

## ARTICLE OPEN



# Finding the semantic similarity in single-particle diffraction images using self-supervised contrastive projection learning

Julian Zimmermann<sup>1</sup>✉, Fabien Beguet<sup>1</sup>, Daniel Guthruf<sup>1</sup>, Bruno Langbehn<sup>2</sup> and Daniela Rupp<sup>1,3</sup>

Single-shot coherent diffraction imaging of isolated nanosized particles has seen remarkable success in recent years, yielding in-situ measurements with ultra-high spatial and temporal resolution. The progress of high-repetition-rate sources for intense X-ray pulses has further enabled recording datasets containing millions of diffraction images, which are needed for the structure determination of specimens with greater structural variety and dynamic experiments. The size of the datasets, however, represents a monumental problem for their analysis. Here, we present an automatized approach for finding semantic similarities in coherent diffraction images without relying on human expert labeling. By introducing the concept of projection learning, we extend self-supervised contrastive learning to the context of coherent diffraction imaging and achieve a dimensionality reduction producing semantically meaningful embeddings that align with physical intuition. The method yields substantial improvements compared to previous approaches, paving the way toward real-time and large-scale analysis of coherent diffraction experiments at X-ray free-electron lasers.

npj Computational Materials (2023)9:24; <https://doi.org/10.1038/s41524-023-00966-0>

## INTRODUCTION

A guiding principle in fundamental condensed matter research is that for understanding function, we have to study structure (The phrase was coined by Nobel laureate Francis Crick in his book *What Mad Pursuit: A personal view of scientific discovery*)<sup>1</sup>. Techniques based on lensless diffractive imaging, like X-ray crystallography and coherent diffraction imaging (CDI), are powerful and widely used tools to discover structures up to atomic resolutions<sup>2</sup>. In the recent past, single-particle coherent diffraction imaging using intense coherent X-ray pulses from free-electron lasers (SP-CDI)<sup>2,3</sup> has revolutionized the field of structural characterization<sup>4–12</sup>. SP-CDI is a technique with which in-situ measurements of isolated and non-fixed nano-scaled targets can be acquired. Depending on the experimental scheme, each recorded diffraction image is a complete and self-contained experiment that needs individual analysis<sup>3</sup>. However, due to the advent of high repetition-rate sources like the European XFEL<sup>13</sup> and LCLS-II<sup>14</sup>, millions of images are typically recorded during one experimental campaign<sup>15</sup>. Manual analysis of such amounts of data represents an enormous problem. It may leave researchers unable to analyze significant amounts of their data comparatively as they have to resort to large-scale averaging, which might wash out or conceal important information, or manually select subsets of the dataset. In this work, we present an embedding technique for diffraction images called contrastive projection learning (CPLR) based on contrastive learning (CLR)<sup>16,17</sup>. CPLR produces a dimensionality-reduced embedding space with which semantic comparisons between diffraction images become possible and, thus, enables human-level comparative analysis on big-data scale datasets.

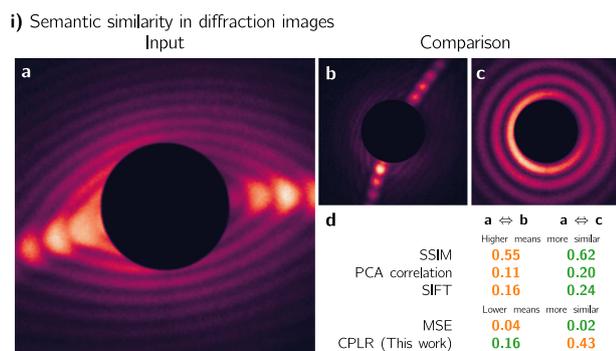
At the very core of every comparative analysis is an assumption establishing a similarity measure between samples. However, current approaches for establishing such a measure for diffraction images cannot compete with the perception of a trained researcher. This perceived similarity, or semantic similarity, is contextually aware; we adopt this term from the domain of natural

language processing. There it is used to differentiate between the *semantic*—contextual—and the *lexicographical*—word for word—similarity<sup>18,19</sup>. The lexicographical measures are in our case pixel-wise or keypoint-based approaches. Please also compare the term *interference* in ref. <sup>20</sup>. In Human-Perceptual-Similarity, interference is the distortion that a judged truth of a property—like the length of a line—can exhibit. Meaning, the perceived length of the line is determined not only by its length, but also by its surrounding), whereas computational methods for diffraction image data currently lack such awareness.

Available strategies for comparative analysis are based on either supervised classifier schemes<sup>21,22</sup> or unsupervised sorting methods<sup>23–27</sup>. However, all these methods come with significant trade-offs: Supervised algorithms align with human perception and produce high-accuracy results<sup>22,28</sup> but are in real-world scenarios unavailable as they require time- and labor-expensive manually fabricated expert labels. Unsupervised routines work in such cases but do not reach comparable accuracy levels and introduce additional restrictions or requirements. For example, traditional cluster techniques produce a lot of unwanted predictions<sup>23</sup>, threshold-based approaches act primarily as hit-finder<sup>24,29</sup>, autocorrelation-based methods only extract particle size information and are computationally costly<sup>25</sup>, auxiliary approaches rely on rarely available additional experimental data<sup>25</sup>, and Fourier-inversion based techniques<sup>15,25,27</sup> are only applicable to diffraction images when the single scattering approximation is valid<sup>30</sup>, meaning, where reconstruction by Fourier inversion is physically meaningful<sup>10,31</sup>.

Our CPLR method can potentially improve all strategies mentioned above; It can serve as an improved similarity measure for unsupervised methods, as in the context of regular self-supervised learning<sup>32–34</sup>, and can act as a powerful pretraining for subsequent supervised or distillation-based training<sup>17</sup>. Furthermore, CPLR directly establishes a way to find semantically similar diffraction images in a fully self-supervised fashion. In self-supervised contrastive learning, a supervised task is constructed by artificially

<sup>1</sup>ETH Zürich, Zürich, Switzerland. <sup>2</sup>Technische Universität Berlin, Berlin, Germany. <sup>3</sup>Max-Born-Institut, Berlin, Germany. ✉email: [zimmermann@phys.ethz.ch](mailto:zimmermann@phys.ethz.ch)



**Fig. 1** Illustrating semantic similarity. **a–c** are diffraction images taken from a publicly available dataset<sup>37</sup> from an SP-CDI experiment on superfluid helium nanodroplets<sup>12,22</sup>. **a, b** are semantically similar while **a** and **c** are not. **d** shows four widely used similarity measures—the structural similarity index (SSIM)<sup>39</sup>, the correlation of the principal components, the mean-squared-error (MSE), and the keypoint-based scale-invariant feature transform (SIFT) distance<sup>40</sup>—that disagree with the human perception, while only our method (labelled CPLR) agrees with it. Color-coding is that the pair of images that is more-similar is green, while the one that is less-similar is orange.

creating label information via domain-specific augmentation strategies<sup>16,35,36</sup>. In this work, we design an augmentation approach for diffraction image data where a deep neural network contrasts images from different coordinate projections.

The quality of the CPLR embedding space is evaluated using a publicly available diffraction image dataset<sup>37</sup> from an SP-CDI experiment on superfluid helium nanodroplets, for which semantically sensitive expert labels are available<sup>12,22</sup>. Using the broadly established linear evaluation protocol<sup>16,17,38</sup>, we show that our method outperforms non-contrastive methods by a large margin while improving the contrastive-learning baseline by 6 to 10%.

