

# Classification of Radio Sources Through Self-Supervised Learning

Nicolas Baron Perez<sup>1</sup>, Marcus Brüggen<sup>1</sup>, Gregor Kasieczka<sup>2</sup>, and Luisa Lucie-Smith<sup>1</sup>

<sup>1</sup> Hamburger Sternwarte, Universität Hamburg, Gojenbergsweg 112, 21029 Hamburg, Germany

<sup>2</sup> Institut für Experimentalphysik, Universität Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany

Received date / Accepted date

## ABSTRACT

**Context.** The morphology of radio galaxies is indicative, for instance, of their interaction with their surroundings. Since modern radio surveys contain a large number of radio sources that will be impossible to analyse and classify manually, it is important to develop automatic schemes. Unlike other fields, which benefit from established theoretical frameworks and simulations, for radio galaxies there are no such comprehensive models, posing unique challenges for data analysis.

**Aims.** In this study, we investigate the classification of radio galaxies from the LOFAR Two-meter Sky Survey Data Release 2 (LoTSS-DR2) using self-supervised learning.

**Methods.** Our classification strategy involves three main steps: (i) self-supervised pre-training; (ii) fine-tuning using a labelled subsample created from the learned representations; and (iii) final classification of the selected unlabelled sample. To enhance morphological information in the representations, we developed an additional random augmentation.

**Results.** Our results demonstrate that the learned representations contain rich morphological information, enabling the creation of a labelled subsample that effectively captures the morphological diversity within the unlabelled sample. Additionally, the classification of the unlabelled sample into 12 morphological classes yields robust class probabilities.

**Conclusions.** We successfully demonstrate that a subset of radio galaxies from LoTSS-DR2, encompassing diverse morphologies, can be classified using self-supervised learning. The methodology developed here bridges the gap left by the absence of simulations and theoretical models, offering a framework that can readily be applied to astronomical image analyses in other wavebands.

**Key words.** Astronomical instrumentation, methods and techniques – Methods: data analysis – Galaxies: jets, nuclei – Radio continuum: galaxies

## 1. Introduction

Deep, wide-area radio surveys are mapping a rapidly increasing number of radio sources. Recent and forthcoming surveys, including the LOFAR Two-Metre Sky Survey (LoTSS; Shimwell et al. (2017)), the Evolutionary Map of the Universe (EMU; Norris et al. (2011)), and surveys utilizing the Square Kilometre Array (SKA), are expected to identify millions of galaxies.

The large number and heterogeneity of radio galaxies makes them difficult to classify and historical schemes may no longer be adequate as they no longer aligned with our physical understanding (Rudnick 2021).

Particularly difficult is the morphological classification that solely rests on the geometric appearance of the source. Morphological classification is very important in order to understand the physical processes that form radio galaxies, e.g. how the jets that are produced by the supermassive black hole are interacting (and have interacted in the past) with their environment. Moreover, it is important for studying the relation between radio galaxies and galaxy evolution, and for assessing the energetics of the sources.

The visual categorization of these radio sources is becoming increasingly time-intensive and will soon be impractical due to the rapidly expanding volumes of data. To address this challenge, citizen science projects, such as the Radio Galaxy Zoo (RGZ; Banfield et al. (2015); Wong et al. (2025)), have been deployed for the classification of astronomical sources.

Radio galaxies can present compact or extended radio morphologies (Miraghaei & Best 2017) and are traditionally classi-

fied into either the FRI (core-bright) or FRII (edge-bright) galaxies (Fanaroff & Riley 1974). There are several differences between the two classes: FRIs show less powerful jets than FRIIs, and these jets are disrupted quite close to the core of the radio galaxy while jets in FRII radio galaxies stay relativistic for much larger distances. In addition to the FRI/FRII classification, radio galaxies have been classified as doubles, double-doubles, triples, narrow-angle tails, wide-angle tails, bent-tails, hybrid, X-shaped, S-shaped, C-shaped, core-dominant, one-sided sources and others (Kempner et al. 2004). Hybrid sources have the potential to reveal the origin of different radio morphologies (Stroe et al. 2022). There is no unified scheme and Rudnick (2021) has advocated for the tagging of radio galaxies instead. This was followed up using a natural language processing algorithm to derive a radio galaxy morphology taxonomy (Bowles et al. 2023).

With the advent of LOFAR, this classification scheme and the resulting luminosity correlation have been revisited using a dataset with a size two orders of magnitude bigger (Mingo et al. 2019). This work found that for this larger sample, the source luminosity could not predict its morphological class, finding a strong overlap of both classes in luminosity. Moreover, different morphologies such as restarting or double-double FRIIs and hybrid sources have also been found (Mingo et al. 2019).

With the excellent performance of convolutional neural networks (CNN) in classifying natural images, there have been several efforts to train CNNs to classify images of radio sources (Maslej-Krešnáková et al. 2021; Lao et al. 2023, e.g.). Wu et al. (2019) developed a radio source morphology classifier based on

a region-based CNN that was trained with FIRST and WISE data using the classification scheme of number of components and flux peaks. More recently, Lao et al. (2025) found thousands of bent-tail radio galaxies in the FIRST survey using deep learning and visual inspection.

Lukic et al. (2019) explored the performance of capsule network architectures against simpler CNN architectures, to classify unresolved, FRI, and FRII morphologies. Brand et al. (2023) have further adapted CNNs for radio galaxy classification by (i) using principal component analysis (PCA) during pre-processing and (ii) by guiding the CNN to search for specific features within the image data. They found that this adaptation led to a more stable training process and could reduce overfitting. See also Ndung'u et al. (2023) for a recent review on the machine-learning-aided classification of radio sources.

Given all these developments, a drawback of supervised classification that is particularly pertinent to astronomy is the relatively small number of labelled training data and the morphological incompleteness when compared to comprehensive surveys. Rustige et al. (2023) explored the augmentation of labelled data using a conditional Wasserstein Generative Adversarial Network (wGAN) to improve the classification accuracy of compact, FRI, FRII and bent sources.

Another drawback is the varying classification schemes that have emerged in recent years, driven by the continuous nature of the diverse morphology of radio galaxies, which can lead to inconsistencies and confusion. Therefore, Rudnick (2021) proposes the use of tags instead of discontinuous boxes. The advantage of using tags is the possibility to combine them and extend the list of tags depending on the observed sources.

In recent years, self-supervised learning (SSL), has demonstrated remarkable results, in particular in the field of computer vision classification, the relevant area for this work (for an overview of recent developments in SSL, refer to Balestrieri et al. (2023); for specific applications in astrophysics, see Huertas-Company et al. (2023)). This training strategy allows models to learn from large amounts of unlabelled data by extracting meaningful representations from the data's inherent structures. Consequently, this approach overcomes the limitations of explicitly labelled datasets, resulting in a more robust model due to its ability to process and learn from a large volume of data.

Subsequently, the model can be “fine-tuned” by continuing its training with a more carefully curated and labelled dataset, using the weights obtained from the “pre-training” phase. This fine-tuning phase helps to adapt the general knowledge that the model learned during pre-training to specific “downstream tasks”. These are particular tasks that the model is designed to assist with, such as image classification. The combination of SSL pre-training followed by fine-tuning is commonly referred to in the literature as semi-supervised learning.

Three distinct types of SSL methods have emerged: invariance-based methods, generative methods, and what we term “context-aware” methods. The first type employs hand-crafted data augmentations to learn representations that remain invariant under these image transformations. The second type involves corrupting or removing parts of images and training the model to predict these missing sections. The third type focuses on predicting representations for specific parts of an image based on representations from other parts. For a more comprehensive discussion on these categories of SSL and the presentation of

the first context-aware method, we direct readers to the work by Assran et al. (2023).

Efforts to use semi-supervised learning techniques in astrophysics include gamma-ray blazar classification (Bhatta et al. 2024), cosmological inference (Akhmetzhanova et al. 2024), semi-supervised classification of FRI and FRII radio galaxies (Hossain et al. 2023), hierarchically fine-tuning of a convolutional autoencoder (AE; Hinton & Salakhutdinov (2006)) to classify six types of radio galaxies (Ma et al. 2019), a comprehensive study on semi-supervised classification of radio galaxies (Sljepcevic et al. 2022), and similarity search for optical galaxies (Stein et al. 2021). Moreover, Sljepcevic et al. (2024) showed that a self-supervised model can be used for high-accuracy classification and similarity search of radio sources. Recently, Cecconello et al. (2024) conducted a benchmark study to evaluate the performance of various invariance-based SSL algorithms on radio astronomy datasets employing different classification schemes. As part of this research, a new labelled dataset was developed using data from multiple radio surveys.

SSL algorithms learn a mapping from image space into an abstract vector space, distributing the dataset according to the features identified by the encoder. If the learned representation captures morphological information - which is heavily influenced by the learning task - the way images are distributed within this vector space should reflect the continuous nature of radio galaxy morphology. This enables to move away from the discontinuous classification boxes at a first level, while being able to search for higher-dense regions containing similar objects at a higher level.

The unsupervised classification has been performed for multiple astronomical datasets. Works based on clustering algorithms include Tohill et al. (2024) for high-redshift galaxies, Pérez-Díaz et al. (2024) for X-ray objects, Luo et al. (2024) for molecular clumps, and Dubois et al. (2022) for galaxy spectra.

Unsupervised classification based on the combination of SSL and clustering algorithms has been applied to galaxy surveys in the near-infrared band (Vega-Ferrero et al. 2024), in the thermal infrared (Guo et al. 2022) and in the visual band (Sarmiento et al. 2021). In the context of radio observations, Mohale & Lochner (2024) classified FRI and FRII radio galaxies with this technique.

Moreover, the unsupervised classification of radio sources exceeding FRI and FRII galaxies with a self-organising map (SOM; Kohonen (2001); Vantyghem et al. (2024)) was studied first in combination with an autoencoder (Sanger 1989) and the clustering algorithm  $k$ -means (Lloyd 1982) by Ralph et al. (2019). Later, the unsupervised classification with a SOM and followed manual grouping of classes was performed within Mostert et al. (2021).

