The Astrophysical Journal Supplement Series, 265:30 (14pp), 2023 March https://doi.org/10.3847/1538-4365/acb440

Spurious Correlations between Galaxies and Multiepoch Image Stacks in the DESI Legacy Surveys

Edgar Eggert and Boris Leistedt
Department of Physics, Imperial College London, Blackett Laboratory, Prince Consort Road, London, SW7 2AZ, UK; b.leistedt@imperial.ac.uk
Received 2022 July 19; revised 2023 January 6; accepted 2023 January 14; published 2023 March 9

Abstract

A nonnegligible source of systematic bias in cosmological analyses of galaxy surveys is the on-sky modulation that is caused by foregrounds and variable image characteristics, such as observing conditions. Standard mitigation techniques perform a regression between the observed galaxy density field and sky maps of the potential contaminants. Such maps are ad hoc lossy summaries of the heterogeneous sets of coadded exposures that contribute to the survey. We present a methodology for addressing this limitation, and we extract spurious correlations between the observed distributions of galaxies and arbitrary stacks of single-epoch exposures. We study four types of galaxies (luminous red galaxies, emission-line galaxies, quasars, and Lyman-break galaxies) in the three regions of the DESI Legacy Surveys (North, South, and Dark Energy Survey), resulting in 12 samples with varying levels and types of contamination. We find that the new technique outperforms the traditional ones in all cases, and is able to remove higher levels of contamination. This paves the way for new methods that extract more information from multiepoch galaxy survey data and mitigate large-scale biases more effectively.

Unified Astronomy Thesaurus concepts: Large-scale structure of the universe (902); Observational cosmology (1146)

1. Introduction

Over the last decades, the study of the statistical properties of galaxies has become a pillar of observational cosmology. Giant surveys of the night sky help us to test increasingly sophisticated models of dark matter, dark energy, gravity, and galaxy formation and evolution. One of the challenges of such surveys is to maintain systematic biases at sufficiently low levels.

One potential bias is the dependency of the data on the observing conditions (e.g., seeing), the calibration (zero-points), and the foregrounds (e.g., stars and Galactic dust). This phenomenon is well known and it often comes into consideration when planning observations, as it is unavoidable that the data will not be exactly uniform (Shafer & Huterer 2015; Awan et al. 2016; Almoubayyed et al. 2020). This is sometimes referred to as the transfer function of the survey; the resulting spurious correlations in the galaxy numbers can mimic cosmological imprints and bias the downstream inference. Hence, they must be mitigated, as demonstrated in recent cosmological analyses by the Dark Energy Survey (DES; see, e.g., Gatti et al. 2021; Wagoner et al. 2021; Rodríguez-Monroy et al. 2022) and the Kilo Degree Survey (see, e.g., Johnston et al. 2021; Loureiro et al. 2022).

One solution to this problem is to map out the potential contaminants on the sky, and ignore the corresponding modes when extracting the cosmological signal from galaxy number counts. This mode (de)projection is effective, but computationally expensive. A more tractable approach is to remove the (best-fit) correlations between the galaxy number counts and the contaminants at the level of maps of the estimated power spectra. While this is in principle a simple regression,[1] deciding on the form of the model and its inputs is critical. Typically, the inputs include any property of the survey that could affect galaxy measurements: astrophysical foregrounds (e.g., stellar density and dust extinction), the observing conditions of the survey (e.g., seeing, airmass, and background noise), etc. The most common form of model is linear in the inputs, and predicts the galaxy number counts or the overdensity field. More complex models (e.g., based on neural networks) can provide significant improvements for some galaxy samples (see, e.g., Rezaie et al. 2020; Chaussidon et al. 2022). Omitting inputs or adopting an overly simplistic model will lead to an incomplete subtraction of the spurious noncosmological correlations. On the contrary, excessively complex models can lead to overfitting and oversubtraction, because of random "chance" correlations between the inputs and the cosmological signal.[2] However, this effect can be predicted analytically, simulated, or mitigated by preventing overfitting (as we do here).

Finally, a powerful avenue for modeling the transfer function is to employ synthetic source injections (SSIs) in images. These play an important role in current surveys (Huang et al. 2017; Everett et al. 2022), and their role will grow, as systematic biases must be modeled at higher accuracy. However, they are very computationally demanding, so they likely complement the regression methodologies covered here.

This paper deals with a specific aspect of this problem: the multiepoch nature of modern surveys. Indeed, to reach sufficient depth, surveys take multiple exposures covering the same sky area, to be coadded to reduce noise. The exposures never align exactly, and are not taken under the exact same observing conditions, which is further complicated by random

[1] This process must be done for each galaxy sample under consideration, since their transfer functions would be different, as illustrated in this paper.
[2] Mode projection is not subject to this effect, since it is equivalent to a Bayesian marginalization of the fit parameters.

artifacts, such as contamination by cosmic rays, stars, or bad exposures generally. As a result, any particular pointing on the sky is covered by a complicated stack of images, further increasing the complexity of the transfer function. All of the aforementioned methodologies and previous works perform a compression (at each sky location) into a set of summary statistics (e.g., the mean), resulting in maps of potential contaminants (abbreviated as "contaminant maps" in the remainder of this paper). Removing correlations between these maps and the galaxy number counts neglects the complexity of the underlying image stacks. For example, coadded exposures are often diverse and include outliers, i.e., single epochs with extreme properties with respect to the rest of the other overlapping exposures or the rest of the survey. To our knowledge, their effect on the uniformity of the transfer function and the accuracy of the aforementioned techniques is unknown. While the propagation of some of the per-epoch summary statistics through coaddition is well understood (e.g., the point-spread function), the end-to-end effect on a galaxy catalog is a complicated function of the input images, and there are almost certainly sources of contamination that are missed by current methods, or for which the propagation through coaddition is not understood. This paper seeks to address one aspect of this issue by employing a model architecture—deep sets—that is able to work with the properties of heterogeneous image stacks instead of coadd summaries.

Deep sets[3] (Zaheer et al. 2017; Korshunova et al. 2018; Murphy et al. 2018; Soelch et al. 2019; Wagstaff et al. 2019; Bronstein et al. 2021; Cranmer et al. 2021) are types of neural networks that are capable of learning functions over unordered sets. They have found a range applications in physics and astronomy (e.g., Komiske et al. 2019; Oladosu et al. 2020). In practice, we will modify the standard deep sets architecture in order to support the geometric setting that we operate under, i.e., a variable number of unaligned exposures. The metadata of the exposures (the seeing, airmass, etc.) form the basis of the data from the stacks that we use to predict the galaxy number counts.

We analyze four types of galaxies (luminous red galaxies, or LRGs; emission-line galaxies, or ELGs; quasars, or QSOs; and g-dropout Lyman-break galaxies, or GLBGs) extracted from DESI Legacy Surveys (DLS) data, and we illustrate the effect of the new technique on clustering measurements and cross-correlations with potential contaminants. This allows us to follow several previous works that have analyzed DLS galaxy samples, e.g., Kitanidis et al. (2020), Rezaie et al. (2020), Chaussidon et al. (2022), Zarrouk et al. (2021). However, we do not extract cosmological parameters from those measurements, since they require more thorough tests (as well as realistic mocks), which must be different for each galaxy sample. We also divide the sky into three areas, corresponding to the North, South, and DES regions of the DLS data, which have significantly different properties. Overall, this gives 12 different populations, each with varying degrees of contamination, offering a wide range of examples to evaluate the performance of the cleaning methods.

With a methodology that is able to work with exposure stacks, one can circumvent the limitations of lossy contaminant maps. One may also be able to perform less stringent sky or data quality cuts. This is particularly relevant for the arrival of data from the Legacy Survey of Space and Time (LSST) at the Vera C. Rubin Observatory, which will run over 10 yr and will have many more exposures than the DLS (The LSST Dark Energy Science Collaboration et al. 2018). While an increased number of exposures can decrease the impacts of outliers on image stacks, it cannot remove the spatial fluctuations altogether. And in spite of optimized observing strategies being developed to minimize their effects on cosmological inference (e.g., Awan et al. 2016; Almoubayyed et al. 2020), the techniques above will remain relevant at the level of precision of LSST.

This paper is structured as follows. In Section 2, we describe the data. Section 3 covers the models. We compare them in Section 4, and conclude in Section 5.

Finally, note that we use HEALPIX to subdivide the sky into pixels of resolution NSIDE = 512, each containing 16 subpixels (resolution NSIDE = 2048).

2. Data

This section describes the data sets that we use and the steps involved in preparing them for our comparison.

2.1. DLS

We use galaxy catalogs and image metadata from Data Release 9 (DR9) of the DLS (Dey et al. 2019). These data are collected with multiple instruments. In the North Galactic Cap, the Beijing–Arizona Sky Survey (BASS) covered 5100 deg^2 in the g and r bands, using the Bok telescope located at Kitt Peak in Arizona (Zou et al. 2017). This was complemented by z-band exposures from the The Mayall z-band Legacy Survey, covering the same area as BASS, with the Mayall telescope, which is also located at the Kitt Peak complex (Dey et al. 2016; Zhou et al. 2018). These two telescopes make up the North area of the DLS.

