Error Analysis of Kilonova Surrogate Models

Kamilė Lukošiūtė
Gravitation Astroparticle Physics Amsterdam (GRAPPA)
Institute for Theoretical Physics Amsterdam
University of Amsterdam, The Netherlands
k.lukosiute@uva.nl

Geert Raaijmakers
Gravitation Astroparticle Physics Amsterdam (GRAPPA)
Institute for Theoretical Physics Amsterdam
University of Amsterdam, The Netherlands

Zoheyr Doctor
Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA)
Northwestern University
1800 Sherman Ave, Evanston, IL 60201, USA

Marcelle Soares-Santos
Department of Phyiscs
University of Michigan
Ann Arbor, MI , USA

Brian Nord
Fermi National Accelerator Laboratory
Kavli Institute for Cosmological Physics,
Department of Astronomy and Astrophysics,
University of Chicago

Abstract

Studies of kilonovae, optical counterparts of binary neutron star mergers, rely on accurate simulation models. The most accurate simulations are computationally expensive; surrogate modelling provides a route to emulate the original simulations and therefore use them for statistical inference. We present a new implementation of surrogate construction using conditional variational autoencoders (cVAE) and discuss the challenges of this method. We additionally present model evaluation methods tailored to the scientific analyses of this field. We find that the cVAE surrogate produces errors well within a standard assumed systematic modelling uncertainty. We also report the results of our parameter inference study, finding our constrained parameters to be comparable with previously published results.

1 Introduction

Binary neutron star (BNS) mergers are an important testbed for an expansive range of phenomena — from particle physics and astrophysics (e.g., neutron star equation of state (EOS) [e.g. 29]) to cosmology (e.g. the value of the Hubble constant [1]). A BNS merger releases gravitational waves (GWs) [21] and neutron-star material [25, 26], powering a kilonova [23] that emits ultraviolet-optical-infrared light [7, 8, 22, 31, 27, 4, 10, 14, 33–35, 38, 2, 15, 32, 3].

To constrain fundamental physics parameters, statistical inference techniques (from traditional Bayesian inference to deep learning) require forward models of BNS mergers and the observable signals they produce. The most accurate forward models of kilonovae and GWs require prohibitively

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

computationally expensive radiative transfer simulations [e.g. 5, 16, 17]. Surrogate modeling provides a route to more efficiently emulate the output of simulations. For example, Gaussian process regression (GPR) has been used in kilonova/GW studies [13, 9, 12]. GPR is unfortunately prohibitively computationally expensive for high-dimensional data sets. Regardless of the type of forward model, it is critical that we fully characterize its accuracy and precision to propagate into statistical inference pipelines.

However, there exist few to no standardized methods for evaluating errors for surrogates in the physical sciences. In this work, we present a surrogate model constructed from conditional variational autoencoders (cVAEs)[19, 30, 36] for kilonova simulations. We quantify accuracy and precision in the raw and derived simulation products of these simulations — i.e., spectra and lightcurves, respective; we also evaluate the performance of the surrogate in a scientific case study.

2 Data

We focus on the set of BNS kilonova spectra first published in [12] because it is publicly available, but our methods can be applied to other kilonova simulation sets[1]. These simulations assume a non-spherical geometry and have four input parameters: mass of the dynamical ejecta $M_{ej,dyn}$, the mass of the post merger ejecta $M_{ej,pm}$, the half opening angle of the lanthanide-rich tidal dynamical ejecta Φ, where above and below the ejecta is lanthanide-free, and $\cos\theta_{obs}$, the cosine of the observer viewing angle. A spectra is calculated in increments of 0.2 days, up to 20 days post-merger, and each spectra is computed at 500 wavelength bins, evenly spaced from 100 Å to 99900 Å. We treat time as a fourth parameter and the input vector becomes $\mathbf{x} = \{M_{ej,dyn}, M_{ej,pm}, \Phi, \cos\theta_{obs}, t\}$. Each output $\mathbf{y}$ is a single spectrum i.e. a vector in $\mathbb{R}^{500}$. In total there are 215600 training examples.

