Interpretable Uncertainty Quantification in AI for HEP

Thomas Y. Chen[1], Biprateep Dey[2], Aishik Ghosh[3,4], Michael Kagan[5], Brian Nord[6,7,8], and Nesar Ramachandra[9]

[1]Fu Foundation School of Engineering and Applied Science, Columbia University, New York, NY 10027, USA
[2]Department of Physics and Astronomy and PITT PACC, University of Pittsburgh, Pittsburgh, PA 15260, USA
[3]Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA
[4]Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[5]SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA
[6]Fermi National Accelerator Laboratory, P.O. Box 500, Batavia, IL 60510, USA
[7]Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA
[8]Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA
[9]Computational Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439, USA

August 10, 2022

Abstract

Estimating uncertainty is at the core of performing scientific measurements in HEP: a measurement is not useful without an estimate of its uncertainty. The goal of uncertainty quantification (UQ) is inextricably linked to the question, "how do we physically and statistically interpret these uncertainties?" The answer to this question depends not only on the computational task we aim to undertake, but also on the methods we use for that task. For artificial intelligence (AI) applications in HEP, there are several areas where interpretable methods for UQ are essential, including inference, simulation, and control/decision-making. There exist some methods for each of these areas, but they have not yet been demonstrated to be as trustworthy as more traditional approaches currently employed in physics (e.g., non-AI frequentist and Bayesian methods).

Shedding light on the questions above requires additional understanding of the interplay of AI systems and uncertainty quantification. We briefly discuss the existing methods in each area and relate them to tasks across HEP. We then discuss recommendations for avenues to pursue to develop the necessary techniques for reliable widespread usage of AI with UQ over the next decade.

1 Requirements for uncertainties in physics

What are the uncertainties we consider in HEP and in AI, and how do they relate or correspond to each other? The following brief text benefits from the excellent discussion on this topic in 2021 PDG review of AI in HEP [1].

1.1 Definitions of Uncertainty in HEP

In HEP, uncertainties are often categorized as "statistical" and "systematic." Statistical uncertainties typically refer to error associated with the finite sizes of a data set. For instance, in collider physics often a feature of the data, whose distribution is sensitive to a parameter of interest, is computed for each collision event and the population is binned for statistical inference. As such, the statistical model of counts in each bin follows a Poisson distribution (e.g. the bins have a Poisson statistical error). Similarly in Astrophysics, photon counting in measure image pixels follows a similar Poisson distribution.

Systematic uncertainties typically arise due to an incomplete knowledge of the true data generation process. When simulations (forward models) or analytic expressions (inverse models) are constructed with this incomplete knowledge, the difference is propagated through inference process to create a systematic error. For example, this could be uncertainty in the theoretical model of particle scattering, in particle-material interactions, detector calibrations, or in the material description of a given detector. Systematic uncertainties can also arise during the use of data-driven models, like those used in AI. Consider, for example, in the case where control samples of the data are selected and used to form a model of a background process that mimics a signal. Systematic uncertainties are often described with nuisance parameter ν, which parameterizes a continuous family of data-generating processes. Predictions depend on these nuisance parameters. When simulations and control samples don't match the original data (which is always), we encounter a domain transition (or adaptation) challenge

In addition to statistical and systematic uncertainties, the process of detecting particles is inherently uncertain and detectors have finite resolution; this resolution can be considered an inherent uncertainty in the detection process.

1.2 Definitions of Uncertainty in AI

In AI, uncertainty is often categorized broadly as "aleatoric" and "epistemic." "Aleatoric" uncertainty is often noted as being due to inherent randomness in the outcome of an experiment. "Epistemic" uncertainty is often described as coming from a lack of knowledge – e.g. a lack of knowledge about the model. In the AI literature, aleatoric uncertainty is often considered irreducible and coming from the data generation process, while epistemic uncertainty is discussed in terms of the uncertainty in model parameters and reduces when more data is added to the training data set.