In this work, we investigate the classification of radio sources from the LoTSS Data Release 2 (LoTSS-DR2; Shimwell et al. (2022)) using an invariance-based SSL method, and we introduce a new data augmentation that we termed “Random Structural View”. The dataset used for pre-training consists of an image-quality filtered subset of the entire release, which we refer to as the “unlabelled sample”, and includes significant morphological diversity. In our approach, we forgo the use of previously labelled datasets to mitigate morphological biases stemming from label incompleteness. Instead, we generate a labelled subsample using the learned representations customised to the morphological diversity of the unlabelled sample. We tackle this classification task with a three-step process: (i) model pre-training, wherein we learn morphologically informed represen-

tations, (ii) model fine-tuning, in which we adjust the model for the classification task, and (iii) classification downstream task, where we classify the unlabelled sample.

This paper is structured as follows: Section 2 provides a detailed description of the data used in this work. Section 3 outlines the methodology for pre-training the model and presents the corresponding results. In Sect. 4, we discuss the fine-tuning process of the model and its associated outcomes. Section 5 highlights the classification results of the unlabelled subsample, serving as our downstream task. In Sect. 6, we interpret our findings and put them in context with existing literature. Finally, in Sect. 7, we summarise our conclusions and insights drawn from the study.

## 2. Data

This section outlines the data utilised in our study. The real data serves multiple purposes: it is employed for model pre-training, for constructing the tailored labelled subsample used for fine-tuning, and represents the unlabelled sample targeted for classification in our downstream task. Meanwhile, the simulated data is used for model evaluation during the pre-training phase.

### 2.1. LOFAR DR2 data

In this work, we used the synthesised image products from LoTSS-DR2 with a resolution of  $6''$ . This data release comprises observations covering 27% of the northern sky within the frequency range of 120 – 168 MHz, for more details, we refer the reader to Shimwell et al. (2022).

We used the publicly available<sup>1</sup> mosaics and the radio-optical cross-matched catalogue (Hardcastle et al. 2023). The full catalogue initially contains 4 116 934 radio sources. We discard the unresolved sources since they are generally simple to classify. Additionally, we excluded the sources that presented errors during image production or where the final image contains NaN values (the image production process is described in detail below). Moreover, we remove sources with largest angular sizes (LAS) smaller than  $30''$  since we focus on morphological classification of resolved sources which span more than 5 beam sizes. This selection leads to 253 242 sources.

We further apply some filters that aid training for this pilot study, resulting in our unlabelled sample:

- We restrict the dataset to sources with available optical coordinates, which resolves the ambiguity introduced when using radio mean positions. For instance, with radio mean positions, images of centre-bright sources are centred at the galaxy core, while images of edge-bright sources with strongly asymmetric hotspot intensity are centred at the most luminous hotspot. This reduces the number of sources to 192 593.
- We apply a filter on the peak flux, selecting sources with  $F_{\text{peak}} > 0.75 \text{ mJy/beam}$ . We were left with 76 235 sources.
- We limit the LAS range to  $30'' < \text{LAS} < 60''$  to reduce the variety of morphologies within a certain established radio galaxy morphology due to source resolution. For instance, a centre-bright radio galaxy with low angular resolution can be described by the linear arrangement of a circularly shaped bright core and two elliptically shaped lobes. Conversely, with a higher angular resolution, the image will show a better-resolved core with jets departing from a perfect

linear shape and lobes that present greater details. We observed that including all sources in  $\text{LAS} > 30''$  leads to considerably worse results. After this cut, we maintain 43 769 sources.

In the following, we describe the production of the images fed into the network. First, we determine the mosaic with its centre closest to the radio source, using the haversine distance to account for mosaic overlaps. Subsequently, we extract a square region centred on the optical position of the host. The cutout has a side length of  $s_L = 1.5 \times \text{LAS}$ . Subsequently, the pixel values below  $v_{\min} = 3 \times \sigma_{\text{img}}$  are clipped to  $v_{\min}$ , where  $\sigma_{\text{img}}$  is the  $\sigma$ -clipped standard deviation of the cutout. This helps to suppress noise and enhance the signal-to-noise ratio by preventing the influence of low-intensity artefacts. Given the varying sizes of cutouts, we resize the images to a common side length of  $s_L = 128$  pixels. Finally, using the min-max normalisation, we rescale the image pixels to values  $v_{\text{pix}} \in [0, 1]$ .

We present several randomly selected examples in Fig. 1. The images illustrate that the morphology of the radio galaxy remains exceptionally complex after filtering.

### 2.2. Simulated Dataset

The morphological complexity of the real dataset motivates the creation of a synthetic dataset based on a very simplified radio galaxy model. This dataset can be used to evaluate the model's basic learning process (see Sect. 3.4).

In this model, we represent a radio galaxy as a source with a core and two lobes. These three components are represented using two-dimensional Gaussian intensity distributions, with the core modelled by a circular Gaussian and the lobes by elliptical Gaussians. We consider two morphological degrees of freedom: (i) centre or edge-bright and (ii) linear or bent-shaped. The former is achieved by adapting the relative intensity of the core and lobes (an invisible core is also included), while the latter is implemented with an angle between the two lobes (where the core is the vertex of the angle). Hence, our simplified synthetic dataset consists of four morphological classes with source bending being a crucial morphological characteristic.

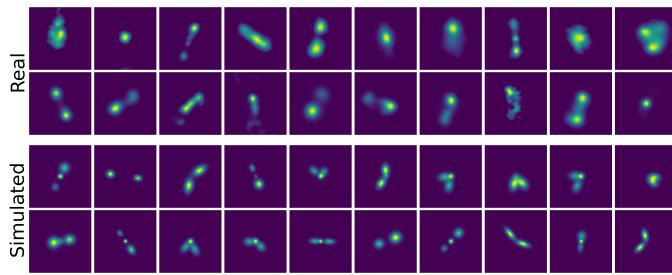
The linearly-shaped mock sources take into account the following morphological parameters: Source size, orientation angle, the four standard deviations of the lobe Gaussian distributions, and if it is centre or edge-bright. Finally, the bent-shaped mock sources have the parameters: Orientation angle, lobe angle, two standard deviations for both identical lobes, and centre or edge-bright. Most of the mock source parameters were sampled from uniform distributions. We generated 5000 examples of each of the four morphological classes to have a balanced dataset. Randomly selected examples are shown in Fig. 1.

## 3. Model Pre-training

This section details the model pre-training process, which uses invariance-based SSL techniques with the unlabelled sample. We describe our model architecture and training strategy, introduce our novel data augmentation approach, discuss the various methods used to evaluate the pre-training performance, and present the results of the pre-training stage.

Self-supervised algorithms are trained by addressing a broader task, the pretext task, rather than focusing on predicting a specific value, the human-annotated label, as in supervised learning. The goal of SSL is to generate useful representations of the data without labels. Here, the representation space should

<sup>1</sup> [https://lofar-surveys.org/dr2\\_release.html](https://lofar-surveys.org/dr2_release.html)



**Fig. 1.** Randomly selected images from both the real and simulated datasets, showcasing their morphological diversity.

group sources together based on morphology. These representations can then be used for downstream tasks. The classification process, which is the downstream task we focus on in this study, is illustrated in Fig. 2.

### 3.1. Self-Supervised Training

The invariance-based SSL task involves generating two augmented views of the same input image and training the model to produce similar representations for these views compared to representations of different images. SimCLR is among the initial SSL algorithms that demonstrated exceptional performance in the linear classification of its extracted representations (the linear evaluation protocol) (Chen et al. 2020). We opted to use SimCLR for representation extraction due to its simplicity. Other models tested included BYOL (Grill et al. 2020), MSF (Abbasi Koohpayegani et al. 2021), ProPos (Huang et al. 2021) and NNCLR (Dwibedi et al. 2021)). Simplicity was prioritised at this stage to focus on developing the pipeline, with the goal of creating a classification scheme that comprehensively encompasses the morphological diversity of the dataset.

In terms of its network architecture, SimCLR consists of an encoder  $f(\cdot)$  responsible for extracting a representation vector  $\mathbf{h}$  from its input  $\mathbf{v}$ , followed by a projector  $g(\cdot)$  that maps the representation  $\mathbf{h}$  into a projection  $\mathbf{z}$  in a lower-dimensional space, see Fig. 3. The input images have dimensions of 128x128 pixels, the representation consists of 512 dimensions, and the projection vectors have 256 dimensions. We employ a ResNet18 (He et al. 2015) for our encoder  $f(\cdot)$ , as has been done in previous works (e.g. Slijepcevic et al. (2024) or Mohale & Lochner (2024)), and a multi-layer perceptron (MLP) with one hidden layer for the projector  $g(\cdot)$ .

To define the learning task of SimCLR, we generate two views  $\mathbf{v}_1, \mathbf{v}_2$  of each input image  $\mathbf{x}$  using random transformations (more details are given below). Thus, given a batch of size  $N$ , we pass the  $2N$  views through the networks, obtaining  $2N$  projections. Then, the learning task is to ensure that projections  $\mathbf{z}_1$  and  $\mathbf{z}_2$  are closer in the learned space than  $\mathbf{z}_1$  is to any of the remaining  $2N - 2$  projections. Mathematically, this is formulated by minimizing the NT-Xent (Normalized Temperature-Scaled Cross Entropy) loss function, which, for a pair of views  $\mathbf{v}_i, \mathbf{v}_j$ , is given by:

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

where  $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / \|\mathbf{u}\| \|\mathbf{v}\|$  represents the cosine similarity,  $\tau$  is a temperature parameter, and  $I_{[k \neq i]}$  is 1 if  $k \neq i$  and 0 otherwise. The temperature parameter adjusts the influence of the data

points within the batch: a smaller value gives greater weight to closer representations, while a larger value increases the influence of representations that are further apart (Frosst et al. 2019).

A key component of SimCLR is the set of random transformations used to generate image views that define the learning task. We employ standard augmentations: random vertical flips with 50% probability, random rotations between  $-360^\circ$  and  $360^\circ$ , random brightness jitter with a factor between 0 and 1.1, and random resize crop with a scale between 0.9 and 1. Additionally, we incorporate a custom augmentation, which is detailed in Sect. 3.2.

