Inductive Domain Transfer In Misspecified Simulation-Based Inference

Ortal Senouf *
EPFL
Lausanne, Switzerland

Antoine Wehenkel
Apple
Zürich, Switzerland

Cédric Vincent-Cuaz
EPFL
Lausanne, Switzerland

Emmanuel Abbé
EPFL, Apple
Lausanne, Switzerland

Pascal Frossard
EPFL
Lausanne, Switzerland

Abstract

Simulation-based inference (SBI) of latent parameters in physical systems is often hindered by model misspecification–the mismatch between simulated and real-world observations caused by inherent modeling simplifications. RoPE, a recent SBI approach, addresses this challenge through a two-stage domain transfer process that combines semi-supervised calibration with optimal transport (OT)-based distribution alignment. However, RoPE operates in a fully transductive setting, requiring access to a batch of test samples at inference time, which limits scalability and generalization. We propose a fully inductive and amortized SBI framework that integrates calibration and distributional alignment into a single, end-to-end trainable model called FRISBI. Our method leverages mini-batch OT with a closed-form coupling to align real and simulated observations that correspond to the same latent parameters, using both paired calibration data and unpaired samples. A conditional normalizing flow is then trained to approximate the OT-induced posterior, enabling efficient inference without simulation access at test time. Across a range of synthetic and real-world benchmarks–including complex medical biomarker estimation–our approach matches or exceeds the performance of RoPE, while offering improved scalability and applicability in challenging, misspecified environments.

1 Introduction

Inference of latent variables that describe important properties of physical systems is a fundamental problem in many domains, including environmental [1, 2], mechanical [3, 4, 5], and physiological [6, 7, 8] systems. Traditionally, this problem has been approached by formulating a mathematical model that relates the observations $\boldsymbol{x}$ to the latent parameters of interest $\boldsymbol{\theta}$, and solving the corresponding inverse problem to infer $\boldsymbol{\theta}$ from $\boldsymbol{x}$ [9, 10].

Modern machine learning (ML) has achieved remarkable success in complex tasks, sparking interest in its application to inferring latent parameters from observations. However, standard supervised learning is often impractical in this context, as ground truth parameter data is typically expensive or infeasible to obtain, such as in medical applications where direct measurement may require invasive procedures. To address this, two prominent approaches have emerged: *simulation-based inference* (SBI) [11] and *hybrid learning* [12, 13, 14]. SBI trains ML models on simulated data to directly estimate parameters while capturing uncertainty through posterior estimation, becoming a

*Corresponding Author, ortal.senouf@epfl.ch

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

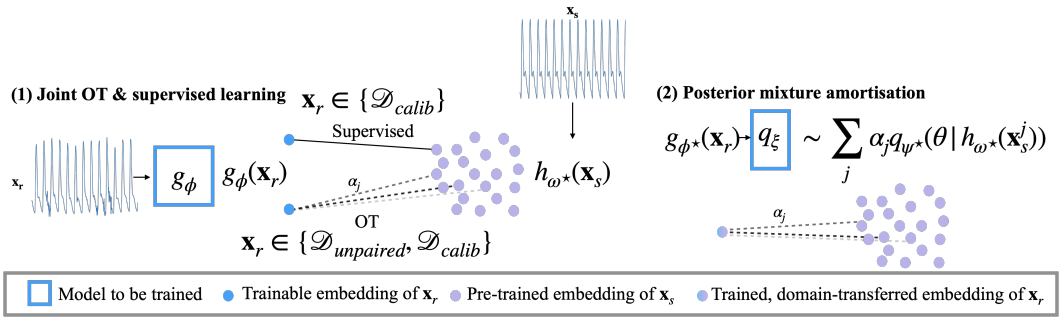

Figure 1: **FRISBI Overview.** Similar to RoPE [22], we assume a trained *neural statistics encoder* (NSE), $h_{\omega^\star}$, that maps simulation data $\boldsymbol{x}_s$ to embeddings $h_{\omega^\star}(\boldsymbol{x}_s)$, and a *neural posterior estimator* (NPE), $q_{\psi^\star}$, which estimates simulated posterior distributions. FRISBI performs: **(1)** Joint optimal transport (OT) and supervised learning. Both paired and unpaired samples contribute to the OT plan (dashed lines), weighted by α_j. Supervised samples from $\mathcal{D}_{calib}$ (solid lines) anchor the OT matching. The real-observations encoder g_ϕ is fine-tuned to optimize representations for both supervised learning and OT-based domain transfer. **(2)** A *conditional density estimator*, q_ξ, approximates the posterior arising from the OT-based mixture of posteriors.

cornerstone in scientific domains [15, 16, 17, 18, 19], though it can suffer from sensitivity to model misspecification [20, 21]. In hybrid learning, the simulator is integrated with ML components to obtain a more accurate model of the system. While this approach provides a useful inductive bias, it often requires simulators that are differentiable and computationally feasible, limiting its applicability to realistic and complex systems.

A recent work introduces RoPE [22], an SBI approach designed to address model misspecifications through a two-stage, semi-supervised domain transfer strategy. The first stage focuses on pointwise calibration using a small set of labeled real observations—i.e., observations for which the corresponding parameters $\boldsymbol{\theta}$ are known, whereas the second stage aligns distributions using optimal transport (OT). While effective, this approach is inherently transductive, it requires a batch of test samples for inference, limiting its applicability in scenarios that require inductive inference. In many real-world settings, access to a batch of test-time observation is unrealistic, and the inferred posterior for a given single test input can vary depending on the batch it is embedded in—undermining stability and reproducibility. Additionally, the strict separation between pointwise and distribution-wise alignment may prevent full exploitation of their complementary strengths.

In this work, we propose a new framework, illustrated in Fig. 1, that builds upon elements of RoPE by amortizing the optimal transport (OT) step through a mini-batch unbalanced OT approach [23, 24]. Similarly to RoPE, our method assumes access to a limited set of ground-truth pairs of observations $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$. It features a closed-form solution for the transport plan and offers two key advantages:

- **Inductive Joint Training of Alignment Steps:** Enables end-to-end training of both pointwise and distribution-wise alignment, better leveraging their complementary strengths.
- **Amortised Posterior Estimation:** Provides a scalable, inductive solution for OT-based posterior estimation, eliminating the need to repeatedly access simualtions during inference.

We show that our method, while fully inductive and applicable to individual test samples, achieves competitive—and often superior—performance compared to the transductive RoPE baseline across a range of benchmarks. This includes both synthetic and real-world datasets, with strong results in terms of accuracy and calibration, even in challenging settings such as complex biomarker estimation.

2 Background

2.1 Simulation-Based Inference and Neural Posterior Estimation

We consider a simulator $S : \boldsymbol{\theta} \longrightarrow \mathcal{X}$ that, given parameters $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$, produces simulated data $\boldsymbol{x}_s = S(\boldsymbol{\theta})$, whose likelihood $p(\boldsymbol{x}_s \mid \boldsymbol{\theta})$ is intractable. Simulation-based inference (SBI) methods

sidestep likelihood evaluation by training a neural conditional density estimator $q_\psi(\boldsymbol{\theta} \mid \boldsymbol{x}_s)$ (e.g. a conditional normalizing flow [25]) to approximate the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x}_s)$. Very often, when $\boldsymbol{x}_s$ is high-dimensional, a *neural statistics encoder* (NSE [22]) h_ω is used to obtain a lower-dimensional representation with sufficient information for the inference task. Eventually, in *neural posterior estimation* (NPE), one minimizes the expected negative log-likelihood over simulator draws w.r.t parameters $\boldsymbol{\theta}$:

$$\mathcal{L}_{\mathrm{NPE}}(\psi, \omega) = \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta}),\, \boldsymbol{x}_s \sim p(\cdot \mid \boldsymbol{\theta})} \big[-\log q_\psi(\boldsymbol{\theta} \mid h_\omega(\boldsymbol{x}_s)) \big].$$

Under sufficient expressiveness of the density estimation model q_ψ and the encoder h_ω, while considering access to arbitrarily large simulated data $\mathcal{D}_{\mathrm{SBI}} = \{(\boldsymbol{\theta}_j, \boldsymbol{x}_s^j)\}_{j=1}^{N_{\mathrm{SBI}}}$, the density $p(\boldsymbol{\theta} \mid \boldsymbol{x}_s)$ can be approximated by sampling from $q_{\psi^\star}(\boldsymbol{\theta} \mid h_{\omega^\star}(\boldsymbol{x}_s))$ [26][27].

