AFRL-RI-RS-TR-2023-023

INTERPRETABLE AND ROBUST ARTIFICIAL INTELLIGENCE

COLUMBIA UNIVERSITY

FEBRUARY 2023

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2023-023 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCRDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
WILMAR SIFRE
Work Unit Manager

/ S /
GREGORY HADYNSKI
Assistant Technical Advisor
Computing & Communications Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
FEBRUARY 2023	FINAL TECHNICAL REPORT	SEPTEMBER 2018	SEPTEMBER 2022

4. TITLE AND SUBTITLE

Interpretable and Robust Artificial Intelligence

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-18-C-0130	N/A	611011E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		R2KM

6. AUTHOR(S)

David M. Blei

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8.PERFORMING ORGANIZATION REPORT NUMBER
COLUMBIA UNIVERSITY SPONSORED PROJECTS ADMINISTRATION 116TH AND BDWY NEW YORK NY 10027	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10.SPONSOR/MONITOR'S ACRONYM(S)	11.SPONSOR/MONITOR'SREPORT NUMBER(S)
Air Force Research Laboratory/RITA 525 Brooks Road Rome NY 13441-4505	AFRL/RI	AFRL-RI-RS-TR-2023-023

12. DISTRIBUTION Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. **SUPPLEMENTARY NOTES**

14. ABSTRACT

We proposed safe AI systems. These AI systems are trusted because they are interpretable and resilient because they are robust. In particular, we pushed the state of the art in modern probabilistic modeling, including probabilistic models for causal inference. Probabilistic models provide a natural way for domain experts to express their assumptions and then to derive algorithms to compute under those assumptions. The results are interpretable because, for each model, we have a clear mathematical understanding of what is assumed, how the structure of the data interacts with the inferences, and the boundaries of what can and cannot be captured. With the methods we developed around causality, the resulting system will also be robust---robust to changes in the world and to interventions. To directly aid scientific discovery, we studied our methods on several problems in medical informatics, cancer therapy analysis, computational biology, and statistical astrophysics.

15. SUBJECT TERMS

Artificial Intelligence, Probabilistic Modeling, Causality

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	23

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
WILMAR SIFRE	N/A

PREVIOUS EDITION IS OBSOLETE.

STANDARD FORM 298 (REV. 5/2020)

Prescribed by ANSI Std. Z39.18

Contents

Approved for public release; distribution is unlimited.

List of Figures

Approved for public release; distribution is unlimited.

1 SUMMARY

Many of the successes of modern artificial intelligence (AI) are built on black-box prediction. In this paradigm, we are given an observed dataset and we use it to learn a distribution of future data. Black-box prediction has seen great successes in natural language processing, computer vision, and many other applications.

But while powerful, black-box prediction is limited. One limitation is that while black-box methods can form good predictions, they do not provide an interpretable basis for those predictions. This limitation does not matter in predictive tasks. However, the full potential for modern AI is not only as a tool for making predictions, but also as a tool for understanding the world through data.

Another limitation is its focus on pure prediction. Pure prediction tacitly assumes that future data comes from the same distribution as the past. But we cannot always make this assumption. In some domains, the distribution of future data is uncertain—we want to build AI systems that are robust to this uncertainty. In other domains, we explicitly want to change that distribution, i.e., to intervene on the system and to be able to predict what will happen under the intervention.

To overcome these limitations, we proposed safe AI systems that tackle these two problems. These AI systems are *trusted*, because they are interpretable, and *resilient*, because they are robust.

In particular, we pushed the state of the art in modern probabilistic modeling (Bishop, 2006; Murphy, 2013), and including probabilistic models for causal inference. Probabilistic models provide a natural way for domain experts to express their assumptions and then to derive algorithms to compute under those assumptions (Blei, 2014). The results are interpretable because, for each model, we have a clear mathematical understanding of what is assumed, how the structure of the data interacts with the inferences, and the boundaries of what can and cannot be captured. With the methods we developed around causality, the resulting system will also be robust—robust to changes in the world and to interventions.

Approved for public release; distribution is unlimited.

2 INTRODUCTION

Many of the successes of modern AI are built on black-box prediction. We are given an observed dataset and we use it to learn a distribution of future data. This basic paradigm has seen great successes in natural language processing, computer vision, and many other applications.

But there are two key limitations to this paradigm. The first is that while black-box methods can form good predictions, they do not provide an interpretable basis for those predictions. This limitation does not matter in predictive tasks. However, the full potential for modern AI is not only as a tool for making predictions, but also as a tool for understanding the world through data. We need scalable, reliable, and interpretable AI to realize this potential.

The second limitation is its focus on pure prediction. Pure prediction tacitly assumes that future data comes from the same distribution as the past. But we cannot always make this assumption. In some domains, the distribution of future data is uncertain and we want to build AI systems that are robust to this uncertainty. In other domains, we explicitly want to change that distribution, i.e., to intervene on the system and to be able to predict what will happen under the intervention.

We proposed safe AI systems that tackle these two problems. These AI systems are *trusted*, because they are interpretable, and *resilient*, because they are robust.

In the following report, we outline our main accomplishments:

- We developed new methodologies and understanding for probabilistic modeling. These include: new algorithms for approximate posterior inference and new theoretical understanding of variational inference; new classes of models for real-world scientific problems, such as in astrophysics and computational biology; and new understanding of long-found mysteries, such as posterior collapse, when combining probabilistic modeling and deep learning.

