Identifying informative features of Hodgkin-Huxley models using simulation-based inference

Thesis
submitted in partial fulfilment of the requirements for the degree

Master of Science

Graduate School of Neural Information Processing

Faculty of Science
Faculty of Medicine

Eberhard-Karls-Universität Tübingen

Presented by

Jonas Beck
from Wiesbaden, Germany

Tübingen, 6 October 2021

Statement of Authorship

Thesis Advisor:	Prof. Dr. Jakob Macke Department of Machine Learning in Science
Second Reader:	Prof. Dr. Philipp Berens Department of Data Science for Vision Research

Disclosures:

- I affirm that I have written the dissertation myself and have not used any sources and aids other than those indicated.
- I affirm that I have not included data generated in one of my laboratory rotations and already presented in the respective laboratory report

Tübingen, 06.10.2021

Jonas Beck

Abstract

In recent years, simulation-based inference has emerged as a powerful tool to study biophysical models of neuronal dynamics. Algorithms such as Neural Posterior and Neural Likelihood Estimation make it possible to swiftly and efficiently obtain full posterior estimates of complex mechanistic models, without putting any limits on their design. This enables to identify not just one, but multiple models consistent with experimental observations. For several models, such as the Hodgkin Huxley model, this process depends to a large degree on the specific summary features used to describe the membrane voltage traces. In this thesis we quantify which data features constrain parameter uncertainty for a 9 parameter HH model and a set of 23 common summary statistics. To achieve this with minimal training costs, we devise three different methods and suggest two complementary metrics to measure the constrainment. Finally, we identify which features constrain which parameters for two synthetic and one experimentally recorded voltage traces. We find that sodium and potassium conductances, as well as threshold voltages are the most strongly constrained parameters, while the strongest constraining summary features are membrane resting potential, action potential amplitude and action potential threshold.

Contents

1 Introduction

Biophysical models of neuronal dynamics are an essential tool in studying the mechanisms by which neural tissue controls the flow of information and enables the emergence of complex behavioural patterns [1]. In a constant interplay between computational modelling and gathering of experimental data, these mechanistic models are continually refined in an effort to make ever more accurate predictions and to provide better explanations of observed brain activity [2]. Thanks in part to the steady increase in computing power, the interest in computational models of neural dynamics has only increased in recent years, as the generation of large amounts of synthetic data [3] is now feasible. Coupled with recent advances in Bayesian inference methods for simulator models [4, 5], the challenge of identifying mechanistic models that reproduce electrophysiological data can now be better addressed than ever before [6]. In this context it has been suggested that simulation-based inference (SBI) [4] provides a tool to constrain mechanistic models by data without compromising model design, greatly expediting progress in neuroscience [7]. Hence, it is not just of interest to study these models themselves, but also the exact mechanisms of how certain features of the data constrain them. Identifying and describing the confounding factors for common mechanistic models of neural dynamics can yield insights of how metabolic [8–11], environmental [12–14] and physical constraints [11] shape biophysical processes and subsequently model selection, as well as how differences arise across species [15] or brain areas [14]. In turn, this can help improve both inference and model predictions, further bridging the gap between data- and theory-driven approaches [7].

SBI [4, 16], or likelihood-free inference (LFI) as it is also known, refers to a collection of algorithms to conduct Bayesian inference on mechanistic models which come in the form of "black box" simulators. The term "black box" hereby refers to the fact that their likelihood function $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ cannot be evaluated, though running the simulator can be considered sampling. By leveraging the ability to run the simulator, SBI aims to obtain an estimate of the parameter posterior $p(\boldsymbol{\theta}|\mathbf{x}_o) \propto p(\mathbf{x}_o|\boldsymbol{\theta})p(\boldsymbol{\theta})$, given an observed data vector $\mathbf{x_o}$ and a prior distribution $p(\boldsymbol{\theta})$ in the absence of a tractable likelihood

function $p(\boldsymbol{x} \mid \boldsymbol{\theta})$. A particularly recent skew of methods achieves this by learning a synthetic likelihood [17–19] or posterior function [20–22] from data and conducts inference on the learned surrogate instead. Given enough data, a sufficiently flexible model, q_ϕ, will eventually approximate the posterior distribution $q_\phi(\boldsymbol{\theta}|\mathbf{x_o}) \approx p(\boldsymbol{\theta}|\mathbf{x_o})$ in the vicinity of the observation arbitrarily well[19]. Because there are very little constraints on the function class of q_ϕ, recent publications [18, 19, 22] have shown that deep neural networks can do this task very well. This has lead to "neural SBI methods" like Neural Likelihood Estimation (NLE) [18, 19] and Neural Posterior Estimation (NPE) [20–22]. For a detailed overview of these methods read Cranmer et al. [4].