2.2 Transductive Semi-Supervised Posterior Estimation (RoPE)

In the presence of model misspecification, the simulator-induced posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x}_s)$ may be biased relative to the true real-world posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x}_r)$. Consequently, both NSE $h_{\omega^\star}$ and the NPE $q_{\psi^\star}$, when trained solely on simulated data, may fail to perform reliably on real observations. RoPE [22], on which this work builds, addresses this in two stages.

NSE Fine-tuning. Once the NSE, $h_{\omega^\star}$, and NPE, $q_{\psi^\star}$ from Section 2.1 are trained and fixed, a small set (*calibration set*) of real observations $\boldsymbol{x}_r^i$ and their corresponding known parameters $\boldsymbol{\theta}_i$, is used to adapt the encoder to the domain shift. Each $\boldsymbol{\theta}_i$ is passed through the simulator to obtain the corresponding $\boldsymbol{x}_s^i = S(\boldsymbol{\theta}_i)$ and the calibration set becomes $\mathcal{D}_{calib} = \{\boldsymbol{x}_r^i, \boldsymbol{x}_s^i\}_{i=1}^{N_{calib}}$. Then, g_ϕ, the NSE for the real observations, initialized as $h_{\omega^\star}$, is fine-tuned on $\mathcal{D}_{calib}$ to minimize the mean squared error between $g_\phi(\boldsymbol{x}_r^i)$ and $h_{\omega^\star}(\boldsymbol{x}_s^i)$ for every pair of $\boldsymbol{x}_r^i, \boldsymbol{x}_s^i \in \mathcal{D}_{calib}$. The outcome is a limited (depending on the size of $\mathcal{D}_{calib}$) domain adaptation of the NSE g_ϕ, enabling it to encode representations of real observations in the same latent space as $h_{\omega^\star}(\mathbf{x}_s)$.

Entropic OT coupling. Since NSE fine-tuning relies on a limited calibration set, its ability to generalize to unseen real observations is constrained, leaving residual uncertainty in the domain transfer. RoPE takes this uncertainty into account by coupling real and simulated observations through entropic OT, which computes a soft assignment matrix between embeddings $\{g_{\phi^\star}(\boldsymbol{x}_r^i)\}$ of test observations $\mathcal{D}_{test} = \{\boldsymbol{x}_r^i\}_{i=1}^{N_{test}}$ and embeddings $\{h_{\omega^\star}(\boldsymbol{x}_s^j)\}$ of a fresh simulation set $\mathcal{D}_{OT} = \{\boldsymbol{x}_s^j\}_{j=1}^{N_{OT}}$. The latter then acts as prototypes to which the matched embeddings $\{g_{\phi^\star}(\boldsymbol{x}_r^i)\}$ must be close in an Euclidean sense, similarly to the soft Kmeans algorithm [28]. Specifically, they propose to solve for the following semi-balanced entropic OT problem [29, 30]:

$$\boldsymbol{P}^\star = \arg \min_{\boldsymbol{P} \in \mathcal{B}(N_{test}, N_{OT})} \langle \boldsymbol{P}, \boldsymbol{C} \rangle + \rho\, \mathrm{KL}\big(\boldsymbol{P}^\top \mathbf{1}_{N_{test}} \,\|\, \tfrac{1}{N_{OT}} \mathbf{1}_{N_{OT}}\big) + \gamma\, \langle \boldsymbol{P}, \log \boldsymbol{P} \rangle. \tag{1}$$

where $\boldsymbol{P}$ is constrained row-wise to $\mathcal{B}(N_{test}, N_{OT}) = \{\boldsymbol{P} \in \mathbb{R}_+^{N_{test} \times N_{OT}} | \boldsymbol{P} \mathbf{1}_{N_{OT}} = \frac{\mathbf{1}}{N_{test}} \mathbf{1}_{N_{test}}\}$ and $\boldsymbol{C}$ is the pairwise euclidean distance matrix between both sets of embeddings. The weight ρ encourages prototypes to be matched to a uniform number of test samples, enabling unbalanced OT to accommodate prior misspecification, while γ controls the entropy regularization strength, thereby tuning the method's sensitivity to model misspecification and to uncertainty introduced during encoder fine-tuning on the calibration set. Problem 1 is commonly solved using an iterative bregman projection solver [31, 32, 33]. As detailed in Appendix, it comes down to actualize along iterations t the transport plan following:

$$\boldsymbol{P}^{(t+1)} \leftarrow \mathrm{diag}\big(\frac{\mathbf{1}_{N_{test}}}{N_{test} \boldsymbol{K}^{(t)} \mathbf{1}}\big) \boldsymbol{K}^{(t)} \quad \text{with} \quad \boldsymbol{K}^{(t)} = \exp\left(\frac{-\boldsymbol{C} - \rho \mathbf{1} \log(N_{OT} \boldsymbol{P}^{(t)\top} \mathbf{1})^\top}{\gamma}\right) \tag{2}$$

Finally, [22] shows that the calibrated posterior for each $\boldsymbol{x}_r^i$ can be approximated by marginalizing over $\boldsymbol{x}_s$, resulting in the following posterior:

$$\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_r^i) := \sum_{j=1}^{N_{OT}} \alpha_{ij}\, q_{\psi^\star}(\boldsymbol{\theta} \mid h_{\omega^\star}(\boldsymbol{x}_s^j)) \quad \text{with} \quad \alpha_{ij} = N_{test}\, P_{ij}^\star. \tag{3}$$

3 Methods

RoPE, while offering robust posterior estimation, requires computing the OT coupling over the entire test batch $\mathcal{D}_{test}$ at once, rendering the approach inherently **transductive**. This limits its ability to generalize to unseen observations without re-computing the transport plan.

We propose a new Framework for Robust Inductive domain transfer in misspecified Simulation-Based Inference named FRISBI. It relies on a unified workflow that achieves a joint distribution-level and point-wise alignment while enabling **inductive** inference, thereby extending the solution to misspecified SBI beyond the limitations described above. The encoder g_ϕ is first trained using a joint objective that combines a variant of entropic OT admitting closed-form solutions, with a supervised calibration loss, as detailed in Section 3.1. Then to avoid the need for accessing simulations at test time, we further amortize this solution using the inductive strategy presented in Section 3.2. A complete description of the full pipeline and training procedure is provided in Algorithm 3.2. For clarity, Appendix A includes a summary table describing the different datasets.

3.1 Balancing Unpaired Alignment and Point-wise Domain Transfer

In addition to the data assumed to be available in RoPE [22], we also assume access to a large, **unpaired** dataset of real observations, denoted as $\mathcal{D}_u = \{\boldsymbol{x}_r^i\}_{i=1}^{N_u}$. Furthermore, we can generate a large set of simulations, separated from the one used to train the NPE, forming the dataset $\mathcal{D}_{OT} = \{\boldsymbol{x}_s^j\}_{j=1}^{N_{OT}}$. We propose to learn an encoder g_ϕ that optimizes a joint objective, denoted $\mathcal{L}_{\text{joint}}$, composed of two terms. The first component coincides with the entropic OT objective used in RoPE (see Eq. (1)), with the column-marginal constraint parameter fixed at $\rho = 0$. It defines a coupling between the encoded real samples $g_\phi(\boldsymbol{x}_r^i)$ and the fixed simulated representations $h_{\omega^\star}(\boldsymbol{x}_s^j)$. The second component operates on the calibration set $\mathcal{D}_{\text{calib}}$ and aims to control the deviation of the embeddings of real samples from those of their paired simulated samples. Formally, g_ϕ is trained by solving the following problem:

$$\arg\min_{\phi} \underbrace{\sum_{\substack{\boldsymbol{x}_r \sim D_r \\ \boldsymbol{x}_s \in D_s}} \left[P_{ij} \|g_\phi(\boldsymbol{x}_r^i) - h_{\omega^\star}(\boldsymbol{x}_s^j)\|^2 + \gamma\, P_{ij} \log P_{ij}\right]}_{\text{Entropic OT}} + \lambda \underbrace{\sum_{\substack{\boldsymbol{x}_r, \boldsymbol{x}_s \\ \in D_{calib}}} \|g_\phi(\boldsymbol{x}_r^i) - h_{\omega^\star}(\boldsymbol{x}_s^i)\|^2}_{\text{Supervised Loss}}, \quad (4)$$