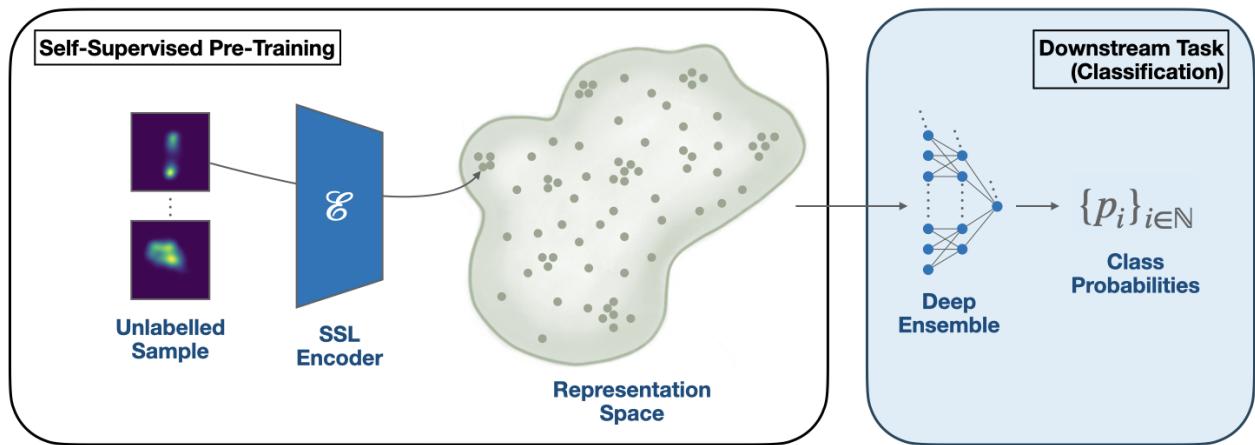
We train the SimCLR model to minimise the NT-Xent loss function with a learning rate (LR) of 0.0003 and a batch size of 1024. The LR schedule starts with a linear warm-up (Goyal et al. 2017) phase of 5 epochs starting at half the LR and is succeeded by a cosine annealing (Loshchilov & Hutter 2016) phase of 200 epochs that finishes at 1/50 of the LR. We use AdamW (Loshchilov & Hutter 2017) as the optimiser with a weight decay of 0.0001 and a temperature of  $\tau = 0.09$  in the NT-Xent loss. The temperature value is chosen empirically. It should be small enough to account for low-level morphological properties, yet large enough to ensure that representations of sources with the same morphology are not separated in representation space due to non-morphological properties. For downstream tasks, we normalize the representations predicted with the encoder to have an Euclidean norm of 1, as is done when computing the NT-Xent loss function during training.

### 3.2. Random Structural View

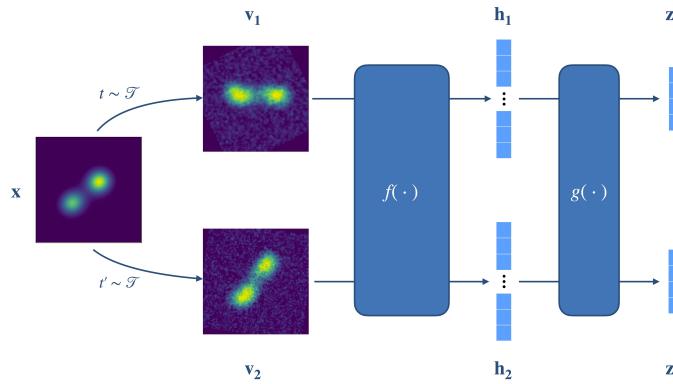
So far, the two image views used to compute the NT-Xent loss were produced from the same input image. These are used as positive pairs, while all other  $2N - 2$  views in the batch are considered negative examples. For our classification purpose, this set-up is not ideal because sources with the same morphology, which appear identical, are treated as negative examples. Furthermore, sources with different morphologies, such as three-component FRI and FRII, can share numerous characteristics, with the only difference being the relative brightness of the core and the lobe hotspots. However, the random augmentations considered so far cannot differentiate between the sources based on this information.

To solve this issue, we use a metric to measure the similarity between images, based on the assumption that images of sources with identical morphology will have a higher similarity value compared to those with different morphologies. Referring back to the previous example, the similarity score between two three-component FRI will be higher than that between a three-component FRI and an FRII. Therefore, for a given source, we can select one of the sources with the highest similarity scores as a positive example. This allows us to create a random augmentation that, with a certain probability, replaces one augmented view of the original source with an augmented view of a “very similar” source. We call this random augmentation *Random Structural View*.

The similarity metric we use is the Structural Similarity Index Measure (SSIM), which was designed to imitate the human visual system under the assumption that the human eye is very sensitive to changes in structural information (Wang et al. 2004). Moreover, we use the following procedure to make the metric semi-invariant (SI-SSIM): For a pair of images  $(\mathbf{x}_i, \mathbf{x}_j)$ , SI-SSIM is defined as the maximum value of SSIM between  $\mathbf{x}_i$  and eight transformed versions of  $\mathbf{x}_j$ . The transformations considered are: the original image, rotations by  $90^\circ$ ,  $180^\circ$  and  $270^\circ$ ; a vertical



**Fig. 2.** The diagram illustrates the classification process. Initially, the encoder maps the input sample into the representation space, where calculations such as evaluating image similarities and determining distances to morphological classes can be performed. Subsequently, the deep ensemble generates robust class probabilities based on these representations.



**Fig. 3.** SimCLR architecture: The input image  $\mathbf{x}$  undergoes random transformation twice, generating two views. These views are then processed through the encoder  $f(\cdot)$  to extract corresponding representations. Subsequently, an MLP  $g(\cdot)$  is employed to project these representations into a lower-dimensional space. The objective of the learning task is to maximize the similarity between these projections.

flip; and a vertical flip combined with each of the aforementioned rotations.

To optimise the computation time, we compute the SI-SSIM matrix  $M^{\text{img}}(\mathbf{x}_i, \mathbf{x}_j)$  before starting the training of the SSL algorithm for a given dataset. For the computation, we make use of GPU power to accelerate the calculation. Depending of the dataset size, we opted to reduce the image resolution to further accelerate the SI-SSIM calculation.

During training, given an image  $\mathbf{x}_i$ , we draw one image from the  $k = 4$  most similar images present in the batch ( $k = 0$  being the image  $\mathbf{x}_i$  itself) and use its transformed version as the second augmented view. The intended choice of a small  $k$  is to ensure that the most similar images present in the batch have almost identical morphology.

### 3.3. Dimensionality Reduction and Clustering of Representations

To examine the morphological information encapsulated in the representations, the most direct approach is to employ clustering algorithms. However, given the extensive processing time

required for clustering algorithms on high-dimensional data, we apply PCA to reduce the dimensionality of the 512-dimensional vectors while preserving 95% of the total variance. This reduction results in 22 principal components, minimising component correlation and organising them according to the variance they explain. We then apply a clustering algorithm to identify high-density regions within the transformed principal component space.

We tested the clustering algorithms  $k$ -means, HDBSCAN, BIRD and agglomerative clustering and found that HDBSCAN<sup>2</sup> (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*; McInnes et al. (2017)) worked best for our application. This clustering algorithm can handle clusters of different densities and accounts for “noise”, i.e. data points that lie far from any dense region that defines the clusters. Since we are working with uncurated data, our dataset will include outliers, corrupted data, and noise. It can also detect non-spherical clusters, which is important since we do not have any reason to assume that the emerging high-density regions in the principal component space should be spherical.

HDBSCAN includes mainly two parameters to adjust the clustering, the parameter `min_samples`, which determines how conservative the clustering should be, and the minimal cluster size `min_cluster_size`. Setting `min_samples = 1` works best for our data.

The clustering analysis involves an iterative application of HDBSCAN on the PCA-processed representations to determine the `min_cluster_size` that yields the highest number of clusters. We found that `min_cluster_size = 19` produced the highest number of clusters, totalling 95. Our objective in identifying the maximum cluster count is to obtain a set of sources within the clusters that captures the morphological diversity of the unlabelled sample as completely as possible.

### 3.4. SSL Evaluation Metrics

In the following, we present a short description of the metrics used to evaluate the learning performance of the pre-trained model.

Before outlining the specific evaluation metrics, we introduce the Soft Nearest Neighbour Loss (SNNL; Salakhutdinov &

<sup>2</sup> <https://hdbSCAN.readthedocs.io/en/latest/>

Hinton (2007); Frosst et al. (2019)), which we used as an evaluation metric during pre-training. SNNL measures the relative distances in the representation space of a given image by comparing the distances between this image and others with the same label in the batch to the distances between this image and all images in the batch. For a more detailed explanation of SNNL, refer to Sect. 4.3.

The evaluation metrics utilised during pre-training are as follows:

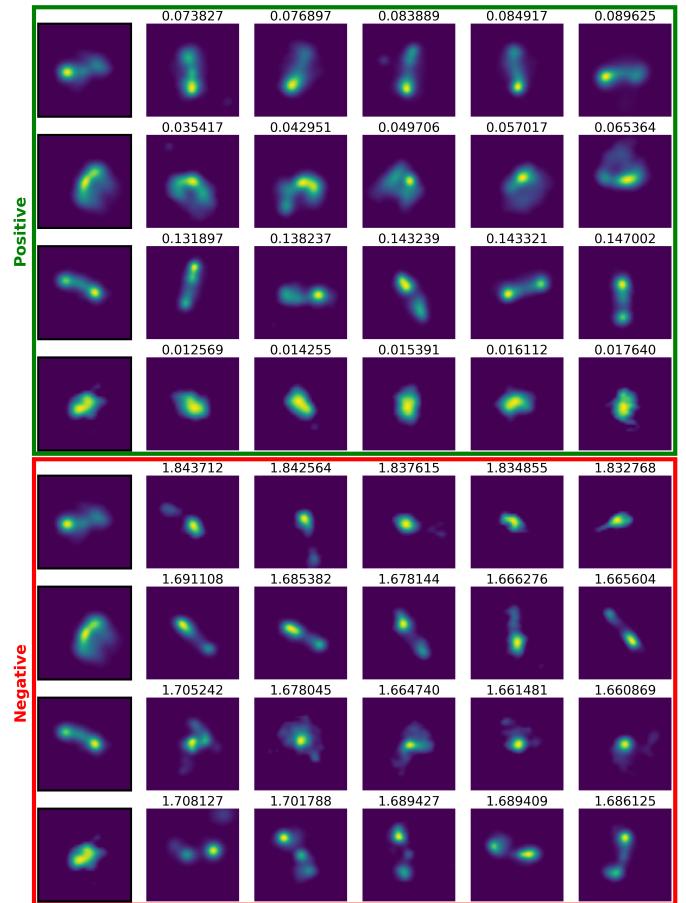
- **Top- $k$  Accuracy:** This metric measures the percentage of input data where the cosine similarity between the projections of both corresponding augmented views ranks among the top  $k$  in the batch.
- **Mean Position Accuracy:** After computing the cosine similarity for all  $2N$  projections in a batch, this metric provides the average rank of the cosine similarity between the augmented view projections for each input image, averaged over all  $N$  images in the batch.
- **SNN Metric:** SNNL is employed to assess how closely neighbouring images, whose distances are calculated using SI-SSIM in image space, are located in the representation space. We use a temperature parameter of  $\tau = 0.07$  and consider  $k^{\text{SNN}} = 10$  SI-SSIM neighbours, which are used as labels for the metric computation.
- **Linear Evaluation Protocol:** This metric consists in the supervised classification of the representations obtained from the self-supervised encoder using a linear model. We use the simulated dataset to calculate the metric.
- **Clustering:** To understand the information encoded in the representations, we perform clustering. Visual inspection of the resulting clusters provides insights into the overall structure of the representation space by contrasting different clusters, while examining the morphological homogeneity within each cluster reveals the finer details of the structure.

