Astronomy & Astrophysics manuscript no. output
May 14, 2025

Deep learning inference with the Event Horizon Telescope

II. The Zingularity framework for Bayesian artificial neural networks

M. Janssen[1,2], C.-k. Chan[3,4,5], J. Davelaar[6,7], and M. Wielgus[8]

[1] Department of Astrophysics, Institute for Mathematics, Astrophysics and Particle Physics (IMAPP), Radboud University, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands
e-mail: M.Janssen@astro.ru.nl
[2] Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, D-53121 Bonn, Germany
[3] Steward Observatory and Department of Astronomy, University of Arizona, 933 N. Cherry Ave., Tucson, AZ 85721, USA
[4] Data Science Institute, University of Arizona, 1230 N. Cherry Ave., Tucson, AZ 85721, USA
[5] Program in Applied Mathematics, University of Arizona, 617 N. Santa Rita Ave., Tucson, AZ 85721, USA
[6] Department of Astrophysical Sciences, Peyton Hall, Princeton University, Princeton, NJ 08544, USA
[7] NASA Hubble Fellowship Program, Einstein Fellow
[8] Instituto de Astrofísica de Andalucía-CSIC, Glorieta de la Astronomía s/n, E-18008 Granada, Spain

Received TBD; accepted TBD

ABSTRACT

Context. In this second paper in our publication series, we present the open-source Zingularity framework for parameter inference with deep Bayesian artificial neural networks. We carried out out supervised learning with synthetic millimeter very long baseline interferometry observations of the Event Horizon Telescope (EHT). Our ground-truth models are based on general relativistic magnetohydrodynamic simulations of Sgr A* and M87* on horizon scales. The models predict the synchrotron emission produced by these accreting supermassive black hole systems.
Aims. We investigated how well Zingularity neural networks are able to infer key model parameters from EHT observations, such as the black hole spin and the magnetic state of the accretion disk, when uncertainties in the data are accurately taken into account.
Methods. Zingularity makes use of the TensorFlow Probability library and is able to handle large amounts of data with a combination of the efficient `TFRecord` data format plus the Horovod framework for distributed deep learning. Our approach is the first analysis of EHT data with Bayesian neural networks, where an unprecedented training data size, under consideration of a closely modeled EHT signal path, and the full information content of the observational data are used. Zingularity infers parameters based on salient features in the data and is containerized for scientific reproducibility.
Results. Through parameter surveys and dedicated validation tests, we identified neural network architectures, that are robust against internal stochastic processes and unaffected by noise in the observational and model data. We give examples of how different data properties affect the network training. We show how the Bayesian nature of our networks gives trustworthy uncertainties and uncovers failure modes for uncharacterizable data.
Conclusions. It is easy to achieve low validation errors during training on synthetic data with neural networks, particularly when the forward modeling is too simplified. Through careful studies, we demonstrate that our trained networks can generalize well so that reliable results can be obtained from observational data.

Key words. methods: data analysis – techniques: high angular resolution – techniques: interferometric

1. Introduction

Low-powered active galactic nuclei (AGNs) are well described by numerical general relativistic magnetohydrodynamics (GRMHD) simulations (e.g., De Villiers & Hawley 2003; McKinney 2006; Dibi et al. 2012; Ryan et al. 2018). GRMHD simulations solve the equations of a magnetohydrodynamic fluid in curved spacetime. An initial setup with a weakly magnetized torus self-consistently evolves into radiatively inefficient accretion flows accompanied by outflows and jets.

In this work, we focus on the low-luminosity AGNs Sagittarius A* (Sgr A*) and Messier 87* (M87*). M87* is a nearby elliptical galaxy in the Virgo cluster and features a prominent extragalactic radio jet that has been resolved by observations in the radio to X-ray bands (e.g., Curtis 1918; Byram et al. 1966; Owen et al. 1989; Sparks et al. 1996; Biretta et al. 1999; Marshall et al. 2002; Hada et al. 2011; Mertens et al. 2016; Walker et al. 2018). Sgr A* is known as the "starving" supermassive black hole in our Galactic Center with a very low accretion rate and no visible jet emission, discovered as a bright radio source (Balick & Brown 1974). Detections of a gravitational redshift (GRAVITY Collaboration et al. 2018; Do et al. 2019) and Schwarzschild precession (GRAVITY Collaboration et al. 2020; Gravity Collaboration et al. 2022) of a star in orbit around the black hole served as important tests of General Relativity and for over two decades Sgr A*'s rich multiwavelength variability has been studied. Here, we refer to the detailed recent studies by Witzel et al. (2021) and Event Horizon Telescope Collaboration et al. (2022a) and references therein.

Electromagnetic observables computed from GRMHD models through general relativistic ray-tracing (GRRT) are overall in good agreement with high-resolution radio observations of Sgr A* and M87*. Selections of a "standard" set of GRMHD-GRRT mod-

els (Event Horizon Telescope Collaboration et al. (2019d, 2022e); see also the overview of the model space given in Janssen et al. (2025a)) have been scored against high-resolution millimeter-wave observations of Sgr A* and M87* with the Event Horizon Telescope (EHT) very long baseline interferometry (VLBI) array in various ways.

In Event Horizon Telescope Collaboration et al. (2019d), a χ^2 scoring of 100-500 snapshots per M87* model against the EHT total intensity data of M87* with the Themis (Broderick et al. 2020) and Gena (Fromm et al. 2019) software packages disfavored magnetically arrested disk (MAD) accretion models, which have strong and organized magnetic fields that can disrupt the accretion flow (e.g., McKinney et al. 2012; Narayan et al. 2012) and highly negative black hole spins (measured with respect to the accretion flow). For this average image scoring analysis, closure phases (Jennison 1958) and total intensity amplitudes were used. The measurements were averaged over minute-long VLBI scan, which can lead to decoherence on low signal-to-noise ratio (S/N) data, where atmospheric phase turbulence cannot be calibrated reliably.

In Event Horizon Telescope Collaboration et al. (2021c), the image-integrated linear and circular polarization, average linear polarization, and a parameter describing the azimuthal electric-vector position angle pattern (Palumbo et al. 2020; Palumbo & Wong 2022; Emami et al. 2023b) derived from the M87* linear polarization maps at 230 GHz (Event Horizon Telescope Collaboration et al. 2021a; Goddi et al. 2021) have been compared against a 200 GRRT image frames subset blurred with a 20 μas beam for each model. Overall, MAD models are in better agreement than standard and normal evolution (SANE) models, where the magnetic fields are weaker and more turbulent. The MAD preference is reinforced when the measured upper limits on resolved circular polarization are taken into account (Event Horizon Telescope Collaboration et al. 2023).

In Event Horizon Telescope Collaboration et al. (2022e), parts of the EHT Sgr A* data were used for source-structural constraints based on the image size, salient features in the measured flux densities, and geometric ring-model fits (Event Horizon Telescope Collaboration et al. 2022c), making use of the new Comrade software (Tiede 2022), among others. Models with negative spin, edge-on inclination angles, and equal ion and electron temperatures are disfavored.

The 2017 EHT Sgr A* polarization data were analyzed in Event Horizon Telescope Collaboration et al. (2024). The KerrBAM semianalytic model (Palumbo et al. 2022) was used for a qualitative exploration of the relations between polarization measurements and physical quantities of Sgr A*. The GRMHD scoring leaves a single passing MAD model at an inclination of 150° with a spin of 0.94 and comparatively strong jet emission, assuming that the observed rotation measure is produced by an external Faraday screen.

Multiwavelength observations of M87* (EHT MWL Science Working Group et al. 2021) rule out spin-zero models (based on jet power) and SANE models, where very hot electrons in the accretion disk produce an X-ray luminosity excess (Event Horizon Telescope Collaboration et al. 2019d). Sgr A* multi-wavelength data (Event Horizon Telescope Collaboration et al. 2022a) favors MAD models and rejects most models with large inclination angles and all models with an equal ion and electron temperature. The majority of the models are more variable than the observed intra-day total-flux light-curve flux variations (Wielgus et al. 2022), while SANE models are overall less variable than MADs (see the discussion in Event Horizon Telescope Collaboration et al. 2022e).

Using one specific image feature, the size of the black hole shadow (Falcke et al. 2000) calibrated with GRMHD-GRRT models, Psaltis et al. (2020), Kocherlakota et al. (2021), and Event Horizon Telescope Collaboration et al. (2022d) performed tests of spacetime metrics. The degree to which the shadow can be used for a clean test of gravity through a robust relation to the photon ring (e.g., Bardeen 1973; Johnson et al. 2020) is being debated and depends on how well we understand low-powered AGN accretion physics (Narayan et al. 2019; Gralla 2021; Chael et al. 2021; Bronzwaer & Falcke 2021; Özel et al. 2022; Wielgus 2021; Vincent et al. 2022; Paugnat et al. 2022).

Recently, further EHT data analysis methods have been proposed. Medeiros et al. (2023a,b) used dictionary learning of a GRMHD-GRRT simulations to reconstruct images from EHT data through a principal components analysis. Emami et al. (2023a) have shown that E and B linear polarization modes are informative regarding GRMHD-GRRT parameters. Conroy et al. (2023) have devised a method to measure rotation speeds in future EHT movie reconstructions using autocorrelations in the image domain. Chael et al. (2023) have identified a relation between horizon-scale polarimetric observables and electromagnetic energy extraction from the black hole spin. Yfantis et al. (2024) have developed a Bayesian ray-tracing parameter estimation framework, related to earlier work by Kim et al. (2016) and Psaltis et al. (2022). New methods for directly reconstructing the source structure for analyses in the image domain have been presented in Müller & Lobanov (2022), Müller et al. (2023), Mus et al. (2024a), Mus et al. (2024b), and Mus & Martí-Vidal (2024).

In this work we present Zingularity,[1] an open-source generic TensorFlow-based (Abadi et al. 2016a,b) framework of Bayesian deep neural networks for astronomical interferometers, which we developed to use the full information content of data from EHT observations for a GRMHD-GRRT data-driven parameter inference without the need to compromise on the number of model images due to computational limitations. This is realized by using large-scale synthetic data libraries that are based on a wide range of Sgr A* and M87* simulations (Janssen et al. 2025a). Here we describe and validate the Zingularity network training for the 2017 EHT observations. Janssen et al. (2025b) shows the parameter inference results from the application of the trained network to observational data alongside predictions of future upgraded EHT observations using models that go beyond the Kerr (1963) metric.

Driven by the vastly increasing data sizes, machine learning methods are becoming increasingly popular in astronomy, mostly in the area of source finding as well as classification (see for example VanderPlas et al. 2012; Baron 2019; Fluke & Jacobs 2020; Kong et al. 2020; Smith & Geach 2023; Djorgovski et al. 2022; Huertas-Company & Lanusse 2023; Moriwaki et al. 2023) and with convolutional neural networks (e.g., LeCun et al. 2015) in particular. Related to our work, Shatskiy & Evgeniev (2019), van der Gucht et al. (2020), Yao-Yu Lin et al. (2020), and Popov et al. (2021) have used neural networks in the EHT total intensity image domain, where the uncertainties and model degeneracies are large for the current EHT (u, v) coverage. Qiu et al. (2023) have trained a random forest machine learning model on a few salient features in the image domain from GRMHD-GRRT simulations, including polarization information. Levis et al. (2024) used neural networks to study the orbital dynamics around Sgr A* in the millimeter wavelength polarimetric observations. Compared to Yao-Yu Lin et al. (2021), where total intensity visibility measurements from M87* were used for the network training,

[1] https://gitlab.com/mjanssen2308/zingularity

we used a larger training dataset for M87*, also considered the Sgr A* data, included more effects for the modeling of the EHT signal path, took the additional polarization information from the measurements into account, and obtained uncertainties on the inferred parameters by using a Bayesian network and bootstrapping of observational data errors.

In terms of more general machine learning applications for astronomical interferometers, Sun et al. (2020) have developed a method for optimizing the telescope locations for VLBI arrays, Sun & Bouman (2020) present a variational deep probabilistic imaging approach, quantifying reconstruction uncertainty, and Sun et al. (2022) use an improved variational inference method to obtain accurate posterior samples for parameter inference tasks. Neural-network-based interferometric imaging methods have been developed by Morningstar et al. (2018, 2019), Gheller & Vazza (2022), Schmidt et al. (2022), Aghabiglou et al. (2024), Thyagarajan et al. (2024), Feng et al. (2024), and Lai et al. (2025). Mohan et al. (2024) and SaraerToosi & Broderick (2024) use neural networks to generate M87* images. Finally, Duarte et al. (2022) have used a deep convolutional neural network as a fast generator of GRMHD simulations, a method that might substantially speed up the generation of synthetic training data in the future.

In Section 2 of this work we introduce machine learning and astrophysical interferometry concepts used throughout this paper. In Section 3, we give our motivation for developing Zingularity (i.e., the usefulness of ML applications for the analysis of data from astronomical interferometers like the EHT). In Section 4, we introduce the GRMHD-GRRT synthetic data library of Sgr A* and M87* EHT observations used as a training dataset for Zingularity in this work. In Section 5, we describe the Zingularity framework together with the specific algorithms used to fit models to EHT data in this paper. In Section 6, we show the suite of validation tests and diagnostics implemented to verify the output of Zingularity, particularly under the aspect of how Zingularity performs Bayesian model parameter inference from uncharacterized data and how much this inference is hindered by instrumental effects. We offer our conclusions about Zingularity and the ability to extract GRMHD-GRRT parameters from current EHT observations in Section 7. We finish with an outlook of planned future studies with Zingularity in Section 8.

2. Machine learning and interferometry concepts

Machine learning (ML) methods automatically learn the characteristics of a training dataset $\widetilde{T}$. In this work, supervised learning is carried out, where labeled training data are used. As such, the ML algorithm optimizes itself based on a known input-output mapping. We are using a fraction of the data contained in $\widetilde{T}$ as validation data, which is used to compute the accuracy of the trained ML model based on data that was not used for the optimization.

Artificial neural networks (ANNs) are ML algorithms that are designed based on neural connections of biological brains. Information in the form of floating point numbers are passed from an input dataset, through connected computing nodes ('neurons') that each transform the data with some function g, up to a final layer, that yields the predictions of the ANN. A network organized in layers, for which the same type of operation $g(x)$ is computed for each neuron and the input x is taken as the output of the previous layer in the case of feedforward networks (Appendix A). Every network has one input and one output layer. ANNs with more than three layers are referred to as deep neural networks. Free parameters of each neuron's g are optimized ('trained') based on a loss function between the network's input and output.

We refer to all data imperfections (i.e., all differences in the data measured between a realistic instrument and a hypothetical perfect measurement device) as data corruption effects along the signal path $\widetilde{C}$. These effects encompass thermal noise and uncorrected systematics from the instrument itself, corruptions that occur along the long signal path of astronomical observatories (e.g., interstellar scattering and the added noise and absorption from Earth's atmosphere for ground-based observatories) as well as uncorrected systematics introduced by assumptions made during the data reduction process.

We denote uncharacterized data, for which we do not know the underlying physical reality as $\widetilde{U}$. Typically, $\widetilde{U}$ is obtained from a measurement and affected by $\widetilde{C}$. We assume $\widetilde{U}$ to be reasonably well described by a model $\widetilde{M}$ and wish to infer the model parameters from the data. For our EHT horizon-scale observations of Sgr A* and M87*, we use GRMHD-GRRT images as $\widetilde{M}$.