- We developed a long thread of new ideas for applied causality, essentially to uncovering interpretable and understandable inferences from real-world scientific data. These innovations include: a new methodology for multiple causal inference, where many possible causes can lead to an outcome; new methods for integrating neural networks and representation learning into causal inference algorithms; real-world scientific applications towards uncovering dose response curves in cancer research.

- We developed new methods for estimating model fitness and model diagnostics. We developed a new method that helps capture the notion of parsimony into the model selection process, one by which we check not only if a model fits the data well but also if a simpler model can "fool"

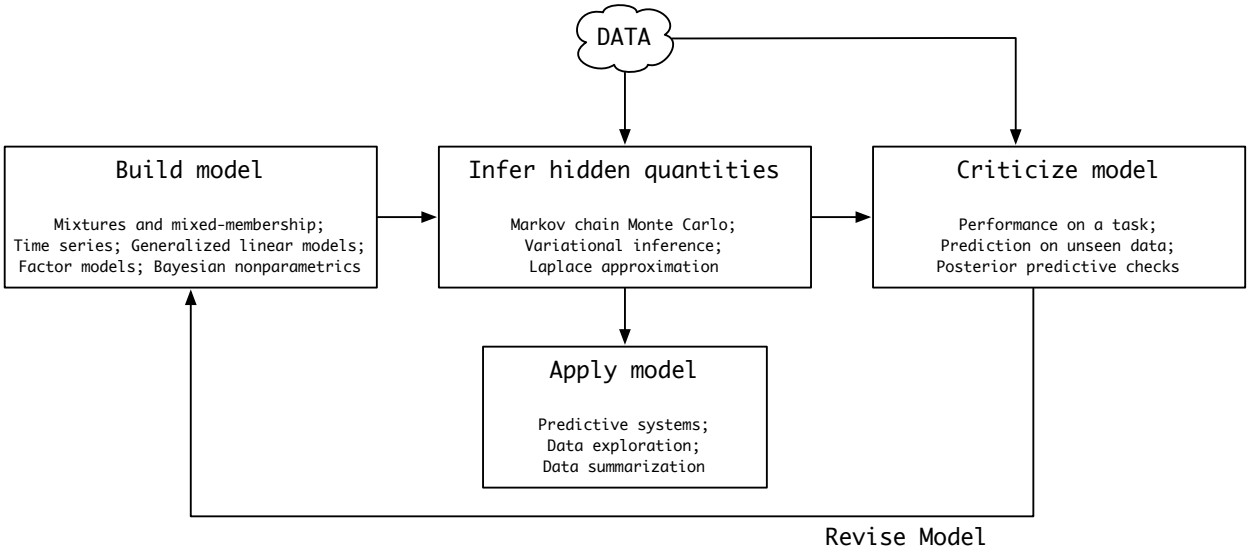


Figure 1: Box's loop

the more complex model. And we developed a new method for addressing the "double counting" problem of classical methods of Bayesian model criticism, one that naturally combines frequentist and Bayesian thinking, and one that is calibrated in a formal sense for better utility in real scientific applications.

We review background in probabilistic models, variational inference, and causality. Then we describe these research accomplishments.

3 METHODS, ASSUMPTIONS, AND PROCEDURES

Our work is in the framework of probabilistic machine learning. We review probabilistic machine learning and its central computational problems.

3.1 Background: Probabilistic Machine Learning

Probabilistic machine learning uses probabilistic models to analyze data. In the process of probabilistic modeling, we first develop a joint distribution of hidden and observed variables that captures our assumptions about how the data arises and how it interacts with structures we cannot observe. We then analyze our data by computing the conditional distribution of the hidden variables given the observations. This distribution, called the posterior, lets us examine the hidden structure that was likely to lead to the observed data, to form a predictive distribution of new data, and to check our model for comparison to others and direction of misfit. We then revise the model and continue with the analysis. We call this cycle "Box's loop." (Blei, 2014) (Figure 1).

The key technical problems to probabilistic modeling are how to compute the posterior, how to assess the quality of a model, and how to revise it based on our observations. As we describe in the next section, we developed fundamental new methods for posterior computation, model fitness, and model diagnostics. We deployed these methods to several applications.

Generically, let $x_{1:N}$ be N observations and divide the hidden variables into *global variables* β and *local variables* $z_{1:N}$. The global variables—like the mixture locations in a Gaussian mixture model—describe something about the whole data set. The local variables—like the component assignments—help govern the distribution of each data point, conditionally independent of the others. (Many models contain distinctive sets of variables like this, though not all. For other models, there may only be global hidden variables.) The posterior distribution is

$$p(z_{1:N}, \beta \mid x_{1:N}) = p(\beta) \prod_{n=1}^{N} p(z_n, x_n \mid \beta) / p(x_{1:N}). \tag{1}$$

The numerator is the joint distribution; the denominator is the marginal probability of the observations. Computing this posterior is the problem of *posterior inference*. For many models of interest, the denominator is not tractable to compute—it usually is construed as a complicated integral that marginalizes out the hidden variables—and we must resort to *approximate inference*.

The posterior is critical in the *predictive distribution* of new data given the observed data. In the predictive distribution, we marginalize out the hidden variables via the posterior,

$$p(x \mid x_{1:N}) = \int \left(\int p(z \mid \beta) p(x \mid z, \beta) dz \right) p(\beta \mid x_{1:N}) d\beta. \tag{2}$$

The inner integral is over the local hidden variables of the new data point; the outer integral is over the posterior of the global variables given the observed data set. This predictive distribution is used for both forming predictions and for implementing our proposed methods for assessing model fitness and developing model diagnostics.

