Variational Autoencoder with Normalizing flow for X-ray spectral fitting

Fiona Redmen
Department of Physics & Astronomy
University of Southampton
Southampton, SO17 1BJ, UK
fr1n21@soton.ac.uk

Ethan Tregidga
Laboratoire d'astrophysique
EPFL
Observatoire de Sauverny, 1290 Versoix, Switzerland
ethan.tregidga@epfl.ch

James Steiner
Harvard-Smithsonian Center for Astrophysics
Cambridge, MA 02138, USA
james.steiner@cfa.harvard.edu

Cecilia Garraffo
Harvard-Smithsonian Center for Astrophysics
Cambridge, MA 02138, USA
cgarraffo@cfa.harvard.edu

Abstract

Black hole X-ray binaries (BHBs) can be studied with spectral fitting to provide physical constraints on accretion in extreme gravitational environments. Traditional methods of spectral fitting such as Markov Chain Monte Carlo (MCMC) face limitations due to computational times. We introduce a probabilistic model, utilizing a variational autoencoder with a normalizing flow, trained to adopt a physical latent space. This neural network produces predictions for spectral-model parameters as well as their full probability distributions. Our implementations result in a significant improvement in spectral reconstructions over a previous deterministic model while performing three orders of magnitude faster than traditional methods.

1 Introduction

Black hole X-ray binaries (BHBs) are composed of a close companion star orbiting, and accreting matter, onto a stellar-mass black hole. The accretion flows in these systems consist of a geometrically-thin and optically-thick disk enshrouded by a corona of hot electrons [14]. These components emit predominantly in the X-rays, and their resulting spectra can be fit to infer physical properties of the system. Such inference provides key insights into the accretion mechanisms at work [4], enabling exploration of General Relativity in extreme environments [7]. The standard analytical software for X-ray astronomy, XSPEC [1], uses Markov Chain Monte Carlo (MCMC) to obtain full posterior distributions on physical parameters. This method of spectral fitting faces limitations due to unwieldy computational times when dealing with a large number of spectra, especially when employing complex models.

This work advances a new approach, expediting the traditional MCMC method. We use a library of observational data collected for several dozen BHBs with the Neutron-star Interior Composition Explorer (NICER) mission on the International Space Station [8], and additionally employ synthetic NICER spectra. Both cover the $0.3 - 10$ keV energy range with 240 spectral bins, and from such data we derive full posterior distributions of five physical parameters, along with spectral reconstructions.

Recent machine learning (ML) work for X-ray astronomy includes representation learning [21], with applications in anomaly detection, unsupervised classification, and similarity searches [3]. Previous ML work in spectral fitting has successfully applied autoencoders (AEs) to this problem, which

achieved parameter-value inference $2\,700$ times faster than traditional techniques [20]. However, that study did not provide uncertainties on the parameter predictions, which is key for scientific interpretability. Moreover, the method required additional iterations of computationally-expensive spectral fitting via XSPEC for parameter predictions to remain comparable in accuracy. The works in Barret & Dupourqué [2] and Dupourqué & Barret [5] discuss the application of simulation-based inference with automatic posterior transformation [9] to this problem and accurately obtain full posteriors.

We present an implementation of the network described in Tregidga et al. [20] that outputs full posterior distributions. Our posterior predictions reproduce spectra with higher fidelity than before, with similar accuracy to traditional MCMC methods, removing the requirement of additional iterations of XSPEC fitting. The architecture takes X-ray spectra as input and outputs the corresponding physical parameter distributions. In addition to maximizing posteriors as in Barret & Dupourqué [2], we include a decoder, a reconstruction loss, and a latent loss to further modulate training. We demonstrate the robustness of our method when applied to real spectra. The code, which utilizes the repository in Tregidga et al. [20], is available at `https://github.com/fi-redmen/fspnet-var.git`.

2 Method

2.1 Datasets

We aggregate the full library of BHB spectra collected by NICER from July 2017 through August 2022, yielding a total of $10\,800$ spectra, from 25 BHBs. Data is preprocessed as in Tregidga et al. [20]. Ground-truth benchmarks, which we call our targets, are obtained as a reference via rigorous fitting of spectra, performed using XSPEC. Synthetic spectra are created using XSPEC, simulating the physical model with the detector response accounted for, but excluding noise. A total of $100\,000$ spectra are generated, spanning a broad distribution of viable physical parameters. We split the data to train on 80% of the spectra and use the remaining 20% as a validation set.