where $\boldsymbol{P} \in \mathcal{B}(N_u, N_{\mathbf{OT}})$ is an optimal coupling between both distributions and (γ, λ) are regularization hyperparameters. We stress that for the supervised loss, all paired samples $(\boldsymbol{x}_r, \boldsymbol{x}_s)$ in the calibration set $\mathcal{D}_{calib}$ are taken. Whereas for the OT loss, the set $\mathcal{D}_r$ is a combination of a batch $\mathcal{B}_t$ sampled from the unpaired dataset $\mathcal{D}_u$ and samples from the calibration set $\mathcal{D}_{calib}$, while the set $\mathcal{D}_s$ consists of the entire simulation dataset $\mathcal{D}_{OT}$ as well as simulated samples $(\boldsymbol{x}_s)$ from the calibration set $\mathcal{D}_{calib}$. Intuitively, when calibration pairs are accurate (i.e., they share the exact same latent parameters $\boldsymbol{\theta}$), minimizing the supervised loss on the calibration set tends to sharpen the transport plan, resulting in alignment between the OT and supervised objectives. In contrast, when calibration pairs are noisy or mismatched, the two objectives may conflict, allowing the OT term to compensate for uncertainties in the calibration set.

We specifically enforce $\rho = 0$ in the entropic OT objective as the resulting optimization problem w.r.t $\boldsymbol{P}$ is naturally well-suited for inductive learning. Indeed, one can see that setting $\rho = 0$ in Eq. (1), implies that this problem admits a closed-form solution $\boldsymbol{P}^\star = \text{diag}(\frac{\mathbf{1}_{N_u}}{N_u \boldsymbol{K}\mathbf{1}})\boldsymbol{K}$ where $\boldsymbol{K} = e^{-\boldsymbol{C}/\gamma}$ and $\boldsymbol{C}$ is the pairwise euclidean distance matrix between embeddings (see also [31, Proposition 1]). This leads to the efficient stochastic gradient descent (SGD) algorithm described in Stage 1 of Algorithm 3.2, which alternates between computing embeddings for real observations and independently computing the corresponding closed-form couplings for each embedding. Setting $\rho = 0$ enables this closed-form computation, requiring N_{OT} operations per sample in D_u, each involving a Euclidean distance in $\mathbb{R}^d$, for an overall complexity of $\mathcal{O}(d)$. In the mini-batch setting, this yields a per-step complexity of $B_t N_{\text{OT}} d$, where B_t is the batch size. Remark that an analogous strategy could be applied when $\rho > 0$, replacing the closed-form computation with iterative updates of Eq. 2 until convergence, as typically done in mini-batch OT [23, 24]. However, this approach is more computationally demanding and prone to bias, with high sensitivity to batch size. In addition,

RoPE, which uses $\rho > 0$ and cannot operate on mini-batches, scales as $\mathcal{O}(\log(N_u N_{\text{OT}}) N_u N_{\text{OT}} d)$, where N_u is the total number of real observations. These considerations further motivate our choice of $\rho = 0$ for improved efficiency and scalability.

Finally, once the model is trained, the posterior mixture coefficients $\alpha_{ij} = N_u P^\star_{ij}$ for a new test sample $\boldsymbol{x}_r^{test}$ can be directly computed by evaluating $C_{test,j}$ using the trained encoder $g_{\phi^\star}$ and the transport plan $\boldsymbol{P}^\star$, obtained in closed form. This yields the posterior mixture as defined in RoPE (Eq. 3).

3.2 Amortization of OT-based Posterior Estimation

Although the pipeline described in Section 3.1 enables inductive posterior estimation, it still relies on access to the same simulations used to train the loss in Eq. 4, at test time. To mitigate that, we propose to fit a conditional normalizing flow (cNF) q_ξ that approximates the OT-based posterior mixture, conditioned directly on the real observation embeddings $g_{\phi^\star}(\boldsymbol{x}_r)$.

To fit q_ξ, we maximize its expected log density under the target mixture,

$$\arg\max_{\xi} \; \mathbb{E}_{\boldsymbol{\theta} \sim p_{\text{target}}(\boldsymbol{\theta}|\boldsymbol{z})} \big[\log q_\xi(\boldsymbol{\theta} \mid \boldsymbol{z})\big].$$

where $p_{target}(\boldsymbol{\theta}|\boldsymbol{z})$ is computed as in Eq. 3. By the linearity of expectation, this is equivalent to

$$\sum_{j=1}^{N} \alpha_{ij} \, \mathbb{E}_{\boldsymbol{\theta} \sim q_{\psi^\star}(\boldsymbol{\theta}|h_{\omega^\star}(\boldsymbol{x}_s^j))} \big[\log q_\xi(\boldsymbol{\theta} \mid \boldsymbol{z})\big].$$

In practice, we approximate each inner expectation by drawing K samples $\{\boldsymbol{\theta}^{(j,k)}\}_{k=1}^K$ from $q_{\psi^\star}(\boldsymbol{\theta} \mid h_{\omega^\star}(\boldsymbol{x}_s^j))$, and train q_ξ with the set of unpaired real observations $\mathcal{D}_u$, and simulations $\mathcal{D}_{OT}$ on $\mathcal{L}_{flow}$:

$$\arg\min_{\xi} \; -\frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \Big[\frac{1}{K} \sum_{j=1}^{N_{OT}} \alpha_{ij} \sum_{k=1}^{K} \log q_\xi\big(\boldsymbol{\theta}^{(j,k)} \mid \boldsymbol{z}\big) \Big]. \tag{5}$$

During inference, for a given test embedding $\boldsymbol{z} = g_{\phi^\star}(\boldsymbol{x}_r^{test})$, $p(\boldsymbol{\theta} \mid \boldsymbol{x}_r)$ can be approximated by sampling directly from $q_{\xi^\star}(\boldsymbol{\theta} \mid \boldsymbol{z})$ without requiring any access to simulations.

Training Procedure

Datasets: $\mathcal{D}_u = \{\boldsymbol{x}_r^i\}_{i=1}^{N_u}, \quad \mathcal{D}_{OT} = \{\boldsymbol{x}_s^j\}_{j=1}^{N_{OT}}, \quad \mathcal{D}_{calib} = \{(\boldsymbol{x}_r^i, \boldsymbol{x}_s^i)\}_{i=1}^{N_{calib}}$

Trained Models: NSE $h_{\omega^\star}$, NPE $q_{\psi^\star}$

Stage 1: Joint supervised & OT Training

$\boldsymbol{w}_j \leftarrow h_{\omega^\star}(\boldsymbol{x}_s^j) \;\; \forall \boldsymbol{x}_s^j \in \mathcal{D}_{\mathcal{OT}}$
$\boldsymbol{w}_{ic} \leftarrow h_{\omega^\star}(\boldsymbol{x}_s^i) \; \forall \boldsymbol{x}_s^i \in \mathcal{D}_{calib}$
for $e = 1$ to epochs **do**
  for batch $\mathcal{B}_t : \{\boldsymbol{x}_r^i\}_{i \in \mathcal{B}_t}$ from $\mathcal{D}_u$ **do**
    $\boldsymbol{z}_i \leftarrow g_\phi(\boldsymbol{x}_r^i) \; \forall i \in \mathcal{B}_t$
    $\boldsymbol{z}_{ic} \leftarrow g_\phi(\boldsymbol{x}_r^i), \; \forall i \in \mathcal{D}_{calib}$
    $P_{ij} = \frac{1}{|\mathcal{B}_t|} \frac{\exp(-\|\boldsymbol{z}_i - \boldsymbol{w}_j\|^2/\gamma)}{\sum_j \exp(-\|\boldsymbol{z}_i - \boldsymbol{w}_j\|^2/\gamma)}$
    Compute $\mathcal{L}_{joint}$ 4 $\forall i \in \mathcal{B}_t, ic, j$
    Update ϕ by gradient step