3.2 Background: Variational Inference

In machine learning, there are two main methods for approximating the conditional—Markov chain Monte Carlo (MCMC) and variational inference. In MCMC, we form a Markov chain whose stationary distribution is the conditional, run the chain until it has "converged" (determining this convergence precisely is not usually possible), and then collect independent samples from which to approximate the posterior. MCMC is powerful, and has been widely studied, especially in Bayesian statistics. It is implemented in most existing probabilistic programming systems.

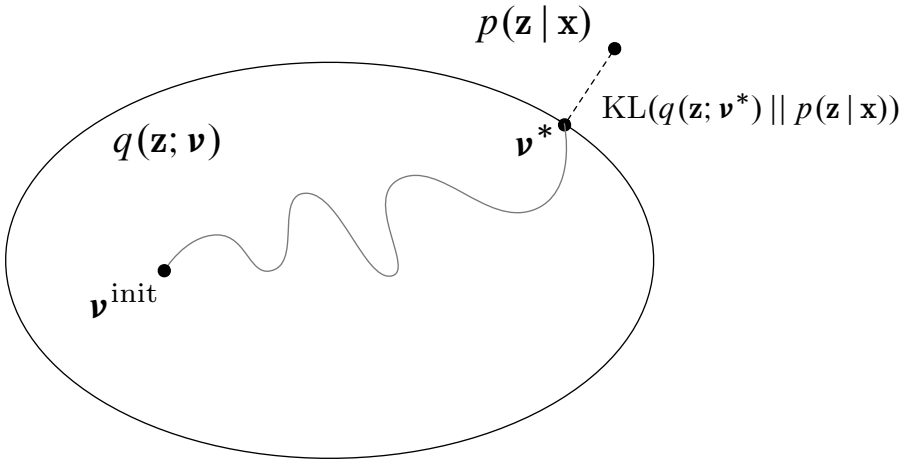


Figure 2: A schematic of variational inference

We built on *variational inference*, a deterministic alternative to MCMC that replaces sampling with optimization. Variational inference has been shown to be empirically faster than MCMC in several settings, though it is difficult to formally compare them. Mean-field variational inference provided the foundation for our research, though also extended beyond this assumption. Building on this method, we dramatically sped up and expanded the scope of generic approximate inference algorithms, and thus made probabilistic programming much more scalable.

Here we review the basics of mean-field variational inference.

The idea is to posit a factorized distribution of the hidden variables that is indexed by free *variational parameters*, $q(z_{1:N}, \beta) = q(\beta \mid \lambda) \prod_{n=1}^{N} q(z_n \mid \phi_n)$. These parameters—the local variational parameters ϕ_n and global variational parameter λ—are fit to make $q(z_{1:N}, \beta)$ close in Kullback-Leibler (KL) divergence to the true posterior $p(\beta, z_{1:N} \mid x_{1:N})$. We then use the fitted q as a proxy for the posterior, e.g., in a predictive distribution of new data or to explore the hidden structure of the observations.

But the KL is not computable. Variational methods optimize the *evidence lower bound* (ELBO),

$$\mathcal{L}(\lambda, \phi_{1:N}) = \mathbb{E}_q[\log p(\beta, z_{1:N})] + \mathbb{H}(q), \tag{3}$$

where $\mathbb{H}(\cdot)$ is the entropy of the distribution q. This objective is equal to the negative KL plus a constant; thus maximizing it is equivalent to minimizing KL. Note that the variational "model" is not a model of data, but rather a flexible family of distributions over the latent variables. The connection to the data and to the posterior is via optimizing the ELBO with respect to that family.

Figure 2 illustrates the main idea behind variational inference (VI). There is a *variational family*

Approved for public release; distribution is unlimited.

of distributions of latent variables. It is indexed by *variational parameters*; each setting of the variational parameters is a distribution of latent variables. We want to approximate the *exact posterior*, which is outside the variational family. VI begins at an *initial setting* of the variational parameters; it then *optimizes* them to find the member of the family that is closest to the exact posterior. Closeness is measured by the *KL divergence*; the KL is the objective of the optimization. In our research on variational inference, we consider and develop each piece of this framework. Indeed, the accomplishments described below can all be seen as improving one aspect of this algorithmic idea.

Typical applications of variational inference optimize the ELBO using coordinate ascent, iteratively optimizing each variational parameter. These updates are in closed form for models where each complete conditional is in the exponential family. (A complete conditional is the distribution of a hidden variable given all the other variables in the model.) But these methods are not useful in probabilistic programming, where the user should be able to express models from a much wider class without regard for the specific form of the complete conditionals. Further, each application of variational inference has required painstaking derivation and mathematics. This goes against the philosophy of a PPS to make modern machine learning accessible to a wide audience of users.

3.3 Background: Causality

Causality is a large field. Interested readers should consult the excellent books on this subject (Hernan and Robins, 2020; Pearl and MacKenzie, 2018; Pearl et al., 2016; Imbens and Rubin, 2015; Morgan and Winship, 2015). Here we provide a brief summary of the main ideas of this field.

Causal inference seeks to estimate a prediction under an intervention. If I make a change to the world, what do I expect to see? The key challenge is that the kinds of conditional expectations that are easy to estimate from observational data may not be predictive of the intervention. Other variables, such as confounders, can mislead the estimates. Thus the main goals of causal inference is to understand what assumptions are needed to make causal inferences and what estimators from the data will produce them.

