A Matter of Structure: Robust Empirical Inference from Theory to Application

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Jonas Bernhard Wildberger
aus Düsseldorf

Tübingen
2026

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	15.07.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Bernhard Schölkopf
2. Berichterstatter:	Prof. Dr. Peter Gehler

Summary

Machine learning's success in idealized settings often fails to translate to real-world science, where complex data from astrophysics to genomics violates standard assumptions. To bridge this gap, this dissertation develops targeted contributions across three interconnected dimensions: theoretical foundations, methodological innovations, and applied challenges.

At the **theoretical** level, this work introduces the Interventional Kullback-Leibler (IKL) divergence, a novel framework for quantifying structural and distributional differences between causal models across multiple interventional environments. Unlike traditional metrics that focus on observational distributions, IKL captures both structural misalignments between causal graphs and parametric differences in observational distributions, guaranteeing similar outcomes under intervention across most environments. This enables principled causal model comparison and iterative discovery in dynamic systems where interventions induce distribution shifts.

On the **methodological** front, Flow Matching Posterior Estimation (FMPE) addresses scalability limitations in simulation-based inference by leveraging continuous normalizing flows. Rather than estimating conditional probability densities directly, FMPE parameterizes posteriors via vector fields. Since these are easier to estimate, computational resources can be focused on interpreting high-dimensional observations. This architectural insight yields substantial training efficiency improvements (30% reduction) while achieving performance comparable to specialized methods incorporating physical symmetries. Its effectiveness is demonstrated on challenging gravitational wave inference tasks and standard benchmarks.

In **applied** domains, this dissertation develops a probabilistic framework for adapting gravitational wave parameter estimation to time-varying detector noise. By modeling Power Spectral Densities (PSDs) through an interpretable latent space that separates broadband components from spectral lines, the approach maintains distributional support over anticipated variations rather than attempting to predict future conditions precisely. This enables real-time inference across extended observing runs without retraining, sustaining accuracy comparable to models with full future knowledge using only historical data and minimal target information.

Beyond individual contributions, this research reveals fundamental principles for robust empirical inference. Notably, *structural understanding*—whether causal graphs, architectural considerations, or physical mechanisms—consistently emerged as key to success across all domains. This understanding enables designing broad (synthetic) training distributions that anticipate future variations rather than attempting precise prediction of specific scenarios, often yielding superior robustness in dynamic environments. The power of such broad, massive training regimes is exemplified by the success of large foundation models, where diverse datasets enable generalization across tasks.

These findings suggest that the future of scientific machine learning lies not in developing increasingly sophisticated individual techniques, but in understanding fundamental principles that make empirical inference robust and reliable in complex, dynamic settings. This dissertation thus contributes both targeted solutions to scientific inference challenges and broader insights for approaching machine learning in scientific discovery. These methods aim to make machine learning more reliable in the messy, non-ideal settings where the most important questions lie.

Zusammenfassung

Der Erfolg des maschinellen Lernens unter idealisierten Bedingungen lässt sich oft nicht auf die reale Wissenschaft übertragen, wo komplexe Daten von der Astrophysik bis zur Genomik Standardannahmen verletzen. Um diese Lücke zu schließen, entwickelt diese Dissertation gezielte Beiträge in drei miteinander verbundenen Dimensionen: theoretische Grundlagen, methodische Innovationen und angewandte Herausforderungen.

Auf **theoretischer** Ebene führt diese Arbeit die Interventionale Kullback-Leibler (IKL) Divergenz ein, ein neuartiges Framework zur Quantifizierung struktureller und distributioneller Unterschiede zwischen kausalen Modellen über mehrere interventionelle Umgebungen hinweg. Im Gegensatz zu traditionellen Metriken, die sich auf Beobachtungsverteilungen konzentrieren, erfasst IKL sowohl Unterschiede in der Struktur kausaler Graphen als auch parametrische Unterschiede in den Beobachtungsverteilungen und garantiert ähnliche Ergebnisse unter Interventionen in den meisten Umgebungen. Dies ermöglicht prinzipiengeleiteten Vergleich kausaler Modelle und iterative Entdeckung in dynamischen Systemen, wo Interventionen Verteilungsverschiebungen verursachen.

Auf **methodischer** Ebene adressiert Flow Matching Posterior Estimation (FMPE) Limitationen in der Skalierung in Simulation-Based-Inference durch die Nutzung von Continuous Normalizing Flows. Anstatt bedingte Wahrscheinlichkeitsdichten direkt zu schätzen, parametrisiert FMPE Posteriorverteilungen über Vektorfelder. Da diese einfacher zu schätzen sind, können die Ressourcen auf die Interpretation hochdimensionaler Beobachtungen fokussiert werden. Diese architektonische Neuerung erzielt erhebliche Verbesserungen der Trainingseffizienz (30% Reduktion), während die Leistung vergleichbar mit spezialisierten Methoden, die physikalische Symmetrien einbeziehen ist. Dies wird an anspruchsvollen Gravitationswellen-Inferenzaufgaben und Standardbenchmarks demonstriert.

In **angewandten** Bereichen entwickelt diese Dissertation ein probabilistisches Framework zur Anpassung der Gravitationswellen-Parameterschätzung an zeitvariantes Detektorrauschen. Durch Modellierung von spektralen Leistungsdichten über einen interpretierbaren latenten Raum, der Breitbandkomponenten von Spektrallinien trennt, behält der Ansatz distributionelle Abdeckung über antizipierte Variationen bei. Dies ermöglicht Echtzeit-Inferenz über erweiterte Beobachtungsläufe ohne zusätzliches Finetuning und erreicht Genauigkeit vergleichbar mit Modellen mit vollständigem Zukunftswissen unter Verwendung nur historischer Daten und minimaler Zielinformation.

Über individuelle Beiträge hinaus enthüllt diese Forschung fundamentale Prinzipien für robuste empirische Inferenz. Insbesondere erwies sich *strukturelles Verständnis*—seien es kausale Graphen, architektonische Überlegungen oder physikalische Mechanismen—konsistent als Schlüssel zum Erfolg in allen Domänen. Dieses Verständnis ermöglicht das Design breiter (synthetischer) Trainingsverteilungen, die zukünftige Variationen antizipieren, anstatt präzise Vorhersagen spezifischer Szenarien zu versuchen, was oft zu besserer Robustheit in dynamischen Umgebungen führt. Die Macht solcher breiten, massiven Trainingsregime wird durch den Erfolg großer Foundation Models exemplifiziert, bei denen riesige Datensätze die Generalisierung über Aufgaben hinweg ermöglichen.

Diese Erkenntnisse deuten darauf hin, dass die Zukunft des wissenschaftlichen maschinellen Lernens nicht in der Entwicklung zunehmend ausgeklügelterer Einzeltechniken liegt, sondern im Verständnis fundamentaler Prinzipien, die empirische Inferenz in komplexen, dynamischen Umgebungen robust und zuverlässig machen. Diese Dissertation trägt sowohl gezielte Lösungen

für Herausforderungen in der wissenschaftlichen Inferenz als auch weitgehendere Erkenntnisse für maschinelles Lernen in der wissenschaftlichen Entdeckung bei. Diese Methoden zielen darauf ab, maschinelles Lernen zuverlässiger in chaotischen, nicht-idealen Umgebungen zu machen, wo die wichtigsten Fragen liegen.

Contents

Introduction 1

On September 14, 2015, humanity detected ripples in spacetime itself for the first time—but GW150914 was as much a triumph of data science as it was of physics. The challenge was not simply finding a signal buried in noise, but characterizing it reliably from data burdened with complex, non-stationary noise that defied conventional analysis [1–3]. This complexity strains traditional methods, making them computationally intensive and prohibitively slow—often taking hours or days when multimessenger astronomy demands actionable insights within minutes to coordinate follow-up observations across the electromagnetic spectrum. This scenario exemplifies a core problem in modern science: the need to extract reliable knowledge from data that deviate dramatically from the idealized statistical assumptions (e.g., stationary, Gaussian noise) for which many classical algorithms are designed.

[1]: Abbott et al. (2016), 'Observation of Gravitational Waves from a Binary Black Hole Merger'
[2]: Abbott et al. (2016), 'Astrophysical Implications of the Binary Black-Hole Merger GW150914'
[3]: Green et al. (2021), 'Complete parameter inference for GW150914 using deep learning'

This tension between idealized assumptions and real-world complexity is not unique to astrophysics. While machine learning excels in static, controlled environments with large, independent, and identically distributed (i.i.d.) samples [4, 5], data from domains as diverse as genomic sequencing and econometrics often defy these conditions. They present heterogeneous distributions, noisy measurements, or high-dimensional challenges that push standard approaches to their limits [6–8]. Empirical inference emerges as the bridge between these messy realities and reliable scientific insights, transforming noisy observations into actionable knowledge across diverse domains and scales.

[4]: LeCun et al. (2015), 'Deep learning'
[5]: Goodfellow et al. (2016), *Deep Learning*

[6]: Peters et al. (2017), *Elements of causal inference: foundations and learning algorithms*
[7]: Zhang et al. (2013), 'Domain adaptation under target and conditional shift'
[8]: Pearl (2009), *Causality: Models, Reasoning and Inference*

This dissertation explores three dimensions of empirical inference, tracing an arc from theoretical foundations—where understanding dynamic systems under change sharpens causal insights—to methodological innovations—where scaling computational techniques enables handling complex data—and finally to applied challenges—where ensuring reliability in dynamic, real-world settings delivers impactful scientific outcomes. Each dimension strengthens this bridge, enhancing our capacity to transform data into meaningful, trustworthy conclusions.

1.1 Empirical Inference in Machine Learning

Empirical inference is "the process of drawing conclusions from observational data," encompassing scientific experiments—where measurements inform laws—and natural systems—where organisms adapt to environmental cues [9]. This dual perspective highlights its role as a unifying principle in statistical learning and artificial intelligence. As a central methodology within machine

[9]: Schölkopf (2011), 'Empirical Inference'

learning, empirical inference underpins most algorithms by focusing on statistical generalization from finite, often noisy data samples under uncertainty, contributing to the overarching goal of simulating general intelligence. Its formalization owes much to Vapnik's Statistical Learning Theory, which introduced Empirical Risk Minimization (ERM) to rigorously address learning from limited samples, forming the theoretical basis for models like Support Vector Machines [10, 11].

[10]: Vapnik (1999), *The nature of statistical learning theory*
[11]: Schölkopf et al. (2002), *Learning with kernels : support vector machines, regularization, optimization, and beyond*

Beyond its theoretical underpinnings, empirical inference serves as the bedrock of supervised learning, where models are fitted to labeled datasets by minimizing empirical loss, and informs Bayesian inference, where prior beliefs are updated with observed data. These principles have driven landmark achievements in scientific discovery, such as AlphaFold's predictions of protein structures from vast sequence data [12] and the characterization of the first gravitational wave event (GW150914), where Bayesian parameter estimation inferred the masses, spins, and distance of the merging black holes from faint signals buried in noisy detector outputs [1]. These successes demonstrate the power of empirical methods to tackle complex, data-driven tasks across diverse fields.

[12]: Jumper et al. (2021), 'Highly accurate protein structure prediction with AlphaFold'

[1]: Abbott et al. (2016), 'Observation of Gravitational Waves from a Binary Black Hole Merger'

However, applying these techniques at the frontiers of scientific discovery or in complex real-world systems often involves navigating inherent challenges—such as dynamic behavior, observational noise, and vast scale—that demand sophisticated, robust methodologies. This motivates this dissertation's exploration of empirical inference across an arc from theoretical insights into causal systems, through methodological advancements in simulation-based inference, to applied solutions for real-world scientific challenges.

1.2 Background

This dissertation builds on advances in three key areas that exemplify the challenges of empirical inference across different scales: causal inference (theoretical foundations), simulation-based inference (methodological innovations), and gravitational wave analysis (applied challenges).

1.2.1 Causal Inference

Causal inference seeks to uncover cause-and-effect relationships from observational and interventional data, moving beyond traditional machine learning's correlative focus [13, 14]. Rooted in graphical models and do-calculus, it formalizes how interventions affect outcome distributions, enabling reasoning about counterfactuals and robustness [6]. At its core, causal inference employs directed acyclic graphs (DAGs) to represent causal relationships,

[13]: Pearl (2003), 'Causality: models, reasoning, and inference'
[14]: Pearl (2009), *Causality*

[6]: Peters et al. (2017), *Elements of causal inference: foundations and learning algorithms*

where nodes denote variables and edges indicate direct causal influences. Building on DAGs, Structural Causal Models (SCMs) provide a more comprehensive framework by explicitly defining the functional relationships between variables through deterministic equations, often of the form $X_i = f_i(\mathbf{PA}_i, U_i)$, where $\mathbf{PA}_i$ are the parents of X_i in the graph and U_i represents exogenous noise or unobserved factors. SCMs allow for the simulation of interventions by replacing specific functions, thus modeling causal effects more comprehensively [14]. The do-calculus, a key mathematical tool introduced by Pearl, provides a framework to compute interventional distributions from observational data, expressed as $P(Y|do(X = x))$, which denotes the probability of outcome Y given an intervention setting X to value x, distinct from conditional probabilities $P(Y|X = x)$ that may include confounding effects [14]. Do-calculus consists of three rules that enable the transformation of interventional queries into observational ones under certain graphical conditions, facilitating causal effect identification even when direct experimentation is infeasible. This mathematical machinery is crucial for empirical inference because it provides a principled way to reason about cause and effect from observational data—a capability essential for scientific discovery where controlled experiments are often impossible. The formalism allows identification of causal effects under specific graph structures, often using adjustment criteria like the back-door criterion to block confounding paths [6].

[14]: Pearl (2009), *Causality*

[14]: Pearl (2009), *Causality*

[6]: Peters et al. (2017), *Elements of causal inference: foundations and learning algorithms*

In machine learning, causal inference enhances interpretability and generalization, as seen in applications like estimating treatment effects in medical trials to disentangle confounders [15]. Yet, interventions often induce distribution shifts, and unknown causal graphs pose challenges, necessitating robust metrics to compare models across environments [16, 17]. Interventions can be categorized into types such as hard interventions, which fix a variable to a specific value (e.g., $do(X = x)$), and soft interventions, which modify the causal mechanism without severing dependencies on parent variables, preserving conditional relationships while altering the distribution. A central problem is identifying whether a causal effect is identifiable from data, which depends on the underlying graph and available observations—a challenge addressed through structural assumptions or experimental design [6]. This theoretical domain underpins tools like the Interventional Kullback-Leibler divergence in this dissertation, extending empirical inference's reach in dynamic systems by providing a principled way to quantify differences under interventional shifts.

[15]: Imbens (2015), 'Matching Methods in Practice: Three Examples'

[16]: Schölkopf (2019), 'Causality for Machine Learning'
[17]: Schölkopf et al. (2021), *Towards Causal Representation Learning*

[6]: Peters et al. (2017), *Elements of causal inference: foundations and learning algorithms*

While causal inference provides the theoretical foundation for understanding interventional differences, practical implementation of these ideas requires scalable computational methods—particularly when dealing with high-dimensional scientific data.

1.2.2 Simulation-Based Inference and Normalizing Flows

Simulation-based inference (SBI) extends Bayesian inference to intractable likelihood settings, using simulations to approximate posteriors when analytical solutions are unavailable [18]. Widely used in cosmology, epidemiology, and other scientific fields, SBI leverages neural density estimators to learn distributions from simulated data, addressing scenarios where the likelihood function $p(x|\theta)$ is computationally infeasible to evaluate directly, yet simulations from the forward model $x \sim p(x|\theta)$ are possible [19]. A key approach within SBI is Neural Posterior Estimation (NPE), which directly fits a conditional density estimator $q(\theta|x)$ to approximate the true posterior $p(\theta|x)$ using a maximum likelihood objective

$$\mathcal{L}_{\text{NPE}} = -\mathbb{E}_{p(\theta)p(x|\theta)}[\log q(\theta|x)]. \tag{1.1}$$

This is achieved by training on simulated pairs (θ, x) drawn from the joint distribution $p(\theta)p(x|\theta)$, enabling amortized inference over multiple observations [20].

A common choice for the density estimator in NPE and other SBI methods is the normalizing flow [19, 21], which defines a probability distribution $q(\theta|x)$ through an invertible mapping $\psi_x : \mathbb{R}^n \to \mathbb{R}^n$ from a simple base distribution $q_0(\theta)$, expressed as

$$q(\theta|x) = q_0(\psi_x^{-1}(\theta)) \det \left| \frac{\partial \psi_x^{-1}(\theta)}{\partial \theta} \right|. \tag{1.2}$$

These flows can be implemented as discrete normalizing flows, which are compositions of simpler mappings with triangular Jacobians, allowing efficient density evaluation and sampling. This makes them suitable for modeling complex posteriors in high-dimensional spaces. Alternatively, continuous normalizing flows [22] offer a different parameterization, defined by a time variable $t \in [0, 1]$ and a vector field $v_{t,x}$ that governs sample trajectories via a neural ordinary differential equation (ODE)

$$\frac{d}{dt}\psi_{t,x}(\theta) = v_{t,x}(\psi_{t,x}(\theta)), \tag{1.3}$$

with the resulting density computed as

$$q(\theta|x) = q_0(\theta) \exp\left(- \int_0^1 \operatorname{div} v_{t,x}(\theta_t)\, \mathrm{d}t \right). \tag{1.4}$$

While continuous flows provide flexibility in modeling, they often face challenges in training due to the computational cost of ODE integration [22]. As a methodological domain, SBI bridges theoretical Bayesian principles to practical application, enabling inference in complex systems where traditional methods fail.

These computational challenges manifest starkly in gravitational wave analysis, where the demands of real-time parameter esti-

[18]: Cranmer et al. (2020), 'The frontier of simulation-based inference'

[19]: Papamakarios et al. (2021), 'Normalizing Flows for Probabilistic Modeling and Inference'

[20]: Papamakarios et al. (2016), 'Fast ε-free Inference of Simulation Models with Bayesian Conditional Density Estimation'

[19]: Papamakarios et al. (2021), 'Normalizing Flows for Probabilistic Modeling and Inference'

[21]: Rezende et al. (2015), 'Variational Inference with Normalizing Flows'

[22]: Chen et al. (2018), 'Neural Ordinary Differential Equations'

[22]: Chen et al. (2018), 'Neural Ordinary Differential Equations'

mation in noisy, dynamic environments push simulation-based methods to their limits.

1.2.3 Gravitational Waves in Machine Learning

Gravitational waves (GWs) are ripples in the fabric of spacetime, predicted by Einstein's General Theory of Relativity [23] and first directly observed on September 14, 2015, with the detection of GW150914 by the Advanced LIGO detectors, marking the dawn of GW astronomy [1]. Produced by cataclysmic events such as the inspiral and merger of binary black holes or neutron stars, GWs carry information about their sources' physical properties, offering a unique window into extreme astrophysical phenomena and tests of fundamental physics [24]. The LIGO, Virgo, and KAGRA observatories detect these minute distortions in spacetime using laser interferometry, measuring strain amplitudes as small as 10^{-21} over kilometer-scale baselines, a feat requiring extraordinary sensitivity amidst complex detector noise [25–27].

[23]: Einstein (1916), 'The foundation of the general theory of relativity.'

[1]: Abbott et al. (2016), 'Observation of Gravitational Waves from a Binary Black Hole Merger'

[24]: Christensen et al. (2022), 'Parameter estimation with gravitational waves'

[25]: Aasi et al. (2015), 'Advanced LIGO'

[26]: Acernese et al. (2015), 'Advanced Virgo: a second-generation interferometric gravitational wave detector'

[27]: Akutsu et al. (2021), 'Overview of KAGRA: Detector design and construction history'

GW parameter estimation seeks to infer source properties—such as masses, spins, and distances—from observed signals, a task complicated by detector noise characterized by its power spectral density (PSD) $S_n(f)$. In traditional approaches, the likelihood of parameters θ is modeled assuming additive, stationary Gaussian noise, expressed as $p(d|\theta, S_n) \propto \exp\left(-\frac{1}{2}(d - h(\theta)|d - h(\theta))_{S_n}\right)$, where $(a|b)_{S_n}$ is a noise-weighted inner product integrating over frequency bands and detectors. PSDs are estimated using off-source data or joint signal-noise modeling, but noise variability over time and across observing runs (e.g., LIGO's O3 run) challenges inference accuracy [24, 28]. Machine learning, particularly simulation-based inference (SBI), offers a transformative approach by training conditional density estimators like normalizing flows $q(\theta|d, S_n)$ on simulated data with varying PSDs, enabling rapid posterior sampling in seconds once a detection is made, as exemplified by frameworks like Dingo that amortize computational costs across events [28].

[24]: Christensen et al. (2022), 'Parameter estimation with gravitational waves'

[28]: Dax et al. (2021), 'Real-Time Gravitational Wave Science with Neural Posterior Estimation'

As an application of empirical inference, GW analysis exemplifies the challenge of extracting faint signals from dynamic, noisy environments, pushing the limits of existing methods. This makes it an ideal testbed for robust inference techniques: the signal-to-noise ratios are often barely detectable, the noise characteristics evolve unpredictably over time, and the scientific stakes—including multimessenger astronomy alerts that must be issued within minutes—demand both accuracy and speed. Using various Bayesian inference techniques like MCMC and nested sampling[29–31], the field has already yielded insights into compact binary merger rates and tests of general relativity from over 50 detected events. Future observations are expected to include core-collapse supernovae and stochastic backgrounds [24]. Yet, evolving noise distributions—due to detector upgrades or environmental factors—demand models

[28]: Dax et al. (2021), 'Real-Time Gravitational Wave Science with Neural Posterior Estimation'

[29]: Veitch et al. (2015), 'Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library'

[30]: Pankow et al. (2015), 'Novel scheme for rapid parallel parameter estimation of gravitational waves from compact binary coalescences'

[31]: Ashton et al. (2019), 'BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy'

[24]: Christensen et al. (2022), 'Parameter estimation with gravitational waves'

that generalize across temporal scales without costly retraining, a critical issue for real-time analysis during extended observing runs.

These areas—causal inference, simulation-based inference, and gravitational wave analysis—collectively span the theoretical, methodological, and applied dimensions of empirical inference. Together, they frame critical problems of quantifying structural differences in causal models, developing efficient inference methods for high-dimensional scientific tasks, and adapting to non-stationary environments in real-world applications. This dissertation addresses these intertwined issues through targeted contributions to extract reliable insights from diverse, challenging data landscapes.

1.3 Problem Statement

The advancement of machine learning and its application to complex scientific domains hinges on robust empirical inference—the capacity to extract meaningful insights from data that often deviate significantly from idealized i.i.d. conditions [32, 33]. While foundational techniques like Empirical Risk Minimization and Bayesian inference provide powerful tools, pushing the frontiers of data-driven discovery requires addressing specific, interconnected obstacles that limit the scope and reliability of current methods across theoretical, methodological, and applied contexts. This dissertation targets three critical challenges within empirical inference, each representing a barrier to progress in understanding and modeling complex systems: First, at the **theoretical** level, validating causal models remains a key challenge. Data often presents itself across diverse environments arising from interventions, offering a natural signal for inferring causal relationships. Yet, leveraging this information is problematic. Existing metrics often fall short: purely structural comparisons like the Structural Hamming Distance (SHD) [34, 35] ignore distributional consequences, while others like the Structural Intervention Distance (SID) [36] require access to the true causal graph, which is rarely available. Fully harnessing multi-environment data thus requires new, principled metrics that can quantify discrepancies between causal hypotheses across the intervention-induced distribution shifts without presupposing the ground truth structure—a foundational step for enhancing causal reasoning in dynamic contexts [6, 16].

Second, on the **methodological** front, simulation-based inference (SBI) techniques encounter scalability limitations. As problem dimensionality increases, training expressive neural density estimators such as normalizing flows becomes computationally prohibitive, often compromising precision in approximating high-dimensional posteriors. This underscores the demand for novel

[32]: Hendrycks et al. (2021), 'The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization'
[33]: Geirhos et al. (2020), 'Shortcut learning in deep neural networks'

[34]: Acid et al. (2003), 'Searching for Bayesian Network Structures in the Space of Restricted Acyclic Partially Directed Graphs'
[35]: Tsamardinos et al. (2006), 'The max-min hill-climbing Bayesian network structure learning algorithm'

[36]: Peters et al. (2013), 'Structural Intervention Distance (SID) for Evaluating Causal Graphs'

[6]: Peters et al. (2017), *Elements of causal inference: foundations and learning algorithms*
[16]: Schölkopf (2019), 'Causality for Machine Learning'

algorithms that enhance efficiency while maintaining accuracy for complex scientific tasks [18, 19].

[18]: Cranmer et al. (2020), 'The frontier of simulation-based inference'
[19]: Papamakarios et al. (2021), 'Normalizing Flows for Probabilistic Modeling and Inference'

Third, in **applied** domains, the reliability of machine learning models is frequently undermined by non-stationary data characteristics. Gravitational wave (GW) parameter estimation exemplifies this issue, where detector noise evolves unpredictably over time (e.g., across LIGO/Virgo observing runs like O3), causing models trained on historical data to underperform on future observations. This highlights the need for adaptive methods that sustain performance in dynamic experimental settings without requiring constant, resource-intensive retraining [37].

[37]: Abbott et al. (2020), 'A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals'

These three challenges — while distinct in their technical manifestations — share a common thread: they arise when real-world complexity pushes beyond the assumptions underlying standard machine learning approaches. This dissertation offers targeted solutions to each problem, contributing novel frameworks and methodologies that help extract reliable insights from diverse, complex data, ultimately broadening the impact of machine learning in scientific discovery.

1.4 Outline

The remainder of this dissertation is structured as follows. Each chapter addresses a distinct challenge within the broader scope of empirical inference, directly corresponding to the critical problems identified in the preceding discussion. These chapters are self-contained, adapted with minor modifications from their original conference or journal versions, and collectively aim to advance our ability to extract robust insights from complex data across theoretical, methodological, and applied domains. For the reader's benefit, this dissertation assumes a working knowledge of machine learning, causal inference, Bayesian statistics, and gravitational wave analysis.

Measuring Causal Differences with IKL Divergence Chapter 2, based on Wildberger et al. [38], introduces the Interventional Kullback-Leibler (IKL) divergence, a novel framework for comparing causal models. Instead of treating intervention-induced distribution shifts as an obstacle, this work leverages them as a source of information for model validation. The IKL divergence is designed to operate directly on multi-environment data with known intervention targets, providing a principled measure of both structural and distributional discrepancies between causal hypotheses. This offers a precise tool to evaluate model alignment where traditional metrics fall short, advancing empirical inference by extending its theoretical reach and enhancing causal interpretation in dynamic systems.

On the Interventional Kullback-Leibler Divergence

Jonas Wildberger, Siyuan Guo, Arnab Bhattacharyya, and Bernhard Schölkopf.

CLeaR 2023: 2nd Conference on Causal Learning and Reasoning.

Scaling Simulation-Based Inference with FMPE Chapter 3, based on Wildberger et al. [39], presents Flow Matching Posterior Estimation (FMPE), a simulation-based inference method utilizing continuous normalizing flows to overcome the scalability limitations of discrete flows. This contribution tackles the methodological challenge of training neural density estimators for high-dimensional scientific problems, achieving a 30% reduction in training time while improving accuracy on GW tasks and SBI benchmarks, supported by a KL bound ensuring mass-covering posteriors. FMPE enhances empirical inference's computational scalability, broadening its applicability to complex scientific domains.

Flow Matching for Scalable Simulation-Based Inference

Jonas Wildberger[*], Maximilian Dax[*], Simon Buchholz[*], Stephen R. Green, Jakob H. Macke, and Bernhard Schölkopf.

NeurIPS 2023: Advances in Neural Information Processing Systems.

[*]Shared first author, equal contribution.

Adapting to Noise Shifts in Gravitational Wave Inference Chapter 4, based on Wildberger et al. [40], develops a probabilistic model for adapting flow-based gravitational wave (GW) inference to time-varying noise distributions, a critical need for real-time astrophysical analysis. Addressing the applied challenge of non-stationarity, it focuses on the DINGO framework to forecast future power spectral densities (PSDs) using historical data and a single reference PSD, enabling accurate inference across 37 real O3 events without retraining, and latency in the order of seconds. By broadening PSD support via latent space operations, it ensures robustness to detector variability, advancing empirical inference's reliability in GW astronomy across extended temporal and noise scales.

Adapting to Noise Distribution Shifts in Flow-Based Gravitational-Wave Inference

Jonas Wildberger, Maximilian Dax, Stephen R. Green, Jonathan

Gair, Michael Pürrer, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf.

Physical Review D: 107, 084046 (2023).

On the Interventional Kullback-Leibler Divergence

2

Abstract

Modern machine learning approaches excel in static settings where a large amount of i.i.d. training data are available for a given task. In a dynamic environment, though, an intelligent agent needs to be able to transfer knowledge and re-use learned components across domains. It has been argued that this may be possible through causal models, aiming to mirror the modularity of the real world in terms of independent causal mechanisms. However, the true causal structure underlying a given set of data is generally not identifiable, so it is desirable to have means to quantify differences between models (e.g., between the ground truth and an estimate), on both the observational and interventional level.

In the present work, we introduce the Interventional Kullback-Leibler (IKL) divergence to quantify both structural and distributional differences between models based on a finite set of multi-environment distributions generated by interventions from the ground truth. Since we generally cannot quantify all differences between causal models for every finite set of interventional distributions, we propose a sufficient condition on the intervention targets to identify subsets of observed variables on which the models provably agree or disagree.

About This Chapter

This chapter has been adapted from Wildberger et al. [38].

On the Interventional Kullback-Leibler Divergence

Jonas Wildberger, Siyuan Guo, Arnab Bhattacharyya and Bernhard Schölkopf.

CLeaR 2023: 2nd Conference on Causal Learning and Reasoning.

2.1 Introduction

Classical machine learning methods are concerned with learning a distribution P (or properties thereof) from a large i.i.d. sample using a simpler model Q. Recent advances in causality have shown how access to heterogeneous data from multiple environments can help uncover causal relationships within the data and address problems like adversarial robustness or domain generalization [17, 41–45]. In this framework, distributional shifts emerge from interventions in some underlying causal model $\mathscr{P}$. Identifying the true causal model, however, is difficult even in the multi-environment regime, requiring many environments and (often unverifiable) assumptions on the type of interventions present. One thus often needs to be satisfied with an approximate model $\mathscr{Q}$, which is *close* to a hypothetical ground truth $\mathscr{P}$.

[17]: Schölkopf et al. (2021), *Towards Causal Representation Learning*
[41]: Peters et al. (2017), *Elements of Causal Inference: Foundations and Learning Algorithms*
[42]: Zhang et al. (2013), 'Domain Adaptation under Target and Conditional Shift'
[43]: Kügelgen et al. (2021), *Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style*
[44]: Eastwood et al. (2022), *Probable Domain Generalization via Quantile Risk Minimization*
[45]: Guo et al. (2022), 'Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data'

In the present work, we propose a quantitative notion of observational and interventional *closeness* between two causal models $\mathscr{P}$ and $\mathscr{Q}$. Extending the classical Kullback-Leibler divergence, our *Interventional Kullback-Leibler divergence* $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q})$ is computed between a finite set of interventional distributions with known intervention targets $((P^e, \mathcal{I}_e))_{e\in\mathcal{E}}$ from $\mathscr{P}$, denoted as $\mathscr{P}_{\mathcal{E}}$, and our model estimate $\mathscr{Q}$, consisting of a single reference distribution estimate Q and graphical model estimate G_Q. Intuitively, we can understand D_{IKL} as quantifying the average deviation between distributions under interventions and our model's predictions, had the interventions also happened on our model estimate. We thus quantify how good our model estimate is beyond observational.

Minimizing the IKL divergence over our estimate $\mathscr{Q}$ with respect to arbitrary interventions is equivalent to finding the true causal model $\mathscr{P} = \mathscr{Q}$. In practice, having access to only a limited amount of multi-environment distributions presents interesting challenges regarding which structural differences are identifiable. We propose a sufficient condition on the intervention targets trading-off a priori knowledge about $\mathscr{P}$ to establish the equivalence $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0 \iff \mathscr{P} = \mathscr{Q}$. If this condition is not known to be satisfied, we further show which differences with respect to subsets of variables can be provably identified.

Previous approaches [34–36] define interventional distances via differences between two causal graphs. In contrast, our IKL divergence measure (1) does not require access to the true causal graph G_P and (2) also takes into account distributional differences. Further, unlike [46, 47], there is no need to experimentally perform (hard-)interventions for the evaluation of our distance; instead we allow for arbitrary unknown mechanism changes, as long as the targets of the interventions are discerned.

[34]: Acid et al. (2003), 'Searching for Bayesian Network Structures in the Space of Restricted Acyclic Partially Directed Graphs'
[35]: Tsamardinos et al. (2006), 'The max-min hill-climbing Bayesian network structure learning algorithm'
[36]: Peters et al. (2013), 'Structural Intervention Distance (SID) for Evaluating Causal Graphs'
[46]: Acharya et al. (2018), *Learning and Testing Causal Models with Interventions*
[47]: Peyrard et al. (2020), *A Ladder of Causal Distances*

The remainder of this work is organized as follows: In Section 2.2, we introduce the problem settings and the assumptions to set the stage for the following theoretical analysis. Section 2.3 contains a

discussion about related work and our main results, studying the Interventional Kullback-Leibler divergence first for the example of a known graph and then in the general setting. Finally, in Section 2.4, we discuss the operational significance of the IKL divergence and various applications related to the area of multi-environment causal learning. Summarizing, our main contributions are:

- We introduce the Interventional Kullback-Leibler divergence D_{IKL} that quantifies distributional and structural differences between a given set of multi-environment distributions $(P^e)_{e\in\mathcal{E}}$ from the ground truth $\mathscr{P}$ and our inferred model $\mathscr{Q}$.
- We discuss various properties and applications of the IKL divergence related to tasks in causal inference.
- We prove a theorem establishing equivalence between $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) = 0$ and equality between causal models $\mathscr{P} = \mathscr{Q}$ under certain conditions on the environments. We thereby trade off (partial) knowledge about $\mathscr{P}$ and access to interventional distributions $(P^e)_{e\in\mathcal{E}}$. This gives rise to a sufficient condition that can be verified using only the inferred model $\mathscr{Q}$.

2.2 Problem Setting and Preliminaries

We begin by reviewing the causal formalism used below. This is based on interventional distribution shifts, modelled as causal graphical models [8].

[8]: Pearl (2009), *Causality: Models, Reasoning and Inference*

2.2.1 Causal Terminology

Definition 2.2.1 (Causal Graphical Model) *A Causal Graphical Model (CGM) (P, G_P) over a variable set $\mathbf{X} = \{X_1, \dots, X_d\}$ consists of a distribution P and a directed acyclic graph (DAG) $G_P = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V} := \{1, \dots, d\}$ is an index set and $\mathbf{E} \subseteq \mathbf{V}^2$ is an edge set, such that $(i, j) \in \mathbf{E}$ if and only if X_i directly causes X_j. We say that two CGMs $(P, G_P), (Q, G_Q)$ are the same $(P, G_P) = (Q, G_Q)$ if and only if $P = Q$ and $G_P = G_Q$.*

Definition 2.2.2 (Markov Factorization) *Given a joint distribution P and a DAG G, we say P satisfies the Markov factorization property[a] with respect to G (or is Markovian with respect to G) if*

$$P(X_1, \dots, X_d) = \prod_i P(X_i \mid \mathbf{PA}_i^G), \tag{2.1}$$

where $\mathbf{PA}_i^G$ denotes the set of parents of X_i in G.

[a] We assume that all distributions have densities with respect to the Lebesgue measure.

While Definition 2.2.2 allows us to deduce properties of the distribution P from the corresponding graph G_P, the following faithfulness property allows us to perform inference in the converse direction, i.e. finding the structure of the graph given the distribution.

Definition 2.2.3 (Causal Faithfulness) *A joint distribution P is called faithful with respect to a DAG G if any conditional independence relationship in P is implied by d-separation in G [41, 48].*

[41]: Peters et al. (2017), *Elements of Causal Inference: Foundations and Learning Algorithms*
[48]: Spirtes et al. (2000), *Causation, Prediction, and Search*

Assumption 2.2.1 *In the following, we consider the set of distributions that are both Markovian and faithful with respect to some causal DAG G, i.e. given an observed distribution P, there exists a compatible graph G_P.*

A justification for the faithfulness assumption is given by Theorem 3.5 [49] stating that almost all distributions P that are Markovian with respect to G are also faithful to G.

[49]: Koller et al. (2009), *Probabilistic Graphical Models: Principles and Techniques*

2.2.2 Multi-environment learning

Throughout this work, we take advantage of data from multiple environments. Specifically, we assume that given a set of multi-environment distributions $(P^e)_{e\in\mathcal{E}}$ over some variable set $\mathbf{X}$, there exists an underlying ground truth CGM $\mathscr{P} := (P, G_P)$ giving rise to $(P^e)_{e\in\mathcal{E}}$. The generative process behind these distributions is assumed to satisfy the below [41, 50]:

[41]: Peters et al. (2017), *Elements of Causal Inference: Foundations and Learning Algorithms*
[50]: Schölkopf et al. (2012), 'On causal and anticausal learning'

Assumption 2.2.2 (Independent Causal Mechanisms (ICM)) *A causal generative process of a system of variables is composed of autonomous modules that do not inform or influence each other. Mathematically speaking, given a Markov factorization of the joint distribution as in* (2.1)*, the causal conditionals (also called* mechanisms*) $P(X_i \mid \mathbf{P}A_i^{G_P})$ represent autonomous modules in the sense that*

(i) intervening on one $P(X_i \mid \mathbf{P}A_i^{G_P})$ will not change other mechanisms $P(X_j \mid \mathbf{P}A_j^{G_P})$, $i \neq j$,
(ii) knowing one mechanism will not provide information about other mechanisms.

Under Assumption 2.2.2, environment shifts act independently on each mechanism $P(X_i \mid \mathbf{P}A_i^{G_P})$. Thus, for each environment, $e \in \mathcal{E}$ there exists a well-defined set of mechanisms that are altered by the environment.

Assumption 2.2.3 (Multi-Environment Distributions) *For each environment $e \in \mathcal{E}$, 'Nature' first chooses a subset of variables to intervene upon $\mathcal{I}_e \subseteq [d]$, where $[d] := \{1, \dots, d\}$. For each intervened variable, 'Nature' then chooses a mechanism shift $\tilde{P}(X_i \mid \mathbf{P}A_i) \in \mathcal{F}$*

to perform, where $\mathcal{F}$ is a family of functions. The transition from P to P^e thus maps $P(X_i \mid \mathbf{P}A_i) \mapsto \tilde{P}(X_i \mid \mathbf{P}A_i)$ if $i \in \mathcal{I}_e$ and $P(X_i \mid \mathbf{P}A_i) \mapsto P(X_i \mid \mathbf{P}A_i)$ otherwise, and we may represent the joint distribution in environment e as:

$$P^e(X_1, \ldots, X_d) = \prod_{i \in \mathcal{I}^e} \tilde{P}(X_i \mid \mathbf{P}A_i) \cdot \prod_{i \in [d] \setminus \mathcal{I}^e} P(X_i \mid \mathbf{P}A_i) \quad (2.2)$$

Such mechanism shifts arise when the context of the observed distribution changes. This may happen for example as the result of different outer circumstances like climate or time [51] or by explicitly intervening on an experiment. Such shifts are generally called *interventions*. Special cases include *soft interventions* where the structure of the graph is left invariant, i.e., X_i does not become conditionally independent of any of its parents, and *hard interventions* where $\tilde{P}(X_i \mid \mathbf{P}A_i) = \delta(X_i - x)$. In any case, P^e will still be Markovian with respect to the ground truth graph G_P. However, it may no longer be faithful to it, unless the interventions are soft.