Stage 2: Conditional NF Amortization

$\mathcal{Z} : \boldsymbol{z}_i \leftarrow g_{\phi^\star}(\boldsymbol{x}_r^i) \; \forall \boldsymbol{x}_r^i \in \mathcal{D}_u$
$\alpha_{ij} = \frac{\exp(-\|\boldsymbol{z}_i - \boldsymbol{w}_j\|^2/\gamma)}{\sum_j \exp(-\|\boldsymbol{z}_i - \boldsymbol{w}_j\|^2/\gamma)} \; \forall \boldsymbol{z}_i, \boldsymbol{w}_j$
for $e = 1$ to epochs **do**
  for batch $\mathcal{B}_t : \{\boldsymbol{z}_i\}_{i \in \mathcal{B}_t}$ from $\mathcal{Z}$ **do**
    Sample $\boldsymbol{\theta}_{j,k} \sim q_{\psi^\star}(\boldsymbol{\theta} \mid \boldsymbol{w}_j)$
    Compute $\mathcal{L}_{flow}$ 5, $\forall i \in \mathcal{B}_t, j, k$
    Update ξ by gradient step

Inference: $\boldsymbol{z} = g_{\phi^\star}(\boldsymbol{x}_r), \quad p(\boldsymbol{\theta} \mid \boldsymbol{x}_r) \approx q_{\xi^\star}(\boldsymbol{\theta} \mid \boldsymbol{z})$ by sampling $\boldsymbol{\theta} \sim q_{\xi^\star}(\boldsymbol{\theta} \mid \boldsymbol{z})$

4 Experiments

4.1 Benchmarks

We evaluated our proposed approach on four benchmarks: a synthetic one, two real but controlled ones, and one complex real-world benchmark. In the synthetic setting, real observations are emulated using a more complex simulator. The two controlled benchmarks involve data sampled from real systems with experimental control. The final benchmark contains real-world observations collected "in the wild," with no control over sample distributions or nuisance variable variations. The first three benchmarks were also used to evaluate RoPE [22], the main baseline method.

Pendulum. A widely-used synthetic test case in hybrid modeling and simulation-based inference literature [13, 12, 14]. The simulator models the displacement of an ideal, frictionless pendulum, determined by its natural frequency ω_0 and initial angle ϕ_0. To emulate real observations, we use a damped pendulum model that introduces friction into the system. The damping is controlled by a friction coefficient $\alpha \in \mathbb{R}^+$. The parameters we aim to infer are $\boldsymbol{\theta} = \{\omega_0 \in [\frac{\pi}{10}, \pi],\ \phi_0 \in [-\pi, \pi]\}$.

Causal Chambers [34]. Two real, controlled datasets collected from experimental rigs—a wind tunnel and a light tunnel—with adjustable parameters. In the wind tunnel, the target parameter is the hatch opening angle $\boldsymbol{\theta} = \{H \in [0, 45°]\}$. We adopt model A2C3 from [34] as the simulator, which captures pressure dynamics and hatch mechanics, while simplifying aerodynamics and omitting sensor noise, actuator delays, and environmental effects. In the light tunnel, the parameters are the RGB light intensities and a polarizer attenuation factor: $\boldsymbol{\theta} = \{R, G, B \in [0, 255],\ \alpha \in [0, 1]\}$. We use model F3, which simulates photodiode and camera responses under varying exposure and gain, but ignores optical aberrations and sensor noise.

Real Hemodynamics Data. This benchmark uses a subset [6, 35] of the MIMIC-II dataset [36], comprising 350 patients who underwent thermodilution—a procedure estimating cardiac output (CO) via cold fluid injection and downstream temperature measurement. Each patient has arterial blood pressure (ABP) signals aligned with CO readings, yielding ~ 2200 valid ABP segments with corresponding CO values. For simulation, we use OpenBF [37], a validated 1D cardiovascular flow simulator supporting fast, multiscale finite-volume simulations. The estimated parameters are $\boldsymbol{\theta} = \{\text{HR}, \text{CO}\}$, where HR is heart rate, obtained from ECG measurements. The empirical means and standard deviations of HR and CO in the dataset are $(87, 12)$ beats/min and $(5.1, 1.6)$ L/min respectively. In contrast, the CO from the simulations is derived by the known connection $\text{CO} = \text{HR} \times \text{SV}$, where SV is the stroke volume. HR and SV are sampled from uniform distributions $\text{U}(50, 150)$ beats/min and $\text{U}(40, 140)$ L/beat respectively. This yields CO in the range of $[2, 20]$ L/min. This use case is inherently affected by label noise, since the pairing of HR and CO values with observed pulse waves depends on temporal alignment through patients' electronic records, which is prone to misalignments.

4.2 Experimental design

Baselines. The primary method we compare against is RoPE [22]. However, a direct comparison on the same test set is unfair, as RoPE is purely transductive. Nevertheless, we include this setting as a baseline, denoted **RoPE full test**. For a fairer inductive comparison, we introduce a single-sample variant of RoPE: for each test point $\boldsymbol{x}_r^{\text{test}}$, we add it to the unpaired training set $\mathcal{D}_u$, compute the OT coupling with simulations $\mathcal{D}_{OT}$, and estimate the posterior using the resulting plan. This is repeated per test point, and we denote this baseline as **RoPE single sample**. We also compare against baselines from [22], including **NPE** (see Section 2.1), applied directly to real observations without domain transfer. To assess the role of the calibration set, we also include an unsupervised domain adaptation (**UDA**) NPE baseline following [38]. In addition, we include OT-only baselines that do not adapt the embedding space. Here, the fixed encoder $h_{\omega^\star}$ is applied to both real and simulated samples, and OT is computed in this space. The full-test variant is denoted **OT-only (full test)**, and the single-sample variant as **OT-only (single sample)**. For all experiments involving OT, we use solvers from the POT Python library [39, 40].

Another baseline is **finetune-only**, where the finetuned encoder $g_{\phi^\star}$ is applied to test samples, and NPE is used to estimate the posterior directly from $g_{\phi^\star}(\boldsymbol{x}_r)$, without OT-based mixing.

Finally, we include two additional baselines: the upper bound **SBI**, where NPE is trained and tested on simulations, and the **prior** estimator.

Metrics. We evaluate our method using the same metrics as in [22], which assess two critical aspects of posterior estimation: accuracy and calibration.

The first metric is the **log-posterior probability (LPP)**, defined as the average log-likelihood of the true parameter values under the estimated posterior distribution. It measures how much density the true parameters receive in their estimated posteriors, effectively capturing the sharpness and accuracy of the model's predictions. Higher LPP indicates a better match between the estimated posterior and true value.

The second metric is the **average coverage area under the curve (ACAUC)**, which reflects how well the model's credible intervals align with the true parameter coverage. It provides a measure of calibration by comparing the fraction of true parameters falling within the estimated credible intervals at different confidence levels. In all experiments, credible intervals are obtained by drawing 1000 samples from the corresponding approximate posterior distribution. A perfectly calibrated model has an ACAUC of zero, while positive values indicate overconfident estimates and negative values indicate underconfident estimates. For a detailed mathematical definition, we refer the reader to [22].

4.3 Results

In this section, we present the key findings of our experiments. A detailed description of the implementation can be found in the supplementary material.

4.3.1 Performance Across Different Calibration Set Sizes

We evaluate methods that rely on calibration data using 5-fold cross-validation. In each fold, a different, randomly sampled subset of the calibration data is used for training and validation, with independent random initialization of model weights. This approach captures both data variability and the effects of random initialization, providing a robust assessment of performance.

We evaluate calibration set sizes of 10, 50, 200, and 1000 samples, while keeping the test set size fixed at 1000 samples across all benchmarks. The simulation set, $\mathcal{D}_{SBI}$, used to train both the NSE and NPE, as well as the unpaired real observations set, $\mathcal{D}_u$, each contain 1000 samples. Similarly, the set of simulations used for the OT in RoPE and our proposed joint training, $\mathcal{D}_{OT}$, also consists of 1000 samples. For training the amortised posterior estimator (Section 3.2), we exclusively use $\mathcal{D}_u$. Following the guidelines in [22], the entropy regularization weight γ is set to 0.5 for all baselines involving OT, including our joint training approach. Finally, since this section focuses on the setting where both simulations and real observations are drawn from the same prior distribution $p(\boldsymbol{\theta})$, we use balanced Sinkhorn OT for RoPE, effectively setting ρ in eq. 1 to a very high value.