One example: Let x be the treatment variable (or causes), y be the outcome, and z be the confounding variables. Suppose these variables appear together in a joint distribution $p(x, y, z)$. A causal estimate of the expectation of y when we intervene on x is provided by the "backdoor adjustment" formula or "g-estimation",

$$\mathbb{E}[Y\,;\,\mathrm{do}(x)] = \int p(z)p(y \mid x, z)\mathrm{d}z. \tag{4}$$

What is important about this formula is that the left side is a causal quantity, a quantity about an intervention, while the right side can be calculated from the observational joint. Note it is different from the conditional $p(y \mid x)$, which would integrate $p(z \mid x)$. This formula is valid when z contains all confounders or, in the graphical models formalism, blocks all backdoor paths.

Below we will discuss our accomplishments around applied causal inference. These works all try to estimate proxies for z from large-scale observed data. Of course, they make additional assumptions. But they are useful in data where we cannot always easily use an existing causal inference method, such as backdoor adjustment.

4 RESULTS AND DISCUSSION

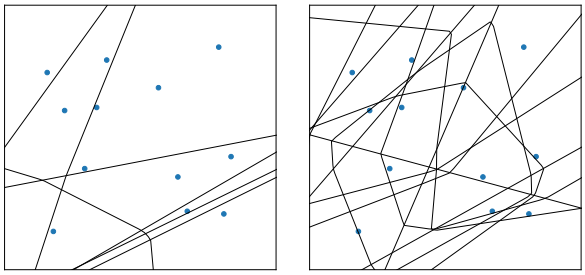
In this section, we detail each of our accomplishments as part of the Defense Advanced Research Project Agency (DARPA) Synergistic Discovery and Design (SD2) project.

4.1 Scalable and Generic Computing for Probabilistic Machine Learning

In a collection of related papers, we developed scalable and generic methods for probabilistic modeling and computation. We developed new models and new algorithms. We studied them both empirically and under a theoretical lens. These include flexible nonparametric probabilistic models, such as those based on deep learning and those with equivalent power.

In Loper et al. (2021) we developed a linear time inference method for Gaussian processes on one dimension. This method significantly speeds up Gaussian processes for important one-dimensional problems, like those found in computational neuroscience. Wu et al. (2021) developed hierarchical inducing points for Gaussian processes, also with an eye on improved computational complexity. To address sequential data, which is a crucial paradigm for scientific data analysis, Schein et al. (2019) develops Poisson-randomized gamma dynamical systems, which allow us to capture non-negative sequential latent variables with an efficient inference algorithm. Finally Ruiz et al. (2018) develops fast algorithms for large-scale categorical variables, which are also prominent in many scientific applications, especially in computational biology and genetics.

In Wang et al. (2021); Dieng et al. (2019) we addressed the problem of *posterior collapse*, where deep generative models do not provide meaningful representations of the data to which they are fit. In Dieng et al. (2019) we show how skip connections in the underlying neural network can help mitigate this issue. In Wang et al. (2021) we show theoretically *why* posterior collapse occurs—it is a symptom of latent-variable non-identifiability—and explain how the methods developed in the research literature perfectly fit this theory. In a closely related paper, we showed how variational



(a) At initialization (b) After NAC training

Figure 3: Distinct linear regions of a simple ReLU network with 2 layers of width 4 on 2D toy data. The lines represent the activation boundaries that divide the input space into distinct linear regions. NAC maximizes the number of linear regions of the network on the data, hence the maximum nonlinear expressivity.

auto-encoders can benefit from empirical Bayes thinking (Wang et al., 2019), again to mitigate posterior collapse.

Looking more deeply at the neural network side of deep probabilistic models, in Park et al. (2021) we present neural activation coding (NAC) as a novel approach for learning deep representations from unlabeled data. We argue that the deep encoder should maximize its nonlinear expressivity on the data for downstream predictors to take full advantage of its representational power. To this end, NAC maximizes the mutual information between activation patterns of the encoder and the data over a noisy communication channel. We show that learning for a noise-robust activation code increases the number of distinct linear regions of ReLU encoders, hence the maximum nonlinear expressivity. See Figure 3. More interestingly, NAC learns both continuous and discrete representations of data, which we respectively evaluate on linear classification and nearest-neighbor retrieval. Our empirical results show that NAC attains better or comparable performance on both tasks over recent baselines. In addition, NAC pre-training provides significant benefits to the training of deep generative models, such as avoiding posterior collapse (see above).

In a thread of related papers, we pushed the state of the art of probabilistic inference, making it both more scalable and more general. Wang and Blei (2019b) shows some of the first formal theoretical results for variational inference; Wang and Blei (2019c) extends the results to the important problem of model misspecification. Naesseth et al. (2020) presents a simple and scalable algorithm for general variational inference problems; Moretti et al. (2021) applies it to the biological problem of Bayesian phylogenetic inference.