A further portion of the DR9 data comes from the Dark Energy Camera Legacy Survey (DECaLS), which covers the sky in all three bands—g, r, and z—using the Dark Energy Camera mounted on the Blanco Telescope at Cerro Tololo in Chile. The area covered by DECaLS covers almost 15,000 deg^2 of sky. However, the part of the DECaLS catalog that was used to generate the DES features more exposures per pixel than the non-DES part of the DECaLS catalog. Hence, the DECaLS catalog is further split into a DES part (4600 deg^2) and a non-DES part (9900 deg^2), with the latter referred to as South from here onward. This separation is made according to the total number of exposures (in all bands) covering each NSIDE = 2048 HEALPIX pixels, with DES having more than 60.

The DLS data are complemented (as part of DR9) with near-IR fluxes that have been extracted from the unWISE catalogs, based on data from the WISE satellite (Wright et al. 2010; Meisner et al. 2017).

2.2. Galaxy Catalogs

We generate galaxy catalogs from the public DLS DR9 brick data. One data brick covers a predefined area of the sky and lists all of the objects detected in it, from the three (g, r, and z) deep images made by coadded single-epoch images in this region. For each detected object, a host of metadata are available, including the location in R.A. and decl., fluxes, and Milky Way transmission, among other statistics. On average, every brick contains around 9000 objects, and there are more

[3] http://akosiorek.github.io/ml/2020/08/12/machine_learning_of_sets.html

than 350,000 unique bricks in DR9. We process every brick with the DESITARGET code.[4] The catalog-making step has already included a sophisticated masking of bad pixels or objects near bright stars or extended galaxies, which has been refined over the years (Kitanidis et al. 2020; Rezaie et al. 2020; Chaussidon et al. 2022; Zarrouk et al. 2021).

We consider four samples of galaxies: LRGs, ELGs, QSOs, and GLBGs. The color cuts for the first three galaxy types can be found in the DESITARGET pipeline, as well as in Kitanidis et al. (2020), Rezaie et al. (2020), and Chaussidon et al. (2022), for example. We only needed to define the cuts for the GLBGs (which are analogous to those for QSOs): signal-to-noise ratios smaller than 3 in the r band (i.e., a nondetection), greater than 4 in the r, z, and W1 bands, and greater than 3 or 2 in the W2 band (in the South or North, respectively). Those are the only detection cuts, and they do not include the color cuts that are typically added to remove contaminants from the Lyman-break selection (Hildebrandt et al. 2009; Harikane et al. 2017, 2022). The same quality mask as for the ELGs ("notinELG") is applied.

We generate number-count sky maps at HEALPIX resolution NSIDE = 512. Higher resolutions are more challenging, due to the larger numbers of pixels, the smaller numbers of galaxies in each pixel, and the associated Poisson noise.

2.3. Image Properties (Exposure-level Contaminants)

The first type of potential contaminant relates to the properties of the exposures that were processed as part of the DLS data (to make the coadded images covering the bricks) and from which the galaxy catalogs are constructed. We extract this information from the seven million annotated exposures of DLS DR9. The metadata of the exposures form the basis of the data from the exposure stacks that we use to predict the galaxy number counts.

Geometrically speaking, exposures cover the sky in a complicated manner; as a result, any sky pointing is described by a stack of exposures (which are different in the three g, r, and z bands). In order to resolve and store this information, we identify which exposures cover (the centers of)[5] HEALPIX NSIDE = 2048 pixels, and store the metadata in a database-like format, since the variable lengths of the stacks make it impossible to store this in rectangular arrays, without resorting to zero-padding.

For each exposure, we extract the following properties: the seeing, airmass, sky surface brightness, and sky counts (the airmass is averaged over bands, as in Kitanidis et al. 2020). These are readily available in the publicly released DLS tables, so no image processing is required. They are also likely to be effective summaries of the images. However, other choices would be possible (including a direct processing of the images). The diversity of the image stacks is illustrated in Figure 1. We further compress them into maps, by averaging the values in the stacks for each pixel. This is the information that is employed by conventional systematics mitigation techniques. We also calculate the standard deviation, minimum, and maximum of each vector, which we will use for additional null tests.

[4] https://github.com/desihub/desitarget

[5] Focusing on the centers of HEALPIX pixels, rather than resolving the full geometry of the problem, is an approximation that is sufficient for this work.

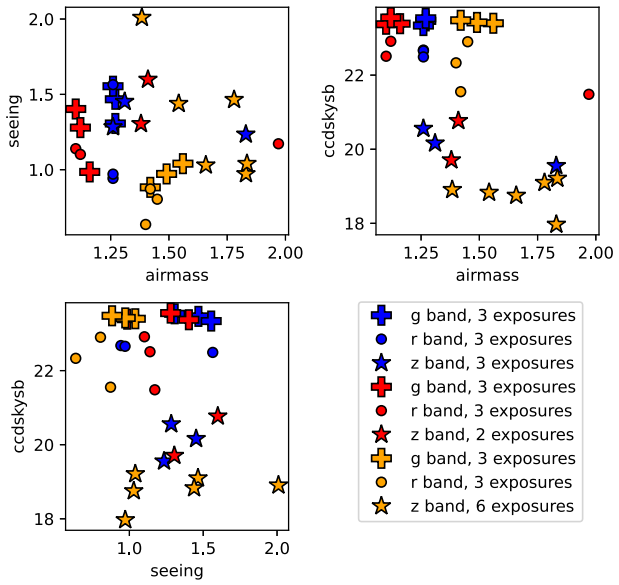

Figure 1. Illustrations of the diversity of the properties for single-epoch exposures, as well as their variable lengths. Subsets of properties are shown for the exposures covering three NSIDE = 2048 subpixels (IDs 21975319, 27803830, and 791280). We develop a method to predict the contamination of galaxy catalogs from the permutation-invariant and variable-length stacks of the exposure properties, rather than from the summary statistics (e.g., mean over exposures) that are employed by conventional methods.

2.4. External Maps of Contaminants (Sky-level Systematics)

Some contaminations do not originate from the observations (exposures) themselves, but rather from Galactic foregrounds, and therefore depend on the sky location of a given pixel. We call these sky-level contaminants, and store them as HEALPIX maps at NSIDE = 512.

We include a map of the galactic dust extinction, extracted using the galactic dust map SFD98 (Schlegel et al. 1998; Green 2018). We also include maps that trace the density of the stars on the sky, as these can be mistaken as galaxies and enter our catalogs as spurious objects, or affect the density or properties of the detected galaxies (e.g., the tails of bright stars). Two maps ("GAIA12" and "GAIA") are based on Gaia data (Gaia Collaboration et al. 2018): the density of the point sources in the magnitude ranges between PHOT_G_MEAN_MAG < 12 and $12 <$ PHOT_G_MEAN_MAG < 17. We also follow the procedure of Chaussidon et al. (2022) and make a map of the Sagittarius stream from the catalogs of Antoja et al. (2020). Finally, we make an additional stellar map that is generated by counting point-source objects from the bricks, following Kitanidis et al. (2020). The outliers were cleaned by removing all objects that were more than three standard deviations away from the stellar locus, calculated as the median color as a function of magnitude. Additionally, we include a map of the neutral hydrogen column density ("HINH"), from Bekhti et al. (2016).

More discussion of the origins of each contaminant (both at the exposure level and the sky level), and how they can affect catalogs, can be found in Kitanidis et al. (2020) and Chaussidon et al. (2022).

Table 1 summarizes all the potential contaminants that are included in this work.

Table 1
Systematic Contaminants and the Level at Which They Impact Galaxy Density

Contaminant	Level	X
Stellar	Pixel	Multiband
$E(B-V)$	Pixel	Multiband
HINH	Pixel	Multiband
GAIA	Pixel	Multiband
GAIA12	Pixel	Multiband
Sagittarius	Pixel	Multiband
Airmass	Exposure	Multiband
Seeing	Exposure	g, r, z
Sky surface brightness	Exposure	g, r, z
CCD sky counts	Exposure	g, r, z

2.5. Sky Cuts

We create DR9 coverage masks (at NSIDE = 512 and 2048) from the random catalogs that are available.[6] Pixels with insufficient coverage (<95%) are removed. Furthermore, all objects that are situated in the Large Magellanic Cloud between R.A. $\in [52°, 100°]$ and decl. $\in [-70°, -50°]$ are cut, since this area is usually too heavily contaminated by stars to draw reliable inferences of galaxy distribution.

We then split the remaining sky into the three different areas (North, South, and DES), using the number exposures (DES having more than 60 exposures at NSIDE = 2048) and cuts on the R.A. and decl.

We subsequently apply further sky cuts, based on the maps of the previous sections, in order to reject the most extreme values. These can be caused by processing issues, or simply by the most extreme image qualities. This also has the benefit of confining the values that are taken by each map to a reasonable range, suitable for machine-learning algorithms once they are mapped to [0, 1]. We cut pixels with values outside the median plus and minus ten times the median absolute deviation.

The final three regions are shown in Figure 2, and the average numbers of objects are given in Table 2. The variations between the regions indicate the differences in the selection cuts, average depths, or spatial fluctuations.