2.2 Physical Model

In this exploratory work, we utilize a highly-simplified spectral model which incorporates three spectral features. Thermal emission from the inner thin disk is described by a multicolor black-body continuum [23]. The Compton up-scattering of thermal-disk's photons by hot electrons in the corona redistributes a portion of those photons to higher energies, manifesting as a power law [18]. Photoelectric absorption by gas along our line of sight to the source cuts off the spectrum at low energies ($\lesssim 1$ keV) [22]. Through this model, we can determine the following physical properties: temperature of the thin disk kT_{disk}, the amplitude of the thin-disk emission N, photon index for the Comptonized emission Γ, the fraction of photons that are Compton up-scattered f_{sc}, and the line-of-sight column density of neutral gas N_{H}.

2.3 Architecture

The network consists of three elements: an encoder, a normalizing flow (NF) [15], and a decoder. We use the network and multi-stage training scheme as in Tregidga et al. [20], as this has been shown to effectively perform deterministic spectral parameter predictions. The encoder contains a series of convolutional layers which downscale the dimensionality of the spectra [17] and learn patterns in spectra across neighboring data points through various scales. A series of linear layers then allows the encoder to map from the spectral space to the parameter space. The output of the encoder is a partially reduced dimension representation of the spectra, referred to as "context". The NF is a neural spline flow [6], parameterized by an autoregressive network [13], which outputs a series of ten monotonic rational-quadratic spline transforms [6], conditioned on the context. The transforms describe the predicted posteriors as a map from a standard normal distribution to the posterior. Draws are sampled from the posteriors to form our latent space, and serve as input to the decoder. The decoder follows a similar but reverse-ordered structure to the encoder; a series of linear layers map the parameter space to spectral space and a bidirectional gated recurrent unit [16] is used to learn sequential relationships from lower to higher energies and vice versa. We use GELU activation functions throughout the network. The architecture is displayed in Figure 1. For each spectrum, a five-dimensional distribution is predicted, with each dimension representing a different parameter. Using a full NF rather than a
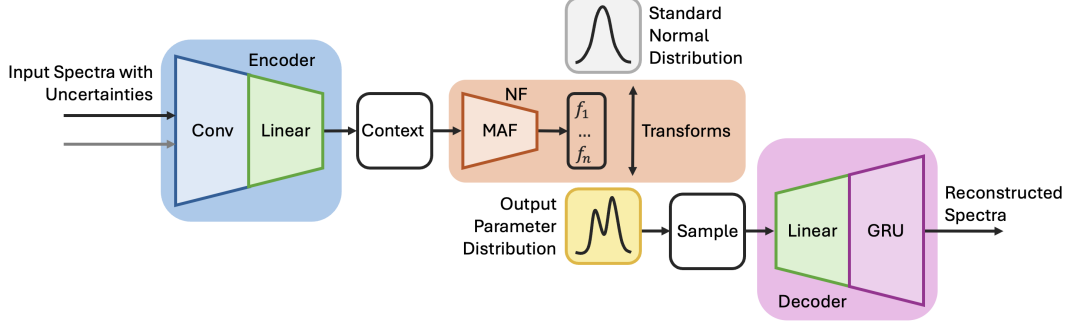
Figure 1: Illustration of AE with NF architecture.

vanilla VAE produces complex posteriors which more closely resemble the true spectral parameter distributions.

We use a Gaussian negative log likelihood [12] reconstruction loss, as this prioritizes data points with lower relative errors. Since our goal is to obtain physical parameters, rather than an arbitrary decomposition of the spectra, we impose a physical latent space. To do this, we use an additional latent loss term which forces the latent values to match with physical quantities, penalizing differences as the mean square error. To learn the correct distributions, we include a third flow loss term that maximizes the log probability at the values seen in the dataset with corresponding spectra. The flow loss is mathematically described as

$$\mathcal{L}_{\mathrm{NF}}(\boldsymbol{\phi}) = -\log q_{\boldsymbol{\phi}}(\boldsymbol{\theta}|\boldsymbol{x}) \tag{1}$$

where $q_{\boldsymbol{\phi}}$ describes the predicted probability distribution of parameters $\boldsymbol{\theta}$ given spectra $\boldsymbol{x}$ and $\boldsymbol{\phi}$ represents the network parameters.