3 Methods

3.1 Challenges of Surrogate Modelling via cVAEs

We employ a standard conditional variational autoencoder (cVAE) for learning a mapping from the input parameters $\mathbf{x}$ to the output spectrum $\mathbf{y}$, where the goal is to learn the distribution $p(\mathbf{y}|\mathbf{x})$ [36, 20]. Theoretically, cVAEs can be used as to generate complex conditional data distributions by choosing an appropriate likelihood function. The Gaussian likelihood would be the appropriate choice in our case, as all of our output parameters are continuous and real. Unfortunately, the maximum likelihood objective is ill-posed for continuous deep latent variable models models, such as those employing Gaussian distributions [24]. Intuitively, if a model can express an optimal set of encoder and decoder parameters ϕ and θ such that the reconstruction $\mu_{\theta,i}(z)$ is very close to the target y_i, then the $-\frac{1}{2}\log 2\pi\sigma^2_{\theta,i}(z)$ term of the Gaussian likelihood will push the variance to zero before the $[2\sigma^2_{\theta,i}(z)]^{-1}$ term can catch up. Therefore, a Gaussian VAE will produce a meaninglessly small variance. For a rigorous discussion of the issue, called "Variance Shrinkage," see [24] and [11].

This means that any cVAE employing a Gaussian likelihood is uninterpretable as a probabilistic model and therefore does not accurately model $p(\mathbf{y}|\mathbf{x})$ when the Gaussian likelihood is used. There are several standard procedures to avoid variance shrinkage [11]. First, it is possible to set a global variance to a constant globally, i.e. $\sigma^2 = 1$, with the consequence that the log-likelihood simply becomes a mean squared error term. This also means that the variance is still uninterpretable. Another standard procedure is to use a Bernoulli distribution, even when the data is not binary. In practice, optimizing a Bernoulli negative log likelihood is considerably easier than using a mean squared error [11]. Nevertheless the output remains uninterpretable as a distribution over the data and errors must still be approximated through other means. We take the approach of building a cVAE-based surrogate model and developing a set of procedures for estimating errors.

3.2 Training

We separate each set into training, validation, and test sets (a split of approx. 80% / 10 % / 10 %). We use the validation scores to choose the values of the hyperparameters: the dimensionality of $\mathbf{z}$

[1] The full dataset details can be found at `https://github.com/mbulla/kilonova_models/tree/master/bns_m3_3comp`

and the dimensionality of the hidden layer of the decoder and encoder. All other hyperparameters of the model, such as the learning rate and batch size, are fixed across all of our tests and models. The final chosen model has a hidden layer size of 1000 neurons and a latent layer size of 20. After choosing a final model setup, we create nine different training/validation data splits and train nine models on each of the data splits. We then evaluate all nine models on the test set. The nine models give us information on the sensitivity of the surrogate model to data re-sampling. We choose one of the models at random for final results.

We use Adam as our gradient-based optimizer [18] and PyTorch for implementation [28]. To optimize for practical ease of training, we use the binary cross entropy loss function with a final sigmoid activation function and transform both the input and output variables into range $[0, 1]$. We make the final model deterministic for final evaluation and production use by always using the same sample for $\mathbf{z}$ in prediction. We train all models for 200 epochs; the final training loss, averaged over the nine data split models is 11.23. The hyperparameter search required approximately 24 hours of training time on a single Nvidia GeForce 1080Ti GPU. The final nine experiment models required approximately 26 hours of training on the same GPU.

3.3 Model Evaluation

Error Sources and Quantifying Data Stochasticity Because a cVAE does not produce a probability distribution, we rely on standard error estimation methods to assess model performance. Our chosen errors are informed by our study of potential sources of uncertainty inherent in the data set. First, there is systematic error due to assumptions of the original simulations; this error will propagate to the surrogate model. Quantifying this error requires theoretical studies of kilonovae, so we do not discuss it further here. Second, a surrogate model will always have some systematic error due to its inherent data-compression nature. Users of surrogate models accept the systematic error because of the speed of prediction with such models.