In many applications, researchers want to extract statistically meaningful patterns from high-dimensional data. In Moran et al. (2022) we developed an interpretable probabilistic model for

finding such patterns in unsupervised settings. In unsupervised settings, one task is to learn low-dimensional representations (or factors) of high-dimensional data. For example, in genomics, many genes may be related to the same biological process; this coordination among the genes can be summarized as a single factor. To learn such factors, many researchers fit deep generative models (DGMs) with the help of variational autoencoders (VAEs). There are a number of challenges with such DGMs, however. Firstly, most DGMs are not identifiable: even with infinite data from the model, we cannot distinguish the true factors. That is, without identifiability, we cannot learn reliable latent representations of our data. Secondly, the latent factors are generally not interpretable; it is unclear what the factors represent with regard to the observed data. To solve these challenges, we introduced the sparse VAE **?**. The sparse VAE learns a set of latent factors which summarize the associations in the observed data features. The underlying model is sparse in that each observed feature (i.e. each dimension of the data) depends on a small subset of the latent factors. We prove such sparse deep generative models are identifiable: with infinite data, the true model parameters can be learned.

Finally we looked at several specific scientific applications of probabilistic modeling, combining domain expertise and ML expertise to build tailored models that solve specific problems.

The first application was in astrophysics. Interstellar dust corrupts nearly every stellar observation and accounting for it is crucial to measuring physical properties of stars. Miller et al. (2022) models the dust distribution as a spatially varying latent field with a Gaussian process and develops a likelihood model and inference method that scales to millions of astronomical observations. Modeling interstellar dust is complicated by two factors. The first is integrated observations. The data come from a vantage point on Earth, and each observation is an integral of the unobserved function along our line of sight, resulting in a complex likelihood and a more difficult inference problem than in classical GP inference. The second com- plication is scale; stellar catalogs have millions of observations. To address these challenges, we developed Ziggy, a scalable approach to GP inference with integrated observations based on stochastic variational inference. We study Ziggy on synthetic data and the Ananke dataset, a high-fidelity mechanistic model of the Milky Way with millions of stars. Ziggy reliably infers the spatial dust map with well-calibrated posterior uncertainties.

The second application was in biological data around cancer. Common approaches to gene signature discovery in single-cell RNA-sequencing (scRNA-seq) depend upon predefined structures like clusters or pseudo-temporal order, require prior normalization, or do not account for the sparsity of single-cell data. Levitin et al. (2019) presents single-cell hierarchical Poisson factorization, a Bayesian factorization method that adapts hierarchical Poisson factorization for de novo discovery

of both continuous and discrete expression patterns from scRNA-seq. Our model does not require prior normalization and captures statistical properties of single-cell data better than other methods in benchmark datasets. Applied to scRNA-seq of the core and margin of a high-grade glioma, scHPF uncovers marked differences in the abundance of glioma subpopulations across tumor regions and regionally associated expression biases within glioma subpopulations. scHFP revealed an expression signature that was spatially biased toward the glioma-infiltrated margins and associated with inferior survival in glioblastoma.

The final application was a model for perovskite crystallization based on a combination of Bayesian nonparametric regression and transfer-learned embeddings as features. This enabled us to more precisely estimate the probability of crystallization for unseen amines, and the model can be improved through active learning. We entered this model in the "perovskite prediction bakeoff." The goal of the bakeoff was to compare different models' predictions on four unseen amines. For each amine, we submitted 18 reagent concentrations that our models predicted will successfully produce a perovskite crystal. The chemists then ran experiments with these reagent concentrations to empirically assess the predictions. Our model used Bayesian nonparametric regression (specifically Bayesian Additive Regression Trees, BART) (Chipman et al., 2010) with transfer-learned embeddings as additional features. The embeddings were extracted from a convolutional molecule model (Chemprop) (Yang et al., 2019) trained on an external E-molecules dataset (N $\sim$ 95,000) to predict features associated with crystallization. Our BART model had the second highest success rate for predicting perovskite crysallization across all amines. Per-amine, BART was always among the top-3 models for all four amines tested; the other models were not as consistent. In contrast to other top-performing methods, BART used all of the input features and historical data, as well as incorporating its own transfer-learned latent representation of the amine molecule structure. This has clear benefits, as BART performs consistently well across all four amines tested, showing the potential benefits of transfer learning for predicting outcomes of chemical reactions.

4.2 Robust Probabilistic Modeling with Scalable Causal Inference

In another thread of research, we developed robust Artificial Intelligence (AI) with probabilistic modeling and scalable causal inference. Causal inference focuses on making predictions under intervention, when something about the world has changed. This is in contrast to classical machine learning methods, which only provide "passive" predictions. Causal inference is crucial to making scientific discoveries and to forming interpretable inferences; each causal inference can be directly seen as the answer to a well-formed question.

Wang and Blei (2019a, 2021) develops new algorithms for causal inference under multiple causes. This methodology helps us understand much of the practice in fields like genome-wide association studies. The main idea is that the multiple causes provide indirect evidence for confounders that are not explicitly measured in the data. More specifically, there is evidence for those confounders in the correlation structure among the causes. Methods for unsupervised probabilistic modeling can help uncover them and account for them. Zhang et al. (2019) further applies this methodology to electronic health records, for discoveries in medical science. Sridhar et al. (2022) applies it to network data, to understand influential nodes in a network of interacting entities. Wang et al. (2020) considers matrix factorization (for recommender systems and other applications) as a multiple causal problem.

In another thread of research on applied causality, we studied how we can use neural networks, along with the idea of invariance, to help answer causal questions. Veitch et al. (2019) corrects for confounders in network data, using network embedding models from deep learning; Veitch et al. (2020) considers a similar idea for text analysis. Shi et al. (2019) looks at specific neural network architectures to estimate treatment effects in scientific settings. Shi et al. (2022) considers how invariance plays a specific role in estimating treatment effects, and Yin et al. (pear) develops a conformal inference methodology for individual effects.

