hill, New York, 2004).
- Hahn, T. *International Tables for Crystallography*. International Tables for Crystallography, Vol. A (International Union of Crystallography: Chester, England, 2006). <http://it.iucr.org/Ab/>.
- Stokes, H. T. & Hatch, D. M. FINDSYM: program for identifying the space group symmetry of a crystal. *J. Appl. Crystallogr.* **38**, 237–238 (2005).
- Spek, A. L. Structure validation in chemical crystallography. *Acta Crystallogr. D* **65**, 148–155 (2009).
- Grosse-Kunstleve, R. W. Algorithms for deriving crystallographic space group information. *Acta Crystallogr. A* **55**, 383–395 (1999).
- Englert, U. Symmetry relationships between crystal structures. Applications of crystallographic group theory in crystal chemistry. By Ulrich Müller. *Angew. Chem. Int. Ed.* **52**, 11 973–11 973 (2013).
- Atsushi T. *SpGlib*, <https://atzotogo.github.io/spglib/> (2009).
- Hicks, D. et al. AFLOW-SYM: platform for the complete, automatic and self-consistent symmetry analysis of crystals. *Acta Crystallogr. Sect. A* **74**, 184–203 (2018).
- NOMAD Laboratory. NOMAD. <https://nomad-coe.eu> (2015).
- Ryan, M. P., Williams, D. E., Chater, R. J., Hutton, B. M. & McPhail, D. S. Why stainless steel corrodes. *Nature* **415**, 770–774 (2002).
- Duarte, M. J. et al. Element-resolved corrosion analysis of stainless-type glass-forming steels. *Science* **341**, 372–376 (2013).
- Gault, B., Moody, M. P., Cairney, J. M. & Ringer, S. P. Atom probe crystallography. *Mater. Today* **15**, 378–386 (2012).
- Park, W. B. et al. Classification of crystal structure using a convolutional neural network. *IUCrj* **4**, 486–494 (2017).
- Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. *Tutorial for Insightful Classification of Crystal Structures Using Deep Learning* [https://doi.org/10.17172/NOMAD\\_TUT/2018.05.28-1](https://doi.org/10.17172/NOMAD_TUT/2018.05.28-1) (2018).
- Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
- Behler, J. & Parrinello, M. Generalized neural network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 1–4 (2007).
- Rupp, M., Tkatchenko, A., Müller, K.-R., Lilienfeld, V. & Anatole, O. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 58301 (2012).
- Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
- Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890 (2017).
- Huo, H. & Rupp, M. Unified representation for machine learning of molecules and crystals. Preprint at <http://arxiv.org/abs/1704.06439> (2017).
- Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
- Isayev, O. et al. Universal fragment descriptors for predicting electronic properties of inorganic crystals. *Nat. Commun.* **8**, 15679 (2016).
- Zhu, L. et al. A fingerprint based metric for measuring similarities of crystalline structures. *J. Chem. Phys.* **144**, 034203 (2016).
- Deringer, V. L. & Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **95**, 094203 (2017).
- Morawietz, T., Singraber, A., Dellago, C. & Behler, J. How van der Waals interactions determine the unique properties of water. *Proc. Natl. Acad. Sci. USA* **113**, 8368–8373 (2016).
- Henke, B., Gullikson, E. & Davis, J. X-ray interactions: photoabsorption, scattering, transmission, and reflection at  $E = 50\text{--}30,000$  eV,  $Z = 1\text{--}92$ . *At. Data Nucl. Data Tables* **54**, 181–342 (1993).
- Friedrich, W., Knipping, P. & Laue, M. Interferenzerscheinungen bei Röntgenstrahlen. *Ann. Phys.* **346**, 971–988 (1913).
- THOMSON, G. P. & REID, A. Diffraction of cathode rays by a thin film. *Nature* **119**, 890–890 (1927).
- Wollan, E. O. & Shull, C. G. The diffraction of neutrons by crystalline powders. *Phys. Rev.* **73**, 830–841 (1948).
- Setyawan, W. & Curtarolo, S. High-throughput electronic band structure calculations: challenges and tools. *Comput. Mater. Sci.* **49**, 299–312 (2010).
- Looijenga-Vos, A. & Buerger, M. J. in *International Tables for Crystallography 44–54* (International Union of Crystallography, Chester, England, 2006).
- De Graef, M. & McHenry, M. E. *Structure of Materials: An Introduction to Crystallography, Diffraction and Symmetry* (Cambridge University Press, Cambridge, UK, 2007).
- Ashcroft, N. W. & Mermin, N. D. *Solid State Physics* (Cengage Learning, London, 2011).
- Bengio, Y. Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2**, 1–127 (2009).
- Schmidhuber, J. Deep learning in neural networks: qn overview. *Neural Netw.* **61**, 85–117 (2015).
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
- Hinton, G. E. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
- Hinton, G. E., Osindero, S. & Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput.* **18**, 1527–1554 (2006).
- LeCun, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551 (1989).
- LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1989).
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNetClassification with deep convolutional neural networks. in *Advances in Neural Information Processing Systems 25* (eds Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, New York, 2012).
- Pámies, P. Auspicious machine learning. *Nat. Biomed. Eng.* **1**, 0036 (2017).
- Curtarolo, S. et al. AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
- Taylor, R. H. et al. A RESTful API for exchanging materials data in the AFLOWLIB.org consortium. *Comput. Mater. Sci.* **93**, 178–192 (2014).
- Mehl, M. J. et al. The AFLOW Library of crystallographic prototypes. *Comput. Mater. Sci.* **136**, S1–S828 (2016).
- Zeiler, M. D., Krishnan, D., Taylor, G. W. & Fergus, R. ImageNetClassification with deep convolutional neural networks. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (eds Ma, B., Su, Y. & Jurie, F.) 2528–2535 (IEEE, San Francisco, CA, 2010).
- Zeiler, M. D. & Fergus, R. *Visualizing and Understanding Convolutional Networks* 818–833, [https://doi.org/10.1007/978-3-319-10590-1\\_53](https://doi.org/10.1007/978-3-319-10590-1_53) (2014).
- Bach, S. et al. On pixel-wise explanations for nonlinear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, e0130140 (2015).
- Kumar, D. & Menkovski, V. Understanding anatomy classification through visualization. In *NIPS Machine Learning for Health*, no. Nips 1–5 <http://arxiv.org/abs/1611.06284> (2016).
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W. & Müller, K.-R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit.* **65**, 211–222 (2017).
- Kumar, D., Wong, A. & Taylor, G. W. Explaining the unexplained: a class-enhanced attentive response (CLEAR) approach to understanding deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1686–1694 (IEEE, Honolulu, HI, 2017).
- Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Human. Behav.* **1**, 0021 (2017).
- Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).
- Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).