Lastly, there is statistical error arising from the Monte Carlo noise of radiative transfer simulations [5, 16]. No surrogate model will be able to perfectly predict the noise within the unseen test set. Computing Monte Carlo noise from simulations involves running independent simulations several times and computing residuals from the mean spectrum [6]. Performing this is expensive and we do not have access to the simulator. We create models for mean spectra by Gaussian smoothing using several values for the standard deviation of the Gaussian kernel. Taking all the sets of Gaussian smoothed spectra, we compute the fractional residuals from the simulated data and use this as an estimate of the fractional Monte Carlo noise. In Section 4, we compare our estimated fractional Monte Carlo noise with the errors of the surrogate predictions.

Our proposed error evaluation methods for kilonova surrogate models are the following:

Errors in Raw Spectra The fractional error between the model prediction in a given wavelength bin $y_{\mathrm{pred},\lambda}$ and the value of the original test data point $y_{\mathrm{test},\lambda}$ is the "spectral error": $\epsilon_{s,\lambda} = y_{\mathrm{pred},\lambda}/y_{\mathrm{test},\lambda} - 1$. We use the mean and median of ϵ_s across all wavelengths to search for a systematic bias, and we use the median of $|\epsilon_s|$ to asses the scale of the errors.

Error in Observables Imaging telescopes observe the flux of photons through a wideband filter. Band observations from simulated spectra are computed by taking the log of the flux at a chosen distance (40 Mpc in this work) and integrating over the wavelengths of a given band (i.e., the $ugrizy$ band set). We compute the error in the band magnitude between the predicted magnitudes and the test magnitudes: $\Delta m = m_{\mathrm{pred,band}} - m_{\mathrm{test,band}}$.

Scientific Use Case We evaluate the performance of the surrogate model by using it within a representative scientific task. We perform a nested sampling fit using the `dynesty` [37] sampler on the GW170817 lightcurve data using our surrogate model. We compare the differences in the best fit parameters between our fit using the same dataset, which was first collected in [9] and previously published fits for the same BNS kilonova model but using a different surrogate construction method.

4 Results and Discussion

The value of the mean and median spectral errors (taken over the entire test dataset) are 6620±3180 and 0.067±0.020, where the range refers to the standard deviation across all nine experiments of the

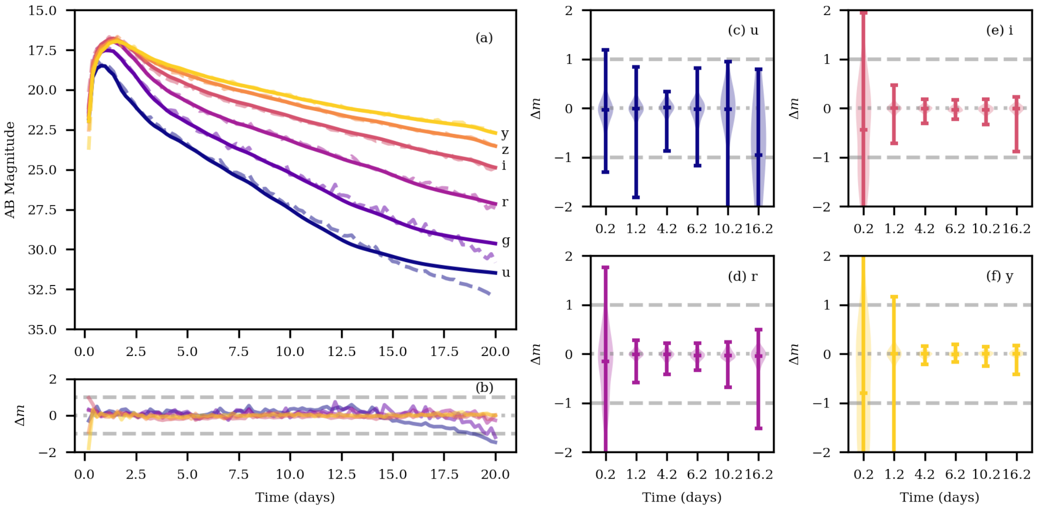

Figure 1: (a) An example true lightcurve (brightness versus time) constructed from the test dataset (solid lines) in the $ugrizy$ bands and its cVAE-based reconstruction (dashed lines) at the same input parameters. The input parameters are $M_{ej,dyn} = 0.01$, $M_{ej,pm} = 0.09$, $\Phi = 30$, $\cos\theta_{obs} = 0.3$. (b) The difference between the cVAE prediction and the example in (a). (c,d,e,f) Distributions of Δm for some of the bands and at six different time steps: the distribution is shown across the entire test set. The three small, horizontal lines for each distribution are the maximum, median, and mean.