Statistical error due to finite sample sizes and inherent randomness in measurements – e.g. due to finite detector resolutions – are concepts well studied in HEP. Model uncertainty in HEP often takes the form of a lack of knowledge in our physics models and is irreducible without further measurements [1], and thus is more akin to a systematic uncertainty, like a domain shift. More generally, the usage of ML uncertainty terminology and their correspondence to HEP uncertainty terminology is often ambiguous. How we use and interpret the various kinds uncertainties in a consistent way across AI-related analyses in HEP remains an unresolved discussion and problem.

The use of AI is growing across many sectors of HEP – e.g., reconstructing detector data for estimating physical objects like particle momenta or galaxy properties; classifying

detection events as signal or background; estimating the density of populations of events; defining low-dimensional powerful features that can be used for inference; constructing surrogate models of large-scale simulations; scheduling experimental data acquisition and operating instruments.

Depending on the task ML is used for and how inference on a data set is performed, different kinds of uncertainties may be needed. In cases where the ML model predictions determine the quality of a statistical model used to compare to data, careful attention to uncertainty is needed. In other cases the quality of the ML model may determine the optimality of an analysis or test statistic, but would would not lead to incorrect inference results; if and what uncertainties are needed in such cases also requires careful consideration. In the case of problems relating to instrument/experiment control, where the goal is predictive modeling of future states and to determine actions for a system to take, the full suite of related systematic uncertainties – AI model parameter uncertainty, resolutions of measurements, statistical, and domain shift – are notably essential to assess risk of and protect the data acquisition process and the instrument itself.

1.3 Inference for Parameter Estimation

A primary goal in HEP is to estimate the parameters θ of a physical model – e.g., the Standard Model of particle physics or the cosmological ΛCDM model – from experimental data, y. In collider physics, frequentist methods are common and the profile likelihood ratio test statistic is typically used for confidence interval estimation. In addition, in collider physics inference often takes the form of unfolding, wherein the corruption of particle properties due to the stochastic nature of detectors is removed (e.g. de-convolved). Bayesian inference has been in widespread use in cosmology since approximately 2010 – almost to the exception of frequentism. Indeed, the field is moving toward the use of Bayesian hierarchical models to simultaneous capture global features of the universe (e.g., cosmic dark matter density) and local features of individual objects (e.g., mass of one galaxy).

When one aims to model high-dimensional feature spaces, the estimation of the likelihood, the likelihood ratio, and the posterior all become extremely challenging. Simulation-Based Inference (SBI; aka, likelihood-free inference or implicit likelihood inference) has emerged as a subfield that uses AI with multi-fidelity simulators to efficiently and effectively perform such inference tasks [2]. Using the likelihood ratio trick, learned classifiers between different parameter hypothesis can be used to approximate likelihood ratios, and several strategies to perform this estimation efficiently have been proposed [3]. Similarly, using powerful normalizing flow models, which are capable of explicitly modeling high dimensional data densities, schemes have been developed to estimate likelihoods and posteriors [4].

2 Modern Approaches to UQ in AI

Predictive uncertainty associated with the output depends on both the AI model and the data. The terms "epistemic uncertainty" (also referred to as "systematic" or "model" uncertainty) and "aleatoric uncertainty" (also known as "statistical" or "data" uncertainty)

are commonly used in AI literature to differentiate these sources of errors [5, 6]. However, while the terms are sometimes used interchangeably, these correspondences don't consistently hold up across all problems and model-building scenarios.

In the context of scientific AI, however, the predictive uncertainties also depend on the mismatch between the data domains. For instance, a deep neural network (DNN) trained solely on numerical astrophysical simulations will have additional uncertainties associated when applied to real telescope observations. Such dataset-dependent uncertainties could be due to a domain transition (when the coverage or distribution of training data does not match with those of the real data) or due to out-of-domain information (training simulations not capturing all the effects of the physical phenomena) [7]. The above classifications are neither universal nor exhaustive, and uncertainty quantification must be specific to a particular AI approach to the scientific problem. The exact representation of uncertainty in an AI approach is often not fully known, due to the caveats mentioned above. Memory usage, convergence or training time, scaling across computing nodes, and other practical considerations play vital roles in the particular choice of a model for a given problem. Hence, AI-based uncertainty approximation remains a popular research avenue with variety of methods in active development [7–9].

