

Amortized Bayesian Multilevel Models

Daniel Habermann^{*}, Marvin Schmitt[†], Lars Kühmichel[‡], Andreas Bulling[§],
Stefan T. Radev[¶], and Paul-Christian Bürkner^{||}

Abstract. Multilevel models (MLMs) are a central building block of the Bayesian workflow. They enable joint, interpretable modeling of data across hierarchical levels and provide a fully probabilistic quantification of uncertainty. Despite their well-recognized advantages, MLMs pose significant computational challenges, often rendering their estimation and evaluation intractable within reasonable time constraints. Recent advances in simulation-based inference offer promising solutions for addressing complex probabilistic models using deep generative networks. However, the utility and reliability of deep learning methods for estimating Bayesian MLMs remain largely unexplored, especially when compared with gold-standard samplers. To this end, we explore a family of neural network architectures that leverage the probabilistic factorization of multilevel models to facilitate efficient neural network training and subsequent near-instant posterior inference on unseen datasets. We test our method on several real-world case studies and provide comprehensive comparisons to Stan’s gold standard sampler, where possible. Finally, we provide an open-source implementation of our methods to stimulate further research in the nascent field of amortized Bayesian inference.

Keywords: Bayesian models, amortized inference, multilevel models.

1 Introduction

Accurate inference and reliable uncertainty quantification in reasonable time is a frontier of today’s statistical research (Cranmer et al., 2020). One major difficulty arising in almost all experimental and observational data is the presence of complex dependency structures, such as natural groupings (e.g., data gathered in different countries) or repeated measurements of the same observational units over time (e.g., patients in clinical trials or students in educational studies; Gelman and Hill, 2006). To reflect these dependency structures, multilevel models (MLMs), also referred to as latent variable, hierarchical, random, or mixed effects models, have become an integral part of modern Bayesian statistics (e.g., Goldstein, 2011). Despite the wide success of Bayesian MLMs across the quantitative sciences, a major challenge is their limited scalability, as estimating the full posterior distribution of all parameters of interest can be very costly Betancourt (2017). For models where the likelihood function is analytically tractable and differentiable with respect to model parameters, Markov chain Monte Carlo (MCMC)

arXiv: 2408.13230

^{*}Department of Statistics, TU Dortmund University, Germany, daniel.habermann@tu-dortmund.de

[†]Cluster of Excellence SimTech, University of Stuttgart, Germany

[‡]Department of Statistics, TU Dortmund University, Germany

[§]Institute for Visualisation and Interactive Systems, University of Stuttgart, Germany

[¶]Department of Cognitive Science, Rensselaer Polytechnic Institute, USA

^{||}Department of Statistics, TU Dortmund University, Germany, paul.buerkner@gmail.com

sampling algorithms as implemented in probabilistic programming languages like Stan (Stan Development Team, 2024) are the current gold standard for generating accurate draws from the posterior distribution. Recently, there has been a growing body of research that is concerned with using pairs of samples from the prior distribution and simulated datasets to train neural networks that can approximate the posterior for unseen data (e.g., Wildberger et al., 2024). The prospect of these neural posterior estimation (NPE) methods is that they can enable *amortized* inference (Gershman and Goodman, 2014), which refers to the property that posterior inference for new data is almost instant after training of the neural networks has been completed (Radev et al., 2020b; Tejero-Cantero et al., 2020).

Expanding neural density estimation to MLMs is challenging because the dimension of the posterior depends on the number of groups and may vary over datasets. Such settings require specialized neural architectures to lift the requirement of fixed-length inputs. In MLMs, amortized inference with varying group sizes is complicated by the fact that estimates of local parameters for one group depend not only on that specific group but also on all other groups in the dataset. Additionally, the number of parameters increases linearly with the number of groups, rendering network training inefficient for datasets with a large number of groups.

In important pioneering work, initial progress has been made in addressing these challenges (Rodrigues et al., 2021; Arruda et al., 2023; Heinrich et al., 2023), but they have so far considered only a narrow range of MLMs and have performed only limited validation against state-of-the-art samplers for Bayesian inference. In particular, existing methods do not consider (1) simultaneous amortization over both the number of groups and the number of observations per group, (2) amortization in models with covariates, (3) simultaneous joint estimation with non-independent and identically distributed (non-IID) response variables, (4) highly correlated group-level parameters, (5) additional (non-hierarchical) shared parameters, as well as (6) systematic calibration and shrinkage analysis to verify inference correctness at each level. Our proposed framework addresses these challenges by making the following contributions:

- We develop neural network architectures that utilize the probabilistic factorization of the likelihood of multilevel models to facilitate efficient neural network training and subsequent near-instant amortized posterior inference.
- We test our method on a wide range of real-world case studies which together cover all of the above mentioned challenges. We also provide comparisons to Stan as a gold-standard method for posterior inference whenever possible.
- We provide an efficient and user-friendly implementation of our algorithm in the `BayesFlow` Python library (Radev et al., 2023c). In this way, users can benefit from amortized inference for otherwise intractable models and generate fast and accurate draws from the posterior distribution.

We start by briefly discussing the limitations of MCMC-based sampling methods and current research avenues (Subsection 1.1). We then follow by introducing neural poste-

rior estimation, highlighting possible advantages over traditional MCMC-based methods as well as current shortcomings (Subsection 2.2). Our novel contributions start from Subsection 2.3 onward, where we extend NPE to multilevel models and address challenges specific to hierarchical data. For readers without a background in machine learning, we provide an overview of the current landscape and introduction to “simulation-based inference”, “amortized Bayesian inference”, and “neural density estimation” in the online supplement (Habermann et al., 2025). Section 3 evaluates multilevel NPE through a variety of experiments, focusing on model classes that were previously not amenable to neural posterior estimation.

1.1 Challenges of Posterior Inference in Bayesian Multilevel Models

Considerable effort has been made to improve the sampling speed of MCMC algorithms, especially for high-dimensional models with rich structure, such as MLMs. This includes step-size adjustments or reparameterizations to work-around or improve the posterior geometry (e.g., Hoffman et al., 2019), enabling within-chain parallelization via parallel evaluation of the likelihood (Lee et al., 2010, e.g.), as well as the development of specialized algorithms that entail faster and more reliable adaptation phases, such that multiple short chains can be run in parallel (e.g., Zhang et al., 2022; Margossian et al., 2024). However, despite these advancements, sampling methods based on MCMC are still too slow for many multilevel settings. Even for models and datasets of moderate size, posterior inference can take days or even weeks, creating a large gap between the models that researchers aim to compute (which may include more complex structures) and the simplified models that are computationally feasible with current methods. Expanding the space of models that can be fit in a reasonable time frame is therefore an important requirement for the broad applicability of Bayesian MLMs.

To address these challenges, approximate inference methods such as variational inference (e.g., Hoffman and Blei, 2015) or integrated Laplace approximation (INLA, e.g., Margossian et al., 2020) have been developed. Such algorithms trade off speed against posterior accuracy or modeling flexibility (e.g., Yao et al., 2018; Margossian and Saul, 2023). In cases where model parameters and their uncertainty are of interest, and not only model predictions, the loss of posterior accuracy is particularly problematic, as it may invalidate the conclusions drawn from the model’s inference – even if the model itself is reasonable.

In addition to sampling speed, there are two fundamental issues with posterior inference for Bayesian MLMs in most applications of scientific interest:

1. The need to rerun computationally expensive sampling algorithms when new data becomes available or when the quality of model fits needs to be assessed.
2. Generating accurate draws from the posterior when the likelihood is intractable or non-differentiable.

Issue (1) is a crucial limiting factor in cases where data are arriving in real-time or many datasets need to be evaluated. In addition, essential steps in the modern

Bayesian workflow (Gelman et al., 2020) like cross-validation (e.g., Vehtari et al., 2017) or simulation-based calibration (SBC, e.g., Talts et al., 2018; Modrák et al., 2023) require many model refits on subsets of the data, rendering careful model testing a time-intensive task for all but the simplest models. Issue (2) is also becoming increasingly common, as many scientific domains nowadays employ models whose output is the result of a complex simulation program (Cranmer et al., 2020). Consequently, closed-form solutions of the likelihood are often unavailable.

Both of these issues have led to the development of novel algorithms that do not rely on MCMC sampling. For models without a tractable likelihood, simulation-based inference (SBI) methods such as approximate Bayesian computation (ABC, Sisson et al., 2018, e.g.,) can sample from the posterior without computing the likelihood directly. The main drawback of ABC compared to neural network based approaches is that it heavily suffers from the curse of dimensionality (e.g., Barber et al., 2015; Beneventano et al., 2020) as the number of parameters increases. This is of particular concern for MLMs as their number of model parameters scales linearly with the number of groups in the data. While ABC algorithms for MLMs have been developed (Sisson and Fan, 2011), they have not been able to fully overcome this major drawback of ABC. Other approaches to Bayesian inference for MLMs with intractable likelihoods include mean-field variational algorithms (e.g., Tran et al., 2017), but this class of algorithms tends to come with the aforementioned loss of posterior accuracy (e.g., Yao et al., 2018).

2 Methods

2.1 Notation and Definitions

We denote the joint distribution over both the parameters θ and observed data x of a parametric Bayesian model as $p(\theta, x)$. We assume that $\theta \in \mathbb{R}^D$ represents all unobservable (latent) quantities, whereas $x \in \mathcal{X}$ denotes all observable quantities. The joint distribution $p(\theta, x)$ factorizes into a data model $p(x|\theta)$ and a prior $p(\theta)$, defining a “generative model” that can be translated into a computer program and sampled from according to a simple ancestral sampling scheme:

$$\begin{aligned}\theta &\sim p(\theta) \\ x &\sim p(x|\theta)\end{aligned}\tag{1}$$

Multilevel models can be expressed as a special case of this general notation, where the prior $p(\theta)$ is factorized in an application-dependent manner. For example, in a two-level model with exchangeable groups j and exchangeable observations within groups i , there are three qualitatively different kinds of parameters: (1) *Local parameters* λ_j indexed by $j \in \{1, \dots, J\}$ that are specific to each group j , such as separate intercepts or slopes per group in a regression model; (2) *hyperparameters* τ , such as means and standard deviations for the local parameters, or correlations of local parameters in the same group; and (3) *shared parameters* ω with a direct effect on the likelihood function of all groups, without any hierarchical component (see Figure 1 for a diagram of the dependency structure, left panel). We use the term *global parameters* to refer to the hyperparameters τ and shared parameters ω jointly.