final hyperparameter setup. The mean is very large because it is sensitive to outliers, especially when the test set value is very close to 0. Because the mean is larger than the median, the distribution of predictions is skewed high with positive outliers. The value of the median of $|\epsilon_s|$ is 0.285±0.004, where the range is the standard deviation across all nine experiments. This value is correlated with wavelength; across wavelengths, the curve of the median of $|\epsilon_s|$ taken across the dataset is seemingly bounded by our estimates of the Monte Carlo noise. We show this in the Appendix. A viable step for future work would be to compute the Monte Carlo noise directly by simulating several times and compare with surrogate model outputs.

The values of the median across the test set of the absolute value of Δm for each band are $u : 0.277 \pm 0.016, g : 0.161 \pm 0.014, r : 0.090 \pm 0.027, i : 0.071 \pm 0.019, z : 0.053 \pm 0.019$, and $z : 0.049 \pm 0.006$ where again the median is taken over every example in the test set and all nine experiments and the reported range is the standard deviation for the experiments. A common value quoted in magnitude for systematic modelling uncertainty is 1 magnitude [9]; the errors produced by the cVAE surrogate are mostly well below this, as also seen in Figures 1c-1f, where we show the distributions of Δm for the bands *u, r, i* and *y*. The surrogate model performs worst when the lightcurves are dim, which can be seen by the growing distributions of Δm with time. $t = 0.2$ is also unreliably predicted, perhaps due to simulation modelling uncertainty. Figure 1a shows an example for a test lightcurve and the prediction via cVAE; Figure 1b shows the lightcurve's Δm. To see an example of a direct prediction of a spectrum, see the Appendix.

Lastly, we report the results of our nested sampling fit for GW170817. We use the flat priors of $\log_{10}(0.001) \leq \log_{10}(M_{ej,dyn}/M_\odot) \leq \log_{10}(0.02)$, $\log_{10}(0.01) \leq \log_{10}(M_{ej,pm}/M_\odot) \leq \log_{10}(0.13)$, $0 \leq \cos(\Phi) \leq 1$, and $0 \leq \cos(\theta) \leq 1$, extending over the entire published data range. We find the best fit parameters of $\log_{10}(M_{ej,dyn}/M_\odot) = -2.31^{+0.21}_{-0.22}$,$\log_{10}(M_{ej,pm}/M_\odot) = -1.13^{+0.11}_{-0.21}$, $\Phi = 47.65^{+19.98}_{-12.89}$, and $\theta_{obs} = 64.46^{+17.12}_{-22.02}$ where the errors refer to the 1σ confidence levels. The parameter inference required 5 minutes to be completed on one Intel® Core™ i7-7700HQ CPU. We aim to compare with the fit previously published in [12], but we cannot do so directly because their fit is only one part of a multi-step analysis and therefore involves different priors for the parameters. Nevertheless, their median values are within 1σ levels of our median values, and the widths of their confidence intervals are comparable to ours (see [12]).

5 Conclusion

We discussed that, while the cVAE is fast, it cannot be used as a surrogate model that returns probabilities over the output data directly. However, we also showed that the cVAE is still useful for mapping from a lower dimensional space to a higher dimensional space. We then constructed and used a suite of metrics and evaluations to assess the surrogate model error performance in a scientific case study. We presented an application for a specific kilonova data set, a specific surrogate construction method, and a prescription for error analysis that can be used in any study to assess the performance of fast forward models.