With synthetic data $\widetilde{S}$, we attempt to create mock observations that sample the data produced by an astronomical observatory as closely as possible. This requires a full forward modeling, where the physical reality is well described by $\widetilde{M}$ and all relevant data corruption effects along the signal paths are modeled correctly. We used synthetic data based on GRMHD-GRRT images for our training input of our ANN as described in Janssen et al. (2025a): $\widetilde{T} = \widetilde{S}(\widetilde{M}, \widetilde{C})$. It is therefore required that the most significant features of $\widetilde{S}$, that have discriminative power for $\widetilde{M}$ in the presence of $\widetilde{C}$, to also be present in $\widetilde{U}$. GRRT images are currently our best models for the interpretation of EHT observations (Event Horizon Telescope Collaboration et al. 2019d, 2021b). We note that that restriction to GRMHD-GRRT makes our analyses model-dependent and that the validity of these models is also being questioned in the literature (e.g., Gralla 2021). While we make use of a large parameter space of the used models, all simulations assume ideal MHD and a simple description of the electron temperature as described in the previous paper in this series. Further limitations of the models used in this initial study are described in the outlook section 8.

Most astronomical instruments measure the electric field emitted by a radiating source with sky brightness distribution I. Interferometers cross-correlate the signals measured by pairs of telescopes to create visibilities $\mathcal{V}$. A Fourier relationship links I and uncorrupted $\mathcal{V}$ (van Cittert 1934; Zernike 1938). However, interferometers provide only an incomplete sampling of $\mathcal{V}$ corresponding to projected vectors connecting all pairs of telescopes (baselines). Hence, obtaining I from $\mathcal{V}$ becomes an ill-posed problem and additional information or assumptions must be used (Thompson et al. 2017). Furthermore, $\widetilde{C}$ must be modeled to correct the $\mathcal{V}$ measurements (Hamaker et al. 1996). Visibilities are measured as a function of time t, frequency ν, telescope baseline vector (u, v, w), and polarization P. Each telescope in the interferometer typically measures two orthogonal polarization states of the radiation (right/left-handed circular polarization or horizontal/vertical linear polarization). Four polarization products P are formed from the two telescopes that form a baseline and the two orthogonal polarization measurements. The four Stokes parameters (Stokes 1851) can be formed from different linear combinations of the four P values.

ML methods build a complex model that tries to capture the most important features of $\widetilde{M}$ through a training dataset $\widetilde{T}$, such that parameters of $\widetilde{M}$ can be retrieved from $\widetilde{U}$ indirectly. We

Table 1. Shorthand for terminology used in this work.

Symbol	Description
$\widetilde{U}$	Uncharacterized data; here the observational data
$\widetilde{C}$	Data corruption effects along the signal path
$\widetilde{M}$	Model data
$\widetilde{S}$	Synthetic data
$\widetilde{T}$	Training data; here $\widetilde{T} = \widetilde{S}(\widetilde{M}, \widetilde{C})$

will be referring to methods that are fitting $\widetilde{M}$ directly to $\widetilde{U}$ as conventional. For conventional interferometer methods, $\widetilde{M}$ are typically simple geometric models or pixel-based images I. The direct fitting of complex models such as GRRT images poses a significant computational challenge. We summarize our shorthand for the machine learning concepts in Table 1.

3. Motivation

In this section, we summarize the motivation for developing Zingularity, which are the use-cases and unique opportunities when employing ML for data from astronomical interferometers and the EHT in particular.

Firstly, the TensorFlow library makes efficient use of the modern `TFRecord`[2] data format and scales well with CPU, GPU, and TPU[3] computing power. This allows Zingularity to be trained on large $\widetilde{T}$ that span a wide range of $\widetilde{M}$ and $\widetilde{C}$ parameters. For the EHT, Zingularity can make inferences based on the full GRMHD-GRRT parameter space, which is not computationally feasible with conventional methods. Furthermore, Zingularity is able to process visibilities to their full extend, without losses that can occur when data are being averaged. Once the network is trained, the application to $\widetilde{U}$ for inference is practically instantaneous. Efficient and scalable analysis software will be needed for the increasing data rates of future instruments. In particular, the EHT plus next-generation EHT (Blackburn et al. 2019; Johnson et al. 2023; Doeleman et al. 2023) will observe with an increased bandwidth as well as more telescopes in the future (The Event Horizon Telescope Collaboration 2024) and exascale computing will be needed to handle the data produced by the Square Kilometre Array (Dewdney et al. 2009) and next-generation Very Large Array (McKinnon et al. 2019). It has already been demonstrated that such big computing tasks are achievable with TensorFlow (Kurth et al. 2018). More details about the computational speed of Zingularity are given in Section 5.4.

Secondly, supervised ML is designed to obtain results with well defined fidelity metrics that describe how accurately model parameters can be recovered in a traceable manner (Section 5.1.5).

Thirdly, ANNs are training on salient and robust features by design. As such, data points that are strongly affected by $\widetilde{C}$ or that are not being distinctive for the underlying $\widetilde{M}$ parameters of interest are not taken into account. In simple χ^2 analyses for the goodness of fit, these data points can lead to poor results. Closure phases (Jennison 1958), log closure amplitudes (Blackburn et al. 2020), and closure traces (Broderick & Pesce 2020) are known robust observables that can be formed from visibilities, but whose errors are no longer Gaussian in the low S/N regime and whose variance does depend on telescope gain errors (Lockhart & Gralla 2022). GRMHD scoring employed in past EHT analyses is dependent on how the method is implemented as well as how observational and model uncertainties are treated. Deep ANNs may uncover more complex robust data combinations, which are discriminative for the $\widetilde{M}$ parameter space. Especially for complex models, such as GRRT images, the parameter-to-feature correspondence is not fully known a priori.

Fourthly, by using large $\widetilde{T}$ consisting of realistic synthetic data, all forward-modeled data corruption effects are taken into account by ML methods. For sparse interferometric measurements, data corruptions (e.g., antenna gain errors, instrumental polarization leakage, and atmospheric effects for high frequency observations, Janssen et al. 2025a) can have a substantial influence on the obtained results. The modeling of these (time-variable) effects is a) convoluted with the reconstructions of (time-variable) source structures during self-calibration (Readhead & Wilkinson 1978; Pearson & Readhead 1984), b) limited by the S/N of the observational data (e.g., Janssen et al. 2022), c) often based on simplifying assumptions (Event Horizon Telescope Collaboration et al. 2019b), and d) introduces additional noise from the precision of determined calibration solutions (e.g., fringe-fitting phase, delay, and rate estimates; Thompson et al. 2017). Some algorithms can marginalize over a range of some $\widetilde{C}$ (e.g., Broderick et al. 2020; Pesce 2021), which requires considerable computational resources. Moreover, a realistic modeling of the data acquisition and processing of $\widetilde{U}$ for the generation of $\widetilde{S}$ ensures that the results are not affected by unknown systematics introduced by a specific data reduction procedure. VLBI-specific data processing methods are often complex and the consequences of some calibration assumptions are not always fully explored and understood. One example is the usage of a point source during the fringe-fititng stage (Natarajan et al. 2020). In analyses published by the EHT collaboration, a small percentage of systematic noise is added to deal with unknown data imperfections (Event Horizon Telescope Collaboration et al. 2019b, 2022a). Here, we assert this to be covered by our forward modeling plus error bootstrapping. The flat addition of a single systematic noise budget to all baselines can lead to source signals being washed out on baselines that measure high correlated flux densities.

Fifthly, ML can easily be used to study the predictive power of $\widetilde{M}$ separately from $\widetilde{C}$, if the signal path and instrument can be modeled accurately. Synthetic training data can be flexibly generated with different combinations of $\widetilde{C}$ and the corresponding accuracy of $\widetilde{M}$ parameter predictions can be studied. The predictive power of $\widetilde{M}$ alone can be studied in the limit of a perfect hypothetical instrument, where no $\widetilde{C}$ is added to the synthetic data. Roelofs et al. (2020) performed a simple image-based fidelity study of how well an upgraded EHT would be able to detect the M87 jet for example. Another example, where the model predictions are straightforward, are self-similar photon-subrings surrounding the shadow of a black hole. These rings can be studied to test GR and to accurately measure black hole parameters (Johnson et al. 2020). It is expected for the photon-ring signals to dominate at long interferometric baselines. With ML, one could for instance study the accuracy of a space-VLBI GR test, marginalized over a range of possible accretion environments around the black hole, as a function of maximally achieved baseline length and calibration uncertainty.

Sixthly, Zingularity makes use of the full information content of $\widetilde{T}$ and $\widetilde{U}$. Conventional mm VLBI methods on the other hand often analyze the visibilities in stages: Using only the total

[2] https://www.tensorflow.org/tutorials/load_data/tfrecord.
[3] https://cloud.google.com/tpu/docs/tpus

Table 2. ZINGULARITY training and application parameters. For Sgr A*, we found two equally viable models with different $N_{\rm ep}$.

Category	Param.	Value		Description
Common ZINGULARITY GRMHD -GRRT EHT parameters	f	Swish (Ramachandran et al. 2017)		Activation function for all hidden layers
	Ξ	RMSProp		Optimization algorithm
	L	Negative log-likelihood		Loss function
	$\eta_{\rm val}$	0.1		Fraction of $\widetilde{T}$ used for validation
	$N_{\rm b}$	256		Training batch size
	l_r	$0.001 \times n/N_{\rm ep}$		Learning rate warm-up for $0 \leq n \leq 0.1 \times N_{\rm ep}$
		$0.0001 \times (1 + \cos(n\pi/N_{\rm ep}))/2$		Learning rate cosine decay for $0.1 \times N_{\rm ep} \leq n \leq N_{\rm ep}$
		M87*	**Sgr A***	
Fiducial models	$N_{\rm vis}$	8×5489	8×13840	Number of data points in a training sample
	$N_{\rm tr}$	600,000	252,000	Number of training samples
	$N_{\rm ep}$	70	50, 60	Number of training epochs
	$\eta_{\rm drop}$	0.01	0	Dropout rate for stochastic neuron deactivation
	$\mathcal{L}_1$	0.01	0.01	L_1 (lasso) regularization hyperparameter
	$\mathcal{L}_2$	0.01	0.01	L_2 (ridge) regularization hyperparameter
	$k_{\rm conv}$	8	8	Receptive field of CNN layers
	$n_{\rm CNNb}$	16	8	Baseline number of neurons for the ResNet CNN layers
	$n_{\rm CNNl}$	128	2048	Neurons in last ResNet CNN layer
	$N_{\rm dense}$	15	12	Number of post-ResNet dense variational layers
	$n_{\rm dense}$	128	1024	Neurons in post-ResNet dense variational layers
	$N_{\rm free}$	1,376,806	135,068,877	Number of free parameters in the network
Bootstrapping errors for the EHT	$\mathcal{D}$	1 – 3 % (EHT et al. 2021a)		Polarization leakage ($\mathcal{D}$-terms)
	$\mathcal{G}_{\rm planet}$	10 % (Janssen et al. 2019a)		Primary calibrator model uncertainties
	$\mathcal{G}_{\rm scatter}$	5 – 35 % (Janssen et al. 2019a)		DPFU uncertainty due to measurement scatter
	$\mathrm{gc_B}$	3.6 – 10.4 % (Janssen et al. 2019a)		Measurement error on gain curve curvature
	$\mathrm{gc_{E0}}$	1 – 2 % (Janssen et al. 2019a)		Measurement error on gain curve peak elevation
	$\sigma_{\rm th}$	$\sim 8.5 \times 10^{-6}\ \sqrt{\mathrm{SEFD_1 SEFD_2}}$		Thermal noise of EHT data used in this work

intensity (Stokes $\mathcal{I}$) information first, followed by linear polarization (Stokes $\mathcal{Q}$ and $\mathcal{U}$) in some cases. The usually negligible circular polarization (Stokes $\mathcal{V}$) signals from the source and temporal evolution of I are also often modeled separately, if at all.

Finally, while previous ANN-based parameter estimation studies for the EHT were based on Stokes $\mathcal{I}$ images (van der Gucht et al. 2020; Yao-Yu Lin et al. 2020), ZINGULARITY works with the visibilities directly and makes use of the full polarization information. The advantages of using the visibilities directly are that the intermediate modeling step necessary to obtain I is removed and that there is a precise description of $\mathcal{V}$ uncertainties. Additionally, there are no constraints for image-specific parameters. Thus, different models, for example with different fields of view and pixel sizes, can be used together.

Before proceeding, the limitations of our current ML approach should be discussed as well. Like in the EHT GRMHD scoring, results have to be interpreted within the $\widetilde{M}$ parameter space. Further, with our current BANN implementations, we are tied to specific (u, v) coverages from specific observations. Here, it is important to note that a proper forward modeling for unbiased inference requires the $\widetilde{T}$ generation to be tailored to the characteristics of specific observations anyway. We thus argue that computational efficiency could be gained by speeding up our $\widetilde{S}$ generation methods, rather than attempting to devise a flexible BANN that can be applied to observations it was not trained on.

4. Training data

In this work, we used the EHT synthetic data library $\widetilde{S}$ that is based on the standard GRMHD-GRRT Sgr A* and M87* models $\widetilde{M}$ from Janssen et al. (2025a) as training data $\widetilde{T}$ for ZINGULARITY. A direct comparison of the models with EHT image reconstructions is not possible, because the observational images have a limited resolution and are not unique (Event Horizon Telescope Collaboration et al. 2019c, 2022b).

$\widetilde{S}$ was created with SYMBA (Roelofs et al. 2020) to model the complete signal path of observational VLBI data and stored as `TFRecord` files. The $\widetilde{M}$ parameters of interest are the MAD/SANE magnetic state of the accretion disk $\phi_{\rm mag}$, the black hole spin a_*, the proportionality between the ion- to electron temperature $R_{\rm high}$ in the accretion disk, and for Sgr A*, also the inclination angle $i_{\rm los}$ and position angle $\theta_{\rm PA}$ of the source. For M87*, we fixed $i_{\rm los} = 17°$ (Mertens et al. 2016) and $\theta_{\rm PA} = 288°$ (Walker et al. 2018). We set black hole masses of $4.14 \times 10^6\ M_\odot$ and $6.2 \times 10^9\ M_\odot$ and distances of 8.127×10^3 parsec (pc) and 16.9×10^6 pc for Sgr A*, respectively M87* (Gebhardt et al. 2011; Gravity Collaboration et al. 2019; Do et al. 2019). Random variations at the 10 % level were added to the mass over distance ratios in the synthetic data generation as described in Section 3.6 of Janssen et al. (2025a). For each $\widetilde{M}$, we have about 1000 images. For M87*, we have multiple $\widetilde{S}$ realizations for each single image. For Sgr A*, we have multiple realizations from groups of 432 images, capturing the variability of the source during an observation.