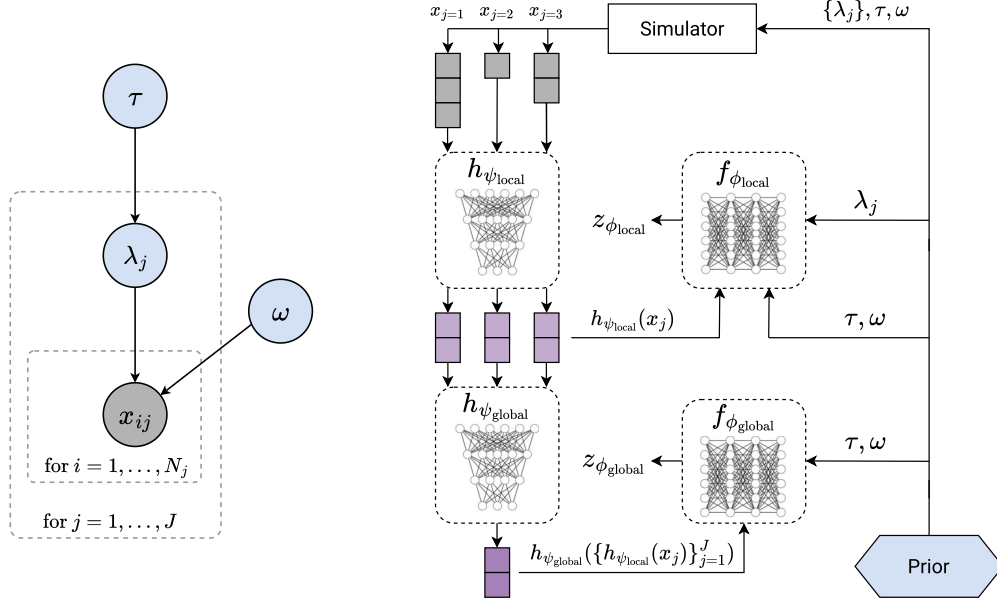


Figure 1: **Left:** Directed graph of a two-level Bayesian model with hyperparameters τ , local parameters λ_j , shared parameters ω and observed data x_{ij} . Gray-shaded circles indicate observed values, whereas non-shaded circles indicate latent variables. Dashed boxes represent exchangeability. **Right:** Summary of the multilevel NPE architecture. Simulated datasets with varying group sizes (grey stacked boxes) are transformed into fixed-length representations by a local summary network $h_{\psi_{\text{local}}}$ (purple squares). A global summary network $h_{\psi_{\text{global}}}$ then converts these local summaries into a fixed-length dataset representation. The local inference network $f_{\phi_{\text{local}}}$ maps prior draws of λ_j to a base distribution $z_{\phi_{\text{local}}}$, conditioned on local summaries and τ and ω . The global inference network $f_{\phi_{\text{global}}}$ similarly maps prior draws of τ, ω to a base distribution $z_{\phi_{\text{global}}}$, conditioned on the global dataset summary.

Thus, in a two-level setting, the full set of parameters is defined as $\theta := \{\{\lambda_j\}_{j=1}^J, \tau, \omega\}$. Consequently, the prior $p(\theta)$ factorizes as:

$$p(\theta) = p(\tau, \omega, \{\lambda_j\}_{j=1}^J) = p(\tau) p(\omega) \prod_{j=1}^J p(\lambda_j | \tau) \quad (2)$$

Let x_{ij} denote the observation with index $i \in 1, \dots, N_j$ within group j . The generative model can then be described as (see Figure 1):

$$\begin{aligned} \tau, \omega &\sim p(\tau)p(\omega) \\ \lambda_j | \tau &\sim p(\lambda | \tau) \quad \text{for } j = 1, \dots, J \\ x_{ij} &\sim p(x | \lambda_j, \omega) \quad \text{for } i = 1, \dots, N_j \end{aligned} \quad (3)$$

To fit these multilevel models in practice, we first introduce the concept of neural posterior estimation in Subsection 2.2, before expanding it to multilevel NPE in Subsection 2.3.

2.2 Neural Posterior Estimation

Recently, algorithms based on neural posterior estimation (NPE, e.g., Greenberg et al., 2019; Radev et al., 2020a) have emerged as a promising alternative to classical MCMC. In NPE, generating draws from the posterior distribution of a Bayesian model is achieved by transforming a random input vector, e.g. sampled from a unit Gaussian, into draws from the target posterior distribution using specialized neural architectures. Below, we focus on NPE with normalizing flows (e.g., Kobyzev et al., 2020) as a concrete example, although other generative neural architectures can be used, such as diffusion models (e.g., Sharrock et al., 2022), consistency models (e.g., Schmitt et al., 2024b), or flow matching (e.g., Wildberger et al., 2024).

A normalizing flow is an invertible transformation between the target (posterior) distribution and a tractable base distribution, such as a unit Gaussian. Let $z \in \mathbb{R}^D$ be a D -dimensional random variable with tractable density, e.g. $p_0(z) = \text{Normal}(z | 0, \mathbb{I})$, and let $\theta \in \mathbb{R}^D$ be a random variable with unknown probability density equivalent to the posterior $p(\theta | x)$ over model parameters θ given observed data x . Let $f : \mathbb{R}^D \rightarrow \mathbb{R}^D$ be an invertible and differentiable function such that $z = f(\theta)$ and $\theta = f^{-1}(z)$. In practice, this function is usually parameterized in terms of a conditional invertible neural network f_ϕ with learnable network weights ϕ .

The *inference network* f_ϕ maps model parameters θ to latent Gaussian variates z using conditional information from the observed dataset x . Conversely, the inverse pass through the network maps z to θ using the same conditional information. Training data $\{(\theta^{(m)}, x^{(m)})\}_{m=1}^M$ for the neural network are generated by first sampling a single realization of model parameters $\theta^{(m)}$ from the prior $p(\theta)$ and then simulating a synthetic dataset $x^{(m)}$ of a specified size via the generative process (i.e., $x^{(m)} \sim p(x | \theta^{(m)})$). Because f_ϕ is differentiable, the density of the approximate posterior $q_\phi(\theta | x)$ can be computed via the change-of-variables formula:

$$q_\phi(\theta | x) = p_0(f_\phi(\theta; x)) \left| \det \frac{\partial f_\phi(\theta; x)}{\partial \theta} \right|, \quad (4)$$

where $f_\phi(\theta; x)$ denotes a forward pass of parameters θ through the neural network f_ϕ , conditional on a dataset x and $p_0(f_\phi(\theta; x))$ is the tractable density of the base distribution, e.g. with $p_0(z) = \text{Normal}(z | 0, \mathbb{I})$. Training of the network weights ϕ is achieved by optimizing

$$\mathcal{L}_{\text{NPE}}(\phi) = -\frac{1}{M} \sum_{m=1}^M \log q_\phi(\theta^{(m)} | x^{(m)}) \quad (5)$$

Minimizing this objective during training for sufficiently large simulation budget M results in minimizing the expected forward Kullback-Leibler (KL) divergence between

the approximate posterior $q_\phi(\theta | x)$ and the true posterior $p(\theta | x)$:

$$\begin{aligned}\hat{\phi} &= \arg \min_{\phi} \mathbb{E}_{x \sim p(x)} [\text{KL}(p(\theta | x) || q_\phi(\theta | x))] \\ &= \arg \min_{\phi} \mathbb{E}_{(\theta, x) \sim p(\theta, x)} [-\log q_\phi(\theta | x)]\end{aligned}\tag{6}$$

One particular advantage of NPE is that, after an initial training phase, it enables near-instantaneous posterior sampling for unseen datasets without any further training required (e.g., Radev et al., 2020a). This property is called *amortization* (Gershman and Goodman, 2014), as the initial computational cost during training is later offset by allowing posterior inference for new datasets typically within a fraction of a second. Amortization is a direct consequence of the fact that a single (global) set of neural network parameters ϕ is used to approximate the posterior distribution of many datasets in the support of $p(\theta, x)$.

Typically, the goal is to allow posterior inference for datasets with varying sample size $N \in \mathbb{N}$, which requires that the inference network is also trained with datasets of varying sample size. As the inference network can only handle fixed-length inputs, a second neural network, called *summary network* is often used in practice (e.g., Radev et al., 2020a). This summary network converts datasets with varying sample size into fixed-length representations, which are then used as conditions for the inference network. The summary network can also be beneficial for other reasons: If the data are high-dimensional, it may improve computational efficiency by using a lower-dimensional representation that is sufficient to inform posterior inference. Second, special neural architectures can be used to reflect the structure of the data. For example, permutation invariant networks can be used for exchangeable data to ensure that posterior inference does not depend on the order of individual data points.

Throughout this work, we denote summary networks by h_ψ , where ψ refers to the learnable network weights. The summary network h_ψ and inference network f_ϕ are optimized in one joint *end-to-end* optimization objective. Expanding (6) with a summary network yields:

$$\hat{\psi}, \hat{\phi} = \arg \min_{\psi, \phi} \mathbb{E}_{(\theta, x) \sim p(\theta, x)} [-\log q_\phi(\theta | h_\psi(x))],\tag{7}$$

which can also be trained via a simulation-based loss as shown in (5). The learned summary statistics are then *approximately sufficient* (Chen et al., 2020; Radev et al., 2020a) for posterior inference. The term *approximately sufficient* describes the fact that the learned summary statistics $h_\psi(x)$ are generally not sufficient to reconstruct the dataset x itself, but are sufficient in the sense that they do not alter the posterior distribution when swapped with the data: $p(\theta | x) \approx p(\theta | h_\psi(x))$.

2.3 Multilevel Neural Posterior Estimation

Standard NPE is not directly suitable for estimating multilevel models without further modifications. This limitation arises from the hierarchical structure of multilevel models,

where parameters are dependent across levels, and observations exhibit varying degrees of exchangeability that need to be accounted for (i.e., observations are exchangeable within groups but not across groups). In particular, the following additions are necessary to enable multilevel NPE:

- **Hierarchical summary networks.** Amortizing over both the number of groups and number of observations requires network training on datasets with varying group sizes and numbers of observations per group. Because observations are nested, hierarchical summary networks are necessary to provide both local information (i.e., characteristics of observations within groups) and global information (i.e., characteristics across groups).
- **Hierarchical inference networks.** In MLMs, the dimension of the joint posterior distribution depends on the number of groups and may also vary between datasets, which requires inference networks with varying number of outputs. This can be realized by hierarchical inference networks.
- **Different types of summary and inference networks.** Multilevel models feature different types of dependencies. For example, hierarchical time series models have non-exchangeable observations within groups, while other model classes may require exchangeability. Neural network architectures for estimating multilevel models therefore need to be flexible enough to account for these differences.

In what follows, we present a neural architecture that is based on NPE with learned summary statistics and enables fully amortized Bayesian inference for MLMs with both tractable and intractable likelihoods. We call this method *Multilevel Neural Posterior Estimation (ML-NPE)*.

Model Structure We consider a two-level hierarchical model as introduced in Subsection 2.1 with local parameters λ_j , hyperparameters τ and shared parameters ω . The observation with index $i \in 1, \dots, N_j$ within group $j \in 1, \dots, J$ is denoted by x_{ij} , and the data of group j is written as x_j . The entire dataset is represented by x with $x = \{x_j\}_{j=1}^J$. For visual clarity, we omit the indexing ranges for local parameters and denote the set of all local parameters $\{\lambda_j\}_{j=1}^J$ simply as $\{\lambda_j\}$.