Broader Impact

Inaccurate error characterization can lead to inaccurate scientific conclusions, and therefore careful error characterization is a fundamental part of responsible science. With this work, we aim to encourage others, including those outside the astrophysics community to consider and quantify the errors that come from assumptions, such as surrogate modelling, carefully. By communicating the importance of error quantification, we aim to avoid potential negative impacts from irresponsibly published science.

Acknowledgments and Disclosure of Funding

We thank Samaya Nissanke for useful discussions and comments.

GR and KL are grateful for financial support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) through the Projectruimte and VIDI grants (Nissanke).

We thank SURFsara (www.surfsara.nl) for the support in using the Lisa Compute Cluster.

Work supported by the Fermi National Accelerator Laboratory, managed and operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

We acknowledge the Deep Skies Lab as a community of multi-domain experts and collaborators who've facilitated an environment of open discussion, idea-generation, and collaboration. This community was important for the development of this project.

References

[1] Abbott B. P., et al., 2017a, Nature, 551, 85

[2] Abbott B. P., et al., 2017b, The Astrophysical Journal, 848, L12

[3] Abbott B. P., et al., 2017c, The Astrophysical Journal, 848, L13

[4] Arcavi I., et al., 2017, Nature, 551, 64

[5] Bulla M., 2019, Monthly Notices of the Royal Astronomical Society, 489, 5037

[6] Bulla M., Sim S. A., Kromer M., 2015, Monthly Notices of the Royal Astronomical Society, 450, 967

[7] Burbidge E. M., Burbidge G. R., Fowler W. A., Hoyle F., 1957, Reviews of Modern Physics, 29, 547

[8] Cameron A. G. W., 1957, Publications of the Astronomical Society of the Pacific, 69, 201

[9] Coughlin M. W., et al., 2018, Monthly Notices of the Royal Astronomical Society, 480, 3871

[10] Coulter D. A., et al., 2017, Science, 358, 1556

[11] Detlefsen N. S., Jørgensen M., Hauberg S., 2019, arXiv:1906.03260 [cs, stat]

[12] Dietrich T., Coughlin M. W., Pang P. T. H., Bulla M., Heinzel J., Issa L., Tews I., Antier S., 2020, Science, 370, 1450

[13] Doctor Z., Farr B., Holz D. E., Pürrer M., 2017, Physical Review D, 96, 123011

[14] Drout M. R., et al., 2017, Science, 358, 1570

[15] Goldstein D. A., et al., 2019, The Astrophysical Journal, 881, L7

[16] Kasen D., Metzger B., Barnes J., Quataert E., Ramirez-Ruiz E., 2017, Nature, 551, 80

[17] Kawaguchi K., Kyutoku K., Shibata M., Tanaka M., 2016, The Astrophysical Journal, 825, 52

[18] Kingma D. P., Ba J., 2017, arXiv:1412.6980 [cs]

[19] Kingma D. P., Welling M., 2014, arXiv:1312.6114 [cs, stat]

[20] Kingma D. P., Welling M., 2019, Foundations and Trends® in Machine Learning, 12, 307

[21] LIGO Scientific Collaboration and Virgo Collaboration et al., 2017, Physical Review Letters, 119, 161101

[22] Lattimer J. M., Schramm D. N., 1974, The Astrophysical Journal Letters, 192, L145

[23] Li L.-X., Paczyński B., 1998, The Astrophysical Journal Letters, 507, L59

[24] Mattei P.-A., Frellsen J., 2018, arXiv:1802.04826 [cs, stat]

[25] Metzger B. D., 2019, Living Reviews in Relativity, 23, 1

[26] Metzger B. D., Berger E., 2012, , 746, 48

[27] Metzger B. D., et al., 2010, Monthly Notices of the Royal Astronomical Society, 406, 2650

[28] Paszke A., et al., 2019, arXiv:1912.01703 [cs, stat]

[29] Radice D., Perego A., Hotokezaka K., Fromm S. A., Bernuzzi S., Roberts L. F., 2018, The Astrophysical Journal, 869, 130

[30] Rezende D. J., Mohamed S., Wierstra D., 2014, arXiv:1401.4082 [cs, stat]