For each realization, data corruption effects $\widetilde{C}$ due to thermal noise, uncertain telescope gains and polarization leakages, Earth's atmosphere, as well as the interstellar scattering screen toward Sgr A* were varied. Given a set bandwidth, integration time, and quantization efficiency of the recorded data, the thermal noise was determined by the System Equivalent Flux Densities (SEFDs)

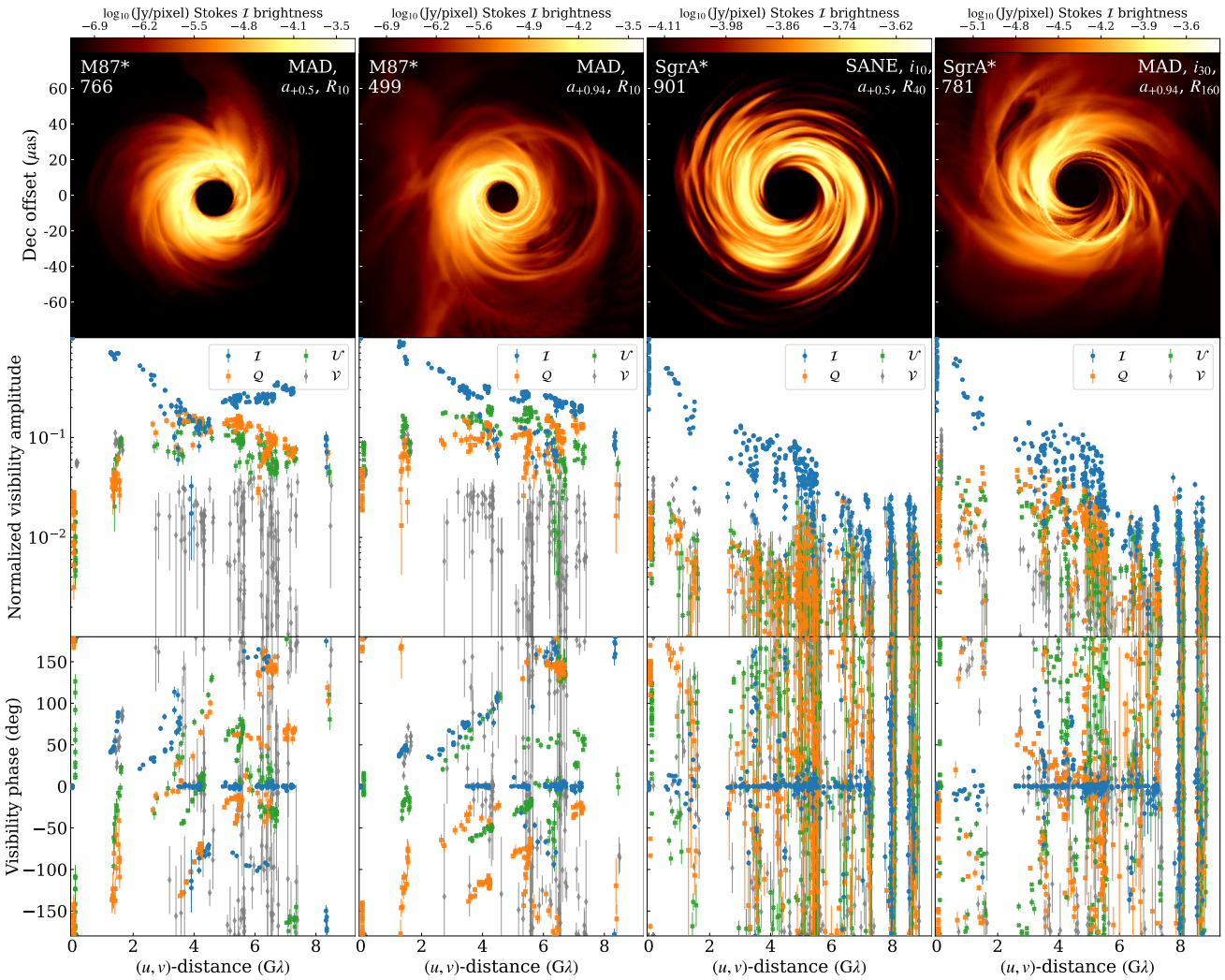

Fig. 1. Four training dataset examples. The top row shows the total intensity ray-traced ground-truth model images on logarithmic scales with varying dynamic ranges. Normalized full-pol visibility amplitudes and phases of corresponding synthetic data realizations are displayed with thermal noise error bars as a function of baseline length in units of the observing wavelength $\lambda \approx 1.3$ mm (see Janssen et al. (2025a) for the (u, v) coverage) in the middle and bottom rows, respectively. The measurements shown can come from different orientations at the same baseline length. For a better readability, the visibilities have been averaged over scan durations, normalized amplitudes lower than 0.001 have been clipped, and the values of the different Stokes parameters are each offset by $50\,\mathrm{M}\lambda$ on the x-axis. In the top left corner of each model image is indicated whether the model and data correspond to M87* or Sgr A* and the GRRT frame number of the image. The strongly time-variable Sgr A* data were generated from multiple GRRT frames, of which a single frame is displayed here. Spin $a_* = s$, $R_{\rm high} = r$, and $i_{\rm los} = l$ parameters are listed in a shorthand notation as a_s, R_r, and i_l in the top right corner of each model image. The Sgr A* models are shown here with $\theta_{\rm PA} = 0$.

of two antennas forming a baseline. The SEFD is the sum of all noise contributions along the signal path. Telescope gains are errors in the measured amplitude and phase of the signal. Polarization leakage is the cross-talk between the two orthogonal telescope receivers, which measure different polarization states of the incoming radiation. The Earth's atmosphere causes an attenuation of the signal, additional noise, and phase errors. The Sgr A* scattering screen leads to a blurred source image with additional induced substructures. These corruption effects are described in detail in Section 4 of Janssen et al. (2025a).

As noted in Section 1, some regions of the GRMHD-GRRT parameter space are disfavored based on simple comparisons with the EHT data and multiwavelength constraints. Nonetheless, we did not apply any a priori cuts on $\widetilde{T}$ and trained on data that covers the full $\widetilde{M}$ parameter space. On the one hand, this serves as a validation for Zingularity, as the posterior probability from the fit should be disjoint from the the parameter space that is strongly disfavored by EHT constraints applied in earlier works. On the other hand, our strategy serves as a test of the models. If GRMHD-GRRT describes the physical reality of Sgr A* and M87* well, the models favored by Zingularity fits should be in agreement with (quasi-)simultaneous multiwavelength constraints in the absence of parameter degeneracies.

Table 2 gives an overview of the training data and neural network parameters used in this work. We have $N_{\rm tr}$ individual $\widetilde{T}$ samples. Each consists of $N_{\rm vis}$ real and imaginary values of the complex correlation coefficients per RR, RL, LR, and LL correlation product. The (u, v) coverage corresponds to the 2017 EHT observations on April 7 for Sgr A* and April 11 for M87* (Event Horizon Telescope Collaboration et al. 2019a; Janssen et al. 2025a). The time cadence of the data within VLBI scans is kept to a short 10 s sampling, to avoid decoherence effects from residual phase errors. Using the correlation outputs directly without long averaging intervals keeps the systematic error budgets low; the determined total intensity error budget can be erroneous in the presence of circularly polarized source signals and long averaging

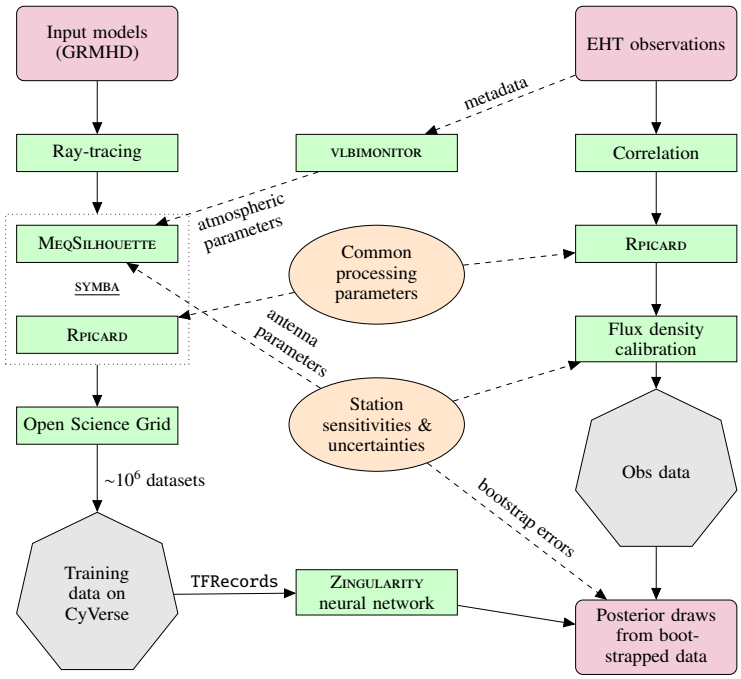

Fig. 2. Flowchart of the Zingularity data streams. The left column shows the pathway from the input theory models $\widetilde{M}$ to the training data $\widetilde{T}$. The right column shows the processing chain for the observational EHT data $\widetilde{U}$. The central column presents the common metadata used.

times likely cause decoherence in mm VLBI observations. For the current EHT data of Sgr A* in particular, both effects need to be taken into account (Event Horizon Telescope Collaboration et al. 2022a). But also for the M87* EHT data, coherence losses will occur when averaging the data in time (Figure 18 in Event Horizon Telescope Collaboration et al. 2019b).

Figure 1 shows a few $\widetilde{T}$ examples alongside the underlying ground-truth $\widetilde{M}$. Unlike the native visibilities used as neural network input, the data shown here are in a physically meaningful Stokes $\mathcal{I}$ (total intensity), $\mathcal{Q}\&\mathcal{U}$ (linear polarization), $\mathcal{V}$ (circular polarization) amplitude and phase representation averaged over VLBI scan durations of a few-minutes. The Sgr A* data have a better (u, v) coverage and exhibit intrinsic variability – while a single static frame made up an M87* dataset, multiple frames were used akin to a movie for Sgr A*, as the hours-long EHT observing track is much longer than the ~ 20 s gravitational timescale of the source.

The two M87* models displayed differ in the black hole spin parameter a_*. The model with the highly spinning $a_* = 0.94$ black hole possesses more extended emission. With relatively low $R_{\rm high} = 10$ values, the ratio of jet to disk emission is comparatively small. Between about 3.5 Gλ and 4.5 Gλ, a clusters of Stokes $\mathcal{I}$ visibility phases are offset by roughly 180° between the two models. Concurrently, slight offsets and a steeper phase evolution with baseline length for the $a_* = 0.5$ model data are present for the $\mathcal{Q}$ and $\mathcal{U}$ phases. The collective differences in phases across multiple Stokes parameters can be possible salient model features that allow for a distinction of the spin parameter in this case study. Other differences in the visibility data could be the result of different $\widetilde{C}$ realizations. For example, telescope gain errors could cause the differences of Stokes $\mathcal{I}$ amplitudes, while polarization leakage could be responsible for the $\mathcal{Q}$ and $\mathcal{U}$ phase differences at other baseline locations, where Stokes $\mathcal{I}$ shows no significant changes.

Out of the few Sgr A* $\widetilde{M}$ that pass most of the multiwavelength and EHT data constraints considered in Event Horizon Telescope Collaboration et al. (2022e), we selected two examples for Figure 1. The SANE $a_* = 0.5$, $R_{\rm high} = 40$, $i_{\rm los} = 10°$ model fails only the 86 GHz source size. The MAD $a_* = 0.94$, $R_{\rm high} = 160$, $i_{\rm los} = 30°$ model fails only the variability constraints (Wielgus et al. 2022; Broderick et al. 2022).

For the Sgr A* data, it is more difficult to identify salient features by eye due to the intrinsic $\widetilde{M}$ variability. A possibly important feature is the higher polarization of the MAD model; across all baseline lengths, the $\mathcal{Q}$ and $\mathcal{U}$ amplitudes are higher and phases more coherent. As noted in Event Horizon Telescope Collaboration et al. (2022e), the Sgr A* SANE model shown here is indeed most likely weakly polarized. However, the polarization can differ between model frames and be affected by polarization leakage, which makes the need for a deep BANN, trained on many $\widetilde{M}$ and $\widetilde{C}$ realizations to identify the robust salient features, evident.

5. Zingularity

Zingularity[4] is a modular open-source framework for the implementation of Bayesian TensorFlow-based neural networks. The input data for $\widetilde{T}$ and $\widetilde{U}$ is converted into the optimized and efficient `TFRecord` format (see Section 5 in Janssen et al. (2025a)). Based on this self-contained uniform data format, any kind of TensorFlow ANN can be trained and used for inference, independent from the type of input data. Zingularity runs on CPUs, a single GPU/TPU, or through distributed computing on multiple GPUs/TPUs. Bootstrapping methods, where $\widetilde{U}$ is resampled based on known $\widetilde{C}$, are implemented as well. For each resampled dataset, the surrogate posterior q_φ can be computed efficiently for inference.

Figure 2 shows an overview of the complete Zingularity processing chain from the input data to posteriors; how the training data are produced from the GRMHD models, how the ob-

[4] https://gitlab.com/mjanssen2308/zingularity

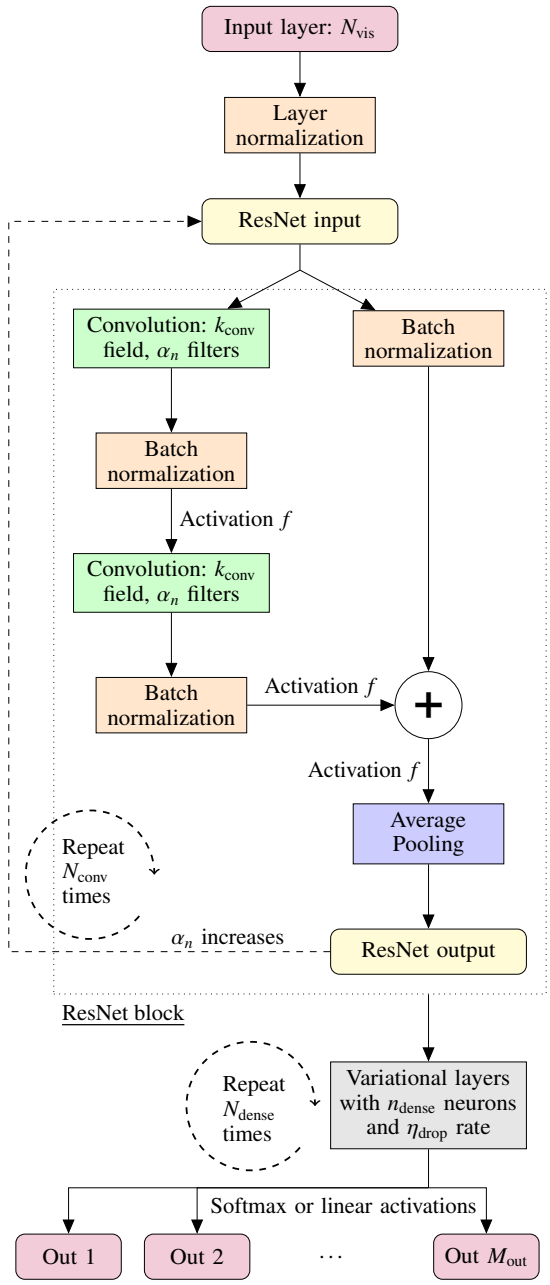

Fig. 3. Layout and data flow of our chosen M87* and Sgr A* BANN architectures. The ResNet and dense variational blocks are repeated $N_{\rm conv}$ and $N_{\rm dense}$ times, respectively. Single variational neurons are used in each output layer.

servational data are processed, and the parallels between the observation and theory processing chains. The integration of the VLBIMONITOR (Event Horizon Telescope Collaboration et al. 2019a), Open Science Grid (Pordes et al. 2007; Sfiligoi et al. 2009), and CyVerse data storage (Goff et al. 2011; Merchant et al. 2016) into a single pipeline with the PEGASUS software (Deelman et al. 2015) is described in Janssen et al. (2025a).