Adopting a similar approach to Heinrich et al. (2023), which focuses on applications in the physical sciences, we extend amortized inference to multilevel models by addressing additional challenges such as varying group sizes and number of observations per group, covariates, and non-IID responses. We leverage the dependency structure implied by the probabilistic graph shown in Figure 1 (left panel). The joint distribution $p(\tau, \omega, \{\lambda_j\}, x)$ factorizes as:

$$p(\tau, \omega, \{\lambda_j\}, x) = p(\tau)p(\omega) \prod_{j=1}^J p(\lambda_j | \tau)p(x_j | \lambda_j, \omega) \quad (8)$$

This means that the joint posterior

$$p(\tau, \omega, \{\lambda_j\} | x) = \frac{p(\tau, \omega, \{\lambda_j\}, x)}{p(x)} \quad (9)$$

can be further factorized into two parts, namely (1) the hyperparameters τ together with the shared parameters ω conditional on all data; and (2) the local parameters λ_j conditional on the global parameters and the group-level data:

$$p(\tau, \omega, \{\lambda_j\} | x) = p(\tau, \omega | x) p(\{\lambda_j\} | \tau, \omega, x) \quad (10)$$

$$= p(\tau, \omega | x) \prod_{j=1}^J p(\lambda_j | \tau, \omega, x_j) \quad (11)$$

This factorization suggests a two-stage approach: First, the global parameters τ and ω are estimated conditionally on all observed data. Then, the local parameters λ_j are estimated conditionally independent of each other given τ , ω and the group-level data x_j . In this way, we divide the problem of learning the joint posterior into two smaller problems of learning the global parameters τ and ω , and the local parameters λ_j separately. The factorization of the posterior in (11) is not unique and we discuss another possible factorization in the online supplement.

Network Architecture To represent this posterior, we introduce a *hierarchical summary network*, which consists of a series of two summary networks of arbitrary architectures. The first summary network $h_{\psi_{\text{local}}}(x_j)$ transforms the group-level data x_j (with potentially different sample sizes) into fixed-length representations, whereas another summary network transforms the output of the first summary network for all groups into a fixed-length representation of the whole dataset: $h_{\psi_{\text{global}}}(\{h_{\psi_{\text{local}}}(x_j)\}_{j=1}^J)$. The architecture of the local summary network $h_{\psi_{\text{local}}}$ can be freely chosen according to the structure of the group-level data (e.g., recurrent neural networks for time series), while the global summary network $h_{\psi_{\text{global}}}$ is typically chosen to account for exchangeability of groups, for example, via permutation-invariant neural networks (e.g., Lee et al., 2019).

Similarly, we also use two inference networks to represent the factorization of the posterior in (11). A global inference network $f_{\phi_{\text{global}}}(\tau, \omega; h_{\psi_{\text{global}}}(\{h_{\psi_{\text{local}}}(x_j)\}_{j=1}^J))$ estimates the global parameters conditional on the output of the global summary network, while a local inference network $f_{\phi_{\text{local}}}(\lambda_j; \tau, \omega, h_{\psi_{\text{local}}}(x_j))$ estimates the local parameter for group j conditional on the global parameters and the output of the local summary network for inputs x_j . Figure 1 (right panel) shows a diagram of the network architecture.

Network Training Writing $\phi = (\phi_{\text{global}}, \phi_{\text{local}})$ for the weights of the inference networks and $\psi = (\psi_{\text{global}}, \psi_{\text{local}})$ for the weights of the summary networks, the optimization criterion of (7) becomes

$$\hat{\phi}, \hat{\psi} = \arg \min_{\phi, \psi} \mathbb{E}_{(\tau, \omega, \theta, x) \sim p(\tau, \omega, \theta, x)} \left[-\log q_{\phi, \psi}(\tau, \omega, \{\lambda_j\} | x) \right] \quad (12)$$

The negative log-density term inside the expectation can now be decomposed into:

$$\begin{aligned}
-\log q_\phi(\tau, \omega, \{\lambda_j\} | x) &= -\log q_{\phi_{\text{global}}}(\tau, \omega | h_{\psi_{\text{global}}}(\{h_{\psi_{\text{local}}}(x_j)\}_{j=1}^J)) \\
&\quad - \sum_{j=1}^J \log q_{\phi_{\text{local}}}(\lambda_j | \tau, \omega, h_{\psi_{\text{local}}}(x_j)),
\end{aligned} \tag{13}$$

which can be optimized via standard backpropagation.

Amortized Sampling from ML-NPE After training, the neural networks can be saved and reused for posterior inference on new datasets with any number of groups J and number of observations per group N_j , under the condition that they remain within the typical range seen during training.

Let $z_{\phi_{\text{global}}}^{(s)}$ and $z_{\phi_{\text{local}}}^{(s)}$ denote random draws with the index $s \in \{1, \dots, S\}$ from the base distribution of the global inference network $f_{\phi_{\text{global}}}$ and local inference network $f_{\phi_{\text{local}}}$, respectively (here: both unit Gaussian, see Subsection 2.2 for details). Let x^{obs} denote an observed dataset. Then, random draws from the approximate posterior $q_{\phi, \psi}(\tau, \omega, \{\lambda_j\} | x^{\text{obs}})$ can be obtained by the following ancestral sampling scheme:

$$\begin{aligned}
z_{\phi_{\text{global}}}^{(s)} &\sim \text{Normal}(z_{\phi_{\text{global}}} | 0, \mathbb{I}) \\
z_{\phi_{\text{local}}}^{(s)} &\sim \text{Normal}(z_{\phi_{\text{local}}} | 0, \mathbb{I}) \\
\tau^{(s)}, \omega^{(s)} &= f_{\phi_{\text{global}}}^{-1}(z_{\phi_{\text{global}}}^{(s)}; h_{\psi_{\text{global}}}(\{h_{\psi_{\text{local}}}(x_j^{\text{obs}})\}_{j=1}^J)) \\
\lambda_j^{(s)} &= f_{\phi_{\text{local}}}^{-1}(z_{\phi_{\text{local}}}^{(s)}; \tau^{(s)}, \omega^{(s)}, h_{\psi_{\text{local}}}(x_j^{\text{obs}})) \quad \text{for } j = 1, \dots, J,
\end{aligned} \tag{14}$$

where $f_{\phi_{\text{global}}}^{-1}$ and $f_{\phi_{\text{local}}}^{-1}$ denote inverse passes through the respective invertible inference networks. This approach ensures efficient sampling at each level, requiring only single passes through the pre-trained invertible networks. Additionally, using only a single inference network for the local parameters $\{\lambda_j\}$ prevents the output dimensionality of the local inference network from scaling linearly with the number of groups J .

3 Empirical Evaluation

We evaluate our neural two-level architecture in three different applications based on real-world datasets that aim to cover a broad spectrum of use cases for multilevel models. All models are implemented in the BayesFlow software for amortized Bayesian workflows (Radev et al., 2023c). Concretely, the applications include the following:

1. An autoregressive time series model that predicts and compares air passenger traffic between European countries and the United States of America.
2. A diffusion decision model to infer latent parameters of a decision-making process from reaction time data.
3. A generative neural network model to infer handwriting styles from high-dimensional image data.

3.1 Evaluation Metrics

For each application, we implement a series of comprehensive inference and model checks:

Simulation-Based Calibration We perform simulation-based calibration (SBC, e.g., Talts et al., 2018; Modrák et al., 2023) to ensure that ML-NPE samples from the correct posterior. SBC assesses correct posterior calibration (i.e., a $q \times 100\%$ posterior interval contains the true value in approximately $q \times 100\%$ of the cases) (e.g., Bürkner et al., 2023). Incorrect calibration hints at issues with computational validity, for example arising from insufficient simulation-based training or lack of expressiveness of the neural networks.

Posterior Predictive Checks To ensure that all models are reasonable descriptions of the observed data, we perform posterior predictive checks (e.g., Gabry et al., 2019). We investigate inferential accuracy by providing graphical comparisons of posterior intervals and ground truth parameter values over a test set. When computationally feasible, we also fit each model in Stan and provide plots of posterior intervals to compare the estimates against a known to be reliable reference algorithm. These model tests on non-simulated data are particularly important in amortized Bayesian inference because of a possible generalization gap: If the simulated data used for training does not lie within the scope of data the model is going to be applied on, amortized Bayesian inference might yield posterior draws that are far off from the true posterior.

Posterior Shrinkage As a final model check, we compare posterior shrinkage of the group-level parameters toward their global mean between our neural framework and Stan, as any subtle algorithmic error would likely first manifest as incorrect posterior shrinkage. Shrinkage is achieved by placing a hierarchical prior on the local parameters, such as $\eta_j \sim \text{Normal}(\mu_\eta, \sigma_\eta)$, where η_j is an individual local parameter and μ_η and σ_η are the hierarchical mean and standard deviation for that parameter, respectively. This prior structure pulls individual estimates η_j closer together, reducing variance and borrowing strength across groups (e.g., Carlin and Louis, 1997). Following Gelman (2006), a useful way to quantify this effect is the pooling factor κ_j :

$$\kappa_j = \frac{\text{Var}(\eta_j - \mu_\eta)^2}{\sigma_\eta^2}, \quad (15)$$

where $\text{Var}(\cdot)$ denotes the variance and σ_η^2 is the squared hierarchical standard deviation.

3.2 Experiment 1: Air Passenger Traffic

We apply ML-NPE to analyze trends in European air passenger traffic data provided by Eurostat (2022a,b,c). We highlight that our neural architecture can correctly and efficiently amortize over both the number of groups and the number of observations per

group, as well as over covariates. By using a summary network that is aligned to the structure of the data, we also show that ML-NPE can easily be extended to estimate non-IID response variables.

Model Description We obtain time series of annual air passenger counts between 16 European countries (departures) and the United States of America (destination) from 2004 to 2019 and fit the following autoregressive multilevel model:

$$y_{t+1,j} \sim \text{Normal}(\alpha_j + y_{t,j}\beta_j + u_{t,j}\gamma_j + v_{t,j}\delta_j, \sigma_j), \quad (16)$$

where the target quantity $y_{t+1,j}$ is the difference in air passenger traffic for country j between time $t+1$ and t , measured in units of 100 thousand passengers. To predict $y_{j,t+1}$, we use two additional predictors: (i) $u_{t,j}$, the annual growth rate of household debt in country j at time t , defined as the year-over-year percentage change in household debt as a share of GDP; and (ii) $v_{t,j}$, the annual growth rate of real GDP per capita in country j , defined as the year-over-year percentage change in real GDP per capita. The parameters α_j are the country-level intercepts, β_j are the autoregressive coefficients, γ_j are the regression coefficients of the household debt predictor, δ_j are the regression coefficients for the GDP per capita predictor, and σ_j is the standard deviation of the noise term. Mapping the notation to the previous sections, the observables x_j are now split into a target $y_j = \{y_{t,j}\}_{t=1}^T$ and predictors $u_j = \{u_{j,t}\}_{t=1}^T$ and $v_j = \{v_{j,t}\}_{t=1}^T$, with groups defined by the country index j . We assume independent priors and use the following marginal prior distributions:

$$\begin{array}{lll} \alpha_j \sim \text{Normal}(\mu_\alpha, \sigma_\alpha) & \mu_\alpha \sim \text{Normal}(0, 0.3) & \sigma_\alpha \sim \text{Normal}^+(0, 0.25) \\ \beta_j \sim \text{Normal}(\mu_\beta, \sigma_\beta) & \mu_\beta \sim \text{Normal}(0, 0.3) & \sigma_\beta \sim \text{Normal}^+(0, 0.25) \\ \gamma_j \sim \text{Normal}(\mu_\gamma, \sigma_\gamma) & \mu_\gamma \sim \text{Normal}(0, 0.3) & \sigma_\gamma \sim \text{Normal}^+(0, 0.25) \\ \delta_j \sim \text{Normal}(\mu_\delta, \sigma_\delta) & \mu_\delta \sim \text{Normal}(0, 0.3) & \sigma_\delta \sim \text{Normal}^+(0, 0.25) \\ \log(\sigma_j) \sim \text{Normal}(\mu_\sigma, \sigma_\sigma) & \mu_\sigma \sim \text{Normal}(-1, 0.5) & \sigma_\sigma \sim \text{Normal}^+(0, 1) \end{array} \quad (17)$$

As commonly done for autoregressive models, we regress on the differences between time periods to reduce issues due to non-stationarity. This is particularly important in an SBI setting, because for $\beta_j > 1$, the time series would exhibit strong exponential growth that would quickly surpass reasonable air traffic volumes, creating highly unrealistic simulations.