The novel technique of implementing a NF within an AE leverages the decoder by informing parameter predictions on degeneracies. For example, the two parameters N and kT both increase with overall spectral flux. Joint over-predictions or under-predictions in N and kT would return inaccurate spectra, penalized strongly through the reconstruction loss.

We use the AdamW optimizer[10] with an initial learning rate of 1×10^{-3} which is reduced to 1×10^{-6} through a reduce on plateau scheduler. The training is split into three stages. First, the decoder is trained on synthetic data, to learn the physically informed mapping between parameters and spectra. The weights of the decoder are then frozen and the network is trained, end to end, on synthetic data. Finally, the network is trained on real data, with the decoder still frozen, and an initial learning rate of 1×10^{-4}. Transfer learning from synthetic to real data allows us to maximize our training samples, while adapting to real data [19]. Training is halted when a plateau is detected across the last 30 epochs and capped at 400 epochs. The three stages are trained for 400, 121 and 216 epochs, respectively. The first stage utilizes only the reconstruction loss, while the last two stages use all three terms with equal weightings. Experimentation shows that reducing the reconstruction loss weight by 1×10^{-3} or removing the latent and reconstruction losses very slightly worsens performance. Longer training of the last stage leads to overfitting past roughly 250 epochs.

3 Evaluation

Figure 2 shows predicted against target values for a subset of 250 real spectra. We plot 50 values sampled from corresponding posteriors to capture the distributions. The data is colored by the total count rate. The Pearson Correlation Coefficient (PCC) between the predicted and target values for each parameter is printed. Best linear regressions are performed over each of the 50 sets of samples and the mean and standard deviation are taken. We overlay that linear fit and an an idealized line (slope of 1). There is a strong correlation across all parameters, with a minimum PCC of 0.842 shown for Γ. Notably, the spectrum is only weakly affected by Γ when either f_{sc} is low or when kT is very large owing to NICER's bandpass. The worst linear fit (slope of 0.877 ± 0.059) is seen for N.
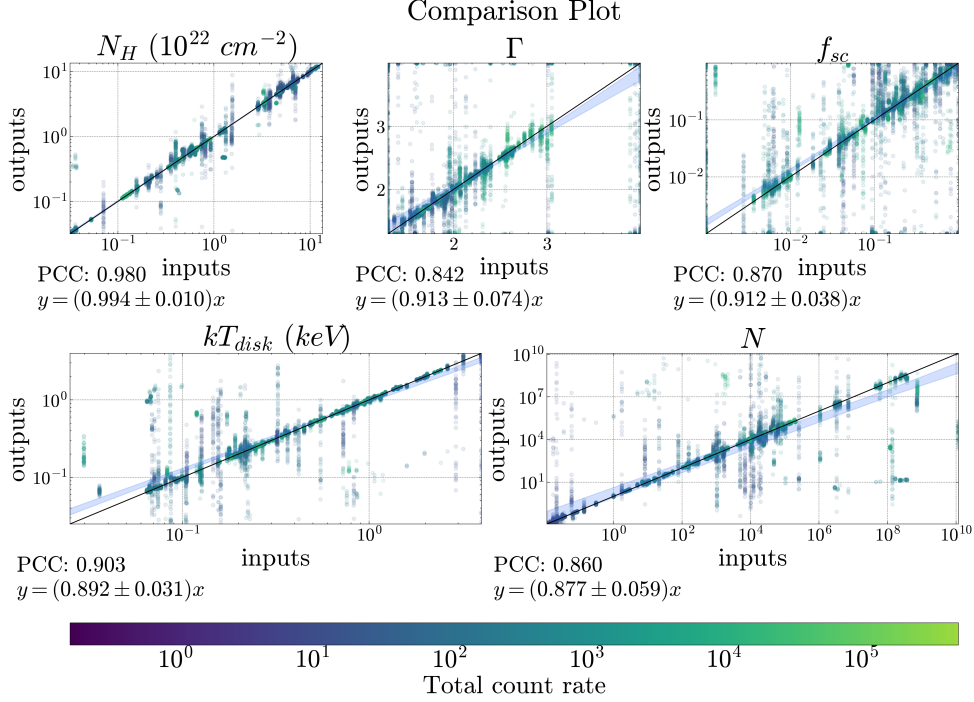
Figure 2: 50 samples from output posteriors plotted against corresponding input values for a total of 250 spectra, and colored by total spectral count rate. PCC and best linear fits are included.