Our goal is to obtain representations that primarily capture morphological information, ensuring that sources with the same morphology have nearly indistinguishable representations. Thus, instead of focusing solely on maximizing top- $k$  accuracy or minimizing mean position accuracy, we aim to improve these metrics to a point of equilibrium, enabled by the high representation similarity among sources within the same class.

Moreover, an improvement in the SNN metric indicates that the fine-grained distribution of the representation space effectively captures morphological similarity. While the improvement of the linear evaluation protocol denotes that the representations encode the notion of centre or edge-bright and linear or bent-shaped. Ultimately, the evaluation of the clustering provides the most significant insights into the information encapsulated within the representations.

### 3.5. Pre-Training Results

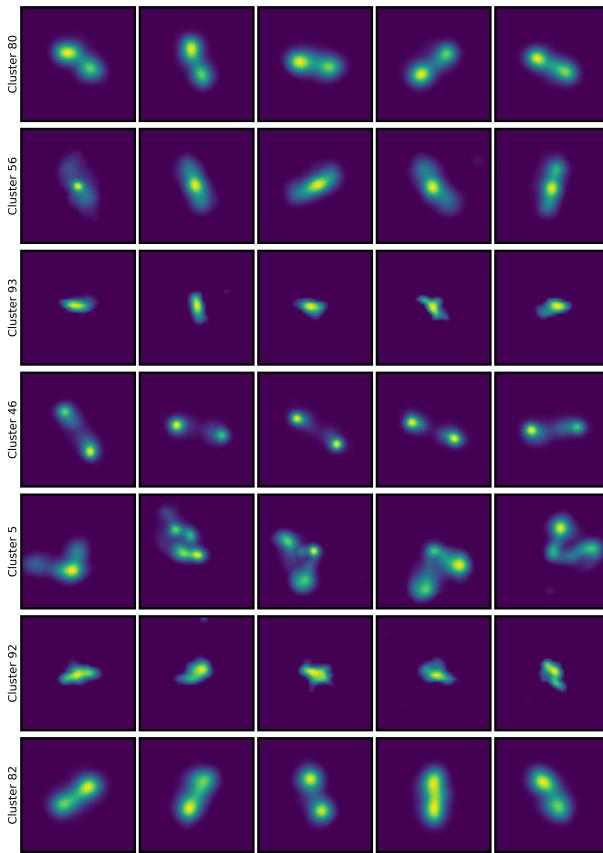
We examine the representations derived from the pre-training phase and the clusters that have been identified. We begin by evaluating the training metrics. The NT-Xent starts at 6.79 and decreases steadily throughout training, eventually reaching 4.48, indicating effective learning. Post-training, we achieve a top-1 accuracy of 18.9%, a top-5 accuracy of 32.7% and a mean position accuracy of 49. Although the accuracy values are not exceptionally high, they show consistent improvement over the course of training. This trend is expected and desirable, as our primary goal is not to distinguish individual sources but to develop clusters of morphologically similar sources that are nearly indistinguishable from each other.



**Fig. 4.** The first column displays the query images. The first four rows show the 5 nearest neighbours based on cosine distance, while the last four rows display the 5 farthest representations. Above each neighbour, the corresponding cosine distance value is shown.

The SNN metric value decreases from 3.47 to 1.68 during training, indicating that the ten most similar sources for each target source, based on the SI-SSIM, are becoming more closely aligned. This trend is evident when examining the images that are closest in representation space for any given source. To demonstrate this, we present the nearest and farthest neighbours of randomly selected images, determined by the cosine distance of the PCA-transformed representations, in Fig. 4. Cosine distance is defined as  $\text{dist}(\mathbf{u}, \mathbf{v}) = 1 - \text{sim}(\mathbf{u}, \mathbf{v})$ . This figure illustrates that the representations capture a wide range of ample morphological information within the PCA-transformed representation space. For closely situated representations, the images associated with them show similar morphologies. Conversely, when comparing more distant representations, the corresponding images exhibit significantly different morphologies. Specially, morphological features, such as bending angle, the relative intensity of radiation peaks, the number of source components, and source extension are effectively encoded within the representations.

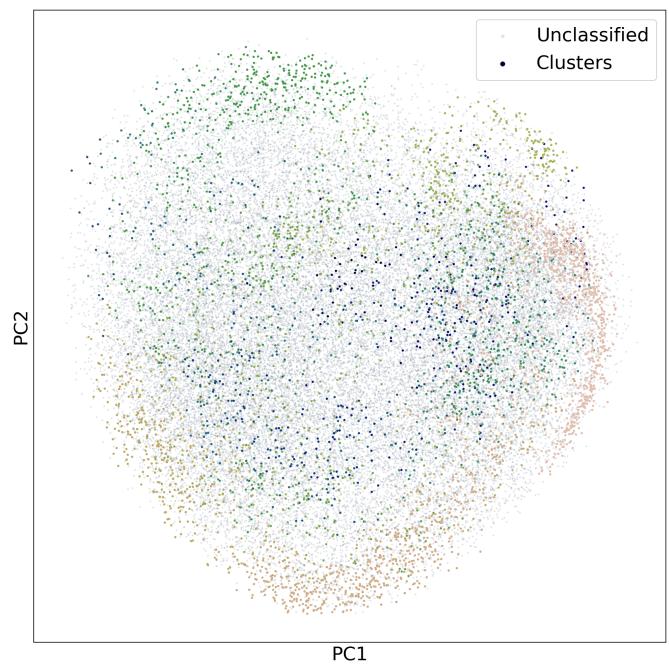
Furthermore, we achieve linear evaluation accuracies of 90%, 89%, and 91% for the simulated training, validation and test datasets, respectively. The linear evaluation protocol involves classifying the representations using a model as simple as a linear one. Therefore, these results indicate that the encoder effectively distinguishes between linear and bent sources, as well as between centre-bright and edge-bright sources. This is evident in the local neighbourhood of any given query rep-



**Fig. 5.** Seven clusters were randomly selected, and from each of these clusters, five members were also chosen at random to be displayed.

resentation. To verify whether this morphological distinction extends to the identified clusters, we present five randomly selected sources from seven randomly selected clusters in Fig. 5. The plot demonstrates that most of the clusters are homogeneous in terms of morphology. Specifically, the clusters shown contain sources with the following morphologies (from top to bottom): symmetric doubles, core-bright sources, amorphous sources, edge-bright sources with visible lobes, and bent sources. The last two clusters again feature amorphous and symmetric double sources, respectively.

We present the raw clustering results in Fig. 6. It shows the distribution of the first two principal components of the representations, with colours indicating cluster pre-labels and gray for unclassified sources. We consistently use the first two principal components throughout the paper to visualise the distribution of the representations in a two-dimensional space, which helps to emphasise the dominant structure of the data. Upon comparison with visualisations obtained using UMAP (McInnes et al. 2018) and t-SNE (van der Maaten & Hinton 2008), no significant differences were observed. The plot reveals a diffuse structure within the PCA-transformed representations, making it challenging to visually distinguish individual clusters without colour differentiation. The clusters are dispersed across the entire representation distribution, indicating that they encompass a wide variety of morphological shapes.



**Fig. 6.** Raw clustering results are depicted, illustrating the distribution of representations in the space defined by the first two principal components. Data points are colour-coded according to their cluster pre-labels with unclassified sources displayed in grey.

## 4. Model Fine-Tuning

In this section, we outline our model fine-tuning procedure, designed to adapt the network for improved performance in predicting cluster labels for each image in the unlabelled sample (our downstream task). We explain how the labelled subsample used for fine-tuning is constructed and describe the final composition of this subsample. Additionally, we introduce the fine-tuning procedure itself and present the results it yields. This procedure involves training a deep ensemble to achieve more accurate class probability predictions.

The necessity of fine-tuning arises from the SSL pre-training, which involves a broad learning task and results in representations that include information not particularly useful for the classification task. Therefore, the goal of fine-tuning is to adapt the representation space for improved classification performance. The fine-tuning task involves predicting the class probability of images with assigned cluster label. Simultaneously, the fine-tuned representation space naturally adapts to accommodate the continuous nature of the radio source morphologies.

### 4.1. Assigning Cluster Labels to Images

Constructing the labelled subsample for the fine-tuning step involves pairing images with their assigned labels, where each label represents the cluster to which the image belongs. The clusters should comprehensively capture the full morphological diversity present in the unlabelled sample.

We start with assigning labels to images based on the unsupervised clustering strategy described in Sect 3.3. The resulting cluster labels are termed “pre-labels” in the subsequent discussion. Using these cluster pre-labels as ground truth class labels for the supervised fine-tuned model would not return meaningful class predictions. In our approach, this is because the clustering

method produces large numbers of redundant classes, which also include imperfections whereby morphologically similar images can be assigned to different clusters while differing images may belong to the same cluster.

Instead, our goal is to design a training set which includes only images with a high probability of belonging to their parent cluster, so that to avoid confusing the model with images which are poorly correlated to their parent cluster. We therefore want to restrict to the purest subset of (images, cluster label) pairs.

Contrary to standard supervised classification tasks, our setting does not suffer from the issue of potentially having a smaller number of training set images. This is thanks to the optimisation conducted during the SSL step using a large dataset. For further details on this characteristic of SSL pre-trained models in the context of radio source classification, see Slijepcevic et al. (2024).

The label assignment procedure involves two steps: (i) reducing the number of redundant clusters, and (ii) assigning labels to the most homogeneous set of images within each cluster.