3. Methods

3.1. Map-based Linear Correction Model

If the contaminants are summarized in HEALPIX pixels (i.e., compressed into maps), which is the standard approach, then modeling the spurious fluctuations amounts to running a simple regression to predict the galaxy number counts from the values of the potential contaminants (in pixels). A linear model is the simplest model that one can employ; this has been widely developed in the literature (e.g., Ross et al. 2011; Elvin-Poole et al. 2018; Kitanidis et al. 2020; Wagoner et al. 2021) and successfully applied to recent surveys. An extensive review is provided in Weaverdyck & Huterer (2021), where connections between different modeling choices and training strategies are also provided. Given that this model is not backed by physics, there is freedom in the choice of scaling for the inputs and outputs, as well as to optimize the loss function. We perform the fit using Python's SCIKIT-LEARN package (Pedregosa et al. 2011). We find that the ridge and lasso regressions do not provide any improvements over the ordinary least squares linear regression, which we thus adopt. Min–max scaling is applied to scale all inputs into [0, 1], to facilitate model training. We do not scale the output: even though it may be slightly more challenging to fit the galaxy number counts, not scaling the output allows us to more easily compare metrics between the models and sky areas.

[6] https://www.legacysurvey.org/dr9/files/

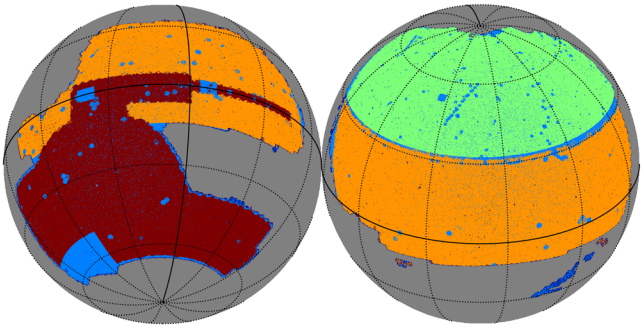

Figure 2. The sky coverage of the three regions used in this work (distinguished by color). Additional sky cuts (indicated in light blue) are performed according to the coverage of the survey as well as the statistics of the various potential contaminants (e.g., pixels with extreme values were cut).

Table 2
Average (Over Each Masked Area) Number of Objects per Surface Area

	North	South	DES
Sky Fraction [%]	0.115	0.211	0.108
N_{LRG}/arcmin^2	610.0	610.0	610.0
N_{ELG}/arcmin^2	2363.9	2440.1	2363.9
N_{QSO}/arcmin^2	305.0	305.0	228.8
N_{GLBG}/arcmin^2	2211.4	1525.1	2135.1

Note. The differences are due to the variations in the depth or observing conditions, or to the color cuts between the North and South regions, respectively.

3.2. Map-based Nonlinear Correction Model

Opting for a nonlinear function of the potential contaminants can provide significant improvements over the linear model, and is routinely employed. Common choices include random forests and neural networks (Rezaie et al. 2020; Chaussidon et al. 2022; Zarrouk et al. 2021). We adopt the latter. This will serve as a baseline to compare the performance of the deep sets. Min–max scaling is also applied to input contaminants, with no scaling to the output predictions.

3.3. Exposure-based Deep Sets Nonlinear Correction Model

The previous methods crucially rely on compressing the contaminants of all the images of a pixel image stack into a 1D feature vector (here, the mean seeing, airmass, CCD counts, and background noise in all three bands, with the airmass averaged over the bands). We now explore whether one can model the contamination directly from the properties of the CCDs that cover a given pixel. This would remove the lossy step of compressing the exposure properties into maps, and potentially extract more information from the full stacks of exposures. Two complications arise: the sets of exposure vectors vary in size, and they have no inherent ordering.

The problem of learning a function on unordered variable-sized sets is not new. Examples from other fields include 3D

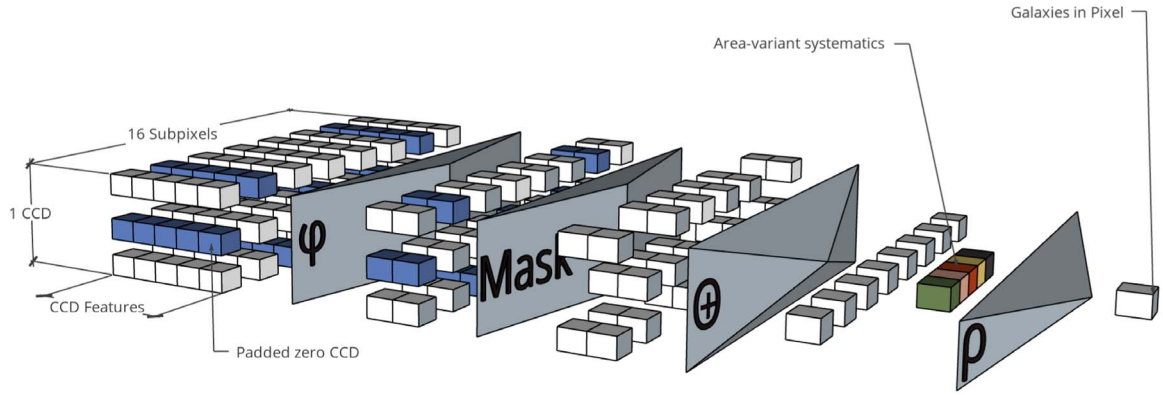

Figure 3. Schematic of the deep sets–inspired architecture employed in this work to support arbitrary image stacks. The CCD features are the summary properties of each CCD (10 in total). The depth corresponds to the 16 NSIDE = 2048 subpixels (some of which are removed by the mask, if they do not fall within the valid region) contributing to each NSIDE = 512 pixel, which are to be aggregated by a permutation-invariant operator (e.g., the sum). The height corresponds to one CCD, which may not cover all the subpixels. Sky-level systematics (the extinction, stellar counts, etc., as given in Table 7) are appended, before a final prediction is calculated for an NSIDE = 512 pixel. φ and ρ are parameterized with neural networks.

point clouds (Qi et al. 2016) or population-level statistics (Zaheer et al. 2017). Earlier attempts to learn functions on sets include the Neural Statistician of Edwards & Storkey (2016). Zaheer et al. (2017) provided the first proof of concept for an architecture that was specifically designed to satisfy the criteria outlined above. Since then, the theory and applications of permutation-invariant layers have blossomed (Ravanbakhsh et al. 2016; Korshunova et al. 2018; Murphy et al. 2018; Soelch et al. 2019; Wagstaff et al. 2019; Bronstein et al. 2021; Cranmer et al. 2021), but deep sets remains the simplest approach. The core idea behind deep sets architecture is the inclusion of layers performing permutation-invariant operations on the inputs, i.e., yielding the same result, regardless of the ordering of the sets. The simplest and most widely used choices for such functions include aggregating (sum) or averaging (mean) over the input elements within a set, or taking the minimum or maximum values. A summary schematic of our final deep sets architecture is shown in Figure 3. We now go through the arguments that led to its construction.

Technically, a deep sets architecture consists of three distinct blocks. First of all, there is a block that applies transformations to all the individual elements in the input set. In the present context, this block will be referred to as a *feature extractor*, and it is simply a fully connected feed-forward neural network. Afterward, the learned representations of every element in the input set are passed into the permutation-invariant layer, a so-called *aggregator*. This layer will learn the weights, to optimally aggregate the function representations that were transformed in the feature extractor. The aggregator compresses the input set to a single-dimensional feature vector of fixed size. This feature vector is then passed to the third block, where it is reduced to a final continuous output, to predict the number of galaxies in the pixel. This last block is another fully connected feed-forward neural network; however, it has a single output neuron and linear activation in the final layer. The final block will be referred to as MLP.

The 16 features of the previous sections are now replaced with five exposure-level features (seeing, sky count, airmass, and surface brightness, as well as a categorical integer variable for the band) and the six other sky-level potential contaminant maps that do not vary between exposures (extinction, stellar maps, etc.). As such, these features are only passed to the MLP block of the network after the exposure-variant systematics are passed through the feature extractor and the aggregator. This is the first difference from a classic deep sets architecture. Furthermore, the task of learning the contamination in the galaxy number counts presents additional challenges that prevent the use of the conventional deep sets architectures of Lee et al. (2019), Zaheer et al. (2017), Soelch et al. (2019), and Qi et al. (2016).

Variable set sizes. First of all, the chosen architecture needs to account for variable-sized inputs. Few implementations in the literature have adapted their networks to the processing of variable-sized inputs, and for those that did, the variability of the input set sizes was handled by zero-padding to a fixed input set size. However, the feature extraction network includes a bias term, meaning that the padded zero vectors are problematically transformed to nonzero vectors. This would impact the outcome of the aggregation block that follows the feature extractor. Consequently, we need to include an additional *masking* operation that tells the aggregator which elements of the input set to aggregate and which ones to ignore for the aggregation. This mask is generated for every input set independently, and passed to the aggregator block during a forward pass through the network.

While zero-padding followed by the masking of the zero vectors usually solves the problem of variable set sizes, a further problem then arises. The generated data sets feature very large variations in the number of exposures that are associated with a given pixel. That is, the set sizes that are fed into the network differ substantially across different pixels. To ensure that a given pixel meets a minimum quality threshold, only pixels that are covered by a minimum of two exposures per band are kept. Since there are three bands (g, r, and z), some pixels on the sky are only covered by six exposures. However, other pixels are covered much more thoroughly, especially in the DES area, with tens of exposures per band per pixel. Using large maximum set sizes substantially increases the training time, since the feature extraction layer is applied to every exposure vector, regardless of whether it represents an actual exposure or whether it was added during zero-padding.