Figure 1 provides a concrete example of three diffraction images from<sup>12</sup>, **a, b**, and **c**. Two images (**b** and **c**) are to be compared to **a**. A human immediately identifies the elongated streak-like feature in **a** as the dominant characteristic and can identify **b** as being more similar to **a** than **c** is to **a**. And, indeed, this is correct from a physical perspective. The nanoparticles' structures that produce **a** and **b** are more similar to each other than those that produce **a** and **c**<sup>12</sup>. However, from an algorithmic perspective, this is not a trivial problem. Figure 1**d** shows for **b** and **c** five similarity measures: the structural similarity index (SSIM)<sup>39</sup>, the correlation of the principal components, the mean-squared-error (MSE), the keypoint-based scale-invariant feature transform (SIFT) distance—calculated using the ratio test as in ref.<sup>40</sup>—and our method (labeled CPLR). Color-coding is that the image that is more similar to **a** is green, while the less similar image is orange. The only measure that agrees with the human perception, and the physics of the problem, is our CPLR method.

Our approach is not limited to data from SP-CDI experiments. Theoretically, CPLR can be applied to diffraction data from all experimental techniques operating in Polar coordinates, including X-ray crystallography and traditional CDI approaches. Ultimately, CPLR provides a path for analyzing the impending amounts of diffraction data where the human perceived similarity is maintained even among millions of diffraction images.

## RESULTS

### Contrastive learning is about augmentation, not architecture

In contrastive learning (CLR), we artificially create label information for supervised learning by designing augmentation pipelines that consider domain and task knowledge<sup>16</sup>. Therefore, CLR is an

instance of self-supervised learning<sup>41</sup>. The fundamental assumption in self-supervised learning is that the input data contain more task-specific information than sparse categorical ground truth data in supervised learning<sup>41</sup>. Consequently, a careful augmentation design should provide better results on downstream tasks than a supervised learning scenario<sup>41,42</sup>. While improvements in accuracy in supervised learning are usually related to architecture modifications, regularization, or loss function, CLR is about domain-specific augmentation strategies above anything else<sup>16,41</sup>. Formally, CLR is a technique to create an embedding space from arbitrary input modalities, enabling comparative analysis. CLR dates back to work done in the nineties<sup>42</sup> but only recently has seen a renaissance, yielding State-of-the-art results in visual-<sup>16,17,38,43</sup>, audio-<sup>44–46</sup>, video-<sup>47–49</sup>, and text-representation<sup>50,51</sup> learning.

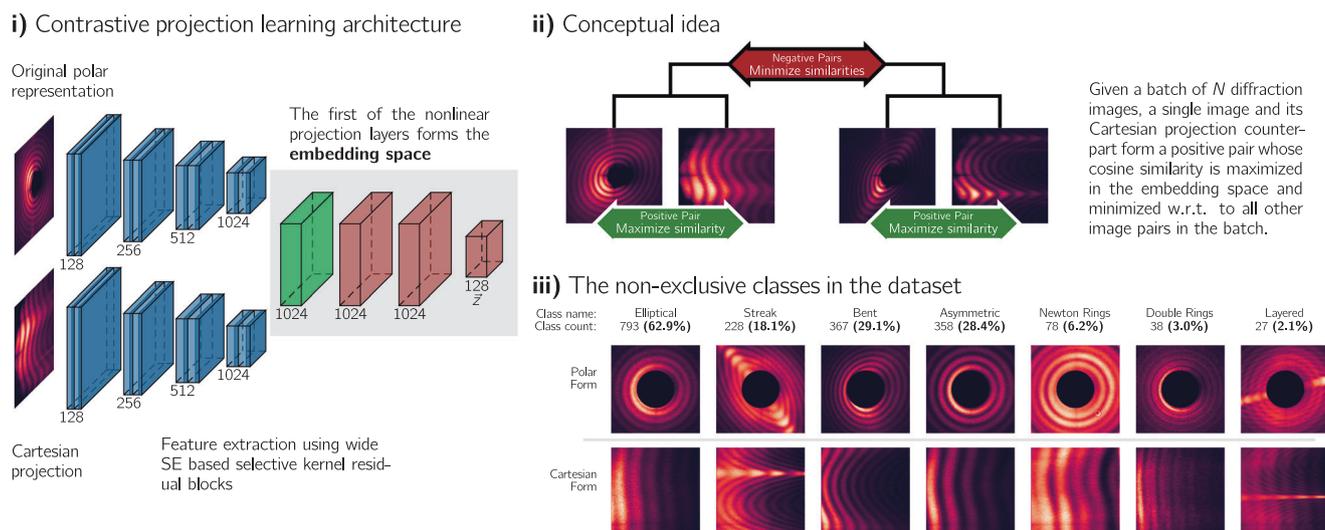
### Contrastive baseline and contrastive projection learning

In this study, we use the experimental design presented in ref.<sup>17</sup>, called SimCLRv2, as a baseline to compare our results. A large encoder and a smaller transformation neural network produce the representations in two stages. First, the encoder acts as a feature extractor; then, the transformation network learns an optimized representation that minimizes the CLR loss, termed normalized temperature-scaled cross-entropy loss (NT-Xent)<sup>52</sup>. Conceptually, a duplicate is produced for each input image where both images are heavily augmented. All duplicates form so-called positive pairs with their originals, while all images with all other images but their duplicates form negative pairs. Then, the network learns to discriminate between positive and negative pairs during training.

In contrastive projection learning (CPLR), we produce the positive pairs not from the same image, as in refs.<sup>16,17</sup>, but project the diffraction images, which are naturally recorded in Polar coordinates, to Cartesian coordinates. Figure 2*i*) and *ii*) provide a schematic overview of the network design and the conceptual idea. Using coordinate projections as an augmentation strategy implicitly penalizes that trivial Polar symmetries are learned and explicitly enforces that learned representations are invariant under rotational and translational changes. A simple example can be constructed with the help of Fig. 2*iii*). There, example images for every class in the dataset are shown. While Fig. 2*iii*) is fully explained in the following subsection, here, we concentrate on the Polar and Cartesian projection of the *Elliptical* class - the first column. The Polar form shows the characteristic Airy rings typical for single-laser-shot and single-particle imaging data<sup>12</sup>. Usually, learning rotational invariance in arbitrary image data is achieved via random rotational transformations during the augmentation stage of training a network. However, such a transformation would yield no, or very little, change with such diffraction images due to the high degree of rotational symmetry. Therefore, we encourage the network to decouple the learned representation from rotational symmetries by associating the Polar form with the Cartesian form. Conceptually, this establishes a causal relation between two fundamentally dissimilar images. In addition, after the coordinate transformation, we leverage a stochastic augmentation pipeline<sup>16,17</sup>. More details can be found in the *Methods* section *Augmentation strategy*.