[51]: Mooij et al. (2016), 'Joint Causal Inference from Multiple Contexts'

Our final assumption concerns the existence of hidden confounder variables. We here relax the causal sufficiency assumption (stating that there are no unobserved confounders) to pseudo causal sufficiency. This ensures in particular that there are no changes in observed distributions that are not explained by the environment and the causal graph G_P.

Assumption 2.2.4 (Pseudo causal sufficiency) *The effect of any unobserved confounder can be completely explained by the environment variable, i.e. in any given environment $e \in \mathcal{E}$ potential unobserved confounders are fixed, and causal sufficiency is satisfied [52].*

[52]: Huang et al. (2019), 'Causal Discovery from Heterogeneous/Nonstationary Data with Independent Changes'

2.3 The Interventional KL divergence

In this section, we introduce our proposed Interventional KL divergence. We begin by reviewing the problem of defining a distance between two CGMs $\mathscr{P} = (P, G_P)$ and $\mathscr{Q} = (Q, G_Q)$ and discussing related work. We then introduce the IKL divergence first for a known true causal graph G_P. This definition is subsequently generalized to an unknown G_P. Finally, we establish the equivalence $\mathscr{P} = \mathscr{Q} \iff D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$, and derive a corollary in a partial information setting. Thereby, $\mathscr{Q}$ takes the role of an 'estimate' for the unknown $\mathscr{P}$, i.e. we know G_Q and have access to and can evaluate the density of Q. For complete proofs of the results in this section, we refer to Appendix B.1.

2.3.1 Related Work

Defining a general distance between two causal graphical models $(P, G_P), (Q, G_Q)$ requires comparing them both with respect to

distributional differences between P, Q and structural differences between their underlying causal graphs G_P, G_Q. Early work addressing this problem considered only structural differences. The Structural Hamming Distance (SHD) [34, 35] counts the number of different edges between two graphs, which has been found to provide limited insights about the effect on interventional distributions [36]. They instead proposed the Structural Intervention Distance (SID) to address this limitation by counting instead the number of different interventional distributions induced by two causal graphs. If the true causal graph G_P is unknown, it is generally unclear how to define a valid distance measure between the causal models. One avenue to solve this problem is leveraging multi-environment distributions that under Assumptions 2.2.3 and 2.2.4 can provide insights about the structural properties of the true causal graph. [46, 47] propose a score for comparing general causal graphical models based on their distributional distance after performing hard interventions on both of them. In section 2.3.3, we show that under these assumptions, the Interventional KL divergence entails their distance as a special case.

[34]: Acid et al. (2003), 'Searching for Bayesian Network Structures in the Space of Restricted Acyclic Partially Directed Graphs'
[35]: Tsamardinos et al. (2006), 'The max-min hill-climbing Bayesian network structure learning algorithm'
[36]: Peters et al. (2013), 'Structural Intervention Distance (SID) for Evaluating Causal Graphs'

Another related area of research is that of multi-environment causal discovery, which seeks to uncover the true causal graph G_P by relaxing the i.i.d. assumption to exchangeable data [45] or interventional data [53–55]. Since these methods usually only provide probabilistic guarantees about recovering the true causal graph, a causal distance metric can be seen as an evaluation tool to track their progress.

[45]: Guo et al. (2022), 'Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data'
[53]: Eaton et al. (2007), 'Exact Bayesian structure learning from uncertain interventions'
[54]: Ghassami et al. (2018), 'Multi-domain Causal Structure Learning in Linear Systems'
[55]: He et al. (2016), *Causal Network Learning from Multiple Interventions of Unknown Manipulated Targets*

Finally, there is related work discussing the identifiability of causal effects and interventional Markov equivalence classes [56–58]. Our work is tangent to this line of research in that we also seek to understand which kind of interventions are necessary in a general multi-environment setting to distinguish causal models and define a valid distance measure between them.

[56]: Tian et al. (2002), 'A General Identification Condition for Causal Effects'
[57]: Hauser et al. (2011), 'Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs'
[58]: Hauser et al. (2015), 'Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs'

2.3.2 The IKL divergence for a known causal graph

Before defining the Interventional KL divergence for general causal models $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ in the next section, we build intuition by first discussing the case of a known ground truth causal graph G_P of $\mathscr{P}$. This means that $G_Q = G_P$, and Q is Markovian with respect to G_P. Furthermore, any differences between $\mathscr{P}, \mathscr{Q}$ are given by distributional differences between P, Q.

Throughout this work, we use the Kullback-Leibler divergence for quantifying distributional differences between P and Q, assuming that P is absolutely continuous with respect to Q:

$$D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) = \int_{\mathcal{X}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}, \tag{2.3}$$

It is known that $D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) = 0$ if and only if $P = Q$. Now suppose that P^e arises from P via a mechanism shift on the variables $\mathbf{X}_{\mathcal{I}_e}$ for a subset $\mathcal{I}_e \subseteq [d]$, i.e.

$$P^e(X_i|\mathbf{P}A_i^{G_P}) = P(X_i|\mathbf{P}A_i^{G_P}) \quad \text{if and only if} \quad i \in [d]\backslash\mathcal{I}_e. \tag{2.4}$$

From the chain rule for relative entropy (see e.g. Thm 2.5.3 [59], [60]), we can deduce the following decomposition to understand how the effect of the mechanism shifts shows up in $D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q(\mathbf{X}))$. From here on, we denote $\mathbb{E}_{x\sim P(x)}[X]$ as $\mathbb{E}_{P(x)}[X]$.

[59]: Cover et al. (2006), *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*

[60]: Budhathoki et al. (2021), 'Why did the distribution change?'

Lemma 2.3.1 *Given causal model $\mathscr{P} = (P, G_P)$ assume that P^e emerges from P via an environment shift according to Assumption 2.2.3. If Q is Markovian with respect to G_P, we have the following decomposition*

$$\begin{aligned} &D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q(\mathbf{X})) = \\ &\quad \sum_{i\in\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{P}a_i^{G_P})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{P}a_i^{G_P})\|Q(X_i \mid \mathbf{P}a_i^{G_P})\right)\right] \\ &\quad + \underbrace{\sum_{i\in[d]\backslash\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{P}a_i^{G_P})}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{P}a_i^{G_P})\|Q(X_i \mid \mathbf{P}a_i^{G_P})\right)\right]}_{=:D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}}\|\mathscr{Q})} \end{aligned} \tag{2.5}$$

As a special case we have for $\mathcal{I}_e = \emptyset$:

$$\begin{aligned} &D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) = \\ &\quad \sum_{i\in[d]} \mathbb{E}_{P(\mathbf{P}a_i^{G_P})}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{P}a_i^{G_P})\|Q(X_i \mid \mathbf{P}a_i^{G_P})\right)\right] \end{aligned} \tag{2.6}$$

The first sum in equation (2.5) goes over the intervened mechanisms and therefore fully characterizes the changes that are directly due to environment shifts. Note that it vanishes if we could perfectly replicate the environment shift on our model estimate Q. The second sum, which is a special case of the Interventional KL divergence to be defined more generally below, goes over the non-intervened variables, quantifying the distributions' distance resulting downstream from intervened variables due to the fact that the expectations over parents are taken with respect to intervened distributions.

Now assume that P is unobserved, but we still want to understand whether Q is a valid model. Since, P, Q are Markovian with respect to the same causal DAG G_P, we can express agreement between P and Q equivalently in terms of vanishing Interventional KL divergence $D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}}\|\mathscr{Q})$. First suppose that $P = Q$. The first sum in equation (2.5) can become arbitrarily large, inflating the KL divergence $D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q(\mathbf{X}))$. However, the second sum in equation (2.5) vanishes, i.e. $D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}}\|\mathscr{Q}) = 0$, since agreement

between the conditional distributions is irrespective of the parent distributions taken in the expectation.* The IKL divergence therefore discounts the additional difference between P^e and Q introduced by the environment shift.

Conversely, assume that $D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}} \| \mathscr{Q}) = 0$. It is clear that this does not generally imply that $P = Q$, since we only sum over a subset of the local divergences composing the difference between the joint distribution P and Q, as shown in equation (2.6). But given multiple interventional distributions $(P^e)_{e \in \mathcal{E}}$, such that each mechanism constituting $P(\mathbf{X})$ remains unchanged in (at least) one environment, we can obtain a valid distance measure between P and Q via averaging. These observations are formalized in the following definition and lemma.

Definition 2.3.1 (Interventional KL divergence for shared causal graphs) *Let $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_P)$ be two causal models sharing the same causal graph. Further, assume that we have access to a set of interventional distributions $\mathscr{P}_{\mathcal{E}} = ((P^e, \mathcal{I}_e))_{e \in \mathcal{E}}$ generated from $\mathscr{P}$. We define the Interventional KL divergence as*

$$D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{p}a_i^{G_P})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{p}a_i^{G_P}) \| Q(X_i \mid \mathbf{p}a_i^{G_P}) \right) \right] \tag{2.7}$$

$$\overset{(2.4)}{=} \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{p}a_i^{G_P})} \left[D_{\mathrm{KL}} \left(P(X_i \mid \mathbf{p}a_i^{G_P}) \| Q(X_i \mid \mathbf{p}a_i^{G_P}) \right) \right] \tag{2.8}$$

If $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$, we say that our model $\mathscr{Q}$ is interventionally equivalent with $\mathscr{P}$ with respect to $\mathcal{E}$, denoted as $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q}$.

Lemma 2.3.2 *Under Assumptions 2.2.3, 2.2.4 and that no variable is always intervened upon, i.e. $\bigcap_{e \in \mathcal{E}} \mathcal{I}_e = \emptyset$, we can conclude that $\mathscr{P}, \mathscr{Q}$ are interventionally equivalent, $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q}$, if and only if $\mathscr{P} = \mathscr{Q}$.*

Note that if the reference distribution $P(\mathbf{X})$ is included in our set of interventional distributions, i.e. there exists $e \in \mathcal{E}$ such that $\mathcal{I}_e = \emptyset$, the KL divergence $D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X}))$ makes up one of the terms in the IKL divergence. Thus, the equivalence $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q} \iff \mathscr{P} = \mathscr{Q}$ is trivially satisfied. We now illustrate an application of the case where we do not have access to the reference distribution P of the CGM (P, G_P).

* Specifically, they need to coincide on the union of support sets $\mathrm{supp}(P(\mathbf{P}A_i^{G_P})) \cup \mathrm{supp}(P^e(\mathbf{P}A_i^{G_P}))$

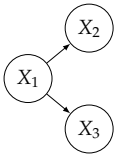

Figure 2.1: A sample DAG over variables X_1, X_2, X_3.

Example 2.3.1 (Partial Observability) In this example, we illustrate how the ideas presented above formalize the common conception of recovering the joint distribution $P(X_1, X_2, X_3)$ without observing all variables at once — given that we know it factorizes in a non-trivial way. So, let the causal graph G_P be given by the DAG in Figure 2.1. By the Markov factorization (2.1), we know that $P(\mathbf{X}) = P(X_1) \cdot P(X_2 \mid X_1) \cdot P(X_3 \mid X_1)$.

Now assume that we do not have access to this joint distribution, but instead we are only provided with two marginal distributions $P(X_1, X_2), P(X_1, X_3)$ and the knowledge of the underlying DAG. Even with no interventions, observing the two marginal distributions is sufficient to recover the joint distribution $P(X_1, X_2, X_3)$, since it is possible to separately estimate $P(X_2 \mid X_1), P(X_3 \mid X_1)$ from the marginals and $P(X_1)$ from either of the marginals. This phenomenon is correctly captured in our IKL metric by modelling unobserved variables as interventions, say $P^{e_1}(\mathbf{X}) = P(X_1, X_2)\tilde{P}(X_3)$, $P^{e_2}(\mathbf{X}) = P(X_1, X_3)\tilde{P}(X_2)$:

$$
\begin{aligned}
&D_{\mathrm{KL}}(P(X_1, X_2) \mid Q(X_1, X_2)) + D_{\mathrm{KL}}(P(X_1, X_3) \mid Q(X_1, X_3)) \\
&\quad \overset{(2.6)}{=} D_{\mathrm{KL}}(P(X_1) \| Q(X_1)) \\
&\qquad + \mathbb{E}_{P(X_1)}[D_{\mathrm{KL}}(P(X_2 \mid X_1) \| Q(X_2 \mid X_1))] \\
&\qquad + D_{\mathrm{KL}}(P(X_1) \| Q(X_1)) \\
&\qquad + \mathbb{E}_{P(X_1)}[D_{\mathrm{KL}}(P(X_3 \mid X_1) \| Q(X_3 \mid X_1))] \\
&\quad \overset{(2.7)}{=} 2 \cdot D_{\mathrm{IKL}}(\mathscr{P}_{\{e_1,e_2\}} \| \mathscr{Q})
\end{aligned}
\tag{2.9}
$$

We can now minimize each of the terms in equation (2.9) with respect to our model distribution Q. By Lemma 2.3.2, this will also minimize the distance between the joint distributions P, Q without observing all variables at once. Note that the term $D_{\mathrm{KL}}(P(X_1) \| Q(X_1))$ occurs twice in equation (2.9). This means that we could still identify the joint distribution in the same way, if an intervention had happened on X_1 in either $P(X_1, X_2)$ or $P(X_1, X_3)$.

2.3.3 The IKL divergence in the general case

In the previous section, we discussed how the Interventional KL divergence quantifies differences between two causal models $\mathscr{P} =$

(P, G_P) and $\mathscr{Q} = (Q, G_P)$ when G_P is known. Since this is generally not the case, we now generalize Definition 2.3.1 to an unknown G_P. To this end, we need to adapt Lemma 2.3.1 determining the decomposition of the KL divergence, since the conditioning sets are also unknown. As it turns out, this decomposition still holds when conditioning on $\mathbf{P}A_i^{G_Q}$ as long as P is Markovian with respect to G_Q. Otherwise, there is a residual term quantifying the difference of P from being Markovian, for which we introduce the notion of a *Markov Projection*.

Definition 2.3.2 (Markov Projection) *Let $\mathcal{A}$ be the space of probability distributions over the variable set $\mathbf{X}$. The* Markov projection *(or M-*projection*) of a probability distribution P onto a DAG G is the mapping $\pi_G : \mathcal{A} \to \mathcal{A}$ defined by:*

$$\pi_G(P) = \underset{Q \text{ Markov w.r.t. } G}{\operatorname{argmin}} D_{\mathrm{KL}}(P \| Q)$$

The following Lemma 2.3.3 explicitly describes the Markov projection onto a DAG G.

Lemma 2.3.3 *Given a distribution P and a DAG G:*

$$\pi_G(P)(\mathbf{X}) = \prod_{i=1}^{d} P(X_i \mid \mathbf{P}A_i^G). \tag{2.10}$$

From the Markov factorization property Definition 2.2.2 it follows that $P = \pi_G(P)$ if and only if P is Markovian with respect to G. Otherwise, we have $P \neq \pi_G(P)$, since there are variables X_i, X_j such that $X_i \not\perp\!\!\!\perp X_j \mid \mathbf{P}A_i^G$ with respect to P, but $X_i \perp\!\!\!\perp X_j \mid \mathbf{P}A_i^G$ in $\pi_G(P)$. If we now replace G_P by G_Q in Lemma 2.3.1 and allow for general (non-Markovian) distributions P, the deviation from being Markovian enters the equation as an additional term:

Lemma 2.3.4 *Given two CGMs $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$, assume that P^e emerges from P via an environment shift according to Assumption 2.2.3. We then have the following decomposition*

$$\begin{aligned} &D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q(\mathbf{X})) = \\ &\quad \sum_{i \in \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{P}a_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{P}a_i^{G_Q}) \| Q(X_i \mid \mathbf{P}a_i^{G_Q}) \right) \right] \\ &\quad + \sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{P}a_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{P}a_i^{G_Q}) \| Q(X_i \mid \mathbf{P}a_i^{G_Q}) \right) \right] \\ &\quad + D_{\mathrm{KL}} \left(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)(\mathbf{X}) \right) \end{aligned} \tag{2.11}$$

As a special case we have for $\mathcal{I}_e = \varnothing$

$$D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) = \sum_{i \in [d]} \mathbb{E}_{P(\mathbf{p}a_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P(X_i \mid \mathbf{p}a_i^{G_Q}) \| Q(X_i \mid \mathbf{p}a_i^{G_Q}) \right) \right] + D_{\mathrm{KL}} \left(P(\mathbf{X}) \| \pi_{G_Q}(P)(\mathbf{X}) \right) \tag{2.12}$$

With this generalized decomposition, we adapt Definition 2.3.1 of the Interventional KL divergence correspondingly to arbitrary CGMs $\mathscr{P}, \mathscr{Q}$.

Definition 2.3.3 (Interventional KL divergence) *Let $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_Q)$ be two causal models. Assume that we have access to a set of interventional distributions $\mathscr{P}_{\mathcal{E}} = ((P^e, \mathcal{I}_e))_{e \in \mathcal{E}}$ generated from $\mathscr{P}$. We then define the Interventional KL divergence as*

$$D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[\sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{p}a_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{p}a_i^{G_Q}) \| Q(X_i \mid \mathbf{p}a_i^{G_Q}) \right) \right] + D_{\mathrm{KL}} \left(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)(\mathbf{X}) \right) \right] \tag{2.13}$$

Note that this definition is consistent with our previous notion of interventional distance in the case where $G_P = G_Q$, see Definition 2.3.1. If $G_P \neq G_Q$, Lemma 2.3.4 implies that we can capture differences between the causal models by comparing the Markov factorizations of P^e and Q with respect to G_Q (first line in equation (2.13)), if we also account for potential deviations from P satisfying this factorization (second line in equation (2.13)).

The definition of D_{IKL} went under the premise that the effect of the interventions that we consider is unknown, i.e. only the joint distributions $P^e(\mathbf{X})$ and intervention targets $\mathcal{I}_e$ are observed, but we don't know the mapping $P(X_i \mid \mathbf{P}A_i^{G_P}) \mapsto \tilde{P}(X_i \mid \mathbf{P}A_i^{G_P})$. If we know the mechanism shifts that happened between the distribution $P(\mathbf{X})$ and the interventional distributions $P^e(\mathbf{X})$ and we are able to accurately reproduce them in our model $\mathscr{Q} = (Q, G_Q)$, then the D_{IKL} collapses to a simpler form, namely the average of the KL divergences between the joint distributions $P^e(\mathbf{X})$ and $Q(\mathbf{X})$.

Theorem 2.3.5 *[Known interventions] Given causal models $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_Q)$ and a set of interventional distributions $\mathscr{P}_{\mathcal{E}} = ((P^e, \mathcal{I}_e))_{e \in \mathcal{E}}$, define Q^e for each environment $e \in \mathcal{E}$ as follows:*

$$Q^e(\mathbf{X}) = \prod_{i \notin \mathcal{I}_e} Q(X_i \mid \mathbf{P}A_i^{G_Q}) \cdot \prod_{j \in \mathcal{I}_e} P^e(X_j \mid \mathbf{P}A_j^{G_Q}) \tag{2.14}$$

where $P^e(X_j \mid \mathbf{PA}_j^{G_Q})$ denotes the mechanisms changed as the result of the environment shift. Then we have that

$$D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q^e(\mathbf{X})) \tag{2.15}$$

Sketch. If we apply the same mechanism shifts to Q^e that appear between P and P^e, we can use the decomposition of the KL divergence in Lemma 2.3.4 to see that these terms cancel out. The sum over the remaining terms then equals our definition of the Interventional KL divergence. □

In particular, this Theorem applies to hard interventions on single variables that were considered by [47].

2.3.4 Interventional differences in the IKL divergence

In the previous sections, we have defined the Interventional KL divergence for two general causal graphical models $\mathscr{P}, \mathscr{Q}$ and related it to the (standard) KL divergence. However, it still remains to show that the IKL divergence defines a valid distance between $\mathscr{P}, \mathscr{Q}$, i.e. that $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q} \iff D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0 \iff \mathscr{P} = \mathscr{Q}$ under suitable conditions on $\mathcal{E}$. This equivalence will be subject to this section. In particular, we will focus on differences beyond Markov equivalence, since statistically identifiable differences between $\mathscr{P}, \mathscr{Q}$ are already encoded in the purely observational distance $D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X}))$. For simplicity, we assume in the following that the empty intervention $\mathcal{I} = \varnothing$ is included in the environment. Further, we make use of our assumptions on the multi-environment distributions (Assumptions 2.2.3, 2.2.4), which imply for Markov equivalent G_P, G_Q that the following statements are equivalent:

(i) For all e with $i \notin \mathcal{I}_e$:

$$\mathbb{E}_{P^e(\mathbf{PA}_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{PA}_i^{G_Q}) \| P(X_i \mid \mathbf{PA}_i^{G_Q}) \right) \right] = 0$$

(i) $\mathbf{PA}_i^{G_Q} = \mathbf{PA}_i^{G_P}$.

As shown in Corollary 2.3.8 and Example 2.3.2, this equivalence does not hold given only a limited set of interventional distributions. A comprehensive characterization of differences between $P^e(X_i \mid \mathbf{PA}_i^{G_Q})$ and $P(X_i \mid \mathbf{PA}_i^{G_Q})$ can be given based on d-separation in the augmented DAG $G_{P,\mathbf{X} \cup \{e\}}$, assuming faithfulness [61]. We adopt the faithfulness assumption in the following but, since G_P is assumed to be generally unknown, we here propose a simpler sufficient condition that trades off knowledge about G_P with access to interventional distributions. It is based on the following definition of (directed) unblocked paths between a variable X_i and intervened variables.

[61]: Perry et al. (2022), *Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis*

Definition 2.3.4 *Let $\mathscr{P} = (P, G_P)$ be a CGM and $e \in \mathcal{E}$ an environment. For some (potentially different) DAG G over the same variables $\mathbf{X}$, we say that there exists a (directed) unblocked path $e \xrightarrow{G} X_i$ in G given some variable set $\mathbf{Z} \subseteq \mathbf{X}$, if there exists $k \in \mathcal{I}_e$ and a directed path $X_k \to \ldots \to X_i$ in G such that no element of $\mathbf{Z}$ intersects with the path. If $k = i$, this condition is trivially satisfied.*

So, in particular, $e \xrightarrow{G_P} X_i$ implies that the variable X_i is d-connected to an intervened variable given conditioning set $\mathbf{Z}$. Assuming faithfulness, we can conclude that the environment has an influence on X_i given $\mathbf{Z}$ and thus $P^e(X_i|\mathbf{Z}) \neq P(X_i|\mathbf{Z})$. If G_Q shares the same skeleton with G_P, this definition yields a sufficient condition for the mechanism $P^e(X_i \mid \mathbf{P}A_i^{G_Q})$ to be different from $P(X_i \mid \mathbf{P}A_i^{G_Q})$.

Lemma 2.3.6 *Let $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_Q)$ be CGMs, such that G_P and G_Q share the same skeleton and $X_i \xrightarrow{G_Q} X_j$, $X_i \xleftarrow{G_P} X_j$ be a flipped edge between G_P and G_Q. Further, let P^e be an interventional distribution generated from $\mathscr{P}$.*

(i) If there exists an unblocked path $e \xrightarrow{G_P} X_i$ given $\mathbf{P}A_j^{G_Q} \backslash X_i$, then $P^e(X_j \mid \mathbf{P}A_j^{G_Q}) \neq P(X_j \mid \mathbf{P}A_j^{G_Q})$.

(ii) If there exists an unblocked path $e \xrightarrow{G_P} X_j$ given $\mathbf{P}A_i^{G_Q}$, then $P^e(X_i \mid \mathbf{P}A_i^{G_Q}) \neq P(X_i \mid \mathbf{P}A_i^{G_Q})$.

This Lemma yields a sufficient condition for structural differences beyond Markov-equivalence to show up in the IKL-divergence. Even if $P = Q$ coincide observationally, the interventional terms $\mathbb{E}_{P^e(\mathbf{P}A_k^{G_Q})}\left[D_{\mathrm{KL}}\left(P^e(X_k \mid \mathbf{P}A_k^{G_Q}) \| Q(X_k \mid \mathbf{P}A_k^{G_Q})\right)\right] > 0 \,\text{for}\, k \in \{i, j\}$ under the conditions of this Lemma. Since these are conditions on the graphical structure of G_P, they cannot be verified if G_P is unknown. But as it turns out, it is sufficient if such paths exist in G_Q. We formalize these observations in the following Theorem:

Theorem 2.3.7 *Let $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_Q)$ be CGMs and $\mathcal{E}$ a set of environments. Now assume that for all (unoriented) edges $X_i \xrightarrow{G_Q} X_j$ there exists an environment $e \in \mathcal{E}$ such that one of the following conditions holds:*

(i) There exists an unblocked path $e \xrightarrow{G_Q} X_i$ given $\mathbf{P}A_j^{G_Q} \backslash X_i$ and $j \notin \mathcal{I}_e$

(ii) There exists an unblocked path $e \xrightarrow{G_Q} X_j$ given $\mathbf{P}A_i^{G_Q}$ and $i \notin \mathcal{I}_e$

Then, $\mathscr{P}$, $\mathscr{Q}$ are interventionally equivalent, $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q}$, if and only if $\mathscr{P} = \mathscr{Q}$.

sketch. If $\mathscr{P} = \mathscr{Q}$, we can deduce $D_{\mathrm{IKL}}(\mathscr{P}\|\mathscr{Q}) = 0$ from Assumptions 2.2.3 and 2.2.4. Conversely, if $P \neq Q$, then $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) \geq 1/|\mathcal{E}| D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) > 0$. So it suffices to show that $P = Q$, but $G_P \neq G_Q$ implies that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) > 0$. This case then follows from Lemma 2.3.6, since an unblocked path in G_Q gives rise to an unblocked path to a mis-oriented edge in G_P and thus one of the conditions of Lemma 2.3.6 is satisfied, leading to a nonzero term in the IKL divergence. □

The less we know about the structure of G_P a priori, the more interventional distributions we need to identify all structural differences between the causal models. Since interventional data can be expensive to collect, we will now show how to identify partial structural differences between $\mathscr{P}$ and $\mathscr{Q}$, given insufficient multi-environmental information to satisfy the conditions of Theorem 2.3.7. In particular, we also allow for environments where only marginal information about a subset of all variables is observed.

Corollary 2.3.8 *Let $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_Q)$ be CGMs with $P = Q$ and $\mathcal{E}$ a set of environments. Assume that we only have partial multi-environmental information in the following ways:*

1) *We only observe a subset of all variables in the causal graph $\mathbf{X}_S \subseteq \mathbf{X}$, i.e. we can only distinguish the graphical sub-models $\mathscr{P}\big|_S, \mathscr{Q}\big|_S$, induced by removing the unobserved variables from distributions and causal graphs.*
2) *We know that the conditions of Theorem 2.3.7 are only satisfied for a subset $E_{\mathcal{E}}$ of all edges in $G_Q\big|_S$.*

We define the restricted IKL divergence via:

$$D^{res}_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\sum_{i\in\mathbf{X}_{E_{\mathcal{E}}}\cap\mathcal{I}_e^C} \mathbb{E}_{P^e(\mathbf{p}a_i^{G_Q|_S})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{p}a_i^{G_Q|_S})\|Q(X_i \mid \mathbf{p}a_i^{G_Q|_S})\right)\right] \tag{2.16}$$

summing only over un-intervened variables X such that $(X, \cdot)$ or $(\cdot, X)$ is contained in $E_{\mathcal{E}}$. Then, $D^{res}_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) = 0$ implies that the graphs G_P, G_Q coincide on the edge set $E_{\mathcal{E}}$. Conversely, if $D^{res}_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) > 0$, there exists a variable $X_i \in \mathbf{X}_{E_{\mathcal{E}}}$ such that $\mathbf{P}A_i^{G_P} \neq \mathbf{P}A_i^{G_Q|_S}$.

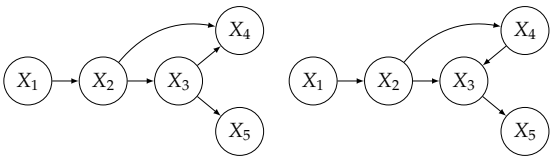

Figure 2.2: Two Markov equivalent DAGs G_P (left) and G_Q (right) over variables $X_1, \ldots, X_5$.

Example 2.3.2 Let $(P, G_P), (Q, G_Q)$ be two CGMs with $P =$

Q and graphs given in Figure 2.2. We want to compute the interventional KL divergence to evaluate the fit of (Q, G_Q) to (P, G_P) using interventional distributions P^{e_1}, P^{e_2}. Assume that in environment e_1 only variables X_2, X_3, X_5 are observed with a mechanism shift $P(X_5|X_3) \mapsto \tilde{P}(X_5|X_3)$, i.e. $E_{e_1} = \{(X_3, X_5)\}$ and $\mathbf{X}_{e_1} = \{X_3, X_5\}$ in Corollary 2.3.8. We can then compute the restricted IKL divergence using only the observed variables:

$$D^{\text{res}}_{\text{IKL}}(\mathscr{P}_{\{e_1\}} \| \mathscr{Q}) = \mathbb{E}_{P^{e_1}(X_2)} D_{\text{KL}}(P^{e_1}(X_3|X_2) \| Q(X_3|X_2)) = 0 \tag{2.17}$$

indicating that the edge (X_3, X_5) is oriented correctly in G_Q. Note that this equality holds even if the unobserved variable X_4 also happened to be intervened, since it is not conditioned upon in the restricted IKL divergence. Next, assume that in another environment e_2, variables X_2, X_3, X_4 are observed with a mechanism shift affecting $P(X_3|X_2) \mapsto \tilde{P}(X_3|X_2)$. Thus, $E_{e_2} = \{(X_4, X_3), (X_2, X_3)\}, \mathbf{X}_{e_2} = \{X_2, X_3, X_4\}$ and the restricted IKL divergence equals

$$\begin{aligned} D^{\text{res}}_{\text{IKL}}(\mathscr{P}_{\{e_2\}} \| \mathscr{Q}) = \mathbb{E}_{P^{e_2}(X_2)} D_{\text{KL}}(P^{e_1}(X_4|X_2) \| Q(X_4|X_2)) \\ + D_{\text{KL}}(P^{e_2}(X_2) \| Q(X_2)). \end{aligned} \tag{2.18}$$

The first term is strictly greater than zero due to the mis-oriented edge between X_3 and X_4, i.e. case (ii) in Lemma 2.3.6 is satisfied. Despite the correct orientation of the edge $X_2 \to X_3$, the second term may also be greater than zero. This happens precisely if the unobserved variable X_1 also happened to be intervened (case (ii) of Lemma 2.3.6 would again be satisfied). Otherwise, this term would vanish, assuring us that the edge is oriented correctly. Thus, in the presence of partial observability, nonzero terms in the interventional KL divergence are not necessarily caused by incorrect (observed) edges. Conversely, however, a vanishing restricted IKL divergence guarantees all identifiable edges $E_{\mathcal{E}}$ to be oriented correctly.

2.4 Discussion

Multi-environment distributions provide a natural setting for machine learning algorithms to both learn causal structures and leverage them to achieve robustness and out-of-distribution generalization. We have assayed how knowledge about the true causal graph can be used to facilitate learning under distribution shifts and generalize from marginal observations to properties of the joint distribution. Both of these are significant problems for learning in real-world settings, where we cannot always afford to collect a large dataset suited to a task. Conversely, multi-domain data can provide a useful learning signal to drive causal discovery.

The focus of the current paper has been a theoretical exploration of

properties of the Interventional KL divergence, and applications are thus beyond our scope. Nevertheless, we would like to discuss some possible use cases to point out directions for future work.

Interventional KL Divergence and Estimation The following result captures the operational significance of low IKL divergence.

Theorem 2.4.1 *Given causal models $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_Q)$ and a set of interventional distributions $\mathscr{P}_\mathcal{E} = ((P^e, \mathcal{I}_e))_{e\in\mathcal{E}}$, define Q^e for each environment $e \in \mathcal{E}$ as in Theorem 2.3.5.*

Suppose $D_{\mathrm{IKL}}(\mathscr{P}_\mathcal{E} \| \mathscr{Q}) \leq \epsilon$. Then, for any bounded function f mapping to $[-B, B]$ and any parameter $\rho > 0$, for at least $1 - \rho$ fraction of the environments $e \in \mathcal{E}$:

$$\left|\mathbb{E}_{P^e(\mathbf{X})}[f(\mathbf{X})] - \mathbb{E}_{Q^e(\mathbf{X})}[f(\mathbf{X})]\right| \leq \frac{B\sqrt{\epsilon}}{\rho}$$

In other words, if the IKL divergence is small, then for most environments, executing the same mechanism changes in 'Nature' and in our model $\mathscr{Q}$ would yield similar statistics. Thus, the more heterogeneous the set of environments, the more similar $\mathscr{P}$ and $\mathscr{Q}$ are in terms of their causal implications.

Evaluation tool for causal discovery methods If data from multi-environment distributions are available, there is a range of methods to recover the true causal structure asymptotically [52, 55, 61–65]. Here, the IKL divergence could help develop methods to verify a learned causal model, provided that intervention targets are known for some of the environments. This can then take into account both the interventional differences, as encoded by the learned DAG, and the observational differences encoded in the learned conditional distributions.

[52]: Huang et al. (2019), 'Causal Discovery from Heterogeneous/Nonstationary Data with Independent Changes'
[55]: He et al. (2016), *Causal Network Learning from Multiple Interventions of Unknown Manipulated Targets*
[61]: Perry et al. (2022), *Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis*
[62]: Janzing et al. (2012), 'Information-geometric approach to inferring causal directions'
[63]: Rojas-Carulla et al. (2015), 'Invariant Models for Causal Transfer Learning'
[64]: Arjovsky et al. (2019), 'Invariant risk minimization'
[65]: Krueger et al. (2021), 'Out-of-distribution generalization via risk extrapolation'

Online Learning of Causal Structures If we do not have a sufficient number of environment distributions a priori to recover the true causal graph, but intervention targets are more broadly available, we have shown in Corollary 2.3.8 that one could evaluate the fit of a model to the true causal model with respect to subsets of variables. These partially learned and verified models could then be used for inference until more interventional data become available to uncover further causal relationships. Indeed, our metric can even serve for orienting edges in the graph: If we compute $D_{\mathrm{IKL}}(\mathscr{P}_{\{e_1\}} \| (P, G_Q))$, then any non-zero term in the IKL divergence is necessarily caused by mis-oriented edges. Specifically, if for some variable $i \in [d] \backslash \mathcal{I}_e$, we have

$$\mathbb{E}_{P^e(\mathbf{P}A_i^{G_Q})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{P}A_i^{G_Q}) \| P(X_i \mid \mathbf{P}A_i^{G_Q})\right)\right] > 0, \qquad (2.19)$$

then $\mathbf{P}A_i^{G_Q} \neq \mathbf{P}A_i^{G_P}$. Thus, we can identify the true parent set by re-computing this term with respect to parent sets $\mathbf{P}A_i^G$ for all graphs G in the Markov equivalence class of G_Q. In this way, the IKL divergence is monotonic w.r.t. both distributional differences and structural differences represented by the number of correctly identified edges.

Flow Matching for Scalable Simulation-Based Inference

3

Abstract

Neural posterior estimation methods based on discrete normalizing flows have become established tools for simulation-based inference (SBI), but scaling them to high-dimensional problems can be challenging. Building on recent advances in generative modeling, we here present flow matching posterior estimation (FMPE), a technique for SBI using continuous normalizing flows. Like diffusion models, and in contrast to discrete flows, flow matching allows for unconstrained architectures, providing enhanced flexibility for complex data modalities. Flow matching, therefore, enables exact density evaluation, fast training, and seamless scalability to large architectures—making it ideal for SBI. We show that FMPE achieves competitive performance on an established SBI benchmark, and then demonstrate its improved scalability on a challenging scientific problem: for gravitational-wave inference, FMPE outperforms methods based on comparable discrete flows, reducing training time by 30% with substantially improved accuracy. Our work underscores the potential of FMPE to enhance performance in challenging inference scenarios, thereby paving the way for more advanced applications to scientific problems.

About This Chapter

This chapter has been adapted from Wildberger et al. [39].

Flow Matching for Scalable Simulation-Based Inference

Jonas Wildberger*, Maximilian Dax*, Simon Buchholz*, Stephen R. Green, Jakob H. Macke, and Bernhard Schölkopf.

NeurIPS 2023: Advances in Neural Information Processing Systems.

3.1 Introduction

The ability to readily represent Bayesian posteriors of arbitrary complexity using neural networks would herald a revolution in scientific data analysis. Such networks could be trained using simulated data and used for amortized inference across observations—bringing tractable inference and speed to a myriad of scientific models. Thanks to innovative architectures such as normalizing flows [19, 21], approaches to neural simulation-based inference (SBI) [18] have seen remarkable progress in recent years. Here, we show that modern approaches to deep generative modeling (particularly flow matching) deliver substantial improvements in simplicity, flexibility and scaling when adapted to SBI.

The Bayesian approach to data analysis is to compare observations to models via the posterior distribution $p(\theta|x)$. This gives our degree of belief that model parameters θ gave rise to an observation x, and is proportional to the model likelihood $p(x|\theta)$ times the prior $p(\theta)$. One is typically interested in representing the posterior in terms of a collection of samples, however obtaining these through standard likelihood-based algorithms can be challenging for intractable or expensive likelihoods. In such cases, SBI offers an alternative based instead on *data simulations* $x \sim p(x|\theta)$. Combined with deep generative modeling, SBI becomes a powerful paradigm for scientific inference [18]. Neural posterior estimation (NPE) [20, 66, 67], for instance, trains a conditional density estimator $q(\theta|x)$ to approximate the posterior, allowing for rapid sampling and density estimation for any x consistent with the training distribution.

The NPE density estimator $q(\theta|x)$ is commonly taken to be a (discrete) normalizing flow [19, 21], an approach that has brought state-of-the-art performance in challenging problems such as gravitational-wave inference [28]. Naturally, performance hinges on the expressiveness of $q(\theta|x)$. Normalizing flows transform noise to samples through a discrete sequence of basic transforms. These have been carefully engineered to be invertible with simple Jacobian determinant, enabling efficient maximum likelihood training, while at the same time producing expressive $q(\theta|x)$. Although many such discrete flows are universal density approximators [19], in practice, they can be challenging to scale to very large networks, which are needed for big-data experiments.