To minimise redundant clusters, we visually inspect random samples from each cluster. Clusters that are determined to contain sources with highly heterogeneous morphologies have their pre-labels replaced with the “unclassified” label. Subsequently, clusters containing sources that appear to have identical morphologies but different pre-labels are assigned a common label.

We then compute the SI-SSIM matrix for each remaining cluster and average over one dimension to obtain a mean SI-SSIM value for each cluster member. Based on the observed morphological homogeneity within each cluster, we retain members at the 90th, 80th, or 60th percentile according to their mean SI-SSIM values. The images falling below these thresholds are reclassified as unclassified, while the remaining images retain their respective cluster labels. In this way, we assign labels to the most homogeneous set of images within each cluster.

In future work, we plan to explore fully automated approaches for constructing training set labels from the representation space.

#### 4.2. Labelled Subsample

We now present the labelled subsample used for fine-tuning, which is specifically designed to capture the morphological diversity inherent in the unlabelled sample targeted for classification in our downstream task.

From the initial 95 cluster pre-labels, we relabelled 26 clusters as “unclassified” due to the heterogeneous morphologies of the images within each cluster (as described in the first step of Sect. 4.1). We identified three potential explanations for the formation of these clusters: Source miscentering, images with poor quality, and sources with common shapes but different intensity peak positions.

Additionally, we consolidated the clusters containing sources with the same morphology, reducing the total number of cluster pre-labels from 69 to 12 (second step in Sect. 4.1). We define and describe these source types as follows:

- **Artifact:** These are mainly sources that are outshined by another source present in the cutout. Due to the other source higher luminosity, the source in question is not visible.
- **Amorphous:** Sources with unclear morphology and small size, although the image size is proportional to the source LAS provided in the catalog.
- **Bright core:** These sources are single Gaussian intensity dots. They might be compact radio sources, also called FR0s (Baldi 2023).

- **Head-tail:** Sources with a bright core and a single lobe connected to the core.
- **Single lobe:** These sources have a bright core and a single diffuse lobe separated from the core.
- **Centre-bright:** These are centre-bright sources with visible core and lobes.
- **Centrally peaked ellipse:** Sources with elongated structure and an intensity peak in the central region, but no clear core and lobe structure.
- **Symmetric double:** These sources have low resolution and appear as two Gaussian intensity regions.
- **Edge-bright:** These sources present bright hotspots; for some, the core and the lobes are visible.
- **WAT FRM:** Wide-Angle Tail FR Mixed sources are bent sources with mixed brightening and a mild bending angle.
- **NAT FRM:** The Narrow-Angle Tail FR Mixed sources are bent sources with mixed brightening (edge and centre) with significant bending angle (not necessarily smaller than 90°).
- **Circular diffuse:** These sources have a relatively circular shape with intensity peaks inside the more diffuse emission. From a morphological point of view, they could be associated with radio halos. This could be confirmed by cross-matching with X-ray observations.

The labelled subsample is highly imbalanced, with the largest cluster containing over 30 times as many sources as the smallest cluster. Fig. 7 presents the members from each cluster with the highest mean SI-SSIM values, highlighting the typical morphology associated with each cluster.

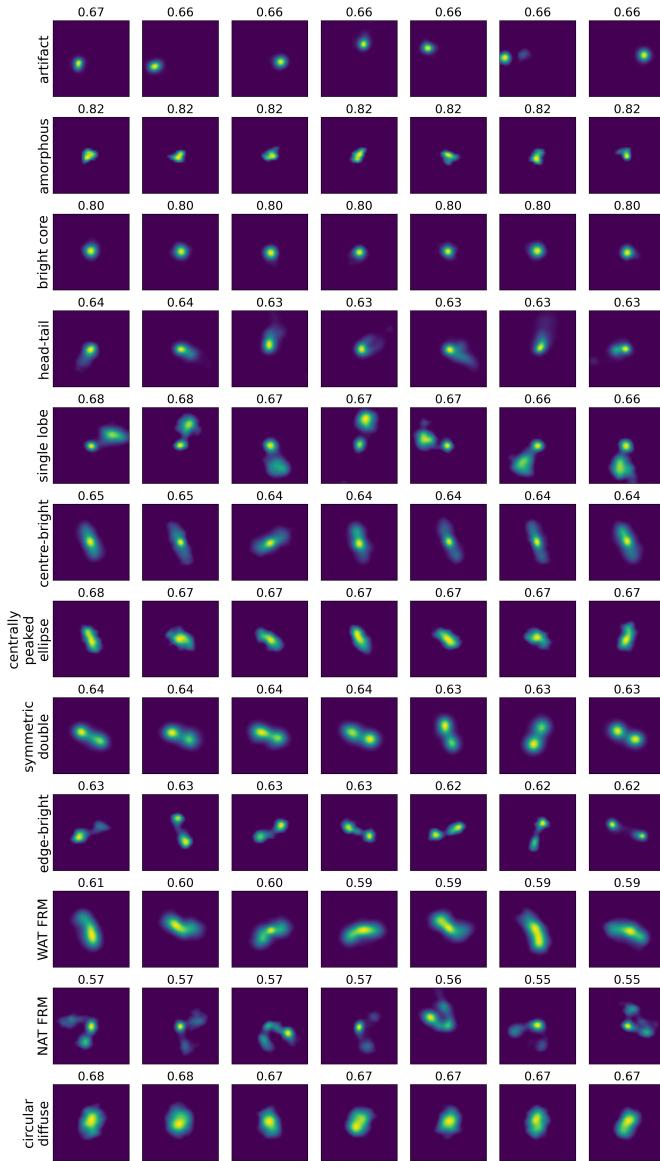
Next, we examine the distribution of sources corresponding to each cluster label within the space defined by the first two principal components of the representations from the pre-training stage. This distribution is illustrated in the left plot of Fig. 8, where each cluster is distinguished by a unique colour. Additionally, a representative image is included to illustrate the typical morphology characteristic of that cluster. Most of the clusters are located within distinct, bounded regions, separated from others. However, some clusters exhibit significant overlap within this two-dimensional space.

Qualitatively, we observe three trends that describe the representation space structure:

- Considering only the clusters at the edge of the distribution, we find that single-component sources are located on the northeastern side while double-component sources are on the southwestern side of the distribution.
- The fraction of active pixels grows anti-clockwise starting at the eastern part (bright core sources) and finishing at the northern part (circular diffuse sources) of the representation distribution.
- The morphological complexity, reflected by the number of source components or bending angle, increases towards the centre of the distribution.

#### 4.3. Adapting the Model to Classification Task

With the constructed labelled subsample, we can proceed to fine-tune the encoder using the cluster labels by adding an MLP on top of the encoder. The supervised learning task employed for classification during this fine-tuning process allows the network to learn features that result in more meaningful representation encoding and more accurate class predictions than those obtained through self-supervision.



**Fig. 7.** Each row corresponds to a distinct cluster, showcasing the seven sources with the highest mean SI-SSIM values to highlight the typical morphology exhibited by the cluster. The mean SI-SSIM value of each source is displayed above each panel.

We use two weighted loss functions during fine-tuning, the cross-entropy loss and SNNL. The loss weights account for the imbalanced nature of the dataset. The weight for class  $i$  is calculated as  $w_i = 1/n_i \times s_d/n_c$ , where  $n_i$  represents the number of sources in class  $i$ ,  $s_d$  is the dataset size, and  $n_c$  is the number of classes.

The cross-entropy loss is a widely used function for multi-class classification tasks and is given as:

$$\mathcal{L}^{\text{Xent}} = - \sum_{c=1}^{n_c} y_c \log(p_c), \quad (2)$$

where  $n_c$  is the number of classes,  $y_c$  is 1 if the image belongs to class  $c$  and 0 otherwise, and  $p_c$  is the predicted class probability.

SNNL was initially introduced in Salakhutdinov & Hinton (2007) to fine-tune a metric learning algorithm and enhance the

$k$ -nearest neighbour classification. Later, Frosst et al. (2019) improved this loss function and applied it to assess the degree of class entanglement in the representation space. A high value of SNNL indicates a significant overlap of the classes, while a low value implies that the classes are well separated and form compact clusters in the representation space Frosst et al. (2019). The minimisation of SNNL decreases the ratio of the sum of distances between the member pairs of class  $i$  over the sum of distances between all data pairs in the batch. Mathematically this is given by:

$$\mathcal{L}_i^{\text{SNN}} = -\log \frac{\sum_{j \neq i, y_j=y_i, j=1,\dots,N} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau^{\text{SNN}})}{\sum_{k \neq i, k=1,\dots,N} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau^{\text{SNN}})}, \quad (3)$$

where  $y_i$  is the class label of the  $i$ -th data point and  $\tau^{\text{SNN}}$  is the temperature of SNNL.

The total fine tuning loss functions is then given by:

$$\mathcal{L}^{\text{FT}} = \alpha \mathcal{L}^{\text{Xent}} + (1 - \alpha) \mathcal{L}^{\text{SNN}}, \quad (4)$$

where  $\alpha$  is the loss function weighting factor.

We fine-tune the encoder optimising the validation accuracy with a batch size of 512 and a learning rate of 0.001 for 100 epochs. We use an SNNL temperature parameter of  $\tau^{\text{SNN}} = 0.2$  and a loss function weighting factor of  $\alpha = 0.5$ . A higher temperature value enhances the compactness of each cluster while increasing the distance between clusters, particularly for those that differ significantly based on the extracted features.