### The dataset

Helium nanodroplets were imaged at XUV photon energies in a single-shot single-particle experiment<sup>12</sup> at the FERMI free-electron laser<sup>53</sup>. The scattering images were recorded using a non-linear MCP-type detector<sup>54</sup> in a so-called wide-angle setting where each diffraction image contains 3D-structural information and cannot be reconstructed via Fourier inversion<sup>10,12,55</sup>. The publicly available and hand-curated dataset<sup>37</sup> contains 7264 diffraction images and is, to the best of our knowledge, the only dataset with



**Fig. 2** The details on the CPLR architecture, the concept, and the classes within the dataset. **i)** Schematic of the used architecture. In this example, a positive pair, consisting of one diffraction image in its natural Polar and its Cartesian form, are passed on to the feature-extractor network, a modified ResNet50-D<sup>51</sup>. The adjacent grey box then shows the transformation network, a three layer MLP, where the first layer—in green—forms the embedding space within the here used SimCLRv2<sup>17</sup> framework. **ii)** The conceptual idea in more detail. The goal during training is to maximize similarity (minimize distance in embedding space) for positive pairs and minimize similarity (maximize distance in embedding space) for negative pairs. **iii)** The seven possible, non-exclusive classes of the here used multiclass and multilabel dataset. The top row shows the diffraction image in Polar coordinates and the bottom row the Cartesian projection. In addition, the absolute and relative class counts are given alongside the class names. Several different features can appear in a single image which then belongs to several classes simultaneously. For example, a *Streak* feature is also present in the example for the *Layered* class.

semantically sensitive labels. It has been extensively benchmarked in a supervised classification task<sup>22</sup>. We discarded 6000 diffraction images as they either exhibited strictly round or no Airy patterns at all. The round Airy patterns are by far the most common class, and we removed them to create a more balanced dataset since they can be reliably sorted using radial slices and a classical peak-finder (see the section *Size distribution of helium nanodroplets* in the supplemental material of ref. <sup>12</sup>). However, more research should go toward adapting recent advances in few- or single-shot learning<sup>56–58</sup> for diffraction imaging as a preprocessing step to tackle inherent class imbalances that such datasets often have with minimal labeling effort.

The provided expert labels can be used for multiclass and multilabel analysis, meaning every diffraction image has binary label information for multiple classes that are often mutually non-exclusive. Figure 2iii) shows this dataset's seven possible classes and their absolute and relative occurrence. For every class, one example is given in Polar and Cartesian form. To illustrate the multilabel property: The characteristic streak-like feature that defines the *Streak* class can also be found, for example, in the image for the *Layered* class. For this reason, the given percentages do not add up to 100%, as multiple images belong to multiple classes and most classes are heavily under-represented. This pronounced multilabel imbalance is typical for diffraction image datasets<sup>9,12</sup>.

### Training and evaluation

As in ref. <sup>17</sup>, we use a  $2x$ -wide<sup>59</sup>, selective-kernel<sup>60</sup> ResNet50-D<sup>61</sup> network with squeeze-excitation blocks<sup>62</sup> as feature extractor, and a three-layer multi-layer-perceptron (MLP) as transformation network. We train for 1000 epochs, with a batch size of 628, using the LARS optimizer<sup>63</sup> with a cosine schedule<sup>64</sup>, ten warmup epochs<sup>65</sup>, and optimizing the NT-Xent loss<sup>52</sup>. Our projection approach adds negligible computational cost to the vanilla CLR baseline; We train for half an hour on four NVIDIA 3090 GPU. Training for 1000 epochs on datasets with more than a million samples can be achieved using, for example, 128 TPU v3 cores for 15 h<sup>16</sup>. The code, pretrained models, and training results are

openly available at [https://gitlab.ethz.ch/nux/machine-learning/contrastive\\_projection\\_learning](https://gitlab.ethz.ch/nux/machine-learning/contrastive_projection_learning).

The quality of the learned embedding space is evaluated using the so-called linear evaluation protocol<sup>16,17,38</sup>, which is carried out as follows: After self-supervised training, we freeze the feature extractor network and use the first layer of the transformation network (this is indicated by the green layer in Fig. 2ii)) as embedding space. Then, for every ground truth class in the dataset, we train a linear classifier on top of the learned representations and calculate the precision and recall score. Both metrics are obtained for every class via 5-fold stratified cross-validation to account for statistical fluctuations from sampling and the dataset's class imbalance. precision is a metric for how many of the non-zero predictions are actually correct and recall gives the percentage how many labels were missed. So, if 5 out of 10 samples have a label of 1 and the rest a label of 0, and we predict one of the five to be 1 but the rest as 0, then precision is 100%, but recall is only 20%, we refer the reader to section 11.1 in<sup>66</sup> for further details. It has turned out to be important to use precision and recall as most classes are very rare, with three out of seven classes appearing in under 7% of all images. In these cases, accuracy would produce very high scores for classifiers predicting every image as not being part of any class. Moreover, we calculate an additional metric called overlap score in order to compare our method with metrics operating on raw images, such as the structural similarity (SSIM) index<sup>39</sup>, the complex wavelet SSIM (CWSSIM) index<sup>67</sup>, and the keypoint-based scale-invariant feature transform (SIFT) distance<sup>40</sup>. Let  $S_{\text{Input}}$  and  $S_{\text{Neighbor}}$  be the sets of labels for the input image and for an image for which the overlap should be calculated. Then, the overlap is the ratio between the cardinality of the intersection of these two sets ( $|S_{\text{Input}} \cap S_{\text{Neighbor}}|$ ) and the minimum of the individual cardinalities ( $\min(|S_{\text{Input}}|, |S_{\text{Neighbor}}|)$ ). Therefore, overlap is a measure of agreement between two images. However, it is based directly on the label information for the given images and not on the actual distance in embedding space. The reported values in Tables 1 and 2 are the global average in overlap for every image and its 13 closest images according to

**Table 1.** Macroaverage (The macroaverage computes the average independently for each class and then takes the global average. Thereby, treating all classes equally. A microaverage aggregates each contribution of all classes first and averages that. For a discussion on macro- and microaveraging see section 13.6 in ref. <sup>68</sup>.) results on the helium nanodroplets dataset.

Method	Linear evaluation		Local similarity	
	Precision	Recall	Measure	Overlap
<b>Contrastive-based</b>				
CLR				
Polar $\leftrightarrow$ Polar	0.49	0.50	Cosine	0.49
Cartesian $\leftrightarrow$ Cartesian	0.49	0.48	Cosine	0.43
CPLR				
Cartesian $\leftrightarrow$ Polar	0.51	0.54	Cosine	<b>0.52</b>
Polar $\leftrightarrow$ Cartesian	<b>0.52</b>	<b>0.55</b>	Cosine	<b>0.52</b>
<b>Continuous latent variables methods</b>				
Factor Analysis	0.28	0.35	Euclidean	0.28
PCA	0.29	0.35	Euclidean	0.40
			Correlation	0.43
Kernel PCA	0.28	0.35	Euclidean	0.43
			Correlation	0.40
UMAP <sup>75</sup>	0.30	0.31	Euclidean	0.41
<b>Variational Bayesian methods</b>				
VAE (A categorical distribution with number of categories $k = 1$ , and with probabilities $p_n = 1/n_{\text{label}}$ ) <sup>76,77</sup>	0.34	0.35	Wasserstein $W_2$	0.42
<b>Random baseline</b>				
Uniform Noise	0.23	0.27	Euclidean	0.28
<b>Direct measures</b>				
SSIM (see section 13.6 in ref. <sup>68</sup> )	N/A	N/A	Custom	0.37
CWSSIM <sup>67</sup>	N/A	N/A	Custom	0.36
SIFT <sup>40</sup>	N/A	N/A	Euclidean	0.32

Contrastive-based are the results for the CLR baseline and the CPLR method, where the arrows indicate with which projections the contrastive task was constructed. Continuous latent variables or variational Bayesian methods list techniques that have been used with diffraction data in the past. Random baseline gives the result for an artificial embedding space built from uniform noise, this is the lowest possible score. Direct measures are applied directly to the images, and cannot be evaluated using the linear evaluation protocol. All methods except for the Direct measures are trained/run for five times were each time the evaluation scores were obtained via 5-fold stratified cross-validation. The standard deviation of all methods is equal or below 0.01. The best result for each score is given in bold letters.

the pairwise-calculated distances. 13 was chosen as it corresponds to 1% of all images in the dataset. We call this global average the overlap score, opposed to simply overlap, and consider it to be a local-neighborhood evaluation of disentanglement in the embedding space. Consequently, an overlap score of 0.5 corresponds to: On average, the 13 most similar images shared 50% of the original image's labels.