Recent studies [68, 69] propose neural posterior score estimation (NPSE), a rather different approach that models the posterior distribution with score-matching (or diffusion) networks. These techniques were originally developed for generative modeling [70–72], achieving state-of-the-art results in many domains, including image generation [73, 74]. Like discrete normalizing flows, diffusion models transform noise into samples, but with trajectories parametrized by a *continuous* "time" parameter t. The trajectories solve a stochastic differential equation [75] (SDE) defined in terms of a vector field v_t, which is trained to match the score of

[19]: Papamakarios et al. (2021), 'Normalizing Flows for Probabilistic Modeling and Inference'
[21]: Rezende et al. (2015), 'Variational Inference with Normalizing Flows'
[18]: Cranmer et al. (2020), 'The frontier of simulation-based inference'

[20]: Papamakarios et al. (2016), 'Fast ε-free Inference of Simulation Models with Bayesian Conditional Density Estimation'
[66]: Lueckmann et al. (2017), 'Flexible statistical inference for mechanistic models of neural dynamics'
[67]: Greenberg et al. (2019), 'Automatic posterior transformation for likelihood-free inference'
[28]: Dax et al. (2021), 'Real-Time Gravitational Wave Science with Neural Posterior Estimation'
[19]: Papamakarios et al. (2021), 'Normalizing Flows for Probabilistic Modeling and Inference'

[68]: Sharrock et al. (2022), 'Sequential Neural Score Estimation: Likelihood-Free Inference with Conditional Score Based Diffusion Models'
[69]: Geffner et al. (2022), 'Score Modeling for Simulation-based Inference'
[70]: Sohl-Dickstein et al. (2015), 'Deep unsupervised learning using nonequilibrium thermodynamics'
[71]: Song et al. (2019), 'Generative modeling by estimating gradients of the data distribution'
[72]: Ho et al. (2020), 'Denoising diffusion probabilistic models'
[73]: Dhariwal et al. (2021), 'Diffusion Models Beat GANs on Image Synthesis'
[74]: Ho et al. (2022), 'Cascaded Diffusion Models for High Fidelity Image Generation'
[75]: Song et al. (2020), 'Score-based generative modeling through stochastic differential equations'

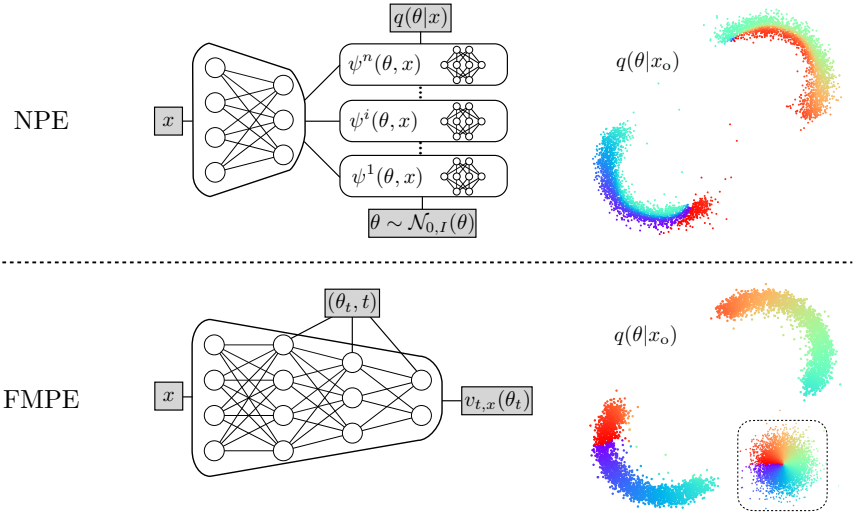

Figure 3.1: Comparison of network architectures (left) and flow trajectories (right). Discrete flows (NPE, top) require a specialized architecture for the density estimator. Continuous flows (FMPE, bottom) are based on a vector field parametrized with an unconstrained architecture. FMPE uses this additional flexibility to put an enhanced emphasis on the conditioning data x, which in the SBI context is typically high dimensional and in a complex domain. Further, the optimal transport path produces simple flow trajectories from the base distribution (inset) to the target.

the intermediate distributions p_t. NPSE has several advantages compared to NPE, including the ability to combine multiple observations at inference time [69] and, importantly, the freedom to use unconstrained network architectures.

We here propose to use flow matching, another recent technique for generative modeling, for Bayesian inference, an approach we refer to as flow-matching posterior estimation (FMPE). Flow matching is also based on a vector field v_t and thereby also admits flexible network architectures (Fig. 3.1). For flow matching, however, v_t directly defines the velocity field of sample trajectories, which solve ordinary differential equations (ODEs) and are deterministic. As a consequence, flow matching allows for additional freedom in designing non-diffusion paths such as optimal transport, and provides direct access to the density [76]. These differences are summarized in Tab. 3.1.

[76]: Lipman et al. (2022), 'Flow Matching for Generative Modeling'

Our contributions are as follows:

- We adapt flow-matching to Bayesian inference, proposing FMPE. In general, the modeling requirements of SBI are different from generative modeling. For the latter, sample quality is critical, i.e., that samples lie in the support of a complex distribution (e.g., images). In contrast, for SBI, $p(\theta|x)$ is typically less complex for fixed x, but x itself can be complex and high-dimensional. We therefore consider pyramid-like

Table 3.1: Comparison of posterior-estimation methods.

	NPE	NPSE	**FMPE (Ours)**
Tractable posterior density	Yes	No	Yes
Unconstrained network architecture	No	Yes	Yes
Network passes for sampling	Single	Many	Many

architectures from x to v_t, with gated linear units to incorporate (θ, t) dependence, rather than the typical U-Net used for images (Fig. 3.1). We also propose an alternative t-weighting in the loss, which improves performance in many tasks.

- Under certain regularity assumptions, we prove an upper bound on the KL divergence between the model and target posterior. This implies that estimated posteriors are mass-covering, i.e., that their support includes all θ consistent with observed x, which is highly desirable for scientific applications [77].
- We perform a number of experiments to investigate the performance of FMPE. Our two-pronged approach, which involves a set of benchmark tests and a real-world problem, is designed to probe complementary aspects of the method, covering breadth and depth of applications. First, on an established suite of SBI benchmarks, we show that FMPE performs comparably—or better—than NPE across most tasks, and in particular exhibits mass-covering posteriors in all cases (Sec. 3.4). We then push the performance limits of FMPE on a challenging real-world problem by turning to gravitational-wave inference (Sec. 3.5). We show that FMPE substantially outperforms an NPE baseline in terms of training time, posterior accuracy, and scaling to larger networks.

[77]: Hermans et al. (2021), 'Averting a crisis in simulation-based inference'

3.2 Preliminaries

Normalizing flows. A normalizing flow [19, 21] defines a probability distribution $q(\theta|x)$ over parameters $\theta \in \mathbb{R}^n$ in terms of an invertible mapping $\psi_x : \mathbb{R}^n \to \mathbb{R}^n$ from a simple base distribution $q_0(\theta)$,

$$q(\theta|x) = (\psi_x)_* q_0(\theta) = q_0(\psi_x^{-1}(\theta)) \det \left| \frac{\partial \psi_x^{-1}(\theta)}{\partial \theta} \right| , \qquad (3.1)$$

where $(\cdot)_*$ denotes the pushforward operator, and for generality we have conditioned on additional context $x \in \mathbb{R}^m$. Unless otherwise specified, a normalizing flow refers to a *discrete* flow, where ψ_x is given by a composition of simpler mappings with triangular Jacobians, interspersed with shuffling of the θ. This construction results in expressive $q(\theta|x)$ and also efficient density evaluation [19].

[19]: Papamakarios et al. (2021), 'Normalizing Flows for Probabilistic Modeling and Inference'
[21]: Rezende et al. (2015), 'Variational Inference with Normalizing Flows'

Continuous normalizing flows. A continuous flow [22] also maps from base to target distribution, but is parametrized by a continuous "time" $t \in [0, 1]$, where $q_0(\theta|x) = q_0(\theta)$ and $q_1(\theta|x) = q(\theta|x)$. For each t, the flow is defined by a vector field $v_{t,x}$ on the sample space.* This corresponds to the velocity of the sample trajectories,

[22]: Chen et al. (2018), 'Neural Ordinary Differential Equations'

$$\frac{d}{dt}\psi_{t,x}(\theta) = v_{t,x}(\psi_{t,x}(\theta)), \qquad \psi_{0,x}(\theta) = \theta. \tag{3.2}$$

We obtain the trajectories $\theta_t \equiv \psi_{t,x}(\theta)$ by integrating this ODE. The final density is given by

$$q(\theta|x) = (\psi_{1,x})_* q_0(\theta) = q_0(\theta) \exp\left(-\int_0^1 \operatorname{div} v_{t,x}(\theta_t)\, \mathrm{d}t\right), \tag{3.3}$$

which is obtained by solving the transport equation $\partial_t q_t + \operatorname{div}(q_t v_{t,x}) = 0$.

The advantage of the continuous flow is that $v_{t,x}(\theta)$ can be simply specified by a neural network taking $\mathbb{R}^{n+m+1} \to \mathbb{R}^n$, in which case Equation (3.2) is referred to as a *neural ODE* [22]. Since the density is tractable via Equation (3.3), it is in principle possible to train the flow by maximizing the (log-)likelihood. However, this is often not feasible in practice, since both sampling and density estimation require many network passes to numerically solve the ODE Equation (3.2).

Flow matching. An alternative training objective for continuous normalizing flows is provided by flow matching [76]. This directly regresses $v_{t,x}$ on a vector field $u_{t,x}$ that generates a target probability path $p_{t,x}$. It has the advantage that training does not require integration of ODEs, however it is not immediately clear how to choose $(u_{t,x}, p_{t,x})$. The key insight of [76] is that, if the path is chosen on a *sample-conditional* basis,† then the training objective becomes extremely simple. Indeed, given a sample-conditional probability path $p_t(\theta|\theta_1)$ and a corresponding vector field $u_t(\theta|\theta_1)$, we specify the sample-conditional flow matching loss as

[76]: Lipman et al. (2022), 'Flow Matching for Generative Modeling'

[76]: Lipman et al. (2022), 'Flow Matching for Generative Modeling'

$$\mathcal{L}_{\text{SCFM}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],\, x\sim p(x),\, \theta_1\sim p(\theta|x),\, \theta_t\sim p_t(\theta_t|\theta_1)} \|v_{t,x}(\theta_t) - u_t(\theta_t|\theta_1)\|^2 . \tag{3.4}$$

Remarkably, minimization of this loss is equivalent to regressing $v_{t,x}(\theta)$ on the *marginal* vector field $u_{t,x}(\theta)$ that generates $p_t(\theta|x)$ [76]. Note that in this expression, the x-dependence of $v_{t,x}(\theta)$ is picked up via the expectation value, with the sample-conditional vector field independent of x.

There exists considerable freedom in choosing a sample-conditional

* In the SBI literature, this is also commonly referred to as "parameter space".

† We refer to conditioning on θ_1 as *sample*-conditioning to distinguish from conditioning on x.

path. Ref. [76] introduces the family of Gaussian paths

$$p_t(\theta|\theta_1) = \mathcal{N}(\theta|\mu_t(\theta_1), \sigma_t(\theta_1)^2 I_n), \tag{3.5}$$

where the time-dependent means $\mu_t(\theta_1)$ and standard deviations $\sigma_t(\theta_1)$ can be freely specified (subject to boundary conditions[‡]). For our experiments, we focus on the optimal transport paths defined by $\mu_t(\theta_1) = t\theta_1$ and $\sigma_t(\theta_1) = 1 - (1 - \sigma_{\min})t$ (also introduced in [76]). The sample-conditional vector field then has the simple form

$$u_t(\theta|\theta_1) = \frac{\theta_1 - (1 - \sigma_{\min})\theta}{1 - (1 - \sigma_{\min})t}. \tag{3.6}$$

Neural posterior estimation (NPE). NPE is an SBI method that directly fits a density estimator $q(\theta|x)$ (usually a normalizing flow) to the posterior $p(\theta|x)$ [20, 66, 67]. NPE trains with the maximum likelihood objective $\mathcal{L}_{\text{NPE}} = -\mathbb{E}_{p(\theta)p(x|\theta)} \log q(\theta|x)$, using Bayes' theorem to simplify the expectation value with $\mathbb{E}_{p(x)p(\theta|x)} \to \mathbb{E}_{p(\theta)p(x|\theta)}$. During training, $\mathcal{L}_{\text{NPE}}$ is estimated based on an empirical distribution consisting of samples $(\theta, x) \sim p(\theta)p(x|\theta)$. Once trained, NPE can perform inference for every new observation using $q(\theta|x)$, thereby *amortizing* the computational cost of simulation and training across all observations. NPE further provides exact density evaluations of $q(\theta|x)$. Both of these properties are crucial for the physics application in section 3.5, so we aim to retain these properties with FMPE.

Related work

Flow matching [76] has been developed as a technique for generative modeling, and similar techniques are discussed in [78–80] and extended in [81, 82]. Flow matching encompasses the deterministic ODE version of diffusion models [70–72] as a special instance. Although to our knowledge flow matching has not previously been applied to Bayesian inference, score-matching diffusion models have been proposed for SBI in [68, 69] with impressive results. These studies, however, use stochastic formulations via SDEs [75] or Langevin steps and are therefore not directly applicable when evaluations of the posterior density are desired (see Tab. 3.1). It should be noted that score modeling can also be used to parameterize continuous normalizing flows via an ODE. Extension of [68, 69] to the deterministic formulation could thereby be seen as a special case of flow matching. Many of our analyses and the practical guidance provided in Section 3.3 therefore also apply to score matching.

[‡] The sample-conditional probability path should be chosen to be concentrated around θ_1 at $t = 1$ (within a small region of size $\sigma_{\min}$) and to be the base distribution at $t = 0$.

[20]: Papamakarios et al. (2016), 'Fast ε-free Inference of Simulation Models with Bayesian Conditional Density Estimation'
[66]: Lueckmann et al. (2017), 'Flexible statistical inference for mechanistic models of neural dynamics'
[67]: Greenberg et al. (2019), 'Automatic posterior transformation for likelihood-free inference'
[76]: Lipman et al. (2022), 'Flow Matching for Generative Modeling'
[78]: Albergo et al. (2022), 'Building normalizing flows with stochastic interpolants'
[79]: Liu et al. (2022), 'Flow straight and fast: Learning to generate and transfer data with rectified flow'
[80]: Neklyudov et al. (2022), 'Action Matching: A Variational Method for Learning Stochastic Dynamics from Samples'
[81]: Davtyan et al. (2022), 'Randomized Conditional Flow Matching for Video Prediction'
[82]: Tong et al. (2023), 'Conditional flow matching: Simulation-free dynamic optimal transport'
[70]: Sohl-Dickstein et al. (2015), 'Deep unsupervised learning using nonequilibrium thermodynamics'
[71]: Song et al. (2019), 'Generative modeling by estimating gradients of the data distribution'
[72]: Ho et al. (2020), 'Denoising diffusion probabilistic models'
[68]: Sharrock et al. (2022), 'Sequential Neural Score Estimation: Likelihood-Free Inference with Conditional Score Based Diffusion Models'
[69]: Geffner et al. (2022), 'Score Modeling for Simulation-based Inference'
[75]: Song et al. (2020), 'Score-based generative modeling through stochastic differential equations'

We here focus on comparisons of FMPE against NPE [20, 66, 67], as it best matches the requirements of the application in section 3.5. Other SBI methods include approximate Bayesian computation [83–87], neural likelihood estimation [88–91] and neural ratio estimation [92–98]. Many of these approaches have sequential versions, where the estimator networks are specifically tuned to a specific observation x_o. FMPE has a tractable density, so it is straightforward to apply the sequential NPE [20, 66, 67] approaches to FMPE. In this case, inference is no longer amortized, so we leave this extension to future work.

[20]: Papamakarios et al. (2016), 'Fast ε-free Inference of Simulation Models with Bayesian Conditional Density Estimation'
[66]: Lueckmann et al. (2017), 'Flexible statistical inference for mechanistic models of neural dynamics'
[67]: Greenberg et al. (2019), 'Automatic posterior transformation for likelihood-free inference'
[83]: Sisson et al. (2018), 'Overview of ABC'
[84]: Beaumont et al. (2002), 'Approximate Bayesian computation in population genetics'
[85]: Beaumont et al. (2009), 'Adaptive approximate Bayesian computation'
[86]: Blum et al. (2010), 'Non-linear regression models for Approximate Bayesian Computation'
[87]: Prangle et al. (2013), *Semi-automatic selection of summary statistics for ABC model choice*

3.3 Flow matching posterior estimation

To apply flow matching to SBI we use Bayes' theorem to make the usual replacement $\mathbb{E}_{p(x)p(\theta|x)} \to \mathbb{E}_{p(\theta)p(x|\theta)}$ in the loss function Equation (3.4), eliminating the intractable expectation values. This gives the FMPE loss

$$\mathcal{L}_{\text{FMPE}} = \mathbb{E}_{t\sim p(t),\, \theta_1\sim p(\theta), x\sim p(x|\theta_1),\, \theta_t\sim p_t(\theta_t|\theta_1)} \left\| v_{t,x}(\theta_t) - u_t(\theta_t|\theta_1) \right\|^2 , \tag{3.7}$$

which we minimize using empirical risk minimization over samples $(\theta, x) \sim p(\theta)p(x|\theta)$. In other words, training data is generated by sampling θ from the prior, and then simulating data x corresponding to θ. This is in close analogy to NPE training, but replaces the log likelihood maximization with the sample-conditional flow matching objective. Note that in this expression we also sample $t \sim p(t)$, $t \in [0, 1]$ (see Sec. 3.3.3), which generalizes the uniform distribution in Equation (3.4). This provides additional freedom to improve learning in our experiments.

3.3.1 Probability mass coverage

As we show in our examples, trained FMPE models $q(\theta|x)$ can achieve excellent results in approximating the true posterior $p(\theta|x)$. However, it is not generally possible to achieve *exact* agreement due to limitations in training budget and network capacity. It is therefore important to understand how inaccuracies manifest. Whereas sample quality is the main criterion for generative modeling, for scientific applications one is often interested in the overall shape of the distribution. In particular, an important question is whether $q(\theta|x)$ is *mass-covering*, i.e., whether it contains the full support of $p(\theta|x)$. This minimizes the risk to falsely rule out possible explanations of the data. It also allows us to use importance sampling if the likelihood $p(x|\theta)$ of the forward model can be evaluated, which can be used for precise estimation of the posterior [99, 100].

[99]: Müller et al. (2019), 'Neural importance sampling'
[100]: Dax et al. (2023), 'Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference'

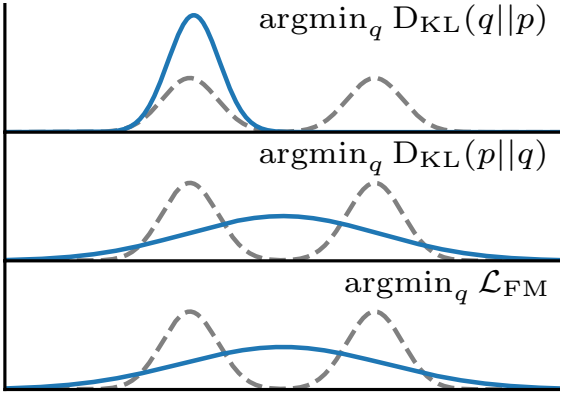

Figure 3.2: A Gaussian (blue) fitted to a bimodal distribution (gray) with various objectives.

Consider first the mass-covering property for NPE. NPE directly minimizes the forward KL divergence $D_{\mathrm{KL}}(p(\theta|x)||q(\theta|x))$, and thereby provides probability-mass covering results. Therefore, even if NPE is not accurately trained, the estimate $q(\theta|x)$ should cover the entire support of the posterior $p(\theta|x)$ and the failure to do so can be observed in the validation loss. As an illustration in an unconditional setting, we observe that a unimodal Gaussian q fitted to a bimodal target distribution p captures both modes when using the forward KL divergence $D_{\mathrm{KL}}(p||q)$, but only a single mode when using the backwards direction $D_{\mathrm{KL}}(q||p)$ (Fig. 3.2).

For FMPE, we can fit a Gaussian flow-matching model $q(\theta) = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ to the same bimodal target, in this case, parametrizing the vector field as

$$v_t(\theta) = \frac{(\sigma_t^2 + (t\hat{\sigma})^2 - \sigma_t)\theta_t + t\hat{\mu} \cdot \sigma_t}{t \cdot (\sigma_t^2 + (t\hat{\sigma})^2)} \tag{3.8}$$

(see Appendix D.1), we also obtain a mass-covering distribution when fitting the learnable parameters $(\hat{\mu}, \hat{\sigma})$ via Equation (3.4). This provides some indication that the flow matching objective induces mass-covering behavior, and leads us to investigate the more general question of whether the mean squared error between vector fields u_t and v_t bounds the forward KL divergence. Indeed, the former agrees up to constant with the sample-conditional loss Equation (3.4) (see Sec. 3.2).

We denote the flows of u_t, v_t, by ϕ_t, ψ_t, respectively, and we set $q_t = (\psi_t)_* q_0$, $p_t = (\phi_t)_* q_0$. The precise question then is whether we can bound $D_{\mathrm{KL}}(p_1||q_1)$ by $\mathrm{MSE}_p(u, v)^\alpha$ for some positive power α. It was already observed in [101] that this is not true in general, and we provide a simple example to that effect in Lemma D.2.1 in Appendix D.2. Indeed, it was found in [101] that to bound the forward KL divergence we also need to control the Fisher divergence, $\int p_t(\mathrm{d}\theta)(\nabla \ln p_t(\theta) - \nabla q_t(\theta))^2$.

[101]: Albergo et al. (2023), 'Stochastic interpolants: A unifying framework for flows and diffusions'

[101]: Albergo et al. (2023), 'Stochastic interpolants: A unifying framework for flows and diffusions'

Here we show instead that a bound can be obtained under sufficiently strong regularity assumptions on p_0, u_t, and v_t. The following statement is slightly informal, and we refer to the supplement for the complete version.

Theorem 3.3.1 *Let $p_0 = q_0$ and assume u_t and v_t are two vector fields whose flows satisfy $p_1 = (\phi_1)_* p_0$ and $q_1 = (\psi_1)_* q_0$. Assume that p_0 is square integrable and satisfies $|\nabla \ln p_0(\theta)| \leq c(1 + |\theta|)$ and u_t and v_t have bounded second derivatives. Then there is a constant $C > 0$ such that (for $\mathrm{MSE}_p(u, v) < 1$))*

$$D_{\mathrm{KL}}(p_1 || q_1) \leq C \, \mathrm{MSE}_p(u, v)^{\frac{1}{2}}. \tag{3.9}$$

The proof of this result can be found in appendix D.2. While the regularity assumptions are not guaranteed to hold in practice when v_t is parametrized by a neural net, the theorem nevertheless gives some indication that the flow-matching objective encourages mass coverage. In Section 3.4 and 3.5, this is complemented with extensive empirical evidence that flow matching indeed provides mass-covering estimates.

We remark that it was shown in [102] that the KL divergence of SDE solutions can be bounded by the MSE of the estimated score function. Thus, the smoothing effect of the noise ensures mass coverage, an aspect that was further studied using the Fokker-Planck equation in [101]. For flow matching, imposing the regularity assumption plays a similar role.

[102]: Song et al. (2021), 'Maximum likelihood training of score-based diffusion models'

[101]: Albergo et al. (2023), 'Stochastic interpolants: A unifying framework for flows and diffusions'

3.3.2 Network architecture

Generative diffusion or flow matching models typically operate on complicated and high dimensional data in the θ space (e.g., images with millions of pixels). One typically uses U-Net [103] like architectures, as they provide a natural mapping from θ to a vector field $v(\theta)$ of the same dimension. The dependence on t and an (optional) conditioning vector x is then added on top of this architecture.

[103]: Ronneberger et al. (2015), 'U-net: Convolutional networks for biomedical image segmentation'

For SBI, the data x is often associated with a complicated domain, such as image or time series data, whereas parameters θ are typically low dimensional. In this context, it is therefore useful to build the architecture starting as a mapping from x to $v(x)$ and then add conditioning on θ and t. In practice, one can therefore use any established feature extraction architecture for data in the domain of x, and adjust the dimension of the feature vector to $n = \dim(\theta)$. In our experiments, we found that the (t, θ)-conditioning is best achieved using gated linear units [104] to the hidden layers of the network (see also Fig. 3.1); these are also commonly used for conditioning discrete flows on x.

[104]: Dauphin et al. (2017), 'Language modeling with gated convolutional networks'

3.3.3 Re-scaling the time prior

The time prior $\mathcal{U}[0, 1]$ in Equation (3.4) distributes the training capacity uniformly across t. We observed that this is not always optimal in practice, as the complexity of the vector field may depend on t. For FMPE we therefore sample t in Equation (3.7) from a power-law distribution $p_\alpha(t) \propto t^{1/(1+\alpha)}$, $t \in [0, 1]$, introducing an additional hyperparameter α. This includes the uniform distribution for $\alpha = 0$, but for $\alpha > 0$, assigns greater importance to the vector field for larger values of t. We empirically found this to improve learning for distributions with sharp bounds (e.g., Two Moons in Section 3.4).

3.4 SBI benchmark

We now evaluate FMPE on ten tasks included in the benchmark presented in [105], ranging from simple Gaussian toy models to more challenging SBI problems from epidemiology and ecology, with varying dimensions for parameters ($\dim(\theta) \in [2, 10]$) and observations ($\dim(x) \in [2, 100]$). For each task, we train three separate FMPE models with simulation budgets $N \in \{10^3, 10^4, 10^5\}$. We use a simple network architecture consisting of fully connected residual blocks [106] to parameterize the conditional vector field. For the two tasks with $\dim(x) = 100$ (B-GLM-Raw, SLCP-D), we condition on (t, θ) via gated linear units as described in Section 3.3.2 (Fig. D.3 in Appendix D.3 shows the corresponding performance gain). For the remaining tasks with $\dim(x) \leq 10$ we concatenate (t, θ, x) instead. We reserve 5% of the simulations for validation. See Appendix D.3 for details.

[105]: Lueckmann et al. (2021), 'Benchmarking simulation-based inference'

[106]: He et al. (2015), *Deep Residual Learning for Image Recognition*

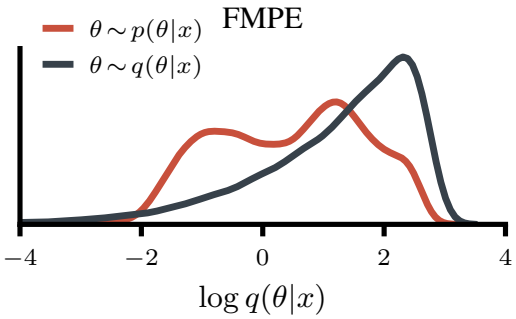

Figure 3.3: Histogram of FMPE densities $\log q(\theta|x)$ for reference samples $\theta \sim p(\theta|x)$ (Two Moons task, $N = 10^3$). The estimate $q(\theta|x)$ clearly covers $p(\theta|x)$ entirely.

For each task and simulation budget, we evaluate the model with the lowest validation loss by comparing $q(\theta|x)$ to the reference posteriors $p(\theta|x)$ provided in [105] for ten different observations x in terms of the C2ST score [107, 108]. This performance metric is

[105]: Lueckmann et al. (2021), 'Benchmarking simulation-based inference'

[107]: Friedman (2003), 'On multivariate goodness–of–fit and two–sample testing'

[108]: Lopez-Paz et al. (2016), 'Revisiting classifier two-sample tests'

computed by training a classifier to discriminate inferred samples $\theta \sim q(\theta|x)$ from reference samples $\theta \sim p(\theta|x)$. The C2ST score is then the test accuracy of this classifier, ranging from 0.5 (best) to 1.0. We observe that FMPE exhibits comparable performance to an NPE baseline model for most tasks and outperforms on several (Fig. 3.4). In terms of the MMD metric (Fig. D.1 in the Appendix), FMPE clearly outperforms NPE (but MMD can be sensitive to its hyperparameters [105]). As NPE is one of the highest ranking methods for many tasks in the benchmark, these results show that FMPE indeed performs competitively with other existing SBI methods. We report an additional baseline for score matching in Fig. D.2 in the Appendix.

As NPE and FMPE both directly target the posterior with a density estimator (in contrast to most other SBI methods), observed differences can be primarily attributed to their different approaches for density estimation. Interestingly, a great performance improvement of FMPE over NPE is observed for SLCP with a large simulation budget ($N = 10^5$). The SLCP task is specifically designed to have a simple likelihood but a complex posterior, and the FMPE performance underscores the enhanced flexibility of the FMPE density estimator.

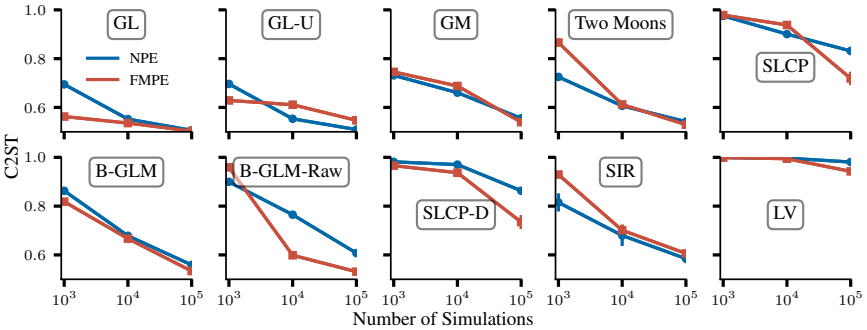

Figure 3.4: Comparison of FMPE with NPE, a standard SBI method, across 10 benchmark tasks [105].

Finally, we empirically investigate the mass coverage suggested by our theoretical analysis in Section 3.3.1. We display the density $\log q(\theta|x)$ of the reference samples $\theta \sim p(\theta|x)$ under our FMPE model q as a histogram (Fig. 3.3). All samples $\theta \sim p(\theta|x)$ fall into the support from $q(\theta|x)$. This becomes apparent when comparing to the density $\log q(\theta|x)$ for samples $\theta \sim q(\theta|x)$ from q itself. This FMPE result is therefore mass covering. Note that this does not necessarily imply conservative posteriors (which is also not generally true for the forward KL divergence [77, 109, 110]), and some parts of $p(\theta|x)$ may still be undersampled. Probability mass coverage, however, implies that no part is entirely missed (compare Fig. 3.2), even for multimodal distributions such as Two Moons. Fig. D.4 in the Appendix confirms the mass coverage for the other

[77]: Hermans et al. (2021), 'Averting a crisis in simulation-based inference'
[109]: Delaunoy et al. (2022), *Towards Reliable Simulation-Based Inference with Balanced Neural Ratio Estimation*
[110]: Delaunoy et al. (2023), 'Balancing Simulation-based Inference for Conservative Posteriors'

benchmark tasks.

3.5 Gravitational-wave inference

3.5.1 Background

Gravitational waves (GWs) are ripples of spacetime predicted by Einstein and produced by cataclysmic cosmic events such as the mergers of binary black holes (BBHs). GWs propagate across the universe to Earth, where the LIGO-Virgo-KAGRA observatories measure faint time-series signals embedded in noise. To-date, roughly 90 detections of merging black holes and neutron stars have been made [111], all of which have been characterized using Bayesian inference to compare against theoretical models.[§] These have yielded insights into the origin and evolution of black holes [112], fundamental properties of matter and gravity [113, 114], and even the expansion rate of the universe [115]. Under reasonable assumptions on detector noise, the GW likelihood is tractable,[¶] and inference is typically performed using tools [29, 31, 116, 117] based on Markov chain Monte Carlo [118, 119] or nested sampling [120] algorithms. This can take from hours to months, depending on the nature of the event and the complexity of the signal model, with a typical analysis requiring up to $\sim 10^8$ likelihood evaluations. The ever-increasing rate of detections means that these analysis times risk becoming a bottleneck. SBI offers a promising solution for this challenge that has thus been actively studied in the literature [3, 28, 100, 121–126]. A fully amortized NPE-based method called DINGO recently achieved accuracies comparable to stochastic samplers with inference times of less than a minute per event [28]. To achieve accurate results, however, DINGO uses group-equivariant NPE [28, 125] (GNPE), an NPE extension that integrates known conditional symmetries. GNPE, therefore, does not provide a tractable density, which is problematic when verifying and correcting inference results using importance sampling [100].

3.5.2 Experiments

We here apply FMPE to GW inference. As a baseline, we train an NPE network with the settings described in [28] with a few

[111]: Abbott et al. (2021), 'GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run'

[112]: Abbott et al. (2021), 'Population Properties of Compact Objects from the Second LIGO-Virgo Gravitational-Wave Transient Catalog'

[113]: Abbott et al. (2018), 'GW170817: Measurements of neutron star radii and equation of state'

[114]: Abbott et al. (2021), 'Tests of general relativity with binary black holes from the second LIGO-Virgo gravitational-wave transient catalog'

[115]: Abbott et al. (2017), 'A gravitational-wave standard siren measurement of the Hubble constant'

[29]: Veitch et al. (2015), 'Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library'

[31]: Ashton et al. (2019), 'BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy'

[116]: Romero-Shaw et al. (2020), 'Bayesian inference for compact binary coalescences with bilby: validation and application to the first LIGO–Virgo gravitational-wave transient catalogue'

[117]: Speagle (2020), 'dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences'

[118]: Metropolis et al. (1953), 'Equation of state calculations by fast computing machines'

[119]: Hastings (1970), 'Monte Carlo sampling methods using Markov chains and their applications'

[120]: Skilling (2006), 'Nested sampling for general Bayesian computation'

[§] BBH parameters $\theta \in \mathbb{R}^{15}$ include black-hole masses, spins, and the spacetime location and orientation of the system (see Tab. D.3 in the Appendix). We represent x in frequency domain; for two LIGO detectors and complex $f \in [20, 512]$ Hz, $\Delta f = 0.125$ Hz, we have $x \in \mathbb{R}^{15744}$.

[¶] Noise is assumed to be stationary and Gaussian, so for frequency-domain data, the GW likelihood $p(x|\theta) = \mathcal{N}(h(\theta)|S_n)(x)$. Here $h(\theta)$ is a theoretical signal model based on Einstein's theory of general relativity, and S_n is the power spectral density of the detector noise.

minor changes (see Appendix D.4).[‖] This uses an embedding network [127] to compress x to a 128-dimensional feature vector, which is then used to condition a neural spline flow [128]. The embedding network consists of a learnable linear layer initialized with principal components of GW simulations followed by a series of dense residual blocks [106]. This architecture is a powerful feature extractor for GW measurements [28]. As pointed out in Section 3.3.2, it is straightforward to reuse such architectures for FMPE, with the following three modifications: (1) we provide the conditioning on (t, θ) to the network via gated linear units in each hidden layer; (2) we change the dimension of the final feature vector to the dimension of θ so that the network parameterizes the conditional vector field $(t, x, \theta) \to v_{t,x}(\theta)$; (3) we increase the number and width of the hidden layers to use the capacity freed up by removing the discrete normalizing flow.

We train the NPE and FMPE networks with $5 \cdot 10^6$ simulations for 400 epochs using a batch size of 4096 on an A100 GPU. The FMPE network ($1.9 \cdot 10^8$ learnable parameters, training takes $\approx$ 2 days) is larger than the NPE network ($1.3 \cdot 10^8$ learnable parameters, training takes $\approx$ 3 days), but trains substantially faster. We evaluate both networks on GW150914 [1], the first detected GW. We generate a reference posterior using the method described in [100]. Fig. 3.5 compares the inferred posterior distributions qualitatively and quantitatively in terms of the Jensen-Shannon divergence (JSD) to the reference.

3.5.3 Discussion

Our results for GW150914 show that FMPE substantially outperforms NPE on this challenging problem. We believe that this is related to the network structure as follows. The NPE network allocates roughly two thirds of its parameters to the discrete normalizing flow and only one third to the embedding network (i.e., the feature extractor for x). Since FMPE parameterizes just a vector field (rather than a collection of splines in the normalizing flow) it can devote its network capacity to the interpretation of the high-dimensional $x \in \mathbb{R}^{15744}$. Hence, it scales better to larger networks and achieves higher accuracy. Remarkably, the performance iscomparable to GNPE, which involves a much simpler learning task with likelihood symmetries integrated by construction. This enhanced performance, comes in part at the cost of increased

[‖] Our implementation builds on the public DINGO code from https://github.com/dingo-gw/dingo.

[‖] We omit the three parameters $\phi_c, \phi_{JL}, \theta_{JN}$ in the evaluation as we use phase marginalization in importance sampling and the reference therefore uses a different basis for these parameters [100]. For GNPE we report the results from [28], which are generated with slightly different data conditioning. Therefore, we do not display the GNPE results in the corner plot, and the JSDs serve only as a rough comparison. The JSD for the t_c parameter is not reported in [28] due to a t_c marginalized reference.

[3]: Green et al. (2021), 'Complete parameter inference for GW150914 using deep learning'

[28]: Dax et al. (2021), 'Real-Time Gravitational Wave Science with Neural Posterior Estimation'

[100]: Dax et al. (2023), 'Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference'

[121]: Cuoco et al. (2020), 'Enhancing Gravitational-Wave Science with Machine Learning'

[122]: Gabbard et al. (2022), 'Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy'

[123]: Green et al. (2020), 'Gravitational-wave parameter estimation with autoregressive neural network flows'

[124]: Delaunoy et al. (2020), 'Lightning-Fast Gravitational Wave Parameter Inference through Neural Amortization'

[125]: Dax et al. (2022), 'Group equivariant neural posterior estimation'

[126]: Chatterjee et al. (2022), 'Rapid localization of gravitational wave sources from compact binary coalescences using deep learning'

[28]: Dax et al. (2021), 'Real-Time Gravitational Wave Science with Neural Posterior Estimation'

[28]: Dax et al. (2021), 'Real-Time Gravitational Wave Science with Neural Posterior Estimation'

[125]: Dax et al. (2022), 'Group equivariant neural posterior estimation'

[100]: Dax et al. (2023), 'Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference'

[127]: Radev et al. (2020), *BayesFlow: Learning complex stochastic models with invertible neural networks*

[128]: Durkan et al. (2019), 'Neural Spline Flows'

[106]: He et al. (2015), *Deep Residual Learning for Image Recognition*

[1]: Abbott et al. (2016), 'Observation of Gravitational Waves from a Binary Black Hole Merger'

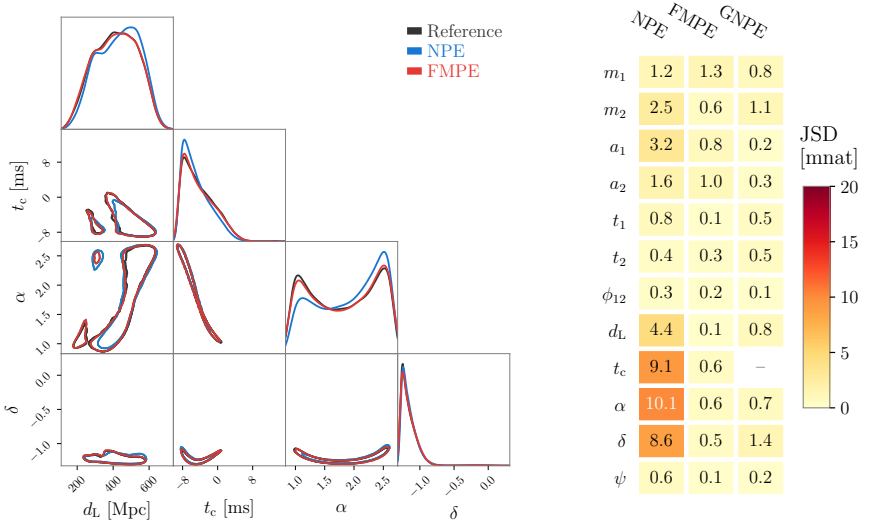

Figure 3.5: Results for GW150914 [1]. Left: Corner plot showing 1D marginals on the diagonal and 2D 50% credible regions. We display four GW parameters (distance d_L, time of arrival t_c, and sky coordinates α, δ); these represent the least accurate NPE parameters. Right: Deviation between inferred posteriors and the reference, quantified by the Jensen-Shannon divergence (JSD). The FMPE posterior matches the reference more accurately than NPE, and performs similarly to symmetry-enhanced GNPE. (We do not display GNPE results on the left due to different data conditioning settings in available networks.)

inference times, typically requiring hundreds of network forward passes. See Appendix D.4 for further details.