5.1. EHT GRMHD-GRRT network implementation

In this work, we used a BANN architecture that couples a residual (ResNet) Convolutional Neural Network (LeCun et al. 1989, 1998; He et al. 2015) with several `DenseVariational` fully connected layers performing variational inference (Figure 3). The implementation in ZINGULARITY is done with TENSORFLOW PROBABILITY (Dillon et al. 2017, Section 5.1.2), combined with multiple regularization methods for an increased model robustness (Section 5.1.3).

A deep architecture makes it possible for the network to pick up complex data combinations as salient features, closure quantities for example. With the use of variational layers with trainable weight distributions as hidden- and output layers, the BANN captures the epistemic and aleatoric uncertainties. These correspond to uncertainties in the model and the data, respectively.

In the following subsections, we describe our BANNs used to infer GRMHD-GRRT parameters from EHT observations as implemented in ZINGULARITY. The numerical values and method implementations for all relevant parameters of the network are listed in Table 2.

5.1.1. Input and output

ZINGULARITY performs the conversion of the input data to `TFRecord` files as a pre-processing step. The flexible conversion pipeline can handle any type of labeled or unlabled data. For full-polarization interferometric data, the real and imaginary component arrays for each correlation product correspond to 8 input 'channels' in our neural network architecture. An initial layer normalization (Lei Ba et al. 2016) is applied to the input, where the normalization parameters are optimized using each single example in a training batch.

The output layers consist of fully connected single Bayesian variational neurons for each inference task: With linear activation functions for a_*, $R_{\rm high}$, and for Sgr A* also $i_{\rm los}$ as well as $\theta_{\rm PA}$. The $\phi_{\rm mag}$ classification is done with a softmax activation. The priors and posterior functions used for the regression output layers are the same as those used by the hidden layers. The hidden layers (Section 5.1.2) model the data between the input and output. We use a single chain of hidden layers and split only in the last output layer. All of our $M_{\rm out}$ output layers are connected to same last hidden layer, so we can infer dependences between $\widetilde{M}$ parameters. We have tested that using individual, distinct networks per parameter does not improve the inference accuracy.

When applied to Sgr A* and M87*, the BANN is trained for $N_{\rm ep}$ epochs (iterations over $\widetilde{T}$). $N_{\rm ep}$ is determined empirically as the value where the loss L saturates. As studied by Genkin & Engel (2020), the best stopping criterion is where the loss curve starts developing a very shallow, plateau-like curve. When training for longer, the network may fit to noise in the training data. In practice, we determine the approximate saturation point by eye for each model and then survey a range of $N_{\rm ep}$ around it. For Sgr A* we find two equally viable $N_{\rm ep}$ of 50 and 60. Unless stated otherwise, we are using the $N_{\rm ep} = 60$ trained network as our fiducial Sgr A* model for the analyses presented in this work.

The hidden layers between the input and output are described below.

5.1.2. Hidden layers

Generally, we need a BANN with sufficiently large trainable parameters $N_{\rm free}$, to be in an overparameterized "double descent" regime (e.g., Belkin et al. 2019; Schaeffer et al. 2023). Additionally, we strive for deeper rather than wider networks by having more layers and fewer neurons per layer. In practice, our network widths are based on the input data size and computational limitations. Network depths are increased until no further improvements in validation errors are gained. We experimented

with different BANN architectures and selected the one with the best performance in our parameter surveys described later in this manuscript.

The combination of convolution operations and skip connections in the ResNet enables an in-depth modeling of the data to pick out the salient features in the data (i.e., which locations in time-baseline space and combinations of the visibility data are the most informative for the $\widetilde{M}$ parameter inference in the presence of $\widetilde{C}$). The Bayesian nature of the variational layers enables us to see parameter dependences and uncertainties in the posteriors formed from the ResNet pre-processed data.

ResNet blocks: The input visibilities are time-baseline sorted into a two-dimensional $(8, N_{\rm vis})$ shape. Initially, the data flows through $N_{\rm conv} = \log_2 N_{\rm vis}$ ResNet blocks.

Each block has the same architecture, where the data are passed through two parallel branches. The first branch consists of two consecutive and identical sub-blocks. In these sub-blocks, the data first flows through a convolution layer with α_n filters/neurons, then a Batch normalization layer (which normalizes the input per training batch), and finally a layer with activation function f. For the ResNet block number $n = 1, 2, ..., N_{\rm conv}$, we have

$$\alpha_n = (n_{\rm CNNb})^{1-\frac{n}{N_{\rm conv}}} \, (n_{\rm CNNl})^{\frac{n}{N_{\rm conv}}} \quad . \tag{1}$$

Each convolution filter α_n has a receptive field of size $k_{\rm conv}$. The purpose of these operations is to extract the most meaningful locations and combinations of the data in terms of information content. We thus refer to this as modeling branch.

The second branch is a skip connection, where the data flows through one layer with α_n convolution filters, each having a receptive field of one (to have a matching dimension of the data with the modeling branch), followed by Batch normalization.

The output of the two branches are then added and passed through another activation f. Finally, the data are downsampled by a factor of two in an average pooling layer. The data in the last ResNet block will have a dimension equal to $n_{\rm CNNl}$.

The initial weight parameters of each convolution layer are drawn from a uniform distribution between $\pm\sqrt{6/N_{\rm I+O}}$, where $N_{\rm I+O}$ is the number of input plus output units (Glorot & Bengio 2010). Bias terms are initialized as zeros.

Bayesian fully connected layers: After the ResNet blocks, the data are passed through $N_{\rm dense}$ fully connected variational layers, each having $n_{\rm dense}$ neurons with activation function f.

Multivariate standard Normal distributions serve as prior for the weights, while no variational inference is performed for the bias terms. The surrogate posteriors are represented with trainable Normal distributions.

We also tried network architectures that consisted only out of Bayesian fully connected layers, without any convolution layers. Overall, these had higher validation errors compared to our fiducial ResNet+fully connected models. Yet, for M87* the parameter inference on observational data gave consistent results (Janssen et al. 2025b). For Sgr A*, it was difficult to find robust network architectures without convolutions through our parameter survey method (Section 6.2).

5.1.3. Generalization

Overfitting is common in ML applications, as a model may train itself on peculiarities of $\widetilde{T}$, which are not generally representative for the underlying features of interest from $\widetilde{M}$. Overfitting in the traditional statistical sense can however be benign in deep learning tasks: Neural networks can perform well, even with perfect fits to training data when the sample size is much smaller than the number of possible directions in parameter space that are unimportant for predictions (e.g., Bartlett et al. 2020; Chatterji & Long 2020). We strive for a model that generalizes well.

We split out a fraction $\eta_{\rm val}$ of $\widetilde{T}$ as validation data. These data are not used for training and can therefore be seen as uncharacterized data $\widetilde{U}$ that is unknown to the network. We check that our performance metrics (Section 5.1.5) give the same answers for the training and validation data.

Additionally, a Dropout regularization is employed, where a fraction of $\eta_{\rm drop}$ of each hidden layers' neurons are randomly deactivated in each forward pass when the network is trained and inferences are made (Hinton et al. 2012; Wan et al. 2013). The stochastic drops of neurons are akin to a random sampling over different network architectures, thereby reducing noise in the trained parameters.

Finally, L_1 and L_2 regularization losses are added to L with rate hyperparameters of $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. By penalizing the absolute value of the weights in the loss function, L_1 steers weights of neurons toward zero. As a result, the sparsity of the network increases, as unimportant data pathways can be deactivated. L_2 penalizes the sum of the weights' squares, steering them toward smaller values. Smaller weights reduce the complexity of the network and single weight outliers that have high values will be strongly reduced.

In this work, we are surveying only small values of $\eta_{\rm drop}$, L_1, and L_2. Larger values typically cause larger training losses without improving validation errors.

5.1.4. Optimization algorithm

We use the `RMSProp` algorithm for the gradient descent along the loss surface of our deep BANN, which is unlikely to get stuck in a bad local minima (Choromanska et al. 2014; Kawaguchi 2016). With the simple stochastic gradient descent, we encountered problematic overfitting, where the training/validation loss was decreasing/increasing. `RMSProp` keeps a moving average of the square gradients and divides the gradient by the root of this average. This enables a dynamic learning rate along each dimension of the gradient descent, which is not significantly slowed down by gradients from past iterations. To increase the stability of the network training, we also use an adaptive overall learning rate with a peak value of l_r. Following Loshchilov & Hutter (2016), Goyal et al. (2017), and He et al. (2018), we use an initial linear learning rate warm-up from $10 l_r/N_{\rm ep}$ at the first epoch to l_r when 10 % of the epochs have been processed, followed by a cosine decay. The rationale is to have an initially low learning rate for numerical stability of the training of a network with initially random parameters and to smoothly decrease the learning rate when the BANN parameters are converging. The gradient descent iteration through $\widetilde{T}$ is done in small batches of size $N_{\rm b}$.

5.1.5. Performance metrics

We use a mean absolute error metric to track the regression performance of the network during the training for both the training and validation data. For the classification, we measure the fraction of correct predictions over the total number of predictions as the network's accuracy.

We typically report only the average performance metrics for all training and validation $\widetilde{T}$. In principle, it would also be possible to look at the network performance for distinct $\widetilde{M}$ parameter regions. Yet, parameter-dependent network performances will be evident in the $\widetilde{U}$ posteriors, which are the final information of interest; if the inferred parameters are in a region of good/poor network performance, the posteriors will be narrow/wide. In the same vein, the BANN can deal with parameter degeneracies.

5.2. Bootstrapping of known data corruption effects

Within the generic ZINGULARITY bootstrapping framework, we resample uncharacterized EHT visibilities with $\widetilde{C}$ in the following order:

1. Per-antenna polarization leakage terms $\mathcal{D}$ (Event Horizon Telescope Collaboration et al. 2021a).
2. Per-antenna amplitude gain errors. Typically, these are assumed to be $O(10\,\%)$ for all stations. Here, we adopt a more precise description. From the gain (DPFU) measurements of each antenna, we take the determined statistical $\mathcal{G}_{\rm scatter}$ uncertainty from the measurement error into account. Separately, we group the APEX, IRAM 30m, LMT, and SPT stations and add a common $\mathcal{G}_{\rm planet}$ uncertainty based on the accuracy of the model used for the common solar system object (Saturn) that served as a primary calibrator. For all other stations, $\mathcal{G}_{\rm planet}$ is added independently as different primary calibrators were used (Janssen et al. 2019a).
3. Per-antenna gain curve uncertainties. For the fitted gain curves as a function of elevation E, $\mathrm{gc}(B, E_0;\, E) = 1 - B\,(E - E_0)^2$, statistical gc_B and gc_{E0} uncertainties from the fit are taken into account (Janssen et al. 2019a). These are usually ignored, but they can become important for data taken at low elevations.
4. Per-baseline thermal noise. These are taken from the CASA `SIGMA` estimator of the visibilities themselves (see, e.g., Section 2 of Janssen et al. 2019b).

The magnitude of these corruptions is given in Table 2. For each resampled dataset, we draw many times from the network's posterior (a forward pass takes only a few seconds of computational time) to take the combined uncertainties into account and obtain conservative results. We find bootstrapping to be useful for ensuring that the trained BANN does not overfit on data corruption effects.

5.3. Software architecture and scientific reproducibility

We host ZINGULARITY on GITLAB.[5] The software is freely available under the GNU General Public License. All configuration options are set with a single YAML input file. We implemented a deterministic random number seeding for TENSORFLOW in ZINGULARITY.

Using a Continuous Integration / Continuous Delivery setup, with every code change we automatically run unit tests, deploy a containerized DOCKER[6] version of the software, and generate documentation with DOXYGEN.[7] A DOCKER container for every git commit hash is available online.[8]

The results shown here are derived with ZINGULARITY version 1.0.0, which is based on TENSORFLOW version 2.5.0. The DOCKER container `36d816a5d063e673f7502a8aa2eaf4a870431a02`[9] can be used to reproduce the results with the `zingularity-EHT2017` configuration files for Sgr A* and M87* that are stored in the container under `/usr/local/src/zingularity/input_examples`. The synthetic data used for the network training in this work are described in Janssen et al. (2025a).

5.4. Computational efficiency

The TENSORFLOW backend of ZINGULARITY is optimized for CPUs, GPUs, and TPUs. Through the TENSORFLOW reader of `TFRecord` files, ZINGULARITY efficiently shuffles the input randomly, caches and pre-fetches data for multiple training epochs, and splits the features into small batches for the training. These optimizations result in an efficient interplay of GPU/TPU and CPU computations.

Moreover, we shard the `TFRecord` input, which enables parallel read access for the training. This is utilized with ZINGULARITY through an HOROVOD (Sergeev & Del Balso 2018) implementation of distributed computing. The HOROVOD parallelization speedup can be configured with the NVIDIA Collective Communication Library version 2 (NCCL 2) or any proprietary MPI implementation for GPUs that support the `allreduce` or `allgather`, `broadcast`, and `reducescatter` operations. The open-source MPI implementations are usually not as fast and, by default, are used only when NCCL is unavailable. The HOROVOD Tensor Fusion offers an additional computational acceleration by batching together as many tensors, that are queued to be processed, as possible into a single `allreduce` operation.

Distributed computing benchmark tests presented in Sergeev & Del Balso (2018) demonstrated the efficient scalability of HOROVOD over native TENSORFLOW on up to 128 GPUs. As a future outlook, we note the ongoing quantum machine learning developments of our underlying TENSORFLOW framework following the latest quantum hardware advancements (Broughton et al. 2020).

GPU-based correlators offer fast and energy-efficient processing platforms for astronomical interferometers, which scale well for a large number of antennas. Novel correlator designs make use of the tensor core technology of new GPUs, which is yet more efficient (Romein 2021; Yu et al. 2023). It is worth noting that such computing platforms would be perfectly suitable for TENSORFLOW-based applications such as ZINGULARITY during telescope downtime.

Taking one of our BANNs with 12 million trainable parameters as a representative example, we found that a full iteration (training epoch) over 600,000 datasets, each with 21,956 visibility data points, takes 30 seconds on a single NVIDIA A100 GPU. Obtaining 100 posterior samples from 100 bootstrapped $\widetilde{U}$ takes 20 seconds. We have used the containerized version of ZINGULARITY through Apptainer (formerly Singularity, Kurtzer et al. 2017) for these tests.