Input dimensionality. To overcome the problem of prohibitively large zero-padded vectors, we employ a subpixelization when dividing each pixel into 16 NSIDE = 2048 subpixels, which are covered by fewer CCDs than the original NSIDE = 512 pixels. This reduces the maximum necessary set

size that needs to be supported by the architecture. Furthermore, fewer zero vectors will be processed in the feature extractor, thus speeding up the model training. However, we still want the output galaxy number counts to be calculated at the resolution NSIDE = 512. As a result, the model includes tensors with another input dimension (subpixels), and a final layer averaging them into pixels, diverging farther from the standard deep sets architecture. The maximum set size is set such that fewer than 0.1% of all subpixels in the footprint of this area are covered by more exposures. For the one in 1000 subpixels that feature more subpixels, a random subset of exposures is drawn. The maximum set sizes are 30, 25, and 40, for the North, South, and DES regions, respectively.

The values for the input CCDs for the deep sets are scaled using a robust scaler utilizing the interquartile range, rather than the minimum and maximum. As before, the output galaxy number counts per pixel are not scaled.

3.4. Hyperparameter Optimization

Hyperparameters can drastically affect the ability of a model to generalize to unseen data. K-fold cross-validation has previously been used on the nonlinear model (e.g., Rezaie et al. 2020). While useful in scenarios with limited or highly imbalanced data, a major drawback of k-fold cross-validation is that all models have to be trained k times. We found that k-fold cross-validation was not needed for the data considered here, as the random assignment of pixels to training and validation folds still generates samples that are representative of one another.

The fully connected feed-forward neural network and the deep sets are built in the PyTorch Python framework (Paszke et al. 2019). The software package Optuna (Akiba et al. 2019) is used to randomly draw the values of the hyperparameters in their respective ranges of interest, run the optimization, and automatically stop poor runs ("pruning") and start new ones in other regions of the prior. This allows us to explore more effectively the high-dimensional space of the hyperparameters.

We now enumerate the various hyperparameters that are relevant here, and what ranges we explore. We first consider the parameters that are relevant to the neural network model.

We use a Poisson loss function (Rezaie et al. 2020), which we optimize using the Adam optimizer (Kingma & Ba 2017). The learning rates for our models are sampled uniformly in $[10^{-5}, 10^{-2}]$. Additionally, we employ a weight decay regularization to improve robustness. This is achieved by adding a function of the weights to the loss function in the model, which enforces weight sparsity and prevents the weights from becoming large. This is controlled by a hyperparameter λ that is passed to the optimizer (Kingma & Ba 2017), sampled uniformly in [0, 0.3]. The multilayer perceptrons trained here are allowed to have between one and six layers. The hidden layers can take on a minimum of eight neurons per layer (approximately half the feature space) all the way up to 256 neurons per layer, to allow for wider networks as well. Those integer parameters are sampled uniformly. We consider the following values for the batch size: 16, 32, 128, or 256. To improve robustness, feed-forward neural network models can be trained by randomly dropping a given percentage of neurons during training (Hinton et al. 2012). The dropout percentage parameter was randomly sampled uniformly in [0, 0.5], meaning that in each dropout layer anywhere between zero to half the neurons are randomly deactivated during training.

For the deep sets, there are additional hyperparameters. First of all, a deep set consists of two fully connected feed-forward neural networks that need to be configured: the feature extractor, at the beginning, and the MLP network, at the end of the model. As such, parameters such as the number of layers, the neurons per layer, and the percentage dropout have to be drawn and ultimately set for each network independently, using the same ranges as above. Furthermore, the deep sets architecture can use different permutation-invariant functions in the aggregator block of the network. The choices here are limited to aggregation by using the maximum value of a given set, aggregation to the mean, or summing the values passed from the feature extractor. Last, the potential depth of a deep sets architecture can lead to a network getting stuck on local minima during optimization. As such, several different initializations are available to the network parameters. The choices here are Xavier Glorot, Kaiming He, and following a uniform or a normal distribution.

3.5. Maps and Power Spectra

To evaluate the performance of the cleaning methods, we will look at maps and angular power spectra. When needed, number-count maps can be converted to overdensity maps by dividing by the average number density (and subtracting one). This must be done in each region separately. These overdensity maps are what we can calculate the angular spectra of, for null tests or comparisons with cosmological models (which we do not do here). Angular power spectra are calculated using NaMaster (Alonso et al. 2019) applied to the NSIDE = 512 maps, with spherical harmonic modes up to $\ell_{\rm max} = 1024$, again in each region separately (even though the sky maps plotted in this paper show them together). The individual multipoles ℓ are binned into bands of size 11. All the power spectra and covariances are calculated for this band size, as implemented in NaMaster. When covariances on the measured (auto- or cross-) power spectra are needed, we use a Gaussian approximation (Alonso et al. 2019). However, the latter requires a model for the underlying true power spectra. We simply feed the measured power spectra (the sum of the measured spherical harmonic coefficients, before deconvolving the effect of the mask) divided by the fraction of sky covered. This is a rough estimate of the underlying power spectrum, which delivers positive and smoother power spectra than those resulting from the full deconvolution of the mask. It is sufficient for the purpose of obtaining covariances that, in our case, are only used (1) to divide the power spectra by their typical uncertainties, to illustrate fractional chances, and (2) to calculate chi-squared statistics when performing null tests with cross-power spectra. In particular, the same covariance is fed to the chi-squared statistics of the three methods, so that they can be compared, and the accuracy of the covariance only affects the absolute values of the chi-squared statistics, which are less critical in our case, since we are concerned with the relative values before and after cleaning.

4. Results

4.1. Results of Each Fit

We randomly split the data in each region into training (60%), validation (20%), and testing (20%) sets of pixels (thus

Table 3
Comparison of the Results of the Fits from the Various Models (with the Best Model in Each Row Highlighted in Bold)

Area	Galaxy Type	Linear	Neural Net	Deep Sets
North	LRG	0.01	0.008	**0.013**
	ELG	0.08	0.139	**0.144**
	QSO	0.104	0.112	**0.127**
	GLBG	0.109	0.21	**0.261**
South	LRG	0.006	0.006	0.006
	ELG	0.061	0.108	**0.129**
	QSO	0.091	0.095	**0.1**
	GLBG	0.046	0.167	**0.189**
DES	LRG	0.011	0.02	**0.026**
	ELG	0.027	**0.048**	**0.048**
	QSO	**0.035**	0.033	0.03
	GLBG	0.075	0.105	**0.134**

Note. The metric is the coefficient of determination R^2 calculated over the full area (thus including the training, validation, and testing data). For each model, a hyperparameter search was performed.

they are not necessarily adjacent). We train each model on the first set, and use the second for stopping, when the model performance no longer improves. For the nonlinear and deep sets models, this is coupled with the exploration of hyperparameters, performed with OPTUNA. We adopt the best models, as measured by the coefficient of determination R^2 calculated on the test data set. Finally, we apply these models to the full data set (reassembling the training, validation, and testing sets in each region), which we will use below. The corresponding R^2 are shown in Table 3. We do not discuss the hyperparameters in detail here,[7] because they are not directly interpretable, so their values (in absolute terms, or relative to the other regions or models) do not provide additional insight.

We see in Table 3 that except for cases with very low levels of contamination (i.e., LRGs), the deep sets model outperforms the linear and neural network models. This is not unexpected, given that it has many more degrees of freedom, and thus should be able to extract any contamination in the data more effectively.

4.2. Original Maps, Corrected Maps, and Angular Power Spectra

Figure 4 shows the original uncorrected galaxy number counts. It can be seen that: (1) except for LRGs, high levels of spurious spatial variations are present; and (2) the depths and spatial variations of some of the samples can be significantly different between the three regions. This follows previous findings (Kitanidis et al. 2020; Chaussidon et al. 2022; Zarrouk et al. 2021): LRGs have little contamination; ELGs are much more highly contaminated, due to their selection near the depth limit; and QSOs also exhibit high contamination, following the stellar density, since stars and QSOs are easily confused.

Figure 5 shows the maps as corrected by the deep sets method. Very few contamination-like patterns can be observed. They all show comparable levels of quasi-uniform and isotropic fluctuations. Figure 6 shows the correlations extracted by the neural network technique, and Figure 7 shows the additional correlations found by the deep sets technique. This allows us to examine more closely which patterns in the number counts have been successfully explained by the regression as a function of the potential contaminants.