In this section, we briefly mention a few popular categories of these UQ approaches.

- **Emulator-based posterior estimation:** A conservative approach to performing Bayesian approximation involves using AI surrogates in forward models or explicit likelihoods in Bayesian inference schemes such as Markhov Chain Monte Carlo (MCMC) techniques. These emulator-based approaches [e.g., 10–14] are now widely used in cosmological survey sciences. Coupled with active sampling techniques, one may also consider SBI methods [15] using neural density estimators. Recent work on differentiable emulators [16–18] also paves the way for faster posterior sampling via Hamiltonian Monte Carlo (HMC) techniques.

- **Sampling over network parameters:** neural networks could be used as inference machines as alternatives to traditional posterior estimation techniques, by sampling over the model parameters of network weights and biases [19]. This involves treating neural network's parameters as something you'd want to do posterior analysis over, by MCMC (or some other technique). The numerical expense of sampling over a large number of model parameters (multiple training experiments?) can be numerically expensive, as is the case with deep neural networks. However, with GPU-accelerators and AI-specific hardware, HMC sampling over a relatively small number network parameters will be within reach in the near future.

- **Variational Inference (VI):** This family of approaches approximate and learn the posterior distribution over network weights and biases. These VI-based methods effectively convert the statistical inference problem to a familiar network optimization problem, where a loss metric is optimized via gradient descent algorithms. A popular network that uses VI is the variational auto-encoder used in HEP problems that require generative modeling or dimension-reduction [20, 21]. Normalizing flow models [22] also use VI to construct complex distributions by transforming a probability density through a series of invertible mappings. Another widely used VI approach

of Monte Carlo Dropout (MC Dropout) uses a regularization technique of dropout layers [23] to an approximate a variational distribution. The trade-off between the quality of uncertainty estimates and the computational complexity is a notable feature of various VI-approaches: versions of MC dropout provice uncertainty estimates with little to no overhead during training time.

- **Deep ensemble methods:** Another approach to predicting UQ from point-estimating DNNs is to train a large number of networks, and use the ensemble distribution to predict the posterior prediction [24]. With an ensemble, each individually trained network has its own loss function. This is in contrast to sampling over NN's, where there aren't individual loss functions The ensemble of networks can be obtained by either on batches of data or different architectures, random weight initializations, MC Dropout, bootstrapping or other techniques. Ensemble-based UQ using sampling techniques like active learning or hyperparameter optimization [25, 26] are also currently explored in massive supercomputing environments. Averaging in the weight space, instead of the predictions, at different stages of training also results in a computationally inexpensive and reliable estimate of UQ [27].

- **Deep Gaussian Processes:** A Conventional Gaussian process (GP) is a non-parametric Bayesian model that provides uncertainty estimates on predictions, even with small number of training datapoints. The Deep GP is a composition of GPs that can effectively scale at large datasets using VI [28]. Deep GPs are demonstrated to perform better than several AI UQ techniques for domain shifts and against adversarial agents [29].

- **Test-time implementations:** Multiple predictions from a point-prediction or a probabilistic network can be augmented to evaluate the error in the predictions [30]. Such approaches are extremely useful in error propagation, wherein the input data from the sensors can be noisy [31].

- **Bayesian Neural Networks (BNNs):** These are typically optimized through VI. These incorporate parameters of a distribution for the the weights of an activation neuron [32].

- **Conformal methods:** A class of algorithms that use past experiences to determine levels of confidence in new predictions. Once a point prediction is made, a non-conformity score is constructed, which is a measure of how unusual the new example is compared to previously seen examples – i.e., a labeled calibration set. The conformal algorithm then converts this non-conformity score into a prediction region. Conformal methods can be used with any estimator, are computationally efficient and guarantee finite sample marginal coverage [33, 34].

- **Simulation-based inference (SBI):** Also known as implicit likelihood inference (or the misnomer, likelihood-free inference), SBI sequesters the model to the simulator only [2]. For a given choice of parameters of interest, a high-fidelity simulation (including the internal/hidden physical mechanisms and the observational signatures)

is compared to observations. If the simulation matches the observation, these parameters are selected as the best parameters to represent reality. Sometimes, these methods use normalizing flows for neural density estimation.