6. Validation

6.1. Training diagnostics

ZINGULARITY launches the TENSORBOARD visualization and diagnostics toolkit from TENSORFLOW. The dashboard shows the evolution of losses, performance metrics, as well as biases and weights for each layer. Additionally, a graph of the neural network can

[5] Accessible under https://gitlab.com/mjanssen2308/zingularity
[6] https://www.docker.com/
[7] https://www.doxygen.nl/
[8] https://hub.docker.com/r/mjanssen2308/zingularity
[9] https://tinyurl.com/3m49tm7e

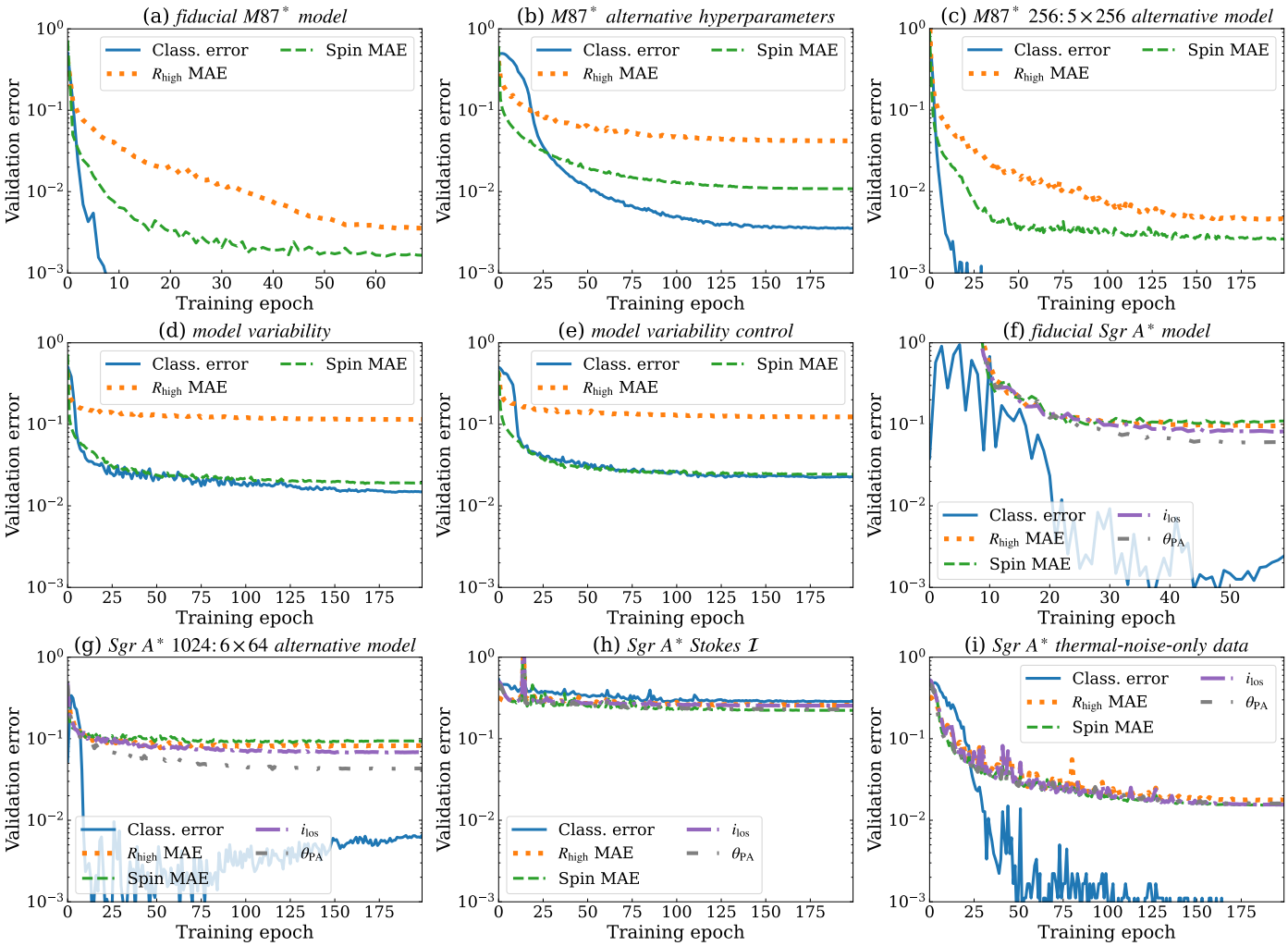

Fig. 4. Performance metrics for the Sgr A* and M87* network training are displayed for various dedicated ZINGULARITY validation tests as described in Section 6. The validation error is computed from normalized labels of validation data not seen by the network during training. The mean absolute error (MAE) is computed as the average over all validation samples for normalized regression labels. The classification error (Class. error) is defined as one minus the network's accuracy, i.e., the fraction of misclassified validation samples. For some studies, the classification errors get numerically close to zero beyond the logarithmic y-axis limit displayed in the figure. Training errors (not shown here) follow the validation curves with a constant (negative) offset, as is typical for neural networks (e.g., Advani & Saxe 2017). For the M87* data, $i_{\rm los}$ and $\theta_{\rm PA}$ are fixed. The training of the fiducial models is shown only up to the determined $N_{\rm ep}$.

be displayed. Together with integrated features such as the WHAT-IF-TOOL (Wexler et al. 2019), TENSORBOARD is an Explainable AI feature, which helps with the identification of salient features in the data and understanding of the choices made by neural networks. We have performed several targeted validation tests with the help of TENSORBOARD data. Unless stated otherwise, we have used the full standard training sets with 600,000 samples for M87* and 252,000 samples for Sgr A*. Figure 4 shows the results of these ZINGULARITY tests:

(a) fiducial M87* model – we show the training performance metrics for our EHT 2017 M87* model described in Section 5.1.

(b) M87* alternative hyperparameters – we show how well our EHT 2017 M87* model performs with a different set of f = ReLU, Ξ = stochastic gradient descent (SGD), $\eta_{\rm drop} = 0, \mathcal{L}_1 = 0, \mathcal{L}_2 = 0.05$ hyperparameters as opposed to those listed in Table 2, using the same $n_{\rm CNNl}: N_{\rm dense} \times n_{\rm dense}$ layout.

(c) M87* 256 : 5 × 256 alternative model – instead of the default ($n_{\rm CNNl} = 128$) : ($N_{\rm dense} = 15$) × ($n_{\rm dense} = 128$) ResNet model, we show how well a 256 : 5 × 256 model performs for the M87* training data in comparison, using the same hyperparameters.

(d) model variability – prompted by intrinsic model variability being the dominating source of noise in our synthetic data (Event Horizon Telescope Collaboration et al. 2019e; Satapathy et al. 2022; Event Horizon Telescope Collaboration et al. 2022e; Janssen et al. 2025a), we have tested if our model overfits to the time-evolution of our training data. We have turned off the random shuffling of M87* training data and used a $\eta_{\rm val} = 0.5$ split across the number of GRMHD-GRRT image frames for each model. The training was done with synthetic data generated from the first 50 % of model frames and validation with the latter 50 %. Here, we have used 450,000 training samples.

(e) model variability control – as a control study and baseline for the model variability, we have done the same as above but with the random shuffling enabled again.

(f) fiducial Sgr A* model – we show the training performance metrics for our EHT 2017 Sgr A* model described in Section 5.1.

(g) Sgr A* 1024 : 6×64 alternative model – we show how well an alternative $n_{\rm CNNl} = 1024, N_{\rm dense} = 6, n_{\rm dense} = 64$ model performs for the Sgr A* training data.

(h) Sgr A* Stokes $\mathcal{I}$ – we show the performance of our fiducial Sgr A* model when trained on only Stokes $\mathcal{I}$ data instead of the full polarization information content from all correlation products. Here, we have used 30,000 training samples.

(i) Sgr A* thermal-noise-only data – instead of doing the full SYMBA forward modeling for $\widetilde{S}$ (Janssen et al. 2025a), we have created 25,000 Sgr A* training samples with only thermal noise added as $\widetilde{C}$.

The training of our fiducial BANNs has converged with low error rates. Most noticeably, the MAD/SANE magnetic states are easily distinguishable for the standard sets of models in our classifiers. The performances of our BANNs are unaffected by small changes in the networks' architectures and hyperparameters. We see smaller validation errors for M87* compared to Sgr A*, which is most likely due to a combination of two effects. Firstly, we have 2.4 times more training data for M87*. Secondly, the intrinsic model variability of Sgr A* within a single training sample translates into additional noise for the parameter inference. For the M87* models, $R_{\rm high}$ shows the highest validation errors, probably because $R_{\rm high}$ has not much influence on the GRRT image morphology for MAD models (Event Horizon Telescope Collaboration et al. 2019d).

The model variability tests show that our network is able to generalize well, as our control study has the same magnitude of validation errors. The robust data features used for the model parameter discrimination are not bound to the particular image snapshot realizations of the time-variable source structure. We note that both the Sgr A* and M87* models show substantial intrinsic time variability, but on different time scales. While Sgr A* varies within EHT observing tracks, M87* allows for a cleaner variability study with a clear cut across single model frames per observing track. Given that our fiducial model performs substantially better than our variability studies, we see that a sufficiently large training dataset of several hundred thousand samples is needed to bring down BANN errors to low levels.

Without the polarization information and with only a few $N_{\rm tr}$, our Sgr A* models perform poorly and are barely being trained at all. The polarization has the biggest impact on the MAD/SANE classification, which is evident from the direct relation to the magnetic field structure. For spin measurements, a recent analysis by Ricarte et al. (2023) has shown the importance of the linear polarization structure.

When only thermal noise is added as $\widetilde{C}$, the validation errors are strongly reduced, even when only a few $N_{\rm tr}$ are used. A neural network trained on synthetic data with lacking noise properties will thus likely overfit on observational data.

6.2. Network hyperparameter survey

Small parameter surveys have been performed to find the fiducial hyper-parameters for M87* and Sgr A* listed in Table 2. We surveyed

- $\eta_{\rm drop} = (0,\ 0.01,\ 0.02)$,
- $\mathcal{L}_1 = (0,\ 0.01,\ 0.02)$,
- $\mathcal{L}_2 = (0,\ 0.01,\ 0.02)$.
- $N_{\rm ep} = (50,\ 60,\ 75,\ 100,\ 200)$.

We ran every parameter combination thrice for each source with different seed values for the BANN initialization at the start of training. We identify networks as viable and stable when the loss is low and the three differently instantiated networks agree within 10 % for all inferred parameters. The parameter inference is computed for many bootstrapping realizations from the observational M87* and Sgr A* $\widetilde{U}$ EHT data plus a few randomly selected validation data samples.

As the bootstrapping $\widetilde{C}$ is already incorporated into $\widetilde{S}$ and by pre-selecting networks with low validation errors, our stability criterion for the validation data is almost always fulfilled by construction. For $\widetilde{U}$, the hurdle for the network to robustly generalize is higher. Typically, we see consistency across hyperparameters and different viable network architectures with some outliers occurring for a single inferred parameter for a particular random seed. We have rejected networks where such stochastic outliers can appear.

6.3. Parameter inferences of test datasets

So far, we have established that our fiducial BANN models give reliable results also on data not seen during training. However, all synthetic observations used so far were based on the same type of KHARMA GRMHD-GRRT models (Wong et al. 2022; Prather et al. 2021). In this section, we describe inferences on test datasets that were obtained from a different kind of simulation model: GRMHD runs from the BHAC code (Porth et al. 2017) ray-traced by RAPTOR (Bronzwaer et al. 2018, 2020). The consistency between our different codes has been established, but we are using different setups and assumptions in our different code libraries. As explored in Event Horizon Telescope Collaboration et al. (2022e, 2024), KHARMA and BHAC models show significant differences in the standard EHT model scoring cuts and different assumptions on adiabatic indices during ray-tracing impact the electron temperature as well as polarimetric quantities. Hence, synthetic data from BHAC-RAPTOR models can be used to test the robustness of our physical parameter inference with respect to nuisance variables in a comparison with the standard EHT model scoring. As an additional check concerning the known issue of GRMHD model variability, we note that the Sgr A* test data model images are sampled with a cadence of 200 s as opposed to the 100 s of our standard models used for training and validation. Synthetic datasets are formed from movies of many frames over the course of a Sgr A* EHT VLBI observing track. The test data can be used to make a final model selection in case the parameter surveys leave multiple equally viable models.

We investigated the following models:

1. M87*, SANE, $a_* = 0$, $R_{\rm high} = 40$.
2. M87*, MAD, $a_* = -0.6$, $R_{\rm high} = 160$; a_* falling outside of the grid of sampled parameters in the models used for the training data.
3. Sgr A*, SANE, $a_* = 0.94$, $R_{\rm high} = 1$, $i_{\rm los} = 70°$.
4. Sgr A*, MAD, $a_* = -0.1$, $R_{\rm high} = 30$, $i_{\rm los} = 40°$; a_*, $R_{\rm high}$, $i_{\rm los}$ falling outside of the training data grid.

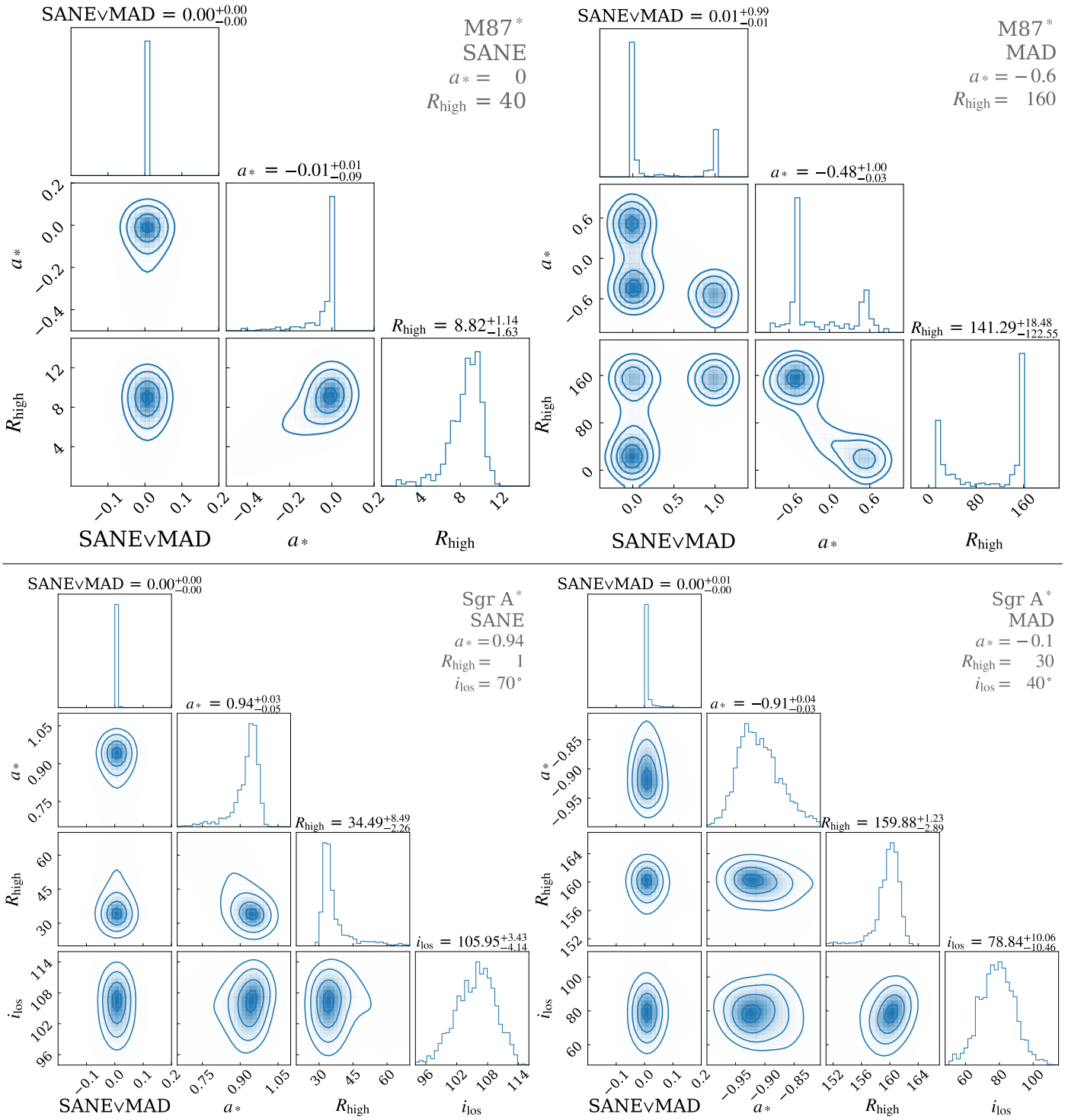

Fig. 5. Inference results on M87* (top row) and Sgr A* (bottom row) test datasets created from differing simulations (described in Section 6.3) with our fiducial BANN. The corner plots give the inferred parameters. Ground truth values are labeled in the top right corners. For the magnetic state, a value of zero corresponds to a certain SANE classification and a value of one to a certain MAD classification.