The results, including mean scores and standard deviations computed across folds, are presented in Fig. 2. Our method (solid red line) consistently outperforms the transductive single-sample RoPE baseline (solid orange line) in terms of LPP, while maintaining ACAUC close to zero across most calibration set sizes. This indicates that it achieves a robust **inductive** parameter estimation compared to RoPE. With larger calibration sets, our method matches or even exceeds the performance of the full-test RoPE (solid green line), which has access to the entire test set—highlighting the effectiveness of combining OT-based domain transfer with supervised calibration. It also highlights the importance of incorporating a calibration set, even a small one, as the UDA baseline completely fails under significant misspecification, consistent with the observations in [41]. However, as the calibration set size increases, our method—like the fine-tuning baselines—shows a decline in confidence (lower ACAUC). In contrast, the RoPE variants are less affected by this trend, suggesting that our joint training approach increasingly relies on the supervised loss over the OT loss as more calibration data becomes available. We also evaluate performance on two additional benchmarks exhibiting more moderate misspecification (Appendix C.1), also considered in [22]. Similar to RoPE, FRISBI shows no significant advantage over baseline methods in simpler and minimally misspecified setting.

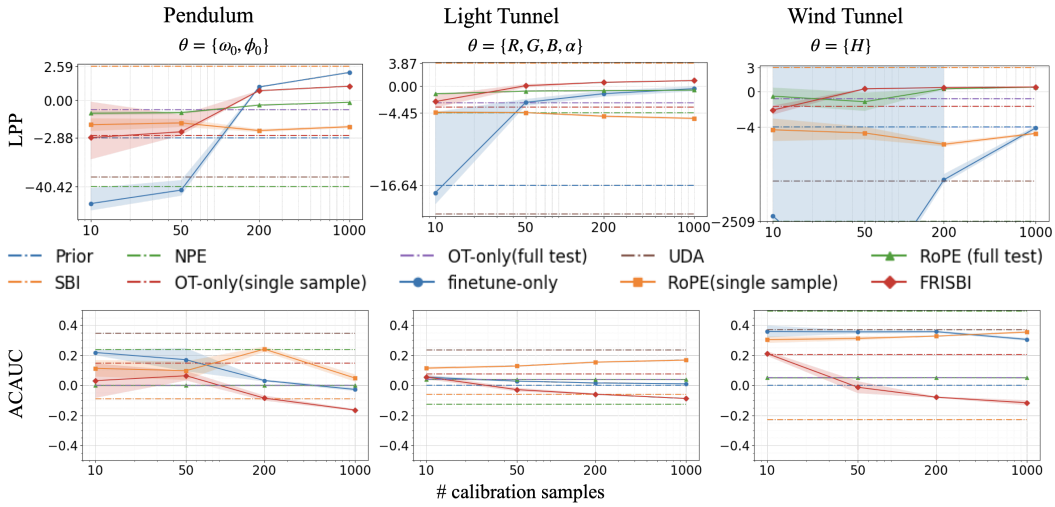

Figure 2: **Results across different calibration set sizes.**. The **top** row displays performance in terms of LPP (↑) while the **bottom** one is the calibration metric ACAUC (→ 0 ←). The horizontal axis indicates the sample size while the vertical one is the metric value. Baselines that do not rely on a calibration set are represented by fixed horizontal dashed lines for easier comparison.

4.3.2 Ablation

The importance of joint training. As described in Section 3, our full pipeline consists of two main components: joint distribution and point-to-point alignment, which we refer to as **joint training only**, and the posterior-mixture amortization step, which we refer to as **amortized solution only**. To assess the individual contributions and necessity of these components, we evaluate their performance separately and in combination. For the **amortized solution only** baseline, we train the amortization step described in Section 3.2 directly to approximate the posteriors estimated by RoPE, using the unpaired real set $\mathcal{D}_u$ and the OT simulations set $\mathcal{D}_{OT}$. For the **joint training only** baseline, we evaluate the posteriors obtained through our joint training approach described in Section 3.1, without the additional amortization step. Finally, we compare these to the **full pipeline**, which combines both components as described in Section 3.

The mean scores and standard deviations for each variant are reported in Fig. 3. Overall, the joint training approach and the full pipeline achieve comparable performance in terms of LPP and ACAUC, both consistently outperforming the amortized version of the transductive RoPE solution (blue). This suggests that the performance gains in our pipeline are largely attributable to the joint training strategy. The full pipeline offers the added benefit of not requiring access to the simulations $\mathcal{D}_{OT}$ at test time. Notable differences arise in specific cases: the full pipeline (green) performs better in the low-calibration regime of the light tunnel benchmark (middle of Fig. 3), suggesting it may act as a regularizer for the posterior mixture when calibration data is limited, while joint training alone (orange) outperforms the full pipeline in the pendulum benchmark. In the latter, we observed that in one fold, the cNF model used to amortize the posterior mixture failed to reduce its training loss (Eq. 5), indicating difficulties in learning the posterior. This could potentially be mitigated by using a more expressive cNF model.

Hyperparameter sensitivity analysis. In Appendix C.2, we present a sensitivity analysis of the hyperparameters γ and λ on the Light Tunnel benchmark. As in RoPE, a larger γ is beneficial when the calibration set is small, producing a more diffuse OT coupling. With larger calibration sets (e.g., 1,000 samples), γ can be reduced to obtain a sharper coupling. The effect of λ is evaluated under 10% label noise, since noise influences the weighting of the supervised objective. Higher λ values decrease confidence as the calibration set grows, while smaller values yield better LPP scores, suggesting that an adaptive tuning of λ based on noise level and data size may be advantageous.

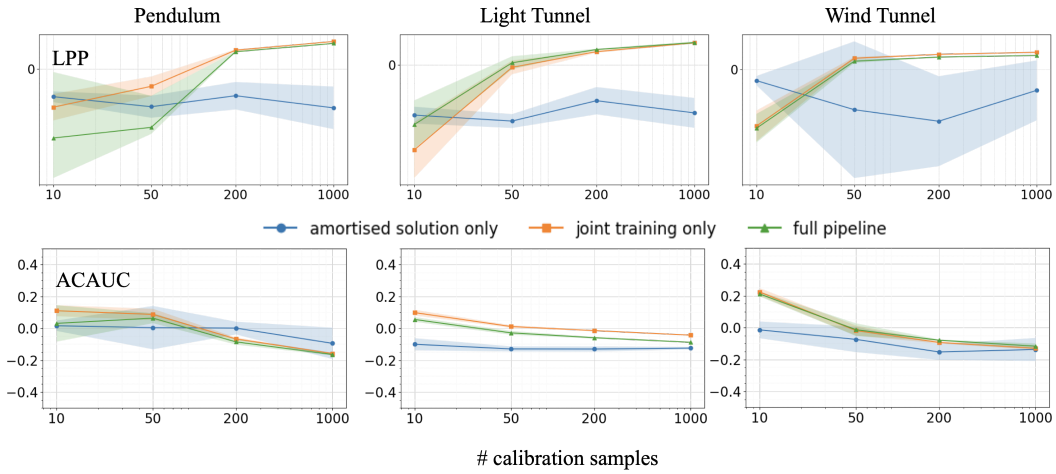

Figure 3: **Ablation Analysis.** Comparison of joint training only (3.1), solution amortization only (3.2), and the full pipeline. The horizontal axis shows the number of calibration samples, while the vertical axis represents the LPP(↑, top) and ACAUC(→ 0 ←, bottom) scores.

4.3.3 Label Noise Robustness

In a more realistic scenario, the calibration set known parameters $\boldsymbol{\theta}$ are themselves measurements and thus inherently noisy. For example, in the case of CO (cardiac output) labels, both the measurement procedure and the temporal alignment with the corresponding ABP (arterial blood pressure) signal can introduce noise. This means that the resulting simulated observations $\boldsymbol{x}_s$ may also be inaccurately paired with the real observations, potentially affecting the quality of the calibration set.

To evaluate the robustness of our proposed method to label noise, we add Gaussian noise to the calibration labels, corresponding to 1% and 10% of the parameter range of the assumed prior in the light tunnel benchmark. The models are trained using these noisy labels, while the test set remains clean for evaluation, as in the previous subsections. In this experiments we still consider the balanced OT settings (no prior mismatch) and keep γ at 0.5. In Fig. 4, our method (solid red line) almost consistently achieves higher LPP scores than the single-sample RoPE baseline (solid orange line), indicating greater robustness to label noise due to its inductive nature. While it is somewhat more sensitive to high noise levels compared to the full-test RoPE baseline (solid green line), this sensitivity diminishes with larger calibration sets. For example, with 200 calibration samples, the performance gap noticeably narrows.