In future work we plan to carry out a more complete analysis of GW inference using FMPE. Indeed, GW150914 is a loud event with good data quality, where NPE already performs quite well. DINGO with GNPE has been validated in a variety of settings [28, 100, 125, 129] including events with a larger performance gap between NPE and GNPE [125]. Since FMPE (like NPE) does not integrate physical symmetries, it would likely need further enhancements to fully compete with GNPE. This may require a symmetry-aware architecture [130], or simply further scaling to larger networks. A straightforward application of the GNPE mechanism to FMPE—GFMPE—is also possible, but less practical due to the higher inference costs of FMPE. Nevertheless, our results demonstrate that FMPE is a promising direction for future research in this field.

3.6 Conclusions

We introduced flow matching posterior estimation, a new simulation-based inference technique based on continuous normalizing flows.

[28]: Dax et al. (2021), 'Real-Time Gravitational Wave Science with Neural Posterior Estimation'

[100]: Dax et al. (2023), 'Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference'

[125]: Dax et al. (2022), 'Group equivariant neural posterior estimation'

[129]: Wildberger et al. (2023), 'Adapting to noise distribution shifts in flow-based gravitational-wave inference'

[125]: Dax et al. (2022), 'Group equivariant neural posterior estimation'

[130]: Cohen et al. (2016), 'Group Equivariant Convolutional Networks'

In contrast to existing neural posterior estimation methods, it does not rely on restricted density estimation architectures such as discrete normalizing flows, and instead parametrizes a distribution in terms of a conditional vector field. Besides enabling flexible path specifications, while maintaining direct access to the posterior density, we empirically found that regressing on a vector field rather than an entire distribution improves the scalability of FMPE compared to existing approaches. Indeed, fewer parameters are needed to learn this vector field, allowing for larger networks, ultimately enabling to solve more complex problems. Furthermore, our architecture for FMPE (a straightforward ResNet with GLU conditioning) facilitates parallelization and allows for cheap forward/backward passes.

We evaluated FMPE on a set of 10 benchmark tasks and found competitive or better performance compared to other simulation-based inference methods. On the challenging task of gravitational-wave inference, FMPE substantially outperformed comparable discrete flows, producing samples on par with a method that explicitly leverages symmetries to simplify training. Additionally, flow matching latent spaces are more naturally structured than those of discrete flows, particularly when using paths such as optimal transport. Looking forward, it would be interesting to exploit such structure in designing learning algorithms. This performance and flexibilty underscores the capability of continuous normalizing flows to efficiently solve inverse problems.

Adapting to noise distribution shifts in flow-based gravitational-wave inference

4

Abstract

Deep learning techniques for gravitational-wave parameter estimation have emerged as a fast alternative to standard samplers—producing results of comparable accuracy. These approaches (e.g., DINGO) enable amortized inference by training a normalizing flow to represent the Bayesian posterior conditional on observed data. By conditioning also on the noise power spectral density (PSD) they can even account for changing detector characteristics. However, training such networks requires knowing in advance the distribution of PSDs expected to be observed, and therefore can only take place once all data to be analyzed have been gathered. Here, we develop a probabilistic model to forecast future PSDs, greatly increasing the temporal scope of trained deep-learning models. Using PSDs from the second LIGO-Virgo observing run (O2)—plus just a *single* PSD from the beginning of the third (O3)—we show that we can train a DINGO network to perform accurate inference *throughout* O3 (on 37 real events). We therefore expect this approach to be a key component to enable the use of deep learning techniques for low-latency analyses of gravitational waves.

About This Chapter

This chapter has been adapted from Wildberger et al. [40].

Adapting to noise distribution shifts in flow-based gravitational-wave inference

Jonas Wildberger, Maximilian Dax, Stephen R. Green, Jonathan Gair, Michael Pürrer, Jakob H. Macke, Alessandra Buonanno and Bernhard Schölkopf.

Physical Review D: 107, 084046 (2023).

4.1 Introduction

Detector noise plays a crucial role in interpreting observations of gravitational waves (GWs). In its simplest form, noise is assumed to be additive, and stationary and Gaussian. This means that it is characterized in frequency domain by its power spectral density (PSD) $S_\mathrm{n}(f)$. The GW likelihood for parameters θ is then the probability that after a proposed signal $h(\theta)$ is subtracted from data d, the residual is noise satisfying these assumptions, i.e.,

$$p(d|\theta, S_\mathrm{n}) \propto \exp\left(-\frac{1}{2}(d-h(\theta)|d-h(\theta))_{S_\mathrm{n}}\right), \tag{4.1}$$

where $(\cdot|\cdot)$ is the noise-weighted inner-product,

$$(a|b)_{S_\mathrm{n}} = 4\sum_I \Re \int_{f_\mathrm{min}}^{f_\mathrm{max}} \frac{\hat{a}_I^*(f)\hat{b}_I(f)}{S_{\mathrm{n},I}(f)} df. \tag{4.2}$$

Here, $\hat{a}$ denotes the Fourier transform of a, and the sum runs over interferometers I.

Although the LIGO [25], Virgo [26] and KAGRA [27, 131, 132] detectors are mostly stable during an observing run, the noise spectrum does vary over time and across detectors. Moreover, between observing runs, detectors are upgraded, resulting in reduced noise levels and increased sensitivity. The particular PSDs at the time of an event must therefore be estimated and taken into account when performing inference. Estimation of a PSD is typically carried out using either data adjacent to an event (off-source, Welch method [133]) or by jointly modeling the signal and noise from the on-source data (e.g., BayesWave [134, 135]). The PSD is then inserted into the likelihood and samples are drawn from the posterior $p(\theta|d, S_\mathrm{n})$ using Markov chain Monte Carlo [29] or nested sampling methods [31, 116, 117].

An emerging alternative to classical likelihood-based inference is simulation-based inference using probabilistic deep learning [3, 28, 122–125, 136–139]. Technologies such as normalizing flows enable neural networks to describe complex conditional probability distributions. A conditional density estimator $q(\theta|d)$ can be trained using simulated data to approximate $p(\theta|d)$ such that once a detection is made, samples can be drawn in seconds. To incorporate varying noise PSDs into these methods, the estimate of S_n is provided as additional context to the network, i.e. $q(\theta|d, S_\mathrm{n})$. During training, random PSDs are drawn from an empirical distribution $p(S_\mathrm{n})$ estimated from signal-free data during an observing run. Simulated data are then constructed based on these PSDs, and the PSD is provided as context. At inference time, the network is effectively "tuned" to the interferometers at the time of detection. This approach (called Dingo [28]) fully amortizes training costs across all detections within an observing run, and has been

[25]: Aasi et al. (2015), 'Advanced LIGO'
[26]: Acernese et al. (2015), 'Advanced Virgo: a second-generation interferometric gravitational wave detector'
[27]: Akutsu et al. (2021), 'Overview of KAGRA: Detector design and construction history'
[131]: Somiya (2012), 'Detector configuration of KAGRA: The Japanese cryogenic gravitational-wave detector'
[132]: Aso et al. (2013), 'Interferometer design of the KAGRA gravitational wave detector'
[133]: Welch (1967), 'The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms'
[134]: Cornish et al. (2015), 'BayesWave: Bayesian Inference for Gravitational Wave Bursts and Instrument Glitches'
[135]: Cornish et al. (2021), 'BayesWave analysis pipeline in the era of gravitational wave observations'
[29]: Veitch et al. (2015), 'Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library'
[31]: Ashton et al. (2019), 'BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy'
[116]: Romero-Shaw et al. (2020), 'Bayesian inference for compact binary coalescences with bilby: validation and application to the first LIGO–Virgo gravitational-wave transient catalogue'
[117]: Speagle (2020), 'dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences'

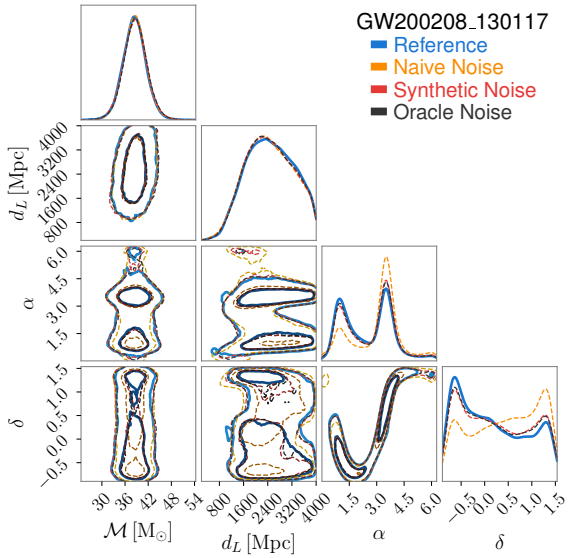

Figure 4.1: Posterior for GW200208_130117. Results from Dingo models trained only with empirically estimated detector-noise PSDs from the beginning of an observing run (orange) may deviate visibly from the reference (blue, obtained using importance sampling [100]). Dingo models trained with our proposed synthetic PSD model (red) achieve much more accurate results, on par with using PSDs estimated throughout the entire observing run (black). Our PSD model therefore enables Dingo to adapt to unseen drifts of the PSDs without retraining.

shown to produce results nearly indistinguishable from classical samplers.

The approach described requires access to the empirical distribution $p(S_n)$ of PSDs during an observing run. This makes it unsuitable for *online* parameter estimation (e.g., to provide alerts to electromagnetic telescopes) since the distribution $p(S_n)$ covering future observations is unavailable at the time of network training. Thus, a network trained with an empirical PSD distribution can only be used for a limited time—once the PSDs change too much, the measured data become out-of-distribution (OOD). Such a disagreement between training and inference distributions can lead to inaccurate results (Fig. 4.1). Here, we address this problem and provide a solution for training Dingo models robustly to better adapt to shifting distributions. Although our main motivation is the Dingo pipeline, the proposed method can in fact be used to improve the robustness of any machine learning method for gravitational wave search or inference that requires noise PSDs for training.

During training, Dingo models $q(\theta|d, S_n)$ estimate a distribution *conditional on* (as opposed to *marginalized over*) the PSD S_n. We are therefore not restricted to using (an approximation to) the real distribution $p(S_n)$ of PSDs; instead, we can use a synthetic distribution $q(S_n)$ whose support *contains* that of $p(S_n)$. In other words, if supp $p(S_n) \subseteq$ supp $q(S_n)$, then the Dingo model trained with $q(S_n)$ can be used for the entire observing run.

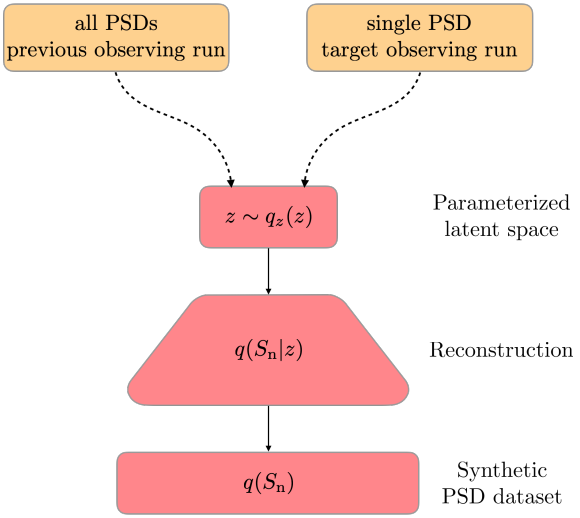

Figure 4.2: Overview of the PSD generation pipeline. The PSDs from the previous observing run and a single PSD from the target observing run are used for fitting the latent distribution $q_z(z)$, as indicated by the dashed arrows. We can then generate synthetic PSDs by first sampling latent variables in $z \sim q_z(z)$. These are then reconstructed back to frequency space via $q(S_n|z)$, obtaining a synthetic PSD distribution $q(S_n)$.

In this work, we develop a parameterized latent variable model for $q(S_n)$, which we fit to PSDs from a previous observing run. It is then straightforward to modify this distribution of PSDs via operations on the latent space. In particular, we use a one-shot observation from an upgraded detector to shift the latent space distribution coarsely to the expected noise level. Further, we broaden the spread of PSDs by blurring the latent space distribution. Our model $q(S_n)$ thereby represents a broad distribution over PSDs, which is capable of capturing variations throughout an observing run.

We evaluate our approach on the third LIGO-Virgo-KAGRA (LVK) observing run (O3) by analyzing 390 simulated and 37 real GW events [140]. We train Dingo with our PSD model and consistently find similar performance to Dingo trained with real O3 PSDs. Since our PSD model only uses O2 data and a single PSD from the beginning of O3, this demonstrates that our approach is indeed capable of preparing Dingo for unseen PSD changes.

4.2 Methods

In probabilistic modeling, a latent variable model [141–143] enables efficient sampling of new data given an empirical distribution. We define a latent variable model for the PSDs as

$$q(S_n) = \int q(S_n|z) q_z(z) dz \tag{4.3}$$

with latent variables z that we describe below. We use a fixed parameterization for $q(S_n|z)$ to integrate knowledge about the data

generating process into the model. PSD data can then be fitted via the distribution $q_z(z)$ over the latent variables. This provides an interpretable framework, enabling systematic interventions on the PSD distribution. Synthetic PSDs can then be generated by sampling $z \sim q_z(z)$ and reconstructing z back to frequency space via $q(S_\mathrm{n}|z)$. This pipeline is visualized in Fig. 4.2; next we explain the individual components in detail.

4.2.1 Parametrization of $q(S_\mathbf{n}|z)$

We aim to design a model $q(S_\mathrm{n})$ with sufficient expressiveness to capture every realistic PSD. On the other hand, the latent space should be low dimensional and maximally disentangled, such that the distribution $q_z(z)$ partially factorizes. In particular, the latent features z should model those PSD characteristics that are likely to change over time. This increases the data efficiency and robustness of our model under distribution shifts. These requirements inform our design of $q(S_\mathrm{n}|z)$. The following computations are carried out in log-space.

We leverage domain knowledge to parameterize the latent feature space explicitly. Each PSD is represented by a n-dimensional vector $S_\mathrm{n} \in \mathbb{R}^n$ for uniformly distributed frequency bins $f_1, \dots, f_n$ in the range $[f_\mathrm{min}, f_\mathrm{max}]$. According to [144], S_n can be decomposed into two main components, the smooth broad-band noise $b \in \mathbb{R}^n$ and a sum of high-power spectral lines $\sum_i s_i \in \mathbb{R}^n$. Improvements to the detectors are intended to reduce the impact of the broad-band noise, making it likely to change over time. The position and shape of the spectral lines may vary on much shorter time-scales. We thus model these components independently. We further assume independent additive Gaussian noise with constant variance σ^2 in each frequency bin.

With these assumptions, our model reads

$$q(S_\mathrm{n}|z) = \mathcal{N}\left(b + \sum_{i=1}^{l} s_i, \sigma^2 I_n\right) \tag{4.4}$$

where I_n denotes the $n \times n$ identity matrix. In latent space, the broad-band noise is represented by k values $y_1, \dots, y_k \in \mathbb{R}_+$ on fixed log-spaced frequency nodes $x_1, \dots, x_k$. The logarithmic distribution accounts for larger fluctuations of the broad-band noise in lower frequency regions. b is then obtained from its latent representation via cubic spline interpolation between the nodes.

Each spectral line s_i is represented by parameters f_{0i}, A_i, Q_i, denoting the center, height, and width of a truncated Cauchy distribution, respectively, i.e.

$$s_i[f_{0i}, A_i, Q_i](f) = \frac{w(f, f_{0i}) A_i f_{0i}^4}{(f_{0i} f)^2 + (Q_i(f_{0i}^2 - f^2))^2}, \tag{4.5}$$

[3]: Green et al. (2021), 'Complete parameter inference for GW150914 using deep learning'
[28]: Dax et al. (2021), 'Real-Time Gravitational Wave Science with Neural Posterior Estimation'
[122]: Gabbard et al. (2022), 'Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy'
[123]: Green et al. (2020), 'Gravitational-wave parameter estimation with autoregressive neural network flows'
[124]: Delaunoy et al. (2020), 'Lightning-Fast Gravitational Wave Parameter Inference through Neural Amortization'
[125]: Dax et al. (2022), 'Group equivariant neural posterior estimation'
[136]: Chua et al. (2020), 'Learning Bayesian posteriors with neural networks for gravitational-wave inference'
[137]: Chatterjee et al. (2019), 'Using Deep Learning to Localize Gravitational Wave Sources'
[138]: Krastev et al. (2021), 'Detection and Parameter Estimation of Gravitational Waves from Binary Neutron-Star Mergers in Real LIGO Data using Deep Learning'
[139]: Shen et al. (2021), 'Statistically-informed deep learning for gravitational wave parameter estimation'
[28]: Dax et al. (2021), 'Real-Time Gravitational Wave Science with Neural Posterior Estimation'
[140]: Abbott et al. (2021), 'Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo'
[141]: Bartholomew et al. (2011), *Latent variable models and factor analysis: A unified approach*
[142]: Kingma et al. (2014), 'Auto-Encoding Variational Bayes'
[143]: Rezende et al. (2014), 'Stochastic Backpropagation and Approximate Inference in Deep Generative Models'
[144]: Littenberg et al. (2015), 'Bayesian inference for spectral estimation of gravitational wave detector noise'

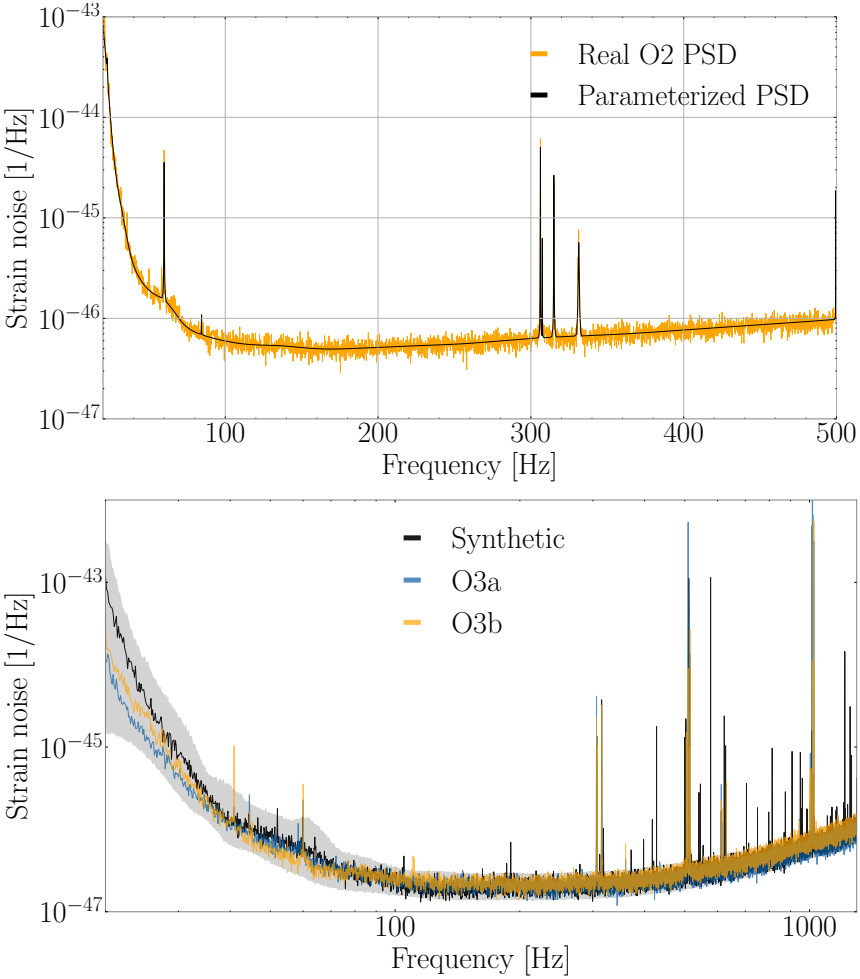

Figure 4.3: Upper panel: Comparison of a real O2 PSD and its representation under the latent variable model. Since $l = 400$ is sufficiently large, we can accurately model spectral lines that are close together (at around 300Hz). Lower panel: Comparison of a synthetic PSD (dark gray) and two real PSDs from O3a (blue) and O3b (orange). The synthetic PSD was sampled from a model trained on O2 noise, rescaled to O3. In the background (light gray) we include the envelope (1st to 99th percentile) of the synthetic broad-band distribution. By shifting the broad-band noise level (Section 4.2.3), we match the overall scale of O3 PSDs. By broadening the distribution, we ensure that PSDs from both O3a and O3b are covered.

with

$$w(f, f_{0i}) = \begin{cases} 1, & \text{if } |f - f_{0i}| \leq \delta, \\ \exp(-\frac{|f-f_{0i}|}{\delta}), & \text{otherwise,} \end{cases} \tag{4.6}$$

similarly to [144]. To efficiently treat a varying number of spectral lines per PSD, we segment the frequency range into l equally-wide intervals and model a single spectral line within each. Absence of a spectral line is modeled with $A_i \approx 0$. We do not find that restricting to one line per interval hinders Dingo performance in practice, provided the number of intervals is sufficiently large; see Fig. 4.3 (upper panel). In our experiments, we used $k = 30$, and $l = 400$ over a frequency range of $[f_{\min}, f_{\max}] = [20, 2048]$ Hz. If desired, the model could be extended to use overlapping intervals. In this case, each spectral distribution would no longer be independent, but rather conditional on the parameters of the preceding frequency interval.

[144]: Littenberg et al. (2015), 'Bayesian inference for spectral estimation of gravitational wave detector noise'

We found that σ^2 does not vary significantly between different PSDs. We therefore use a fixed value for σ^2 (estimated from the variance of the observed broad-band noise), and do not consider it as part of our latent space. Summarizing, our mapping from the latent space to frequency space reads

$$z = \begin{pmatrix} y_1, \dots, y_k, \\ f_{01}, A_1, Q_1, \\ \vdots \\ f_{0l}, A_l, Q_l \end{pmatrix} \mapsto S_{\mathrm{n}}. \tag{4.7}$$

The dimensionality of our latent space is $d = k + 3l$.

4.2.2 Fitting the latent distribution $q_z(z)$

We fit $q_z(z)$ based on an empirical distribution of PSDs estimated from detector data. We estimate this collection of PSDs using Welch's method applied to signal-free data stretches during the previous observing run. We then project these onto the latent space using maximum likelihood estimation and obtain $q_z(z)$ as the product of Gaussian kernel density estimates (KDEs) of these projections. This gives a tractable approximation to the true latent distribution. Importantly, having an analytic KDE provides flexibility to widen $q_z(z)$ by increasing the KDE bandwidth.

We perform the maximum likelihood projection onto the latent space in two steps. We first estimate the spline parameters $y_1, \dots, y_k$ for the broad-band noise as the sample mean of the PSD in the vicinity of the corresponding x_i. Since the broad-band noise should not model the spectral lines, we apply a filter before this step.* Once the broad-band noise is fitted, we subtract it from the

* We compute a running median over the neighborhood sets of each x_i. Every data point that deviates by more than three standard deviations from the running median is marked as an outlier and ignored for the fit of the broad-band noise.

PSD to fit the spectral lines. Specifically, we obtain (f_{0i}, A_i, Q_i) by solving the least-squares problem

$$\begin{aligned}(f_{0i}, A_i, Q_i) &= \underset{(\tilde{f},\tilde{A},\tilde{Q})}{\operatorname{argmin}} \sigma^{-2} r r^\top \\ r &= r(f, A, Q) = S_\mathrm{n} - (b + s_i(f, A, Q)),\end{aligned} \tag{4.8}$$

over each interval i. Together, these two steps correspond to a maximum-likelihood estimate of z for Equation (4.4) with fixed σ^2. On eight CPU cores, it takes less than a minute to obtain the latent maximum likelihood estimate z_MLE for each PSD, and this procedure is straightforwardly parallelizable. Indeed, this fast calculation (compared to ≈ 1 hour for BayesLine [144]) is one reason for choosing this simpler approach when building our PSD model.

Having computed z_MLE for each PSD from the empirical distribution, we next use KDEs to obtain $q_z(z)$. We use a separate KDE for each "independent" noise source, and then combine them multiplicatively. For the broad-band noise, we partition the frequency range into three intervals according to the most dominant noise source: $\mathcal{F}_1 = [f_\mathrm{min}, 30\,\mathrm{Hz}]$ for seismic noise, $\mathcal{F}_2 = [30\,\mathrm{Hz}, 100\,\mathrm{Hz}]$ for thermal noise and $\mathcal{F}_3 = [100\,\mathrm{Hz}, f_\mathrm{max}]$ for shot noise [144]. Since these noise sources should be largely independent of each other, we use individual KDEs for each subset $\{y_i \mid x_i \in \mathcal{F}_j\}$. These subsets do not overlap, so the reconstructed spline naturally agrees on the boundary between intervals $\mathcal{F}_i$. Similarly, the spectral lines are modeled independently, so we use a separate KDE for each set of parameters (f_{0i}, A_i, Q_i). The latent distribution is then given as the product of the $3 + l$ individual KDEs.

The independence assumption made above, and the corresponding partial factorization of $q_z(z)$ into $3 + l$ independent distributions, serves two purposes. First, it provides broader coverage of the latent space compared to an unfactorized KDE. Second, it decorrelates factors of variation that are independent so that trained networks do not learn to expect spurious correlations that exist in the empirical PSDs. In particular, this allows the PSD model to capture a greater variety of noise curves. Both of these effects ensure that trained networks are generally more robust with respect to changes in the PSD so that they can adapt to unseen data.

4.2.3 Interventions in latent space

The latent space is much lower dimensional than the raw PSD space and has features that correspond to the main components comprising a PSD. By intervening in this space (i.e., changing the PSD distribution at the level of the latent space) we can therefore naturally adapt the PSD distribution to changing detector noise characteristics. We perform two such types of interventions: (1) we shift the broad-band noise to account for improved detector

sensitivity, and (2) we broaden the distribution to account for uncertainty in the estimated broad-band shift and other unmodeled detector changes.

For (1), we rescale the latent distribution over $y_1, \dots, y_k$ according to an estimated PSD from early in the target run. Given a single PSD S_n^{target} from the target run, we shift the distribution over y_i so that its mean corresponds to the target PSD, i.e.,

$$y_i \mapsto y_i - \mathbb{E}[y_i] + y_i^{\text{target}}. \tag{4.9}$$

Here, y_i^{target} refers to the latent features corresponding to S_n^{target}. The resulting distribution corresponds to variations in all latent variables inferred from past data, but with broad-band noise level shifted to that of the target observing run. Note that this works even when S_n^{target} is not close to the mean of the target run, as long as the difference is small compared to the variance of the estimated distribution. Such shifts are particularly useful to flexibly change the shape of the noise curves, which may be required after major detector upgrades.

The broadening (2) is controlled by the KDE bandwidth parameter, and it improves the robustness of Dingo networks trained with the synthetic noise distribution. A larger bandwidth parameter can compensate for more significant shifts in the PSD distribution, although it may require higher learning capacity.

In Fig. 4.3, we illustrate how these interventions on the latent space ensure that synthetically generated PSDs match the broad-band noise level of O3 PSDs (lower panel), despite the difference in scale between O2 and O3.

4.3 Results

We prepare Dingo networks with the settings from [28]. In particular, we use the waveform model IMRPhenomPv2 [145–147] with frequency-domain data in the range $[20, 1024]$ Hz, $\Delta f = 0.125$ Hz, and the same prior. We train three networks, each with a different noise PSD dataset: (1) the *Oracle* dataset consists of PSDs estimated from real O3a and O3b detector noise (∼5,000 PSDs per detector); (2) the *Synthetic* dataset is sampled with our proposed method, using only O2 data and a single PSD from the start of O3 (50,000 PSDs per detector); and (3) the *Naive* baseline dataset consists of real PSDs from the first four days of O3a (100 PSDs per detector). The *Oracle* dataset encompasses the real PSDs that Dingo encounters at inference time, so this should be an upper bound for the performance of our method. On the other hand, the *Naive* dataset is based only on data available at the beginning of O3, so to be useful our method should significantly outperform this baseline.

[28]: Dax et al. (2021), 'Real-Time Gravitational Wave Science with Neural Posterior Estimation'

[145]: Khan et al. (2016), 'Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era'

[146]: Hannam et al. (2014), 'Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms'

[147]: Bohé et al. (2017), 'Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors'

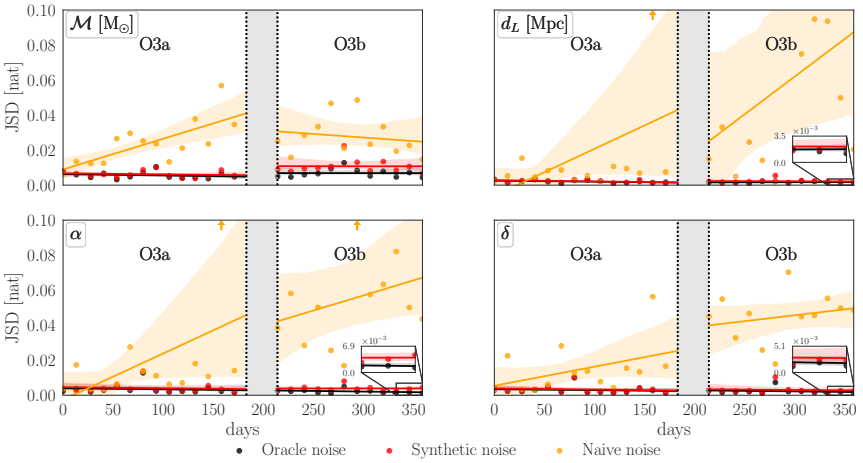

Figure 4.4: JSDs between Dingo and reference samples. Injections use noise PSDs estimated at various times throughout O3 (indicated on horizontal axis). Results are averaged over injections with five random sets of source parameters (fixed for all PSDs). Day 0 marks the beginning of O3a. The gap indicates the period when detectors were offline between O3a and O3b.

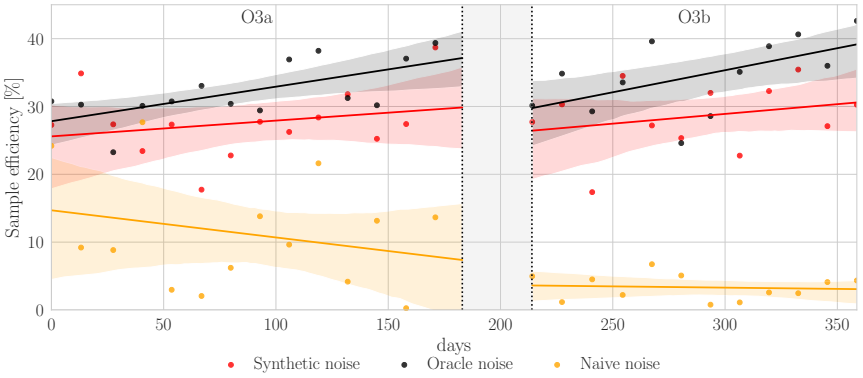

Figure 4.5: Importance sampling efficiency ϵ for Dingo models. Injections use noise PSDs estimated at various times throughout O3 (indicated on horizontal axis). Results are averaged over injections with five random sets of source parameters (fixed for all PSDs). Day 0 marks the beginning of O3a. The gap indicates the period when detectors were offline between O3a and O3b.

We evaluate the three DINGO networks on simulated and real strain data. To assess performance, we compare against reference posteriors. Since we analyze more than 400 events, generating these using stochastic samplers is not feasible. Instead, we generate reference posteriors by importance-weighting the DINGO results produced with the *Oracle* PSD dataset, which provides verified inference results at low computational cost (DINGO-IS) [100]. To save computational time, we further use phase marginalization when calculating the likelihoods [29, 148, 149]. We note that with precessing waveforms this is only an approximation, but for IMRPhenomPv2 the error introduced is small. See [100] for an alternative approach to generate the phase in the presence of higher modes. For a comparison of DINGO-IS against BILBY DYNESTY see Appendix B.

[100]: Dax et al. (2023), 'Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference'
[29]: Veitch et al. (2015), 'Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library'
[148]: Veitch et al. (2013), 'Analytic marginalisation of phase parameter'
[149]: Thrane et al. (2019), 'An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models'
[100]: Dax et al. (2023), 'Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference'

4.3.1 Simulated data

We first evaluate inference results when solely the PSD is varied in the strain data. To this end, we sample gravitational-wave parameters from the prior and inject the signal into identical noise realizations that are scaled with different PSDs. Specifically, we choose 26 real PSDs from O3, evenly spaced in time over the course of one year. The injection strains are then analyzed using each of the three DINGO networks.

Fig. 4.4 shows the mean Jensen-Shannon divergence (JSD) [150] between DINGO and DINGO-IS posteriors for chirp mass ($\mathcal{M}$), luminosity distance (d_L) and sky position (α, δ). These parameters have been found to be most significantly impacted by conditioning the network on an incorrect PSD at inference time [28]. (We find a similar trend for training with incomplete PSD information.) Results for all 15 parameters are provided in the appendix, showing a similar qualitative behavior to the parameters reported here.

At the beginning of the observing run, we observe similar accuracy for all three networks. This is not surprising, as the PSDs are still captured by the *Naive* dataset at this point. However, as PSDs drift away from their initial distribution, the performance of the *Naive* baseline decreases substantially, with JSDs increasing by one to two orders-of-magnitude. The effect is particularly striking at the transition from O3a to O3b. This demonstrates that the *Naive* baseline indeed fails to generalize well to unseen PSDs. The *Oracle* network on the other hand shows excellent performance throughout the entire duration of O3, with JSDs around $2 \cdot 10^{-3}$ nat.[†] This is also expected, since the *Oracle* network has access to PSDs from the entire duration of O3 during training.

Finally, with the *Synthetic* PSD dataset, DINGO accuracy approaches that of the *Oracle* network—even for PSDs recorded almost a year after the PSD used to recalibrate the *Synthetic* dataset (zoomed-in

[†] For comparison, a maximum JSD of $2 \cdot 10^{-3}$ nat has been established as a bound for indistinguishability in [31].

parts in Fig. 4.4). It far outperforms the *Naive* baseline, demonstrating that with our proposed method, DINGO can indeed be trained to generalize to unseen PSDs, with at most a small decrease in accuracy.

In the same experiment, we study the sample efficiency ϵ when using the inferred DINGO posterior as a proposal distribution for importance sampling [100] (Fig. 4.5). Small values of ϵ have been found to flag failure cases of the inference network, e.g. caused by out-of-distribution data, establishing it as another quality measure of the DINGO posteriors. For the *Naive* baseline, ϵ decreases with time and plummets after the transition from O3a to O3b. In contrast, the *Oracle* and *Synthetic* networks provide a consistently high sample efficiency of $\epsilon \approx 30\%$. As a performance metric, ϵ is sensitive to deviations in full parameter space, including high-dimensional correlations. Fig. 4.5 thus confirms the trend observed for the JSDs, and demonstrates that the *Synthetic* PSD dataset can be used to ensure a high efficiency of DINGO-IS throughout an observing run.

4.3.2 Real data

[140]: Abbott et al. (2021), 'Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo'

We now perform a large study on real data [140], analyzing all 37 BBH events from GWTC-2 and GWTC-3 that are consistent with the training prior. Here, we use four different DINGO models with distance priors $[0.1, 2]$ Gpc, $[0.1, 4]$ Gpc, $[0.1, 6]$ Gpc and $[0.1, 12]$ Gpc. The mean JSDs over all 15 parameters are reported in Table 4.1. Generally, the scores fluctuate more compared to the highly-controlled simulated data, in that the noise realizations vary and can be non-stationary or non-Gaussian, and real signals do not precisely match models. GW190527_092055, for example, is more difficult for the *Oracle* network even though the PSD is covered by the training dataset. We see that the trend observed for simulated data translates to real events and all 15 parameters. The decreasing accuracy for the *Naive* baseline becomes particularly apparent after the transition from O3a to O3b. With our proposed *Synthetic* noise model, however, the accuracy is maintained throughout the entire observing run, and on par with the *Oracle* performance. This showcases, once again, that our approach is indeed capable of forecasting shifts in the PSD distribution and to enable robust inference.

GW190517_055101 has a high mean JSD for all three datasets, suggesting that the poor performance is not due to an out-of-distribution PSD. Across all events (except GW190517_055101) and parameters, we obtain average JSDs of $1.2 \cdot 10^{-3}$ nat for the *Oracle* noise dataset, $1.4 \cdot 10^{-3}$ nat with the *Synthetic* noise dataset and $2.7 \cdot 10^{-3}$ nat with the *Naive* noise dataset. We thus conclude that our approach serves as a convincing replacement to the empirical PSD distribution when full PSD information is unavailable.

Event	Noise datasets Oracle	Synthetic	Naive
GW190408_181802	2.1	4.0	3.1
GW190413_052954	0.2	0.5	0.4
GW190413_134308	0.9	0.6	2.0
GW190421_213856	0.3	0.4	0.5
GW190503_185404	1.1	2.7	1.6
GW190513_205428	2.1	3.7	2.9
GW190514_065416	0.5	0.5	0.7
GW190517_055101	21.3	25.4	14.2
GW190519_153544	1.2	1.4	1.2
GW190521_074359	1.4	2.4	2.7
GW190527_092055	5.5	2.1	1.4
GW190602_175927	1.9	1.9	3.3
GW190701_203306	1.2	1.4	1.4
GW190719_215514	0.7	0.3	0.4
GW190727_060333	0.3	0.6	0.6
GW190731_140936	0.3	0.4	1.1
GW190803_022701	0.6	0.3	1.4
GW190805_211137	0.3	0.4	1.1
GW190828_063405	0.7	2.1	2.6
	Noise datasets		
Event	Oracle	Synthetic	Naive
GW190909_114149	0.4	0.5	0.9
GW190915_235702	0.8	0.7	1.4
GW190926_050336	1.7	1.8	1.8
O3b ↓			
GW191127_050227	3.5	1.7	3.2
GW191204_110529	3.1	4.4	6.1
GW191215_223052	0.8	1.1	2.8
GW191222_033537	0.4	0.5	1.4
GW191230_180458	0.3	0.4	0.7
GW200128_022011	0.6	1.0	2.9
GW200129_065458	2.6	1.6	5.3
GW200208_130117	0.5	0.6	5.5
GW200208_222617	2.3	2.3	3.1
GW200209_085452	1.3	1.8	2.2
GW200216_220804	0.4	0.9	3.5
GW200219_094415	0.8	1.0	4.0
GW200220_124850	0.2	0.3	0.7
GW200224_222234	0.7	1.4	14.5
GW200311_115853	1.6	2.6	10.2

Table 4.1: 37 binary black hole events from GWTC-2 and GWTC-3 analyzed with DINGO trained with three different PSD datasets. We report the JSD (in units of 10^{-3} nat) between DINGO and reference posteriors, averaged over all inferred source parameters.