We compare our method to the SimCLRv2 baseline and other approaches that have been used in the past with diffraction images in the following sections. More details on training and evaluation can be found in the *Methods* section *Training and linear evaluation strategy*.

### The embedding space is linearly separable into semantic features

The evaluation scores after training are provided in Table 1. The second and third columns show the precision and recall score of

the linear evaluation protocol, and the last two columns show the used metric for calculating the pairwise distances to calculate the overlap score, which is given in the last column. *Contrastive-based* shows the results for the CLR baseline and our CPLR method. The arrows indicate the coordinate projections used, where the first term is used for inference and the second is used for constructing the contrastive task. Consequently, *Polar*  $\leftrightarrow$  *Polar* and *Cartesian*  $\leftrightarrow$  *Cartesian* are cases of the standard CLR framework with either purely unmodified (*Polar*) or purely projected (*Cartesian*) diffraction images. *Cartesian*  $\leftrightarrow$  *Polar* and *Polar*  $\leftrightarrow$  *Cartesian* are cases of our CPLR method, where the difference between the two is that we changed the input for inference at evaluation time to either the *Cartesian* or the *Polar* form. *Continuous latent variables* and *variational Bayesian* methods are techniques used previously in the context of diffraction images. More details are given in the *Methods* section *Non-contrastive-based methods*. *Random baseline* gives the result for an artificial embedding space built from uniform noise; this is the lowest possible score. This baseline is equivalent to a random guesser with no learned information about the dataset. *Direct measures* are methods applied directly to the images, which cannot be evaluated using the linear evaluation protocol.

All methods except for the *Direct measures* are trained/run five times using five different integer random\_state keys where the evaluation scores were each time obtained via 5-fold stratified cross-validation. The macro average see section 13.6 in ref. <sup>68</sup>) over all classes for all train and cross-validation runs is given in the table. The standard deviation of all methods is equal to or below 0.01.

Compared to the CLR baseline, the CPLR method yields significant improvements on all metrics improving precision relative by 6% (from 0.49 to 0.52) and recall by 10% (from 0.50 to 0.55). Relative to the best non-CLR methods (VAE for precision and all but UMAP for recall), the CPLR method improves precision by 35% (from 0.34 to 0.52) and recall by 36% (from 0.35 to 0.55). In addition, the overlap score is relatively improved by about 6% (from 0.49 to 0.52) compared to the CLR baseline and 17% (from 0.43 to 0.52) compared to the two best PCA-based approaches. CPLR is the only method that achieves precision and overlap scores above 0.50. To put these results into perspective, a fully supervised ResNet architecture can achieve a global Precision score of 0.922 and a Recall of 0.870 on this dataset<sup>22</sup>. So, while the CPLR method provides a significant improvement over all other methods, the gap in performance highlights the difficulty that diffraction imaging datasets represent, as contrastive learning using non-scientific data already achieves accuracies comparable to supervised learning scenarios<sup>16,17</sup>.

### CPLR is more robust with fewer samples

The general idea of using the linear evaluation protocol for evaluation is to look for linearly separable regions in the embedding space. Therefore, this method only applies to one-hot (a multinomial distribution where the number of trials ( $n$ ) is equal to the number of possible labels ( $n_{\text{label}}$ ) and with probabilities  $p_n = 1/n_{\text{label}}$ ) ground truth data, meaning multiclass but single label. However, the helium nanodroplets dataset has multiclass and multilabel (a categorical distribution with number of categories  $k = 1$ , and with probabilities  $p_n = 1/n_{\text{label}}$ ) ground truth data, where each image has multiple associated labels. Moreover, as typical in datasets on helium nanodroplets, the dataset is heavily unbalanced, where simpler shapes, like *Elliptical*, dominate other classes<sup>9</sup>.

It is, therefore, instructive to look at the individual averages for every class, which are given in Table 2. The CPLR method performs significantly better than the CLR baseline and non-contrastive methods in linear evaluation. The most significant improvement is with rarely occurring classes that appear only in  $\leq 7\%$  of all images.

**Table 2.** Results for every class in the helium nanodroplets dataset.

Class	$n_{\text{abs}}$	$n_{\text{rel}}$	CPLR			CLR			VAE			Kernel PCA		
			Linear eval.		Similarity	Linear eval.		Similarity	Linear eval.		Similarity	Linear eval.		Similarity
			Precision	Recall	Overlap	Precision	Recall	Overlap	Precision	Recall	Overlap	Precision	Recall	Overlap
Elliptical	793	62.9 %	<b>0.79</b>	<b>0.77</b>	0.65	0.78	<b>0.77</b>	0.65	0.71	0.71	<b>0.66</b>	0.68	0.64	0.61
Streak	228	18.1 %	<b>0.91</b>	<b>0.90</b>	0.66	0.89	0.86	<b>0.67</b>	0.83	0.72	0.61	0.46	0.47	0.57
Bent	367	29.1 %	<b>0.51</b>	<b>0.53</b>	0.48	<b>0.51</b>	0.51	<b>0.49</b>	0.40	0.50	0.44	0.35	0.41	0.41
Asymmetric	358	28.4 %	<b>0.42</b>	0.38	0.55	0.38	<b>0.40</b>	0.55	0.33	0.33	<b>0.56</b>	0.29	0.37	0.52
Newton Rings	78	6.2 %	<b>0.32</b>	<b>0.34</b>	<b>0.51</b>	0.24	0.25	0.48	0.06	0.04	0.48	0.09	0.21	0.44
Double Rings	38	3.0 %	<b>0.34</b>	<b>0.40</b>	<b>0.45</b>	<b>0.34</b>	<b>0.40</b>	0.43	0.05	0.04	0.43	0.06	0.22	0.41
Layered	27	2.1 %	<b>0.37</b>	<b>0.37</b>	<b>0.33</b>	0.31	0.32	0.30	0.02	0.09	0.26	0.08	0.25	0.29

Compared to table 1, we only show the four best-performing methods, namely our CPLR (Polar  $\leftrightarrow$  Cartesian) method along with the CLR (Polar  $\leftrightarrow$  Polar) baseline, and the VAE, and Kernel PCA (Using the euclidean metric) approaches. All methods have been trained/run for five times where each time the evaluation scores were obtained via 5-fold stratified cross validation. The standard deviation of all methods and for all classes is equal or below 0.01. The best result for each score is highlighted in bold letters.