There are also methods in Reinforcement Learning and other areas of ML that are of interest to physics, which may be used in observation scheduling and instrument control. Some of these are discussed in this recent review [35].

Numerous implementations of these UQ algorithm families have started to shift the landscape of AI from deterministic models to those that are more probabilistic. Depending on the research problem, each of these methods have advantages and drawbacks. For instance, the MC dropout and test-time approaches are computationally inexpensive compared to the methods that involve sampling. We restrict the discussion of the UQ methods to the above, but note again that numerous implementations within each family tackle the issue of achieving robust UQ predictions using neural networks.

2.1 Calibration of Uncertainty Estimates

AI-based algorithms quote uncertainty estimates as probability distributions: they may be class probabilities for a classification task, or a credibility interval or a full probability distribution function for regression tasks. Calibration refers to the property of such probabilistic estimates to be consistent with the frequency of occurrence of the target variable in observed populations. For example, if an algorithm predicts a class probability of 0.8 or produces an 80% credibility interval, then the true value should fall in the predicted class or the credibility interval 80% of the time for any arbitrary test set. Though calibration seems like a simple property, most common machine learning algorithms provide mis-calibrated probabilistic estimates [36–40]. Though the term "calibration" is widely used to refer to this property in the AI community, the term "coverage" is more popular in the statistics community.

We want models to produce calibrated uncertainty estimates, because we would like to interpret probabilistic estimates as long run frequencies. Mis-calibrated uncertainty estimates used as input for a physical analysis will result in biased estimates of physical quantities.

Calibration performance of AI-based methods is generally assessed by comparing the predicted probability of a prediction belonging to a class (or a credibility interval) against their empirical probabilities calculated using a test set: ideally both the quantities should be equal. This is quantified using metrics like Expected Calibration Error [ECE; 41], Brier score [42], Probability Integral Transform [PIT; 43], etc. All of these tests fail to condition on the class and input features, but rather asses the calibration of probabilistic estimates as an ensemble and not on an individual basis [44, 45]. Moreover some of these metrics are optimized through completely un-physical probability estimates[46]. Therefore, metrics and tests should be developed to measure the calibration of individual uncertainty estimates (i.e., local calibration/conditional coverage) and not just for ensembles (i.e., global calibration/marginal coverage) as inconsistencies in various regions of the feature space cancel out to produce deceptively optimal results when looked as an ensemble[45, 47, 48].

Generally, it is easier to use generic re-calibration methods on outputs of existing algorithms than to develop custom workarounds for each of them. Therefore, methods like Plat scaling [49] and Isotonic regression [50] are widely used to re-calibrate outputs of AI-based uncertainty estimates. However, they suffer from the same problems as the metrics discussed immediately above in terms of their inability to ensure local calibration (conditional coverage) and therefore new methods should be developed to perform re-calibration of uncertainty estimates while ensuring local calibration [51–54].

2.2 Bias Mitigation

It is often very expensive to get labeled data from experiments and observations that also span the entire range of feature space of interest, but it is often that we have a complete set of unlabeled or simulated data. One approach to minimizing biases in our training data would be to train on simulations and then use transfer learning (or domain adaptation) methods on observed data. Similarly, we can also use self-supervised learning on an unlabeled data set and then fine tune to labeled data.

HEP heavily relies on comparing unlabeled (observed) data to labeled (simulation or observed) data) to infer parameters of mathematical models that describe our universe: e.g., AI models are often trained on simulations and then applied to data. Systematic differences between the simulations and observed data can thus bias physics results. The origin of such differences could be due to limitations in computational power, the limited power of some statistical technique applied, our ignorance about the exact state of a detector, or our inability to correctly model certain physics processes accurately within the simulation. The nature of the systematic uncertainty informs the appropriate mitigation/quantification mechanism [55].