#### 4.4. Deep Ensemble Training

We selected deep ensembles to obtain more reliable probability estimates for our class predictions. Lakshminarayanan et al. (2016) investigated how the training of a deep ensemble should be conducted to obtain well-calibrated (predicted uncertainties and actual results observed in repeated experiments agree) and generalisable predicted probabilities. They concluded that training an ensemble of independent, randomly initialised networks with cross-entropy loss and averaging the predicted probabilities leads to the desired result. In this case, the predicted probabilities are given by:

$$p(y|\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M p_{\theta_m}(y|\mathbf{x}, \theta_m), \quad (5)$$

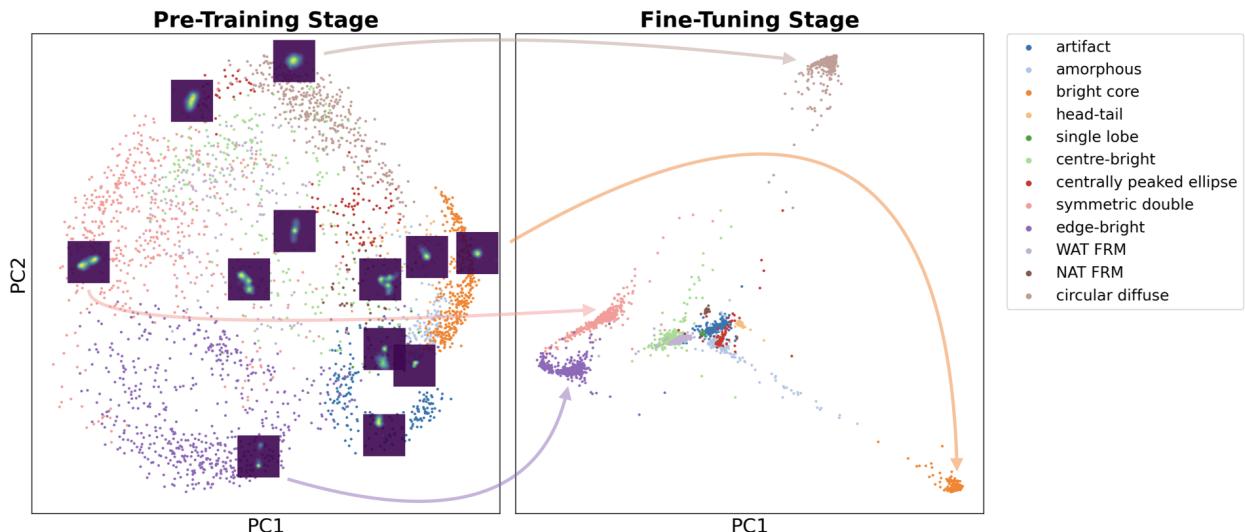
where  $y$  is the class label,  $\mathbf{x}$  is the image,  $M$  is the number of networks that constitute the ensemble, and  $\theta_m$  represents the parameters of the  $m$ -th network.

We constructed our deep ensemble by fine-tuning the encoder 50 times, with the weights of the classification head being randomly initialized for each iteration. The ensemble representations are derived by averaging the representations predicted by each member of the ensemble.

#### 4.5. Fine-Tuning Results

We now analyse the results of the fine-tuning process. Utilising the labelled subsample during fine-tuning achieves classification accuracies of 99.2%, 99.2%, and 97.2% for training, validation, and test datasets, respectively.

The distribution of the PCA-transformed representations post fine-tuning is depicted in the right panel of Fig. 8. Most clusters are situated near the centre of the space defined by



**Fig. 8.** This figure illustrates the representations of sources from the labelled subsample within the space defined by the first two principal components, with colours corresponding to cluster labels as listed in the legend. **Left:** Representations from the pre-training stage are displayed, showing a dispersed distribution pattern. Additionally, a representative image from each cluster is provided to highlight the morphological structure within the distribution. **Right:** Post fine-tuning representations are presented, demonstrating compact clustering according to the labels. Arrows are included to depict the evolution of selected clusters from the pre-training to the fine-tuning stage.

the first two principal components, while four clusters are positioned farther away. The most distant clusters are the *bright core* sources and the *circular diffuse* sources. These classes exhibit the most distinct morphologies, being single-component objects. The other two clusters that are notably distant from the others are the *symmetric double* and *edge-bright* sources, which are located close to each other. Both of these have unique morphologies characterised by two nearly identical components.

We find that the sources are organised into more compact clusters (right panel of Fig. 8) compared to their distribution at the pre-training stage (left panel of Fig. 8), with inter-cluster distances more strongly influenced by the morphological similarity of the clusters. The improved separation of clusters in the fine-tuned space is an anticipated outcome of the model incorporating class label information. In general, cluster compactness and inter-cluster distance can be further adjusted by tuning the temperature parameter of the SNNL. Despite this, some clusters still overlap. Visual inspection shows that overlapping clusters share morphological properties such as hotspot distance, source size, and resolution.

## 5. Results

In this section, we present the classification of all sources in the unlabelled sample, reflecting the predictions made with the deep ensemble.

The number of sources in each class are listed in Table 1. The table includes counts for the entire unlabelled sample as well as for sources with a class probability greater than 90%. Focusing on the latter, *double sources* are the most prevalent, comprising 33.8% of this set. Additionally, 14.9% of the sources are *edge-bright*, while 10.6% are *centre-bright*. Furthermore, 19.2% of the sources exhibit a bending angle. Lastly, images classified as contamination, i.e. *artifact* and *amorphous* sources, account for 5.8% of the dataset.

In Fig. 9, we display the five sources with the highest and lowest class probabilities for each class. By showcasing these images, we illustrate the model's confidence in its predictions.

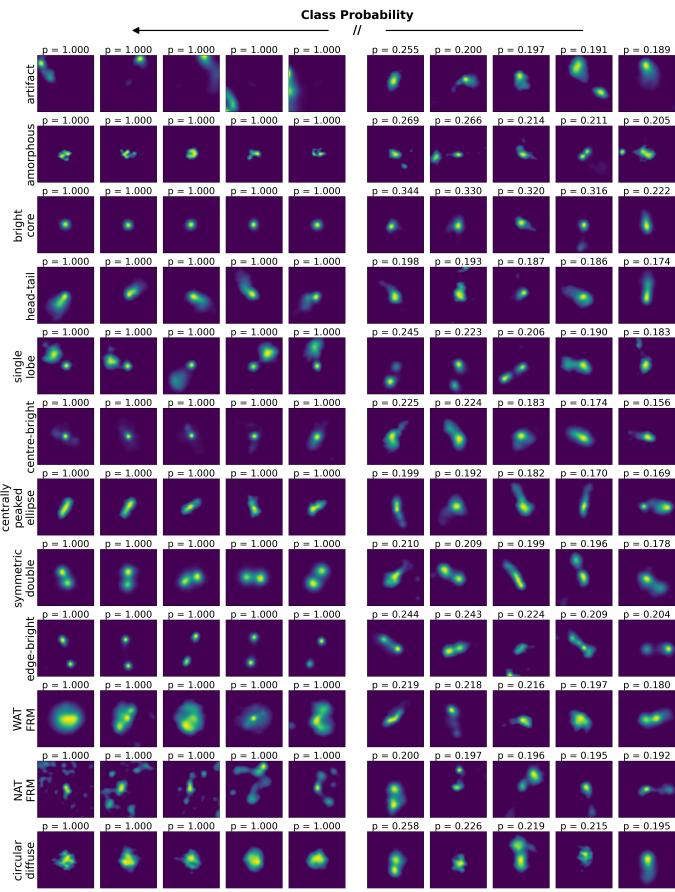
	Class Name	Whole Sample	p ≥ 90%
1	artifact	1258	910
2	amorphous	873	555
3	bright core	1130	919
4	head-tail	2441	1292
5	single lobe	607	228
6	centre-bright	4386	2682
7	centrally peaked ellipse	1506	616
8	symmetric double	12544	8564
9	edge-bright	6314	3774
10	WAT FRM	7255	3721
11	NAT FRM	2491	1151
12	circular diffuse	1425	939
	Total	42230	25351

**Table 1.** Number of sources in each class, presented for the entire unlabelled sample and for the subset of sources with a class probability greater than 90%.

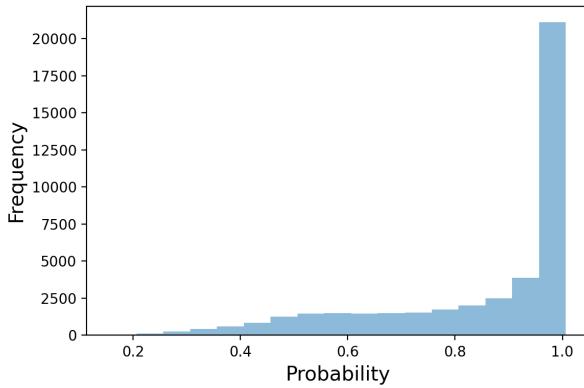
The sources with the highest probabilities serve as prototypical examples that effectively represent most classes, particularly those with simpler morphology, helping to confirm the model's reliability in these categories. Meanwhile, examining the sources with the lowest probabilities helps identify ambiguous or challenging cases, highlighting areas where the model may struggle and providing insights into the potential areas for further improvement.

We display the distribution of class probabilities in Fig. 10. These probabilities range from 0.16 to 1, with 60% of the sources having a class probability greater than 0.9.

Next, we analyse the distribution of these sources in the space defined by the first two principal component of the ensemble representations. This distribution is shown in Fig. 11, where



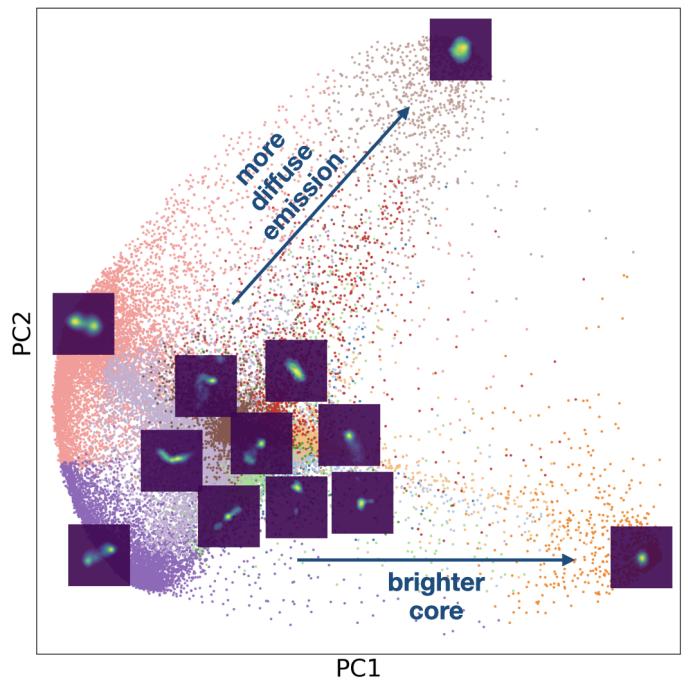
**Fig. 9.** Each row corresponds to a different class. The first five columns display the sources with the highest class probabilities, while the last five columns show those with the lowest class probabilities. This figure illustrates the model’s confidence in its predictions. The class probability for each source is indicated at the top of each panel.



**Fig. 10.** Distribution of the deep ensemble class probabilities for all 42 230 images in the unlabelled sample, using a bin width of 0.05.

a prototypical image is provided for each class. The clusters containing *bright core* and *circular diffuse* sources remain being the more distant (as observed with a single fine-tuned model), while the gap between double-component sources and the other clusters has been partially filled by the previously unlabelled sources.