In the CO estimation experiment, we explicitly account for the increased uncertainty in the calibration set by setting the entropic regularization parameter γ to a higher value of 1.5 for the RoPE and OT baselines and in our proposed method. Additionally, in the RoPE baselines we use unbalanced OT settings to handle known prior mismatches, as suggested in [22]. The calibration set size in these experiments is 200, and we report results across 5 random splits to assess robustness. It is important to note that clean test-set $\boldsymbol{\theta}$ labels are not available in this setting. As shown in Fig. 5, Our method (in red) significantly outperforms all baselines, including RoPE, in terms of LPP, while showing slight overconfidence (ACAUC < 0.2), highlighting the inference benefits of our amortized inductive framework in real complex settings.

5 Discussion and Conclusion

In this work, we introduce an amortized and inductive posterior estimator for misspecified simulation-based inference. Our method leverages mini-batch optimal transport to enable joint training over both unpaired and small paired calibration sets. The final posterior, approximated by an OT-based mixture, is amortised by a conditional normalizing flow, eliminating the need for additional transport computations or access to simulations at test time – a key limitation of transductive approaches like RoPE. Our approach demonstrates competitive performance across a

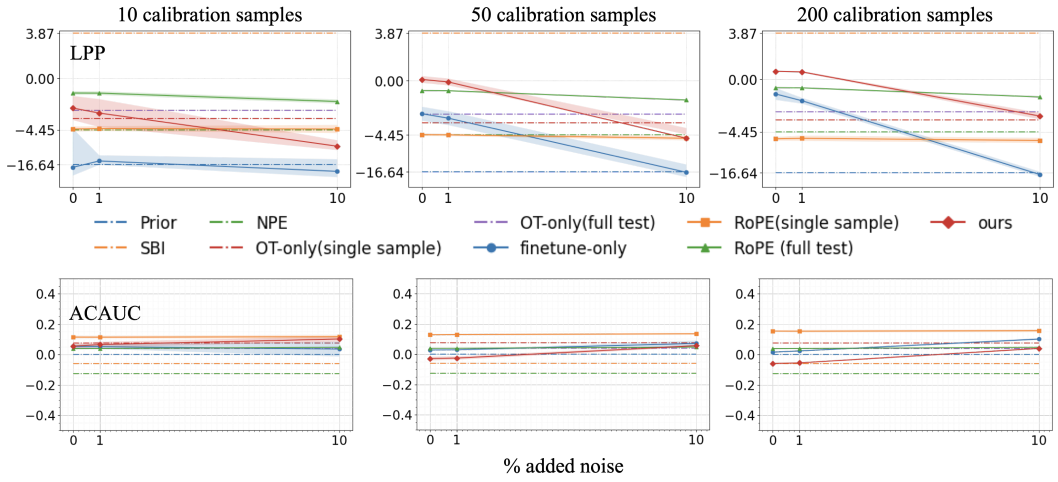

Figure 4: **Label Noise Robustness.** Impact of increasing label noise on performance, measured by LPP (↑, top) and ACAUC (→ 0 ←, bottom), across three calibration set sizes (noted above each panel) on the light tunnel benchmark. The horizontal axis represents the noise rate, while the vertical axis shows the metric score.

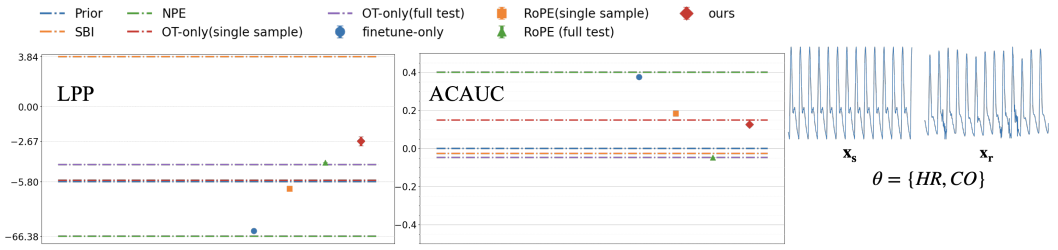

Figure 5: **Cardiac Biomarker Estimation.** Performance comparison across all baselines for heart rate (HR) and cardiac output (CO) estimation, using a calibration set of 200 samples. On the right, an example of real and simulated arterial pulse waveforms is shown.

range of benchmarks, including both synthetic and complex real-world datasets.

Dimensionality and scalability. While our experiments focus on relatively low-dimensional cases representative of many real-world scenarios, our framework naturally extends to higher-dimensional settings. Two main challenges arise: (i) the mismatch between simulated and real embeddings tends to increase with dimensionality, requiring larger calibration sets, and (ii) the embedding dimension needed to capture sufficient statistics typically grows with the parameter dimension, complicating the coupling. Following Chen et al. [42], we use an overparameterized embedding dimension of 16 to mitigate this issue. The joint optimization of supervised and OT objectives further alleviates the limitations of small calibration sets by exploiting unlabeled data. As discussed in Section 3.1, FRISBI is expected to scale more efficiently than RoPE in high-dimensional settings.

Limitations and future directions. Sensitivity to label noise, as observed in our experiments, suggests that training data quality significantly impacts posterior accuracy. While this could be partially mitigated with a higher entropy regularization weight, more adaptive joint loss formulations should be explored and specifically adaptive updates of the supervised loss weight λ. Additionally, active learning or uncertainty-aware sampling strategies could be investigated to guide calibration set selection under a fixed budget. Finally, the training process of our proposed method still involves two separate stages: the joint OT-supervised training and the subsequent amortization of the induced posterior mixture. A more holistic alternative could involve learning the transport mapping directly through a neural OT framework, potentially integrating the matching and inference stages into a single unified process.

References

[1] Shamil Maksyutov, Tomohiro Oda, Makoto Saito, Rajesh Janardanan, Dmitry Belikov, Johannes W Kaiser, Ruslan Zhuravlev, Alexander Ganshin, Vinu K Valsala, Arlyn Andrews, et al. A high-resolution inverse modelling technique for estimating surface co 2 fluxes based on the nies-tm–flexpart coupled transport model and its adjoint. *Atmospheric Chemistry and Physics Discussions*, 2020:1–33, 2020.

[2] Hamed Sahranavard, Ali Mohtashami, Ehsan Mohtashami, and Abolfazl Akbarpour. Inverse modeling application for aquifer parameters estimation using a precise simulation–optimization model. *Applied Water Science*, 13(2):58, 2023.

[3] A Dolev, S Davis, and I Bucher. Noncontact dynamic oscillations of acoustically levitated particles by parametric excitation. *Physical Review Applied*, 12(3):034031, 2019.

[4] Pei Pei, Yongbo Peng, and Canxing Qiu. An improved semi-active structural control combining optimized fuzzy controller with inverse modeling technique of mr damper. *Structural and Multidisciplinary Optimization*, 65(9):272, 2022.

[5] Salvatore Sessa, Nicoló Vaiana, Massimo Paradiso, and Luciano Rosati. An inverse identification strategy for the mechanical parameters of a phenomenological hysteretic constitutive model. *Mechanical Systems and Signal Processing*, 139:106622, 2020.

[6] JX Sun, AT Reisner, M Saeed, and RG Mark. Estimating cardiac output from arterial blood pressurewaveforms: a critical evaluation using the mimic ii database. In *Computers in Cardiology, 2005*, pages 295–298. IEEE, 2005.

[7] Arun J Sanyal, Laurent Castera, and Vincent Wai-Sun Wong. Noninvasive assessment of liver fibrosis in nafld. *Clinical Gastroenterology and Hepatology*, 21(8):2026–2039, 2023.

[8] Ida Marie Hauge-Iversen, Einar S Nordén, Arne Olav Melleby, Linn Espeland, Lili Zhang, Ivar Sjaastad, and Emil Knut Stenersen Espe. Non-invasive estimation of left ventricular chamber stiffness using cardiovascular magnetic resonance and echocardiography. *Journal of Cardiovascular Magnetic Resonance*, 27(1):101849, 2025.

[9] Lennart Ljung. *System identification*. Univ., 1995.