2.2.1 Transfer Learning (TL) and Domain Adaptation (DA)

When an AI algorithm is trained on data that is not representative of, or biased in relation to, the testing data, this results in differences in the distribution between training and testing sets. Training on unbiased data is often not possible (or it is computationally burdensome) due to limitations in generation of the data. Domain adaptation and transfer learning are nascent research directions under the umbrella of AI that focus on resolving these differences, aiding in generalization of a HEP events classifier from a source dataset to a target domain. For example, TL has been used to fine-tune a DNN trained on generic imagery to suit the particular task of neutrino interaction classification, as an insufficient number of simulated events were available as data to train the specific use-case model directly [56]. Additionally, an adversarial learning-based DA approach was proposed to successfully mitigate bias in simulated event classification at the LHC [57, 58].

2.2.2 De-correlation methods to reduce bias in a statistical technique

A frequently used technique in HEP is to fit a smoothly falling parametric function to a background distribution (in some feature, like mass or momentum) to estimate its contribution in the signal region. While applying AI models to filter away uninteresting events

improves the sensitivity – therefore shrinking the statistical uncertainty on the final result – it can reshape the background distribution in undesired ways, resulting in larger systematic uncertainties. If the background distribution is no longer smoothly falling, it could result in a large bias. Several recently proposed AI methods are able to reduce the correlation between the output of the model and some given feature to avoid sculpting the background distribution, or between the output of the model and nuisance parameters defining systematic uncertainties [57, 59]. This sacrifices classification power to reduce the overall bias, but therefore improves the total (statistical and systematic) uncertainty for the final physics result.

2.2.3 Uncertainties with a statistical origin

Some systematic uncertainties have a statistical origin, for example the state of a detector, z, may be measured or calibrated using some auxiliary data and reported as a maximum likelihood estimate, z_0, along with an uncertainty. This uncertainty provides a probabilistic prior $P(z)$ on z. If an AI model is trained on simulations that assume the detector is in state z_0, the uncertainty can then be estimated by evaluating the bias of the model on other simulations that assume the detector in some other states, z_-, z_+. The final physics result may be reported such that the total uncertainty accounts for every possible bias in z, as well as its associated prior in $P(z)$.

AI offers an opportunity to take it a step further by reducing the final uncertainty of such a physics result. De-correlation methods may also be applied in this case to make the AI model insensitive to change in z, again sacrificing the classification power of the model in favor invariance to z, which results in increased statistical uncertainties but reduces systematic uncertainties. For example, a classifier can be forced to perform similarly on simulated and observed data in the control region to ensure that it does not learn any artifacts arising from inaccuracies in the simulation. However, for these kinds of uncertainties, a more appropriate solution is precisely the opposite idea, adaptive risk minimisation. An uncertainty-aware model is made maximally sensitive to the origin of the bias by parameterizing it on z. This model provides an optimal classifier for every value of z. This reduces the statistical component of the final uncertainty and allows us to better extract information about the true value of z from data, which can be combined with the prior $P(z)$ to further improve the final result.

While classifiers are trained for all the above discussed methods, the classifier objective acts only as a proxy for the final objective of a physics analysis. Inference-aware learning, where the true objective is directly optimized with differentiable programming, provides another tool to directly reduce the total uncertainty of a physics result. A model that is both inference-aware and uncertainty-aware is likely to provide close to optimal results. Since differentiable programming requires encoding the analysis-specific objective into differentiable code, this is a new skill that physicists will have to acquire in the future.

2.2.4 Uncertainties without a statistical origin

In contrast to the kind of uncertainties discussed above, certain systematic uncertainties lack a statistical origin. There is no probabilistic prior and it is often impossible to simu-

late the full range of possible biases coming from them. For example consider theoretical uncertainties that arise due to our inability to correctly model and compute QFT. Different hadronization models [60, 61] predict slightly different results but do not span the full range of possible mismodelling errors. The differences between two available hadronization models are often used as an ad hoc quantification of the uncertainty on hadronization modelling. The systematic uncertainties that arise because we are unable to compute QFT at infinite order are also estimated using similar well-motivated guesswork. Unphysical scales in the computation are varied in the simulation and the sensitivity of a physics result to these variations is treated as a systematic uncertainty. While this helps ascertain a rough estimate of the potential bias, the uncertainty band does not come from any rigorous statistical method for estimating an uncertainty.