Parameter inferences are obtained with 1000 bootstrapping realizations times 1000 posterior draws. The test results are shown in Figure 5. The two left panels show models with parameters that fall within the training data grid. We ascribe the small $R_{\rm high}$ and $i_{\rm los}$ errors to the aforementioned differences in ray-tracing.

The M87* $a_* = -0.6$ model shows a multimodal posterior as the network cannot unambiguously associate the data features with the closest matching models from the training data. Even though the ground truth parameters are in the posterior, this test presents a failure mode of the network. The Sgr A* MAD, $a_* = -0.1$, $R_{\rm high} = 30$, $i_{\rm los} = 40°$ model has been misidentified as SANE, $a_* = -0.9$, $R_{\rm high} = 160$, $i_{\rm los} = 80°$ model. The Figure 6 visibility data comparison for three Sgr A* models explains this discrepancy and elucidates the fiducial data features identified by the BANN. We denote the studied synthetic data as follows: A for the SANE, $a_* = -0.94$, $R_{\rm high} = 160$, $i_{\rm los} = 70°$ data, B for

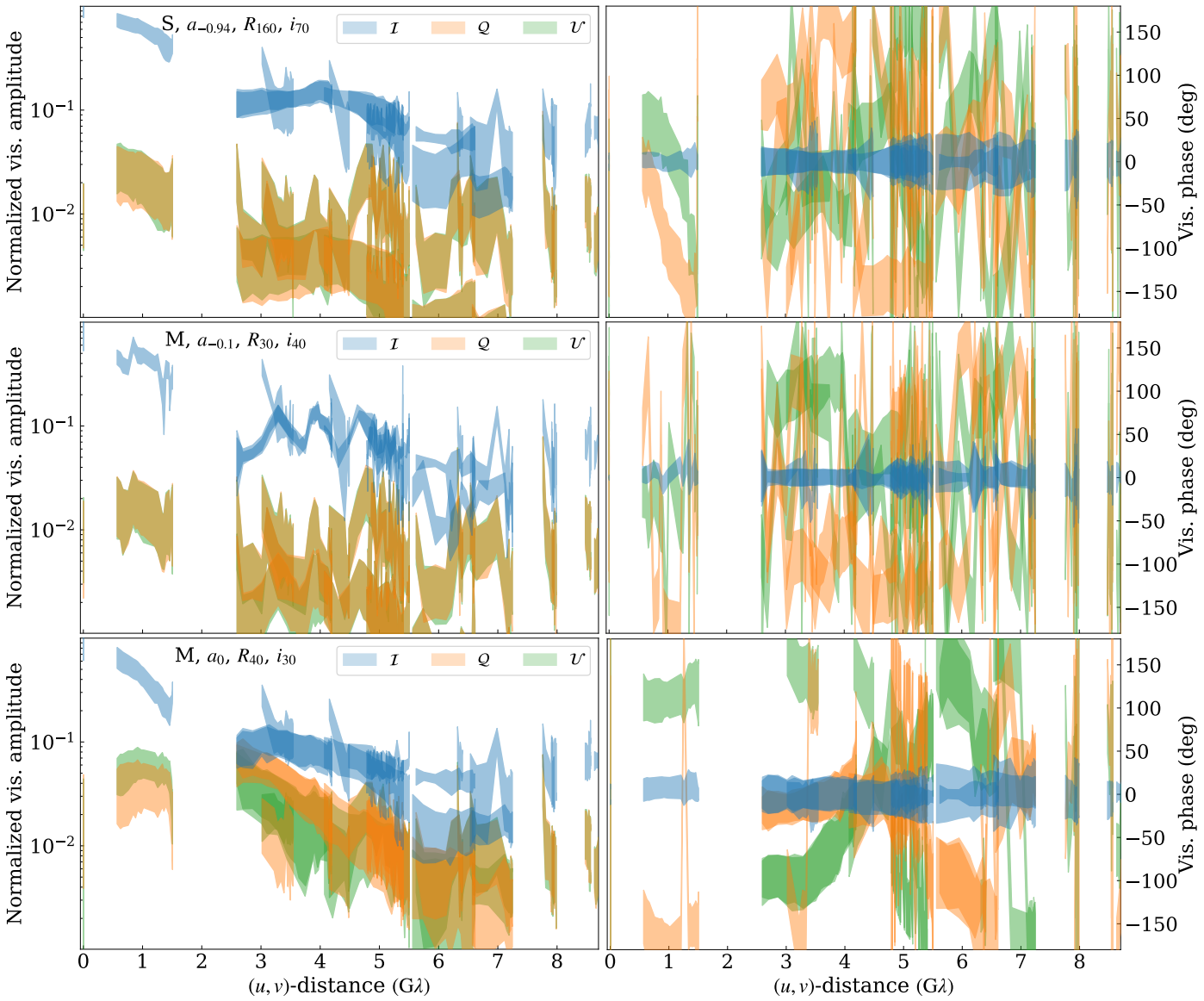

Fig. 6. Normalized visibility amplitudes and phases in degrees (deg) color-coded by $\mathcal{I}$, $\mathcal{Q}$, $\mathcal{U}$ Stokes parameters with standard deviation error bands computed from 100 synthetic data realizations of 2017 Sgr A* EHT observations with corresponding (u, v) coverage are displayed. The parameters for the three models considered here are given in the top left corners of the amplitude plots in the left column: a S(ANE) model in the top row and two different M(AD) models in the middle and bottom rows. The model in the middle is from test data, which has smaller error bars because the underlying simulation was run for a shorter time, leading to a smaller model variability.

MAD, $a_* = -0.1$, $R_{\rm high} = 30$, $i_{\rm los} = 40°$, and C for MAD, $a_* = 0$, $R_{\rm high} = 40$, $i_{\rm los} = 30°$. B is our test data, C is training data with the closest matching parameters to B, and A is training data with the closest matching parameters to the posterior obtained from B (Figure 5). Both A and B have matching $\mathcal{Q}$ and $\mathcal{U}$ amplitudes, while C shows clear offsets. The amplitude differences between $\mathcal{Q}$, $\mathcal{U}$ and $\mathcal{I}$ are also similar for A and B, but much smaller for C. For the phases, $\mathcal{Q}$ and $\mathcal{U}$ are constant at the short baseline around 1 Gλ for C, while A and B display variability. Similarly, the $\mathcal{U}$ and particularly $\mathcal{Q}$ phases vary much more with baseline length for A and B compared to C. All in all, it is not surprising that model A parameters are inferred from B, given the similarity of the visibilities.

Hence, the B misidentification is due to the limited grid of model parameters in the training data, which in turn is a result of the computational cost of GRMHD simulations. This exercise highlights the model-dependence of our method. The inference results point to an interpolated parameter space of the GRMHD-GRRT training data models. Note that the standard EHT model scoring suffers from the same problem, being restricted to the same type of sparsely sampled model libraries. Ideally, we would have a finer model parameter grid to include data like B in $\widetilde{T}$. A better trained BANN could then either be sufficiently complex to be able to distinguish between A and B or the learned similarity between models would show up as uncertainty in the posterior, similar to the multimodal posterior of the M87* $a_* = -0.6$ model.

Through our hyperparameter surveys and test data inferences, we have identified the best BANNs for Sgr A* and M87*. In Janssen et al. (2025b), we will use these fiducial models for inference rather than averaging results over many models that would be acceptable but have worse performance. As opposed to the very small validation errors from the training diagnostics, the posteriors shown in this section give more realistic indications of the true inference uncertainties of our BANNs, which in the end depend on the specific input data $\widetilde{U}$.

7. Summary and conclusions

In this second manuscript from a series, we presented our open-source Zingularity framework for computationally efficient BANN training, validation, and inference. As a first Zingularity application, we used a comprehensive GRMHD-GRRT library of synthetic mm VLBI observations, to study how well physical parameters of the Sgr A* and M87* AGN systems can be constrained with observations of the EHT. We described our end-to-end pipeline, from the theoretical source models and observational data to the final posteriors. For scientific reproducibility, we made use of an open-source workflow management system and containerized versions of our synthetic data generation and machine learning software.

Compared to similar studies in the literature, we included a wider range of theoretical models and ensured that the intrinsic model variability was properly managed. From these models, we created a comprehensive library of mock EHT observations. We considered a wide range of data corruption effects, both in the realistic synthetic data generation and the bootstrapping of uncertainties in the parameter inference process. Furthermore, we circumvented coherence losses and influences of data calibration methods on the measurements. The influence of coherence losses in the EHT data is described in Section 3.3 of Event Horizon Telescope Collaboration et al. (2022a), for example. Here, we could use the full-Stokes information content of visibilities averaged in short (10 s) time bins with small computational costs, thanks to efficient algorithms and data formats within the TensorFlow framework. Influences from the use of particular data reduction strategies are visible in the geometric modeling results of the April 6, 2017 Sgr A* EHT data shown in Figure 30 of Event Horizon Telescope Collaboration et al. (2022c), for example. Depending on the chosen modeling parameters, the fits converged to different values for the ring diameter, asymmetry, and position angle for the same data when it is calibrated with different methods. Here, we incorporated the same, (upgraded, Janssen et al. 2025a) calibration process in the training and observational data, thus ensuring that calibration-specific biases on the visibilities were not erroneously being picked up as model-dependent features.

Through hyperparameter surveys and dedicated inference runs on test datasets, which also include actual observational data to bridge the synthetic gap, we were able to i) weed out BANN architectures that yield spurious results from problematic overfitting on the training data, ii) uncover the inner workings of our networks, and iii) identify shortcomings in the training data sampling and the associated uncertainties on our parameter inferences. We found the training of our final selection of fiducial BANNs to be well converged with low validation errors and robust against variations in the network (hyper-)parameters. We can deal with the intrinsic variability of our models through our networks' ability to generalize well.

We demonstrated the importance of utilizing the full visibility data content for the BANN training – the polarization information is essential for the GRMHD parameter inference, particularly for the MAD–SANE magnetic field configuration. Additionally, showed that a sufficiently large training dataset is needed to achieve low validation errors and that realistic forward modeling is required for the training data generation. When the multitude of additional data corruption processes affecting EHT data are ignored, artificially low errors are obtained. Thus, a network trained on data where only thermal noise is added would be sensitive to data corruption effects present in observational data, leading to incorrect parameter inferences. Shortcomings in previous machine-learning-based EHT analyses can likely be explained by a combination of these three effects: not enough training data, not utilizing the full information content of the data, and not taking into account all relevant data corruption effects.

8. Outlook

The parameter posteriors obtained by fitting the trained BANNs to observational EHT data are presented in Janssen et al. (2025b). Beyond the models studied in this work, one could consider the effects of particle acceleration (e.g., Dexter et al. 2012; Davelaar et al. 2018, 2019, 2020; Yao et al. 2021; Chatterjee et al. 2021; Cruz-Osorio et al. 2022; Fromm et al. 2022; Zhao et al. 2023), alternatives to the $R_{\rm high}$ electron temperature prescription and heating plus cooling effects (e.g., Dibi et al. 2012; Sądowski et al. 2013; Ressler et al. 2015; Sądowski et al. 2017; Chael et al. 2018; Ryan et al. 2018; Chael et al. 2019; Anantua et al. 2020; Yoon et al. 2020; Mizuno et al. 2021; Salas et al. 2025; Mościbrodzka 2025), nonideal MHD and magnetic reconnection (e.g., Ripperda et al. 2019, 2020; Chashkina et al. 2021; Ripperda et al. 2022; Nathanail et al. 2022; Crinquand et al. 2022), pair production (e.g., Mościbrodzka et al. 2011; Crinquand et al. 2020), gas compositions other than pure hydrogen (e.g., Wong & Gammie 2022), tilted accretion disks (e.g., Fragile et al. 2007; McKinney et al. 2013; Morales Teixeira et al. 2014; White et al. 2019; Liska et al. 2019; Chatterjee et al. 2023), different boundary conditions of the accretion flow (e.g., Shcherbakov & Baganoff 2010; Ressler et al. 2018; Olivares et al. 2023), improved approximations for synchrotron radiative transfer calculations (Davelaar 2025), and further non-Kerr models included in more advanced GRMHD-GRRT simulations in the future. Moreover, by employing frequency-resolved data, we could use spectral indices as well as rotation measures as discriminating model features, and would benefit from multifrequency synthesis (Conway et al. 1990). We also note the possibility of training on data from multiple observing days. Finally, the Zingularity application to EHT data described here can easily be extended to AGN jets observed at larger scales (e.g., Fromm et al. 2016, 2019; MacDonald & Nishikawa 2021). Incorporating jet emission and higher energy emission observations will lead to an improved inference, particularly for the electron distribution function.

Acknowledgements. We thank the anonymous referee for their insight and helpful suggestions that have improved this paper. This publication is part of the M2FINDERS project which has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No 101018682). JD is supported by NASA through the NASA Hubble Fellowship grant HST-HF2-51552.001A, awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555. MW is supported by a Ramón y Cajal grant RYC2023-042988-I from the Spanish Ministry of Science and Innovation. This material is based upon work supported by the National Science Foundation under Award Numbers DBI-0735191, DBI-1265383, and DBI-1743442. URL: `www.cyverse.org`. This research was done using resources provided by the Open Science Grid, which is supported by the National Science Foundation award #2030508. This research used the Pegasus Workflow Management Software funded by the National Science Foundation under grant #1664162. Computations were performed on the HPC system Cobra at the Max Planck Computing and Data Facility This research made use of the high-performance computing Raven-GPU cluster of the Max Planck Computing and Data Facility. Corner plots of posteriors were created with `corner.py` (Foreman-Mackey 2016).