[10] Curtis R Vogel. *Computational methods for inverse problems*. SIAM, 2002.

[11] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[12] Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel De Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124012, 2021.

[13] Naoya Takeishi and Alexandros Kalousis. Physics-integrated variational autoencoders for robust and interpretable generative modeling. *Advances in Neural Information Processing Systems*, 34:14809–14821, 2021.

[14] Antoine Wehenkel and Jens Behrmann. Robust hybrid learning with expert augmentation. *Transaction Machine Learning Research*, 2023.

[15] Johann Brehmer. Simulation-based inference in particle physics. *Nature Reviews Physics*, 3(5):305–305, 2021.

[16] Meysam Hashemi, Anirudh N Vattikonda, Jayant Jha, Viktor Sip, Marmaduke M Woodman, Fabrice Bartolomei, and Viktor K Jirsa. Simulation-based inference for whole-brain network modeling of epilepsy using deep neural density estimators. *medRxiv*, pages 2022–06, 2022.

[17] Jan-Matthis Lückmann. *Simulation-based inference for neuroscience and beyond*. PhD thesis, Universität Tübingen, 2022.

[18] Nicholas Tolley, Pedro LC Rodrigues, Alexandre Gramfort, and Stephanie R Jones. Methods and considerations for estimating parameters in biophysically detailed neural models with simulation based inference. *PLOS Computational Biology*, 20(2):e1011108, 2024.

[19] Grace Avecilla, Julie N Chuong, Fangfei Li, Gavin Sherlock, David Gresham, and Yoav Ram. Neural networks enable efficient and accurate simulation-based inference of evolutionary parameters from adaptation dynamics. *PLoS biology*, 20(5):e3001633, 2022.

[20] Patrick Cannon, Daniel Ward, and Sebastian M Schmon. Investigating the impact of model misspecification in neural simulation-based inference. *arXiv preprint arXiv:2209.01845*, 2022.

[21] Marvin Schmitt, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T Radev. Detecting model misspecification in amortized bayesian inference with neural networks. In *DAGM German Conference on Pattern Recognition*, pages 541–557. Springer, 2023.

[22] Antoine Wehenkel, Juan L. Gamella, Ozan Sener, Jens Behrmann, Guillermo Sapiro, Marco Cuturi, and Jörn-Henrik Jacobsen. Addressing misspecification in simulation-based inference through data-driven calibration. *arXiv preprint arXiv:2405.08719*, 2024.

[23] Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021.

[24] Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR, 2021.

[25] Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

[26] George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, pages 1028–1036, 2016.

[27] Jan-Matthis Lueckmann, Pedro Gonçalves, Gianmaria Bassetto, David Greenberg, and Jakob H. Macke. Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems*, 2017.

[28] Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. *Advances in neural information processing systems*, 25, 2012.

[29] Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed gromov-wasserstein divergence and applications on graphs. In *International Conference on Learning Representations*.

[30] Hugues Van Assel and Randall Balestriero. A graph matching approach to balanced data subsampling for self-supervised learning. In *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice*, 2024.

[31] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[32] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28, 2015.

[33] Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. *Handbook of Numerical Analysis*, 24:407–471, 2023.

[34] Juan L Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence*, pages 1–12, 2025.

[35] Ary L Goldberger, Luis A N Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[36] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii: a public-access intensive care unit database. *Critical care medicine*, 39(5):952–960, 2011.

[37] I. Benemerito, A. Melis, A. Wehenkel, and A. Marzo. openbf: an open-source finite volume 1d blood flow solver. *Physiological Measurement*, 45(12):125002, 2024.

[38] Paxson Swierc, Marcos Tamargo-Arizmendi, Aleksandra Ćiprijanović, and Brian D Nord. Domain-adaptive neural posterior estimation for strong gravitational lens analysis. *arXiv preprint arXiv:2410.16347*, 2024.

[39] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[40] Rémi Flamary, Cédric Vincent-Cuaz, Nicolas Courty, Alexandre Gramfort, Oleksii Kachaiev, Huy Quang Tran, Laurène David, Clément Bonet, Nathan Cassereau, Théo Gnassounou, Eloi Tanguy, Julie Delon, Antoine Collas, Sonia Mazelet, Laetitia Chapel, Tanguy Kerdoncuff, Xizheng Yu, Matthew Feickert, Paul Krzakala, Tianlin Liu, and Eduardo Fernandes Montesuma. Pot python optimal transport (version 0.9.5), 2024.

[41] Lasse Elsemüller, Valentin Pratz, Mischa von Krause, Andreas Voss, Paul-Christian Bürkner, and Stefan T Radev. Does unsupervised domain adaptation improve the robustness of amortized bayesian inference? a systematic evaluation. *arXiv preprint arXiv:2502.04949*, 2025.

[42] Yanzhi Chen, Dinghuai Zhang, Michael Gutmann, Aaron Courville, and Zhanxing Zhu. Neural approximate sufficient statistics for implicit models. *arXiv preprint arXiv:2010.10079*, 2020.

[43] Daniel Ward, Patrick Cannon, Mark Beaumont, Matteo Fasiolo, and Sebastian Schmon. Robust neural posterior estimation and statistical model criticism. *Advances in Neural Information Processing Systems*, 35:33845–33859, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the intro are informative about the proposed approach and the validating experiments

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in all experiments and in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: This is not a theoretical work, but the simplifications that allow our closed-form transport plan solution are reasoned and cited in the methods part and in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our work proposes a new framework and a detailed pipeline of the framework is provided in section 3. In addition, implementation details are provided in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is available for all experiments. Data and data processing is available for the pendulum and causal chambers experiments. For the biomarker experiment it is partially available, as the access to MIMIC-II data set requires credentialing and agreement to terms of use .

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: High level description in the paper, implementation details in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars are reported in all experiments across 5-splits of the training data.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Not in the body of the paper, but in the supplementary material

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The only medical dateset we use is an open one. No expected negative impact.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: It is a technical work that tries to solve a specific technical problem. It does not have a broader societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We mostly used standard open-source Python libraries (e.g., PyTorch, NumPy, scikit-learn, pot) under permissive licenses. We did not redistribute or modify any third-party code or pre-trained models, and all licenses were respected. We used the openBF (Apache 2.0) library for arterial pressure waves simulations and cited it. We used the MIMIC-II dataset, which is distributed under a specific data use agreement (PhysioNet Credentialed Health Data License). Access to the dataset requires credentialing and agreement to terms of use, including proper citation. We obtained access in compliance with these terms and cite the dataset as per the official guidelines

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assests.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human subjects (only existing datasets)

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects (only existing datasets)

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Not an LLM-related work.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Datasets clarification

Table 1: Summary of datasets used in the pipeline.

Dataset	Observations	θ Labels	Usage
D_{SBI}	simulated ($\boldsymbol{x}_s$)	yes	train NPE, NSE
D_{OT}	simulated ($\boldsymbol{x}_s$)	no	train over the OT objective in 3.1, amortization of posterior mixture in 3.2
D_{calib}	real ($\boldsymbol{x}_r$)	yes	train 3.1, MSE objective (and OT)
D_u	real ($\boldsymbol{x}_r$)	no	train 3.1, OT objective, amortization of posterior mixture in 3.2

B Semi-balanced Optimal transport solvers.