While it may be tempting to apply AI-based de-correlation techniques to make a model insensitive to these biases, there is a danger in this case of significantly underestimating the true bias and uncertainties as discussed in Ref. [62]. Since the uncertainty quantification is not based on a robust statistical technique, attempting to reduce the uncertainty using AI may result simply in reducing our ad hoc estimate of the uncertainty, rather than the true uncertainty.

The best way to make concrete progress for such uncertainties is to develop a better understanding of their origin so that they can eventually be treated using a robust statistical model.

3 Recommendations

Below, we propose a list of guidelines in the context of interpretable UQ for our community to be better equipped to drive forward the progress in the field of AI-augmented HEP research. While some of these guidelines apply generally to good approaches in data-driven science research [63], we focus on recommendations pertinent to AI. **Overall, we emphasize that UQ for AI must become physically interpretable for usage in physics context.**

- Develop rigorous tests of local calibration.

- The HEP, Statistics, and AI communities should increase interaction to develop language that accurately reflects concepts relevant to both communities. Within 10 years, we should have a united lexicon for discussing and assessing interpretable uncertainties in AI for HEP.

- The HEP-Stats-AI community should create benchmark data sets for rigorous testing and comparison of approaches to physically interpretable AI UQ for physics.

- Funding agencies should endorse challenges and competitions within funding opportunity announcements to create and compare methods of UQ, including bias mitigation.

- Develop and compare methods of feature visualisation and interpretability, explainability, physics-awareness tools – e.g., Local interpretable model-agnostic explanations (LIME) values, Shapley values, partial dependency plots, or Accumulated local effects (ALE) [64].

- Bring physically interpretable AI UQ to non-physics problems where appropriate (e.g., social science) to help improve risk assessment.

- Common AI UQ methods should be embedded into deep learning software suites, similar to SKLearn, to enable widespread usage, testing, and comparison in the HEP community.

- AI adoption in HEP continues to lag behind latest methodologies. The HEP community should prioritize finding, testing, and comparing new methods generated by the AI community as soon as they emerge.

References

[1] P.A. Zyla et al. Review of Particle Physics. *PTEP*, 2020(8):083C01, 2020.

[2] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proc. Nat. Acad. Sci.*, 117(48):30055–30062, 2020.

[3] Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10):5242–5249, 2020.

[4] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[5] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

[6] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[7] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

[8] Yiping Li, Jianwen Chen, and Ling Feng. Dealing with uncertainty: A survey of theories and practices. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2463–2482, 2012.

[9] Niclas Ståhl, Göran Falkman, Alexander Karlsson, and Gunnar Mathiason. Evaluation of uncertainty quantification in deep learning. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 556–568. Springer, 2020.

[10] Katrin Heitmann, Earl Lawrence, Juliana Kwan, Salman Habib, and David Higdon. The coyote universe extended: precision emulation of the matter power spectrum. *The Astrophysical Journal*, 780(1):111, 2013.

[11] Thomas McClintock, Eduardo Rozo, Arka Banerjee, Matthew R. Becker, Joseph DeRose, Sean McLaughlin, Jeremy L. Tinker, Risa H. Wechsler, and Zhongxu Zhai. The Aemulus Project IV: Emulating Halo Bias. *arXiv e-prints*, page arXiv:1907.13167, July 2019.

[12] Zhongxu Zhai, Jeremy L. Tinker, Matthew R. Becker, Joseph DeRose, Yao-Yuan Mao, Thomas McClintock, Sean McLaughlin, Eduardo Rozo, and Risa H. Wechsler. The Aemulus Project. III. Emulation of the Galaxy Correlation Function. *The Astrophysical Journal*, 874(1):95, March 2019.

[13] Yosuke Kobayashi, Takahiro Nishimichi, Masahiro Takada, Ryuichi Takahashi, and Ken Osato. Accurate emulator for the redshift-space power spectrum of dark matter halos and its application to galaxy power spectrum. *Phys. Rev. D*, 102(6):063504, September 2020.