Arguably the most prominent mechanistic description of single cell neural dynamics is the Hodgkin-Huxley (HH) model [23]. By modelling the electric conductivity in neural tissue as an analogous electric circuit, it has contributed towards a multitude of discoveries [24] in its almost 70 year history and despite its age, remains one of the most widely studied mechanistic descriptions of neural membrane dynamics to date. With such a track-record it is not surprising that a number of algorithms have been proposed for fitting HH models to electrophysiological data [25–29] already. However, obtaining a full posterior estimate has remained difficult [30]. With the success of neural SBI methods for this application being repeatedly demonstrated [7, 21] though, this is starting to change. SBI has now been used in the study of the HH model multiple times, serving as a challenging benchmark to evaluate the performance of novel inference techniques [18, 19], being used to study neural behaviour and dynamics [7, 21] or to conduct stimulus optimisation for retinal neuroprosthetics [31]. Regardless of the intended use case however, a common factor in the constrainment of HH models is the reliance on summary statistics to describe its electrophysiological properties. In the case of Approximate Bayesian Computation (ABC) [20, 32] and classical density estimation-based inference methods [33] this was done mainly to alleviate the impact of the curse of dimensionality [4]. Though, more modern methods, which scale better to higher dimensions can still rely on them to reduce computation costs, isolate specific behaviours or to aid interpretability. The latter is especially true when summary statistics are carefully selected or hand-crafted (i.e. by domain specialists). In any case, the use of such statistics leads to a critical dependence of parameter estimates on the choice of represented data features [34–36] and it is thus important to quantify how individual or combinations of said features impact posterior uncertainty. Understanding how these features constrain model parameters during inference can also lead to new insights about the emergence

of recognisable characteristics in the recordings of membrane voltages, compensatory behaviour and how system dynamics arise from the interaction of system components [15].

It is therefore surprising that the ability of summary features to constrain HH models has so far only been evaluated explicitly in a limited fashion and for very few commonly used hand-crafted summary features [7]. Hence, a more comprehensive study and framework to gauge parameter constrainment of summary features for HH models is still lacking. Such an analysis could not only provide future guidance for the selection of summary statistics during inference in HH models, but also explain how certain features constrain model parameters in the first place. Additionally, it should be enough to conduct this analysis only for one inference method, since conclusions drawn should be equally applicable to other SBI algorithms [16] under the assumption that different procedures converge to the same "true" posterior function when supplied with the same set of summary features [4].

The focus of this work will therefore be to investigate how commonly used summary statistics constrain HH models in SBI and to identify which of these features are particularly good at doing so. To achieve this we devise and experiment with three different methods that augment either Neural Posterior Estimation [22] or Neural Likelihood [19] estimation to more efficiently acquire posteriors conditioned on different feature sets. Although, the methods we introduce are applicable to any study that seeks to quantify posterior constrainment with respect to individual summary features. In order to demonstrate the effectiveness of each method we first apply them to a simple toy problem with access to ground truth data. Subsequently, we pick one of the methods to study an implementation of the HH-model [26] with the goal of evaluating the informativeness of commonly used, hand-crafted summary features. We do this for simulated observations, as well as for electrophysiological recordings. We then outline the benefits and drawbacks of this approach, suggest use cases and present ideas on how it might be improved further. Finally, we conclude with an evaluation of the methods, a summary of this work and a brief outlook on its potential impact. All code and data will be made available at *github.com/mackelab*.

2 Methods

In Sec. 1 we have established the potential impact of simulation-based inference to the field of mechanistic models in neuroscience and the importance of understanding how posterior estimates are constrained by summary features. In this section we will lay out the process by which we intend to identify such features for SBI in general and for Hodgkin-Huxley models concretely. Furthermore, we will construct a suitable toy example to validate our methods and outline the HH model, that we will be using.

2.1 Identifying informative features

The contribution of a particular feature or subset of features $\boldsymbol{x}_1$ towards constraining the final posterior estimate $\hat{p}(\boldsymbol{\theta} \mid \boldsymbol{x})$ is hard to gauge in absolute terms, since we do not have access to the ground truth posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ usually. This means in order to identify informative features we have to consider the impact they have on the posterior estimate relative to each other, i.e. whether the removal of features $\boldsymbol{x}_2$ leads to increased uncertainty of the posterior estimate $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)$ with respect to a reference posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ (see Fig. [2.1]). Where $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2]$. This reference posterior can in theory be any other posterior estimate of interest, although it makes sense to use one that has been conditioned on the set of all available features $\boldsymbol{x}$. Under the assumptions of model convergence and a sufficiently flexible density estimator, differences in constrainment can be ascribed to differences in information content of features or noise in the data. Since we can reasonably assume that noise in the data remains constant, relative differences in posterior uncertainty provide a good way of assessing the contributions of summary features. A set of summary statistics is hence considered "informative", if it is able to constrain the uncertainty of the posterior estimate with respect to a reference posterior estimate. Our goal is therefore to obtain a posterior estimate $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ learned on some set of features $\boldsymbol{x} = (x_0, ..., x_n)$ and compare it against posterior estimates $\bar{p}_i = (\boldsymbol{\theta} \mid \bar{\boldsymbol{x}}_i)$ learned on a subset of these features $\bar{\boldsymbol{x}}_i(x_0, ..., x_{i-1}, x_{i+1}, ..., x_n)$. We will refer to them as full p and partially conditioned or just partial posterior $\bar{p}_i$ estimates respectively. These posterior estimates can then be used to

deduce the relative importance of features x_i for inference.