We develop next how to solve for semi-balanced entropic optimal transport problems discussed in Sections 2.2 and 3. The overall problem reads as

$$\boldsymbol{P}^\star = \arg\min_{\boldsymbol{P}\in\mathcal{B}(N_{test},N_{OT})} \mathcal{L}(\boldsymbol{P}) := \langle \boldsymbol{P},\, \boldsymbol{C}\rangle + \rho\,\mathrm{KL}\big(\boldsymbol{P}^\top \mathbf{1}_{N_{test}} \,\|\, \tfrac{1}{N_{OT}}\mathbf{1}_{N_{OT}}\big) + \gamma\,\langle \boldsymbol{P}, \log \boldsymbol{P}\rangle. \quad (6)$$

This type of problem can be generally solved using an iterative bregman projection solver [31, 32, 29, 33], or equivalently mirror-descent algorithms following the KL geometry. It comes down to the following steps:

Step 1. Compute the gradient of the objective function

$$\nabla\mathcal{L}(\boldsymbol{P}^{(t)}) = \boldsymbol{C} + \rho\mathbf{1}\log(N_{OT}\boldsymbol{P}^{(t)\top}\mathbf{1})^\top + \gamma\log\boldsymbol{P}^{(t)} \quad (7)$$

step 2. Then for a given learning rate τ, one has to solve the problem

$$\boldsymbol{P}^{(t+1)} \leftarrow \operatorname*{arg\,min}_{\boldsymbol{P}\in\mathcal{B}(N_{test},N_{OT})} \langle\nabla\mathcal{L}(\boldsymbol{P}^{(t)}), \boldsymbol{P}\rangle + \tau KL(\boldsymbol{P}|\boldsymbol{P}^{(t)}) \quad (8)$$

we have

$$\begin{aligned}
&\langle\nabla\mathcal{L}(\boldsymbol{P}^{(t)}), \boldsymbol{P}\rangle + \tau KL(\boldsymbol{P}|\boldsymbol{P}^{(t)})\\
=&\langle \boldsymbol{C} + \rho\mathbf{1}\log(N_{OT}\boldsymbol{P}^{(t)\top}\mathbf{1})^\top + \gamma\log\boldsymbol{P}^{(t)}, \boldsymbol{P}\rangle + \tau\langle\boldsymbol{P}, \log\boldsymbol{P} - \log\boldsymbol{P}^{(t)}\rangle\\
(\text{setting } \gamma=\tau) =&\langle \boldsymbol{C} + \rho\mathbf{1}\log(N_{OT}\boldsymbol{P}^{(t)\top}\mathbf{1})^\top, \boldsymbol{P}\rangle + \gamma\langle\boldsymbol{P}, \log\boldsymbol{P}\rangle\\
=&\langle -\log e^{\frac{-\boldsymbol{C}-\rho\mathbf{1}\log(N_{OT}\boldsymbol{P}^{(t)\top}\mathbf{1})^\top}{\gamma}}, \boldsymbol{P}\rangle + \gamma\langle\boldsymbol{P}, \log\boldsymbol{P}\rangle\\
=&\gamma KL(\boldsymbol{P}|\boldsymbol{K}_\rho^{(t)})
\end{aligned} \quad (9)$$

with $\boldsymbol{K}_\rho^{(t)} = e^{\frac{-\boldsymbol{C}-\rho\mathbf{1}\log(N_{OT}\boldsymbol{P}^{(t)\top}\mathbf{1})^\top}{\gamma}}$. Hence this problem comes down to a KL projection on the set $\mathcal{B}(N_{test}, N_{OT})$ of the Gibbs kernel $\boldsymbol{K}_\rho^{(t)}$. As detailed in Proposition 1 in [31]) this problem admits a close-form solution detailed in Equation 2.

C Additional experiments

C.1 CS and SIR benchmarks

We follow the evaluation protocol of [22], originally introduced by [43].

CS. The cancer-stromal cell simulator models 2D cell growth with three Poisson rate parameters $(\lambda_c, \lambda_p, \lambda_d)$. Each sample includes cell counts and the mean and maximum distance between stromal and nearest cancer cells. Misspecification is induced by removing cancer cells located too close to their parent.

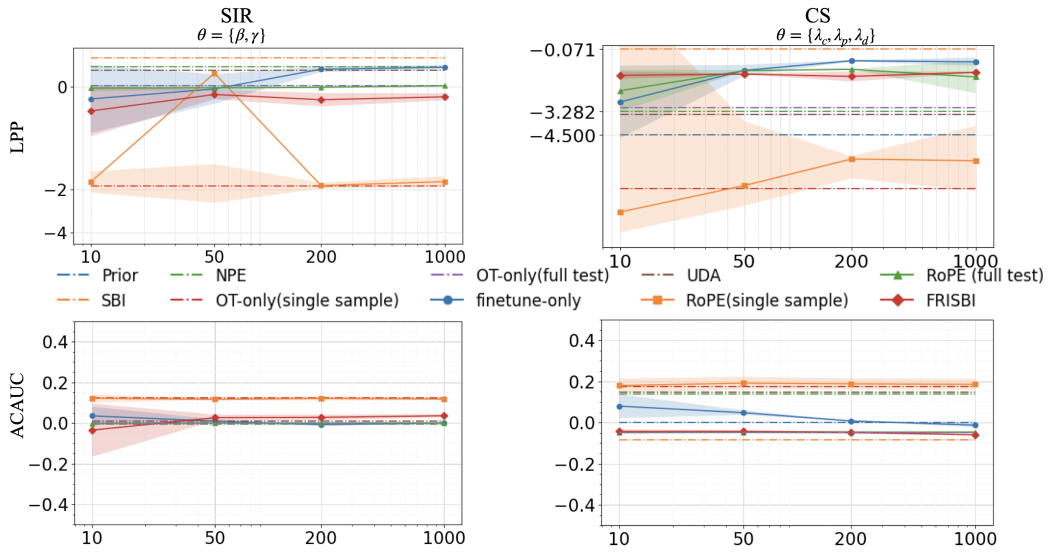

Figure 6: **Minimally misspecified benchmarks (CS and SIR)**. The horizontal axis shows the number of calibration samples, while the vertical axis represents the LPP(↑, top) and ACAUC(→ 0 ←, bottom) scores.

SIR. The stochastic epidemic model simulates infection and recovery dynamics with rates (β, γ). Observations comprise summary statistics of infection counts, timing, and autocorrelation. Misspecification is introduced by delaying weekend infections, adding 5% of them to the following Monday.

We observe that both FRISBI (red) and RoPE underperform relative to the vanilla NPE baseline (dashed green) in the SIR benchmark and achieve performance comparable to the finetuning-only baseline (blue) in the CS case. This suggests that, under minimal misspecification, a standard NPE is sufficient, and in simpler systems, even a small calibration set can adequately capture the mapping between observations and parameters.

C.2 Sensitivity analysis w.r.t hyperparameters γ and λ

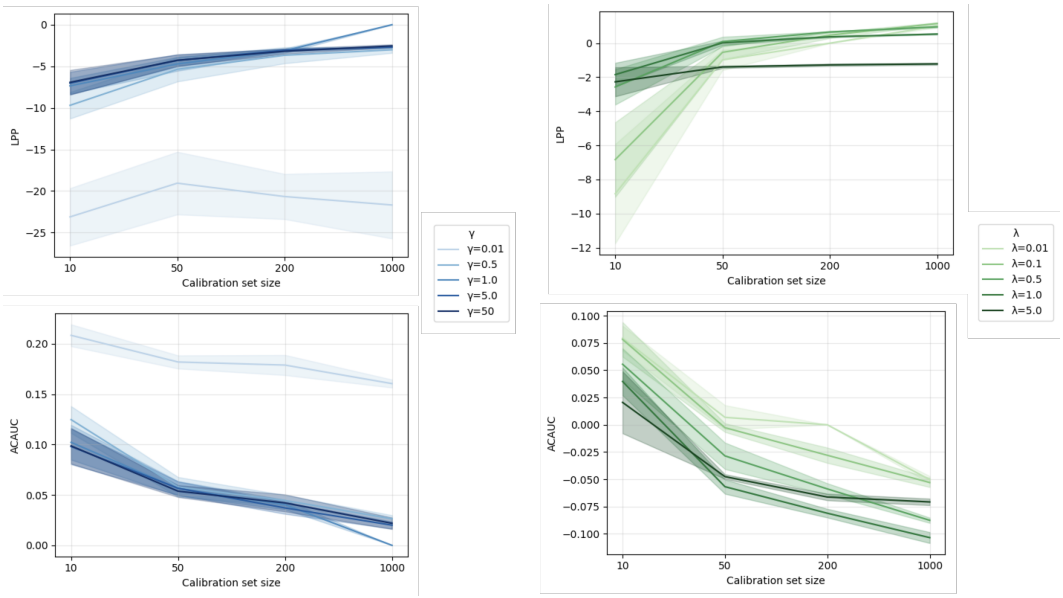

Figure 7: **Sensitivity analysis of hyperparameters γ and λ.** Both experiments are conducted on the Light Tunnel benchmark. The λ sensitivity analysis (right) is performed under a 10% label noise setting. The horizontal axis shows the number of calibration samples, while the vertical axis reports the LPP ($\uparrow$, top) and ACAUC ($\rightarrow 0 \leftarrow$, bottom) scores. The legend indicates the different γ and λ values considered and their corresponding shades.