VAE and PCA-based techniques fail entirely to place these diffraction images in a linearly separable region of the embedding space, resulting in poor precision and recall scores. However, the CLR baseline also yields limited success in the case of radial symmetry-breaking features like the *Newton Rings* and *Layered* class. There, the diffraction images contain features that either break radial symmetry (*Layered*) or introduce a second radially symmetric feature (*Newton Rings*), cf. Fig. 2iii), which in combination with a low class-count brings the CLR method to its limits. The symmetry-breaking projection of the CPLR method helps in those cases and yields better results when fewer images are available.

A qualitative comparison of the CPLR, CLR, and VAE results is given in Fig. 3i) to iii). In each plot, the column *Input image* shows the same three diffraction images, randomly chosen out of the three classes *Streak*, *Layered*, and *Double Rings*. Next to each input image are the four diffraction images belonging to the four closest embeddings in the embedding space. Additionally, every image is augmented in the top left corner by the class names given in the ground truth data and, for all images but the input image, by the overlap score with its corresponding input image.

The images for the *Streak/Bent* class combination in the first row show strong overlap scores for all three frameworks, where the averages of the shown four nearest neighbors are 1.00 for CPLR, 0.92 for CLR, and 0.79 for VAE. However, only the contrastive-based methods placed those embeddings of images next to each other where the characteristic *Streak* feature is orientated and elongated differently than in the input image. We consider this a strength highlighting that both contrastive methods focus more on the semantics within a diffraction image than the pixel-wise similarity. This can also be seen in the *Streak/Bent/Layered* class combination in the second row of Fig. 3, where the direction of the characteristic *Streak* and *Bent* features vary substantially in size and orientation for both CLR-based methods but are identically aligned within the nearest neighbors of the VAE framework.

As already discussed above, our CPLR method outperforms the baseline CLR methods, especially in scenarios with low sample counts and symmetry-breaking features. This behavior is best seen in the second row for the *Streak/Bent/Layered* class combination. Only 27 images in the dataset have a *Layered* label assigned to them, and the CLR method fails to learn this characteristic having an average overlap score of 0.25 for this example and only placing one additional image with a *Layered* label near the input image in the embedding space. On the other hand, the CPLR method performs significantly better, with an average overlap score of 0.67

and placing two images with a *Layered* label, two with a *Streak*, and one with a *Bent* label next to the input image.

This observation also holds for the third example with the *Elliptical/Double Rings* example, where the CPLR method reaches an average overlap score of 0.58, compared to 0.54 and 0.38 for the CLR and the VAE method, respectively.

These qualitative observations, along with the quantitative results presented above, show conclusively that the CPLR method is introducing significant improvements compared to previous methods for finding the semantic similarity in diffraction images and the baseline CLR method.

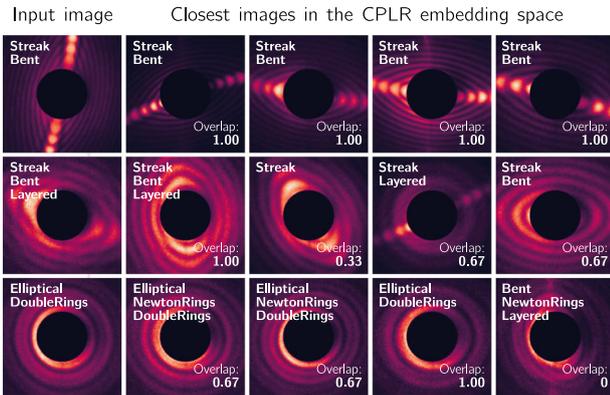
## DISCUSSION

We have introduced a method for finding the semantic similarities in diffraction images without relying on expert labeling. Based on contrastive learning (CLR), we introduced contrastive projection learning (CPLR), where the contrastive learning task is constructed from coordinate-projections of an input diffraction image and not from the same image as in CLR. This relatively easy alternation of the learning scenario substantially improves the quality of the learned embedding space on all metrics and scores. CPLR, therefore, provides a much-needed pathway for the upcoming big-data challenges within the coherent diffraction imaging (CDI) community since similarity calculations are at the core of almost every segmentation, classification, and clustering algorithm. Consequently, CPLR can be implemented as a stand-in-replacement for other similarity metrics in all so-far published classification and clustering approaches for diffraction images<sup>21–27,69–71</sup>, potentially improving a wide range of long-established working routines in research groups.

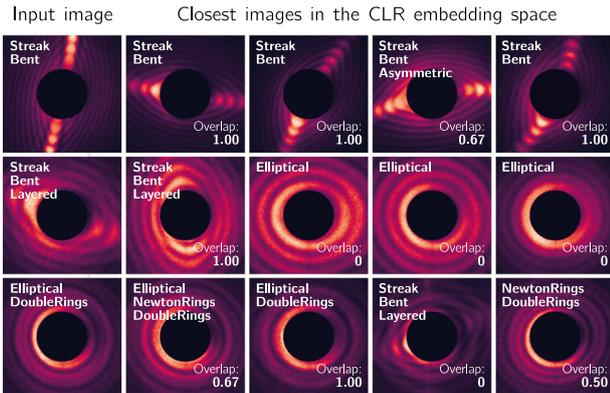
In addition, our method can, theoretically, also be applied to all data that inherit Polar symmetry, such as in X-ray crystallography.

Our results have the potential to enable multiple future possibilities. For example, currently, 3D reconstruction via CDI methods can either be done in the small-angle regime, where reconstruction by Fourier inversion is possible<sup>10,31</sup>, using the Expand-Maximize-Compress (EMC) algorithm<sup>72</sup>, or in the wide-angle regime, via a recursive forward-fitting Multi-Slice-Fourier-Transform (MSFT) approach<sup>10,31,55</sup>. In both cases, the similarity between diffraction images needs to be calculated. As of today, the EMC method can be applied to datasets on the order of millions of images<sup>15,27</sup>. However, the similarity calculation is currently done using the cross-correlation between radial intensity profile lines of the diffraction images at different angles<sup>15,27</sup>, which is computationally costly, and, as the authors in<sup>27</sup> pointed out,

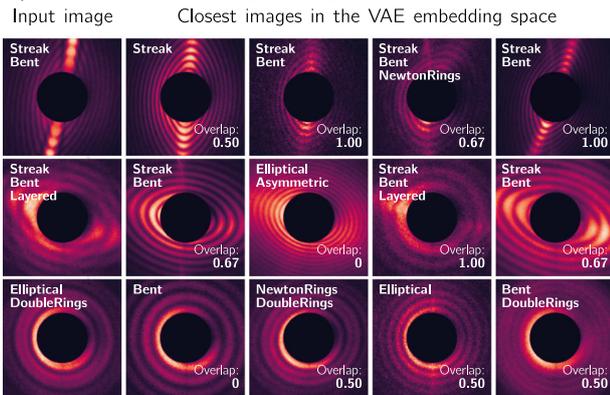
## i) The most similar images using the CPLR framework



## ii) The most similar images using the CLR framework



## iii) The most similar images using the VAE framework



**Fig. 3** Qualitative results on the helium nanodroplets dataset. In i) to iii) the column *Input image* shows three randomly chosen images—from a pre-defined class combination. Next to the input images are the four closest diffraction images according to calculated pairwise distances for the CPLR, CLR, and VAE method using the metric given in Table 1. Every image is augmented in the top left corner by the class names given in the ground truth data and—for all images but the input image—the overlap score with its corresponding input image.

may not be sufficient for more complex patterns. As with the MSFT method, similarity calculations are currently done using either the MSE or even manual estimation by researchers<sup>10,12,31,73</sup>.