Table 4.2: Importance sampling efficiency ϵ for DINGO networks trained with three different PSD datasets, based on 100,000 samples per analysis. Since the variance of ϵ between identical importance sampling runs can be high, we here report median scores over 10 runs for each event.

	Noise datasets		
Event	Oracle	Synthetic	Naive
GW190408_181802	28.5	19.4	17.3
GW190413_052954	44.6	45.0	41.9
GW190413_134308	41.0	49.3	43.6
GW190421_213856	51.1	46.2	44.7
GW190503_185404	46.9	40.4	35.4
GW190513_205428	28.4	28.0	16.0
GW190514_065416	34.5	37.5	37.3
GW190517_055101	20.5	13.6	23.0
GW190519_153544	41.0	41.7	25.3
GW190521_074359	30.9	27.8	18.9
GW190527_092055	25.8	33.6	28.6
GW190602_175927	41.7	46.0	38.8
GW190701_203306	27.5	33.5	26.4
GW190719_215514	24.3	22.9	23.4
GW190727_060333	42.7	48.7	31.1
GW190731_140936	46.5	47.7	33.9
GW190803_022701	51.8	45.5	48.0
GW190805_211137	50.1	51.8	39.2
GW190828_063405	50.8	35.4	34.2
	Noise datasets		
Event	Oracle	Synthetic	Naive
GW190909_114149	32.9	17.6	22.5
GW190915_235702	49.3	48.7	45.1
GW190926_050336	30.8	22.1	27.7
O3b ↓			
GW191127_050227	2.8	3.9	2.5
GW191204_110529	19.3	20.1	17.1
GW191215_223052	45.0	44.1	28.9
GW191222_033537	46.6	47.6	39.7
GW191230_180458	41.3	40.3	40.5
GW200128_022011	33.9	29.8	29.8
GW200129_065458	27.4	30.9	8.2
GW200208_130117	48.7	47.8	41.7
GW200208_222617	9.5	15.5	13.9
GW200209_085452	16.7	13.1	16.6
GW200216_220804	33.7	32.5	30.6
GW200219_094415	25.4	22.6	23.3
GW200220_124850	49.9	47.7	46.6
GW200224_222234	49.5	36.0	9.5
GW200311_115853	43.9	41.8	16.3

We report the importance sampling efficiency for all analyzed events in Table 4.2. We see that high overall scores are achieved with all PSD datasets, although scores are usually lowest with the *Naive* noise dataset. Compared to the simulated events of Section 4.3.1, the downward trend in sampling efficiency for the *Naive* noise dataset is somewhat less clear. This is likely due to varying noise realizations and source parameters across real events, which introduce additional complications when comparing sampling efficiencies. For visual comparison, we include corner plots of selected events in the appendix.

4.3.3 Discussion

We compared the performance of DINGO inference networks trained with empirical and synthetic PSD distributions. In training, these distributions are represented as a finite set of PSD samples, and the empirical distributions consist of far fewer samples per detector (*Naive*: 100 PSDs, *Oracle*: ~ 5,000 PSDs) than the *Synthetic* distribution (50,000 PSDs). The unbalanced nature of these datasets may impact our results, since more training samples naturally lead to better generalization. However, the effectively unlimited number of synthetic samples available should be viewed as an advantage, and in practice, the number of available samples will always depend on the estimation method. Indeed, for the *Naive* dataset, which consists of PSDs from the first four days of O3, the data scarcity severely limits the number of available PSDs. Therefore, using unbalanced datasets correctly reflects the practical considerations. Generally, however, our results showing a time-dependent performance decay for the *Naive* dataset indicate that the distribution shift has a far more significant effect on performance than the size of the dataset.

4.4 Conclusions

We developed a probabilistic model for detector noise PSDs to improve training for flow-based GW inference and enable low-latency analyses. The PSD model can be fitted to empirical distributions and then rapidly sampled to obtain synthetic PSDs. By design, the PSD model operates on an interpretable latent space, such that samples can be modified in a physically meaningful way. This allows for data-efficient modelling of distribution shifts, as may occur between LVK observing runs.

In our experiments, we fitted a PSD distribution based on data from the second observing run (O2), and predicted the updated distribution for O3 based on only a single PSD from the beginning of O3. We demonstrated on simulated and real GW events that DINGO models trained with these synthetic O3 PSDs perform accurate inference throughout the entire run (~ 1 year). We found

comparable performance to DINGO models that have direct access to the entire O3 PSD distribution. This is in stark contrast to DINGO models trained with PSDs from the first few days of O3, for which the accuracy strongly degrades over time.

Our approach crucially hinges on the *conditioning* of DINGO models on the PSD. Indeed, we do not need to predict specific distribution shifts; instead, it is merely necessary that PSDs encountered at inference time lie within the *support* of the training distribution. Anticipated distribution shifts are addressed in two ways: first, the broad-band noise is shifted to better match the target; second, the PSD distribution is artificially broadened to account for unknown (future) variations. Our synthetic distribution thereby captures unseen distribution shifts, at the cost of being somewhat broader than necessary. While this may generally require increased learning capacity, we have not observed this as a limiting factor in our experiments.

These systematic interventions on the PSD distribution are enabled by the design of our PSD model. We disentangle independent factors of variations by separating spectral lines from the broad-band noise, both of which are represented in a latent space.

Our framework further provides a tractable evidence of PSDs under the estimated distribution, which can be used to quantify distribution shifts. Going forward, this metric could be used to verify that our PSD model continues to work well between future observing runs, e.g. O3 and O4. A low evidence would then flag a distribution shift and indicate that re-training the network may be required. Moreover, for individual events, results can be validated using importance sampling [100].

[100]: Dax et al. (2023), 'Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference'

Techniques for parametrizing and estimating PSDs, such as BAYESWAVE [134] and BAYESLINE [144] have been successfully applied to spectral estimation and accurate posterior inference, and our method indeed draws inspiration from these. However, our goal is to model a *distribution* of PSDs, whereas these methods are designed to estimate a *single* PSD. Our work in fact complements these methods: by setting $\sigma^2 = 0$, we can model smooth broad-band noise, such that our synthetic PSDs closely resemble those of BAYESWAVE (as opposed to noisier Welch PSDs). By training DINGO networks using $\sigma^2 = 0$ synthetic PSDs, we therefore hope to enable inference with BAYESWAVE PSDs as context.

[134]: Cornish et al. (2015), 'BayesWave: Bayesian Inference for Gravitational Wave Bursts and Instrument Glitches'

[144]: Littenberg et al. (2015), 'Bayesian inference for spectral estimation of gravitational wave detector noise'

Our PSD model represents a key component in training flow-based inference networks for low-latency analysis of GW data—enabling complete parameter estimation in real time, with no retraining even for unseen PSD distribution shifts (as expected during an observing run). Once DINGO is extended to binary neutron stars, this will play a crucial role in delivering rapid and accurate multimessenger alerts.

Conclusion 5

5.1 Recap

This dissertation began with a deceptively simple premise: that extracting reliable knowledge from complex, real-world data requires more than applying standard machine learning techniques. Through this research, a deeper understanding of empirical inference emerged, revealing it as fundamentally about navigating the inherent tension between theoretical rigor (e.g., ensuring statistical consistency in divergence metrics) and practical necessity (e.g., adapting to noisy signals in gravitational wave detection). This tension, while manifesting differently across domains, reveals consistent underlying principles and illuminates unexpected connections that span causal inference, simulation-based methods, and real-world signal analysis.

Most notably, the concept of *structural understanding* proved central to all three contributions. In developing the IKL divergence, the key insight was that meaningful comparisons between causal models require metrics sensitive to both graph structure and parametric differences. Similarly, FMPE's success stemmed from architectural choices that better aligned network capacity with the structural demands of high-dimensional conditioning data. The gravitational wave noise adaptation framework succeeded precisely because it incorporated an implicit causal model of noise generation, disentangling broadband components from spectral lines in ways that reflected the physical processes producing detector noise.

This structural perspective reveals a recurring theme: *successful empirical inference often requires understanding not just what to measure, but how systems are organized*. Traditional approaches focus on distributional properties—means, variances, correlations—but the challenges addressed in this dissertation demanded methods that respect underlying causal, architectural, or physical structures.

Another unexpected discovery was the fundamental importance of *deliberately designing distributional coverage* for robust inference. While the need for in-distribution data is well-established, the key insight was being conscious and proactive about what constitutes "in-distribution" rather than hoping training data naturally covers relevant variations. Whether ensuring that the IKL divergence captures differences across intervention environments, designing FMPE to maintain mass-covering posteriors, or expanding the PSD support in the noise adaptation framework, the recurring strategy was deliberately broadening coverage over anticipated variations rather than pursuing point-wise accuracy. This insight challenges

conventional approaches that focus on local optimization, suggesting that consciously designed distributional breadth often matters more than precision for real-world empirical inference.

5.2 Key Contributions

Interventional Kullback-Leibler Divergence for Causal Model Comparison Chapter 2 introduced the Interventional Kullback-Leibler (IKL) divergence, a theoretical framework for quantifying differences between causal models across multiple interventional environments. Unlike traditional metrics that focus on observational distributions, IKL captures both structural differences in causal graphs and distributional differences in mechanisms. The framework established key theoretical properties, including operational significance through Theorem 2.4.1, which guarantees that models with small IKL divergence produce similar statistical outcomes under intervention across most environments. Additionally, IKL's monotonic relationship with structural correctness enables its use for iterative causal discovery, providing a principled method to identify mis-oriented edges and guide causal structure learning from multi-environment data.

Flow Matching Posterior Estimation for Scalable Inference Chapter 3 developed Flow Matching Posterior Estimation (FMPE), a simulation-based inference method that leverages continuous normalizing flows to improve scalability for high-dimensional scientific problems. Rather than parameterizing the posterior distribution directly through discrete normalizing flows, FMPE directly learns a conditional vector field, allowing more network capacity to focus on interpreting complex conditioning data. The method demonstrated substantial improvements over traditional neural posterior estimation on gravitational wave inference tasks, achieving performance comparable to specialized techniques that incorporate physical symmetries. FMPE also ensures mass-covering posteriors through a KL bound and enables flexible path specifications while maintaining direct access to posterior densities.

Probabilistic Noise Modeling for Robust Gravitational Wave Inference Chapter 4 addressed the challenge of non-stationary detector noise in gravitational wave parameter estimation by developing a probabilistic model for Power Spectral Densities (PSDs). The framework incorporates an interpretable latent space that separates broadband noise components from spectral lines, enabling data-efficient adaptation to anticipated distribution shifts between observing runs. Rather than predicting future noise precisely, the approach maintains distributional support over anticipated variations by conditioning inference networks on synthetically broadened PSD distributions. Experiments demonstrated that

DINGO models trained with this approach using only historical data and a single reference PSD sustained high accuracy throughout the O3 observing run. Specifically, they performed comparably to models with full future knowledge of the noise PSD drifts, while enabling real-time parameter estimation without retraining.

5.3 Broader Context and Future Directions

This dissertation revealed several fundamental insights about empirical inference that extend well beyond the specific contributions. In particular, it demonstrated that effective approaches to complex inference problems often challenge conventional machine learning wisdom, suggesting new design principles that could reshape how we approach scientific inference more broadly.

5.3.1 The Primacy of Structural Understanding

A consistent theme across all three contributions was how *structural understanding* proved central to success—though the relevant structure differed in each case: causal graphs for IKL, neural architecture allocation for FMPE, and physical noise generation mechanisms for gravitational wave adaptation.

The IKL divergence revealed that meaningful causal model comparison requires metrics sensitive to structural misalignments rather than just distributional differences. Traditional approaches focus on summary statistics, but structural errors—mis-oriented edges—create cascading effects that manifest differently across interventional environments.

FMPE's architectural design similarly stemmed from structural considerations. Traditional NPE methods allocate substantial network capacity to the density architecture, leaving less for interpreting observations—a potentially suboptimal allocation for problems with high-dimensional conditioning data. By inverting this priority—parameterizing just a vector field while focusing network capacity on high-dimensional conditioning—FMPE achieved results comparable to methods with sophisticated physical symmetries.

The gravitational wave noise adaptation framework succeeded precisely because it incorporated an implicit causal model of noise generation, disentangling broadband components from spectral lines based on their physical origins. This structural decomposition enabled effective adaptation to unseen noise variations—something difficult to achieve without modeling the physical processes behind detector noise.

These discoveries point toward a broader principle: *empirical inference problems often require understanding system structure, not just*

system behavior. Future work should prioritize developing methods that explicitly incorporate structural knowledge, whether causal graphs, physical symmetries, or mechanistic decompositions.

5.3.2 Conscious Coverage Design Over Predictive Precision

The second major insight challenges machine learning's traditional emphasis on predictive accuracy. Across all three contributions, success came from *consciously designing for coverage*—though what needed to be covered differed: intervention environments for IKL, parameter space for FMPE, and input noise distributions for gravitational wave adaptation.

This principle proved most dramatic in the gravitational wave work, where training Dingo with synthetically broadened PSD distributions—prioritizing coverage of possible variations over point predictions of future noise—performed comparably to models with perfect future knowledge. The somewhat counterintuitive success of "preparing for many possibilities" over "predicting the right answer" suggests that robust inference often requires fundamentally different optimization objectives than traditional machine learning.

Notably, this coverage need not be exhaustive to be effective. The IKL framework explicitly guarantees similar outcomes across *most* environments rather than all—accepting that complete coverage is often neither achievable nor necessary. This suggests a pragmatic principle: designing for sufficient coverage over the most relevant variations often yields robust performance.

This insight reveals a guideline for scientific machine learning, where models must often perform reliably under conditions that differ from training. Rather than pursuing ever-more-precise predictions of known targets, it could be worthwhile to invest more time in understanding and modeling the space of possible variations—a principle that may also partly explain the success of large language models, where training on diverse, massive datasets effectively brings many downstream applications "in-distribution" [151–153].

[151]: Kaplan et al. (2020), 'Scaling Laws for Neural Language Models'
[152]: Hoffmann et al. (2022), 'Training Compute-Optimal Large Language Models'
[153]: Brown et al. (2020), 'Language Models are Few-Shot Learners'

5.3.3 Emerging Trade-offs in Scientific Inference

This work revealed several important trade-offs that deserve consideration in future method development.

FMPE highlighted a specific training-vs-inference efficiency trade-off: architectural innovations that dramatically improve training efficiency (30% reduction in training time) can come at the cost of increased inference time due to ODE solvers requiring hundreds of network forward passes. This trade-off is particularly relevant

for scientific applications where training may be done offline but inference must often be performed in real-time, as in gravitational wave parameter estimation for multimessenger alerts.

Separately, the noise adaptation work demonstrated a robustness-vs-complexity trade-off: broadening training distributions to improve robustness to distribution shifts may require additional learning capacity to handle the expanded input space effectively. However, experiments suggested this is often a worthwhile investment, as the benefits of sustained performance under varying conditions typically outweigh the computational costs.

Future work should investigate whether these can be circumvented through alternative designs. For FMPE specifically, can faster sampling techniques or specialized ODE solvers reduce inference costs while maintaining training benefits? Can symmetry-aware architectures be integrated with flow matching to achieve scalability without sacrificing speed?

5.3.4 Toward Unified Scientific Inference Frameworks

A promising future direction involves integrating these insights into unified frameworks for scientific inference. Rather than developing methods for individual problem types, the patterns revealed in this work suggest the possibility of more general approaches that leverage structural understanding, prioritize support coverage, and navigate architecture-efficiency-robustness trade-offs systematically.

Such frameworks might begin with explicit causal or physical models of the scientific system (following the noise adaptation approach), use these models to guide architecture design (following FMPE's capacity allocation insights), and employ metrics like IKL divergence to validate structural correctness across multiple environments. The goal would be creating scientific inference systems that are simultaneously theoretically principled, computationally efficient, and robustly applicable across varying experimental conditions.

This vision extends beyond the specific domains explored in this dissertation. Climate modeling, drug discovery, materials science, and other data-intensive scientific fields face similar challenges: complex high-dimensional observations, evolving experimental conditions, and the need for both accuracy and interpretability. The principles identified here—structural understanding, conscious coverage design, and principled trade-off navigation—could inform method development across these domains.

This vision also intersects with the rapid development of foundation models[154]. As large-scale models trained on scientific data become more capable, the principles identified here—structural

[154]: Bommasani et al. (2021), 'On the Opportunities and Risks of Foundation Models'

understanding and conscious coverage design—will likely be essential for building, training, and validating them effectively. The contribution of this work may therefore extend beyond the specific algorithms to the articulation of these underlying principles for robust scientific inference.

This underscores the dissertation's central argument: the future of scientific machine learning lies not in developing ever-more-sophisticated individual techniques, but in understanding the fundamental principles that make empirical inference robust and reliable in complex, dynamic settings. The ultimate goal is transforming machine learning from a tool that works well under ideal conditions into one that enables reliable scientific discovery in the messy, non-ideal settings where we are yet to discover answers to the most important questions.

Appendix

Related Projects A

This chapter presents additional research projects conducted during my doctoral studies that, while not forming the core contributions of this dissertation, demonstrate the broader applicability and impact of the methods and principles developed herein. These projects illustrate how advances in empirical inference extend across diverse scientific domains and reinforce the practical value of robust, scalable inference techniques.

A.1 Neural Importance Sampling for Gravitational-Wave Inference

Neural Importance Sampling for Gravitational-Wave Inference

Maximilian Dax, Stephen R. Green, Jonathan Gair, Michael Pürrer, **Jonas Wildberger**, Jakob H. Macke, Alessandra Buonanno and Bernhard Schölkopf.

Dax et al. [155]

Abstract

We combine amortized neural posterior estimation with importance sampling for fast and accurate gravitational-wave inference. We first generate a rapid proposal for the Bayesian posterior using neural networks, and then attach importance weights based on the underlying likelihood and prior. This provides (1) a corrected posterior free from network inaccuracies, (2) a performance diagnostic (the sample efficiency) for assessing the proposal and identifying failure cases, and (3) an unbiased estimate of the Bayesian evidence. By establishing this independent verification and correction mechanism we address some of the most frequent criticisms against deep learning for scientific inference. We carry out a large study analyzing 42 binary black hole mergers observed by LIGO and Virgo with the SEOBNRv4PHM and IMRPhenomXPHM waveform models. This shows a median sample efficiency of $\approx 10\%$ (two orders-of-magnitude better than standard samplers) as well as a ten-fold reduction in the statistical uncertainty in the log evidence. Given these advantages, we expect a significant impact on gravitational-wave inference, and for this approach to serve as a paradigm for harnessing deep learning methods in scientific applications.

Summary and connection to this dissertation

This research addresses the reliability and trustworthiness of machine learning methods in scientific discovery. By combining neural posterior estimation with importance sampling, it provides a means to verify the accuracy of posterior estimates, complementing the methods in Chapters 3 and 4. This approach tackles common criticisms of deep learning in science by demonstrating how rapid inference can be achieved without sacrificing mathematical rigor, reinforcing the practical applicability of methods like FMPE and Dingo through independent validation mechanisms.

A.2 Flow matching for atmospheric retrieval of exoplanets: Where reliability meets adaptive noise levels

Flow matching for atmospheric retrieval of exoplanets: Where reliability meets adaptive noise levels

Timothy D. Gebhard, **Jonas Wildberger**, Maximilian Dax, Annalena Kofler, Daniel Angerhausen, Sascha P. Quanz and Bernhard Schölkopf.

Gebhard, Timothy D. et al. [156]

Abstract

Context. Inferring atmospheric properties of exoplanets from observed spectra is key to understanding their formation, evolution, and habitability. Since traditional Bayesian approaches to atmospheric retrieval (e.g., nested sampling) are computationally expensive, a growing number of machine learning (ML) methods such as neural posterior estimation (NPE) have been proposed.
Aims. We seek to make ML-based atmospheric retrieval (1) more reliable and accurate with verified results, and (2) more flexible with respect to the underlying neural networks and the choice of the assumed noise models.
Methods. First, we adopted flow matching posterior estimation (FMPE) as a new ML approach to atmospheric retrieval. FMPE maintains many advantages of NPE, but provides greater architectural flexibility and scalability. Second, we used importance sampling (IS) to verify and correct ML results, and to compute an estimate of the Bayesian evidence. Third, we conditioned our ML models on the assumed noise level of a spectrum (i.e., error bars), and thus made them adaptable to different noise models.
Results. Both our noise-level-conditional FMPE and NPE models perform on a par with nested sampling across a range of noise levels when tested on simulated data. FMPE trains about three times faster than NPE and yields higher IS efficiencies. IS successfully corrects inaccurate ML results, identifies model failures via low efficiencies, and provides accurate estimates of the Bayesian evidence.
Conclusions. FMPE is a powerful alternative to NPE for fast, amortized, and parallelizable atmospheric retrieval. IS can verify results, helping to build confidence in ML-based approaches, while also facilitating model comparison via the evidence ratio. Noise level conditioning allows design studies for future instruments to be scaled up; for example, in terms of the range of signal-to-noise ratios.

Summary and connection to this dissertation

This research provides direct validation of FMPE's generalizability beyond gravitational wave analysis, successfully applying the method to exoplanet atmospheric retrieval. Key findings reinforce FMPE's training efficiency advantages (3x faster than NPE) and architectural flexibility demonstrated in Chapter 3. The conditioning on noise levels extends the adaptation principles from Chapter 4 to different data characteristics. These results underscore FMPE's potential as a general-purpose tool for scalable scientific inference across diverse domains.

A.3 From Structured Data to Clinical Notes: Robust Clinical Decision Support with Fine-Tuned LLMs

From Structured Data to Clinical Notes: Robust Clinical Decision Support with Fine-Tuned LLMs

Frederike Lübeck, **Jonas Wildberger**, Frederik Träuble, Maximilian Mordig, Sergios Gatidis,

Andreas Krause and Bernhard Schölkopf

Lübeck et al. [157]

Abstract

Clinical machine learning models are typically trained on highly structured and consistent datasets but deployed in real-world settings dominated by unstructured clinical text, creating a fundamental challenge for practical adoption. In this work, we investigate whether large language models (LLMs), fine-tuned on structured patient data, can generalize effectively to unstructured clinical notes at inference time. Using the UK Biobank dataset for cardiovascular disease (CVD) risk prediction, we demonstrate that LLMs trained on structured representations achieve performance comparable to specialized tabular machine learning models. More importantly, we show that these models maintain strong predictive accuracy when applied to unstructured inputs, such as clinical notes, in both zero-shot and few-shot scenarios.

Summary and connection to this dissertation

This work demonstrates how principles of robust empirical inference extend to healthcare settings, where real-world deployment often entails heterogeneous data sources, missingness, and population or hospital-specific distribution shifts. A central challenge here is the mismatch between training data that is structured and consistent and deployment inputs that are unstructured clinical text; handling such modality and schema shifts requires models that are both flexible and robust. The emphasis on maintaining performance under changing input representations closely parallels the dissertation's focus on reliable inference under non-ideal conditions, and the ability to adapt to new settings with limited additional supervision connects directly to the adaptation strategies discussed in Chapter 4.

A.4 Real-time inference for binary neutron star mergers using machine learning

Real-time inference for binary neutron star mergers using machine learning

Maximilian Dax, Stephen R. Green, Jonathan Gair, Nihar Gupte, Michael Pürrer, Vivien Raymond, **Jonas Wildberger**, Jakob H. Macke, Alessandra Buonanno and Bernhard Schölkopf.

Dax et al. [158]

Abstract

Mergers of binary neutron stars (BNSs) emit signals in both the gravitational-wave (GW) and electromagnetic (EM) spectra. Famously, the 2017 multi-messenger observation of GW170817 [159, 160] led to scientific discoveries across cosmology [115], nuclear physics [161–163], and gravity [164]. Central to these results were the sky localization and distance obtained from GW data, which, in the case of GW170817, helped to identify the associated EM transient, AT 2017gfo [165], 11 hours after the GW signal. Fast analysis of GW data is critical for directing time-sensitive EM observations; however, due to challenges arising from the length and complexity of signals, it is often necessary to make approximations that sacrifice accuracy. Here, we present a machine learning framework that

performs complete BNS inference in just one second without making any such approximations. Our approach enhances multi-messenger observations by providing (i) accurate localization even before the merger; (ii) improved localization precision by ∼ 30% compared to approximate low-latency methods; and (iii) detailed information on luminosity distance, inclination, and masses, which can be used to prioritize expensive telescope time. Additionally, the flexibility and reduced cost of our method open new opportunities for equation-of-state studies. Finally, we demonstrate that our method scales to extremely long signals, up to an hour in length, thus serving as a blueprint for data analysis for next-generation ground- and space-based detectors.

Summary and connection to this dissertation

This research demonstrates a crucial principle that connects to the dissertation's emphasis on leveraging structural understanding: the strategic use of physical domain knowledge to enhance inference efficiency. The key innovation—prior conditioning based on chirp mass estimates—allows a single neural network to be instantly tuned to event-specific priors, dramatically improving performance by incorporating system-specific information. This approach mirrors the structural insights from the dissertation: rather than treating all events identically, the method adapts to the specific characteristics of each binary system. The data compression technique, which factors out predominant phase evolution based on chirp mass estimates, similarly leverages physical understanding to reduce computational complexity while preserving essential information. This exemplifies how incorporating domain knowledge—whether causal structure (Chapter 2), architectural insights (Chapter 3), or noise characteristics (Chapter 4)—can transform seemingly intractable problems into manageable ones, enabling real-time scientific discovery.

A.5 Out-of-Variable Generalisation for Discriminative Models

Out-of-Variable Generalisation for Discriminative Models

Siyuan Guo, **Jonas Wildberger** and Bernhard Schölkopf.

Guo, Wildberger, and Schölkopf [166]

Abstract

The ability of an agent to do well in new environments is a critical aspect of intelligence. In machine learning, this ability is known as *strong* or *out-of-distribution* generalization. However, merely considering differences in distributions is inadequate for fully capturing differences between learning environments. In the present paper, we investigate *out-of-variable* generalization, which pertains to an agent's generalization capabilities concerning environments with variables that were never jointly observed before. This skill closely reflects the process of animate learning: we, too, explore Nature by probing, observing, and measuring proper *subsets* of variables at any given time. Mathematically, *oov* generalization requires the efficient re-use of past marginal information, i.e., information over subsets of previously observed variables. We study this problem, focusing on prediction tasks across environments that contain overlapping, yet distinct, sets of causes. We show that after fitting a classifier, the residual distribution in one environment reveals the partial derivative of the true generating function with respect to the unobserved causal parent in that environment. We leverage this information and propose a method that exhibits non-trivial out-of-variable generalization performance when facing an overlapping, yet distinct, set of causal predictors.

Summary and connection to this dissertation

This research extends the theoretical foundations of empirical inference by addressing a distinct generalization challenge: adapting when the set of observed variables changes between environments. While Chapter 2 focuses on comparing causal models under interventions, this work tackles generalization across different variable sets using causal assumptions to guide the reuse of marginal information. This complements the dissertation's emphasis on moving beyond standard i.i.d. assumptions, demonstrating how causal reasoning can enable adaptation in partially observed environments—a different but related aspect of robust empirical inference.

A.6 Evidence for eccentricity in the population of binary black holes observed by LIGO-Virgo-KAGRA

Evidence for eccentricity in the population of binary black holes observed by LIGO-Virgo-KAGRA

Nihar Gupte, Antoni Ramos-Buades, Alessandra Buonanno, Jonathan Gair, M. Coleman Miller, Maximilian Dax, Stephen R. Green, Michael Pürrer, **Jonas Wildberger**, Jakob Macke, Isobel M. Romero-Shaw, Bernhard Schölkopf.

Gupte et al. [167]

Abstract

Binary black holes (BBHs) in eccentric orbits produce distinct modulations in the emitted gravitational waves (GWs). The measurement of orbital eccentricity can provide robust evidence for dynamical binary formation channels. We analyze 57 GW events from the first, second and third observing runs of the LIGO-Virgo-KAGRA (LVK) Collaboration using a multipolar aligned-spin inspiral-merger-ringdown waveform model with two eccentric parameters: eccentricity and relativistic anomaly. This is made computationally feasible with the machine-learning code `DINGO`, which accelerates inference by 2-3 orders of magnitude compared to traditional inference techniques. First, when using a uniform prior on the eccentricity, we find eccentric aligned-spin against quasi-circular aligned-spin $\log_{10}$ Bayes factors of 1.84 to 4.75 (depending on the glitch mitigation) for GW200129, 3.0 for GW190701 and 1.77 for GW200208_22. We measure $e_{\mathrm{gw,\,10Hz}}$ to be $0.27^{+0.10}_{-0.12}$ to $0.17^{+0.14}_{-0.13}$ for GW200129, $0.35^{+0.32}_{-0.11}$ for GW190701 and $0.35^{+0.18}_{-0.21}$ for GW200208_22. Second, we find $\log_{10}$ Bayes factors between the eccentric aligned-spin versus quasi-circular precessing-spin hypothesis between 1.43 and 4.92 for GW200129, 2.61 for GW190701 and 1.23 for GW200208_22. Third, our analysis does not show evidence for eccentricity in GW190521, which has an eccentric aligned-spin against quasi-circular aligned-spin $\log_{10}$ Bayes factor of 0.04. Fourth, we estimate that if we neglect the spin-precession and use an astrophysically-motivated prior on the rate of eccentric BBHs, the probability of one out of the 57 events being eccentric is greater than 99.5% or $(100 - 8.4 \times 10^{-4})\%$ (depending on the glitch mitigation). Fifth, we study the impact on parameter estimation when neglecting either eccentricity in quasi-circular models or higher modes in eccentric models for GW events. These results underscore the importance of including eccentric parameters in the characterization of BBHs for the upcoming observing runs of the LVK Collaboration and for future detectors on the ground and in space, which will probe a more diverse BBH population.

Summary and connection to this dissertation

This study exemplifies how computational efficiency unlocks new scientific discoveries. Analyzing 57 gravitational-wave events for subtle orbital eccentricity effects would be computationally prohibitive with traditional methods but becomes feasible with the 2-3 orders of magnitude speedup provided by Dingo (incorporating methods related to Chapters 3 and 4). Finding statistically significant evidence for eccentricity across this population provides crucial insights into binary black hole formation channels, demonstrating that advances in scalable empirical inference directly enable new scientific questions to be addressed at scale.

A.7 Flexible Gravitational-Wave Parameter Estimation with Transformers

Flexible Gravitational-Wave Parameter Estimation with Transformers

Annalena Kofler, Maximilian Dax, Stephen R. Green, **Jonas Wildberger**, Nihar Gupte, Jakob H. Macke, Jonathan Gair, Alessandra Buonanno and Bernhard Schölkopf.

Kofler et al. [168]

Abstract

Gravitational-wave data analysis relies on accurate and efficient methods to extract physical information from noisy detector signals, yet the increasing rate and complexity of observations represent a growing challenge. Deep learning provides a powerful alternative to traditional inference, but existing neural models typically lack the flexibility to handle variations in data analysis settings. Such variations accommodate imperfect observations or are required for specialized tests, and could include changes in detector configurations, overall frequency ranges, or localized cuts. We introduce a flexible transformer-based architecture paired with a training strategy that enables adaptation to diverse analysis settings at inference time. Applied to parameter estimation, we demonstrate that a single flexible model—called Dingo-T1—can (i) analyze 48 gravitational-wave events from the third LIGO-Virgo-KAGRA Observing Run under a wide range of analysis configurations, (ii) enable systematic studies of how detector and frequency configurations impact inferred posteriors, and (iii) perform inspiral-merger-ringdown consistency tests probing general relativity. Dingo-T1 also improves median sample efficiency on real events from a baseline of 1.4% to 4.2%. Our approach thus demonstrates flexible and scalable inference with a principled framework for handling missing or incomplete data—key capabilities for current and next-generation observatories.

Summary and connection to this dissertation

This project extends the dissertation's central theme of reliable inference under non-ideal conditions to a setting where the analysis configuration itself becomes a source of distribution shift. The flexible architecture and training strategy enable a single model to adapt at inference time to changes in detector availability and frequency coverage, similar to the adaptation mechanisms developed in Chapter 4 for handling noise and data shifts.

Appendix to Chapter 2 B

B.1 Full proofs

B.1.1 Proof of Lemmas 2.3.1 and 2.3.4

Lemma 2.3.1 *Given causal model $\mathscr{P} = (P, G_P)$ assume that P^e emerges from P via an environment shift according to Assumption 2.2.3. If Q is Markovian with respect to G_P, we have the following decomposition*

$$
\begin{aligned}
&D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q(\mathbf{X})) = \\
&\quad \sum_{i \in \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{pa}_i^{G_P}) \| Q(X_i \mid \mathbf{pa}_i^{G_P}) \right) \right] \\
&\quad + \underbrace{\sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})} \left[D_{\mathrm{KL}} \left(P(X_i \mid \mathbf{pa}_i^{G_P}) \| Q(X_i \mid \mathbf{pa}_i^{G_P}) \right) \right]}_{=: D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}} \| \mathscr{Q})}
\end{aligned} \tag{2.5}
$$

As a special case we have for $\mathcal{I}_e = \varnothing$:

$$
\begin{aligned}
&D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) = \\
&\quad \sum_{i \in [d]} \mathbb{E}_{P(\mathbf{pa}_i^{G_P})} \left[D_{\mathrm{KL}} \left(P(X_i \mid \mathbf{pa}_i^{G_P}) \| Q(X_i \mid \mathbf{pa}_i^{G_P}) \right) \right]
\end{aligned} \tag{2.6}
$$

Lemma 2.3.4 *Given two CGMs $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_Q)$, assume that P^e emerges from P via an environment shift according to Assumption 2.2.3. We then have the following decomposition*

$$
\begin{aligned}
&D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q(\mathbf{X})) = \\
&\quad \sum_{i \in \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}) \right) \right] \\
&\quad + \sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}) \right) \right] \\
&\quad + D_{\mathrm{KL}} \left(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)(\mathbf{X}) \right)
\end{aligned} \tag{2.11}
$$

As a special case we have for $\mathcal{I}_e = \varnothing$

$$
\begin{aligned}
&D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) = \\
&\quad \sum_{i \in [d]} \mathbb{E}_{P(\mathbf{pa}_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}) \right) \right] \\
&\quad + D_{\mathrm{KL}} \left(P(\mathbf{X}) \| \pi_{G_Q}(P)(\mathbf{X}) \right)
\end{aligned} \tag{2.12}
$$

Proof. Both of these Lemmas are immediate consequences of Lemma B.1.1, noting that $\pi_{G_Q}(P)$ satisfies the condition of $\hat{P}$ by Lemma 2.3.3. The decomposition for $D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q(\mathbf{X}))$ then just follows by splitting up the sums.