Model Training To generate training data for the neural approximator, we utilize an ancestral sampling scheme: We define the set of hierarchical means as $\mu = \{\mu_\alpha, \mu_\beta, \mu_\gamma, \mu_\delta, \mu_\sigma\}$ and the set of hierarchical standard deviations as $\sigma = \{\sigma_\alpha, \sigma_\beta, \sigma_\gamma, \sigma_\sigma\}$, which together form the set of hyperparameters $\tau = \{\mu, \sigma\}$. The local parameters are given by $\lambda_j = \{\alpha_j, \beta_j, \gamma_j, \delta_j, \sigma_j\}$ and the model does not have any shared parameters.

We generate training data indexed by m , where $m \in \{1, \dots, M\}$. For each training sample m , we first draw a set of hyperparameters from the prior distribution:

$$\tau^{(m)} \sim p(\tau). \quad (18)$$

Next, we sample J_m group-level parameters from the conditional distribution:

$$\{\lambda_j^{(m)}\}_{j=1}^{J_m} | \tau^{(m)} \sim p(\lambda_j | \tau^{(m)}), \quad (19)$$

where J_m is the total number of groups in the m th training sample, and each index j corresponds to a country. For each country, we then simulate a time series according to (16). To facilitate training on simulated data and without loss of generality, we assume that each time series begins with an initial value drawn from the stationary distribution

$$y_{0,j}^{(m)} \sim \text{Normal}(0, \sigma_j / (1 - \beta_j^2)). \quad (20)$$

Additionally, the predictor variables $u_j^{(m)}$ and $v_j^{(m)}$ are sampled from a standard normal distribution and then standardized to have a mean of 0 and standard deviation 1. Defining

$$x_j^{(m)} = \{y_j^{(m)}, u_j^{(m)}, v_j^{(m)}\} \quad (21)$$

and suppressing the indexing ranges from $j = 1$ to J_m for visual clarity, the final training dataset consists of

$$\{\tau^{(m)}, \{\lambda_j^{(m)}\}, \{x_j^{(m)}\}\}_{m=1}^M. \quad (22)$$

To amortize over the number of countries and observed time points, we vary $J_m \sim \text{DiscreteUniform}(10, 30)$ and $T_m \sim \text{DiscreteUniform}(5, 30)$ for each training sample m . The neural approximator can therefore generate draws from the posterior $p(\tau, \{\lambda_j\} | \{x_j\})$ for datasets with a varying number of countries and a varying number of time points without further training.

Network configuration details are provided in the online supplement. Briefly, the local summary network has an initial long short-term memory (LSTM; Hochreiter and Schmidhuber, 1997) layer to capture time dependencies at the group level. For the global summary network, we employ a set transformer (Lee et al., 2018) to ensure exchangeability of groups. The inference networks are implemented as neural spline flows (Durkan et al., 2019). The neural networks are trained for 50 epochs with a batch size of 64 and 128 iterations per epoch using the Adam optimizer (Kingma and Ba, 2014).

Posterior Predictive Checks To confirm that the simple AR(1) model is adequate enough to describe the observed data, Figure 2 shows the observed time series overlaid with draws from the posterior predictive distribution obtained by our ML-NPE method as well as a reference implementation in Stan. While predictive uncertainty is high, presumably in part because of the relatively simple model, short time series and predictors that are only weakly associated with the response variable, the predictions from the AR(1) model are consistent with the observed data.

Posterior Inference Figure 3 compares marginal posterior estimates with simulated ground truths (posterior recovery) and Figure 4 shows posterior calibration on 100 simulated validation datasets. We observe good posterior recovery and calibration for all parameters. Further, to assess posterior recovery on real data, we also apply our

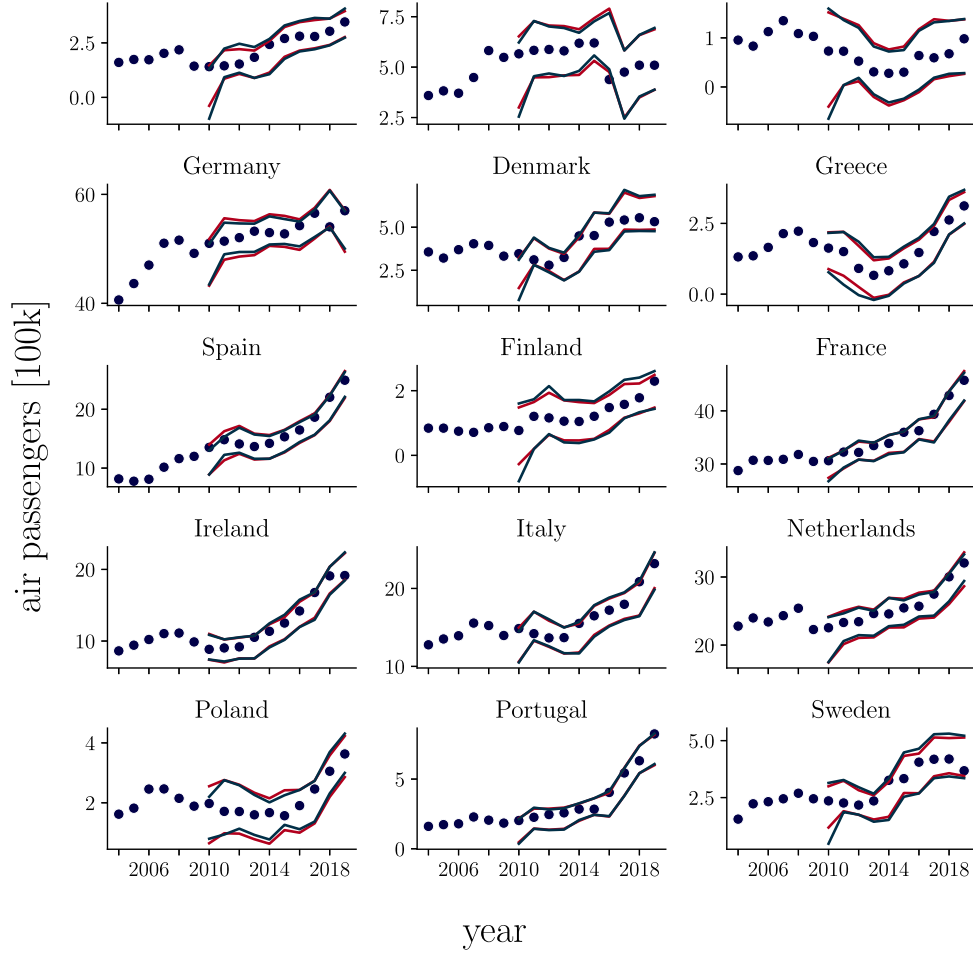


Figure 2: Annual air passenger traffic volume in 100 thousand passengers between different European countries and the United States of America. Blue points show observed data, blue lines show 95% uncertainty intervals of the 1-step ahead posterior predictive distribution. For reference, red lines show draws from the posterior predictive distribution of the same model fitted in Stan. The posterior predictive checks show that the model predictions are consistent with the observed data and intervals of the posterior predictive distribution obtained by the reference implementation in Stan are mostly indistinguishable from prediction intervals obtained by our amortized ML-NPE method.

model to the Eurostat (2022a,b,c) dataset and compare the results to the posterior intervals obtained by the Stan reference model implementation (see Figure 5). The plot shows that the results obtained by ML-NPE are consistent with the results obtained by Stan.

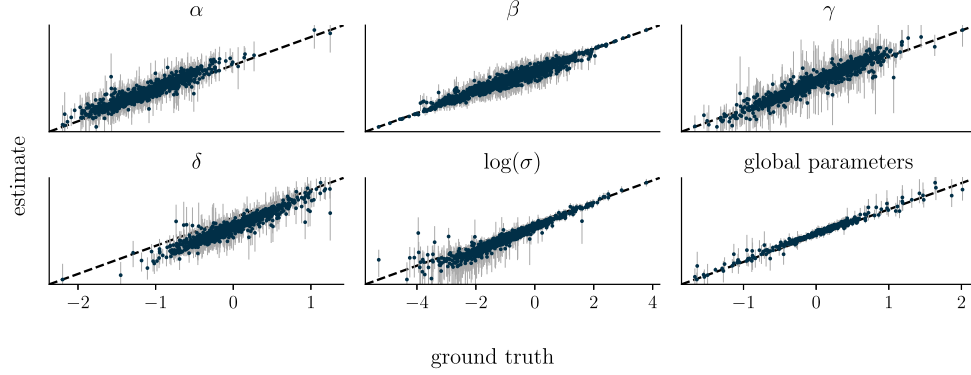


Figure 3: Marginal posterior estimates vs. ground truth on 50 simulated validation datasets. All parameters show good recovery of the true values, indicating that our ML-NPE is able to recover true parameter values without bias. Intervals represent 2.5%-97.5% posterior quantiles.