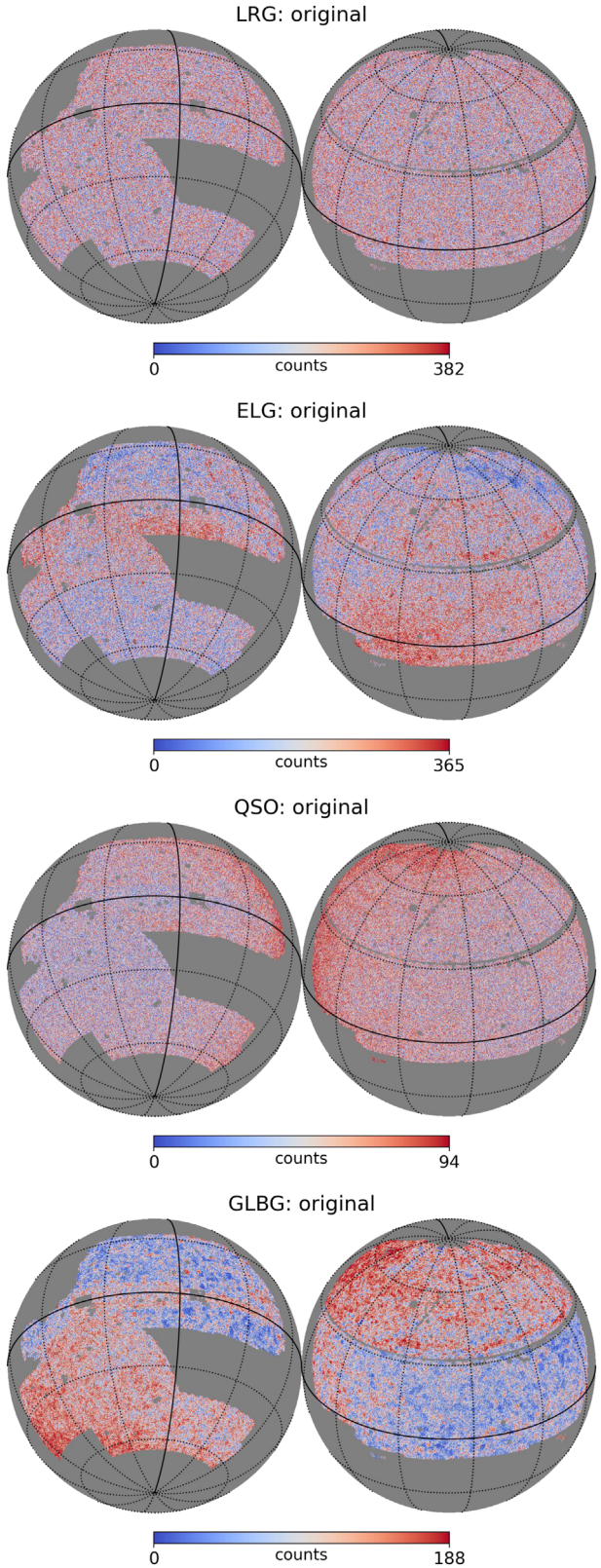

Figure 4. Original uncorrected galaxy catalogs (number counts in HEALPIX pixels NSIDE = 512) constructed from DLS DR9. Suspicious features on a variety of scales are present and most likely noncosmological.

[7] They can be found online at https://github.com/elleggert/astrostatistics/blob/master/models/deep_set/final_run.py.

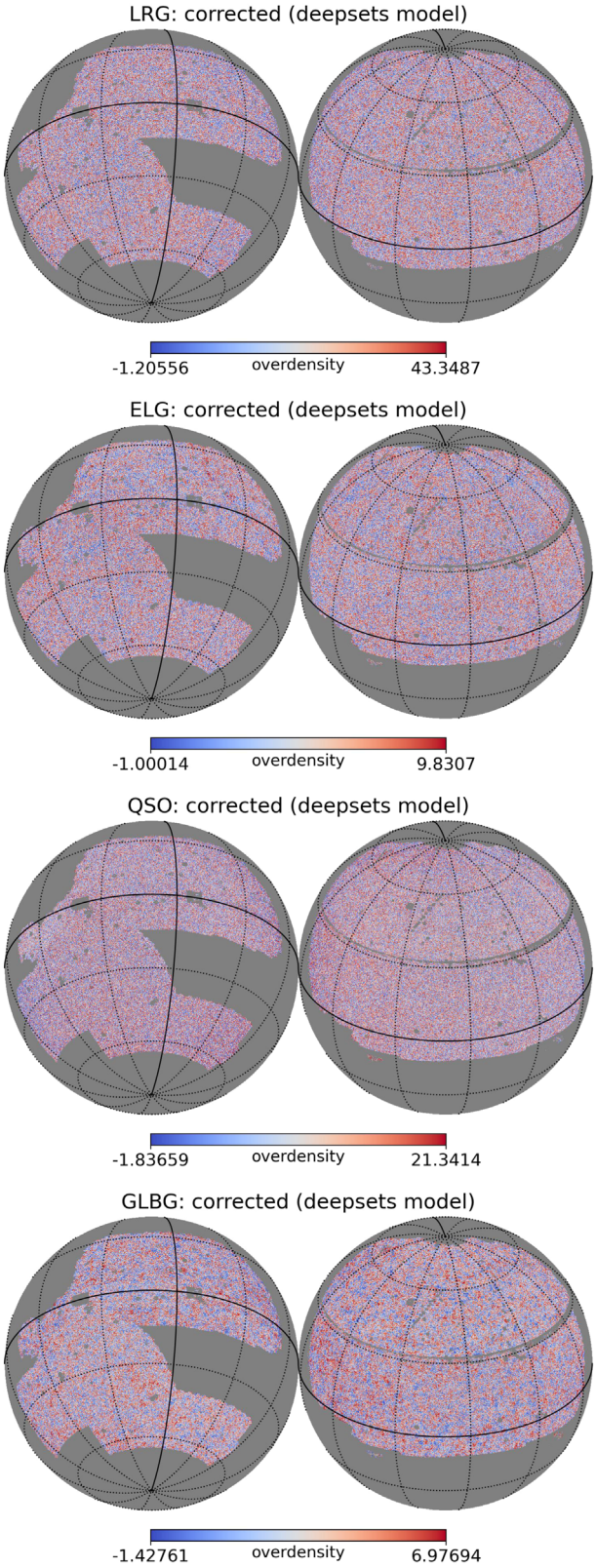

Figure 5. Number-count maps obtained after correcting for correlations with potential contaminants using the new deep sets–based technique. Residual fluctuations are nearly isotropic and uniform across the sky.

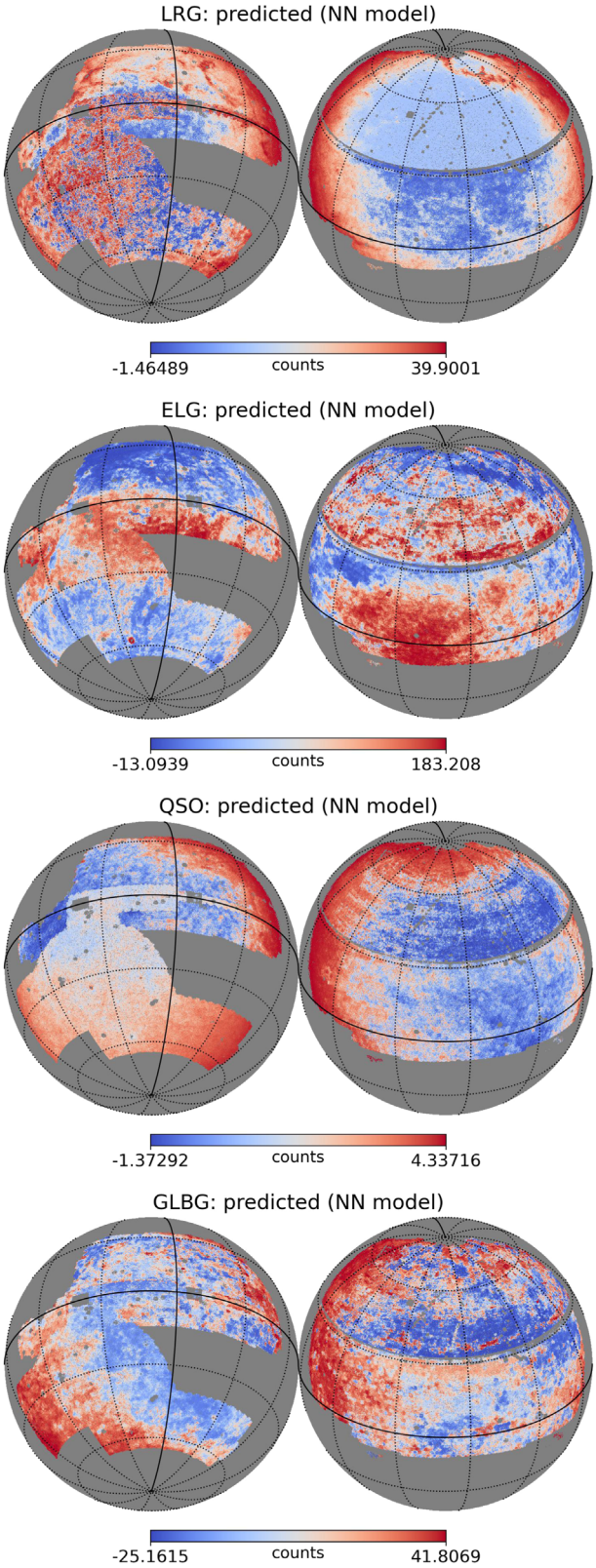

Figure 6. Correlations found by the neural network model between the original number-count maps of Figure 4 and the maps of potential contaminants (which only include the mean seeing, airmass, and CCD noise over the exposure stacks).

In order to gain further insight, we examine the angular power spectra (band powers), since they are a summary statistic from which cosmological information is typically extracted, with harmonic modes (the spherical coordinates equivalent to Fourier modes) helping to separate the effects of different physical scales in terms of signals on the sphere.