Ultimately, CPLR can provide a path to apply the EMC algorithm on more complex datasets, get better results on simpler datasets, and replace the MSE metric in MSFT-based approaches.

Furthermore, it enables quick and reliable statistical reasoning on the variability and occurrence of features within diffraction image datasets, as was done in ref.<sup>12</sup>, for example.

Finally, recent research on contrastive methods in computer vision<sup>43,74</sup> promises accuracies comparable to supervised methods or surpassing them and can be easily implemented into our framework.

Therefore, this manuscript stands as a stepping stone for adapting self-supervised learning to the domain of diffraction imaging.

All code for the discussed experiments, pretrained models, and extracted embedding spaces are available at our ETH Gitlab repository ([https://gitlab.ethz.ch/nux/machine-learning/contrastive\\_projection\\_learning](https://gitlab.ethz.ch/nux/machine-learning/contrastive_projection_learning)).

## METHODS

### Augmentation strategy

A well-defined augmentation strategy is critical in contrastive learning<sup>16</sup>. As pointed out by<sup>16,17</sup>, the essential parts of constructing this strategy are random cropping and random color distortion transformations. The latter is targeted towards histogram and color-channel correlation-based overfitting of the network. Since diffraction data is monochrome, we replace the channel-independent RGB distortion with a single-channel jitter distortion. Furthermore, as in<sup>16,17</sup>, we use a probabilistic augmentation strategy that includes flip, rotation, crop & resize, jitter, fill, and translation transformations on all input patches, see also figure 4 in ref.<sup>16</sup>. However, our crop & resize routine is not changing the aspect ratio, as is usually done in other contrastive learning augmentation pipelines. Changing the aspect ratio would break the causal relation between the *Polar* and *Cartesian* projections. Every transformation has a fixed probability of 50% for being applied at every invocation. We found that no single augmentation strategy is sufficient for achieving the here presented performance, but the composition of all operations is needed, which is in line with the findings in<sup>16</sup>. We implemented the entire pipeline using TensorFlow augmentation layers placed on the GPU itself. Code is available in the official repository ([https://gitlab.ethz.ch/nux/machine-learning/contrastive\\_projection\\_learning](https://gitlab.ethz.ch/nux/machine-learning/contrastive_projection_learning)).

### Training and linear evaluation strategy

The NT-Xent loss that is minimized during training is given by:

$$l_{ij} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}, \quad (1)$$

where  $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T \mathbf{v} / (\|\mathbf{u}\| \|\mathbf{v}\|)$  denotes the cosine similarity between two vectors  $\mathbf{u}$  and  $\mathbf{v}$ ,  $\mathbf{1}_{[k \neq i]} \in \{0, 1\}$  is an indicator function evaluating to 1 if, and only if,  $k \neq i$ , and  $\tau$  denotes a temperature parameter. We performed extensive hyper-parameter optimization to obtain the best possible values for the temperature parameter  $\tau$ , which are 0.200 for Polar  $\leftrightarrow$  Polar, 0.200 for Cartesian  $\leftrightarrow$  Cartesian, 0.075 for Cartesian  $\leftrightarrow$  Polar, 0.100 for Polar  $\leftrightarrow$  Cartesian. The results of this hyper-search, as well as scripts to re-run it, can be found in the official repository ([https://gitlab.ethz.ch/nux/machine-learning/contrastive\\_projection\\_learning](https://gitlab.ethz.ch/nux/machine-learning/contrastive_projection_learning)).

The linear classifier we used for linear evaluation was a single-layer perceptron with an inverse-scaling learning rate schedule and a l2 penalty of 0.001. We used the implementation provided by the *sklearn* Python package. Code is available in the official repository ([https://gitlab.ethz.ch/nux/machine-learning/contrastive\\_projection\\_learning](https://gitlab.ethz.ch/nux/machine-learning/contrastive_projection_learning)).

### Non-contrastive-based methods

Listed in Table 1 are the Factor-Analysis (FA), the Principal-Component-Analysis (PCA), the Kernel-PCA, the Uniform Manifold Approximation & Projection (UMAP)<sup>75</sup>, and the Variational Autoencoder (VAE)<sup>76,77</sup> methods. All of these have been used

with various forms of spectrographic image data. FA- and PCA-based methods are parameter-free dimensionality reduction techniques that are regularly used within all scientific disciplines; while FA considers the dataset's variance, PCA considers the covariance of the data. FA and PCA-based methods have been used with powder diffraction data<sup>78,79</sup> and X-ray diffraction phase analysis<sup>80</sup> and as a dimensionality reduction for subsequent classification<sup>81</sup> and clustering<sup>23</sup>.

A VAE is a generative variational Bayesian model where the input information is encoded to a low dimensional representation via an encoder function and then recreated by a decoder function. The loss function, called the evidence lower bound, is a lower bound on the marginal likelihood<sup>76</sup>. VAEs have been used with diffraction images in various tasks, such as anomaly-detection<sup>82</sup>, dimensionality reduction<sup>83</sup>, phase reconstruction<sup>84,85</sup>, and modeling the continuous 3D shape transition in heterogeneous samples<sup>27</sup>. We train a  $\beta$ -VAE with  $\beta = 1$  and with a controlled capacity increase to  $C = 25$ , as described in<sup>77</sup>, using the code from <https://gitlab.ethz.ch/nux/machine-learning/disentangling-vae>.

UMAP is a dimensionality reduction technique based on manifold learning and topological data analysis and has been used with other spectrographic image data, such as Ronchigrams<sup>86</sup> and Audio spectrograms<sup>87,88</sup>. We use UMAP with the default parameters and a fixed integer random\_state for reproducibility.

The size of the low-dimensional representation for all mentioned methods was set to 1024, identical to the dimensionality of the CLR-based representation space.

## DATA AVAILABILITY

All code for the discussed experiments, pretrained models, data for the figures, and extracted embedding spaces of all models are available at our ETH Gitlab repository ([https://gitlab.ethz.ch/nux/machine-learning/contrastive\\_projection\\_learning](https://gitlab.ethz.ch/nux/machine-learning/contrastive_projection_learning)).

## CODE AVAILABILITY

The code, pretrained models, and training results are openly available at ([https://gitlab.ethz.ch/nux/machine-learning/contrastive\\_projection\\_learning](https://gitlab.ethz.ch/nux/machine-learning/contrastive_projection_learning)).