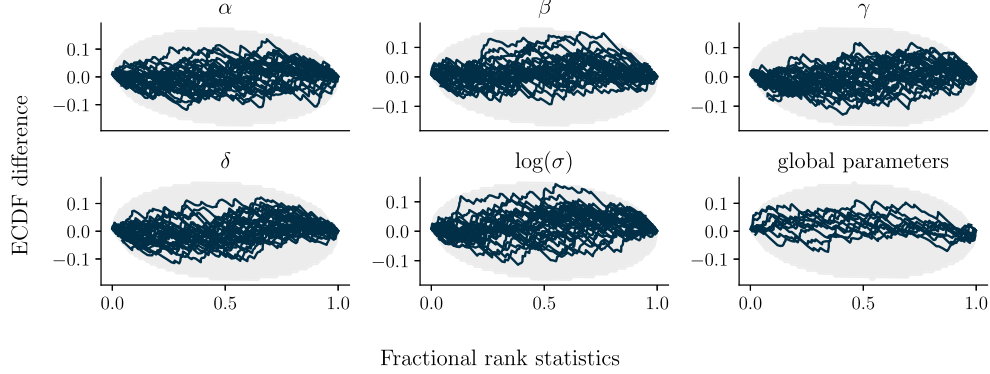


Figure 4: Simulation-based calibration plots based on 100 simulated validation sets. ECDF difference (empirical cumulative distribution function difference) is computed as the difference between the ECDF of the rank distribution (obtained by comparing the rank of prior draws with their corresponding posterior draws) and the uniform cumulative distribution function (see Säilynoja et al., 2022, for details). For the parameters α , β , γ , δ , and $\log(\sigma)$, each line represents a group-level parameter (i.e., the country-specific estimate). The panel *global parameters* contains the estimated mean and standard deviation for each of the 5 group-level parameters. All lines lie within the shaded 99% simultaneous confidence bands (Säilynoja et al., 2022), indicating well-calibrated marginal posterior distributions.

Posterior Shrinkage To verify correct shrinkage behavior of the ML-NPE approach, we compare posterior shrinkage to a reference implementation in Stan. Figure 6 shows scatter plots of shrinkage factors for each parameter obtained on 100 simulated datasets

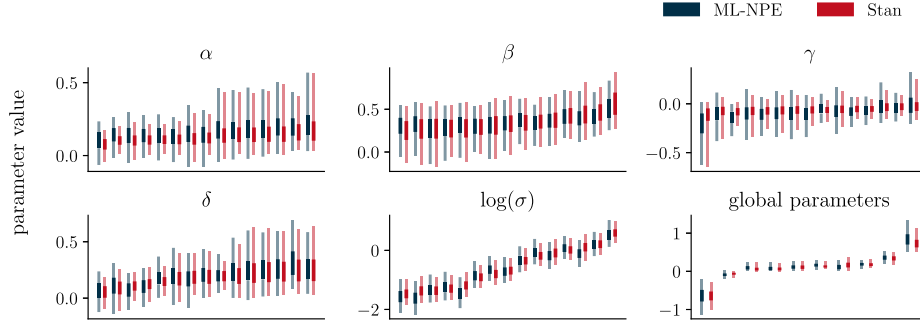


Figure 5: Posterior intervals obtained from amortized ML-NPE (ours) and the gold-standard Hamiltonian Monte Carlo (HMC) reference, as implemented via Stan (computed on non-simulated data). Plots show the central 50% (dark) and 95% (light) posterior credible intervals based on quantiles. For the parameters α , β , γ , δ , and $\log(\sigma)$, each interval pair refers to a single group-level parameter (one for each country). For the global parameters, each interval pair refers to a hyperparameter (one mean and one standard deviation for each of the group-level parameters, so 10 hyperparameters in total). The parameters are sorted by increasing mean (as per Stan) to ease interpretation of shrinkage towards a common mean.

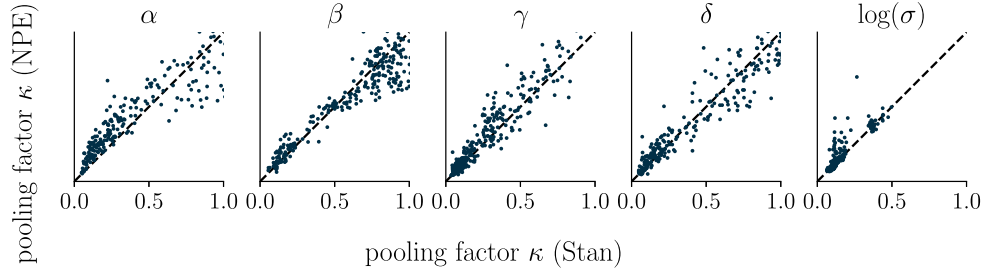


Figure 6: Posterior shrinkage observed in 100 simulated datasets via our ML-NPE methods to the results obtained by Stan. The scatter points lie along the diagonal, indicating that our method is able to correctly shrink group-level regression coefficients towards their common mean.

via our ML-NPE method compared to shrinkage factors obtained on the same data via Stan. The scatter points lie along the diagonal, indicating correct shrinkage behavior of our ML-NPE method.

3.3 Experiment 2: Diffusion Decision Model

As a second experiment, we apply ML-NPE to a popular decision-making model in psychology, neuroscience, and the cognitive sciences. We use this experiment to highlight

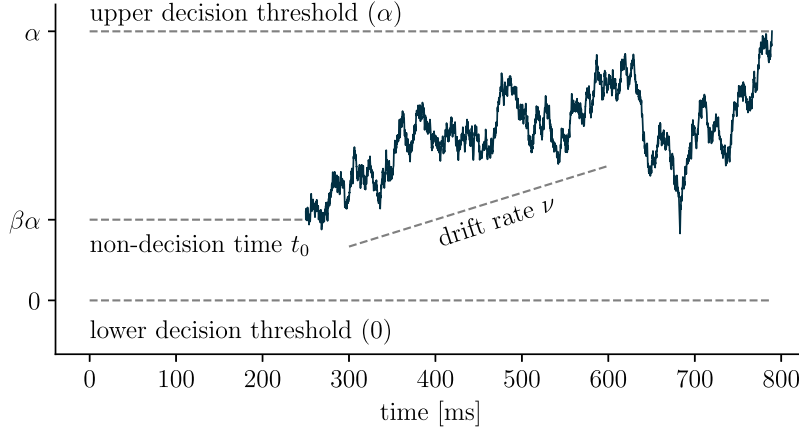


Figure 7: Example of a latent evidence trajectory for a single subject. The decision making process is modeled as a random walk that is governed by 4 parameters: The non-decision time t_0 is the time until evidence accumulation begins and captures components that are not directly related to the decision, such as time spent on processing of sensory information. The drift rate ν corresponds to the rate of information uptake. The decision threshold α can be interpreted as the amount of evidence necessary to make a decision and is thus a measurement of response caution. The relative starting point $\beta \in (0, 1)$ quantifies a subject’s tendency to prefer one choice over the other.

that NPE is not only well-suited for standard use cases like real-time analysis or fitting the same model across multiple datasets, but also for following a traditional Bayesian workflow, e.g. when performing model checking using cross-validation. Additionally, we show that ML-NPE can reliably amortize over both the number of groups and observations, enabling fast and reliable posterior inference for almost arbitrary datasets in this model class.

Model Description Consider a decision task in which participants are presented with sequences of letters and asked to differentiate between words and non-words (i.e., a lexical decision task). The Diffusion Decision Model (DDM; e.g., Ratcliff et al., 2016) simultaneously models this binary decision and the response time via a continuous evidence accumulation process: After an initial non-decision time t_0 , evidence accumulates following a noisy diffusion process with a certain drift rate ν , starting from a point β , until one of two decision thresholds $\{0, \alpha\}$ corresponding to the two choices is hit (see Figure 7). The nested structure of individual observations within subjects lends itself to a multilevel model where subject-specific estimates share information via common hyper-priors. For the first part of this experiment, we consider subject-specific estimates for the non-decision time $t_{0,j}$, drift rate ν_j and decision threshold α_j (local parameters), while estimating a global starting point β for all subjects (shared parameter). The gen-

erative model is summarized as:

$$\begin{aligned}
\nu_j &\sim \text{Normal}(\mu_\nu, \sigma_\nu) & \mu_\nu &\sim \text{Normal}(0.5, 0.3) & \log \sigma_\nu &\sim \text{Normal}(-1, 1) \\
\alpha_j &\sim \text{Normal}(\mu_\alpha, \sigma_\alpha) & \mu_\alpha &\sim \text{Normal}(0, 0.05) & \log \sigma_\alpha &\sim \text{Normal}(-3, 1) \\
t_{0,j} &\sim \text{Normal}(\mu_{t_0}, \sigma_{t_0}) & \mu_{t_0} &\sim \text{Normal}(-1, 0.3) & \log \sigma_{t_0} &\sim \text{Normal}(-1, 0.3) \\
\beta &\sim \text{Beta}(50, 50) \\
y_j &\sim \text{DDM}(\nu_j, \alpha_j, t_{0,j}, \beta),
\end{aligned} \tag{23}$$

where y_j is the vector containing tuples of binary decisions and corresponding reaction times for subject j , DDM is the likelihood of the 4-parameter diffusion decision model, β is the shared bias for all subjects, Beta is the Beta distribution with shape parameters a and b , Normal is the normal distribution with mean μ and standard deviation σ , and $\nu_j, \alpha_j, t_{0,j}$ are the subject-specific drift rate, boundary threshold and non-decision time parameters, respectively. The priors were chosen based on prior predictive checks to match the range of typical outcomes for such decision-making tasks.

Model Training To generate data for network training, we use the same ancestral sampling scheme as for the air traffic experiment in Subsection 3.2. First, we generate $m \in \{1 \dots, M\}$ random draws from the hyperparameters $\tau = \{\mu_\nu, \mu_\alpha, \mu_{t_0}, \sigma_\nu, \sigma_\alpha, \sigma_{t_0}\}$ and shared parameters $\omega = \{\beta\}$. For each of these M random draws, we then draw J_m draws from the corresponding subject specific local parameters $\lambda_j = \{\nu_j, \alpha_j, t_{0,j}\}$. For each subject, we simulate N_m observations from the diffusion process. To facilitate amortization over different group and observation sizes, we randomly draw values of J_m from a discrete uniform distribution with lower and upper bounds of 10 and 30, and N_m from a discrete uniform distribution with lower and upper bounds of 1 and 100. Details of the network configuration are shown in the online supplement. Briefly, we use set transformers (Lee et al., 2018) for both the local and global summary networks and neural spline flows (Durkan et al., 2019) for both inference networks. The simulation budget is set to $M = 20\,000$ and we train for 200 epochs using the Adam optimizer (Kingma and Ba, 2014).

Posterior Inference The online supplement presents posterior recovery and calibration results based on 100 simulated validation datasets, showing good posterior recovery and well-calibrated posteriors for all parameters.

To validate that we also observe accurate posterior inference on non-simulated data, we contrast these results to model fits on experimental data published by Wagenmakers et al. (2008). In total, 17 subjects were asked to make 32 word/non-word decisions in each of the 20 trial blocks. The experiment was repeated in two conditions: once after asking the subjects to focus on accuracy and once after asking the subjects to focus on speed. Figure 8 shows marginal posterior densities for the model fit on a subset of the data (observations in the accuracy condition in trial block 4) compared to the results obtained by Stan.