References

Abadi, M., Agarwal, A., Barham, P., et al. 2016a, arXiv e-prints, arXiv:1603.04467
Abadi, M., Barham, P., Chen, J., et al. 2016b, arXiv e-prints, arXiv:1605.08695
Advani, M. S. & Saxe, A. M. 2017, arXiv e-prints, arXiv:1710.03667

Aghabiglou, A., Chu, C. S., Dabbech, A., & Wiaux, Y. 2024, ApJS, 273, 3
Anantua, R., Ressler, S., & Quataert, E. 2020, MNRAS, 493, 1404
Balick, B. & Brown, R. L. 1974, ApJ, 194, 265
Bardeen, J. M. 1973, in Black Holes (Les Astres Occlus), 215–239
Baron, D. 2019, arXiv e-prints, arXiv:1904.07248
Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. 2020, Proceedings of the National Academy of Science, 117, 30063
Belkin, M., Hsu, D., Ma, S., & Mandal, S. 2019, Proceedings of the National Academy of Science, 116, 15849
Biretta, J. A., Sparks, W. B., & Macchetto, F. 1999, ApJ, 520, 621
Blackburn, L., Doeleman, S., Dexter, J., et al. 2019, arXiv e-prints (Astro2020 APC White Paper), arXiv:1909.01411
Blackburn, L., Pesce, D. W., Johnson, M. D., et al. 2020, ApJ, 894, 31
Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. 2015, Proceedings of the 32nd International Conference on Machine Learning, 37, 1613
Broderick, A. E., Gold, R., Georgiev, B., et al. 2022, ApJ, 930, L21
Broderick, A. E., Gold, R., Karami, M., et al. 2020, ApJ, 897, 139
Broderick, A. E. & Pesce, D. W. 2020, ApJ, 904, 126
Bronzwaer, T., Davelaar, J., Younsi, Z., et al. 2018, A&A, 613, A2
Bronzwaer, T. & Falcke, H. 2021, ApJ, 920, 155
Bronzwaer, T., Younsi, Z., Davelaar, J., & Falcke, H. 2020, A&A, 641, A126
Broughton, M., Verdon, G., McCourt, T., et al. 2020, arXiv e-prints, arXiv:2003.02989
Byram, E. T., Chubb, T. A., & Friedman, H. 1966, Science, 152, 66
Chael, A., Johnson, M. D., & Lupsasca, A. 2021, ApJ, 918, 6
Chael, A., Lupsasca, A., Wong, G. N., & Quataert, E. 2023, ApJ, 958, 65
Chael, A., Narayan, R., & Johnson, M. D. 2019, MNRAS, 486, 2873
Chael, A., Rowan, M., Narayan, R., Johnson, M., & Sironi, L. 2018, MNRAS, 478, 5209
Chashkina, A., Bromberg, O., & Levinson, A. 2021, MNRAS, 508, 1241
Chatterjee, K., Liska, M., Tchekhovskoy, A., & Markoff, S. 2023, arXiv e-prints, arXiv:2311.00432
Chatterjee, K., Markoff, S., Neilsen, J., et al. 2021, MNRAS, 507, 5281
Chatterji, N. S. & Long, P. M. 2020, arXiv e-prints, arXiv:2004.12019
Choromanska, A., Henaff, M., Mathieu, M., Ben Arous, G., & LeCun, Y. 2014, arXiv e-prints, arXiv:1412.0233
Conroy, N. S., Bauböck, M., Dhruv, V., et al. 2023, ApJ, 951, 46
Conway, J. E., Cornwell, T. J., & Wilkinson, P. N. 1990, MNRAS, 246, 490
Crinquand, B., Cerutti, B., Dubus, G., Parfrey, K., & Philippov, A. 2022, Phys. Rev. Lett., 129, 205101
Crinquand, B., Cerutti, B., Philippov, A., Parfrey, K., & Dubus, G. 2020, Phys. Rev. Lett., 124, 145101
Cruz-Osorio, A., Fromm, C. M., Mizuno, Y., et al. 2022, Nature Astronomy, 6, 103
Curtis, H. D. 1918, Publications of Lick Observatory, 13, 9
Davelaar, J. 2025, ApJ, 978, L10
Davelaar, J., Mościbrodzka, M., Bronzwaer, T., & Falcke, H. 2018, A&A, 612, A34
Davelaar, J., Olivares, H., Porth, O., et al. 2019, A&A, 632, A2
Davelaar, J., Philippov, A. A., Bromberg, O., & Singh, C. B. 2020, ApJ, 896, L31
De Villiers, J.-P. & Hawley, J. F. 2003, ApJ, 592, 1060
Deelman, E., Vahi, K., Juve, G., et al. 2015, Future Generation Computer Systems, 46, 17, funding Acknowledgements: NSF ACI SDCI 0722019, NSF ACI SI2-SSI 1148515 and NSF OCI-1053575
Dewdney, P. E., Hall, P. J., Schilizzi, R. T., & Lazio, T. J. L. W. 2009, IEEE Proceedings, 97, 1482
Dexter, J., McKinney, J. C., & Agol, E. 2012, MNRAS, 421, 1517
Dibi, S., Drappeau, S., Fragile, P. C., Markoff, S., & Dexter, J. 2012, MNRAS, 426, 1928
Dillon, J. V., Langmore, I., Tran, D., et al. 2017, arXiv e-prints, arXiv:1711.10604
Djorgovski, S. G., Mahabal, A. A., Graham, M. J., Polsterer, K., & Krone-Martins, A. 2022, arXiv e-prints, arXiv:2212.01493
Do, T., Hees, A., Ghez, A., et al. 2019, Science, 365, 664
Doeleman, S. S., Barrett, J., Blackburn, L., et al. 2023, Galaxies, 11, 107
Duarte, R., Nemmen, R., & Navarro, J. P. 2022, MNRAS, 512, 5848
EHT MWL Science Working Group, Algaba, J. C., Anczarski, J., et al. 2021, ApJ, 911, L11
Emami, R., Doeleman, S. S., Wielgus, M., et al. 2023a, ApJ, 955, 6
Emami, R., Ricarte, A., Wong, G. N., et al. 2023b, ApJ, 950, 38
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2024, ApJ, 964, L26
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2023, ApJ, 957, L20
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2022a, ApJ, 930, L13
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2022b, ApJ, 930, L14
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2022c, ApJ, 930, L15
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2022d, ApJ, 930, L17
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2019a, ApJ, 875, L2
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2019b, ApJ, 875, L3
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2019c, ApJ, 875, L4
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2019d, ApJ, 875, L5
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2019e, ApJ, 875, L6
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2021a, ApJL, 910, 48
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2021b, ApJ, 875, 5
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2022e, ApJ, 930, L16
Event Horizon Telescope Collaboration, Akiyama, K., Algaba, J. C., et al. 2021c, ApJ, 910, L13
Falcke, H., Melia, F., & Agol, E. 2000, ApJ, 528, L13
Feng, B. T., Bouman, K. L., & Freeman, W. T. 2024, ApJ, 975, 201
Fluke, C. J. & Jacobs, C. 2020, WIREs Data Mining and Knowledge Discovery, 10, e1349
Foreman-Mackey, D. 2016, The Journal of Open Source Software, 1, 24
Fragile, P. C., Blaes, O. M., Anninos, P., & Salmonson, J. D. 2007, ApJ, 668, 417
Fromm, C. M., Cruz-Osorio, A., Mizuno, Y., et al. 2022, A&A, 660, A107
Fromm, C. M., Perucho, M., Mimica, P., & Ros, E. 2016, A&A, 588, A101
Fromm, C. M., Younsi, Z., Baczko, A., et al. 2019, A&A, 629, A4
Gebhardt, K., Adams, J., Richstone, D., et al. 2011, ApJ, 729, 119
Genkin, M. & Engel, T. A. 2020, Nature Machine Intelligence, 2, 674–683
Gheller, C. & Vazza, F. 2022, MNRAS, 509, 990
Glorot, X. & Bengio, Y. 2010, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 9, 249
Goddi, C., Martí-Vidal, I., Messias, H., & et al. 2021, ApJ, 910, L14
Goff, S. A., Vaughn, M., McKay, S., et al. 2011, Frontiers in Plant Science, 2, 34
Goyal, P., Dollár, P., Girshick, R., et al. 2017, arXiv e-prints, arXiv:1706.02677
Gralla, S. E. 2021, Phys. Rev. D, 103, 024023
Graves, A. 2011, Practical Variational Inference for Neural Networks, Vol. 24 (Curran Associates, Inc.)
Gravity Collaboration, Abuter, R., Aimar, N., et al. 2022, A&A, 657, L12
GRAVITY Collaboration, Abuter, R., Amorim, A., et al. 2018, A&A, 615, L15
Gravity Collaboration, Abuter, R., Amorim, A., et al. 2019, A&A, 625, L10
GRAVITY Collaboration, Abuter, R., Amorim, A., et al. 2020, A&A, 636, L5
Hada, K., Doi, A., Kino, M., et al. 2011, Nature, 477, 185
Hahnloser, R. H. R., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., & Seung, H. S. 2000, Nature, 405, 947
Hamaker, J. P., Bregman, J. D., & Sault, R. J. 1996, A&AS, 117, 137
He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv e-prints, arXiv:1512.03385
He, T., Zhang, Z., Zhang, H., et al. 2018, arXiv e-prints, arXiv:1812.01187
Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. 2012, arXiv e-prints, arXiv:1207.0580
Hinton, G. E. & van Camp, D. 1993, Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights, COLT '93 (New York, NY, USA: Association for Computing Machinery), 5–13
Huertas-Company, M. & Lanusse, F. 2023, PASA, 40, e001
Janssen, M., Blackburn, L., Issaoun, S., et al. 2019a, EHT Memo Series, 2019-CE-01 (https://eventhorizontelescope.org/for-astronomers/memos)
Janssen, M., Chan, C.-k., Davelaar, J., et al. 2025a, A&A, appected (Paper I)
Janssen, M., Chan, C.-k., Davelaar, J., et al. 2025b, A&A, appected (Paper III)
Janssen, M., Goddi, C., van Bemmel, I. M., et al. 2019b, A&A, 626, A75
Janssen, M., Radcliffe, J. F., & Wagner, J. 2022, Universe, 8, 527
Jennison, R. C. 1958, MNRAS, 118, 276
Johnson, M. D., Akiyama, K., Blackburn, L., et al. 2023, Galaxies, 11, 61
Johnson, M. D., Lupsasca, A., Strominger, A., et al. 2020, Science Advances, 6, eaaz1310
Kawaguchi, K. 2016, arXiv e-prints, arXiv:1605.07110
Kerr, R. P. 1963, Phys. Rev. Lett., 11, 237
Kim, J., Marrone, D. P., Chan, C.-K., et al. 2016, ApJ, 832, 156
Kocherlakota, P., Rezzolla, L., Falcke, H., et al. 2021, Phys. Rev. D, 103, 104047
Kong, L., Huang, T., Zhu, Y., & Yu, S. 2020, Big Data in Astronomy: Scientific Data Processing for Advanced Radio Telescopes (Elsevier)
Kullback, S. & Leibler, R. A. 1951, The Annals of Mathematical Statistics, 22, 79
Kurth, T., Treichler, S., Romero, J., et al. 2018, arXiv e-prints, arXiv:1810.01993
Kurtzer, G. M., Sochat, V., & Bauer, M. W. 2017, PLoS ONE, 12, e0177459
Lai, S., Thyagarajan, N., Wong, O. I., Diakogiannis, F., & Hoefs, L. 2025, MNRAS, 536, 446
LeCun, Y., Bengio, Y., & Hinton, G. 2015, Nature, 521, 436
LeCun, Y., Boser, B., Denker, J. S., et al. 1989, Neural Computation, 1, 541