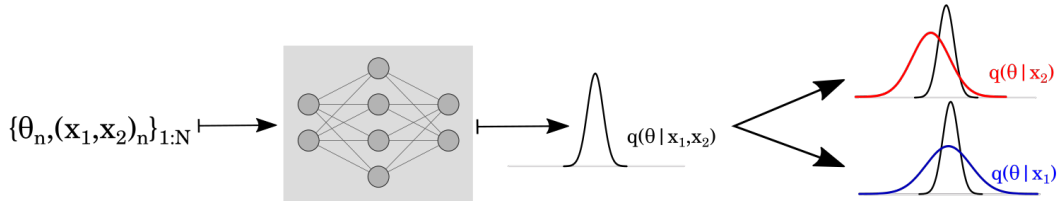

Figure 2.1: Pairs of parameter $\boldsymbol{\theta}$ and feature vectors $\boldsymbol{x}$ are used to train a neural density estimator to approximate the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ (black) of a "black box" model. Removing certain features from $\boldsymbol{x}$ leads to less constraint posterior estimates (blue, red). Some features (x_1) constrain the posterior distribution more than others (x_2) and are thus more informative to the neural network.

In order to efficiently measure feature constrainment we first devise three schemes for fast and flexible acquisition of posteriors conditioned on different feature (sub-)sets. Secondly, we suggest two methods for quantification of their relative uncertainties.

2.1.1 Neural Likelihood Estimation

The first SBI framework we will be considering for posterior acquisition is Neural Likelihood Estimation (NLE) [19]. NLE is a method to conduct Bayesian inference in simulator models with an intractable likelihood function. It works by training a conditional neural density estimator on data generated by the simulator and learns to approximate its likelihood $p(\boldsymbol{x} \mid \boldsymbol{\theta})$. Using Bayes rule the likelihood estimate can then be used to obtain an estimate of the posterior density.

In practice, this involves running the simulator on parameter samples from a prior distribution $\boldsymbol{\theta}_n \sim p(\boldsymbol{\theta})$, which can be equated to sampling the model's likelihood distribution according to $\mathbf{x}_n \sim p(\mathbf{x}|\boldsymbol{\theta}_n)$. The resulting training data $\{\mathbf{x}_n, \boldsymbol{\theta}_n\}_{1:N} \sim p(\mathbf{x}, \boldsymbol{\theta})$ can then be used to train a conditional density estimator $q_\phi(\boldsymbol{\theta}|\mathbf{x})$, parameterised by ϕ to approximate the likelihood distribution in the support of $p(\boldsymbol{\theta})$. q_ϕ is optimised by maximising the total log probability $\sum_n \log\, q_\phi(\boldsymbol{\theta}|\mathbf{x})$ over the training data with respect to ϕ, which, for a large number of training examples, is equivalent to maximising the expectation of the negative Kullback-Leibler (KL) divergence in the support of the prior [19].

$$\mathbb{E}_{p(\boldsymbol{\theta},\mathbf{x})}(\log\, q_\phi(\mathbf{x}|\boldsymbol{\theta})) = -\mathbb{E}_{p(\boldsymbol{\theta})}(\mathcal{D}_{KL}(p(\mathbf{x}|\boldsymbol{\theta})||\, q_\phi(\mathbf{x}|\boldsymbol{\theta}))) + const. \tag{2.1}$$

Since the negative KL divergence is maximal at 0, this implies that with enough simulations, a sufficiently flexible conditional neural density estimator will eventually approximate the true likelihood in the support of the prior

$\hat{p}|_{p(\boldsymbol{\theta})}(\boldsymbol{x} \mid \boldsymbol{\theta}) \approx q_\phi(\boldsymbol{x} \mid \boldsymbol{\theta})$. After obtaining such a tractable likelihood surrogate $q_\phi(\boldsymbol{x} \mid \boldsymbol{\theta}) \approx p(\boldsymbol{x} \mid \boldsymbol{\theta})$, Bayes rule can be used to obtain an estimate of the posterior conditioned on an observation $\boldsymbol{x}_o$ (see Eq. 2.2) via sampling with Markov Chain Monte Carlo (MCMC).

$$\hat{p}(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto q_\phi(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{2.2}$$

Compared to direct posterior approaches this extra sampling step is inconvenient and costly in terms of compute, however it enables the following trick: Because we have access to the intermediate tractable likelihood estimate, it can be marginalised to only include certain subsets of observational features. Via Eq . 2.2 this enables posteriors contextualised on these subsets with just a single trained surrogate likelihood. We call this post-hoc adjustment of the density estimate, as opposed to directly learning a density estimator for every possible feature combination of concern. This results in a posterior estimate $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)$ that can be evaluated and sampled from based on any subset of $\boldsymbol{x}$, using Eq. 2.3.

$$\hat{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_1) \propto q_\phi(\boldsymbol{x}_1 \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) = \int q_\phi(\boldsymbol{x}_1, \boldsymbol{x}_2 \mid \boldsymbol{\theta}) d\boldsymbol{x}_2 \, p(\boldsymbol{\theta}) \tag{2.3}$$