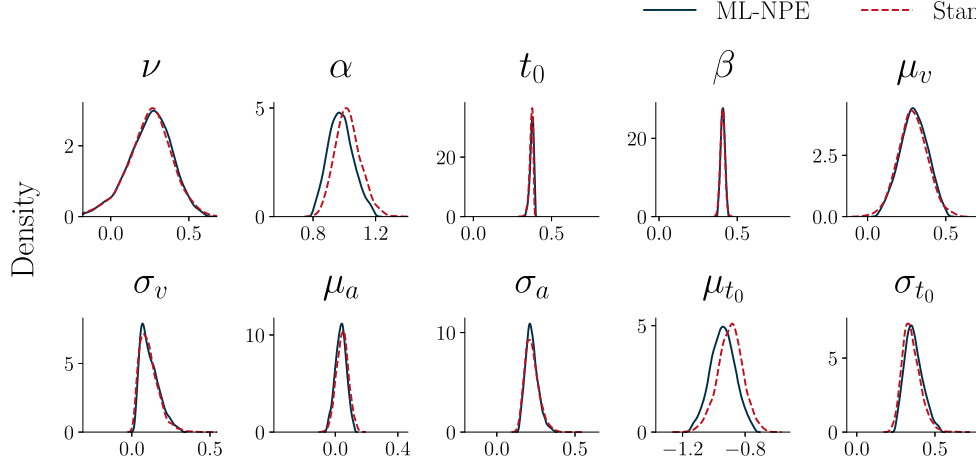


Figure 8: Marginal posterior distributions of the drift diffusion model fitted on experimental data. Blue kernel density plots were obtained by our ML-NPE method, red dashed lines show the results obtained by Stan as a gold standard for reliable posterior inference. The plots show that marginal posteriors obtained by ML-NPE are highly similar to the marginal posteriors obtained by Stan.

Leave-One-Group-Out Cross-Validation Amortized Bayesian inference is not only useful when performing posterior inference on many different datasets. Refitting the model multiple times on different subsets of the data is also required for many essential steps in the modern Bayesian workflow (Gelman et al., 2020), such as cross-validation (e.g., Vehtari et al., 2017). For multilevel models, it is often of practical interest to estimate and compare the predictive performance of models on new groups, leading to leave-one-group-out (LOGO) cross-validation. We define the leave-one-group-out posterior as

$$p(\theta | \{x_{-j}\}) = \frac{p(\{x_{-j}\} | \theta)p(\theta)}{p(\{x_{-j}\})}, \quad (24)$$

where $\{x_{-j}\} = x \setminus x_j$ denotes the data with the j th group removed.

As the LOGO posterior is often vastly different from the full posterior because the local parameters for the left-out group are only identified via their hyperparameters, approximation methods like Pareto-smoothed importance sampling (Vehtari et al., 2024) almost always fail. Performing LOGO cross-validation then requires refitting the model once for each group in the dataset, quickly rendering model comparison computationally infeasible when relying on, for example, MCMC algorithms for posterior inference. With amortized posterior inference, obtaining the LOGO posterior is almost instant (a fraction of a second for our architectures), as it does not require refitting the model for each data subset.

To show that ML-NPE enables model comparison for multilevel models that are otherwise computationally infeasible, we contrast the previously described diffusion decision

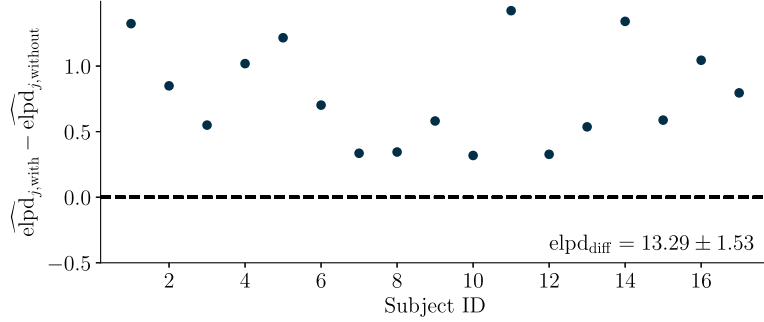


Figure 9: LOGO cross-validation on trial block 4 of the experimental dataset published by Wagenmakers et al. (2008). Scatter points show subject-wise elpd differences between the models with and without subject-specific β_j as defined in (26). The elpd difference $\text{elpd}_{\text{diff}} = \sum_{j=1}^J (\widehat{\text{elpd}}_{j,\text{with}} - \widehat{\text{elpd}}_{j,\text{without}})$ indicates a substantially better fit of the more flexible model with 4 local parameters per subject.

model in Section 3.3 with another model variant that also estimates β hierarchically. Instead of a single shared β for all subjects, this model variant estimates subject-specific β_j that share common hyper-priors. We make the following adjustments compared to the previous description of the diffusion decision model:

$$\beta_j \sim \text{BetaProportion}(\mu_\beta, \xi_\beta) \quad \mu_\beta \sim \text{Beta}(50, 50) \quad \xi \sim \text{Gamma}(5, 3), \quad (25)$$

where β_j is the bias for subject j , BetaProportion is the Beta distribution parameterized by a mean μ_β and concentration parameter ξ_β . The Beta proportion distribution is linked to the (standard) shape parameterization of the Beta distribution via the relation $\alpha = \mu\xi$ and $\beta = (1 - \mu)\xi$. This adapted version of the diffusion decision model also illustrates that ML-NPE can estimate non-normal hyper-prior distributions.

Figure 9 shows expected log predictive density (elpd; Vehtari et al., 2017) values for each subject obtained by first removing subject j from the dataset, generating samples from the LOGO posterior, and then evaluating the predictive performance on the left-out subject:

$$\widehat{\text{elpd}}_j = \log\left(\frac{1}{S} \sum_{s=1}^S p(x_j | \theta^{(s)})\right), \quad (26)$$

where $\theta^{(s)}$ denotes $s = 1, \dots, S$ draws from the leave-one-group-out posterior $p(\theta | \{x_{-j}\})$.

The elpd difference $\text{elpd}_{\text{diff}} = \sum_{j=1}^J (\widehat{\text{elpd}}_{j,\text{with}} - \widehat{\text{elpd}}_{j,\text{without}})$ between the models with and without subject-specific β_j evaluates to $\text{elpd}_{\text{diff}} \approx 13.29$, which is several times larger than its standard error $\text{SE}_{\text{diff}} \approx 1.53$ (Vehtari et al., 2017). That is, there is strong statistical evidence that the model with varying β_j has better predictive performance on data from a previously unseen group.

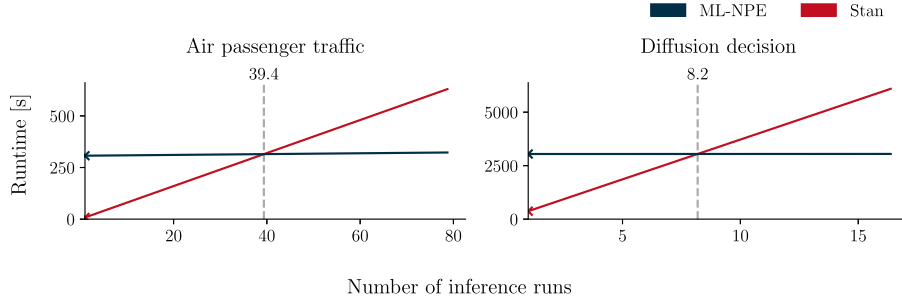


Figure 10: Runtime comparison between Stan (4 parallel chains, 1000 iterations after warm-up) and ML-NPE for the air traffic (left) and diffusion decision (right) case studies, extrapolated from a single run. MCMC scales linearly with the number of inference runs, ML-NPE incurs a higher upfront training cost but enables near-instant inference thereafter. For the air traffic model, Stan takes 8 s per run versus 307 s for training ML-NPE; for the diffusion decision model, Stan takes 372 s per run versus 3044 s for ML-NPE; and for the style inference model, only ML-NPE is applicable, with training taking 5404 s.

Performing such a model comparison is computationally efficient in an amortized Bayesian inference setting, but quickly becomes infeasible when the model is fit using MCMC-based algorithms. Figure 10 shows a brief overview of runtime comparisons between Stan and ML-NPE. For the diffusion decision model, using Stan to draw 1000 samples (after 1000 warm-up draws) from 4 chains in parallel takes about 6 minutes on a standard desktop processor, resulting in a total runtime of about 1.5 hours for all subjects. In contrast, training the amortized model using ML-NPE takes about 1 hour on a standard desktop graphics card, subsequently allowing almost instant inference (a fraction of a second) of the LOGO posteriors. Typical applications of the DDM often feature hundreds of subjects and observations. In these settings, traditional HMC methods can quickly become computationally infeasible even on large compute clusters, but can still be performed on a standard desktop computer using NPE. The online supplement includes a brief discussion of use cases where users can benefit from NPE rather than MCMC-based workflows.

Complementary to cross-validation, amortized deep methods for Bayesian model comparison can also estimate posterior model probabilities efficiently in hierarchical settings (Elsemüller et al., 2024).

3.4 Experiment 3: Style Inference of Hand-Drawn Digits

As a final experiment, we demonstrate the efficacy of our amortized multilevel approximator on unstructured, high-dimensional observations. Concretely, the observations are images that are generated from a pre-trained generative neural network which we treat as a black-box simulator (i.e. an implicit data model), see Figure 11 for an overview.

The simulator is a so-called general incompressible flow network (GIN, Sorrenson

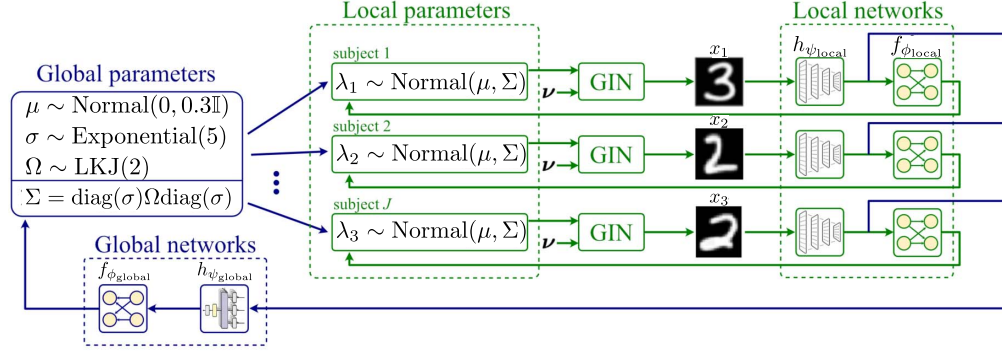


Figure 11: **Experiment 3.** Overview of the hierarchical model of digit style inference.

et al., 2020), which has been trained on a dataset consisting of handwritten digits that are represented as labeled gray-scale images with a dimension of 28×28 pixels. GINs have a special architecture that allows them to recover latent dimensions from some observed data x , given that the data are also annotated with some conditions u (e.g., class labels or time indices of a time series). The main advantage of GINs over traditional methods for identifying latent variables from observed data such as independent component analysis (Comon, 1994) is that the relationship between latent variables and data can be non-linear, while the method is still theoretically well justified. Similarly to the normalizing flows introduced in Subsection 2.2, GINs learn an invertible mapping g from the latent variables $w \in \mathbb{R}^d$ to the data space $x = g(w; \phi)$, where g is parameterized by network weights ϕ . In their experiment, Sorrenson et al. (2020) trained a GIN on 240 000 images of handwritten digits and corresponding labels from the EMNIST (Extended Modified National Institute of Standards and Technology) dataset (Cohen et al., 2017). They identified 22 broadly interpretable latent space variables, of which the first 8 variables encode general style features such as slant and line thickness, while the remaining 14 features encode styles of individual digits (e.g., writing a seven with or without a middle stroke).