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proceedings of the IEEE, 86, 2278
Lei Ba, J., Kiros, J. R., & Hinton, G. E. 2016, arXiv e-prints, arXiv:1607.06450
Levis, A., Chael, A. A., Bouman, K. L., Wielgus, M., & Srinivasan, P. P. 2024, Nature Astronomy, 8, 765
Liska, M., Tchekhovskoy, A., Ingram, A., & van der Klis, M. 2019, MNRAS, 487, 550
Lockhart, W. & Gralla, S. E. 2022, MNRAS, 509, 3643
Loshchilov, I. & Hutter, F. 2016, arXiv e-prints, arXiv:1608.03983
MacDonald, N. R. & Nishikawa, K. I. 2021, A&A, 653, A10
MacKay, D. J. C. 1992, Neural Computation, 4, 448
Marshall, H. L., Miller, B. P., Davis, D. S., et al. 2002, ApJ, 564, 683
McKinney, J. C. 2006, MNRAS, 368, 1561
McKinney, J. C., Tchekhovskoy, A., & Blandford, R. D. 2012, MNRAS, 423, 3083
McKinney, J. C., Tchekhovskoy, A., & Blandford, R. D. 2013, Science, 339, 49
McKinnon, M., Beasley, A., Murphy, E., et al. 2019, in BAAS, Vol. 51, 81
Medeiros, L., Psaltis, D., Lauer, T. R., & Özel, F. 2023a, ApJ, 943, 144
Medeiros, L., Psaltis, D., Lauer, T. R., & Özel, F. 2023b, ApJ, 947, L7
Merchant, N., Lyons, E., Goff, S., et al. 2016, PLOS Biology, 14, e1002342
Mertens, F., Lobanov, A. P., Walker, R. C., & Hardee, P. E. 2016, A&A, 595, A54
Mizuno, Y., Fromm, C. M., Younsi, Z., et al. 2021, MNRAS, 506, 741
Mohan, A., Protopapas, P., Kunnumkai, K., et al. 2024, MNRAS, 527, 10965
Morales Teixeira, D., Fragile, P. C., Zhuravlev, V. V., & Ivanov, P. B. 2014, ApJ, 796, 103
Moriwaki, K., Nishimichi, T., & Yoshida, N. 2023, Reports on Progress in Physics, 86, 076901
Morningstar, W. R., Hezaveh, Y. D., Perreault Levasseur, L., et al. 2018, arXiv e-prints, arXiv:1808.00011
Morningstar, W. R., Perreault Levasseur, L., Hezaveh, Y. D., et al. 2019, ApJ, 883, 14
Mościbrodzka, M. 2025, ApJ, 981, 145
Mościbrodzka, M., Gammie, C. F., Dolence, J. C., & Shiokawa, H. 2011, ApJ, 735, 9
Müller, H. & Lobanov, A. P. 2022, A&A, 666, A137
Müller, H., Mus, A., & Lobanov, A. 2023, A&A, 675, A60
Mus, A. & Martí-Vidal, I. 2024, MNRAS, 528, 5537
Mus, A., Müller, H., & Lobanov, A. 2024a, A&A, 688, A100
Mus, A., Müller, H., Martí-Vidal, I., & Lobanov, A. 2024b, A&A, 684, A55
Narayan, R., Johnson, M. D., & Gammie, C. F. 2019, ApJ, 885, L33
Narayan, R., SÄ dowski, A., Penna, R. F., & Kulkarni, A. K. 2012, MNRAS, 426, 3241
Natarajan, I., Deane, R., van Bemmel, I., et al. 2020, MNRAS, 496, 801
Nathanail, A., Mpisketzis, V., Porth, O., Fromm, C. M., & Rezzolla, L. 2022, MNRAS, 513, 4267
Olivares, H. R., Mościbrodzka, M. A., & Porth, O. 2023, A&A, 678, A141
Owen, F. N., Hardee, P. E., & Cornwell, T. J. 1989, ApJ, 340, 698
Özel, F., Psaltis, D., & Younsi, Z. 2022, ApJ, 941, 88
Palumbo, D. C. M., Gelles, Z., Tiede, P., et al. 2022, ApJ, 939, 107
Palumbo, D. C. M. & Wong, G. N. 2022, ApJ, 929, 49
Palumbo, D. C. M., Wong, G. N., & Prather, B. S. 2020, ApJ, 894, 156
Paugnat, H., Lupsasca, A., Vincent, F. H., & Wielgus, M. 2022, A&A, 668, A11
Pearson, T. J. & Readhead, A. C. S. 1984, ARA&A, 22, 97
Pesce, D. W. 2021, AJ, 161, 178
Popov, A. A., Strokov, V. N., & Surdyaev, A. A. 2021, Astronomy and Computing, 36, 100467
Pordes, R., Petravick, D., Kramer, B., et al. 2007, in 78, Vol. 78, J. Phys. Conf. Ser., 012057
Porth, O., Olivares, H., Mizuno, Y., et al. 2017, Computational Astrophysics and Cosmology, 4, 1
Prather, B., Wong, G., Dhruv, V., et al. 2021, The Journal of Open Source Software, 6, 3336
Psaltis, D., Medeiros, L., Christian, P., et al. 2020, Phys. Rev. Lett., 125, 141104
Psaltis, D., Özel, F., Medeiros, L., et al. 2022, ApJ, 928, 55
Qiu, R., Ricarte, A., Narayan, R., et al. 2023, MNRAS, 520, 4867
Ramachandran, P., Zoph, B., & Le, Q. V. 2017, arXiv e-prints, arXiv:1710.05941
Readhead, A. C. S. & Wilkinson, P. N. 1978, ApJ, 223, 25
Ressler, S. M., Quataert, E., & Stone, J. M. 2018, MNRAS, 478, 3544
Ressler, S. M., Tchekhovskoy, A., Quataert, E., Chandra, M., & Gammie, C. F. 2015, MNRAS, 454, 1848
Ricarte, A., Tiede, P., Emami, R., Tamar, A., & Natarajan, P. 2023, Galaxies, 11, 6
Ripperda, B., Bacchini, F., & Philippov, A. A. 2020, ApJ, 900, 100
Ripperda, B., Liska, M., Chatterjee, K., et al. 2022, ApJ, 924, L32
Ripperda, B., Porth, O., Sironi, L., & Keppens, R. 2019, MNRAS, 485, 299
Roelofs, F., Janssen, M., Natarajan, I., et al. 2020, A&A, 636, A5
Romein, J. W. 2021, A&A, 656, A52
Ryan, B. R., Ressler, S. M., Dolence, J. C., Gammie, C., & Quataert, E. 2018, ApJ, 864, 126
Salas, L. D. S., Liska, M. T. P., Markoff, S. B., et al. 2025, MNRAS, 538, 698
SaraerToosi, A. & Broderick, A. E. 2024, ApJ, 967, 140
Satapathy, K., Psaltis, D., Özel, F., et al. 2022, ApJ, 925, 13
Schaeffer, R., Khona, M., Robertson, Z., et al. 2023, arXiv e-prints, arXiv:2303.14151
Schmidt, K., Geyer, F., Fröse, S., et al. 2022, A&A, 664, A134
Sergeev, A. & Del Balso, M. 2018, arXiv e-prints, arXiv:1802.05799
Sfiligoi, I., Bradley, D. C., Holzman, B., et al. 2009, in 2, Vol. 2, 2009 WRI World Congress on Computer Science and Information Engineering, 428–432
Shatskiy, A. A. & Evgeniev, I. Y. 2019, Soviet Journal of Experimental and Theoretical Physics, 128, 592
Shcherbakov, R. V. & Baganoff, F. K. 2010, ApJ, 716, 504
Sądowski, A., Narayan, R., Tchekhovskoy, A., & Zhu, Y. 2013, MNRAS, 429, 3533
Sądowski, A., Wielgus, M., Narayan, R., et al. 2017, MNRAS, 466, 705
Smith, M. J. & Geach, J. E. 2023, Royal Society Open Science, 10, 221454
Sparks, W. B., Biretta, J. A., & Macchetto, F. 1996, ApJ, 473, 254
Stokes, G. G. 1851, Transactions of the Cambridge Philosophical Society, 9, 399
Sun, H. & Bouman, K. L. 2020, arXiv e-prints, arXiv:2010.14462
Sun, H., Bouman, K. L., Tiede, P., et al. 2022, ApJ, 932, 99
Sun, H., Dalca, A. V., & Bouman, K. L. 2020, arXiv e-prints, arXiv:2003.10424
The Event Horizon Telescope Collaboration. 2024, arXiv e-prints, arXiv:2410.02986
Thompson, A. R., Moran, J. M., & Swenson, George W., J. 2017, Interferometry and Synthesis in Radio Astronomy, 3rd Edition (Springer)
Thyagarajan, N., Hoefs, L., & Wong, O. I. 2024, RAS Techniques and Instruments, 3, 437
Tiede, P. 2022, The Journal of Open Source Software, 7, 4457
Valentin Jospin, L., Buntine, W., Boussaid, F., Laga, H., & Bennamoun, M. 2020, ACM Comput. Surv., 1, arXiv:2007.06823
van Cittert, P. H. 1934, Physica, 1, 201
van der Gucht, J., Davelaar, J., Hendriks, L., et al. 2020, A&A, 636, A94
VanderPlas, J., Connolly, A. J., Ivezic, Z., & Gray, A. 2012, in Proceedings of Conference on Intelligent Data Understanding (CIDU, 47–54
Vincent, F. H., Gralla, S. E., Lupsasca, A., & Wielgus, M. 2022, A&A, 667, A170
Walker, R. C., Hardee, P. E., Davies, F. B., Ly, C., & Junor, W. 2018, ApJ, 855, 128
Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., & Fergus, R. 2013, Proceedings of the 30th International Conference on Machine Learning, 28, 1058
Wexler, J., Pushkarna, M., Bolukbasi, T., et al. 2019, arXiv e-prints, arXiv:1907.04135
White, C. J., Quataert, E., & Blaes, O. 2019, ApJ, 878, 51
Wielgus, M. 2021, Phys. Rev. D, 104, 124058
Wielgus, M., Marchili, N., Martí-Vidal, I., et al. 2022, ApJ, 930, L19
Witzel, G., Martinez, G., Willner, S. P., et al. 2021, ApJ, 917, 73
Wong, G. N. & Gammie, C. F. 2022, ApJ, 937, 60
Wong, G. N., Prather, B. S., Dhruv, V., et al. 2022, ApJS, 259, 64
Yao, P. Z., Dexter, J., Chen, A. Y., Ryan, B. R., & Wong, G. N. 2021, MNRAS, 507, 4864
Yao-Yu Lin, J., Pesce, D. W., Wong, G. N., et al. 2021, arXiv e-prints, arXiv:2110.07185
Yao-Yu Lin, J., Wong, G. N., Prather, B. S., & Gammie, C. F. 2020, arXiv e-prints, arXiv:2007.00794
Yfantis, A. I., Zhao, S., Gold, R., Mościbrodzka, M., & Broderick, A. E. 2024, MNRAS, 535, 3181
Yoon, D., Chatterjee, K., Markoff, S. B., et al. 2020, MNRAS, 499, 3178
Yu, W., Romein, J. W., Dursi, L. J., et al. 2023, Galaxies, 11, 13
Zernike, F. 1938, Physica, 5, 785
Zhao, S.-S., Huang, L., Lu, R.-S., & Shen, Z. 2023, MNRAS, 519, 340

Appendix A: Theory of a supervised Bayesian feedforward neural network

In a feedforward ANN, information flows only in one direction, from the first, to the second, to the third until the final output layer. Here, we give a brief mathematical description of such networks. We make use of the following naming conventions:

- M are the total numbers of layers in the network. Individual layers are numbered as $m = 0, 1, 2, ..., M-1$. Each layer consists of a number of neurons.
- X_m is the output data of layer $m-1$ and input data of layer m. $X_0 \in \widetilde{T}$ is the input data of the network (i.e., a tensor of features that the network trains on). X_M are the output predictions of the network. We denote the output of a specific neuron j in the mth layer with $x_m^{(j)}$.
- Y are the labeled targets of the training data $\widetilde{T}$ that will be compared against X_M for supervised learning. These are integer representations when X_0 belongs to specific classes that are to be identified and/or real numbers for regression tasks of continuous variables of interest belonging to X_0.
- L is the loss function used to compute the error E between the target Y and predicted X_M outputs as $E = L(Y, X_M)$. Together with a learning rate l_r, this error is used to update (train) the weights of the network.
- f_m are activation functions used for layer m. These functions have to be nonlinear if the network is to learn nonlinear behavior in the input data X_m. In recent years, the rectified linear unit $f_m(z) = \max(z, 0)$ (ReLU, Hahnloser et al. 2000) has frequently been used.

Depending on the dimensionality of the data as it is passed through the network, X_m can be a highly dimensional tensor. With the above conventions, the output of layer m of an ANN can be written as

$$X_{m+1} = f_m\left(W_m \cdot X_m + B_m\right) . \tag{A.1}$$

Here, W_m and B_m are the tunable weight and bias tensors of layer m that are being fitted while the network is being trained. The weights are multiplied with the output of the previous layer and therefore depend on the dimensionality of X_m and the specified output dimension $D_m \equiv \dim(X_{m+1})$. Single bias terms are added to the output of each neuron. The dimensionality of B therefore depends only on D_m. We denote $W_m \cdot X_m + B_m$ as the weighted input Z_m of the layer m. The input to a specific neuron j will be written as $z_m^{(j)}$.

We can now write the ANN algorithm in functional form as

$$\mathcal{F}(\{W\}, \{B\}; X_0) = f_{M-1}\left(W_{M-1} \cdot f_{M-2}\left(W_{M-2} \cdot f_{M-3}\left(\cdots f_1\left(W_1 \cdot f_0\left(W_0 \cdot X_0 + B_0\right) + B_1\right) + ... + B_{M-3}\right) + B_{M-2}\right) + B_{M-1}\right) .$$

Here, $\{W\}$ and $\{B\}$ denote the sets of all weights and biases that we will be optimizing, respectively. We denote individual weights of the connection between the k^{th} neuron in the $(m-1)^{\text{th}}$ layer to the j^{th} neuron in the m^{th} layer as $w_m^{(jk)}$. Similarly, $b_m^{(j)}$ will be the bias of the j^{th} neuron in the m^{th} layer. Two commonly used types of layers, are 1) fully connected layers, where each neuron j in the layer is connected to each neuron k in the previous layer and 2) convolution layers (LeCun et al. 1989, 1998), which can be seen as having a small receptive field where most entries of W are zero: each neuron j is only connected to a few neurons in the previous layer. Convolution layers form the basis of convolutional neural networks (CNNs).

For multidimensional prediction problems, individual parallel output layers are commonly used for each individual regression and classification task. All of these output layers are connected to the same second to last layer in a feedforward ANN. Typically, linear activation functions are used for regression and the softmax function σ_s is used for classification. Given a vector $\mathbf{s} = (s_1, s_2, ..., s_{N_c})$ for N_c different classes, $\sigma_s : \mathbb{R}^{N_c} \to [0, 1]^{N_c}$ computes the probability of each class as

$$\sigma_s(\mathbf{s}_i) = \frac{e^{s_i}}{\sum_{j=1}^{N_c} e^{s_j}} \quad \text{for} \quad i = 1, 2, ..., N_c . \tag{A.2}$$

For multidimensional prediction problems, the target labels Y have to be normalized to ensure an equal weighting for all predictions in the loss L.

Below, we describe the backpropagation algorithm, which is widely used to update $\left\{w_m^{(jk)}\right\}$ and $\left\{b_m^{(j)}\right\}$ during the training of ANNs.

Initially, $\{W\}$ and $\{B\}$ are set based on some a priori assumptions or randomly. In a forward pass through the network, $\mathcal{F}$ is computed to determine $L(Y, X_M)$.

For a neuron j of the last layer, we can compute

$$\delta_{M-1}^{(j)} = \frac{\partial L}{\partial x_{M-1}^{(j)}} f'_{M-1}\left(z_{M-1}^{(j)}\right) . \tag{A.3}$$

Here, we use the Lagrange prime ($'$) notation for a derivative. Subsequently, we can form the equivalent quantity for the complete last layer:

$$\delta_{M-1} = \nabla_X(L) \cdot f'_{M-1}\left(Z_{M-1}\right) . \tag{A.4}$$

Starting with the last layer, we can propagate backwards through the network layer by layer to compute

$$\delta_m = f'_m\left(Z_m\right) \cdot W_{m+1} \cdot \delta_{m+1} \tag{A.5}$$

for $m = M-2, M-3, ..., 0$. The individual $\delta_m^{(j)}$ are computed in the same way. Using a learning rate $l_r \leq 1$ to control the step size, each individual weight and bias parameters are then updated with

$$l_r \frac{\partial L}{\partial w_m^{(jk)}} = l_r x_{m-1}^{(k)} \delta_m^{(j)} , \tag{A.6}$$

$$l_r \frac{\partial L}{\partial b_m^{(j)}} = l_r \delta_m^{(j)} . \tag{A.7}$$

In practice, a stochastic gradient descent is used for computational speed. Instead of computing $\mathcal{F}$ and L for the full dataset, we train on batches of size N_b at a time.

Appendix A.1: Bayesian variational inference

Starting with Bayes' theorem, the probability P as a measure of belief for a set of neural network parameters H given data X can be written as

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} = \frac{P(X|H)P(H)}{\int_H P(X|h)P(h)\mathrm{d}h} \,. \tag{A.8}$$

Here, $P(X|H)$, $P(H)$, $P(X)$, $P(H|X)$ are the likelihood, prior, evidence, and posterior, respectively. The difficulty in sampling the posterior arises from the computational cost of evaluating the evidence integral. Following Valentin Jospin et al. (2020), we will now describe the variational inference method that is implemented in the Bayesian ANN (BANN) framework Zingularity and used to sample from a surrogate variational distribution $q_\varphi(H) \sim P(H|X)$ instead of the exact posterior.

For non-Bayesian ANNs, $w_m^{(jk)}$ and $b_m^{(j)}$ are trained as single numbers. As such, $\mathcal{F}$ provides point estimates for inference and forward passes during training. Bayesian networks fit trainable distributions for $w_m^{(jk)}$ and $b_m^{(j)}$ (MacKay 1992; Hinton & van Camp 1993; Graves 2011; Blundell et al. 2015). For inference and forward passes, values are then drawn stochastically from the weight and bias posteriors. The variational distribution is given by the combined distributions of all layers, parameterized by φ. During training, φ are optimized by minimizing the Kullback–Leibler divergence D_{KL} (Kullback & Leibler 1951), such that $q_\varphi(H)$ approximates $P(H|X)$ as closely as possible:

$$D_{\mathrm{KL}}(q_\varphi||P) = \int_H q_\varphi(h) \log\left(\frac{q_\varphi(h)}{P(h|X)}\right)\mathrm{d}h \,. \tag{A.9}$$

Computing D_{KL} directly is expensive due to the integral over $P(h|X)$. Instead the evidence lower bound $\mathcal{E}$ is commonly evaluated:

$$\mathcal{E}_X(q_\varphi||P) \equiv \int_H q_\varphi(h) \log\left(\frac{P(h,X)}{q_\varphi(h)}\right)\mathrm{d}h \tag{A.10}$$

$$= \log\left(P(X)\right) - D_{\mathrm{KL}}(q_\varphi||P) \,. \tag{A.11}$$

Here, $P(h, X)$ is the joint probability of h and X. As the evidence $P(X)$ does not depend on q, maximizing $\mathcal{E}_X(q_\varphi||P)$ is equivalent to minimizing $D_{\mathrm{KL}}(q_\varphi||P)$.