By opting to approximate the likelihood $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ using a Mixture Density Network (MDN) [37], the marginalisation of the likelihood $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ can even be done analytically. For a given partitioning of $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2]$ and a K-component MDN, each mixture component can be partitioned equivalently according to Bishop [38, Chapter 2.3.2], resulting in the following partial posterior distribution

$$\hat{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_1) \propto \sum_{k=1}^{K} \pi_k \mathcal{N}_k(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_{k,1}, \boldsymbol{\Sigma}_{k,11}) \, p(\boldsymbol{\theta}). \tag{2.4}$$

In our experiments we will sample from this posterior distribution using slice sampling [39] and use a uniform prior that covers all biologically relevant parameter ranges. We implement this procedure using python 3.8 and build on top of the already existing sbi library [40].

2.1.2 Neural Posterior Estimation

The next method to gauge feature dependent posterior constrainment augments Neural Posterior Estimation (NPE) [20–22]. Contrary to NLE, NPE directly targets the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ by training a neural density estimator from the simulated data pairs $(\boldsymbol{\theta}_n, \boldsymbol{x}_n)$. This approach carries the advantage

of circumventing additional inference steps and thus is particularly suited for inference tasks that require the accumulation of large amounts of samples.

NPE works very similarly to NLE. However a conditional neural density estimator $q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$ is trained to approximate $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ directly, based on data $\{(\boldsymbol{\theta}_n), \boldsymbol{x}_n\}$ obtained by running the simulator on parameter samples from a prior $\boldsymbol{\theta}_n \sim p(\boldsymbol{\theta})$. Then $\mathcal{L}(\phi) = \sum_n \log\, q_\phi(\boldsymbol{\theta}|\mathbf{x})$ is maximised with respect to the network's parameters ϕ. This yields a tractable posterior estimate without a tractable likelihood intermediary. Conducting inference on the resulting posterior estimate is therefore much easier. However, due to a lack of a tractable likelihood, post-hoc marginalisation is impossible. We therefore instead suggest to approximate the following integral

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}_1) = \int p(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{x}_2 \tag{2.5}$$

using a Monte Carlo (MC) approximation. Even though not exact, it would still allow to obtain good partial posterior estimates $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)$ without a repeated retraining of the posterior estimator given a sufficient amount of samples in the MC approximation.

Before applying the approximation directly, we can rewrite the integral in Eq . 2.5 to include the conditional evidence $p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1)$.

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}_1) &= \int \frac{p(\boldsymbol{\theta}, \boldsymbol{x}_1, \boldsymbol{x}_2)}{p(\boldsymbol{x}_1)} d\boldsymbol{x}_2 && (2.6) \\ &= \int \frac{p(\boldsymbol{\theta} \mid \boldsymbol{x}_1, \boldsymbol{x}_2) p(\boldsymbol{x}_2, \boldsymbol{x}_1)}{p(\boldsymbol{x}_1)} d\boldsymbol{x}_2 && (2.7) \\ &= \int p(\boldsymbol{\theta} \mid \boldsymbol{x}_1, \boldsymbol{x}_2) p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1) d\boldsymbol{x}_2 && (2.8) \end{aligned}$$

This enables us to replace $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$ with the parametric estimate of the posterior density $q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$. Provided the conditional evidence distribution can be estimated or made tractable, the integral can be approximated using a MC sum, which yields

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}_1) \approx \frac{1}{N} \sum_{i=1}^{N} p(\boldsymbol{\theta} \mid \boldsymbol{x}_1, \boldsymbol{x}_{2,i}), \quad \boldsymbol{x}_{2,i} \sim p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1). \tag{2.9}$$

With a tractable estimate of $p(\boldsymbol{x})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, we are thus able to approximate $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)$ for arbitrary subsets of features without any retraining.

In order to estimate the conditional evidence distributions as efficiently as possible, we suggest fitting a Mixture of Gaussians (MoG) to $p(\boldsymbol{x})$ via Expectation-Maximisation (EM). Using Bishop [38, Chapter 2.3.1-2.3.2], $p(\boldsymbol{x})$ can then be

conditioned analytically for every possible partitioning of $\boldsymbol{x}$. This results in the conditional MoG:

$$p(\boldsymbol{x}_2|\boldsymbol{x}_1) = \sum_{k=1}^{n} \frac{\pi_k \mathcal{N}(\boldsymbol{x}_1|\boldsymbol{\mu}_{k,1}, \boldsymbol{\Sigma}_{k,11})}{\sum_{l=1}^{n} \pi_l \mathcal{N}(\boldsymbol{\theta}_1|\boldsymbol{\mu}_{l,1}, \boldsymbol{\Sigma}_{l,11})} \mathcal{N}(\boldsymbol{\theta}_2|\boldsymbol{\mu}_{k,2|1}, \boldsymbol{\Sigma}_{k,2|1}) \tag{2.10}$$

$\boldsymbol{\mu}_{k,ij}$ and $\boldsymbol{\Sigma}_{k,ij}$ are again partitioned according to Bishop [38, Chapter 2.3.2].