□

B.1.2 Proof of Theorem 2.3.2

Lemma 2.3.2 *Under Assumptions 2.2.3, 2.2.4 and that no variable is always intervened upon, i.e.* $\bigcap_{e\in\mathcal{E}} \mathcal{I}_e = \varnothing$*, we can conclude that* $\mathscr{P}, \mathscr{Q}$ *are interventionally equivalent,* $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q}$*, if and only if* $\mathscr{P} = \mathscr{Q}$*.*

Proof. Since P, Q are both Markovian with respect to G_P, it follows by Lemma 2.3.1 that

$$P = Q \Leftrightarrow D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) = 0 \tag{B.1}$$

$$\overset{(2.6)}{\Leftrightarrow} \mathbb{E}_{P(\mathbf{p}a_i^{G_P})}\left[D_{\mathrm{KL}}(P(X_i \mid \mathbf{p}a_i^{G_P})\|Q(X_i \mid \mathbf{p}a_i^{G_P}))\right] = 0 \quad \text{for any } i \in [d] \tag{B.2}$$

$$\Leftrightarrow P(X_i \mid \mathbf{p}a_i^{G_P}) = Q(X_i \mid \mathbf{p}a_i^{G_P}) \quad \text{for any } i \in [d] \tag{B.3}$$

By our assumption on the multi-environment distributions, we further have that

$$P^e(X_i \mid \mathbf{P}A_i^{G_P}) = P(X_i \mid \mathbf{P}A_i^{G_P}) \quad \text{if and only if} \quad i \in [d]\backslash\mathcal{I}_e \tag{B.4}$$

Concluding, $P = Q$ implies that $P^e(X_i \mid \mathbf{P}A_i^{G_P}) = Q(X_i \mid \mathbf{P}A_i^{G_P})$ for all $i \in [d]\backslash\mathcal{I}_e$ leading to $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) = 0$. Conversely, assume that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) = 0$. Since $\bigcap_{e\in\mathcal{E}} \mathcal{I}_e = \varnothing$, it follows that $\bigcup_e [d]\backslash\mathcal{I}_e = [d]$, i.e. each term $\mathbb{E}_{P^e(\mathbf{p}a_i^{G_P})}\left[D_{\mathrm{KL}}(P(X_i \mid \mathbf{p}a_i^{G_P})\|Q(X_i \mid \mathbf{p}a_i^{G_P}))\right]$ is included in $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q})$ for some environment $e \in \mathcal{E}$. From $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) = 0$, we can thus deduce that each local divergence vanishes, and $P(X_i \mid \mathbf{p}a_i^{G_P}) = Q(X_i \mid \mathbf{p}a_i^{G_P})$ for any $i \in [d]$ implying that $P = Q$. □

B.1.3 Proof of Lemma 2.3.3

Lemma 2.3.3 *Given a distribution* P *and a DAG* G*:*

$$\pi_G(P)(\mathbf{X}) = \prod_{i=1}^{d} P(X_i \mid \mathbf{P}A_i^G). \tag{2.10}$$

Proof. To prove the identity, we make use of the following technical Lemma:

Lemma B.1.1 *For distributions* P, Q *such that* Q *is Markovian w.r.t.* G*, we have the following decomposition:*

$$D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) = \sum_{i=1}^{d} \mathbb{E}_{P(\mathbf{p}a_i^G)}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{p}a_i^G)\|Q(X_i \mid \mathbf{p}a_i^G)\right)\right] + D_{\mathrm{KL}}(P(\mathbf{X})\|\hat{P}(\mathbf{X})) \tag{B.5}$$

where $\hat{P}$ *is a distribution satisfying*

$$\hat{P}(\mathbf{X}) = \prod_{i=1}^{d} P(X_i \mid \mathbf{P}A_i^G) \tag{B.6}$$

of Lemma B.1.1.

$$\begin{aligned} D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) &= \mathbb{E}_{P(\mathbf{X})}\left[\log\frac{P\hat{P}}{Q\hat{P}}\right] \\ &= \mathbb{E}_{P(\mathbf{X})}\left[\log\frac{\hat{P}}{Q}\right] + \mathbb{E}_{P(\mathbf{X})}\left[\log\frac{P}{\hat{P}}\right] \\ &= \mathbb{E}_{P(\mathbf{X})}\left[\log\frac{\hat{P}}{Q}\right] + D_{\mathrm{KL}}(P(\mathbf{X})\|\hat{P}(\mathbf{X})) \end{aligned}$$

We now proceed to analyze the remaining term using the definition of $\hat{P}$.

$$\begin{aligned} \mathbb{E}_{P(\mathbf{x})}\left[\log\frac{\hat{P}}{Q}\right] &= \mathbb{E}_{P(\mathbf{x})}\left[\log\prod_{i=1}^{d}\frac{\hat{P}(X_i \mid \mathbf{P}A_i)}{Q(X_i \mid \mathbf{P}A_i)}\right] \\ &= \sum_{i=1}^{d}\mathbb{E}_{P(\mathbf{x})}\left[\log\frac{P(X_i \mid \mathbf{P}A_i)}{Q(X_i \mid \mathbf{P}A_i)}\right] \\ &= \sum_{i=1}^{d}\sum_{x_i,\mathbf{p}a_i}\log\frac{P(x_i \mid \mathbf{p}a_i)}{Q(x_i \mid \mathbf{p}a_i)} \\ &\quad \cdot \sum_{\mathbf{x}_{[d]}\setminus\{x_i,\mathbf{p}a_i\}} P(x_i,\mathbf{p}a_i)\cdot P(\mathbf{x}_{[d]}\setminus\{x_i,\mathbf{p}a_i\} \mid x_i,\mathbf{p}a_i) \\ &= \sum_{i=1}^{d}\sum_{x_i,\mathbf{p}a_i}\log\frac{P(x_i \mid \mathbf{p}a_i)}{Q(x_i \mid \mathbf{p}a_i)}\cdot P(x_i \mid \mathbf{p}a_i)\cdot P(\mathbf{p}a_i) \\ &\quad \cdot \underbrace{\sum_{\mathbf{x}_{[d]}\setminus\{x_i,\mathbf{p}a_i\}} P(\mathbf{x}_{[d]}\setminus\{x_i,\mathbf{p}a_i\} \mid x_i,\mathbf{p}a_i)}_{=1 \text{ for every choice of } (x_i,\mathbf{p}a_i)} \\ &= \sum_{i=1}^{d}\mathbb{E}_{P(\mathbf{p}a_i^G)}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{p}a_i^G)\|Q(X_i \mid \mathbf{p}a_i^G)\right)\right] \end{aligned}$$

To simplify notation, we here leave away the supscript G for the parents and use sums for the expected values, but the continuous case follows with integrals analogously. □

We can now apply the decomposition above to obtain the result:

$$\begin{aligned} \pi_G(P)(\mathbf{X}) = &\operatorname*{argmin}_{\tilde{P}\text{ Markovian w.r.t. }G} D_{\mathrm{KL}}(P(\mathbf{X})\|\tilde{P}(\mathbf{X})) \\ = &\operatorname*{argmin}_{\tilde{P}\text{ Markovian w.r.t. }G} \sum_{i=1}^{d}\mathbb{E}_{P(\mathbf{p}a_i^G)}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{p}a_i^G)\|\tilde{P}(X_i \mid \mathbf{p}a_i^G)\right)\right] \\ &+ D_{\mathrm{KL}}(P(\mathbf{X})\|\hat{P}(\mathbf{X})) \\ = &\operatorname*{argmin}_{\tilde{P}\text{ Markovian w.r.t. }G} \sum_{i=1}^{d}\mathbb{E}_{P(\mathbf{p}a_i^G)}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{p}a_i^G)\|\tilde{P}(X_i \mid \mathbf{p}a_i^G)\right)\right] \end{aligned}$$

Note that the second term in line 2 does not depend on $\tilde{P}$. The claim can then be deduced from noting that $\tilde{P}(\mathbf{X}) = \prod_{i=1}^{d} P(X_i|\mathbf{P}A_i^G)$ sets the sum to zero. □

B.1.4 Proof of Theorem 2.3.5

Theorem 2.3.5 *[Known interventions] Given causal models $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_Q)$ and a set of interventional distributions $\mathscr{P}_{\mathcal{E}} = ((P^e, \mathcal{I}_e))_{e\in\mathcal{E}}$, define Q^e for each environment $e \in \mathcal{E}$ as follows:*

$$Q^e(\mathbf{X}) = \prod_{i\notin\mathcal{I}_e} Q(X_i \mid \mathbf{P}A_i^{G_Q}) \cdot \prod_{j\in\mathcal{I}_e} P^e(X_j \mid \mathbf{P}A_j^{G_Q}) \tag{2.14}$$

where $P^e(X_j \mid \mathbf{P}A_j^{G_Q})$ denotes the mechanisms changed as the result of the environment shift. Then we have that

$$D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q^e(\mathbf{X})) \tag{2.15}$$

Proof. We show the identity using Lemma 2.3.4 noting that Q^e is still Markovian with respect to G_Q (although it may no longer be faithful to G_Q, e.g. if an intervention was hard). Thus, we get the following decomposition

$$\begin{aligned}
&\frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q^e(\mathbf{X})) \\
&= \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} \left[\sum_{i=1}^{d} \mathbb{E}_{P^e(\mathbf{p}a_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{p}a_i^{G_Q}) \| Q^e(X_i \mid \mathbf{p}a_i^{G_Q}) \right) \right] \right. \\
&\qquad \left. + D_{\mathrm{KL}}(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)) \right]
\end{aligned}$$

By assumption, we have that

$$Q^e(X_i \mid \mathbf{P}A_i^{G_Q}) = \begin{cases} P^e(X_i \mid \mathbf{P}A_i^{G_Q}), & \text{if } i \in \mathcal{I}_e \\ Q(X_i \mid \mathbf{P}A_i^{G_Q}) & \text{if } i \in [d] \backslash \mathcal{I}_e \end{cases} \tag{B.7}$$

Thus, the terms corresponding to $i \in \mathcal{I}_e$ vanish yielding

$$\begin{aligned}
&= \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} \left[\sum_{i\in[d]\backslash\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{p}a_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{p}a_i^{G_Q}) \| Q(X_i \mid \mathbf{p}a_i^{G_Q}) \right) \right] \right. \\
&\qquad \left. + D_{\mathrm{KL}} \left(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e) \right) \right] \\
&= D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q})
\end{aligned}$$

□

B.1.5 Proof of Lemma 2.3.6

Lemma 2.3.6 *Let $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_Q)$ be CGMs, such that G_P and G_Q share the same skeleton and $X_i \overset{G_Q}{\to} X_j$, $X_i \overset{G_P}{\leftarrow} X_j$ be a flipped edge between G_P and G_Q. Further, let P^e be an interventional distribution generated from $\mathscr{P}$.*

(i) If there exists an unblocked path $e \xrightarrow{G_P} X_i$ *given* $\mathbf{P}A_j^{G_Q} \backslash X_i$, *then* $P^e(X_j \mid \mathbf{P}A_j^{G_Q}) \neq P(X_j \mid \mathbf{P}A_j^{G_Q})$.

(ii) If there exists an unblocked path $e \xrightarrow{G_P} X_j$ *given* $\mathbf{P}A_i^{G_Q}$, *then* $P^e(X_i \mid \mathbf{P}A_i^{G_Q}) \neq P(X_i \mid \mathbf{P}A_i^{G_Q})$.

Proof. This Lemma follows from Corollary 4.4. by [61] using $\mathbf{Z} = \mathbf{P}A_i^{G_Q}$. If case (i), then there exists an unblocked path to a conditioned child of X_j and thus e is d-connected to X_j resulting in a change in the corresponding mechanism by faithfulness. Otherwise, if case (ii), there is an unblocked path to an unconditioned parent of X_i and therefore e is d-connected to X_i again resulting in a change in the corresponding mechanism.

□

B.1.6 Proof of Theorem 2.3.7

Theorem 2.3.7 *Let* $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_Q)$ *be CGMs and* $\mathcal{E}$ *a set of environments. Now assume that for all (unoriented) edges* $X_i \overset{G_Q}{\rightarrow} X_j$ *there exists an environment* $e \in \mathcal{E}$ *such that one of the following conditions holds:*

(i) There exists an unblocked path $e \xrightarrow{G_Q} X_i$ *given* $\mathbf{P}A_j^{G_Q} \backslash X_i$ *and* $j \notin \mathcal{I}_e$

(ii) There exists an unblocked path $e \xrightarrow{G_Q} X_j$ *given* $\mathbf{P}A_i^{G_Q}$ *and* $i \notin \mathcal{I}_e$

Then, $\mathscr{P}$, $\mathscr{Q}$ *are interventionally equivalent,* $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q}$, *if and only if* $\mathscr{P} = \mathscr{Q}$.

Proof. Let's first assume that $(P, G_P) = (Q, G_Q)$. By our assumption on the multi-environment distributions 2.2.3 and 2.2.4, we have for all $e \in \mathcal{E}$

$$P^e(X_i \mid \mathbf{P}A_i^{G_Q}) = P^e(X_i \mid \mathbf{P}A_i^{G_P}) = P(X_i \mid \mathbf{P}A_i^{G_P}) = Q(X_i \mid \mathbf{P}A_i^{G_Q}) \tag{B.8}$$

if and only if $i \in [d] \backslash \mathcal{I}_e$. Thus, all the conditional divergences in the IKL divergence vanish. Since $G_P = G_Q$, P is Markovian w.r.t. G_Q and so are all interventional distributions. Thus, all the residual terms $D_{\mathrm{KL}}(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)(\mathbf{X})) = 0$ and therefore $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$.

Conversely, assume that $(P, G_P) \neq (Q, G_Q)$. If $P \neq Q$, we have that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) \geq 1/|\mathcal{E}| D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) > 0$. So it suffices to show that $P = Q$, but $G_P \neq G_Q$ implies that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) > 0$. If $P = Q$, then G_Q is necessarily Markov equivalent to G_P and thus G_Q, G_P share the same skeleton and v-structures. Since $G_P \neq G_Q$, there exist adjacent variables X_i, X_j such that $X_j \xrightarrow{G_P} X_i$, but $X_i \xrightarrow{G_Q} X_j$. If condition (i) is satisfied and there is also an unblocked path in G_P $e \xrightarrow{G_P} X_i$ given $\mathbf{P}A_j^{G_Q} \backslash X_i$, we have by Lemma 2.3.6 that $P^e(X_j \mid \mathbf{P}A_j^{G_Q}) \neq P(X_j \mid \mathbf{P}A_j^{G_Q}) = Q(X_j \mid \mathbf{P}A_j^{G_Q})$ which gives us

$$\mathbb{E}_{P^e(\mathbf{P}A_j^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_j \mid \mathbf{P}A_j^{G_Q}) \| Q(X_j \mid \mathbf{P}A_j^{G_Q}) \right) \right] > 0$$

so that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) > 0$, since $j \notin \mathcal{I}_e$. If otherwise, every such path is blocked in G_P, let $(e, X_0, \ldots, X_i)$ be the shortest of such paths in G_Q, which implies that

- None of the intermediate variables is intervened
- Each variable on the path has exactly one parent in G_Q on the path.

If $X_0 = X_i$, the path is also unblocked in G_P, since X_i is directly intervened upon. So, the length of the path is at least 2. Since this path is blocked in G_P, there has to be a mis-oriented edge on the path in G_P compared to G_Q. Let $X_k \overset{G_Q}{\to} X_l$ and $X_l \overset{G_P}{\to} X_k$ denote the first of such mis-oriented edges on the path. But then $(e, X_0, \dots, X_k)$ yields an unblocked path $e \overset{G_P}{\to} X_k$ given $\mathbf{P}A_l^{G_Q} \backslash X_k$ and $X_l \notin \mathcal{I}_e$. Therefore, Lemma 2.3.6 applies, leading to a nonzero term in the IKL divergence. If condition (ii) is satisfied, and there is an unblocked path in G_p $e \overset{G_P}{\to} X_j$ given $\mathbf{P}A_i^{G_Q}$, we have

$$\mathbb{E}_{P^e(\mathbf{P}A_i^{G_Q})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{P}A_i^{G_Q}) \| Q(X_i \mid \mathbf{P}A_i^{G_Q})\right)\right] > 0$$

and consequently $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) > 0$. Otherwise, we get an unblocked path to an intermediate node in the same way as for case (i).

□

B.1.7 Proof of Corollary 2.3.8

Corollary 2.3.8 *Let $\mathscr{P} = (P, G_P)$, $\mathscr{Q} = (Q, G_Q)$ be CGMs with $P = Q$ and $\mathcal{E}$ a set of environments. Assume that we only have partial multi-environmental information in the following ways:*

1) *We only observe a subset of all variables in the causal graph $\mathbf{X}_S \subseteq \mathbf{X}$, i.e. we can only distinguish the graphical sub-models $\mathscr{P}\big|_S, \mathscr{Q}\big|_S$, induced by removing the unobserved variables from distributions and causal graphs.*
2) *We know that the conditions of Theorem 2.3.7 are only satisfied for a subset $E_{\mathcal{E}}$ of all edges in $G_Q\big|_S$.*

We define the restricted IKL divergence via:

$$D^{res}_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \sum_{i \in \mathbf{X}_{E_{\mathcal{E}}} \cap \mathcal{I}_e^C}$$
$$\mathbb{E}_{P^e(\mathbf{p}a_i^{G_Q|_S})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{p}a_i^{G_Q|_S}) \| Q(X_i \mid \mathbf{p}a_i^{G_Q|_S})\right)\right] \tag{2.16}$$

summing only over un-intervened variables X such that $(X, \cdot)$ or $(\cdot, X)$ is contained in $E_{\mathcal{E}}$. Then, $D^{res}_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$ implies that the graphs G_P, G_Q coinicde on the edge set $E_{\mathcal{E}}$. Conversely, if $D^{res}_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) > 0$, there exists a variable $X_i \in \mathbf{X}_{E_{\mathcal{E}}}$ such that $\mathbf{P}A_i^{G_P} \neq \mathbf{P}A_i^{G_Q|_S}$.

Proof. Since $P = Q$, we have that G_P, G_Q are Markov equivalent, so they share the same skeleton. In particular, the subgraphs $G_P|_S, G_Q|_S$ have the same skeleton. By definition, $E_{\mathcal{E}}$ contains exactly those edges for which one of the conditions in Theorem 2.3.7 is satisfied. If one of the edges in $E_{\mathcal{E}}$ was mis-oriented between $G_P|_S, G_Q|_S$ (and therefore G_P, G_Q), this would lead to a nonzero term in the restricted IKL divergence by the same arguments as in proof of Theorem 2.3.7.

Conversely, if we had $\mathbf{P}A_i^{G_P} = \mathbf{P}A_i^{G_Q|_S}$ for all variables X_i that occur in the sum of the IKL divergence, by assumptions 2.2.3 and 2.2.4, all terms in the IKL would vanish.

□

B.1.8 Proof of Theorem 2.4.1

Proof. By Theorem 2.3.5:

$$\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q^e(\mathbf{X})) \le \epsilon$$

Using Pinsker's inequality and the concavity of the square-root function:

$$\begin{aligned}\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} D_{\mathrm{TV}}(P^e(\mathbf{X}), Q^e(\mathbf{X})) &\le \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \sqrt{D_{\mathrm{KL}}(P^e(\mathbf{X}), Q^e(\mathbf{X}))/2} \\ &\le \sqrt{\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}), Q^e(\mathbf{X}))} \\ &\le \sqrt{\epsilon}\end{aligned}$$

By Markov's inequality, for at least $1-\rho$ fraction of e's,

$$D_{\mathrm{TV}}(P^e(\mathbf{X}), Q^e(\mathbf{X})) \le \frac{\sqrt{\epsilon}}{\rho}.$$

The claim now follows from the boundedness of f. □

Appendix to Chapter 3 C

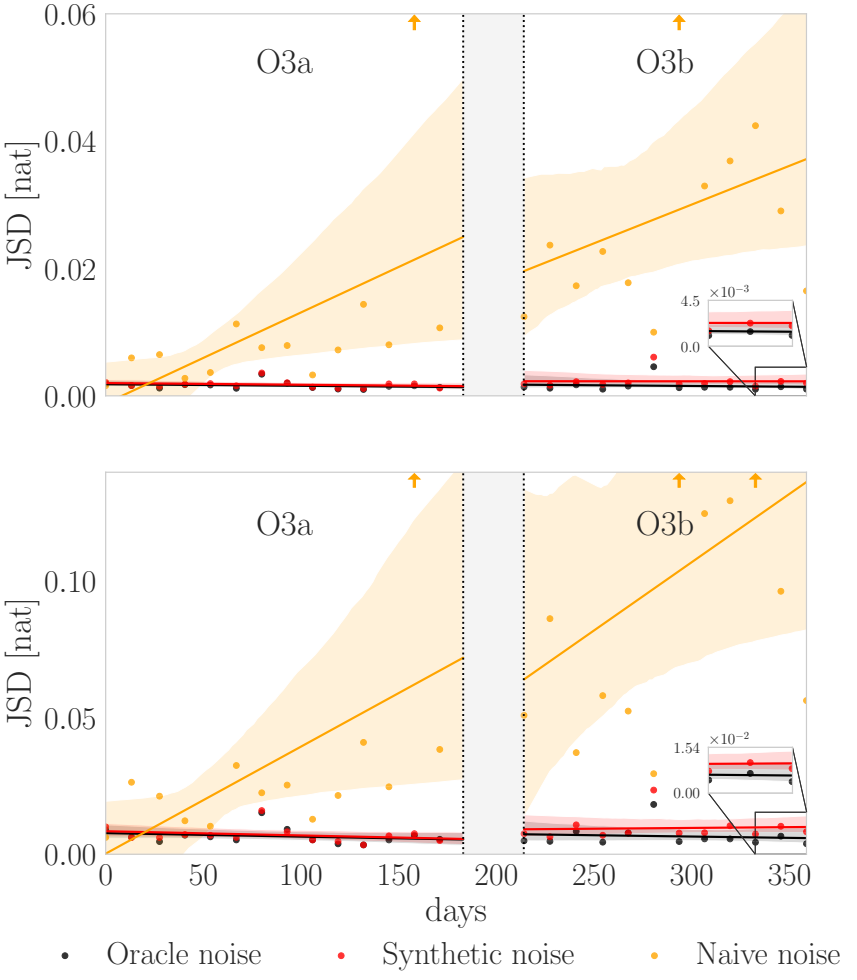

Figure C.1: Average (top) and maximum (bottom) JS-divergences over all parameters between Dingo samples and reference samples. Injections use noise PSDs estimated at various times throughout O3 (indicated on horizontal axis). Results are averaged over injections with five random sets of source parameters (fixed for all PSDs). Day 0 marks the beginning of O3a. The gap indicates the period when detectors were offline between O3a and O3b.

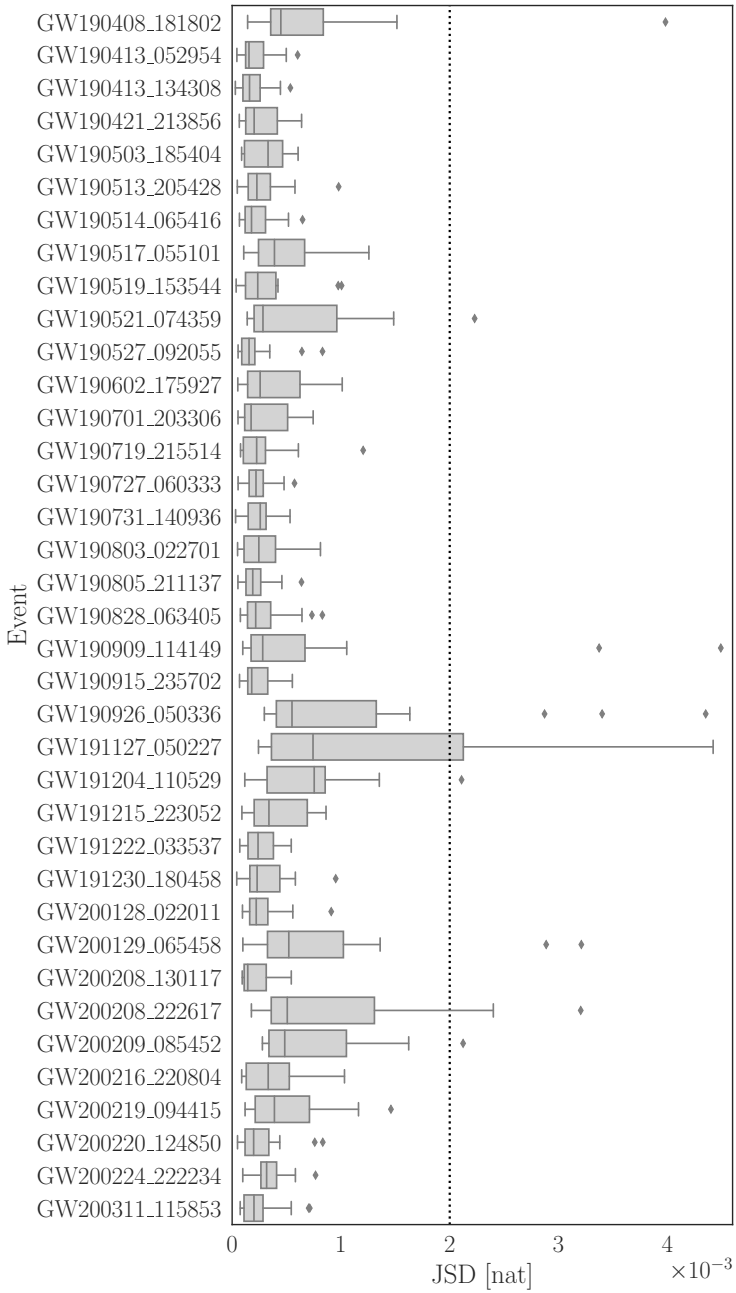

Figure C.2: Comparison of posterior distributions obtained with Dingo-IS and Bilby Dynesty. We here report the median, quartiles and 1.5 IQR whiskers of the JSDs between all one-dimensional marginal distributions. The dotted line marks the indistinguishabilty threshold for the maximum JSD established in [31].

C.1 Comparison to other samplers

In our experiments, we used importance sampling to verify the accuracy of DINGO networks trained with various noise PSD datasets. This approach, called DINGO-IS, has been shown to generate accurate posterior distributions, given a high effective sample size [100]. To validate this claim, and to compare our proposed method with nested sampling, we ran BILBY [169] with the DYNESTY [117] sampler for all 37 real events in section 4.3.2. The box-and-whisker plot of the JSDs between all one-dimensional marginal distributions is presented in Figure C.2. In most cases, the maximum divergences between DINGO-IS and BILBY are well below the indistinguishability threshold of $2 \cdot 10^{-3}$ nat. For events exceeding this threshold, only a small number of parameters are problematic, and the effective sample size is relatively low; see Table 4.2. Overall, we obtain an average maximum JSD of $1.3 \cdot 10^{-3}$ nat and an average mean JSD of $0.4 \cdot 10^{-3}$ nat between DINGO-IS and BILBY, confirming the effectiveness of DINGO-IS for producing reference posterior distributions. Figure C.3, row 3, further compares typical- and worst-case deviations qualitatively. In the typical case, the deviation between DINGO-IS and BILBY is barely visible. While results with DINGO-IS are not perfectly aligned with BILBY's prediction in the worst case, our proposed method still produces distributions that are very close to those of BILBY.

C.2 Additional Results

Figure C.1 shows the mean (top) and maximum (bottom) JSDs between various DINGO posteriors and reference samples, across all inferred parameters, for the study on simulated data in Section 4.3.1. We see that the network trained with *Synthetic* PSDs on average performs similarly to that trained with full PSD information (*Oracle*), with minor deviations for those parameters that are estimated the worst. The average JSD for the *Oracle* dataset is $(1.6 \pm 0.7) \cdot 10^{-3}$ nat, for *Synthetic* is $(2.0 \pm 0.9) \cdot 10^{-3}$ nat, and for *Naive* is $(19.1 \pm 18.8) \cdot 10^{-3}$ nat.

In Figure C.3, we compare posterior marginal distributions of selected O3 events between the reference posterior and our DINGO models. The events in the top row correspond to small and average JSDs between DINGO trained with the *Synthetic* noise PSD dataset and the reference posterior in Table 4.1. With the *Synthetic* PSD distributions, we obtain excellent visual agreement with the reference, even improving upon the *Oracle* PSD distribution in the case of GW190413_134308. By contrast, the *Naive* noise model can lead to significant deviations. In the middle row, we selected two events for which DINGO's predictions with the *Synthetic* noise dataset deviate the most from the reference distributions. However, we do not believe that these inaccuracies are related to the PSD dataset since they are present even when using *Oracle*.

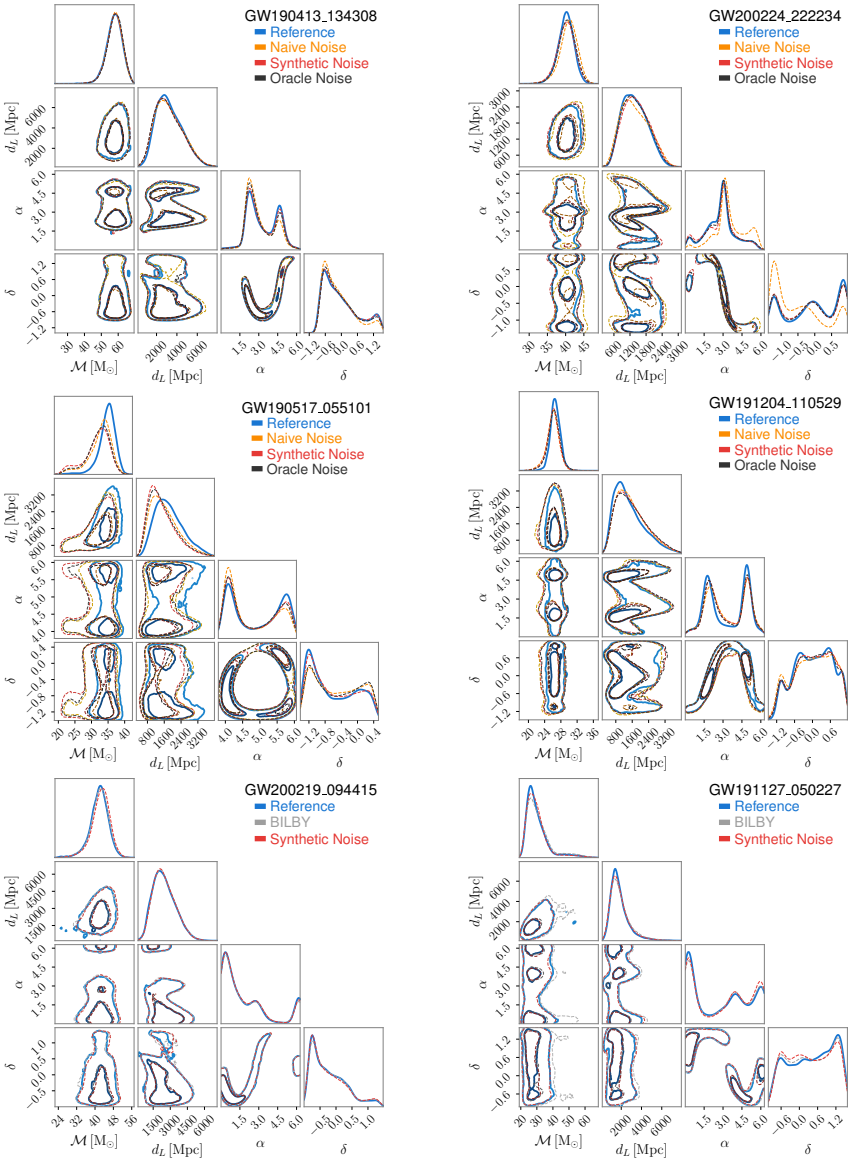

Figure C.3: Marginalized one- and two-dimensional posterior distributions for selected O3 events. Contours represent 50% and 90% credible regions. In the first two rows, we compare DINGO models trained with different PSD datasets to the reference posterior (blue, obtained with importance sampling [100]). Events were chosen to represent small to average deviations between *Synthetic* and the reference (top row) and worst case deviations (middle row). In the bottom row, we further compare our reference results to BILBY (gray), showing an event with average deviation (left) and the event with the largest deviation (right).

Appendix to Chapter 4 D

D.1 Gaussian flow

We here derive the form of a vector field $v_t(\theta)$ that restricts the resulting continuous flow to a one dimensional Gaussian with mean $\hat{\mu}$ variance $\hat{\sigma}^2$. With the optimal transport path $\mu_t(\theta) = t\theta_1$, $\sigma_t(\theta) = 1 - (1 - \sigma_{\min})t \equiv \sigma_t$ from [76], the sample-conditional probability path (3.5) reads

$$p_t(\theta|\theta_1) = \mathcal{N}[t\theta_1, \sigma_t^2](\theta). \tag{D.1}$$

We set our target distribution

$$q_1(\theta_1) = \mathcal{N}[\hat{\mu}, \hat{\sigma}^2](\theta_1). \tag{D.2}$$

To derive the marginal probability path and the marginal vector field we need two identities for the convolution $*$ of Gaussian densities. Recall that the convolution of two function is defined by $f * g(x) = \int f(x-y)g(y)\,\mathrm{d}y$. We define the function

$$g_{\mu,\sigma^2}(\theta) = \theta \cdot \mathcal{N}\left[\mu, \sigma^2\right](\theta). \tag{D.3}$$

Then the following holds

$$\mathcal{N}[\mu_1, \sigma_1^2] * \mathcal{N}[\mu_2, \sigma_2^2] = \mathcal{N}[\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2] \tag{D.4}$$

$$g_{0,\sigma_1^2} * \mathcal{N}[\mu_2, \sigma_2^2] = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\left(g_{\mu_2,\sigma_1^2+\sigma_2^2} - \mu_2 \mathcal{N}[\mu_2, \sigma_1^2 + \sigma_2^2]\right) \tag{D.5}$$

Marginal probability paths

Marginalizing over θ_1 in (D.1) with (D.2), we find

$$\begin{aligned} p_t(\theta) &= \int p_t(\theta|\theta_1) q(\theta_1)\, d\theta_1 \\ &= \int \mathcal{N}\left[t\theta_1, \sigma_t^2\right](\theta)\, \mathcal{N}\left[\hat{\mu}, \hat{\sigma}^2\right](\theta_1) d\theta_1 \\ &= \int \mathcal{N}\left[0, \sigma_t^2\right](\theta - t\theta_1)\, \mathcal{N}(t\hat{\mu}, (t\hat{\sigma})^2)(t\theta_1) \cdot t\, d\theta_1 \\ &= \int \mathcal{N}\left[0, \sigma_t^2\right](\theta - \theta_1^t)\, \mathcal{N}\left[t\hat{\mu}, (t\hat{\sigma})^2\right](\theta_1^t)\, d\theta_1^t \\ &= \mathcal{N}\left[t\hat{\mu}, \sigma_t^2 + (t\hat{\sigma})^2\right](\theta) \end{aligned} \tag{D.6}$$

where we defined $\theta_1^t = t\theta_1$ and used (D.4).

Marginal vector field

We now calculate the marginalized vector field $u_t(\theta)$ based on equation (8) in [76]. Using the sample-conditional vector field (3.6) and the distributions (D.1) and (D.2) we find

$$\begin{aligned}
u_t(\theta) &= \int u_t(\theta|\theta_1)\frac{p_t(\theta|\theta_1)q(\theta_1)}{p_t(\theta)}\,d\theta_1 \\
&= \frac{1}{p_t(\theta)}\int \frac{(\theta_1-(1-\sigma_{\min})\theta)}{\sigma_t}\cdot\mathcal{N}\left[t\theta_1,\sigma_t^2\right](\theta)\cdot\mathcal{N}\left[\hat{\mu},\hat{\sigma}^2\right](\theta_1)\,d\theta_1 \\
&= \frac{1}{p_t(\theta)}\int \frac{(\theta_1-(1-\sigma_{\min})\theta)}{\sigma_t}\cdot\mathcal{N}\left[0,\sigma_t^2\right](\theta-t\theta_1)\cdot\mathcal{N}\left[t\hat{\mu},(t\hat{\sigma})^2\right](t\theta_1)\cdot t\,d\theta_1 \\
&= \frac{1}{p_t(\theta)}\int \frac{\left(\theta_1'-(1-\sigma_{\min})t\cdot\theta\right)}{\sigma_t\cdot t}\cdot\mathcal{N}\left[0,\sigma_t^2\right](\theta-\theta_1')\cdot\mathcal{N}\left[t\hat{\mu},(t\hat{\sigma})^2\right](\theta_1')\cdot d\theta_1' \\
&= \frac{1}{p_t(\theta)}\int \frac{\left(-\theta_1''+(1-(1-\sigma_{\min})t)\cdot\theta\right)}{\sigma_t\cdot t}\cdot\mathcal{N}\left[0,\sigma_t^2\right](\theta_1'')\cdot\mathcal{N}\left[t\hat{\mu},(t\hat{\sigma})^2\right](\theta-\theta_1'')\cdot d\theta_1'' \\
&= \frac{1}{p_t(\theta)}\int \frac{\left(-\theta_1''+\sigma_t\cdot\theta\right)}{\sigma_t\cdot t}\cdot\mathcal{N}\left[0,\sigma_t^2\right](\theta_1'')\cdot\mathcal{N}\left[t\hat{\mu},(t\hat{\sigma})^2\right](\theta-\theta_1'')\cdot d\theta_1''
\end{aligned} \tag{D.7}$$

where we used the change of variables $\theta_1' = t\theta_1$ and $\theta_1'' = \theta - \theta_1'$. Now we evaluate this expression using (D.3), then the identities (D.4) and (D.5) and the marginal probability (D.6)

$$\begin{aligned}
u_t(\theta) &= \frac{-1}{p_t(\theta)\cdot\sigma_t\cdot t}\left(g_{0,\sigma_t^2}*\mathcal{N}\left[t\hat{\mu},(t\hat{\sigma})^2\right]\right)(\theta) + \frac{\theta}{p_t(\theta)\cdot t}\left(\mathcal{N}\left[0,\sigma_t^2\right]*\mathcal{N}\left[t\hat{\mu},(t\hat{\sigma})^2\right]\right)(\theta) \\
&= \frac{-1}{p_t(\theta)\cdot\sigma_t\cdot t}\frac{(\theta-t\hat{\mu})\cdot\sigma_t^2}{\sigma_t^2+(t\hat{\sigma})^2}\cdot\mathcal{N}\left[t\hat{\mu},(\sigma_t^2+(t\hat{\sigma})^2)\right](\theta) + \frac{\theta}{p_t(\theta)\cdot t}\mathcal{N}\left[t\hat{\mu},(\sigma_t^2+(t\hat{\sigma})^2)\right](\theta) \\
&= \frac{(\sigma_t^2+(t\hat{\sigma})^2)\theta-(\theta-t\hat{\mu})\cdot\sigma_t}{p_t(\theta)\cdot t\cdot(\sigma_t^2+(t\hat{\sigma})^2)}\cdot p_t(\theta) \\
&= \frac{(\sigma_t^2+(t\hat{\sigma})^2-\sigma_t)\theta+t\hat{\mu}\cdot\sigma_t}{t\cdot(\sigma_t^2+(t\hat{\sigma})^2)}.
\end{aligned} \tag{D.8}$$

By choosing a vector field v_t of the form (D.8) with learnable parameters $\hat{\mu}, \hat{\sigma}^2$, we can thus define a continuous flow that is restricted to a one dimensional Gaussian.

D.2 Mass covering properties of flows

In this supplement, we investigate the mass covering properties of continuous normalizing flows trained using mean squared error and in particular prove Theorem 3.3.1. We first recall the notation from the main part. We always assume that the data is distributed according to $p_1(\theta)$. In addition, there is a known and simple base distribution p_0 and we assume that there is a vector field $u_t : [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ that connects p_0 and p_1 in the following sense. We denote by ϕ_t the flow generated by u_t, i.e., ϕ_t satisfies

$$\partial_t\phi_t(\theta) = u_t(\phi_t(\theta)). \tag{D.9}$$

Then we assume that $(\phi_1)_* p_0 = p_1$ and we also define the interpolations $p_t = (\phi_t)_* p_0$.

We do not have access to the ground truth distributions p_t and the vector field u_t but we try to learn a vector field v_t approximating u_t. We denote its flow by ψ_t and we define $q_t = (\psi_t)_* q_0$ and $q_0 = p_0$. We are interested in the mass covering properties of the learned approximation q_1 of p_1. In particular, we want to relate the KL-divergence $D_{\mathrm{KL}}(p_1||q_1)$ to the mean squared error,

$$\mathrm{MSE}_p(u, v) = \int_0^1 \mathrm{d}t \int p_t(\mathrm{d}\theta)(u_t(\theta) - v_t(\theta))^2, \tag{D.10}$$

of the generating vector fields. The first observation is that without any regularity assumptions on v_t it is impossible to obtain any bound on the KL-divergence in terms of the mean squared error.

Lemma D.2.1 *For every $\varepsilon > 0$ there are vector field u_t and v_t and a base distribution $p_0 = q_0$ such that*

$$\mathrm{MSE}_p(u, v) < \varepsilon \text{ and } D_{\mathrm{KL}}(p_1||q_1) = \infty. \tag{D.11}$$

In addition we can construct u_t and v_t such that the support of p_1 is larger than the support of q_1.

Proof. We consider the uniform distribution $p_0 = q_0 \sim \mathcal{U}([-1, 1])$ and the vector fields

$$u_t(\theta) = 0 \tag{D.12}$$

and

$$v_t(\theta) = \begin{cases} \varepsilon & \text{for } 0 \le \theta < \varepsilon, \\ 0 & \text{otherwise.} \end{cases} \tag{D.13}$$

As before, let ϕ_t denote the flow of the vector field u_t and similarly ψ_t denote the flow of v_t. Clearly $\phi_t(\theta) = \theta$. On the other hand

$$\psi_t(\theta) = \begin{cases} \min(\theta + \varepsilon t, \varepsilon) & \text{if } 0 \le \theta < \varepsilon, \\ \theta & \text{otherwise.} \end{cases} \tag{D.14}$$

In particular

$$\psi_1(\theta) = \begin{cases} \varepsilon & \text{if } 0 \le \theta < \varepsilon, \\ \theta & \text{otherwise.} \end{cases} \tag{D.15}$$

This implies that $p_1 = (\phi_1)_* p_0 \sim \mathcal{U}([-1, 1])$. On the other hand $q_1 = (\psi_1)_* q_0$ has support in $[-1, 0] \cup [\varepsilon, 1]$. In particular, the distribution of q_1 is not mass covering with respect to p_1 and $D_{\mathrm{KL}}(p_1||q_1) = \infty$. Finally, we observe that the MSE can be arbitrarily small

$$\mathrm{MSE}_p(u, v) = \int_0^1 \mathrm{d}t \int p_t(\mathrm{d}\theta)|u_t(\theta) - v_t(\theta)|^2 = \int_0^1 \int_0^\varepsilon \frac{1}{2}\varepsilon^2 = \frac{\varepsilon^3}{2}. \tag{D.16}$$

Here we used that the density of $p_t(\mathrm{d}\theta)$ is $1/2$ for $-1 \le \theta \le 1$. □

We see that an arbitrary small MSE-loss cannot ensure that the probability distribution is mass covering and the KL-divergence is finite. On a high level this can be explained by the fact that for vector fields v_t that are not Lipschitz continuous the flow is not necessarily continuous, and

we can generate holes in the distribution. Note that we chose p_0 to be a uniform distribution for simplicity, but the result extends to any smooth distribution, in particular the result does not rely on the discontinuity of p_0.