The distribution indicates that there are sources linking the main cluster block with the two more distant clusters. Visual in-



**Fig. 11.** For the entire unlabelled sample, we present the distribution of deep ensemble representations in the space defined by the first two principal components, with colours indicating predicted classes. An example image for each class is included to illustrate the morphological evolution within this space.

spection reveals that these sources exhibit morphologies that become more similar to those of the distant clusters as they draw nearer. For the sources approaching the *bright core* cluster, a bright circular structure begins to dominate their morphology. Meanwhile, sources moving closer to the *circular diffuse* cluster increasingly show a bright intensity peak surrounded by diffuse emission.

We describe the morphological evolution in the distribution as follows:

- The active pixel fraction increases in clockwise direction from the southeast to the northeast with the double-component sources located midway.
- At the centre of the distribution, we find the bent sources, which include both centre-bright and edge-bright sources.
- To their right, three core-dominated clusters form a diagonal. This starts as the lower cluster, which contains centre-bright sources with weak lobes; in the middle are single lobe sources with a single more diffuse lobe; and the upper cluster consists of ellipsoidal centre-dominated sources.
- Below this diagonal, the artifacts and amorphous sources are situated very close to each other.

## 6. Discussion

In this project, we explored how SSL can classify sources in radio astronomical surveys, focusing specifically on LoTSS-DR2.

We found that pre-training an encoder using the learning task proposed in Chen et al. (2020) to predict representations that primarily capture morphological information of radio galaxies, beyond centre-bright and edge-bright examples, is very challenging. The random augmentations suggested in Chen et al. (2020)

are insufficient for the model to distinguish the diverse morphological properties of radio galaxies. Consequently, we developed a new random augmentation, *Random Structural View* (see Sect. 3.2), based on image-level similarity, to tailor the contrastive learning task to the radio astronomical context.

The random augmentation *Random Structural View* uses SI-SSIM to measure source similarity at the image-level. This metric could be enhanced to be size invariant, allowing, for example, two *symmetric double* sources to have the same SI-SSIM value regardless of their apparent size in the image. Additionally, exploring alternative metrics could yield options that are potentially more suitable for this task.

We found that we were unable to fully handle the complete morphological heterogeneity present in the survey, particularly when considering all resolved sources in the catalogue. Consequently, we had to apply stringent cuts to the dataset, which both improved the clarity of source appearance in the images and slightly reduced the morphological diversity of the sources (see Sect. 2). After reducing the size of the unlabelled sample, we pre-trained the SSL model with randomly initialised model weights. For future work, it could be interesting to first pre-train on the whole data release and afterwards continue the pre-training on the filtered dataset. Furthermore, We do not rule out that a different method or additional random augmentations could accommodate the complete morphological diversity. Methodological modifications could include using vision transformers (Dosovitskiy et al. 2020), potentially with increased model size, in place of convolution-based encoders, as well as adopting a generative learning framework, such as the one proposed by Bao et al. (2021), instead of the invariance-based framework used in this work.

We would like to emphasise that the classification scheme used in the linear evaluation protocol can significantly influence the effectiveness of capturing morphological information within the representations. Therefore, it is advantageous to have a clean, large, labelled dataset that encompasses a diverse range of morphological classes.

The resulting representation space successfully incorporates the morphological information we aimed to capture. Among the morphological properties encoded in the representation vectors are data quality (including artefacts and amorphous sources), source type (e.g. radio halo, bright core, or radio galaxy), the number of source components, the relative brightness of core and hotspots, and the degree of source bending. Some properties reflected in the representations, which are not intrinsically related to the source morphology, include the quality of observational calibration, source size, and source resolution.

Chen et al. (2020) reported the emergence of distinct clusters for different ImageNet (Russakovsky et al. 2014) classes within the representation space (see Fig. B.4 in their original paper). Therefore, we anticipated achieving a similar representation space structure for the morphological classes of radio galaxy. Under this assumption, applying a clustering algorithm to the representations should facilitate the identification of radio galaxy classes, as suggested by Mohale & Lochner (2024). However, the obtained representation space structure is more diffuse than required for that approach (see Fig. 6). This necessitated the creation of a labelled subsample tailored to the morphological diversity of the unlabelled sample to fine-tune the encoder and perform the classification as a downstream task (see Sect. 4). We acknowledge that a more suitable model and a better-designed learning task could potentially lead to the desired emergence of morphological clusters.

The most significant difference between SimCLR (Chen et al. 2020) and our work is likely the imbalanced nature of our dataset and the absence of class labels. It is known that SimCLR, along with other SSL algorithms, performs poorly when trained on imbalanced data due to its hidden uniform prior (Assran et al. 2022). For future work, we propose using models that do not have this limitation, such as the extended version of the Masked Siamese Networks model presented in Assran et al. (2022).

Avoiding the use of class labels at the initial stage is a crucial aspect of our approach. However, the availability of class labels representing the entire class diversity of the dataset makes a significant difference, as it allows for a comprehensive evaluation to determine whether the information contained in the representations sufficiently describes the whole dataset with the linear evaluation protocol. This evaluation allows for a more efficient design of the learning problem, such as by developing new random augmentations or combining loss functions.

Additionally, objects in natural images are often more distinctly quantised compared to those in radio sources. For instance, while a cat and an aeroplane may share some features, they are categorically distinct objects. In contrast, a centre-bright bent source and a linear edge-bright source are variations of the same object, influenced by different environmental conditions. Furthermore, the appearance of the radio galaxies in images is influenced by their redshift and the resolution achieved, posing a significant challenge when attempting to map these sources into an abstract representation space.

Mingo et al. (2019) created a dataset of FRI and FRII radio galaxies using an automated algorithm complemented by visual analysis from the LoTSS-DR1 (Shimwell et al. 2019). They identified 1256 FRI and 423 FRII galaxies, with a subsample of 459 FRI sources exhibiting a bent morphology. Interestingly, our sample has fewer FRI sources than FRII sources with a class probability higher than 90%. Specifically, we find 2682 FRI and 3774 FRII sources. Additionally, we found 4872 sources exhibiting bent morphology, which are classified as different categories in our dataset, in contrast to the classification approach used by Mingo et al. (2019). The differing class definitions between our work and theirs may explain the discrepancy in the dominant FR type observed in our samples. Another factor contributing to the discrepancy is the different selection functions employed in each study, which can result in significantly diverging populations. Our findings suggest the importance of exploring classifications beyond the traditional FRI and FRII categories to fully capture the diversity of radio sources from sky surveys.

We would like to discuss potential applications of this work. Firstly, interesting clusters could serve as the basis for astrophysical analyses. Anomalies or sources with uncommon morphologies can be identified among images with the lowest class probabilities. Once an intriguing object is identified, the representation space from the pre-training phase could be used to search for similar objects, potentially uncovering a missing class. With a subset of approximately 30 sources exhibiting this morphology, this new “cluster” could be used to fine-tune the encoder and reclassify the dataset to identify all instances of this morphology. This process could be performed iteratively, leading to comprehensive and efficient morphological classification and representation space. In the next step, the morphological classes that we found in this work will be plotted against other observables from the value added catalogue, such as stellar masses, or derived quantities such as galaxy overdensity to study the dependence on the environment. Finally, the ability to explicitly incorporate complex morphological properties into the representation space while encompassing a wide variety of radio source mor-

phologies is a significant step toward developing a foundation model for radio surveys.

### 6.1. Comparison to other methods

In this work, we used SSL methods to study the morphological classification of radio galaxies that include a broader range of classes than those typically found in existing labelled datasets. Related works are the studies of SSL methods for radio astronomy and unsupervised classification.

Studies of SSL methods for radio astronomy include: The SSL pre-training with large amounts of unlabelled radio source images and the fine-tuning with FRI and FRII labelled images showing an improvement in classification accuracy and the advantage of reducing the number of required labelled images (Hossain et al. 2023; Slijepcevic et al. 2024). In our work, we use a similar SSL model as done in these studies and take advantage of the small number of labelled data required for fine-tuning this models.

Mohale & Lochner (2024) tested the unsupervised classification of FRI and FRII radio galaxies by combining an SSL model and clustering algorithm. We have tested this approach for our classification. However, the representation space structure is not clustered enough to obtain reasonable results with numerous classes.

Another way of classifying radio sources is by number of radio components and number of intensity peaks, as has been done in the RGZ with the help of citizen scientists. Riggi et al. (2024) studied among others the use of an SSL algorithm to perform this task. Separately, the transfer of these labels with the help of an SOM and a random forest (Breiman 2001) was studied in Galvin et al. (2019).

Cecconello et al. (2024) performed a benchmark study comparing different SSL models and different classification schemes providing the linear evaluation for the different combinations. In contrast, we are unable to provide classification accuracies because we do not have true labels for the classified dataset.

The unsupervised classification of radio sources by their visible morphology was tested by Ralph et al. (2019) using the combination of an autoencoder (Sanger 1989), a SOM, and the clustering algorithm  $k$ -means (Lloyd 1982). Later, Mostert et al. (2021) performed the same task using a rotation and flipping invariant SOM followed by manual grouping and labelling of the identified morphological classes. The later was done within the search of the rarest morphologies in LoTSS-DR1.

A great advantage of SOMs over SSL algorithms is their interpretability. We attempted to identify the morphological features detected by the SSL algorithm in each cluster through visual inspection. In contrast, SOMs learn prototypes that are optimized to best fit the training dataset. These prototypes show directly which features are being taking into account by the algorithm. The downside of SOMs is that they describe the dataset using a discrete number of prototypes, which must be selected empirically. Meanwhile, our algorithm first learns a continuous representation of the input data, having more degrees of freedom to describe the dataset. Additionally, the use of representations has the advantage of enabling the use of other data types, such as time-series data, spectra, or language data, e.g. verbal descriptions of the morphology (Bowles et al. 2023).

Galvin et al. (2019) used, additionally to the radio frequency maps, infrared observations. This is an essential improvement, since infrared or optical maps enable the identification of the host galaxy location. This helps to determine the associated components and improves the quality of the morphological class es-

timation. This could also be implement in the SSL approach, by adding these maps as an additional channel to the input images, e.g. by extracting the same sky region from the corresponding survey. Galvin et al. (2020) showed with this approach that a SOM could group radio components and identify the host galaxy.