We implemented this again in python 3.8, using the sbi library [40].

2.1.3 Transfer Learning

As a third and final strategy we will investigate how transfer learning [41] can potentially speed up the retraining process of the conditional density estimator for different subsets of features. The main purpose of this method is to serve as a comparison and benchmark for our two previous methods.

We approach the implementation of a transfer learning scheme in a simplistic manner, by attaching a fully connected layer to the neural density estimator q_ϕ (see Fig. [2.2]). The number of input dimensions of this layer is equal to the number of features and ReLU activation functions are used. In the first iteration of training the modified density estimator then learns a full posterior estimate as in Sec. 2.1.2. However, after training and thanks to the embedding layer the same pre-trained network can learn partial posterior estimates as well. We achieve this by setting input weights in the first layer to zero during subsequent training runs if the weights are connected to missing input features. In this way the neural net retains knowledge about the full posterior distribution from its initial training cycle and should therefore only have to learn the difference between $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)$ in the following training episodes. The amount of training time should therefore be reduced significantly, when compared to training on feature subsets directly.

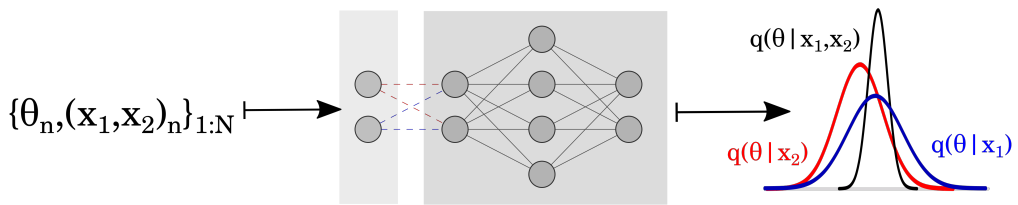

Figure 2.2: An extra fully connected layer is added the the neural density estimator. During the first training iteration the neural network is trained on all features. To quickly obtain partial posterior estimates, the pre-trained state can then be used for subsequent training runs, by dropping out specific weights (dashed red / blue lines) during training.

The implementation of this also made use of python 3.8 and the sbi library [40].

2.1.4 Quantifying informative features

Now that we have established how to obtain posteriors conditioned on arbitrary feature subsets, we turn to the question of what makes a feature "informative" and how to quantify this. As we have alluded to earlier, we consider a feature informative to the SBI procedure, if it is able to constrain the posterior estimate with respect to a reference posterior. In order to quantify the contributions of features well enough for qualitative judgements, we suggest the use of two complementary measures.

The first measure we consider is the ratio of variances of the marginals of $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)$ and $p(\boldsymbol{\theta} \mid \boldsymbol{x})$. An increase in the marginal variances of $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)$ compared to $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ hence indicates that the removed feature set $\boldsymbol{x}_2$ constrains the marginal distributions by the same amount. This lets us deduce if specific features x_i influence the uncertainty of some parameters θ_j more than others. This is very interpretable. However, the downside of relying solely on this measurement is its ignorance of any differences in covariance. Differences in the overall shape of the posterior distributions or covarying changes in uncertainty are not picked up by this indicator. But instead of measuring all covariances, which can be quite complicated to interpret visually, we believe it is a better idea to complement the first measure by a second one. For this purpose we employ an estimate of the KL divergence between $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)$ and $p(\boldsymbol{\theta} \mid \boldsymbol{x})$. Since we can reasonably assume that a posterior estimate does not become more constraint if features are removed, increases in the KL estimate should indicate an increase in uncertainty overall. Looking at the correlation between both of these measures also offers a good sanity check, whether the underlying posterior distributions $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)$ are sensible.

In order to estimate the KL-divergence between $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)$ and $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, we employ a purely sampled based estimator [42, 43]. This is necessary because the posterior distributions obtained via NLE and NPE are not normalised and hence circumvents having to evaluate them explicitly. The estimator makes use of 1-nearest neighbour search and is implemented using k-d trees [44], to reduce the 2^n comparison operations to a manageable number. With samples $\boldsymbol{X} = \{X_i\}_{i=1}^n$ and $\boldsymbol{Y} = \{Y_i\}_{i=1}^n$ from $p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)$ and $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ respectively, the

KL-divergence $KL(p(\boldsymbol{\theta} \mid \boldsymbol{x}_1)||p(\boldsymbol{\theta} \mid \boldsymbol{x}))$ is estimated as:

$$\mathcal{D}_{KL} = \frac{d}{n}\sum_{i=1}^{n} ln\frac{\min_j ||X_i - Y_i||}{\min_{j\neq i}^{n} ||X_i - Y_i||} + ln\frac{m}{n-1}. \tag{2.11}$$

Where d is the dimensionality of the samples. And $||\cdot||$ denotes the ℓ_2-norm.

As for the features, we compiled a total of 23 summary statistics (see Fig. 2.1), that can be extracted from voltage traces with the help of the Allen SDK. They comprise a mixture of whole membrane voltage and individual action potential (AP) related statistics. We have chosen this set specifically, as it is comprised of commonly used and easily interpretable statistics and have provided good posterior estimates for both simulated and experimentally measured data during previous testing.