[14] Nesar Ramachandra, Georgios Valogiannis, Mustapha Ishak, Katrin Heitmann, LSST Dark Energy Science Collaboration, et al. Matter power spectrum emulator for f (r) modified gravity cosmologies. *Physical Review D*, 103(12):123525, 2021.

[15] Justin Alsing, Tom Charnock, Stephen Feeney, and Benjamin Wandelt. Fast likelihood-free cosmology with neural density estimators and active learning. *Monthly Notices of the Royal Astronomical Society*, 488(3):4440–4458, 2019.

[16] Chirag Modi, Francois Lanusse, and Uros Seljak. Flowpm: Distributed tensorflow implementation of the fastpm cosmological n-body solver. *Astronomy and Computing*, 37:100505, 2021.

[17] Andrew P Hearin, Jonás Chaves-Montero, Matthew R Becker, and Alex Alarcon. A differentiable model of the assembly of individual and populations of dark matter halos. *arXiv preprint arXiv:2105.05859*, 2021.

[18] Andrew P Hearin, Nesar Ramachandra, Matthew R Becker, and Joseph DeRose. Differentiable predictions for large scale structure with shamnet. *arXiv preprint arXiv:2112.08423*, 2021.

[19] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[20] Olmo Cerri, Thong Q Nguyen, Maurizio Pierini, Maria Spiropulu, and Jean-Roch Vlimant. Variational autoencoders for new physics mining at the large hadron collider. *Journal of High Energy Physics*, 2019(5):1–29, 2019.

[21] Andrew J Larkoski, Ian Moult, and Benjamin Nachman. Jet substructure at the large hadron collider: a review of recent advances in theory and machine learning. *Physics Reports*, 841:1–63, 2020.

[22] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[23] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[24] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *arXiv e-prints*, page arXiv:1612.01474, December 2016.

[25] Prasanna Balaprakash, Michael Salim, Thomas D Uram, Venkat Vishwanath, and Stefan M Wild. Deephyper: Asynchronous hyperparameter search for deep neural networks. In *2018 IEEE 25th international conference on high performance computing (HiPC)*, pages 42–51. IEEE, 2018.

[26] Romain Egele, Romit Maulik, Krishnan Raghavan, Prasanna Balaprakash, and Bethany Lusch. Autodeuq: Automated deep ensemble with uncertainty quantification. *arXiv preprint arXiv:2110.13511*, 2021.

[27] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[28] Andreas Damianou and Neil D Lawrence. Deep gaussian processes. In *Artificial intelligence and statistics*, pages 207–215. PMLR, 2013.

[29] John Bradshaw, Alexander G de G Matthews, and Zoubin Ghahramani. Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. *arXiv preprint arXiv:1707.02476*, 2017.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Nesar Ramachandra, Jonás Chaves-Montero, Alex Alarcon, Arindam Fadikar, Salman Habib, and Katrin Heitmann. Machine learning synthetic spectra for probabilistic redshift estimation: Syth-z. *arXiv preprint arXiv:2111.12118*, 2021.

[32] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[33] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421, 2008.

[34] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[35] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, dec 2021.

[36] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In Luc De Raedt and Stefan Wrobel, editors, *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 625–632. ACM, 2005.

[37] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

[38] Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Empirical frequentist coverage of deep learning uncertainty quantification procedures. *Entropy*, 23(12):1608, 2021.

[39] Andrey Malinin and Mark J. F. Gales. Predictive uncertainty estimation via prior networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7047–7058, 2018.

[40] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413, 2017.

[41] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2901–2907. AAAI Press, 2015.

[42] GLENN W. BRIER. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950.

[43] F. F. Gan and K. J. Koehler. Goodness-of-fit tests based on p-p probability plots. *Technometrics*, 32(3):289–303, 1990.

[44] Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 38–41. Computer Vision Foundation / IEEE, 2019.