The reduced "PGStat" [11] is a fit statistic for Poisson-distributed data with Gaussian-distributed background. An ideal fit would correspond to a reduced PGStat of 1 and increases with worse fits. We compute the reduced PGStats of reconstructions produced by inputting our parameters to the physical model in XSPEC. We take the median value across the set of 2 160 spectra in our validation data, for robustness against outliers. The first reduced PGStat value, used as a reference point, is calculated using our target parameters. The second and third values are computed with parameters obtained through XSPEC fitting and the NF respectively. Ensuring these values match each other, we allow for a direct comparison in computation time for each method to achieve similar accuracy. XSPEC uses 130 iterations of fitting to fulfill this requirement, obtaining one effective posterior sample. The final value, computed using the same network but without the decoder−trained on synthetic and real data for 121 and 216 epochs respectively−demonstrates the impact of including the decoder. To obtain multiple samples in XSPEC, we run an MCMC chain.[1] The NF samples are obtained in one forward pass of the network, irrespective of sample size. Computation times were measured over predictions run on an Apple M2 chip.

We show reduced PGStats, along with computational times for parameter predictions over this set of spectra in Table 1. Due to the oversimplifications in our model, we do not expect to see a reduced PGStat of 1 from our pre-calculated targets. The reduced PGStat values resulting from the NF and 130 fit iterations with XSPEC are both comparable to the reference value. Therefore, the parameters determined by our NF are able to reconstruct spectra through the physical model with similar accuracy to traditional methods, despite the deviations shown from target values in Figure 2. Computing one sampled parameter prediction is ~ 640 times faster than XSPEC fitting. Computing 1000 samples for full posterior evaluation is ~ 2000 times faster than using MCMC chains. Training the encoder and NF by itself results in worse XSPEC reconstructions, indicating that the decoder does inform the physical parameter predictions. The reduced PGStat resulting from the Tregidga et al. [20] model

---

[1]We estimate the computation time for 1000 effective samples across the 2 160 spectra in our validation set by running a representative MCMC chain for one spectrum. We approximate an autocorrelation length, τ, and use $t_{\text{total}} = \frac{(N_{\text{samples}}\tau + b)t}{L+b}N_{\text{val}}$, where L is the overall chain length, b is the burn in, t is the representative chain run time, N_{samples} is the number of effective samples we aim to obtain, and N_{val} is our validation set size.

Table 1: Reduced PGStats and computation times

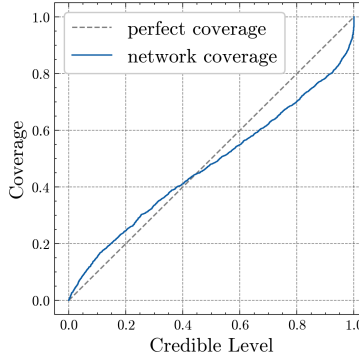| Scenario	Computation time (s)	Reduced PGStat
Pre-calculated targets (for reference)	N/A	3.801
XSPEC	1 sample 1279 ± 7 1000 samples: $\sim 100\,000\,1$	3.804
NF in AE	1 sample: 2.11 ± 0.05 1000 samples: 51.8 ± 0.9	3.796
NF	Same as NF in AE	6.034

Figure 3: Plot of coverage versus credible level for distributions predicted from real spectra (blue, solid). The ideal coverage is also plotted (grey, dashed).

is 62.7 with a baseline value of 4.44, a factor of ~ 14 worse than both traditional methods and the model we present.

Fig. 3 is a coverage plot for distributions predicted from real data. At low credible levels, the network is slightly under-confident, while at high credible levels it is slightly over confident.

4   Conclusion

This work describes the implementation of a NF in an AE network to infer full posterior predictions of physical parameters from BHB spectra. We demonstrate the robustness of this method by performing evaluations on real data predictions. The NF implementation obtains accurate parameter predictions, with most inaccuracies and imprecisions attributable to low count rates in spectra. While remaining comparable in accuracy, posterior predictions are ~ 640 times faster than 130 iterations of XSPEC fitting. This computational speed-up is extended dramatically to a factor of ~ 2000 when comparing to MCMC posterior estimation in XSPEC.

This method can already be applied to perform population studies across the NICER archive, however, further training optimization may improve performance, including more thorough investigations into loss function weightings. Currently, our target parameters are determined using an oversimplified model in XSPEC, limiting the accuracy of network predictions versus true physical values. Future work could apply the network on more complex, physically accurate models, which also contain more information on physics within the systems.