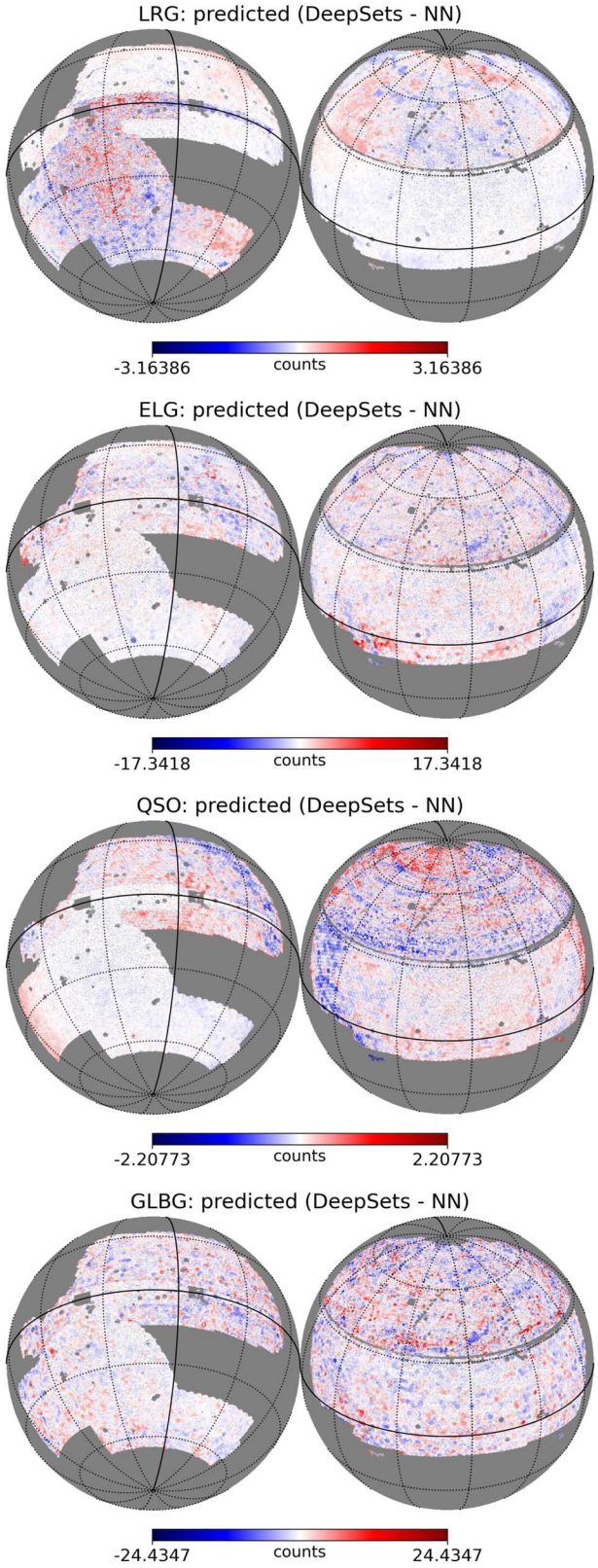

Figure 7. Additional spurious correlations found (on top of those shown in Figure 6) in the number-count maps by the new deep sets model from the full-exposure stacks.

Figure 8 shows the angular power spectra of the original and corrected galaxy catalogs (with number counts converted to overdensity maps). The subpanels show the power subtracted by each method, normalized by the Gaussian (diagonal) error. We see that:

1. Overall, power is removed by all three methods. This is consistent with the expectation that the intrinsic cosmological signal is isotropic and close to uniform, with little power on large scales, and that any contaminant will add large amounts of excess power.
2. The more complex and flexible the model, the more power is identified and removed, as one would expect in the presence of significant contamination. This accords with the results of Table 3.

4.3. Chance Correlations

Chance correlations are random alignments between the cosmological signal and the possible contaminants. They are a potential issue for all the methods discussed here. For example, the power spectrum (additive) bias that results from fitting the linear model to n contaminant maps is $b_\ell \sim -n/(f_{\rm sky}^2(2\ell+1))$, as derived in Elsner et al. (2015). This bias increases with the numbers of degrees of freedom of the model and the inputs, and it is therefore expected to be larger for the nonlinear and deep sets models. However, the chance correlation bias resulting from fitting each model mostly depends on the inputs (the contaminant maps for the linear and neural network models, and the stacks for the deep sets) and the statistical properties of the sample under consideration, mostly its power spectrum. Given that the power spectra of the galaxy samples are similar, and that the samples themselves exhibit some correlation due to their overlap in redshift, we can in fact obtain a rough limit for the chance correlation bias in LRG power spectra. Indeed, these are the cleanest data, by construction, and thus the bias is likely to be most visible in this case. In all other cases, the subtracted power is so much larger that it is not likely to be due to chance correlations, but rather to contamination. However, simulations would be needed to confirm this upper limit on the bias.

4.4. Null Tests of Cross-power Spectra

In order to gain more intuition about the sources of power removed by each technique, we calculate cross-power spectra between the contaminant maps and the number density maps. However, the original contaminant maps used in the neural network technique are, by construction, lossy compressions of the exposure stacks used in the deep sets technique. Cross-correlations with these maps (e.g., the mean seeing) will reveal if the deep sets fits are indeed able to remove more contamination from those sources, as expected. It would also be interesting to measure by how much it beats the neural network technique for information that the latter does not have access to. For this purpose, we build another set of contaminant maps, with summary statistics (the minimum, maximum, and standard deviation) of the exposure stacks. We also add a map of the numbers of exposures (these are not fed directly into any of the methods). While they could in theory be fed into the linear and neural network methods, it is unclear whether they are good summary statistics, plus the point of the deep sets is to directly capture all the information from the stacks and to circumvent the need for picking summary statistics.

We compute cross-angular (band) power spectra $\{\mathcal{C}_b\}_{b=1,\cdots,B}$ between the galaxy overdensity maps and the new extended set of contaminant maps. We then calculate covariance matrices

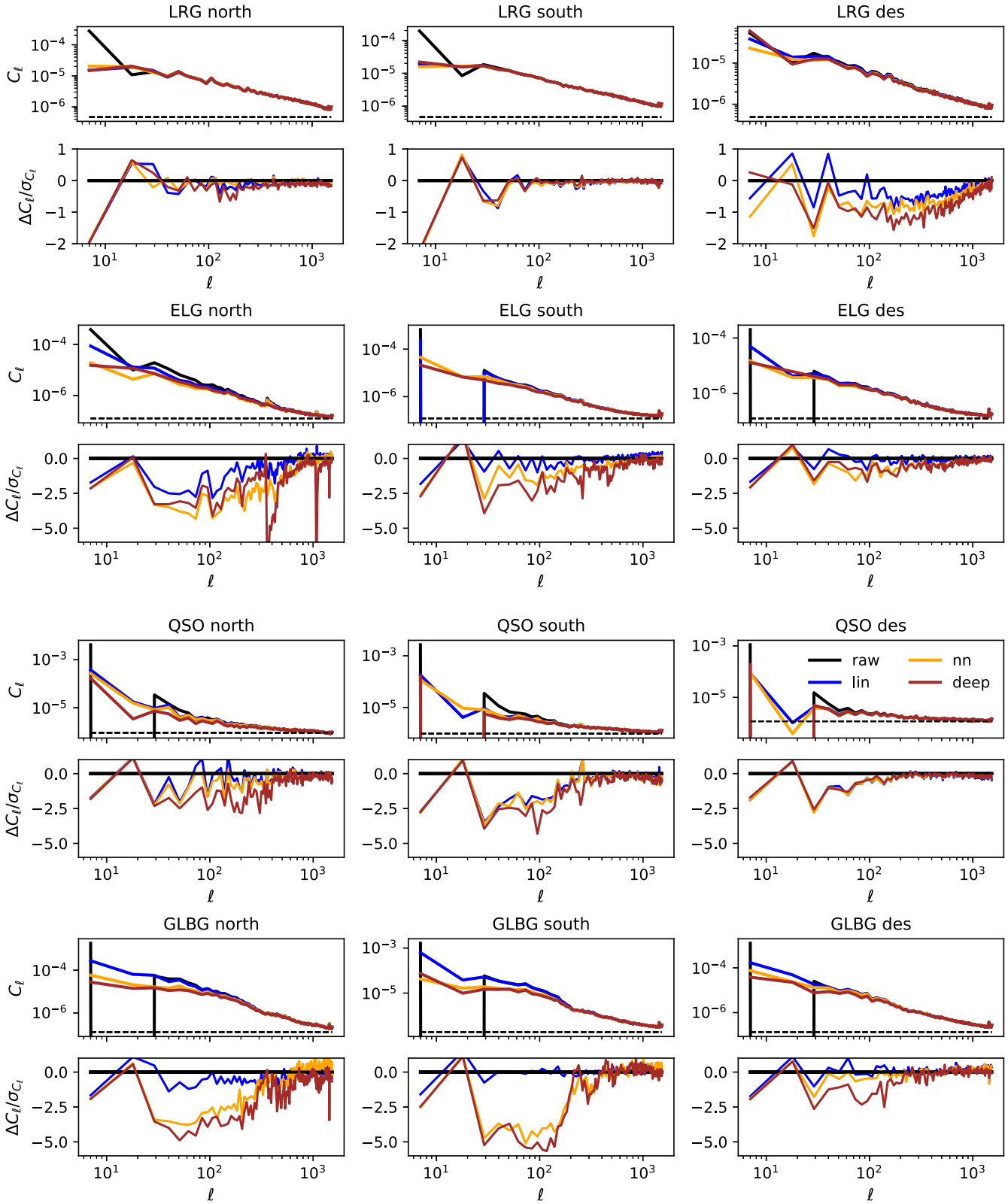

Figure 8. Comparisons of the angular power spectra of the uncorrected and corrected galaxy overdensity maps obtained with the different cleaning methods. The bottom panels show the fractional changes, i.e the corrected spectra minus the uncorrected one, divided by the Gaussian uncertainties defined in Section 3.5. While the changes on large scales are the most pronounced, the cleaning results in a reduction of power on almost all scales, as well as the removal of sharp contamination features (e.g., $\ell \sim 300$).