Tansey et al. (2022a) develops a model of dose response, to form causal estimates of patients' response to cancer treatments. This work was further extended to a fully Bayesian model in Tansey et al. (2022b). To expand on this scientific work, exploratory cancer drug studies test multiple tumor cell lines against multiple candidate drugs. The goal in each paired (cell line, drug) experiment is to map out the dose-response curve of the cell line as the dose level of the drug increases. In these two papers, we developed hierarchical models for dose-response modeling in multisample, multi-treatment cancer drug studies. We used low-dimensional embeddings to share statistical strength between similar drugs and similar cell lines, and involves structured shrinkage priors to encourage smoothness in the dose-response curves while remaining adaptive to sharp jumps when the data call for it. In benchmarks, we outperformed state-of-the-art methods.

In a related study of cancer experiments, also to estimate causal inferences, Tansey et al. (2020) presents Black Box False Discovery Rate (BB-FDR), an empirical-Bayes method for analyzing multi-experiment studies when many covariates are gathered per experiment. BB-FDR learns a series of black box predictive models to boost power and control the false discovery rate (FDR) at two stages of study analysis. In Stage 1, it uses a deep neural network prior to report which experiments yielded significant outcomes. In Stage 2, a separate black box model of each covariate is used to select features that have significant predictive power across all experiments. See Figure 4.

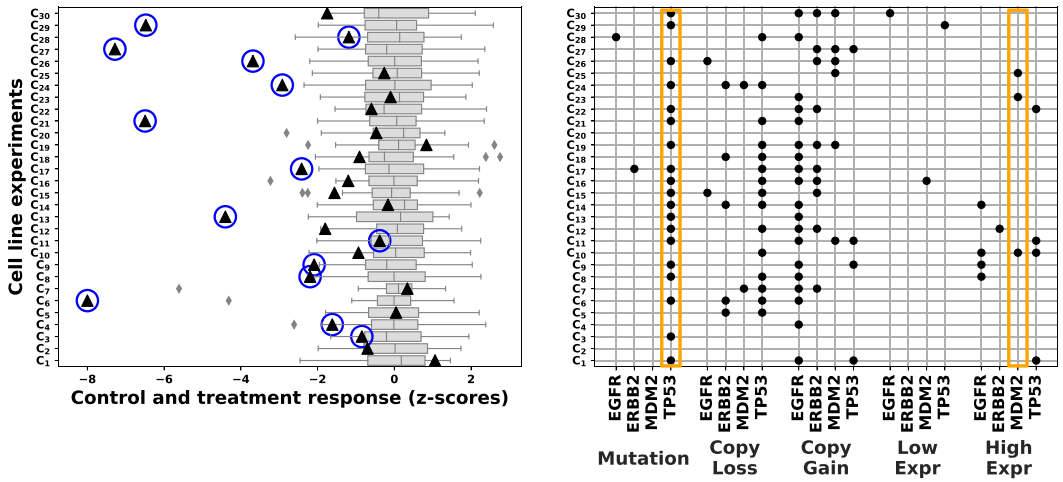


Figure 4: Left: a subset of 30 cell line experiments from the Nutlin-3 case study. Control replicates (grey box plots) and cell line responses (black triangles) are measured as z-scores relative to mean control values. Right: a subset of the corresponding molecular features for each experiment; black dots indicate a cell line has a recurrent mutation in the gene labelled on the x-axis. The goal in Stage 1 analysis is to select cell lines that showed a significant response (double empirical Bayes testing selections are circled in blue). In Stage 2, the molecular features are analysed to understand the mutations driving drug response (double empirical Bayes testing selections are circled in orange).

In benchmarks, BB-FDR outperforms state-of-the-art methods in both stages of analysis.

4.3 Model Criticism: Posterior Predictive Checks and Population Predictive Checks

In our last thread of research, we examined the problem of model criticism. Any probabilistic model will "fit" the data in that it will provide an inference under its own assumptions. How to check these assumptions is a crucial activity when using probabilistic models for important inferences, such as in scientific research and causality. To this end, we addressed some of the crucial issues in Bayesian model criticism. These two new forms of Bayesian model criticism could be important in validating and evaluating new probabilistic models of scientific data.

Traditionally, model criticism methods have been based on the predictive check, an adaptation of goodness-of-fit testing to Bayesian modeling and an effective method to understand how well a model captures the distribution of the data. In modern practice, however, researchers iteratively build and develop many models, exploring a space of models to help solve the problem at hand. While classical predictive checks can help assess each one, they cannot help the researcher understand how the models relate to each other. Moran et al. (pear) introduces the posterior predictive null (PPN), a method for Bayesian model criticism that helps characterize the relationships between models. The idea behind the PPN is to check whether data from one model's predictive distribution