Received: 23 September 2022; Accepted: 10 January 2023;

Published online: 16 February 2023

## REFERENCES

- Crick, F. & Anderson, P. W. *What Mad Pursuit: A Personal View of Scientific Discovery* (Pumyang, 1991).
- Miao, J., Ishikawa, T., Robinson, I. K. & Murnane, M. M. Beyond crystallography: diffractive imaging using coherent x-ray light sources. *Science* **348**, 530–535 (2015).
- Chapman, H. N. & Nugent, K. A. Coherent lensless x-ray imaging. *Nat. Photonics* **4**, 833–839 (2010).
- Seibert, M. M. et al. Single mimivirus particles intercepted and imaged with an x-ray laser. *Nature* **470**, 78–81 (2011).
- Bostedt, C. et al. Clusters in intense FLASH pulses: ultrafast ionization dynamics and electron emission studied with spectroscopic and scattering techniques. *J. Phys. B At. Mol. Opt. Phys.* **43**, 194011 (2010).
- Loh, N. D. et al. Fractal morphology, imaging and mass spectrometry of single aerosol particles in flight. *Nature* **486**, 513–517 (2012).
- Xu, R. et al. Single-shot three-dimensional structure determination of nanocrystals with femtosecond x-ray free-electron laser pulses. *Nat. Commun.* **5**, 4061 (2014).
- Gorkhover, T. et al. Nanoplasma dynamics of single large xenon clusters irradiated with superintense x-ray pulses from the linac coherent light source free-electron laser. *Phys. Rev. Lett.* **108**, 245005 (2012).
- Gomez, L. F. et al. Helium superfluidity, shapes and vorticities of superfluid helium nanodroplets. *Science* **345**, 906–909 (2014).
- Barke, I. et al. The 3d-architecture of individual free silver nanoparticles captured by x-ray scattering. *Nat. Commun.* **6**, 6187 (2015).

- Ekeberg, T. et al. Three-dimensional reconstruction of the giant mimivirus particle with an x-ray free-electron laser. *Phys. Rev. Lett.* **114**, 098102 (2015).
- Langbehn, B. et al. Three-dimensional shapes of spinning helium nanodroplets. *Phys. Rev. Lett.* **121**, 255301 (2018).
- Tschentscher, T. et al. Photon beam transport and scientific instruments at the European XFEL. *NATO Adv. Sci. Inst. Ser. E Appl. Sci.* **7**, 592 (2017).
- Stohr, J. Linac coherent light source II (LCLS-II) conceptual design report. Tech. Rep. SLAC-R-978, SLAC National Accelerator Laboratory (United States). Funding organisation: US Department of Energy (United States) (2011).
- Ayyer, K. et al. 3D diffractive imaging of nanoparticle ensembles using an x-ray laser. *Optica* **8**, 15 (2021).
- Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, Volume 119 in PMLR'20, 1597–1607 (2020).
- Chen, T., Kornblith, S., Swersky, K., Norouzi, M. & Hinton, G. Big self-supervised models are strong semi-supervised learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Article 1865 in NIPS'20, 22243–22255 (2020).
- Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I. & Specia, L. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 1–14 (Association for Computational Linguistics, Vancouver, Canada, 2017).
- Chandrasekaran, D. & Mago, V. Evolution of semantic Similarity—A survey. *ACM Comput. Surv.* **54**, 1–37 (2021).
- Santini, S. & Jain, R. Similarity measures. *IEEE Trans. Pattern Anal. Mach. Intell.* **21**, 871–883 (1999).
- Bobkov, S. A. et al. Sorting algorithms for single-particle imaging experiments at x-ray free-electron lasers. *J. Synchrotron Radiat.* **22**, 1345–1352 (2015).
- Zimmermann, J. et al. Deep neural networks for classifying complex features in diffraction images. *Phys Rev E* **99**, 063309 (2019).
- Yoon, C. H. et al. Unsupervised classification of single-particle x-ray diffraction snapshots by spectral clustering. *Opt. Express* **19**, 16542–16549 (2011).
- Park, H. J. et al. Toward unsupervised single-shot diffractive imaging of heterogeneous particles using x-ray free-electron lasers. *Opt. Express* **21**, 28729–28742 (2013).
- Andreasson, J. et al. Automated identification and classification of single particle serial femtosecond x-ray diffraction data. *Opt. Express* **22**, 2497–2510 (2014).
- Rose, M. et al. Single-particle imaging without symmetry constraints at an x-ray free-electron laser. *IUCr* **5**, 727–736 (2018).
- Zhuang, Y. et al. Unsupervised learning approaches to characterizing heterogeneous samples using x-ray single-particle imaging. *IUCr* **9**, 204–214 (2022).
- Ribeiro, M. T., Singh, S. & Guestrin, C. “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144) (2016, August).
- Barty, A. et al. Cheetah: software for high-throughput reduction and analysis of serial femtosecond x-ray diffraction data. *J. Appl. Crystallogr.* **47**, 1118–1131 (2014).
- Bohren, C. F. & Huffman, D. R. *Absorption and Scattering of Light by Small Particles* (Wiley, 1983).
- Colombo, A. et al. The scatman: an approximate method for fast wide-angle scattering simulations. *J. Appl. Crystallogr.* **55**, 1232–1246 (2022).
- Zhuang, C., Zhai, A. L. & Yamins, D. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, vol. 2019-October, 6002–6012 (Institute of Electrical and Electronics Engineers Inc., 2019).
- Caron, M. et al. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Article 831 in NIPS'20, 9912–9924 (Curran Associates Inc., Red Hook, NY, USA, 2020).
- Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M. & Van Gool, L. SCAN: Learning to classify images without labels. In *Computer Vision – ECCV 2020*, 268–285 (Springer International Publishing, 2020).
- Robinson, J. D. et al. Can contrastive learning avoid shortcut solutions? Conference on Neural Information Processing Systems (2021).
- Chen, X., Hsieh, C.-J. & Gong, B. When vision transformers outperform ResNets without pre-training or strong data augmentations. In *International Conference on Learning Representations* (2022).
- Coherent X-ray Imaging Data Bank. CXIDB ID 94. <https://doi.org/10.11577/1496209>.
- van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at <https://arxiv.org/abs/1807.03748> (2018).
- Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).