## 7. Conclusions

In this work, we investigated the morphological classification of a subset of resolved sources from LoTSS-DR2 using an SSL algorithm. The final classification consists of 12 classes, including known classes of radio galaxies, such as NATs and WATs, as well as new classes. Our classification scheme incorporates distinctions among core-bright, edge-bright and bent sources, as used in previous classification studies, by considering the overall morphological shape and symmetry of the radio emission.

Our approach's characteristic feature is the creation of a small labelled subsample using the representation space learned during pre-training. This strategy aims to capture the morphological diversity of the unlabelled sample within the identified classes as exhaustively as possible, thereby improving the significance of the final source classification. To create the labelled subsample, we required a representation space that encompassed extensive morphological information. To enhance the morphological information within the representations, we designed the random augmentation *Random Structural View*. This augmentation uses the most similar images, evaluated at the image level with a modified version of SSIM, in order to create the second view for the SSL learning task.

Our training strategy involved pre-training using SSL with all images contained in the unlabelled sample, fine-tuning the deep ensemble with the created labelled subsample, and then predicting class probabilities for all images in the unlabelled sample to achieve the final classification.

The deep ensemble can be used to infer the class probabilities of radio source images unseen by the model. We also study the meaning of the position in the fine-tuned representation space. Additionally, the pre-trained model can be used to search for images that are morphologically similar to a query image within its neighbouring regions in the representation space.

The identification of the *single lobe* class in the unlabelled sample suggests that our technique has the potential to uncover morphological classes that may not be readily apparent to the eye. Finally, the improvement of the learned representations regarding morphological information is a step towards the unsupervised classification of enormous astronomical surveys and the development of foundation models.

Obtaining high-quality representations could enable the identification of anomalies, i.e. sources with uncommon morphologies, in radio survey data with SSL algorithms. Lochner et al. (2023) used active learning as an extension of the pipeline to successfully identify anomalies, leading to the discovery of a ring-like radio source in the MeerKAT Galaxy Cluster Legacy Survey.

**Acknowledgements.** Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 460248186 (PUNCH4NFDI). GK, MB, and LLS acknowledge funding by the DFG under Germany's Excellence Strategy – EXC 2121 "Quantum Universe" – 390833306.

## References

- Abbasi Koohpayegani, S., Tejankar, A., & Pirsavash, H. 2021, arXiv e-prints, arXiv:2105.07269
- Akhmetzhanova, A., Mishra-Sharma, S., & Dvorkin, C. 2024, MNRAS, 527, 7459
- Assran, M., Balestrieri, R., Duval, Q., et al. 2022, arXiv e-prints, arXiv:2210.07277
- Assran, M., Duval, Q., Misra, I., et al. 2023, arXiv e-prints, arXiv:2301.08243
- Baldi, R. D. 2023, A&A Rev., 31, 3
- Balestrieri, R., Ibrahim, M., Sobal, V., et al. 2023, arXiv e-prints, arXiv:2304.12210
- Banfield, J. K., Wong, O. I., Willett, K. W., et al. 2015, MNRAS, 453, 2326
- Bao, H., Dong, L., Piao, S., & Wei, F. 2021, arXiv e-prints, arXiv:2106.08254
- Bhatta, G., Gharat, S., Borthakur, A., & Kumar, A. 2024, MNRAS, 528, 976
- Bowles, M., Tang, H., Vardoulaki, E., et al. 2023, MNRAS, 522, 2584
- Brand, K., Grobler, T. L., Kleynhans, W., et al. 2023, MNRAS, 522, 292
- Breiman, L. 2001, Machine learning, 45, 5
- Cecconello, T., Riggi, S., Becciani, U., et al. 2024, arXiv e-prints, arXiv:2411.14078
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020, arXiv e-prints, arXiv:2002.05709
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. 2020, arXiv e-prints, arXiv:2010.11929
- Dubois, J., Fraix-Burnet, D., Moulata, J., Sharma, P., & Burgarella, D. 2022, A&A, 663, A21
- Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., & Zisserman, A. 2021, arXiv e-prints, arXiv:2104.14548
- Fanaroff, B. L. & Riley, J. M. 1974, MNRAS, 167, 31P
- Frosst, N., Papernot, N., & Hinton, G. 2019, arXiv e-prints, arXiv:1902.01889
- Galvin, T. J., Huynh, M., Norris, R. P., et al. 2019, PASP, 131, 108009
- Galvin, T. J., Huynh, M. T., Norris, R. P., et al. 2020, MNRAS, 497, 2730
- Goyal, P., Dollár, P., Girshick, R., et al. 2017, arXiv e-prints, arXiv:1706.02677
- Grill, J.-B., Strub, F., Alché, F., et al. 2020, arXiv e-prints, arXiv:2006.07733
- Guo, X., Liu, C., Qiu, B., et al. 2022, MNRAS, 517, 1837
- Hardcastle, M. J., Horton, M. A., Williams, W. L., et al. 2023, A&A, 678, A151
- He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv e-prints, arXiv:1512.03385
- Hinton, G. E. & Salakhutdinov, R. R. 2006, Science, 313, 504
- Hossain, M. S., Roy, S., Asad, K. M. B., et al. 2023, Procedia Computer Science, 222, 601
- Huang, Z., Chen, J., Zhang, J., & Shan, H. 2021, arXiv e-prints, arXiv:2111.11821
- Huertas-Company, M., Sarmiento, R., & Knapen, J. H. 2023, RAS Techniques and Instruments, 2, 441
- Kempner, J. C., Blanton, E. L., Clarke, T. E., et al. 2004, in The Riddle of Cooling Flows in Galaxies and Clusters of galaxies, ed. T. Reiprich, J. Kempner, & N. Soker, 335
- Kohonen, T. 2001, Self-Organizing Maps
- Lakshminarayanan, B., Pritzel, A., & Blundell, C. 2016, arXiv e-prints, arXiv:1612.01474
- Lao, B., Andernach, H., Yang, X., et al. 2025, arXiv e-prints, arXiv:2501.09883
- Lao, B., Jaiswal, S., Zhao, Z., et al. 2023, Astronomy and Computing, 44, 100728
- Lloyd, S. 1982, IEEE transactions on information theory, 28, 129
- Lochner, M., Rudnick, L., Heywood, I., Knowles, K., & Shabala, S. S. 2023, MNRAS, 520, 1439
- Loshchilov, I. & Hutter, F. 2016, arXiv e-prints, arXiv:1608.03983
- Loshchilov, I. & Hutter, F. 2017, arXiv e-prints, arXiv:1711.05101
- Lukic, V., Brüggen, M., Mingo, B., et al. 2019, MNRAS, 487, 1729
- Luo, X., Zheng, S., Jiang, Z., et al. 2024, A&A, 683, A104
- Ma, Z., Xu, H., Zhu, J., et al. 2019, ApJS, 240, 34
- Maslej-Krešnáková, V., El Boucifry, K., & Butka, P. 2021, MNRAS, 505, 1464
- McInnes, L., Healy, J., & Astels, S. 2017, The Journal of Open Source Software, 2, 205
- McInnes, L., Healy, J., & Melville, J. 2018, arXiv e-prints, arXiv:1802.03426
- Mingo, B., Croston, J. H., Hardcastle, M. J., et al. 2019, MNRAS, 488, 2701
- Miraghaei, H. & Best, P. N. 2017, MNRAS, 466, 4346
- Mohale, K. & Lochner, M. 2024, MNRAS[arXiv:2311.14157]
- Mostert, R. I. J., Duncan, K. J., Röttgering, H. J. A., et al. 2021, A&A, 645, A89
- Ndung'u, S., Grobler, T., Wijnholds, S. J., Karastoyanova, D., & Azzopardi, G. 2023, New A Rev., 97, 101685
- Norris, R. P., Hopkins, A. M., Afonso, J., et al. 2011, PASA, 28, 215
- Pérez-Díaz, V. S., Martínez-Galarza, J. R., Caicedo, A., & D'Abrusco, R. 2024, MNRAS, 528, 4852
- Ralph, N. O., Norris, R. P., Fang, G., et al. 2019, PASP, 131, 108011
- Riggi, S., Cecconello, T., Palazzo, S., et al. 2024, PASA, 41, e085
- Rudnick, L. 2021, Galaxies, 9, 85
- Russakovskiy, O., Deng, J., Su, H., et al. 2014, arXiv e-prints, arXiv:1409.0575
- Rustige, L., Kummer, J., Griese, F., et al. 2023, RAS Techniques and Instruments, 2, 264
- Salakhutdinov, R. & Hinton, G. 2007, in Proceedings of Machine Learning Research, Vol. 2, Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, ed. M. Meila & X. Shen (San Juan, Puerto Rico: PMLR), 412–419
- Sanger, T. D. 1989, Neural Networks, 2, 459
- Sarmiento, R., Huertas-Company, M., Knapen, J. H., et al. 2021, ApJ, 921, 177
- Shimwell, T. W., Hardcastle, M. J., Tasse, C., et al. 2022, A&A, 659, A1
- Shimwell, T. W., Röttgering, H. J. A., Best, P. N., et al. 2017, A&A, 598, A104
- Shimwell, T. W., Tasse, C., Hardcastle, M. J., et al. 2019, A&A, 622, A1
- Slijepcevic, I. V., Scaife, A. M. M., Walmsley, M., et al. 2022, MNRAS, 514, 2599
- Slijepcevic, I. V., Scaife, A. M. M., Walmsley, M., et al. 2024, RAS Techniques and Instruments, 3, 19
- Stein, G., Harrington, P., Blaum, J., Medan, T., & Lukic, Z. 2021, arXiv e-prints, arXiv:2110.13151
- Stroe, A., Catlett, V., Harwood, J. J., Vernstrom, T., & Mingo, B. 2022, ApJ, 941, 136
- Tohill, C., Bamford, S. P., Conselice, C. J., et al. 2024, ApJ, 962, 164
- van der Maaten, L. & Hinton, G. 2008, Journal of Machine Learning Research, 9, 2579
- Vantyghem, A. N., Galvin, T. J., Sebastian, B., et al. 2024, Astronomy and Computing, 47, 100824
- Vega-Ferrero, J., Huertas-Company, M., Costantin, L., et al. 2024, ApJ, 961, 51
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. 2004, IEEE Transactions on Image Processing, 13, 600
- Wong, O. I., Garon, A. F., Alger, M. J., et al. 2025, MNRAS, 536, 3488
- Wu, C., Wong, O. I., Rudnick, L., et al. 2019, MNRAS, 482, 1211