Acknowledgments and Disclosure of Funding

References

[1] Arnaud, K. A. 1996, in Astronomical Society of the Pacific Conference Series, Vol. 101, Astronomical Data Analysis Software and Systems V, ed. G. H. Jacoby & J. Barnes, 17

[2] Barret, D., & Dupourqué, S. 2024, 686, A133, doi: `10.1051/0004-6361/202449214`

[3] Dillmann, S., Martínez-Galarza, J. R., Soria, R., Stefano, R. D., & Kashyap, V. L. 2025, Monthly Notices of the Royal Astronomical Society, 537, 931, doi: `10.1093/mnras/stae2808`

[4] Done, C., Gierliński, M., & Kubota, A. 2007, The Astronomy and Astrophysics Review, 15, 1, doi: `10.1007/s00159-007-0006-1`

[5] Dupourqué, S., & Barret, D. 2025, doi: `10.1051/0004-6361/202555215`

[6] Durkan, C., Bekasov, A., Murray, I., & Papamakarios, G. 2019, Neural Spline Flows, arXiv, doi: `10.48550/arXiv.1906.04032`

[7] Fabian, A. C., Iwasawa, K., Reynolds, C. S., & Young, A. J. 2000, Publications of the Astronomical Society of the Pacific, 112, 1145, doi: `10.1086/316610`

[8] Gendreau, K. C., Arzoumanian, Z., & Okajima, T. 2012, in Space Telescopes and Instrumentation 2012: Ultraviolet to Gamma Ray, ed. T. Takahashi, S. S. Murray, & J.-W. A. d. Herder, Vol. 8443 (SPIE), 844313, doi: `10.1117/12.926396`

[9] Greenberg, D. S., Nonnenmacher, M., & Macke, J. H. 2019, Automatic Posterior Transformation for Likelihood-Free Inference, arXiv, doi: `10.48550/arXiv.1905.07488`

[10] Loshchilov, I., & Hutter, F. 2019, Decoupled Weight Decay Regularization. `https://arxiv.org/abs/1711.05101`

[11] NASA. n.d., Appendix B: Statistics in XSPEC. `https://heasarc.gsfc.nasa.gov/xanadu/xspec/manual/XSappendixStatistics.html`

[12] Nix, D., & Weigend, A. 1994, in Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94), Vol. 1, 55–60 vol.1, doi: `10.1109/ICNN.1994.374138`

[13] Papamakarios, G., Pavlakou, T., & Murray, I. 2018, Masked Autoregressive Flow for Density Estimation, arXiv, doi: `10.48550/arXiv.1705.07057`

[14] Remillard, R. A., & McClintock, J. E. 2006, Annual Review of Astronomy and Astrophysics, 44, 49, doi: `10.1146/annurev.astro.44.051905.092532`

[15] Rezende, D. J., & Mohamed, S. 2016, Variational Inference with Normalizing Flows. `https://arxiv.org/abs/1505.05770`

[16] Schuster, M., & Paliwal, K. 1997, IEEE Transactions on Signal Processing, 45, 2673, doi: `10.1109/78.650093`

[17] Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. 2015, Striving for Simplicity: The All Convolutional Net. `https://arxiv.org/abs/1412.6806`

[18] Steiner, J. F., García, J. A., Eikmann, W., et al. 2017, The Astrophysical Journal, 836, 119, doi: `10.3847/1538-4357/836/1/119`

[19] Tan, C., Sun, F., Kong, T., et al. 2018, A Survey on Deep Transfer Learning, arXiv, doi: `10.48550/arXiv.1808.01974`

[20] Tregidga, E., Steiner, J., Garraffo, C., Rhea, C., & Aubin, M. 2024, Monthly Notices of the Royal Astronomical Society, 529, 1654, doi: `10.1093/mnras/stae629`

[21] Vago, N. O. P., Martínez-Galarza, J. R., & Amato, R. 2025, Extracting Latent Representations from X-ray Spectra. Classification, Regression, and Accretion Signatures of Chandra Sources, arXiv, doi: `10.48550/arXiv.2510.14102`

[22] Wilms, J., Allen, A., & McCray, R. 2000, The Astrophysical Journal, 542, 914, doi: `10.1086/317016`

[23] Zimmerman, E. R., Narayan, R., McClintock, J. E., & Miller, J. M. 2005, The Astrophysical Journal, 618, 832, doi: `10.1086/426071`