Because the learned mapping between latent style features and observations is invertible, we can use the trained GIN from Sorrenson et al. (2020) to simulate images conditioned on the style parameters. For this experiment, we restrict ourselves to the first 4 general style parameters (width of top half of the digit, slant, height, and bend through center), treating all other style parameters as random noise. Our multilevel setting features ten subjects $j = 1, \dots, 10$. Each subject has a 4-dimension style vector $\lambda_j = (\lambda_{j,1}, \lambda_{j,2}, \lambda_{j,3}, \lambda_{j,4})$ which constitute the local parameters $\{\lambda_j\}_{j=1}^J$. The global parameters are an average style $\mu \in \mathbb{R}^4$, a style dispersion $\sigma \in \mathbb{R}_{>0}^4$, and the 6 unique elements of a correlation matrix $\Omega \in [-1, 1]^{4 \times 4}$ between styles. The full multilevel model is given by:

$$\begin{aligned}
 \mu &\sim \text{Normal}(0, 0.3 \cdot \mathbb{I}), \quad \sigma \sim \text{Exponential}(5), \quad \Omega \sim \text{LKJ}(2), & (\text{global parameters}) \\
 \lambda_j &\sim \text{Normal}(\mu, \Sigma) \quad \text{with } \Sigma = \text{diag}(\sigma) \Omega \text{diag}(\sigma) \text{ and } j = 1, \dots, J, & (\text{local parameters}) \\
 x_j &= \text{GIN}(\lambda_j, \nu) \text{ with } j = 1, \dots, J, & (\text{data model})
 \end{aligned}
 \tag{27}$$

where $\text{Normal}(\mu, \Sigma)$ is the multivariate normal distribution with mean vector μ and covariance Σ , \mathbb{I} is the 4×4 identity matrix, $\text{Exponential}(5)$ is the exponential distribution with rate 5, $\text{LKJ}(2)$ is the Lewandowski-Kurowicka-Joe distribution (Lewandowski et al., 2009) with shape 2, and $\text{GIN} : (\lambda_j, \nu) \mapsto x_j$ is the forward pass of the pre-trained global incompressible-flow network from Sorrenson et al. (2020) with style vector λ_j and unmodeled style parameters $\nu \sim \text{Normal}(\mu, 0.5)$.

Model Training The local summary network $h_{\psi_{\text{local}}}$ transforms 28×28 images to a summary vector of dimension 32 through a series of residual convolutional blocks followed by average pooling. The global summary network $h_{\psi_{\text{global}}}$ is a set transformer (Lee et al., 2019) that learns a 32-dimensional summary vector from the local summary vectors of all subjects. The online supplement contains details about the neural networks and training settings.

Results We draw 1000 approximate posterior samples from our multilevel neural posterior estimator conditioned on simulated data of $J = 10$ subjects and compare posterior estimates to simulated ground truths. We also report posterior calibration of local and global parameters. The estimates of the global parameters μ , σ , Ω show excellent calibration (see Figure 12b). While the average styles μ and dispersion σ can be recovered with near-perfect accuracy, the estimates of Ω show much greater uncertainty due to the limited sample size of $J = 10$ subjects. All local parameters of all subjects can be recovered with good precision and calibration (see Figure 12a).

4 Conclusion

In this work, we developed a framework for amortized Bayesian inference of multilevel models using deep generative neural networks. By utilizing the dependency structure of multilevel models and mirroring them in the neural architectures, we show efficient training of the neural approximators and flexible amortization over both the number of groups and the number of observations within groups. This means that, once trained, the neural approximators provide near instant inference of multilevel models on any number of new datasets or data subsets. We demonstrated amortized Bayesian inference on multilevel models in three realistic experiments, where we show both the amortization capabilities and the accuracy of the resulting Bayesian inference as validated by simulation-based calibration and comparison with Stan. To foster practical application, we implement our methods and architectures in the Python library BayesFlow (Radev et al., 2023a), which provides efficient and user-friendly workflows for amortized Bayesian inference.

4.1 Future Work

So far, research on amortized multilevel modeling has mainly focused on two-level models. This already covers a lot of use cases, but still leaves out many practically relevant scenarios of inference on datasets with more than two levels, which should be considered

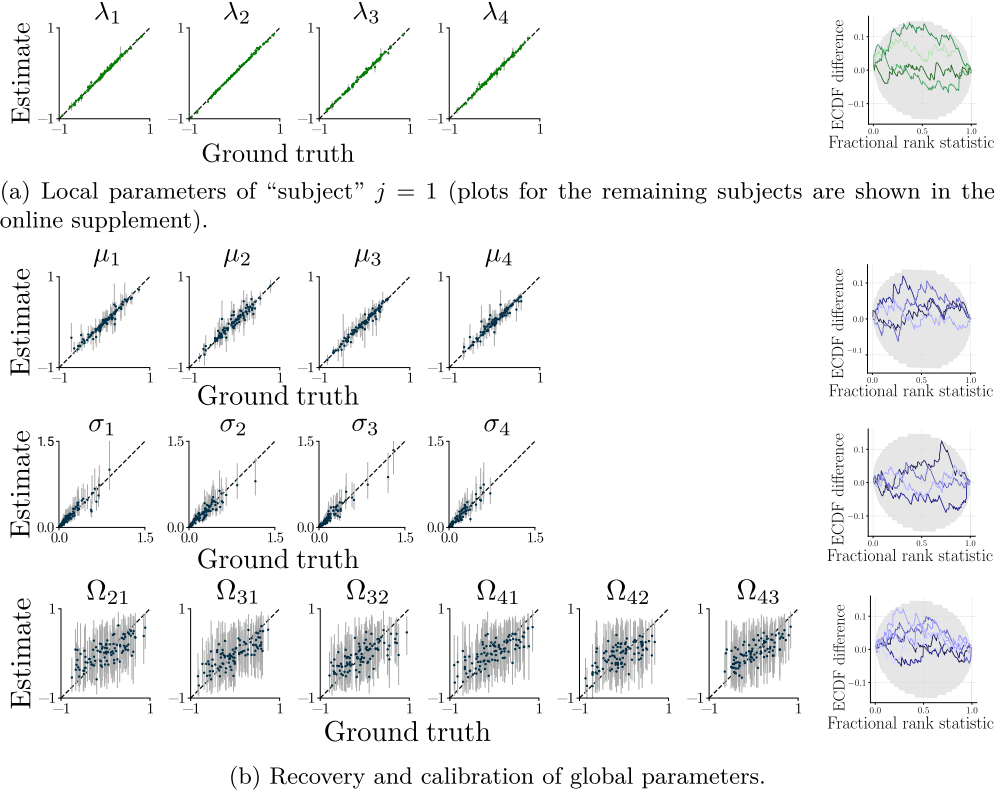


Figure 12: **Experiment 3.** The recovery of the local parameters (a, left) and global parameters (b, left) is at the upper limit given the epistemic and aleatoric uncertainty in the probabilistic model. Dots indicate posterior means, vertical bars represent the symmetric 95% posterior credible intervals. All local and global parameters are well-calibrated (right, SBC-ECDF plots).

in future work. For this, future work on amortization for arbitrary hierarchy structures is needed to: (i) provide algorithms that utilize the dependency structure implied by a probabilistic graph to automatically identify an amortization-friendly factorization of the joint posterior, (ii) implement software to connect the individual network components in an easy-to-use manner. Another challenge is presented by models with very expensive simulators, such that only very few (say, a maximum of a couple of hundred) simulations are available for training the neural approximators. In this low data setting, fully amortized methods may not achieve acceptable levels of inference accuracy and calibration (Lueckmann et al., 2021; Schmitt et al., 2024a; Radev et al., 2023b). Instead, sequential neural methods, which enable targeted inference only for a single dataset (Greenberg et al., 2019) may be better suited but have yet to be developed and validated for multilevel models.

Funding

Daniel Habermann, Lars Kühmichel, Stefan Radev, and Paul Bürkner acknowledge support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Projects 508399956 and 528702768. Paul Bürkner acknowledges support of DFG Collaborative Research Center 391 (Spatio-Temporal Statistics for the Transition of Energy and Transport) – 520388526. Marvin Schmitt and Paul Bürkner acknowledge support of Cyber Valley Project CyVy-RF- 2021-16, the DFG under Germany’s Excellence Strategy – EXC-2075 - 390740016 (the Stuttgart Cluster of Excellence SimTech). Marvin Schmitt acknowledges travel support from the European Union’s Horizon 2020 research and innovation programme under grant agreements No 951847 (ELISE) and No 101070617 (ELSA), and the Aalto Science-IT project.

Supplementary Material

Supplementary Material for “Amortized Bayesian Multilevel Models” (DOI: [10.1214/25-BA1570SUPP](https://doi.org/10.1214/25-BA1570SUPP); .pdf). The Supplementary Material provides a brief overview of the related topics *simulation-based inference*, *amortized Bayesian inference* and *neural density estimation*. It also includes a discussion of an alternative factorization of the joint posterior presented in (11), a short note when neural posterior estimation may be preferable to traditional MCMC sampling, and detailed descriptions of the network training procedures used in all case studies.