$\{C_{bb'}\}$, using the procedure detailed in Section 3.5. We compute chi-squared values $\chi^2 = \sum_{b=1}^{B}\sum_{b'=1}^{B} C_{bb'}^{-1}\mathcal{C}_b\mathcal{C}_{b'}$ that correspond to the hypothesis that the power spectra are consistent with zero, i.e., there is no detectable contamination signal. We reduce the chi-squared values by dividing them by the number of harmonic multipole bands.

Figure 9 shows the resulting reduced chi-squared values. The vertical lines indicate the value of 1, corresponding to the hypothesis (zero-power spectrum) for the limit of no contamination. Even though this calculation relies on the assumption that the fields are Gaussian, and also on fiducial theoretical power spectra, this figure can be used to obtain a rough quantification of which sources of contamination are the most significant in each sample, and how well each method performs.

Across the board, the deep sets method removes more correlations than the other methods. We should consider the contaminants that were accessible to the neural network and the linear models, i.e., the mean maps. In this case, the performance of the neural network and the deep sets is generally comparable. However, there are a few cases in which the deep sets method performs better.

When it comes to the contaminant maps that were not given to the linear or neural network models (standard deviation, minimum, and maximum), the deep sets method substantially reduces the measured residual correlation. This is particularly dramatic for some of the samples, giving an indication of the possible sources of contamination. In cases where significant correlations are present in the raw data, however, the improvements provided by the deep sets method are significant, but do not systematically bring the reduced chi-squared values close to unity.

There is significant residual contamination in a range of cases, which the deep sets method should have been able to extract, in theory. However, this could be because the models were only trained on a fraction of the data/sky. Training the model on the full data set may resolve this. But it could result in increased overfitting and the removal of cosmological power. Thus this may require a more in-depth study of chance correlations and overfitting. In order to reduce this bias, we have taken the approach of fitting the models on a fraction of the data, split into training and testing sets. Other recent works have shown that this is unbiased in the cases that were studied (Kitanidis et al. 2020; Rezaie et al. 2020; Zarrouk et al. 2021). This approach warrants a more robust fitting and optimization of our hyperparameters, and a fairer comparison between the models. However, it would also mean that correlations with the potential contaminants would not necessarily be captured at their full extent, which may then leave some residual systematic biases. Another way of improving the robustness of the correction technique, and to prevent overfitting and chance correlations, is to add terms to the loss function. For example, one could account for the expected covariances (between pixels) due to cosmological fluctuations (Wagoner et al. 2021). We defer these points to future work.

5. Discussion and Conclusion

Modern galaxy surveys are plagued by observational systematics, which bias cosmological analyses if they are not mitigated. One major source of such systematic bias is the spurious modulation of galaxy properties (e.g., number counts), due to varying observing conditions. While this is now routinely treated with linear and nonlinear models that are based on contaminant template maps, such an approach neglects the multiepoch nature of surveys. This work has been motivated by the fact that this has not been studied for wide-field surveys. Extracting this information could lead to cleaner galaxy catalogs and more accurate cosmological analyses.

We have presented a new method that can tap into the information in image stacks. To predict the galaxy number counts, we have restricted our attention to the metadata of the exposures (the seeing, airmass, etc.) in the exposure stacks. However, learning of the contamination from the raw images themselves should be possible, and has been left for future work.

The new method relies on deep sets architectures, modified to better suit pixelized galaxy number counts, variable-length image stacks, and additional contaminant sky maps. We have applied this method to four types of galaxies, extracted from catalogs and images from DLS. These exhibit different levels of contamination. The new model is capable of significantly reducing this contamination, to a greater extent than conventional linear or nonlinear models. Since we do not go all the way to cosmological parameters, we have not performed more exhaustive null tests with other metrics.

Another promising avenue for optimizing the analysis of galaxy surveys is to define sky cuts, achieving an optimal balance between the loss of information and the accuracy of the mitigation of the systematics. This could be done with the results presented here.

In DLS, there are a few (to tens) of single-epoch images per band, on average, contributing to each sky location. This number will dramatically increase for LSST. This may improve survey uniformity, since more images are contributing to each detection, and average survey properties are likely to offer better summary statistics. But this may also increase the likelihood of extreme outliers contributing. The results may therefore go either way in terms of the effective contamination on the sky. Thus, a dedicated study of deep sets models would be needed for LSST, to clarify if methods that work with image stacks directly are indeed necessary.

DESI will soon provide follow-up spectroscopy for a lot of the objects analyzed in this work. This will lead to secure measurements of galaxy type and redshift, and it could even fix erroneous photometry. However, it is known that spurious correlations in the target photometry can propagate into the spectroscopic catalogs (e.g., Ross et al. 2011; Rezaie et al. 2021), and therefore the techniques explored here are relevant to both photometric and spectroscopic analyses.

Finally, we note that the injection of additional sources (SSIs) into single-epoch or coadded images is a powerful technique for simulating contamination, and for validating techniques for mitigating systematics. It has been key in recent cosmological analyses, and it will play an increased role in the LSST era (Huang et al. 2017; Everett et al. 2022). However, because it is computationally expensive, it may only be possible to perform rigorous SSIs on small areas of the sky. Thus, machine-learning methods that are capable of exploiting complicated inputs (such as deep sets) and additionally leveraging the available SSIs will be key to revealing how the contamination entering cosmological analyses should be accurately modeled.

E.E. and B.L. thank the members of the cosmology and quasar groups at Imperial College for useful discussions and feedback during the course of this project.

B.L. is supported by the Royal Society through a University Research Fellowship.

The Legacy Surveys consist of three individual and complementary projects: the Dark Energy Camera Legacy

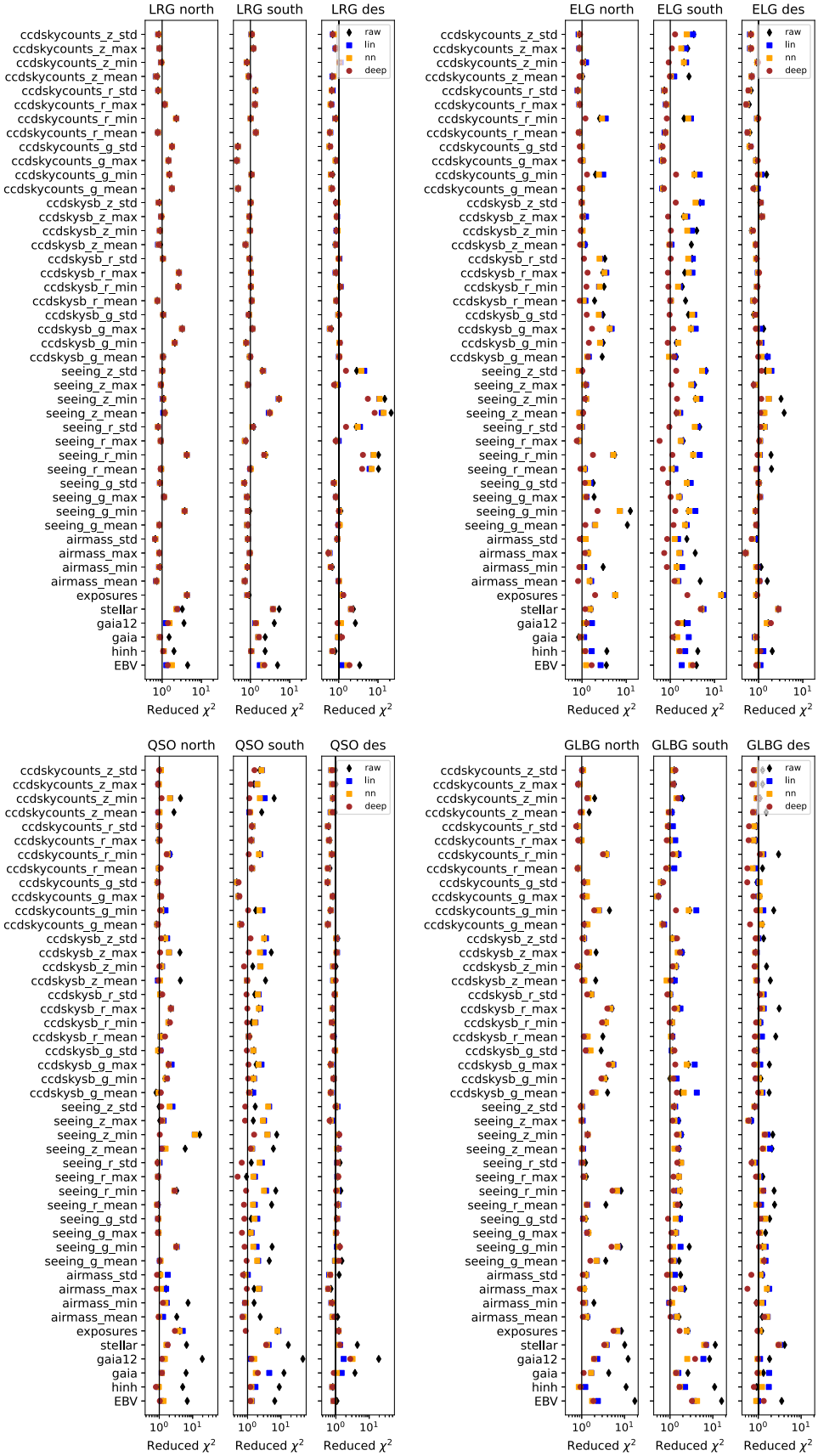

Figure 9. Residual contamination as measured in the cross-power spectra between the cleaned galaxy overdensity maps and the contaminant maps. The reduced chi-squared values are calculated by comparing the cross-spectra (up to $\ell_{\rm max} = 1024$) with the zero-power spectrum, as a hypothesis, using the procedure detailed in Section 3.5.