Action Potential Statistics	**Membrane Voltage Statistics**
Mean threshold of APs (APT)	Mean of the resting potential (V_{rest})
Mean Amplitude of APs (APA)	Standard deviation of V_{rest}
Average width of APs (APW)	Mean of the membrane voltage (V_m)
Average afterhyperpolarisation (AHP)	Standard deviation of V_m
Threshold of the 1^{st} and 3^{rd} AP	
Amplitude of the 1^{st} and 3^{rd} AP	
Width of the 1^{st} and 3^{rd} AP	
AHP of the 1^{st} and 3^{rd} AP	
Adaptation of AP amplitudes	
Average adaptation of AP amplitudes	
Adaptation of interspike intervals (ISI)	
Coefficient of variation (CV) of ISIs	
CV for AP heights	
First spike latency	
Spike count (APC) / firing rate overall	
Spike count of ith quantiles	

Table 2.1: Summary statistics to gauge the similarities between model outputs and observations. Statistics may be used several times, i.e. spike count in first and second half of a voltage trace.

2.2 The Hodgkin-Huxley Model

The HH equations describe the generation of action potentials in the membrane of a giant squid axon by modelling the currents in a resistor capacitor circuit and by approximating the stochastic opening dynamics of voltage gated ion channels with a set of empirically measured differential equations [23]. In our experiments we opt to use a single compartment Hodgkin-Huxley model, Eq . 2.12, derived from Pospischil et al. [26]. This model goes beyond the standard

Sodium, Potassium and leak currents to generate spiking and also considers a slow non-inactivating K^+ current to model spike-frequency adaptation and a high-threshold Ca^{2+} current to generate bursting. In order to account for different experimental environments, we extend the model further by adding two constants, $k_{T_{adj}}$ and r_{SS}, which adjust for the ambient temperature of the recording and for different rate scaling respectively.

$$\begin{aligned} I_t = C\, A_{Soma} \frac{dV_t}{dt} &+ \bar{g}_{Na} m^3 h (V_t - E_{Na}) \\ &+ \bar{g}_K n^4 (V_t - E_K) + \bar{g}_{leak} (V_t - E_{leak}) \\ &+ \bar{g}_M p (V_t - E_K) + \bar{g}_L q^2 r (V_t - E_{Ca}) \end{aligned} \tag{2.12}$$

Eq. 2.12 models the evolution of the membrane voltage $V_t = V(t)$ given a stimulus $I(t) = I_t$. $\bar{g}_i,\ i \in \{Na, K, l, M, L\}$ are the maximum conductances of the Sodium, Potassium, leak, adaptive Potassium and Calcium ion channels respectively and E_i are the associated reversal potentials. I_t denotes the current per unit area, C the membrane capacitance, A_{Soma} the compartment area and n, m, h, q, r and p represent the fraction of independent gates in the open state, based on Hodgkin and Huxley [23]. The gating dynamics can be expressed through Eq. 2.13, 2.14, where $\alpha_z(V_t)$ and $\beta_z(V_t)$ denote the rate constants for one of the gating variables $z \in \{m, n, h, q, r\}$. They model the different voltage dependencies for each channel type and their parameterisation depends on the specific model that is used. In this work we stick to the kinetics of $\alpha_z(V_t, V_T)$, $\beta_z(V_t, V_T)$, $p_\infty(V_t)$ and $\tau_p(V_t, \tau_{max})$ as modelled in Pospischil et al. [26].

$$\frac{dz_t}{dt} = \left(\alpha_z(V_t)\,(1 - z_t) - \beta_z(V_t) z_t\right) \frac{k_{T_{adj}}}{r_{SS}} \tag{2.13}$$

$$\frac{dp_t}{dt} = \left((p_\infty(V_t) - p_t)/\tau_p(V_t)\right) k_{T_{adj}} \tag{2.14}$$

In our experiments we simulate response of the membrane voltage to the injection of a square wave of depolarising current of $200pA$. Since the system of equations is not solvable analytically, we compute the membrane voltage $V_t = V(t)$ using a numerical procedure of choice to iterate over the time axis in steps of $dt = 0.04ms$. We opt for the exponential Euler method [45] as it provides a good trade-off between computational complexity and accuracy for our application. Provided with an initial voltage V_0, we are now able to simulate the evolution of the membrane voltage $V(t)$ for a given time interval [0,T].

2.3 Toy Example

In order to demonstrate that the method works reliably, we construct a simple toy problem. The model that we consider is a Generalised Linear Model (GLM). In our case we consider a three dimensional parameter space $\boldsymbol{\theta} = [\theta_0, \theta_1, \theta_2]$ and four dimensional observation space $\boldsymbol{x} = [x_0, x_1, x_2, x_3]$. $\boldsymbol{\theta}$ is drawn from a uniform prior $\theta_i \sim \mathcal{U}(-5, 5), \quad i \in \{0, 1, 2\}$ and $\boldsymbol{x}$ is then sampled from a Gaussian distribution with mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance $\boldsymbol{\Sigma}$,

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}), \tag{2.15}$$

where $\boldsymbol{\mu}(\boldsymbol{\theta})$ is a linear transformation of the parameter vector $\boldsymbol{\theta}$ and diagonal noise $\boldsymbol{\Sigma}$.