59. Sakurai, J. J. & Napolitano, J. *Modern Quantum Mechanics* (Addison-Wesley, Waltham, MA, 2011).
60. Bunge, H.-J. H. J. *Texture Analysis in Materials Science: Mathematical Methods* (Butterworths, London, 1982).
61. Britton, T. et al. Tutorial: Crystal orientations and EBSD—or which way is up?. *Mater. Charact.* **117**, 113–126 (2016).
62. Hantke, M. F., Ekeberg, T. & Maia, F. R. N. C. Condor: a simulation tool for flash X-ray imaging. *J. Appl. Crystallogr.* **49**, 1356–1362 (2016).
63. Kingma, D. & Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations* 1–13. Preprint at <http://arxiv.org/abs/1412.6980> (2014).
64. Martin, A. et al. TensorFlow: large-scale machine learning on heterogeneous systems, <https://www.tensorflow.org/> (2015).
65. Chollet, F. Keras, <https://github.com/fchollet/keras> (2015).
66. Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (eds Fürnkranz, J. & Joachims, T.) 807–814 (Omnipress, Madison, WI, 2010).
67. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
68. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).

### Acknowledgements

A.Z., L.M.G., and M.S. acknowledge funding from the European Union's Horizon 2020 research and innovation programme, Grant Agreement No. 676580 through the Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence (<https://www.nomad-coe.eu>). D.K. would like to thank Dr. Vlado Menkovski for helpful discussions regarding visualization.

### Author contributions

A.Z., M.S. and L.M.G. conceived the project. A.Z. performed the calculations. A.Z. and D. K. carried out the classification model visualization. A.Z., M.S., and L.M.G. wrote the manuscript. All authors reviewed and commented on the manuscript.

### Additional information

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-018-05169-6>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018