40. Lowe, D. G. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1150–1157 vol.2 (IEEE, 1999).
41. Liu, X. et al. Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* 1–1 (2021).
42. Becker, S. & Hinton, G. E. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature* **355**, 161–163 (1992).
43. Tomasev, N. et al. Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet? Preprint at <https://arxiv.org/abs/2201.05119> (2022).
44. Al-Tahan, H. & Mohsenzadeh, Y. CLAR: Contrastive learning of auditory representations. In *International Conference on Artificial Intelligence and Statistics* (pp. 2530–2538). PMLR, (2021, March).
45. Wang, L. & van den Oord, A. Multi-Format contrastive learning of audio representations. Preprint at <https://arxiv.org/abs/2103.06508> (2021).
46. Saeed, A., Grangier, D. & Zeghidour, N. Contrastive learning of General-Purpose audio representations. In *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3875–3879 (2021).
47. Liu, Y., Wang, K., Liu, L., Lan, H. & Lin, L. TCGL: temporal contrastive graph for Self-Supervised video representation learning. *IEEE Trans. Image Process.* **31**, 1978–1993 (2022).
48. Dave, I., Gupta, R., Rizve, M. N. & Shah, M. TCLR: temporal contrastive learning for video representation. *Comput. Vis. Image Underst.* **219**, 103406 (2022).
49. Pan, T., Song, Y., Yang, T., Jiang, W. & Liu, W. VideoMoCo: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11205–11214) (2021).
50. Gao, T., Yao, X. & Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2021).
51. Rethmeier, N. & Augenstein, I. A primer on contrastive pretraining in language processing: Methods, lessons learned and perspectives. Preprint at <https://arxiv.org/abs/2102.12982> (2021).
52. Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. vol. 29, 1857–1865 (Curran Associates, Inc., 2016).
53. Allaria, E. et al. Highly coherent and stable pulses from the FERMI seeded free-electron laser in the extreme ultraviolet. *Nat. Photonics* **6**, 699–704 (2012).
54. Bostedt, C. et al. Ultrafast x-ray scattering of xenon nanoparticles: imaging transient states of matter. *Phys. Rev. Lett.* **108**, 093401 (2012).
55. Rupp, D. et al. Coherent diffractive imaging of single helium nanodroplets with a high harmonic generation source. *Nat. Commun.* **8**, 493 (2017).
56. Hu, S. X., Li, D., Stühmer, J., Kim, M. & Hospedales, T. M. Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9068–9077) (2022).
57. Singh, A. & Jamali-Rad, H. Transductive decoupled variational inference for Few-Shot classification. Preprint at <https://arxiv.org/abs/2208.10559> (2022).
58. Chen, X. & Wang, G. Few-shot learning by integrating spatial and frequency representation. In *2021 18th Conference on Robots and Vision (CRV)* (pp. 49–56). IEEE, (2021, May).
59. Zagoruyko, S. & Komodakis, N. Wide residual networks. In Richard C. Wilson, E. R. H. & Smith, W. A. P. (eds.) *Proceedings of the British Machine Vision Conference 2016*, 87, 87.1–87.12 (British Machine Vision Association, 2016).
60. Li, X., Wang, W., Hu, X. & Yang, J. Selective kernel networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2019-June, 510–519 (2019).
61. He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J. & Li, M. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 558–567) (2019).
62. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141) (2018).
63. You, Y., Gitman, I. & Ginsburg, B. Large batch training of convolutional networks. Preprint at <https://arxiv.org/abs/1708.03888> (2017).
64. Loshchilov, I. & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations* (2017).
65. Goyal, P. et al. Accurate, large minibatch SGD: Training ImageNet in 1 hour. Preprint at <https://arxiv.org/abs/1706.02677> (2017).
66. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
67. Wang, Z. & Simoncelli, E. P. Translation insensitive image similarity in complex wavelet domain. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 2, ii/573–ii/576 Vol. 2 (2005).
68. Manning, C. D., Raghavan, P. & Schütze, H. *Introduction to Information Retrieval* (Cambridge University Press, 2008).
69. Martineau, B. H., Johnstone, D. N., van Helvoort, A. T. J., Midgley, P. A. & Eggeman, A. S. Unsupervised machine learning applied to scanning precession electron diffraction data. *Advanced Structural and Chemical Imaging* **5**, 1–14 (2019).
70. Agar, J. C. et al. Machine detection of enhanced electromechanical energy conversion in PbZr0.2 Ti0.8 O3 thin films. *Adv. Mater.* **30**, e1800701 (2018).
71. Shi, C. et al. Uncovering material deformations via machine learning combined with four-dimensional scanning transmission electron microscopy. *npj Comput. Mater.* **8**, 1–9 (2022).
72. Loh, N.-T. D. & Elser, V. Reconstruction algorithm for single-particle diffraction imaging experiments. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **80**, 026705 (2009).
73. Colombo, A. et al. Three-Dimensional coherent diffractive imaging of isolated faceted nanostructures. Preprint at <https://arxiv.org/abs/2208.04044> (2022).
74. Grill, J.-B. et al. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Article 1786 in NIPS'20, 21271–21284 (Curran Associates Inc., Red Hook, NY, USA, 2020).
75. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. open source softw.* **3**, 861 (2018).
76. Kingma, D. P. & Welling, M. Auto-Encoding variational bayes. Preprint at <https://arxiv.org/abs/1312.6114v10> (2013).
77. Burgess, C. P. et al. Understanding disentangling in  $\beta$ -VAE. Preprint at <https://arxiv.org/abs/1804.03599> (2018).
78. Westphal, T., Bier, T. A., Takahashi, K. & Wahab, M. Using exploratory factor analysis to examine consecutive in-situ x-ray diffraction measurements. *Powder Diffr.* **30**, 340–348 (2015).
79. Chernyshov, D., Dovgaliuk, I., Dyadkin, V. & van Beek, W. Principal component analysis (PCA) for powder diffraction data: towards unblinded applications. *Crystals* **10**, 581 (2020).
80. Camara, A. H. The importance of factor analysis in quantitative and qualitative x-ray diffraction phase analysis. *KOM – Corrosion and Material Protection Journal* **58**, 52–58 (2014).
81. Matos, C. R. S., Xavier, M. J., Barreto, L. S., Costa Jr, N. B. & Gimenez, I. F. Principal component analysis of x-ray diffraction patterns to yield morphological classification of brucite particles. *Anal. Chem.* **79**, 2091–2095 (2007).
82. Banko, L., Maffettone, P. M., Naujoks, D., Olds, D. & Ludwig, A. Deep learning for visualization and novelty detection in large x-ray diffraction datasets. *npj Comput. Mater.* **7**, 1–6 (2021).
83. Ruiz Vargas, J. C. et al. Shedding light on variational autoencoders. In *2018 XLIV Latin American Computer Conference (CLEI)*, 294–298 (Institute of Electrical and Electronics Engineers Inc., 2018).
84. Yao, Y. et al. AutoPhaseNN: Unsupervised physics-aware deep learning of 3D nanoscale Bragg coherent diffraction imaging. Preprint at <https://arxiv.org/abs/2109.14053> (2021).
85. Cherukara, M. J., Nashed, Y. S. G. & Harder, R. J. Real-time coherent diffraction inversion using deep generative networks. *Sci. Rep.* **8**, 16520 (2018).
86. Li, X. et al. Manifold learning of four-dimensional scanning transmission electron microscopy. *npj Comput. Mater.* **5**, 1–8 (2019).
87. Sainburg, T., Thielk, M. & Gentner, T. Q. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Comput. Biol.* **16**, e1008228 (2020).
88. Thomas, M. et al. A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations. *J. Anim. Ecol.* **91**, 1567–1581 (2022).

## ACKNOWLEDGEMENTS

Excellent support has been provided by the ISG team of DPHYS at ETH Zürich. Funding is acknowledged from the SNF via Grant No. 200021E\_193642 and the NCCR MUST. Further funding was provided by the Leibniz Society via Grant No. SAW/2017/MBI4, and from the DFG via Grant No. MO 719/14-2.

## AUTHOR CONTRIBUTIONS

Earlier versions of the contrastive-based diffraction imaging approach were tested by F.B. and D.G. under supervision of J.Z. and D.R. Training and evaluation of all models were written by J.Z. and performed on the NUX-Noether GPU high-performance-computer at ETH Zürich (NUX-Noether is a single multi-GPU node with two 16 Core AMD EPYC 7302 CPUs, 256 GB DDR4 Ram, and four NVIDIA RTX 3090 24GB GPUs). B.L. and D.R. contributed to discussing the results. The manuscript was written by J.Z. with input from all authors.

## COMPETING INTERESTS

The authors declare no competing financial or non-financial interests.

## ADDITIONAL INFORMATION

**Correspondence** and requests for materials should be addressed to Julian Zimmermann.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**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) 2023