References

- Arruda, J., Schälte, Y., Peiter, C., Teplytska, O., Jaehde, U., and Hasenauer, J. (2023). “An amortized approach to non-linear mixed-effects modeling based on neural posterior estimation.” *bioRxiv*. 2
- Barber, S., Voss, J., and Webster, M. (2015). “The rate of convergence for approximate Bayesian computation.” *Electronic Journal of Statistics*, 9(1). MR3306571. doi: <https://doi.org/10.1214/15-EJS988>. 4
- Beneventano, P., Cheridito, P., Jentzen, A., and von Wurstemberger, P. (2020). “High-dimensional approximation spaces of artificial neural networks and applications to partial differential equations.” *arXiv preprint arXiv:2012.04326*. 4
- Betancourt, M. (2017). “A conceptual introduction to Hamiltonian Monte Carlo.” *arXiv preprint arXiv:1701.02434*. MR1699395. doi: <https://doi.org/10.1017/CB09780511470813.003>. 1
- Bürkner, P.-C., Scholz, M., and Radev, S. T. (2023). “Some models are useful, but how do we know which ones? Towards a unified Bayesian model taxonomy.” *Statistics Surveys*, 17(none): 216–310. MR4671148. doi: <https://doi.org/10.1214/23-ss145>. 11
- Carlin, B. P. and Louis, T. A. (1997). “Bayes and empirical Bayes methods for data analysis.” MR1427749. 11

- Chen, Y., Zhang, D., Gutmann, M., Courville, A., and Zhu, Z. (2020). “Neural approximate sufficient statistics for implicit models.” *arXiv preprint*. 7
- Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. (2017). “EMNIST: Extending MNIST to handwritten letters.” In *2017 international joint conference on neural networks (IJCNN)*, 2921–2926. IEEE. 22
- Comon, P. (1994). “Independent component analysis, a new concept?” *Signal processing*, 36(3): 287–314. 22
- Cranmer, K., Brehmer, J., and Louppe, G. (2020). “The frontier of simulation-based inference.” *Proceedings of the National Academy of Sciences*, 117(48): 30055–30062. MR4263287. doi: <https://doi.org/10.1073/pnas.1912789117>. 1, 4
- Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. (2019). “Neural Spline Flows.” *arXiv preprint*. 13, 18
- Else Müller, L., Schnuerch, M., Bürkner, P.-C., and Radev, S. T. (2024). “A deep learning method for comparing Bayesian hierarchical models.” *Psychological Methods*. 21
- Eurostat (2022a). “Household debt, consolidated including Non-profit institutions serving households - % of GDP.” 11, 14
- Eurostat (2022b). “International extra-EU air passenger transport by reporting country and partner world regions and countries.” 11, 14
- Eurostat (2022c). “Real GDP per capita.” 11, 14
- Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). “Visualization in Bayesian Workflow.” *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(2): 389–402. MR3902665. doi: <https://doi.org/10.1111/rssa.12378>. 11
- Gelman, A. (2006). “Multilevel (Hierarchical) Modeling: What It Can and Cannot Do.” *Technometrics*, 48(3): 432–435. MR2252307. doi: <https://doi.org/10.1198/004017005000000661>. 11
- Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. 1
- Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). “Bayesian Workflow.” *arXiv preprint*. MR4298989. doi: <https://doi.org/10.1214/20-ba1221>. 4, 19
- Gershman, S. and Goodman, N. (2014). “Amortized inference in probabilistic reasoning.” In *Proceedings of the annual meeting of the cognitive science society*, volume 36. 2, 7
- Goldstein, H. (2011). *Multilevel statistical models*. John Wiley & Sons. 1
- Greenberg, D., Nonnenmacher, M., and Macke, J. (2019). “Automatic Posterior Transformation for Likelihood-Free Inference.” In *Proceedings of the 36th International Conference on Machine Learning*, 2404–2414. PMLR. ISSN: 2640-3498. 6, 24

- Habermann, D., Schmitt, M., Kühmichel, L., Bulling, A., Radev, S. T., and Bürkner, P.-C. (2025). “Supplementary Material for “Amortized Bayesian Multilevel Models”.” *Bayesian Analysis*. doi: <https://doi.org/10.1214/25-BA1570SUPP>. 3
- Heinrich, L., Mishra-Sharma, S., Pollard, C., and Windischhofer, P. (2023). “Hierarchical Neural Simulation-Based Inference Over Event Ensembles.” 2, 8
- Hochreiter, S. and Schmidhuber, J. (1997). “Long Short-Term Memory.” *Neural Computation*, 9(8): 1735–1780. 13
- Hoffman, M. and Blei, D. (2015). “Stochastic Structured Variational Inference.” In Lebanon, G. and Vishwanathan, S. V. N. (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, 361–369. San Diego, California, USA: PMLR. 3
- Hoffman, M., Sountsov, P., Dillon, J. V., Langmore, I., Tran, D., and Vasudevan, S. (2019). “Neutralizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport.” *arXiv preprint*. 3
- Kingma, D. P. and Ba, J. (2014). “Adam: A Method for Stochastic Optimization.” *arXiv preprint*. 13, 18
- Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2020). “Normalizing flows: An introduction and review of current methods.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11): 3964–3979. 6
- Lee, A., Yau, C., Giles, M. B., Doucet, A., and Holmes, C. (2010). “On the Utility of Graphics Cards to Perform Massively Parallel Simulation of Advanced Monte Carlo Methods.” *Journal of Computational and Graphical Statistics*, 19(4): 769–789. 3
- Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. (2019). “Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks.” In *Proceedings of the 36th International Conference on Machine Learning*, 3744–3753. 9, 23
- Lee, J., Lee, Y., Kim, J., Kosiorek, A. R., Choi, S., and Teh, Y. W. (2018). “Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks.” *arXiv preprint*. 13, 18
- Lewandowski, D., Kurowicka, D., and Joe, H. (2009). “Generating random correlation matrices based on vines and extended onion method.” *Journal of Multivariate Analysis*, 100(9): 1989–2001. MR2543081. doi: <https://doi.org/10.1016/j.jmva.2009.04.008>. 23
- Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. (2021). “Benchmarking Simulation-Based Inference.” In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 343–351. PMLR. 24
- Margossian, C., Vehtari, A., Simpson, D., and Agrawal, R. (2020). “Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for

- latent Gaussian models and beyond.” *Advances in Neural Information Processing Systems*, 33: 9086–9097. [MR4361132](#). 3
- Margossian, C. C., Hoffman, M. D., Sountsov, P., Riou-Durand, L., Vehtari, A., and Gelman, A. (2024). “Nested \hat{R} : Assessing the Convergence of Markov Chain Monte Carlo When Running Many Short Chains.” *Bayesian Analysis*, 1–28. 3
- Margossian, C. C. and Saul, L. K. (2023). “The shrinkage-delinkage trade-off: An analysis of factorized gaussian approximations for variational inference.” In *Uncertainty in Artificial Intelligence*, 1358–1367. PMLR. 3
- Modrák, M., Moon, A. H., Kim, S., Bürkner, P., Huurre, N., Faltejsková, K., Gelman, A., and Vehtari, A. (2023). “Simulation-based calibration checking for Bayesian computation: The choice of test quantities shapes sensitivity.” *Bayesian Analysis*, 1(1): 1–28. [MR4905015](#). doi: <https://doi.org/10.1214/23-ba1404>. 4, 11
- Radev, S. T., D’Alessandro, M., Mertens, U. K., Voss, A., Köthe, U., and Bürkner, P.-C. (2023a). “Amortized Bayesian Model Comparison With Evidential Deep Learning.” *IEEE Transactions on Neural Networks and Learning Systems*, 34(8): 4903–4917. 23
- Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., and Köthe, U. (2020a). “BayesFlow: Learning Complex Stochastic Models With Invertible Neural Networks.” *IEEE Transactions on Neural Networks and Learning Systems*, 33(4): 1452–1466. [MR4516681](#). doi: <https://doi.org/10.1109/tnnls.2021.3089134>. 6, 7
- Radev, S. T., Schmitt, M., Pratz, V., Picchini, U., Köthe, U., and Bürkner, P.-C. (2023b). “JANA: Jointly Amortized Neural Approximation of Complex Bayesian Models.” In Evans, R. J. and Shpitser, I. (eds.), *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, 1695–1706. PMLR. 24
- Radev, S. T., Schmitt, M., Schumacher, L., Else Müller, L., Pratz, V., Schälte, Y., Köthe, U., and Bürkner, P.-C. (2023c). “BayesFlow: Amortized Bayesian Workflows With Neural Networks.” *Journal of Open Source Software*, 8(89): 5702. 2, 10
- Radev, S. T., Voss, A., Wieschen, E. M., and Bürkner, P.-C. (2020b). “Amortized Bayesian Inference for Models of Cognition.” *arXiv preprint*. 2
- Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). “Diffusion decision model: Current issues and history.” *Trends in cognitive sciences*, 20(4): 260–281. 17
- Rodrigues, P. L. C., Moreau, T., Louppe, G., and Gramfort, A. (2021). “HNPE: Leveraging Global Parameters for Neural Posterior Estimation.” 2
- Säilynoja, T., Bürkner, P.-C., and Vehtari, A. (2022). “Graphical test for discrete uniformity and its applications in goodness-of-fit evaluation and multiple sample comparison.” *Statistics and Computing*, 32(2): 32. [MR4402179](#). doi: <https://doi.org/10.1007/s11222-022-10090-6>. 15
- Schmitt, M., Habermann, D., Köthe, U., Bürkner, P.-C., and Radev, S. T. (2024a). “Leveraging Self-Consistency for Data-Efficient Amortized Bayesian Inference.” *Proceedings of the International Conference on Machine Learning (ICML)*. 24

- Schmitt, M., Pratz, V., Köthe, U., Bürkner, P.-C., and Radev, S. T. (2024b). “Consistency Models for Scalable and Fast Simulation-Based Inference.” *arXiv preprint*. 6
- Sharrock, L., Simons, J., Liu, S., and Beaumont, M. (2022). “Sequential Neural Score Estimation: Likelihood-Free Inference with Conditional Score Based Diffusion Models.” 6
- Sisson, S. A. and Fan, Y. (2011). “Likelihood-free MCMC.” *Handbook of Markov Chain Monte Carlo*, 313–335. MR2858454. 4
- Sisson, S. A., Fan, Y., and Beaumont, M. A. (2018). “Overview of ABC.” MR3889278. 4
- Sorrenson, P., Rother, C., and Köthe, U. (2020). “Disentanglement by Nonlinear ICA with General Incompressible-flow Networks (GIN).” In *International Conference on Learning Representations*. 21, 22, 23
- Stan Development Team (2024). “Stan Modeling Language Users Guide and Reference Manual, 2.31.0.” URL <https://mc-stan.org> 2
- Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). “Validating Bayesian Inference Algorithms with Simulation-Based Calibration.” *arXiv preprint*. 4, 11
- Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P. J., Greenberg, D. S., and Macke, J. H. (2020). “sbi: A toolkit for simulation-based inference.” *Journal of Open Source Software*, 5(52): 2505. 2
- Tran, D., Ranganath, R., and Blei, D. (2017). “Hierarchical implicit models and likelihood-free variational inference.” *Advances in Neural Information Processing Systems*, 30. 4
- Vehtari, A., Gelman, A., and Gabry, J. (2017). “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.” *Statistics and Computing*, 27(5): 1413–1432. MR3647105. doi: <https://doi.org/10.1007/s11222-016-9696-4>. 4, 19, 20
- Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2024). “Pareto Smoothed Importance Sampling.” *Journal of Machine Learning Research*, 25(72): 1–58. MR4749108. 19
- Wagenmakers, E.-J., Ratcliff, R., Gomez, P., and McKoon, G. (2008). “A diffusion model account of criterion shifts in the lexical decision task.” *Journal of Memory and Language*, 58(1): 140–159. 18, 20
- Wildberger, J., Dax, M., Buchholz, S., Green, S., Macke, J. H., and Schölkopf, B. (2024). “Flow Matching for Scalable Simulation-Based Inference.” *Advances in Neural Information Processing Systems*, 36. 2, 6
- Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). “Yes, but Did It Work?: Evaluating Variational Inference.” In *Proceedings of the 35th International Conference on Machine Learning*, 5581–5590. PMLR. ISSN: 2640-3498. 3, 4

Zhang, L., Carpenter, B., Gelman, A., and Vehtari, A. (2022). “Pathfinder: Parallel quasi-Newton variational inference.” *Journal of Machine Learning Research*, 23(306): 1–49. [MR4577745](#). 3