Next, we investigate the mass covering property for Lipschitz continuous flows. When the flows u_t and v_t are Lipschitz continuous (in θ) this ensures that the flows ψ_1 and ϕ_1 are continuous in x and it is not possible to create holes in the distribution as shown above for non-continuous vector fields. We show a weaker bound in this setting.

Lemma D.2.2 *For every $0 \leq \delta \leq 1$ there is a base distribution $p_0 = q_0$ and the are Lipschitz-continuous vector fields u_t and v_t such that $\mathrm{MSE}_p(u, v) = \delta$ and*

$$D_{\mathrm{KL}}(p_1||q_1) \geq \frac{1}{2}\mathrm{MSE}_p(u, v)^{1/3}. \tag{D.17}$$

Proof. We consider p_0, q_0 and u_t as in Lemma D.2.1, and we define

$$v_t(\theta) = \begin{cases} 2\theta & \text{for } 0 \leq \theta < \varepsilon, \\ 2\varepsilon - \theta & \text{for } \varepsilon \leq \theta < 2\varepsilon, \\ 0 & \text{otherwise.} \end{cases} \tag{D.18}$$

Then we can calculate for $0 \leq \theta \leq e^{-2}\varepsilon$ that

$$\psi_t(\theta) = \theta e^{2t}. \tag{D.19}$$

Similarly we obtain for $\varepsilon \leq \theta \leq 2\varepsilon$ (solving the ODE $f' = 2f$)

$$\psi_t(\theta) = 2\varepsilon - (2\varepsilon - \theta)e^{-2t}. \tag{D.20}$$

We find

$$\psi_1(0) = 0,\ \psi_1(e^{-2}\varepsilon) = \varepsilon,\ \psi_1(\varepsilon) = 2 - \varepsilon e^{-2};\ \psi_2(2\varepsilon) = 2\varepsilon. \tag{D.21}$$

Next we find for the densities of q_1 that

$$q_1(\psi_1(\theta)) = q_0(\theta)|\psi_1'(\theta)|^{-1} = \frac{1}{2}\begin{cases} e^{-2} & \text{for } 0 \leq \theta \leq e^{-2}\varepsilon, \\ e^2 & \text{for } \varepsilon \leq \theta \leq 2\varepsilon. \end{cases} \tag{D.22}$$

Together with (D.21) this implies that the density of q_1 is given by

$$q_1(\theta) = \frac{1}{2}\begin{cases} e^{-2} & \text{for } 0 \leq \theta \leq \varepsilon, \\ e^2 & \text{for } 2\varepsilon - \varepsilon e^{-2} \leq \theta \leq 2\varepsilon. \end{cases} \tag{D.23}$$

Note that $p_1(\theta) = 1/2$ for $-1 \leq \theta \leq 1$ and therefore

$$\int_0^\varepsilon \ln \frac{p_1(\theta)}{q_1(\theta)} p_1(\mathrm{d}\theta) = \int_0^\varepsilon \ln(e^2)\frac{1}{2}\mathrm{d}\theta = \varepsilon, \tag{D.24}$$

and

$$\int_{2\varepsilon-\varepsilon e^{-2}}^{2\varepsilon} \ln \frac{p_1(\theta)}{q_1(\theta)} p_1(\mathrm{d}\theta) = \int_{2\varepsilon-\varepsilon e^{-2}}^{2\varepsilon} \ln(e^{-2}) \frac{1}{2} \mathrm{d}\theta = -\varepsilon e^{-2}. \tag{D.25}$$

Moreover we note

$$\int_{\varepsilon}^{2\varepsilon-\varepsilon e^{-2}} q_1(\mathrm{d}\varepsilon) = \int_{e^{-2}\varepsilon}^{\varepsilon} q_0(\mathrm{d}\varepsilon) = \frac{1}{2}\varepsilon(1-e^{-2}) = \int_{\varepsilon}^{2\varepsilon-\varepsilon e^{-2}} p_1(\mathrm{d}\varepsilon), \tag{D.26}$$

which implies (by positivity of the KL-divergence) that

$$\int_{\varepsilon}^{2\varepsilon-\varepsilon e^{-2}} \ln\left(\frac{p_1(\theta)}{q_1(\theta)}\right) p_1(\mathrm{d}\theta) \geq 0. \tag{D.27}$$

We infer using also $p_1(\theta) = q_1(\theta) = 1/2$ for $\theta \in [-1,0] \cap [2\varepsilon, 1]$ that

$$D_{\mathrm{KL}}(p_1||q_1) = \int \ln\left(\frac{p_1(\theta)}{q_1(\theta)}\right) p_1(\mathrm{d}\theta) \geq \varepsilon(1-e^{-2}). \tag{D.28}$$

On the other hand we can bound

$$\int_0^1 \mathrm{d}t \int p_t(\mathrm{d}\theta)|v_t(\theta)-u_t(\theta)|^2 = \frac{1}{2}\int_0^1 \mathrm{d}t \int_0^{2\varepsilon} |u_t(\theta)|^2 = \int_0^{\varepsilon} s^2 \,\mathrm{d}s = \frac{\varepsilon^3}{3}. \tag{D.29}$$

We conclude that

$$D_{\mathrm{KL}}(p_1||q_1) \geq \frac{1}{2}\left(\mathrm{MSE}_p(u,v)\right)^{1/3}. \tag{D.30}$$

In particular, it is not possible to bound the KL-divergence by the MSE even when the vector fields are Lipschitz continuous. □

Let us put this into context. It was already shown in [101] that we can, in general, not bound the forward KL-divergence by the mean squared error and our Lemmas D.2.1 and D.2.2 are concrete examples. On the other hand, when considering SDEs the KL-divergence can be bounded by the mean squared error of the drift terms as shown in [102]. Indeed, in [101] the favorable smoothing effect was carefully investigated.

Here we show that we can alternatively obtain an upper bound on the KL-divergence when assuming that u_t, v_t, and p_0 satisfy additional regularity assumptions. This allows us to recover the mass covering property from bounds on the means squared error for sufficiently smooth vector fields. The scaling is nevertheless still weaker than for SDEs.

We now state our assumptions. We denote the gradient with respect to θ by $\nabla = \nabla_\mu$ and second derivatives by $\nabla^2 = \nabla^2_{\mu\nu}$. When applying the chain rule, we leave the indices implicit. We denote by $|\cdot|$ the Frobenius norm $|A| = \left(\sum_{ij} A_{ij}^2\right)^{1/2}$ of a matrix. The Frobenius norm is submultiplicative, i.e., $|AB| \leq |A| \cdot |B|$ and directly generalizes to higher order tensors.

Assumption D.2.1 *We assume that*

$$|\nabla u_t| \leq L,\ |\nabla v_t| \leq L,\ |\nabla^2 u_t| \leq L',\ |\nabla^2 v_t| \leq L'. \tag{D.31}$$

We require one further assumption on p_0.

Assumption D.2.2 *There is a constant C_1 such that*

$$|\nabla \ln p_0(\theta)| \leq C_1(1 + |\theta|). \tag{D.32}$$

We also assume that

$$\mathbb{E}_{p_0} |\theta|^2 < C_2 < \infty. \tag{D.33}$$

Note that (D.32) holds, e.g., if p_0 follows a Gaussian distribution but also for smooth distribution with slower decay at ∞. If we assume that $|\nabla \ln p_0(\theta)|$ is bounded the proof below simplifies slightly. This is, e.g., the case if $p_0(\theta) \sim e^{-|\theta|}$ as $|\theta| \to \infty$.

We need some additional notation. It is convenient to introduce $\phi_t^s = \phi_t \circ (\phi_s)^{-1}$, i.e., the flow from time s to t (in particular $\phi_t^0 = \phi_t$) and similarly for ψ. We can now restate and prove Theorem 3.3.1.

Theorem D.2.3 *Let $p_0 = q_0$ and assume u_t and v_t are two vector fields whose flows satisfy $p_1 = (\phi_1)_* p_0$ and $q_1 = (\psi_1)_* q_0$. Assume that p_0 satisfies Assumption D.2.2 and u_t and v_t satisfy Assumption D.2.1. Then there is a constant $C > 0$ depending on L, L', C_1, C_2, and d such that (for $\mathrm{MSE}_p(u, v) < 1$))*

$$D_{\mathrm{KL}}(p_1 || q_1) \leq C \, \mathrm{MSE}_p(u, v)^{\frac{1}{2}}. \tag{D.34}$$

Remark D.2.1 We do not claim that our results are optimal, it might be possible to find similar bounds for the forward KL-divergence with weaker assumptions. However, we emphasize that Lemma D.2.2 shows that the result of the theorem is not true without the assumption on the second derivative of v_t and u_t.

Proof. We want to control $D_{\mathrm{KL}}(p_1 || q_1)$. It can be shown that (see equation above (25) in [102] or Lemma 2.19 in [101])

$$\partial_t D_{\mathrm{KL}}(p_t || q_t) = -\int p_t(\mathrm{d}\theta)(u_t(\theta) - v_t(\theta)) \cdot (\nabla \ln p_t(\theta) - \nabla \ln q_t(\theta)). \tag{D.35}$$

Using Cauchy-Schwarz we can bound this by

$$\partial_t D_{\mathrm{KL}}(p_t || q_t) \leq \left(\int p_t(\mathrm{d}\theta) |u_t(\theta) - v_t(\theta)|^2 \right)^{\frac{1}{2}} \left(\int p_t(\mathrm{d}\theta) |\nabla \ln p_t(\theta) - \nabla \ln q_t(\theta)|^2 \right)^{\frac{1}{2}}. \tag{D.36}$$

We use the relation (see (3.3))

$$\ln(p_t(\phi_t(\theta_0)) = \ln(p_0(\theta_0)) - \int_0^t (\operatorname{div} u_s)(\phi_s(\theta_0)) \mathrm{d}s, \tag{D.37}$$

which can be equivalently rewritten (setting $\theta = \phi_t \theta_0$) as

$$\ln(p_t(\theta)) = \ln(p_0(\phi_0^t \theta)) - \int_0^t (\operatorname{div} u_s)(\phi_s^t \theta) \mathrm{d}s. \tag{D.38}$$

We use the following relation for $\nabla\phi_s^t$

$$\nabla\phi_s^t(\theta) = \exp\left(\int_t^s \mathrm{d}\tau\, (\nabla u_\tau)(\phi_\tau^t(\theta))\right). \tag{D.39}$$

This relation is standard and can be directly deduced from the following ODE for $\nabla\phi_s^t$

$$\partial_s \nabla\phi_s^t(\theta) = \nabla\partial_s\phi_s^t(\theta) = \nabla(u_s(\phi_s^t(\theta))) = \left((\nabla u_s)(\phi_s^t(\theta))\right)\cdot\nabla\phi_s^t(\theta). \tag{D.40}$$

We can conclude that for $0 \le s, t \le 1$ the bound

$$|\nabla\phi_s^t(\theta)| \le e^L \tag{D.41}$$

holds. We find

$$\begin{aligned} |\nabla \ln(p_t(\theta))| &= \left|\nabla \ln(p_0)(\phi_0^t\theta)\cdot\nabla\phi_0^t(\theta) - \int_0^t (\nabla \operatorname{div} u_s)(\phi_s^t\theta)\cdot\nabla\phi_s^t(\theta)\mathrm{d}s\right| \\ &\le |\nabla \ln(p_0)(\phi_0^t\theta)|e^L + L'e^L, \end{aligned} \tag{D.42}$$

and a similar bound holds for q_t. In words, we have shown that the score of p_t at θ can be bounded by the score of p_0 of theta transported along the vector field u_t minus a correction which quantifies the change of score along the path. We now bound using the definition $p_t = (\phi_t)_* p_0$ and the assumption (D.32)

$$\begin{aligned} \int p_t(\mathrm{d}\theta)|\nabla \ln p_0(\phi_0^t(\theta))|^2 &= \int p_0(\mathrm{d}\theta_0)|\nabla \ln p_0(\phi_0^t\phi_t(\theta_0))|^2 = \mathbb{E}_{p_0}|\nabla \ln p_0(\theta_0)|^2 \\ &\le \mathbb{E}_{p_0}(C_1(1+|\theta_0|)^2) \le 2C_1^2(1+\mathbb{E}_{p_0}|\theta_0|^2) \le 2C_1^2(1+C_2^2). \end{aligned} \tag{D.43}$$

Similarly we obtain using $q_0 = p_0$

$$\int p_t(\mathrm{d}\theta)|\nabla \ln q_0(\psi_0^t\theta)|^2 = \int p_0(\mathrm{d}\theta_0)|\nabla \ln q_0(\psi_0^t\phi_t\theta_0)|^2. \tag{D.44}$$

In words, to control the score of q integrated with respect to p_t we need to control the distortion we obtain when moving forward with u and backwards with v. We investigate $\psi_0^t\phi_t(\theta_0)$. We now show

$$\partial_h \psi_t^{t+h}\phi_{t+h}^t(\theta)|_{h=0} = u_t(\theta) - v_t(\theta). \tag{D.45}$$

First, by definition of ϕ, we find

$$\partial_h \phi_{t+h}^t(\theta)|_{h=0} = \partial_h \phi_{t+h}\phi_t^{-1}(\theta)|_{h=0} = u_t(\phi_t\phi_t^{-1}(\theta)) = u_t(\theta). \tag{D.46}$$

To evaluate the second contribution we observe

$$\begin{aligned} 0 = \partial_h\theta|_{h=0} &= \partial_h\psi_{t+h}^{t+h}(\theta)|_{h=0} = \partial_h\psi_{t+h}\psi_{t+h}^{-1}(\theta)|_{h=0} \\ &= (\partial_h\psi_{t+h})\psi_t^{-1}(\theta)|_{h=0} + \psi_t(\partial_h\psi_{t+h}^{-1})(\theta)|_{h=0} = v_t(\psi_t\psi_t^{-1}(\theta)) + \partial_h\psi_t\psi_{t+h}^{-1}(\theta)|_{h=0} \\ &= v_t(\theta) + \partial_h\psi_t^{t+h}(\theta)|_{h=0} \end{aligned} \tag{D.47}$$

Now (D.45) follows from (D.46) and (D.47) together with $\phi_t^t = \psi_t^t = \mathrm{Id}$. Using (D.45) we find

$$\partial_t(\psi_0^t\phi_t)(\theta_0) = \partial_h(\psi_0^t\psi_t^{t+h}\phi_{t+h}^t\phi_t)(\theta_0)|_{h=0} = (\nabla\psi_0^t)(\phi_t(\theta_0))\cdot\left((u_t - v_t)(\phi_t(\theta_0))\right). \tag{D.48}$$

Using (D.41) we conclude that

$$\begin{aligned} |\psi_0^t\phi_t(\theta_0) - \theta_0| &\le \left| \int_0^t \partial_s \psi_0^s \phi_s(\theta_0)\, \mathrm{d}s \right| \le \int_0^t |(\nabla \psi_0^s)(\phi_s(\theta_0))| \cdot |u_s - v_s|(\phi_s(\theta_0))\, \mathrm{d}s \\ &\le e^L \int_0^t |u_s - v_s|(\phi_s(\theta_0))\, \mathrm{d}s. \end{aligned} \tag{D.49}$$

We use this and the assumption (D.32) to continue to estimate (D.44) as follows

$$\begin{aligned} \int p_t(\mathrm{d}\theta) |\nabla \ln q_0(\psi_0^t\theta)|^2 &= \int p_0(\mathrm{d}\theta_0) |\nabla \ln q_0(\psi_0^t\phi_t(\theta_0))|^2 \\ &\le C_1^2 \int p_0(\mathrm{d}\theta_0)(1 + |\psi_0^t\phi_t(\theta_0)|)^2 \\ &\le C_1^2 \int p_0(\mathrm{d}\theta_0)(1 + |\psi_0^t\phi_t(\theta_0) - \theta_0| + |\theta_0|)^2 \\ &\le 3C_1^2 + 3C_1^2 \int p_0(\mathrm{d}\theta_0) \left(|\psi_0^t\phi_t(\theta_0) - \theta_0|^2 + |\theta_0|^2 \right) \\ &\le 3C_1^2(1 + \mathbb{E}_{p_0}|\theta_0|^2) + 3C_1^2 e^{2L} \int p_0(\mathrm{d}\theta_0) \left(\int_0^t \mathrm{d}s\, |u_s - v_s|(\phi_s(\theta_0)) \right)^2. \end{aligned} \tag{D.50}$$

Here we used $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ in the second to last step. We bound the remaining integral using Cauchy-Schwarz as follows

$$\begin{aligned} \int p_0(\mathrm{d}\theta_0) \left(\int_0^t |u_s - v_s|(\phi_s(\theta_0)) \right)^2 &\le \int p_0(\mathrm{d}\theta_0) \left(\int_0^t \mathrm{d}s\, |u_s - v_s|^2(\phi_s(\theta_0)) \right) \left(\int_0^t \mathrm{d}s\, 1^2 \right) \\ &\le t \int_0^t \mathrm{d}s \int p_0(\mathrm{d}\theta_0) |u_s - v_s|^2(\phi_s(\theta_0)) \\ &= t \int_0^t \mathrm{d}s \int p_s(\mathrm{d}\theta_s) |u_s - v_s|^2(\theta_s) \\ &\le \int_0^1 \mathrm{d}s \int p_s(\mathrm{d}\theta_s) |u_s - v_s|^2(\theta_s) = \mathrm{MSE}_p(u, v). \end{aligned} \tag{D.51}$$

The last displays together imply

$$\int p_t(\mathrm{d}\theta) |\nabla \ln q_0(\psi_0^t\theta)|^2 \le 3C_1^2 \left(1 + \mathbb{E}_{p_0}|\theta_0|^2 + e^{2L} \mathrm{MSE}_p(u, v) \right). \tag{D.52}$$

Now we have all the necessary ingredients to bound the derivative of the KL-divergence. We control the second integral in (D.36) using (D.42) (and again $(\sum_{i=1}^4 a_i)^2 \le 4 \sum a_i^2$) as follows,

$$\begin{aligned} &\int p_t(\mathrm{d}\theta) |\nabla \ln p_t(\theta) - \nabla \ln q_t(\theta)|^2 \\ &\qquad \le 2 \cdot 2^2 \cdot L'^2 e^{2L} + 4e^{2L} \int p_t(\mathrm{d}\theta) \left(|\nabla \ln q_0(\psi_0^t)\theta)|^2 + |\nabla \ln p_0(\phi_0^t)\theta)|^2 \right). \end{aligned} \tag{D.53}$$

Using (D.43) and (D.52) we finally obtain

$$\begin{aligned}\int p_t(\mathrm{d}\theta)|\nabla \ln p_t(\theta) - \nabla \ln q_t(\theta)|^2 &\leq 8 \cdot L'^2 e^{2L} + C_1^2 e^{2L} \left(20(1 + C_2^2) + 12\,\mathrm{MSE}_p(u, v)\right) \\ &\leq C(1 + \mathrm{MSE}_p(u, v))\end{aligned} \tag{D.54}$$

for some constant $C > 0$. Finally, we obtain

$$\begin{aligned}D_{\mathrm{KL}}(p_1||q_1) &= \int_0^1 \mathrm{d}t\, \partial_t D_{\mathrm{KL}}(p_t||q_t) \\ &\leq (C(1 + \mathrm{MSE}_p(u, v)))^{\frac{1}{2}} \int_0^1 \mathrm{d}t \left(\int p_t(\mathrm{d}\theta)|u_t(\theta) - v_t(\theta)|^2\right)^{\frac{1}{2}} \\ &\leq (C(1 + \mathrm{MSE}_p(u, v)))^{\frac{1}{2}} \left(\int_0^1 \mathrm{d}t \int p_t(\mathrm{d}\theta)|u_t(\theta) - v_t(\theta)|^2\right)^{\frac{1}{2}} \\ &\leq (C(1 + \mathrm{MSE}_p(u, v)))^{\frac{1}{2}}\, \mathrm{MSE}_p(u, v)^{\frac{1}{2}}.\end{aligned} \tag{D.55}$$

□

D.3 SBI Benchmark

In this section, we collect missing details and additional results for the analysis of the SBI benchmark in Section 3.4.

D.3.1 Network architecture and hyperparameters

For each task and simulation budget in the benchmark, we perform a mild hyperparameter optimization. We sweep over the batch size and learning rate (which is particularly important as the simulation budgets differ by orders of magnitudes), the network size and the α parameter for the time prior defined in Section 3.3.3 (see Tab. D.1 for the specific values). We reserve 5% of the simulation budget for validation and choose the model with the best validation loss across all configurations.

D.3.2 Additional results

We here provide various additional results for the SBI benchmark. First, we compare the performance of FMPE and NPE when using the Maximum Mean Discrepancy metric (MMD). The results can be found in Fig. D.1. FMPE shows superior performance to NPE for most tasks and simulation budgets. Compared to the C2ST scores in Fig. 3.4 the improvement shown by FMPE in MMD is more substantial.

Fig. D.2 compares the FMPE results with the optimal transport path from the main text with a comparable score matching model using the Variance Preserving diffusion path [75]. The score matching results were obtained using the same batch size, network size and learning rate as the FMPE network, while optimizing for $\beta_{\min} \in \{0.1, 1, 4\}$ and $\beta_{\max} \in \{4, 7, 10\}$. FMPE with the optimal transport path clearly outperforms the score-based model on almost all configurations.

Table D.1: Sweep values for the hyperparamters for the SBI benchmark. We split the configurations according to simulation budgets, e.g. for 1000 simulations, we only swept over smaller values for network size and batch size. The network architecture has a diamond shape, with increasing layer width from smallest to largest and then decreasing to the output dimension. Each block consists of two fully-connected residual layers.

hyperparameter	sweep values
hidden dimensions	2^n for $n \in \{4, \dots, 10\}$
number of blocks	$10, \dots, 18$
batch size	2^n for $n \in \{2, \dots, 9\}$
learning rate	1.e-3, 5.e-4, 2.e-4, 1.e-4
α (for time prior)	-0.25, -0.5, 0, 1, 4

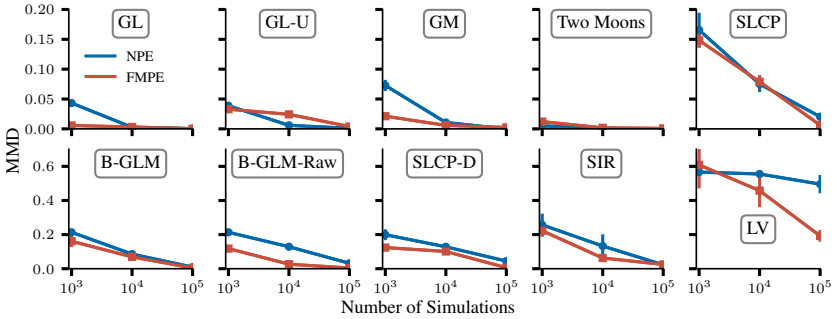

Figure D.1: Comparison of FMPE and NPE performance across 10 SBI benchmarking tasks [105]. We here quantify the deviation in terms of the Maximum Mean Discrepancy (MMD) as an alternative metric to the C2ST score used in Fig. 3.4. MMD can be sensitive to its hyperparameters [105], so we use the C2ST score as a primary performance metric.

In Fig. D.3 we compare FMPE using the architecture proposed in Section 3.3.2 with (t, θ)-conditioning via gated linear units to FMPE with a naive architecture operating directly on the concatenated (t, θ, x) vector. For the two displayed tasks the context dimension $\dim(x) = 100$ is much larger than the parameter dimension $\dim(\theta) \in \{5, 10\}$, and there is a clear performance gain in using the GLU conditioning. Our interpretation is that the low dimensionality of (t, θ) means that it is not well-learned by the network when simply concatenated with x.

Fig. D.4 displays the densities of the reference samples under the FMPE model as a histogram for all tasks (extended version of Fig. 3.3). The support of the learned model $q(\theta|x)$ covers the reference samples $\theta \sim p(\theta|x)$, providing additional empirical evidence for the mass-covering behavior theoretically explored in Thm. 3.3.1. However, samples from the true posterior distribution may have a small density under the learned model, especially if the deviation between model and reference is high; see Lotka-Volterra (bottom right panel). Fig. D.5 displays P–P plots for two selected tasks.

Finally, we study the impact of our time prior re-weighting for one example task in Fig. D.6. We clearly see that our proposed re-weighting leads to increased performance by up-weighting samples for t closer to 1 during training.

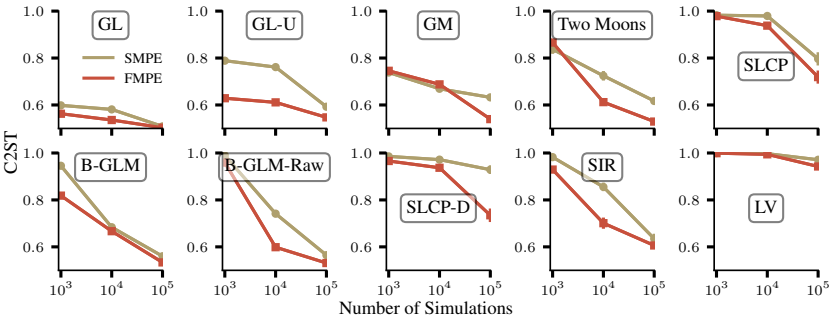

Figure D.2: Comparison of FMPE with the optimal transport path (as used throughout the main paper) with comparable models trained with a Variance Preserving diffusion path [75] by regressing on the score (SMPE). Note that the SMPE baseline shown here is not directly comparable to NPSE [68, 69], as this method uses Langevin steps, which reduces the dependence of the results on the vector field for small t (at the cost of a tractable density).

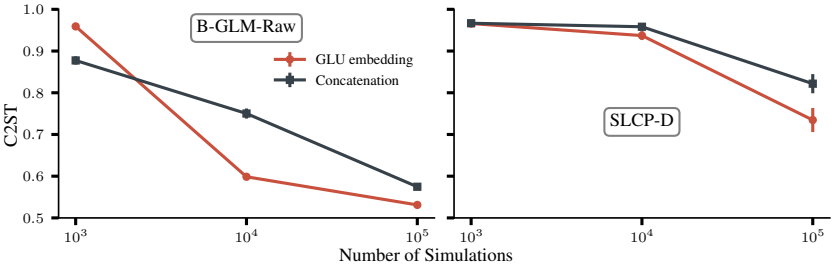

Figure D.3: Comparison of the architecture proposed in Section 3.3.2 with gated linear units for the (t, θ)-conditioning (red) and a naive architecture based on a simple concatenation of (t, θ, x) (black). FMPE with the proposed architecture performs substantially better.

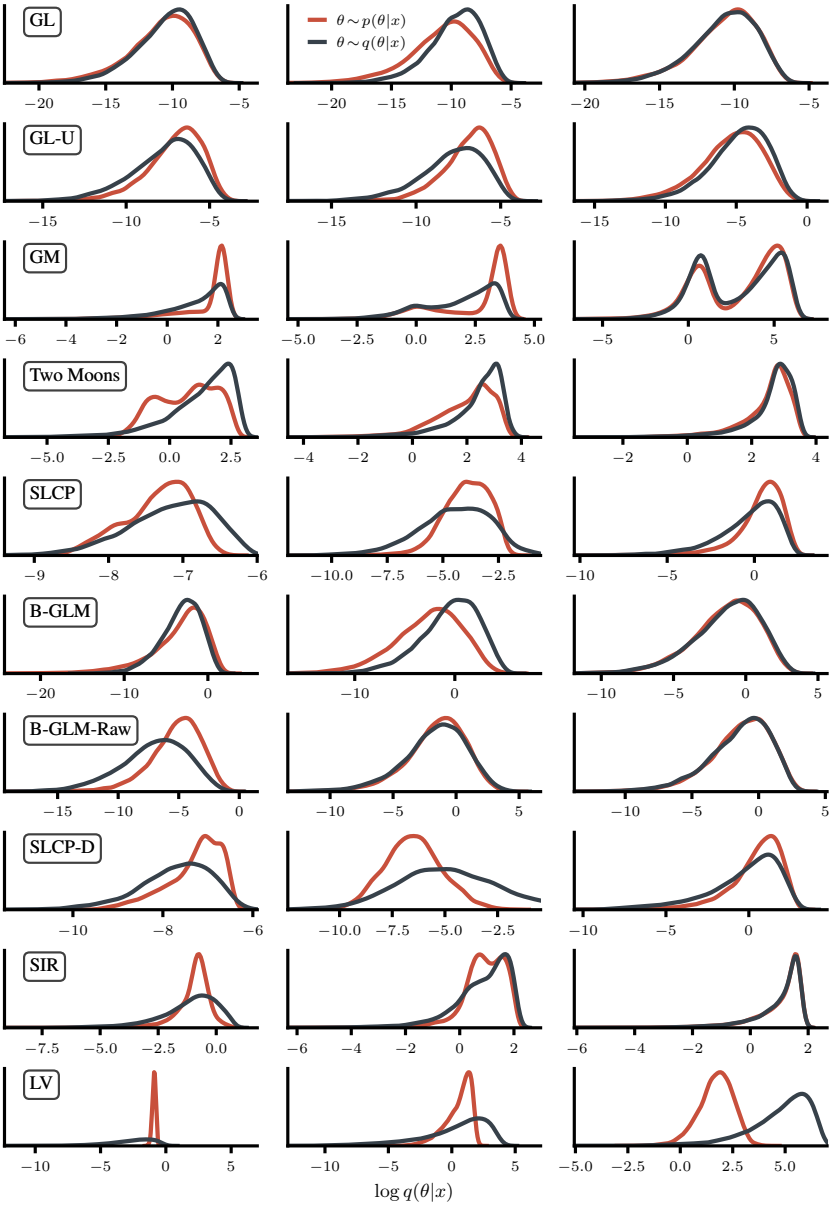

Figure D.4: Histogram of FMPE densities $\log q(\theta|x)$ for samples $\theta \sim q(\theta|x)$ and reference samples $\theta \sim p(\theta|x)$ for simulation budgets $N = 10^3$ (left), $N = 10^4$ (center) and $N = 10^5$ (right). The reference samples $\theta \sim p(\theta|x)$ are all within the support of the learned model $q(\theta|x)$, indicating mass covering FMPE results. Nonetheless, reference samples may have a small density under $q(\theta|x)$, if the validation loss is high, see Lotka-Volterra (LV).

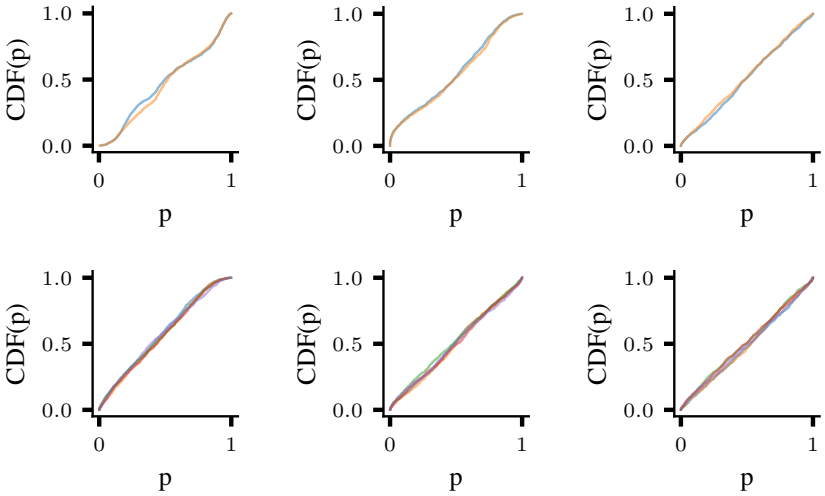

Figure D.5: P-P plot for the marginals of the FMPE-posterior for the Two Moons (upper) and SLCP (lower) tasks for training budgets of 10^3 (left), 10^4 (center), and 10^5 (right) samples.

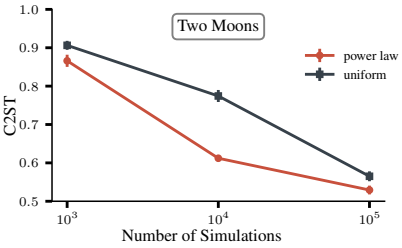

Figure D.6: Comparison of the time prior re-weighting proposed in Section 3.3.3 with a uniform prior over t on the Two Moons task (Section 3.4). The network trained with the re-weighted prior clearly outperforms the reference on all simulation budgets.

D.4 Gravitational-wave inference

We here provide the missing details and additional results for the gravitational wave inference problem analyzed in Section 3.5.

D.4.1 Network architecture and hyperparameters

Compared to NPE with normalizing flows, FMPE allows for generally simpler architectures, since the output of the network is simply a vector field. This also holds for NPSE (model also defined by a vector) and NRE (defined by a scalar). Our FMPE architecture builds on the embedding network developed in [28], however we extend the network capacity by adding more residual blocks (Tab. D.2, top panel). For the (t, θ)-conditioning we use gated linear units applied to each residual block, as described in Section 3.3.2. We also use a small residual network to embed (t, θ) before applying the gated linear units.

In this Appendix we also perform an ablation study, using the *same* embedding network as the NPE network (Tab. D.2, bottom panel). For this configuration, we additionally study the effect of conditioning on (t, θ) starting from different layers of the main residual network.

D.4.2 Data settings

We use the data settings described in [28], with a few minor modifications. In particular, we use the waveform model IMRPhenomPv2 [145, 146, 170] and the prior displayed in Tab. D.3. Generation of the training dataset with 5,000,000 samples takes around 1 hour on 64 CPUs. Compared to [28], we reduce the frequency range from $[20, 1024]$ Hz to $[20, 512]$ Hz to reduce the computational load for data preprocessing. We also omit the conditioning on the detector noise power spectral density (PSD) introduced in [28] as we evaluate on a single GW event. Preliminary tests show that the performance with PSD conditioning is similar to the results reported in this paper. All changes to the data settings have been applied to FMPE and the NPE baselines alike to enable a fair comparison.

D.4.3 Additional results

Tab. D.4 displays the inference times for FMPE and NPE. NPE requires only a single network pass to produce samples and (log-)probabilities, whereas many forwards passes are needed for FMPE to solve the ODE with a specific level of accuracy. A significant portion of the additional time required for calculating (log-)probabilities in conjunction with the samples is spent on computing the divergence of the vector field, see Eq. (3.3).

Fig. D.7 presents a comparison of the FMPE performance using networks of the same hidden dimensions as the NPE embedding network (Tab. D.2 bottom panel). This comparison includes an ablation study on the timing of the (t, θ) GLU-conditioning. In the top-row network, the (t, θ) conditioning is applied only after the 256-dimensional blocks. In contrast, the middle-row network receives (t, θ) immediately after the initial residual block. With FMPE we can achieve performance comparable to NPE, while having only $\approx 1/3$ of the network size (most of the NPE network parameters are in the flow). This suggests that parameterizing the target distribution in terms of a vector field requires less learning capacity, compared to directly learning its density. Delaying the (t, θ) conditioning until the final layers impairs performance. However, the number of FLOPs at inference is considerably reduced, as the context embedding can be cached and a network pass

Table D.2: Hyperparameters for the FMPE models used in the main text (top) and in the ablation study (bottom, see Fig. D.7). The network is composed of a sequence of residual blocks, each consisting of two fully-connected hidden layers, with a linear layer between each pair of blocks. The ablation network is the same as the embedding network that feeds into the NPE normalizing flow.

hyperparameter	values
residual blocks	$2048, 4096\times3, 2048\times3, 1024\times6, 512\times8, 256\times10, 128\times5, 64\times3, 32\times3, 16\times3$
residual blocks (t,θ) embedding	16, 32, 64, 128, 256
batch size	4096
learning rate	5.e-4
α (for time prior)	1
residual blocks	$2048\times2, 1024\times4, 512\times4, 256\times4, 128\times4, 64\times3, 32\times3, 16\times3$
residual blocks (t,θ) embedding	16, 32, 64, 128, 256
batch size	4096
learning rate	5.e-4
α (for time prior)	1

only involves the few layers with the (t,θ) conditioning. Consequently, there's a trade-off between accuracy and inference speed, which we will explore in a greater scope in future work.

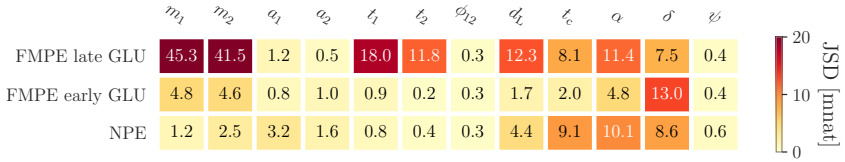

Figure D.7: Jensen-Shannon divergence between inferred posteriors and the reference posteriors for GW150914 [1]. We compare two FMPE models with the same architecture as the NPE embedding network, see Tab. D.2 bottom panel. For the model in the first row, the GLU conditioning of (θ,t) is only applied before the final 128-dim blocks. The model in the middle row is given the context after the very first 2048 block.

Table D.3: Priors for the astrophysical binary black hole parameters. Priors are uniform over the specified range unless indicated otherwise. Our models infer the mass parameters in the basis (M_c, q) and marginalize over the phase parameter ϕ_c.

Description	Parameter	Prior
component masses	m_1, m_2	$[10, 120]\,\mathrm{M}_\odot$, $m_1 \geq m_2$
chirp mass	$M_c = (m_1 m_2)^{\frac{3}{5}}/(m_1 + m_2)^{\frac{1}{5}}$	$[20, 120]\,\mathrm{M}_\odot$ (constraint)
mass ratio	$q = m_2/m_1$	[0.125, 1.0] (constraint)
spin magnitudes	a_1, a_2	$[0, 0.99]$
spin angles	$\theta_1, \theta_2, \phi_{12}, \phi_{JL}$	standard as in [171]
time of coalescence	t_c	$[-0.03, 0.03]$ s
luminosity distance	d_L	[100, 1000] Mpc
reference phase	ϕ_c	$[0, 2\pi]$
inclination	θ_{JN}	$[0, \pi]$ uniform in sine
polarization	ψ	$[0, \pi]$
sky position	α, β	uniform over sky

Table D.4: Inference times per batch for FMPE and NPE on a single Nvidia A100 GPU, using the training batch size of 4096. We solve the ODE for FMPE using the `dopri5` discretization [172] with absolute and relative tolerances of 1e-7. For FMPE, generation of the (log-)probabilities additionally requires the computation of the divergence, see equation (3.3). This needs additional memory and therefore limits the maximum batch size that can be used at inference.