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu}_0 + \boldsymbol{L}\boldsymbol{\theta} \tag{2.16}$$

$$\boldsymbol{\Sigma} = \sigma^2 \mathbb{1} \tag{2.17}$$

$\boldsymbol{L}$ was specifically chosen to probe, whether the posterior reacts and interacts with missing features as expected. For this reason $\boldsymbol{L}$ facilitates that feature x_0 solely depends on θ_0, x_1 only depends on θ_1, x_2 depends on both θ_1 and θ_2, and x_3 has no influence on any model parameter.

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{2.18}$$

3 Results

In the following section we present the results that each method achieves on the toy problem introduced in Sec. 2.3 and inspect feature importance in HH models for both synthetic and experimental data.

3.1 Toy Example

As a first proof of principle, we verify that each of the methods outlined in Sec. 2 correctly predicts individual feature contributions for our toy problem. We compare samples from the post-hoc, i.e. after training adjusted, posterior estimate to samples from the analytic partial posterior distribution. For this purpose we train a MDN with 10 mixture components to approximate the full posterior for 100,000 training samples. We then remove single features x_i from $\boldsymbol{x}$ and draw 10,000 samples for each partial posterior estimate. In order to compensate for differences in training and sampling initialisation we repeat this procedure for 20 different pairs of training and sampling seeds. We admit that 10,000 samples are excessive for the purpose of this analysis, but it leads to smoother and better interpretable contour plots in the associated figures.

3.1.1 Neural Likelihood Estimation

We begin with the post-hoc adjusted posterior estimate obtained using NLE. When one feature is removed at a time, we can observe that samples obtained from the post-hoc adjusted posterior surrogate match those from the analytic ground truth very accurately. This is not only true for the marginal, but also the pairwise distributions, as can be observed in Fig. [3.1]. Here we show a representative example of a partial posterior estimate, where x_2 has been dropped. In this specific case, we can observe, that $p(\boldsymbol{\theta} \mid \bar{\boldsymbol{x}}_2)$ and $\hat{p}(\boldsymbol{\theta} \mid \bar{\boldsymbol{x}}_2)$ not only are less constraint compared to $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, but that $\hat{p}(\theta_2) = p(\theta_2) = \mathcal{U}(-5, 5)$, which is what we expect from the relationship of $\boldsymbol{\theta}$ and $\boldsymbol{x}$ through $\boldsymbol{L}$. We can therefore conclude that the full posterior estimate has converged to the true posterior distribution in a way that post-hoc adjustment of the distribution yields the correct partial posteriors.

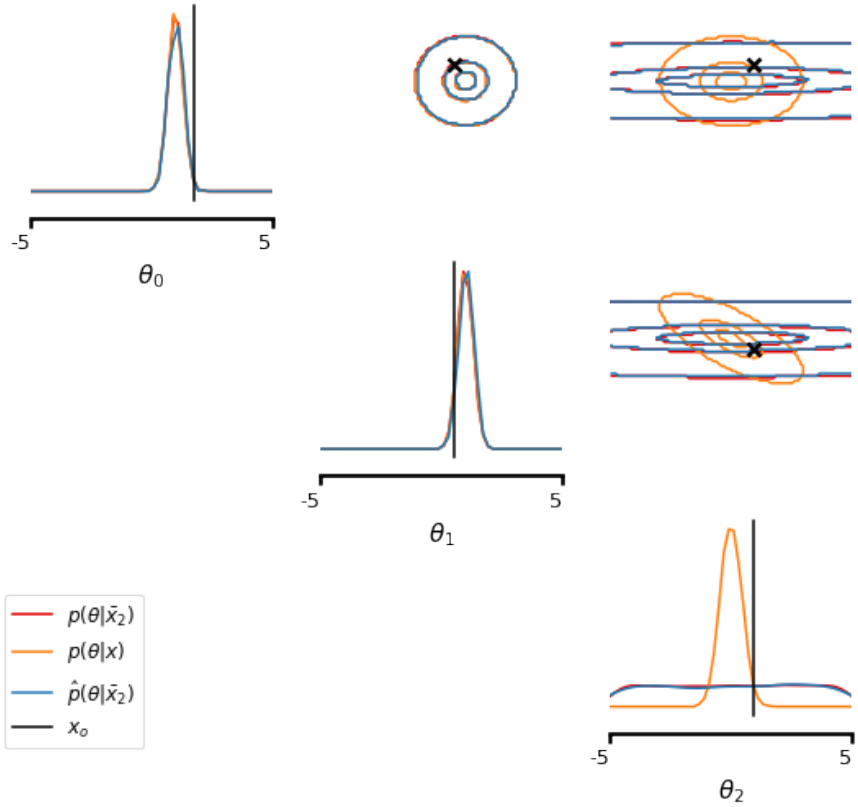

Figure 3.1: Pairwise contour plots (upper triangle) and kernel density estimates of the marginal distributions (diagonal) for analytic ground truth $p(\boldsymbol{\theta} \mid x_0, x_1, x_3)$ and post-hoc estimate of $\hat{p}(\boldsymbol{\theta} \mid x_0, x_1, x_3)$ obtained using NLE. As a reference the full analytic posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ is also included.