[31] Rosswog S., Liebendoerfer M., Thielemann F.-K., Davies M. B., Benz W., Piran T., 1998, arXiv:astro-ph/9811367

[32] Savchenko V., et al., 2017, The Astrophysical Journal, 848, L15

[33] Shappee B. J., et al., 2017, Science, 358, 1574

[34] Smartt S. J., et al., 2017, Nature, 551, 75

[35] Soares-Santos M., et al., 2017, The Astrophysical Journal, 848, L16

[36] Sohn K., Yan X., Lee H., 2015, in Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. NIPS'15. MIT Press, Cambridge, MA, USA, pp 3483–3491

[37] Speagle J. S., 2020, Monthly Notices of the Royal Astronomical Society, 493, 3132

[38] Valenti S., et al., 2017, The Astrophysical Journal, 848, L24

A Appendix

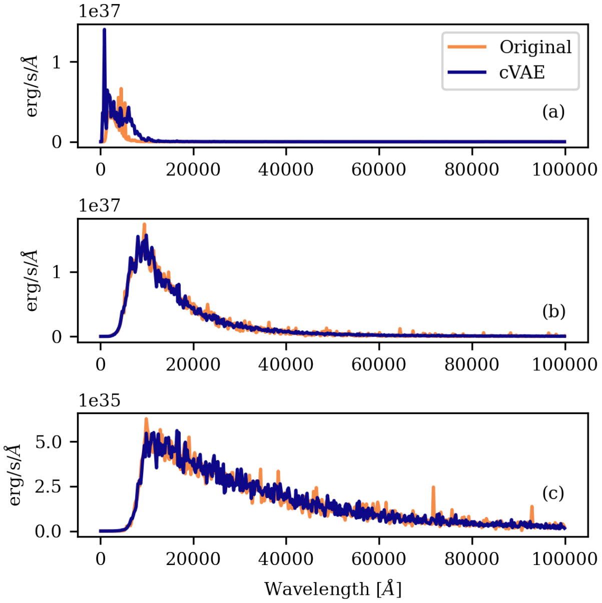

Figure 2: Three original spectra (orange) and corresponding cVAE predictions (blue) for physical parameters $M_{ej,dyn}/M_{\odot} = 0.02$, $M_{ej,pm}/M_{\odot} = 0.05$, $\Phi = 45.0°$, $\cos\,\Theta_{obs} = 0.8$, and times (a) 0.2 days, (b) 4.2 days, and (c) 14.2 days. The corresponding median spectral errors across the whole spectra (a) 12.91 , (b) 0.37, and (c) 0.20.

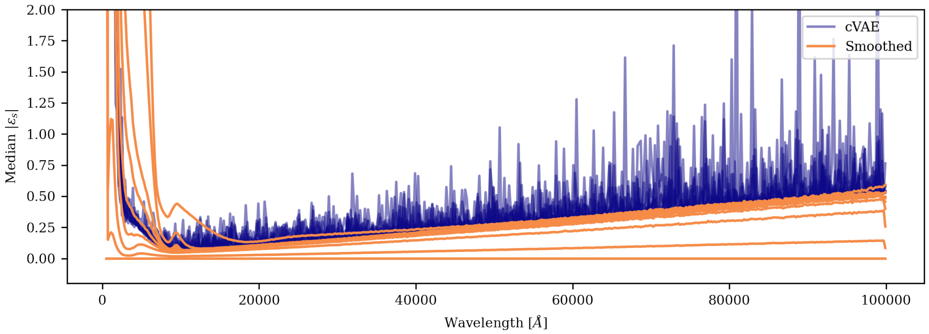

Figure 3: Median of the absolute spectral error $|\epsilon_s|$, where the median is taken over all the spectra in the test data set, as a function of wavelength for the predictions of the cVAE (blue), along with the absolute spectral error from the Gaussian smoothed spectra for nine different settings of the Gaussian kernel parameter. Starting from the flat line and moving upwards, these error lines correspond to the Gaussian smoothed spectra using sigma parameters of 0.1, 0.5, 1,2,3,4,5,10, and 30.