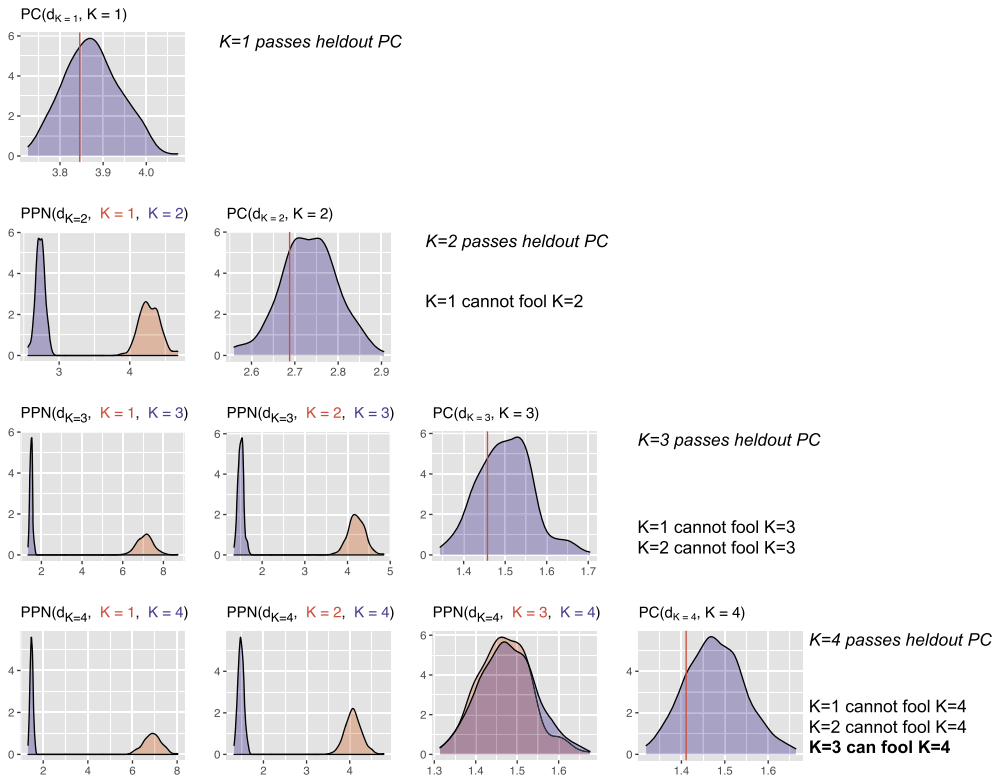


Figure 5: A PPN study of mixture models which suggests $K = 3$ is consistent with the data (no further mixture components are needed). In the data, the true value of K is 3. Along the diagonal are heldout predictive checks; every value of K passes the check. To the left of the diagonal are PPNs, each one checking if a simpler model can fool the model under study. While $K = 1$ and $K = 2$ pass their checks, the PPN shows that they cannot fool $K = 3$, which also passes. On the other hand, $K = 3$ can fool the check for $K = 4$.

can pass a predictive check designed for another model. This form of criticism complements the classical predictive check by providing a comparative tool. A collection of PPNs, which we call a PPN study, can help us understand which models are equivalent and which models provide different perspectives on the data. With mixture models, we demonstrated how a PPN study, along with traditional predictive checks, can help select the number of components by the principle of parsimony. See **??**. With probabilistic factor models, we demonstrated how a PPN study can help understand relationships between different classes of models, such as linear models and models based on neural networks. Finally, we analyzed data from the literature on predictive checks to show how a PPN study can improve the practice of Bayesian model criticism.

In a second thread of research around model criticism, we developed the population predictive check (pop-PC) (Moran et al., 2019). The pop-PC is built from the classical posterior predictive check (PPC), a seminal method that checks a model by assessing the posterior predictive distribution on

the observed data. However, PPC uses the data twice—both to calculate the posterior predictive and to evaluate it—which can lead to overconfident assessments of the quality of a model. Pop PCs, in contrast, compare the posterior predictive distribution to a draw from the population distribution, which in practice is a held-out dataset. We proved this strategy, which blends Bayesian modeling with frequentist assessment, is calibrated, unlike the PPC. Moreover, we demonstrate that calibrating PPC p-values post-hoc does not resolve the "double use of the data" problem.

5 CONCLUSIONS

We described our successes in pushing forward the state of the art of robust and interpretable probabilistic machine learning. Our contributions have changed the landscape of probabilistic inference, applied causality, and real-world scientific analysis.

To conclude, we will discuss some of the important directions for further progress.

- Probabilistic modeling and variational inference have come a long way towards scalable and generic inference. However, these important attributes are not completely together. Often we obtain generic methods—methods usable for any model—at the price of computational complexity. Finding automatic ways to scale generic inference is an important direction for this field.
- In causality, we have made good progress towards handling confounders that are not explicitly observed by adding assumptions and working with observational data. But we need new methods for assessing sensitivity to the required assumptions, and methods for understanding the finite-sample properties of our algorithms. How can we reduce variance while staying unbiased? Should we trade off bias for variance?
- Model criticism is an important activity, especially as inference methods have become more automated and widespread. But understanding the *practice* of model criticism, and deploying it to real-world applications is an important task for future analyses of scientific data.
- In summary, we accomplished many goals around probabilistic modeling, exploratory causal inference, and probabilistic model criticism, with applications to discovery from scientific data. But our ork is not done. In coming years, our vision is that probabilistic modeling, applied causality, and AI will become even more robust, scalable, interpretable, and essential to scientific progress.

Approved for public release; distribution is unlimited.

6 REFERENCES

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Blei, D. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232.

Chipman, H., George, E., and McCulloch, R. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Dieng, A., Kim, Y., Rush, A., and Blei, D. (2019). Avoiding latent variable collapse with generative skip models. In *Artificial Intelligence and Statistics*.

Hernan, M. and Robins, J. (2020). *Causal Inference: What If?* Chapman & Hall/CRC.

Imbens, G. and Rubin, D. (2015). *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press.

Levitin, H., Yuan, J., Cheng, Y., Ruiz, F., Bush, E., Bruce, J., Canoll, P., Iavarone, A., Lasorella, A., Blei, D., and Sims, P. (2019). De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. *Molecular Systems Biology*, 15(e8557).