	Network Passes	Inference Time (per batch)
FMPE (sample only)	248	26s
FMPE (sample and log probs)	350	352s
NPE (sample and log probs)	1	1.5s

List of Figures

List of Tables

Bibliography

[1] B.P. Abbott et al. 'Observation of Gravitational Waves from a Binary Black Hole Merger'. In: *Phys. Rev. Lett.* 116.6 (2016), p. 061102. DOI: 10.1103/PhysRevLett.116.061102 (cited on pages 1, 2, 5, 41, 42, 103).

[2] B. P. Abbott et al. 'Astrophysical Implications of the Binary Black-Hole Merger GW150914'. In: *Astrophys. J. Lett.* 818.2 (2016), p. L22. DOI: 10.3847/2041-8205/818/2/L22 (cited on page 1).

[3] Stephen R. Green and Jonathan Gair. 'Complete parameter inference for GW150914 using deep learning'. In: *Mach. Learn. Sci. Tech.* 2.3 (2021), 03LT01. DOI: 10.1088/2632-2153/abfaed (cited on pages 1, 40, 41, 46, 49).

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 'Deep learning'. In: *nature* 521.7553 (2015) (cited on page 1).

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016 (cited on page 1).

[6] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017 (cited on pages 1–3, 6).

[7] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. 'Domain adaptation under target and conditional shift'. In: *International Conference on Machine Learning (ICML)*. PMLR. 2013, pp. 819–827 (cited on page 1).

[8] Judea Pearl. *Causality: Models, Reasoning and Inference*. 2nd. USA: Cambridge University Press, 2009 (cited on pages 1, 13).

[9] B. Schölkopf. 'Empirical Inference'. In: *International Journal of Materials Research* 2011.7 (July 2011), pp. 809–814 (cited on page 1).

[10] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999 (cited on page 2).

[11] Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002 (cited on page 2).

[12] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 'Highly accurate protein structure prediction with AlphaFold'. In: *Nature* 596.7873 (2021), pp. 583–589 (cited on page 2).

[13] Judea Pearl. 'Causality: models, reasoning, and inference'. In: *Econometric Theory* 19.675-685 (2003), p. 46 (cited on page 2).

[14] Judea Pearl. *Causality*. Cambridge university press, 2009 (cited on pages 2, 3).

[15] Guido W. Imbens. 'Matching Methods in Practice: Three Examples'. In: *Journal of Human Resources* 50.2 (2015), pp. 373–419. DOI: 10.3368/jhr.50.2.373 (cited on page 3).

[16] Bernhard Schölkopf. 'Causality for Machine Learning'. In: *arXiv preprint arXiv:1911.10500* (2019) (cited on pages 3, 6).

[17] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. *Towards Causal Representation Learning*. 2021. DOI: 10.48550/ARXIV.2102.11107. URL: https://arxiv.org/abs/2102.11107 (cited on pages 3, 12).

[18] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. 'The frontier of simulation-based inference'. In: *Proc. Nat. Acad. Sci.* 117.48 (2020), pp. 30055–30062. DOI: 10.1073/pnas.1912789117 (cited on pages 4, 7, 30).

[19] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 'Normalizing Flows for Probabilistic Modeling and Inference'. In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64 (cited on pages 4, 7, 30, 32).

[20] George Papamakarios and Iain Murray. 'Fast ε-free Inference of Simulation Models with Bayesian Conditional Density Estimation'. In: *Advances in neural information processing systems*. 2016 (cited on pages 4, 30, 34, 35).

[21] Danilo Rezende and Shakir Mohamed. 'Variational Inference with Normalizing Flows'. In: *International Conference on Machine Learning*. 2015, pp. 1530–1538 (cited on pages 4, 30, 32).

[22] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 'Neural Ordinary Differential Equations'. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018 (cited on pages 4, 33).

[23] Albert Einstein. 'The foundation of the general theory of relativity.' In: *Annalen Phys.* 49.7 (1916). Ed. by Jong-Ping Hsu and D. Fine, pp. 769–822. DOI: 10.1002/andp.19163540702 (cited on page 5).

[24] Nelson Christensen and Renate Meyer. 'Parameter estimation with gravitational waves'. In: *Rev. Mod. Phys.* 94.2 (2022), p. 025001. DOI: 10.1103/RevModPhys.94.025001 (cited on page 5).

[25] J. Aasi et al. 'Advanced LIGO'. In: *Class. Quant. Grav.* 32 (2015), p. 074001. DOI: 10.1088/0264-9381/32/7/074001 (cited on pages 5, 46).

[26] F. Acernese et al. 'Advanced Virgo: a second-generation interferometric gravitational wave detector'. In: *Class. Quant. Grav.* 32.2 (2015), p. 024001. DOI: 10.1088/0264-9381/32/2/024001 (cited on pages 5, 46).

[27] T. Akutsu et al. 'Overview of KAGRA: Detector design and construction history'. In: *PTEP* 2021.5 (2021), 05A101. DOI: 10.1093/ptep/ptaa125 (cited on pages 5, 46).

[28] Maximilian Dax, Stephen R. Green, Jonathan Gair, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. 'Real-Time Gravitational Wave Science with Neural Posterior Estimation'. In: *Phys. Rev. Lett.* 127.24 (2021), p. 241103. DOI: 10.1103/PhysRevLett.127.241103 (cited on pages 5, 30, 40–42, 46, 49, 53, 55, 102).

[29] J. Veitch, V. Raymond, B. Farr, W Farr, P. Graff, S. Vitale, et al. 'Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library'. In: *Phys. Rev.* D91.4 (2015), p. 042003. DOI: 10.1103/PhysRevD.91.042003 (cited on pages 5, 40, 46, 55).

[30] C. Pankow, P. Brady, E. Ochsner, and R. O'Shaughnessy. 'Novel scheme for rapid parallel parameter estimation of gravitational waves from compact binary coalescences'. In: *Phys. Rev. D* 92.2 (2015), p. 023002. DOI: 10.1103/PhysRevD.92.023002 (cited on page 5).

[31] Gregory Ashton et al. 'BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy'. In: *Astrophys. J. Suppl.* 241.2 (2019), p. 27. DOI: 10.3847/1538-4365/ab06fc (cited on pages 5, 40, 46, 55, 85).

[32] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 'The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization'. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 8320–8329. DOI: 10.1109/ICCV48922.2021.00823 (cited on page 6).

[33] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 'Shortcut learning in deep neural networks'. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673 (cited on page 6).

[34] S. Acid and L. M. De Campos. 'Searching for Bayesian Network Structures in the Space of Restricted Acyclic Partially Directed Graphs'. In: *Journal of Artificial Intelligence Research* 18 (2003), pp. 445–490. DOI: 10.1613/jair.1061 (cited on pages 6, 12, 16).

[35] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. 'The max-min hill-climbing Bayesian network structure learning algorithm'. In: *Machine learning* 65.1 (2006), pp. 31–78 (cited on pages 6, 12, 16).

[36] Jonas Peters and Peter Bühlmann. 'Structural Intervention Distance (SID) for Evaluating Causal Graphs'. In: (2013). DOI: 10.48550/ARXIV.1306.1043 (cited on pages 6, 12, 16).

[37] Benjamin P Abbott et al. 'A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals'. In: *Class. Quant. Grav.* 37.5 (2020), p. 055002. DOI: 10.1088/1361-6382/ab685e (cited on page 7).

[38] Jonas Bernhard Wildberger, Siyuan Guo, Arnab Bhattacharyya, and Bernhard Schölkopf. 'On the Interventional Kullback-Leibler Divergence'. In: *Proceedings of the Second Conference on Causal Learning and Reasoning*. Ed. by Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing. Vol. 213. Proceedings of Machine Learning Research. PMLR, 2023, pp. 328–349 (cited on pages 7, 11).

[39] Jonas Wildberger, Maximilian Dax, Simon Buchholz, Stephen Green, Jakob H Macke, and Bernhard Schölkopf. 'Flow Matching for Scalable Simulation-Based Inference'. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 16837–16864 (cited on pages 8, 29).

[40] Jonas Wildberger, Maximilian Dax, Stephen R. Green, Jonathan Gair, Michael Pürrer, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. 'Adapting to noise distribution shifts in flow-based gravitational-wave inference'. In: *Phys. Rev. D* 107 (8 Apr. 2023), p. 084046. DOI: 10.1103/PhysRevD.107.084046 (cited on pages 8, 45).

[41] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017 (cited on pages 12, 14).

[42] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. 'Domain Adaptation under Target and Conditional Shift'. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML'13. Atlanta, GA, USA: JMLR.org, 2013, III–819–III–827 (cited on page 12).

[43] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. *Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style*. 2021. DOI: 10.48550/ARXIV.2106.04619. URL: https://arxiv.org/abs/2106.04619 (cited on page 12).

[44] Cian Eastwood, Alexander Robey, Shashank Singh, Julius von Kügelgen, Hamed Hassani, George J. Pappas, and Bernhard Schölkopf. *Probable Domain Generalization via Quantile Risk Minimization*. 2022. DOI: 10.48550/ARXIV.2207.09944. URL: https://arxiv.org/abs/2207.09944 (cited on page 12).

[45] Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. 'Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data'. In: *arXiv e-prints*, arXiv:2203.15756 (Mar. 2022), arXiv:2203.15756 (cited on pages 12, 16).

[46] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. *Learning and Testing Causal Models with Interventions*. 2018. DOI: 10.48550/ARXIV.1805.09697. URL: https://arxiv.org/abs/1805.09697 (cited on pages 12, 16).

[47] Maxime Peyrard and Robert West. *A Ladder of Causal Distances*. 2020. DOI: 10.48550/ARXIV.2005.02480. URL: https://arxiv.org/abs/2005.02480 (cited on pages 12, 16, 22).

[48] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. 2nd. MIT press, 2000 (cited on page 14).

[49] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009 (cited on page 14).

[50] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. 'On causal and anticausal learning'. In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML'12. Edinburgh, Scotland: Omnipress, 2012, pp. 459–466 (cited on page 14).

[51] Joris M. Mooij, Sara Magliacane, and Tom Claassen. 'Joint Causal Inference from Multiple Contexts'. In: (2016). DOI: 10.48550/ARXIV.1611.10351 (cited on page 15).

[52] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. 'Causal Discovery from Heterogeneous/Nonstationary Data with Independent Changes'. In: (2019). DOI: 10.48550/ARXIV.1903.01672 (cited on pages 15, 26).

[53] Daniel Eaton and Kevin Murphy. 'Exact Bayesian structure learning from uncertain interventions'. In: *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*. Ed. by Marina Meila and Xiaotong Shen. Vol. 2. Proceedings of Machine Learning Research. San Juan, Puerto Rico: PMLR, 21–24 Mar 2007, pp. 107–114 (cited on page 16).

[54] AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. 'Multi-domain Causal Structure Learning in Linear Systems'. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018 (cited on page 16).

[55] Yango He and Zhi Geng. *Causal Network Learning from Multiple Interventions of Unknown Manipulated Targets*. 2016. DOI: 10.48550/ARXIV.1610.08611. URL: https://arxiv.org/abs/1610.08611 (cited on pages 16, 26).

[56] Jin Tian and Judea Pearl. 'A General Identification Condition for Causal Effects'. In: *Eighteenth National Conference on Artificial Intelligence*. Edmonton, Alberta, Canada: American Association for Artificial Intelligence, 2002, pp. 567–573 (cited on page 16).

[57] Alain Hauser and Peter Bühlmann. 'Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs'. In: (2011). DOI: 10.48550/ARXIV.1104.2808 (cited on page 16).

[58] Alain Hauser and Peter Bühlmann. 'Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs'. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 77.1 (2015), pp. 291–318. (Visited on 10/28/2022) (cited on page 16).

[59] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006 (cited on page 17).

[60] Kailash Budhathoki, Dominik Janzing, Patrick Bloebaum, and Hoiyi Ng. 'Why did the distribution change?' In: *AISTATS 2021*. 2021 (cited on page 17).

[61] Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. *Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis*. 2022. DOI: 10.48550/ARXIV.2206.02013. URL: https://arxiv.org/abs/2206.02013 (cited on pages 22, 26, 79).

[62] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. 'Information-geometric approach to inferring causal directions'. In: *Artificial Intelligence* 182-183 (2012), pp. 1–31. DOI: https://doi.org/10.1016/j.artint.2012.01.002 (cited on page 26).

[63] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. 'Invariant Models for Causal Transfer Learning'. In: (2015). DOI: 10.48550/ARXIV.1507.05333 (cited on page 26).

[64] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 'Invariant risk minimization'. In: *arXiv preprint 1907.02893* (2019) (cited on page 26).

[65] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 'Out-of-distribution generalization via risk extrapolation'. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5815–5826 (cited on page 26).

[66] Jan-Matthis Lueckmann, Pedro J Gonçalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. 'Flexible statistical inference for mechanistic models of neural dynamics'. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 1289–1299 (cited on pages 30, 34, 35).

[67] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. 'Automatic posterior transformation for likelihood-free inference'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2404–2414 (cited on pages 30, 34, 35).

[68] Louis Sharrock, Jack Simons, Song Liu, and Mark Beaumont. 'Sequential Neural Score Estimation: Likelihood-Free Inference with Conditional Score Based Diffusion Models'. In: *arXiv preprint arXiv:2210.04872* (2022) (cited on pages 30, 34, 99).

[69] Tomas Geffner, George Papamakarios, and Andriy Mnih. 'Score Modeling for Simulation-based Inference'. In: *arXiv preprint arXiv:2209.14249* (2022) (cited on pages 30, 31, 34, 99).

[70] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 'Deep unsupervised learning using nonequilibrium thermodynamics'. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2256–2265 (cited on pages 30, 34).

[71] Yang Song and Stefano Ermon. 'Generative modeling by estimating gradients of the data distribution'. In: *Advances in neural information processing systems* 32 (2019) (cited on pages 30, 34).

[72] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 'Denoising diffusion probabilistic models'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851 (cited on pages 30, 34).

[73] Prafulla Dhariwal and Alexander Nichol. 'Diffusion Models Beat GANs on Image Synthesis'. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794 (cited on page 30).

[74] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. 'Cascaded Diffusion Models for High Fidelity Image Generation'. In: *J. Mach. Learn. Res.* 23 (2022), 47:1–47:33 (cited on page 30).

[75] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 'Score-based generative modeling through stochastic differential equations'. In: *arXiv preprint arXiv:2011.13456* (2020) (cited on pages 30, 34, 97, 99).

[76] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 'Flow Matching for Generative Modeling'. In: *CoRR* abs/2210.02747 (2022). DOI: 10.48550/arXiv.2210.02747 (cited on pages 31, 33, 34, 89, 90).

[77] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, and Gilles Louppe. 'Averting a crisis in simulation-based inference'. In: *arXiv preprint arXiv:2110.06581* (2021) (cited on pages 32, 39).

[78] Michael S Albergo and Eric Vanden-Eijnden. 'Building normalizing flows with stochastic interpolants'. In: *arXiv preprint arXiv:2209.15571* (2022) (cited on page 34).

[79] Xingchao Liu, Chengyue Gong, and Qiang Liu. 'Flow straight and fast: Learning to generate and transfer data with rectified flow'. In: *arXiv preprint arXiv:2209.03003* (2022) (cited on page 34).

[80] Kirill Neklyudov, Daniel Severo, and Alireza Makhzani. 'Action Matching: A Variational Method for Learning Stochastic Dynamics from Samples'. In: *arXiv preprint arXiv:2210.06662* (2022) (cited on page 34).

[81] Aram Davtyan, Sepehr Sameni, and Paolo Favaro. 'Randomized Conditional Flow Matching for Video Prediction'. In: *arXiv preprint arXiv:2211.14575* (2022) (cited on page 34).

[82] Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. 'Conditional flow matching: Simulation-free dynamic optimal transport'. In: *arXiv preprint arXiv:2302.00482* (2023) (cited on page 34).

[83] Scott A Sisson, Yanan Fan, and Mark A Beaumont. 'Overview of ABC'. In: *Handbook of approximate Bayesian computation*. Chapman and Hall/CRC, 2018, pp. 3–54 (cited on page 35).

[84] Mark A Beaumont, Wenyang Zhang, and David J Balding. 'Approximate Bayesian computation in population genetics'. In: *Genetics* 162.4 (2002), pp. 2025–2035 (cited on page 35).

[85] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. 'Adaptive approximate Bayesian computation'. In: *Biometrika* 96.4 (2009), pp. 983–990 (cited on page 35).

[86] Michael G. B. Blum and Olivier François. 'Non-linear regression models for Approximate Bayesian Computation'. In: *Stat. Comput.* 20.1 (2010), pp. 63–73. DOI: 10.1007/s11222-009-9116-0 (cited on page 35).

[87] Dennis Prangle, Paul Fearnhead, Murray P. Cox, Patrick J. Biggs, and Nigel P. French. *Semi-automatic selection of summary statistics for ABC model choice*. 2013 (cited on page 35).

[88] Simon Wood. 'Statistical inference for noisy nonlinear ecological dynamic systems'. In: *Nature* 466 (Aug. 2010), pp. 1102–4. DOI: 10.1038/nature09319 (cited on page 35).

[89] Christopher C Drovandi, Clara Grazian, Kerrie Mengersen, and Christian Robert. *Approximating the Likelihood in Approximate Bayesian Computation*. 2018 (cited on page 35).

[90] George Papamakarios, David Sterratt, and Iain Murray. 'Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows'. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 837–848 (cited on page 35).

[91] Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H Macke. 'Likelihood-free inference with emulator networks'. In: *Symposium on Advances in Approximate Bayesian Inference*. PMLR. 2019, pp. 32–53 (cited on page 35).

[92] Rafael Izbicki, Ann Lee, and Chad Schafer. 'High-dimensional density ratio estimation with extensions to approximate likelihood computation'. In: *Artificial intelligence and statistics*. PMLR. 2014, pp. 420–429 (cited on page 35).

[93] Kim Pham, David Nott, and Sanjay Chaudhuri. 'A note on approximating ABC-MCMC using flexible classifiers'. In: *Stat* 3 (Mar. 2014). DOI: 10.1002/sta4.56 (cited on page 35).

[94] Kyle Cranmer, Juan Pavez, and Gilles Louppe. 'Approximating likelihood ratios with calibrated discriminative classifiers'. In: *arXiv preprint arXiv:1506.02169* (2015) (cited on page 35).

[95] Joeri Hermans, Volodimir Begy, and Gilles Louppe. 'Likelihood-free MCMC with approximate likelihood ratios'. In: *arXiv preprint arXiv:1903.04057* 10 (2019) (cited on page 35).

[96] Conor Durkan, Iain Murray, and George Papamakarios. 'On contrastive learning for likelihood-free inference'. In: *International conference on machine learning*. PMLR. 2020, pp. 2771–2781 (cited on page 35).

[97] Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U. Gutmann. *Likelihood-free inference by ratio estimation*. 2020 (cited on page 35).

[98] Benjamin K Miller, Christoph Weniger, and Patrick Forré. 'Contrastive neural ratio estimation'. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 3262–3278 (cited on page 35).

[99] Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. 'Neural importance sampling'. In: *ACM Transactions on Graphics (TOG)* 38.5 (2019), pp. 1–19 (cited on page 35).

[100] Maximilian Dax, Stephen R. Green, Jonathan Gair, Michael Pürrer, Jonas Wildberger, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. 'Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference'. In: *Phys. Rev. Lett.* 130.17 (2023), p. 171403. DOI: 10.1103/PhysRevLett.130.171403 (cited on pages 35, 40–42, 47, 55, 56, 60, 86, 87).

[101] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. 'Stochastic interpolants: A unifying framework for flows and diffusions'. In: *arXiv preprint arXiv:2303.08797* (2023) (cited on pages 36, 37, 93, 94).

[102] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. 'Maximum likelihood training of score-based diffusion models'. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 1415–1428 (cited on pages 37, 93, 94).

[103] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 'U-net: Convolutional networks for biomedical image segmentation'. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241 (cited on page 37).

[104] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 'Language modeling with gated convolutional networks'. In: *International conference on machine learning*. PMLR. 2017, pp. 933–941 (cited on page 37).

[105] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. 'Benchmarking simulation-based inference'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 343–351 (cited on pages 38, 39, 98).

[106] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*. 2015 (cited on pages 38, 41).

[107] Jerome H Friedman. 'On multivariate goodness–of–fit and two–sample testing'. In: *Statistical Problems in Particle Physics, Astrophysics, and Cosmology* 1 (2003), p. 311 (cited on page 38).

[108] David Lopez-Paz and Maxime Oquab. 'Revisiting classifier two-sample tests'. In: *arXiv preprint arXiv:1610.06545* (2016) (cited on page 38).

[109] Arnaud Delaunoy, Joeri Hermans, François Rozet, Antoine Wehenkel, and Gilles Louppe. *Towards Reliable Simulation-Based Inference with Balanced Neural Ratio Estimation*. 2022 (cited on page 39).

[110] Arnaud Delaunoy, Benjamin Kurt Miller, Patrick Forré, Christoph Weniger, and Gilles Louppe. 'Balancing Simulation-based Inference for Conservative Posteriors'. In: *arXiv preprint arXiv:2304.10978* (2023) (cited on page 39).

[111] R. Abbott et al. 'GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run'. In: *arXiv preprint arXiv:2111.03606* (Nov. 2021) (cited on page 40).

[112] R. Abbott et al. 'Population Properties of Compact Objects from the Second LIGO-Virgo Gravitational-Wave Transient Catalog'. In: *Astrophys. J. Lett.* 913.1 (2021), p. L7. DOI: 10.3847/2041-8213/abe949 (cited on page 40).

[113] B. P. Abbott et al. 'GW170817: Measurements of neutron star radii and equation of state'. In: *Phys. Rev. Lett.* 121.16 (2018), p. 161101. DOI: 10.1103/PhysRevLett.121.161101 (cited on page 40).

[114] R. Abbott et al. 'Tests of general relativity with binary black holes from the second LIGO-Virgo gravitational-wave transient catalog'. In: *Phys. Rev. D* 103.12 (2021), p. 122002. DOI: 10.1103/PhysRevD.103.122002 (cited on page 40).

[115] B. P. Abbott et al. 'A gravitational-wave standard siren measurement of the Hubble constant'. In: *Nature* 551.7678 (2017), pp. 85–88. DOI: 10.1038/nature24471 (cited on pages 40, 71).

[116] I. M. Romero-Shaw et al. 'Bayesian inference for compact binary coalescences with bilby: validation and application to the first LIGO–Virgo gravitational-wave transient catalogue'. In: *Mon. Not. Roy. Astron. Soc.* 499.3 (2020), pp. 3295–3319. DOI: 10.1093/mnras/staa2850 (cited on pages 40, 46).

[117] Joshua S Speagle. 'dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences'. In: *Monthly Notices of the Royal Astronomical Society* 493.3 (Feb. 2020), pp. 3132–3158. DOI: 10.1093/mnras/staa278 (cited on pages 40, 46, 86).

[118] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 'Equation of state calculations by fast computing machines'. In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092 (cited on page 40).

[119] W. K. Hastings. 'Monte Carlo sampling methods using Markov chains and their applications'. In: *Biometrika* 57.1 (Apr. 1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97 (cited on page 40).

[120] John Skilling. 'Nested sampling for general Bayesian computation'. In: *Bayesian Analysis* 1.4 (2006), pp. 833–859. DOI: 10.1214/06-BA127 (cited on page 40).

[121] Elena Cuoco, Jade Powell, Marco Cavaglià, Kendall Ackley, Michał Bejger, Chayan Chatterjee, Michael Coughlin, Scott Coughlin, Paul Easter, Reed Essick, et al. 'Enhancing Gravitational-Wave Science with Machine Learning'. In: *Machine Learning: Science and Technology* 2.1 (May 2020), p. 011002. DOI: 10.1088/2632-2153/abb93a (cited on pages 40, 41).

[122] Hunter Gabbard, Chris Messenger, Ik Siong Heng, Francesco Tonolini, and Roderick Murray-Smith. 'Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy'. In: *Nature Phys.* 18.1 (2022), pp. 112–117. DOI: 10.1038/s41567-021-01425-7 (cited on pages 40, 41, 46, 49).

[123] Stephen R. Green, Christine Simpson, and Jonathan Gair. 'Gravitational-wave parameter estimation with autoregressive neural network flows'. In: *Phys. Rev. D* 102.10 (2020), p. 104057. DOI: 10.1103/PhysRevD.102.104057 (cited on pages 40, 41, 46, 49).

[124] Arnaud Delaunoy, Antoine Wehenkel, Tanja Hinderer, Samaya Nissanke, Christoph Weniger, Andrew R. Williamson, and Gilles Louppe. 'Lightning-Fast Gravitational Wave Parameter Inference through Neural Amortization'. In: *Third Workshop on Machine Learning and the Physical Sciences*. Oct. 2020 (cited on pages 40, 41, 46, 49).

[125] Maximilian Dax, Stephen R. Green, Jonathan Gair, Michael Deistler, Bernhard Schölkopf, and Jakob H. Macke. 'Group equivariant neural posterior estimation'. In: *International Conference on Learning Representations*. Nov. 2022 (cited on pages 40–42, 46, 49).

[126] Chayan Chatterjee, Linqing Wen, Damon Beveridge, Foivos Diakogiannis, and Kevin Vinsen. 'Rapid localization of gravitational wave sources from compact binary coalescences using deep learning'. In: *arXiv preprint arXiv:2207.14522* (July 2022) (cited on pages 40, 41).

[127] Stefan T. Radev, Ulf K. Mertens, Andreass Voss, Lynton Ardizzone, and Ullrich Köthe. *BayesFlow: Learning complex stochastic models with invertible neural networks*. 2020 (cited on page 41).

[128] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. 'Neural Spline Flows'. In: *Advances in Neural Information Processing Systems*. 2019, pp. 7509–7520 (cited on page 41).

[129] Jonas Wildberger, Maximilian Dax, Stephen R. Green, Jonathan Gair, Michael Pürrer, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. 'Adapting to noise distribution shifts in flow-based gravitational-wave inference'. In: *Phys. Rev. D* 107.8 (2023), p. 084046. DOI: 10.1103/PhysRevD.107.084046 (cited on page 42).

[130] Taco Cohen and Max Welling. 'Group Equivariant Convolutional Networks'. In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Ed. by Maria-Florina Balcan and Kilian Q. Weinberger. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 2990–2999 (cited on page 42).

[131] Kentaro Somiya. 'Detector configuration of KAGRA: The Japanese cryogenic gravitational-wave detector'. In: *Class. Quant. Grav.* 29 (2012). Ed. by Mark Hannam, Patrick Sutton, Stefan Hild, and Chris van den Broeck, p. 124007. DOI: 10.1088/0264-9381/29/12/124007 (cited on page 46).

[132] Yoichi Aso, Yuta Michimura, Kentaro Somiya, Masaki Ando, Osamu Miyakawa, Takanori Sekiguchi, Daisuke Tatsumi, and Hiroaki Yamamoto. 'Interferometer design of the KAGRA gravitational wave detector'. In: *Phys. Rev. D* 88.4 (2013), p. 043007. DOI: 10.1103/PhysRevD.88.043007 (cited on page 46).

[133] Peter Welch. 'The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms'. In: *IEEE Transactions on audio and electroacoustics* 15.2 (1967), pp. 70–73 (cited on page 46).

[134] Neil J. Cornish and Tyson B. Littenberg. 'BayesWave: Bayesian Inference for Gravitational Wave Bursts and Instrument Glitches'. In: *Class. Quant. Grav.* 32.13 (2015), p. 135012. DOI: 10.1088/0264-9381/32/13/135012 (cited on pages 46, 60).

[135] Neil J. Cornish, Tyson B. Littenberg, Bence Bécsy, Katerina Chatziioannou, James A. Clark, Sudarshan Ghonge, and Margaret Millhouse. 'BayesWave analysis pipeline in the era of gravitational wave observations'. In: *Phys. Rev. D* 103.4 (2021), p. 044006. DOI: 10.1103/PhysRevD.103.044006 (cited on page 46).

[136] Alvin J. K. Chua and Michele Vallisneri. 'Learning Bayesian posteriors with neural networks for gravitational-wave inference'. In: *Phys. Rev. Lett.* 124.4 (2020), p. 041102. DOI: 10.1103/PhysRevLett.124.041102 (cited on pages 46, 49).

[137] Chayan Chatterjee, Linqing Wen, Kevin Vinsen, Manoj Kovalam, and Amitava Datta. 'Using Deep Learning to Localize Gravitational Wave Sources'. In: *Phys. Rev. D* 100.10 (2019), p. 103025. DOI: 10.1103/PhysRevD.100.103025 (cited on pages 46, 49).

[138] Plamen G. Krastev, Kiranjyot Gill, V. Ashley Villar, and Edo Berger. 'Detection and Parameter Estimation of Gravitational Waves from Binary Neutron-Star Mergers in Real LIGO Data using Deep Learning'. In: *Phys. Lett. B* 815 (2021), p. 136161. DOI: 10.1016/j.physletb.2021.136161 (cited on pages 46, 49).

[139] Hongyu Shen, E. A. Huerta, Eamonn O'Shea, Prayush Kumar, and Zhizhen Zhao. 'Statistically-informed deep learning for gravitational wave parameter estimation'. In: (Mar. 2021) (cited on pages 46, 49).

[140] Rich Abbott et al. 'Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo'. In: *SoftwareX* 13 (2021), p. 100658. DOI: 10.1016/j.softx.2021.100658 (cited on pages 48, 49, 56).

[141] David J Bartholomew, Martin Knott, and Irini Moustaki. *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons, 2011 (cited on pages 48, 49).

[142] Diederik P. Kingma and Max Welling. 'Auto-Encoding Variational Bayes'. In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. 2014 (cited on pages 48, 49).

[143] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 'Stochastic Backpropagation and Approximate Inference in Deep Generative Models'. In: *International Conference on Machine Learning*. 2014, pp. 1278–1286 (cited on pages 48, 49).

[144] Tyson B. Littenberg and Neil J. Cornish. 'Bayesian inference for spectral estimation of gravitational wave detector noise'. In: *Phys. Rev. D* 91 (8 Apr. 2015), p. 084034. DOI: 10.1103/PhysRevD.91.084034 (cited on pages 49, 51, 52, 60).

[145] Sebastian Khan, Sascha Husa, Mark Hannam, Frank Ohme, Michael Pürrer, Xisco Jiménez Forteza, and Alejandro Bohé. 'Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era'. In: *Phys. Rev.* D93.4 (2016), p. 044007. DOI: 10.1103/PhysRevD.93.044007 (cited on pages 53, 102).

[146] Mark Hannam, Patricia Schmidt, Alejandro Bohé, Leïla Haegel, Sascha Husa, Frank Ohme, Geraint Pratten, and Michael Pürrer. 'Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms'. In: *Phys. Rev. Lett.* 113 (15 Oct. 2014), p. 151101. DOI: 10.1103/PhysRevLett.113.151101 (cited on pages 53, 102).

[147] Alejandro Bohé et al. 'Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors'. In: *Phys. Rev. D* 95.4 (2017), p. 044028. DOI: 10.1103/PhysRevD.95.044028 (cited on page 53).

[148] John Veitch and Walter Del Pozzo. 'Analytic marginalisation of phase parameter'. In: *URL: https://dcc. ligo. org/LIGO-T1300326/public* (2013) (cited on page 55).

[149] Eric Thrane and Colm Talbot. 'An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models'. In: *Publ. Astron. Soc. Austral.* 36 (2019). [Erratum: Publ.Astron.Soc.Austral. 37, e036 (2020)], e010. DOI: 10.1017/pasa.2019.2 (cited on page 55).

[150] J. Lin. 'Divergence measures based on the Shannon entropy'. In: *IEEE Trans. Info. Theor.* 37.1 (1991), pp. 145–151. DOI: 10.1109/18.61115 (cited on page 55).

[151] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 'Scaling Laws for Neural Language Models'. In: *CoRR* abs/2001.08361 (2020) (cited on page 64).

[152] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 'Training Compute-Optimal Large Language Models'. In: *CoRR* abs/2203.15556 (2022). DOI: 10.48550/ARXIV.2203.15556 (cited on page 64).

[153] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 'Language Models are Few-Shot Learners'. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. 2020 (cited on page 64).

[154] Rishi Bommasani et al. 'On the Opportunities and Risks of Foundation Models'. In: *ArXiv* (2021) (cited on page 65).

[155] Maximilian Dax, Stephen R. Green, Jonathan Gair, Michael Pürrer, Jonas Wildberger, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. 'Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference'. In: *Phys. Rev. Lett.* 130.17 (2023), p. 171403. DOI: 10.1103/PhysRevLett.130.171403 (cited on page 69).

[156] Gebhard, Timothy D., Wildberger, Jonas, Dax, Maximilian, Kofler, Annalena, Angerhausen, Daniel, Quanz, Sascha P., and Schölkopf, Bernhard. 'Flow matching for atmospheric retrieval of exoplanets: Where reliability meets adaptive noise levels'. In: *A&A* 693 (2025), A42. DOI: 10.1051/0004-6361/202451861 (cited on page 70).

[157] Frederike Lübeck, Jonas Bernhard Wildberger, Frederik Träuble, Maximilian Mordig, Sergios Gatidis, Andreas Krause, and Bernhard Schölkopf. 'From Structured Data to Clinical Notes: Robust Clinical Decision Support with Fine-Tuned LLMs'. In: *1st ICML Workshop on Foundation Models for Structured Data*. 2025 (cited on page 71).

[158] Maximilian Dax, Stephen R. Green, Jonathan Gair, Nihar Gupte, Michael Pürrer, Vivien Raymond, Jonas Wildberger, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. 'Real-time inference for binary neutron star mergers using machine learning'. In: *Nature* 639.8053 (2025), pp. 49–53. DOI: 10.1038/s41586-025-08593-z (cited on page 71).

[159] B. P. Abbott et al. 'GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral'. In: *Phys. Rev. Lett.* 119.16 (2017), p. 161101. DOI: 10.1103/PhysRevLett.119.161101 (cited on page 71).

[160] B. P. Abbott et al. 'Multi-messenger Observations of a Binary Neutron Star Merger'. In: *Astrophys. J. Lett.* 848.2 (2017), p. L13. DOI: 10.3847/2041-8213/aa920c (cited on page 71).

[161] B. P. Abbott et al. 'Gravitational Waves and Gamma-rays from a Binary Neutron Star Merger: GW170817 and GRB 170817A'. In: *Astrophys. J. Lett.* 848.2 (2017), p. L13. DOI: 10.3847/2041-8213/aa920c (cited on page 71).

[162] B. P. Abbott et al. 'Properties of the binary neutron star merger GW170817'. In: *Phys. Rev.* X9.1 (2019), p. 011001. DOI: 10.1103/PhysRevX.9.011001 (cited on page 71).

[163] B. P. Abbott et al. 'GW170817: Measurements of neutron star radii and equation of state'. In: *Phys. Rev. Lett.* 121.16 (2018), p. 161101. DOI: 10.1103/PhysRevLett.121.161101 (cited on page 71).

[164] B. P. Abbott et al. 'Tests of General Relativity with GW170817'. In: *Phys. Rev. Lett.* 123.1 (2019), p. 011102. DOI: 10.1103/PhysRevLett.123.011102 (cited on page 71).

[165] D. A. Coulter et al. 'Swope Supernova Survey 2017a (SSS17a), the Optical Counterpart to a Gravitational Wave Source'. In: *Science* 358 (2017), p. 1556. DOI: 10.1126/science.aap9811 (cited on page 71).

[166] Siyuan Guo, Jonas Bernhard Wildberger, and Bernhard Schölkopf. 'Out-of-Variable Generalisation for Discriminative Models'. In: *The Twelfth International Conference on Learning Representations*. 2024 (cited on page 72).

[167] Nihar Gupte, Antoni Ramos-Buades, Alessandra Buonanno, Jonathan Gair, M. Coleman Miller, Maximilian Dax, Stephen R. Green, Michael Pürrer, Jonas Wildberger, Jakob Macke, Isobel M. Romero-Shaw, and Bernhard Schölkopf. 'Evidence for eccentricity in the population of binary black holes observed by LIGO-Virgo-KAGRA'. In: *Phys. Rev. D* 112.10 (2025), p. 104045. DOI: 10.1103/PhysRevD.112.104045 (cited on page 73).

[168] Annalena Kofler, Maximilian Dax, Stephen R. Green, Jonas Wildberger, Nihar Gupte, Jakob H. Macke, Jonathan Gair, Alessandra Buonanno, and Bernhard Schölkopf. 'Flexible Gravitational-Wave Parameter Estimation with Transformers'. In: (Dec. 2025) (cited on page 74).

[169] I. M. Romero-Shaw, C. Talbot, S. Biscoveanu, V. D'Emilio, G. Ashton, C. P. L. Berry, S. Coughlin, S. Galaudage, C. Hoy, M. Hübner, K. S. Phukon, M. Pitkin, M. Rizzo, N. Sarin, R. Smith, S. Stevenson, A. Vajpeyi, M. Arène, K. Athar, S. Banagiri, N. Bose, M. Carney, K. Chatziioannou, J. A. Clark, M. Colleoni, R. Cotesta, B. Edelman, H. Estellés, C. García-Quirós, Abhirup Ghosh, R. Green, C. -J. Haster, S. Husa, D. Keitel, A. X. Kim, F. Hernandez-Vivanco, I. Magaña Hernandez, C. Karathanasis, P. D. Lasky, N. De Lillo, M. E. Lower, D. Macleod, M. Mateu-Lucena, A. Miller, M. Millhouse, S. Morisaki, S. H. Oh, S. Ossokine, E. Payne, J. Powell, G. Pratten, M. Pürrer, A. Ramos-Buades, V. Raymond, E. Thrane, J. Veitch, D. Williams, M. J. Williams, and L. Xiao. 'Bayesian inference for compact binary coalescences with BILBY: validation and application to the first LIGO-Virgo gravitational-wave transient catalogue'. In: *MNRAS* 499.3 (Dec. 2020), pp. 3295–3319. DOI: 10.1093/mnras/staa2850 (cited on page 86).

[170] Alejandro Bohé, Mark Hannam, Sascha Husa, Frank Ohme, Michael Pürrer, and Patricia Schmidt. 'PhenomPv2 – technical notes for the LAL implementation'. In: *LIGO Technical Document, LIGO-T1500602-v4* (2016) (cited on page 102).

[171] Benjamin Farr, Evan Ochsner, Will M. Farr, and Richard O'Shaughnessy. 'A more effective coordinate system for parameter estimation of precessing compact binaries from gravitational waves'. In: *Phys. Rev. D* 90.2 (2014), p. 024018. DOI: 10.1103/PhysRevD.90.024018 (cited on page 104).

[172] J.R. Dormand and P.J. Prince. 'A family of embedded Runge-Kutta formulae'. In: *Journal of Computational and Applied Mathematics* 6.1 (1980), pp. 19–26. DOI: https://doi.org/10.1016/0771-050X(80)90013-3 (cited on page 104).