[45] David Zhao, Niccolò Dalmasso, Rafael Izbicki, and Ann B. Lee. Diagnostics for conditional density models and bayesian inference algorithms. In Cassio P. de Campos, Marloes H. Maathuis, and Erik Quaeghebeur, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, volume 161 of *Proceedings of Machine Learning Research*, pages 1830–1840. AUAI Press, 2021.

[46] S. J. Schmidt, A. I. Malz, J. Y. H. Soo, I. A. Almosallam, M. Brescia, S. Cavuoti, J. Cohen-Tanugi, A. J. Connolly, J. DeRose, P. E. Freeman, M. L. Graham, K. G. Iyer, M. J. Jarvis, J. B. Kalmbach, E. Kovacs, A. B. Lee, G. Longo, C. B. Morrison, J. A. Newman, E. Nourbakhsh, E. Nuss, T. Pospisil, H. Tranin, R. H. Wechsler, R. Zhou, R. Izbicki, and LSST Dark Energy Science Collaboration. Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST). *Monthly Notices of the Royal Astronomical Society*, 499(2):1587–1606, December 2020.

[47] Wittawat Jitkrittum, Heishiro Kanagawa, and Bernhard Schölkopf. Testing goodness of fit of conditional density models with kernels. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pages 221–230. AUAI Press, 2020.

[48] Rachel Luo, Aadyot Bhatnagar, Yu Bai, Shengjia Zhao, Huan Wang, Caiming Xiong, Silvio Savarese, Stefano Ermon, Edward Schmerling, and Marco Pavone. Local Calibration: Metrics and Recalibration. *arXiv e-prints*, page arXiv:2102.10809, February 2021.

[49] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[50] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pages 694–699. ACM, 2002.

[51] Biprateep Dey, Jeffrey A. Newman, Brett H. Andrews, Rafael Izbicki, Ann B. Lee, David Zhao, Markus Michael Rau, and Alex I. Malz. Re-calibrating Photometric Redshift Probability Distributions Using Feature-space Regression. *arXiv e-prints*, page arXiv:2110.15209, October 2021.

[52] Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local temperature scaling for probability calibration. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6869–6879. IEEE, 2021.

[53] Niccolo Dalmasso, David Zhao, Rafael Izbicki, and Ann B Lee. Likelihood-free frequentist inference: Bridging classical statistics and machine learning in simulation and uncertainty quantification. *arXiv preprint arXiv:2107.03920*, 2021.

[54] Biprateep Dey, David Zhao, Jeffrey A Newman, Brett H Andrews, Rafael Izbicki, and Ann B Lee. Calibrated predictive distributions via diagnostics for conditional coverage. *arXiv preprint arXiv:2205.14568*, 2022.

[55] Nesar Ramachandra, Jonás Chaves-Montero, Alex Alarcon, Arindam Fadikar, Salman Habib, and Katrin Heitmann. Machine learning synthetic spectra for probabilistic redshift estimation: SYTH-Z. *Monthly Notices of the Royal Astronomical Society*, 06 2022. stac1790.

[56] Andrew Chappell and Leigh H Whitehead. Application of transfer learning to neutrino interaction classification. *arXiv preprint arXiv:2207.03139*, 2022.

[57] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[58] Jose M Clavijo, Paul Glaysher, Jenia Jitsev, and JM Katzy. Adversarial domain adaptation to reduce sample bias of a high energy physics event classifier. *Machine learning: science and technology*, 3(1):015014, 2021.

[59] Gregor Kasieczka and David Shih. Robust jet classifiers through distance correlation. *Phys. Rev. Lett.*, 125:122001, Sep 2020.

[60] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008.

[61] M. Bahr et al. Herwig++ Physics and Manual. *Eur. Phys. J. C*, 58:639–707, 2008.

[62] Aishik Ghosh and Benjamin Nachman. A cautionary tale of decorrelating theory uncertainties. *Eur. Phys. J. C*, 82(1):46, 2022.

[63] Alyssa Goodman, Alberto Pepe, Alexander W Blocker, Christine L Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, Yolanda Gil, Paul Groth, Margaret Hedstrom, et al. Ten simple rules for the care and feeding of scientific data. *PLoS computational biology*, 10(4):e1003542, 2014.

[64] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.