Loper, J., Blei, D., Cunningham, J., and Paninski, L. (2021). A general linear-time inference method for Gaussian processes on one dimension. *Journal of Machine Learning Research*, 22(234):1–36.

Miller, A., Anderson, L., Leistedt, B., Cunningham, J., Hogg, D., and Blei, D. (2022). Mapping interstellar dust with Gaussian processes. *Annals of Applied Statistics*, 16(4):2672–2692.

Moran, G., Blei, D., and Ranganath, R. (2019). Population predictive checks. *arXiv:1908.00882*.

Moran, G., Cunningham, J., and Blei, D. (to appear). The posterior predictive null. *Bayesian Analysis*.

Moran, G., Sridhar, D., Wang, Y., and Blei, D. (2022). Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*.

Moretti, A., Zhang, L., Naesseth, C., Venner, H., Blei, D., and Pe'er, I. (2021). Variational combinatorial sequential Monte Carlo methods for Bayesian phylogenetic inference. In *Uncertainty in Artificial Intelligence*.

Morgan, S. and Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press, 2nd edition.

Murphy, K. (2013). *Machine Learning: A Probabilistic Approach*. MIT Press.

Naesseth, C., Lindsten, F., and Blei, D. (2020). Markovian score climbing: Variational inference with KL(p || q). In *Neural Information Processing Systems*.

Park, Y., Lee, S., Kim, G., and Blei, D. (2021). Unsupervised representation learning via neural activation coding. In *International Conference on Machine Learning*.

Pearl, J., Glymour, M., and Jewell, N. (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Pearl, J. and MacKenzie, D. (2018). *The Book of Why*. Basic Books.

Ruiz, F., Titsias, M., Dieng, A., and Blei, D. (2018). Augment and reduce: Stochastic inference for large categorical distributions. In *International Conference on Machine Learning*.

Schein, A., Linderman, S., Zhou, M., Blei, D., and Wallach, H. (2019). Poisson-randomized gamma dynamical systems. In *Neural Information Processing Systems*.

Shi, C., Blei, D., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. In *Neural Information Processing Systems*.

Shi, C., Veitch, V., and Blei, D. (2022). Invariant representation learning for treatment effect estimation. In *Uncertainty in Artificial Intelligence*.

Sridhar, D., Bacco, C. D., and Blei, D. (2022). Estimating social influence from observational data. In *Causal Learning and Reasoning*.

Tansey, W., Li, K., Zhang, H., Linderman, S., Blei, D., Rabadan, R., and Wiggins, C. (2022a). Dose-response modeling in high-throughput cancer drug screenings: An end-to-end approach. *Biostatistics*, 23(2):643–665.

Tansey, W., Tosh, C., and Blei, D. (2022b). A Bayesian model of dose-response for cancer drug studies. *Annals of Applied Statistics*, 16(2):680–705.

Tansey, W., Wang, Y., Rabadan, R., and Blei, D. (2020). Double empirical Bayes testing. *International Statistical Review*, 88.

Veitch, V., Sridhar, D., and Blei, D. (2020). Adapting text embeddings for causal inference. In *Uncertainty in Artificial Intelligence*.

Approved for public release; distribution is unlimited.

Veitch, V., Wang, Y., and Blei, D. (2019). Using embeddings to correct for unobserved confounding in networks. In *Neural Information Processing Systems*.

Wang, Y. and Blei, D. (2019a). The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596.

Wang, Y. and Blei, D. (2019b). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Wang, Y. and Blei, D. (2019c). Variational Bayes under model misspecification. In *Neural Information Processing Systems*.

Wang, Y. and Blei, D. (2021). A proxy variable view of shared confounding. In *International Conference on Machine Learning*.

Wang, Y., Blei, D., and Cunningham, J. (2021). Posterior collapse and latent variable non-identifiability. In *Neural Information Processing Systems*.

Wang, Y., Liang, D., Charlin, L., and Blei, D. (2020). Causal inference for recommender systems. In *ACM Conference on Recommender Systems*.

Wang, Y., Miller, A., and Blei, D. (2019). Comment: Variational autoencoders as empirical Bayes. *Statistical Science*, 34(2):229–233.

Wu, L., Miller, A., Anderson, L., Pleiss, G., Blei, D., and Cunningham, J. (2021). Hierachical inducing point Gaussian process for inter-domian observations. In *Artificial Intelligence and Statistics*.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. (2019). Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388.

Yin, M., Shi, C., Wang, Y., and Blei, D. (to appear). Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*.

Zhang, L., Wang, Y., Ostropolets, A., Mulgrave, J., Blei, D., and Hripcsak, G. (2019). The medical deconfounder: Assessing treatment effects with electronic health records. In *Machine Learning for Health Care*.

Approved for public release; distribution is unlimited.

7 LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

AI	Artificial Intelligence
BART	Bayesian additive regression trees
BB-FDR	Black-box false discovery rate
DARPA	Defense Advanced Research Project Agency
DGM	Deep generative model
ELBO	Evidence lower bound
KL	Kullback-Leibler
MCMC	Markov chain Monte Carlo
NAC	Neural activation coding
PPC	Posterior predictive check
PPN	Posterior predictive null
Pop-PC	Population predictive check
ReLU	Recified linear unit
SD2	Synergistic Discovery and Design
scRNA-seq	Single-cell RNA-sequencing
VAE	Variational autoencoder