Survey (DECaLS; Proposal ID 2014B-0404; PIs: David Schlegel and Arjun Dey), the Beijing–Arizona Sky Survey (BASS; NOAO Prop. ID 2015A-0801; PIs: Zhou Xu and Xiaohui Fan), and the Mayall *z*-band Legacy Survey (MzLS; Prop. ID 2016A-0453; PI: Arjun Dey). DECaLS, BASS, and MzLS together include data obtained, respectively, at the Blanco telescope, Cerro Tololo Inter-American Observatory, NSF's NOIRLab; the Bok telescope, Steward Observatory, University of Arizona; and the Mayall telescope, Kitt Peak National Observatory, NOIRLab. The Legacy Surveys project is honored to be permitted to conduct astronomical research on Iolkam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation.

NOIRLab is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

This project has used data obtained with the Dark Energy Camera (DECam), which was constructed by the Dark Energy Survey (DES) collaboration. Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundacao Carlos Chagas Filho de Amparo, Financiadora de Estudos e Projetos, Fundacao Carlos Chagas Filho de Amparo a Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Cientifico e Tecnologico and the Ministerio da Ciencia, Tecnologia e Inovacao, the Deutsche Forschungsgemeinschaft, and the Collaborating Institutions in the Dark Energy Survey. The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energeticas, Medioambientales y Tecnologicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenossische Technische Hochschule (ETH) Zurich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciencies de l'Espai (IEEC/CSIC), the Institut de Fisica d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig Maximilians Universitat Munchen and the associated Excellence Cluster Universe, the University of Michigan, NSF's NOIRLab, the University of Nottingham, the Ohio State University, the University of Pennsylvania, the University of Portsmouth, the SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, and Texas A&M University.

BASS is a key project of the Telescope Access Program (TAP), which has been funded by the National Astronomical Observatories of China, the Chinese Academy of Sciences (the Strategic Priority Research Program "The Emergence of Cosmological Structures," grant No. XDB09000000), and the Special Fund for Astronomy from the Ministry of Finance. BASS is also supported by the External Cooperation Program of the Chinese Academy of Sciences (grant No. 114A11KYSB20160057) and the Chinese National Natural Science Foundation (grant No. 11433005).

The Legacy Survey team makes use of data products from the Near-Earth Object Wide-field Infrared Survey Explorer (NEOWISE), which is a project of the Jet Propulsion Laboratory/California Institute of Technology. NEOWISE is funded by the National Aeronautics and Space Administration.

The Legacy Surveys imaging of the DESI footprint is supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy, under Contract No. DE-AC02-05CH1123; by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility, under the same contract; and by the U.S. National Science Foundation, Division of Astronomical Sciences, under Contract No. AST-0950945 to NOAO.

ORCID iDs

Boris Leistedt https://orcid.org/0000-0002-3962-9274

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. 2019, arXiv:1907.10902
Almoubayyed, H., Mandelbaum, R., Awan, H., et al. 2020, MNRAS, 499, 1140
Alonso, D., Sanchez, J., & Slosar, A. 2019, MNRAS, 484, 4127
Antoja, T., Ramos, P., Mateu, C., et al. 2020, A&A, 635, L3
Awan, H., Gawiser, E., Kurczynski, P., et al. 2016, ApJ, 829, 50
Bekhti, N. B., Flöer, L., Keller, R., et al. 2016, A&A, 594, A116
Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. 2021, arXiv:2104.13478
Chaussidon, E., Yèche, C., Palanque-Delabrouille, N., et al. 2022, MNRAS, 509, 3904
Cranmer, M., Kreisch, C., Pisani, A., et al. 2021, ICLR 2021 SimDL Workshop, https://simdl.github.io/files/40.pdf
Dey, A., Rabinowitz, D., Karcher, A., et al. 2016, Proc. SPIE, 9908, 99082C
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168
Edwards, H., & Storkey, A. 2016, arXiv:arXiv:1606.02185
Elsner, F., Leistedt, B., & Peiris, H. V. 2015, MNRAS, 456, 2095
Elvin-Poole, J., Crocce, M., Ross, A., et al. 2018, PhRvD, 98, 042006
Everett, S., Yanny, B., Kuropatkin, N., et al. 2022, ApJS, 258, 15
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1
Gatti, M., Sheldon, E., Amon, A., et al. 2021, MNRAS, 504, 4312
M. Green, G. 2018, JOSS, 3, 695
Harikane, Y., Ouchi, M., Ono, Y., et al. 2017, PASJ, 70, S11
Harikane, Y., Ono, Y., Ouchi, M., et al. 2022, ApJS, 259, 20
Hildebrandt, H., Pielorz, J., Erben, T., et al. 2009, A&A, 498, 725
Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. 2012, arXiv:1207.0580
Huang, S., Leauthaud, A., Murata, R., et al. 2017, PASJ, 70, S6
Johnston, H., Wright, A. H., Joachimi, B., et al. 2021, A&A, 648, A98
Kingma, D. P., & Ba, J. 2017, arXiv:1412.6980
Kitanidis, E., White, M., Feng, Y., et al. 2020, MNRAS, 496, 2262
Komiske, P. T., Metodiev, E. M., & Thaler, J. 2019, JHEP, 2019, 121
Korshunova, I., Degrave, J., Huszár, F., et al. 2018, arXiv:1802.07535
Lee, J., Lee, Y., Kim, J., et al. 2019, Proc. Machine Learning Research 97, Proc. of the 36th Int. Conf. on Machine Learning, 3744, https://proceedings.mlr.press/v97/lee19d.html
Loureiro, A., Whittaker, L., Mancini, A. S., et al. 2022, A&A, 665, A56
Meisner, A. M., Lang, D., & Schlegel, D. J. 2017, AJ, 154, 161
Murphy, R. L., Srinivasan, B., Rao, V., & Ribeiro, B. 2018, arXiv:1811.01900
Oladosu, A., Xu, T., Ekfeldt, P., et al. 2020, arXiv:2007.04459
Paszke, A., Gross, S., Massa, F., et al. 2019, Advances in Neural Information Processing Systems 32, ed. H. Wallach et al. (Red Hook, NY: Curran Associates, Inc.), 8024, https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Qi, C. R., Su, H., Mo, K., & Guibas, L. J. 2016, arXiv:1612.00593
Ravanbakhsh, S., Schneider, J., & Poczos, B. 2016, arXiv:1611.04500
Rezaie, M., Seo, H.-J., Ross, A. J., & Bunescu, R. C. 2020, MNRAS, 495, 1613
Rezaie, M., Ross, A. J., Seo, H.-J., et al. 2021, MNRAS, 506, 3439
Rodríguez-Monroy, M., Weaverdyck, N., Elvin-Poole, J., et al. 2022, MNRAS, 511, 2665

Ross, A. J., Ho, S., Cuesta, A. J., et al. 2011, MNRAS, 417, 1350
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Shafer, D. L., & Huterer, D. 2015, MNRAS, 447, 2961
Soelch, M., Akhundov, A., van der Smagt, P., & Bayer, J. 2019, in Artificial Neural Networks and Machine Learning – ICANN 2019: Theoretical Neural Computation, ed. I. Tetko et al. (Cham: Springer), 444
The LSST Dark Energy Science Collaboration, Mandelbaum, R., Eifler, T., et al. 2018, arXiv:1809.01669
Wagoner, E. L., Rozo, E., Fang, X., et al. 2021, MNRAS, 503, 4349
Wagstaff, E., Fuchs, F., Engelcke, M., Posner, I., & Osborne, M. A. 2019, Proc. Machine Learning Research 97, Proc. of the 36th Int. Conf. on Machine Learning, 6487, https://proceedings.mlr.press/v97/wagstaff19a.html
Weaverdyck, N., & Huterer, D. 2021, MNRAS, 503, 5061
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Zaheer, M., Kottur, S., Ravanbakhsh, S., et al. 2017, arXiv:1703.06114
Zarrouk, P., Rezaie, M., Raichoor, A., et al. 2021, MNRAS, 503, 2562
Zhou, Z., Zhou, X., Zou, H., et al. 2018, PASP, 130, 085001
Zou, H., Zhou, X., Fan, X., et al. 2017, PASP, 129, 064101