Next, we evaluate posterior constrainment for each feature. We compare samples from the analytic ground truth, direct and the post-hoc adjusted partial posteriors. Looking first at the KL divergence between the full and partial posteriors (Fig. [3.2, left]) we can observe that the KL of the post-hoc adjusted partial posteriors matches those predicted by both direct and analytic partial posteriors. We can further observe that the KL estimates of the post-hoc adjusted partial posteriors are only slightly more dispersed than those of the direct method. Additionally, the graph reflects what we already know from constructing the toy model, namely, that dropping x_4 has no effect on the shape of the posterior estimate, hence the KL remains 0. Similar to the KL estimates, the median increase in variance of the marginal posterior distributions (Fig. [3.2, right]) also shows good agreement between measures for post-hoc adjusted, direct and ground truth partial posteriors. More interestingly though, the increase in marginal variances reflects the relationship of parameters and features determined by $\boldsymbol{L}$ in Sec. 2.3 very precisely. Fig. [3.2,

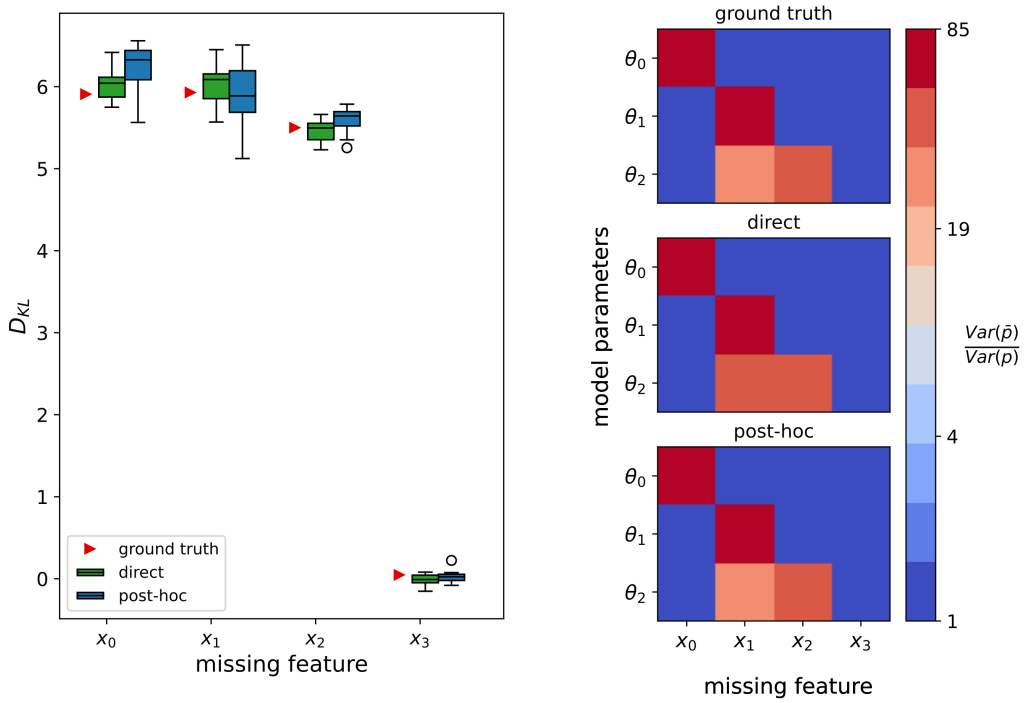

Figure 3.2: Predicted single feature constrainment according to analytic, direct and post-hoc adjusted NLE posterior estimates. left: KL-divergence between the full and partial posterior distributions $\mathcal{D}_{KL}[p(\theta \mid x)||p(\theta \mid \bar{x}_i)]$. right: Median increase in variance of the marginal posterior distributions per dropped feature.

right] reveals the dependence of feature x_1 on both θ_1 and θ_2, which is not evident solely from Fig. [3.2, left].

3.1.2 Neural Posterior Estimation

In order to compare the methods, we repeat the analysis conducted in Sec. 3.1.1 for NPE obtained partial posteriors. Here we see that the post-hoc adjusted partial posteriors predict feature constrainment almost as well as their directly trained counterpart when compared to the ground truth (see Fig. [3.3, 3.4]). Furthermore, as was already the case with the NLE obtained partial estimates, the dispersion of the KL is larger for the post-hoc adjusted case than for its direct counterpart. Compared to the post-hoc adjusted NLE posteriors, the partial posteriors obtained from the MC estimates also have a higher dispersion (see Fig. [3.2, left]). Considering the median increase in marginal variance, Fig. [3.3, left], shows really good agreement again though. All contributions are clearly visible and as predicted by the ground truth.

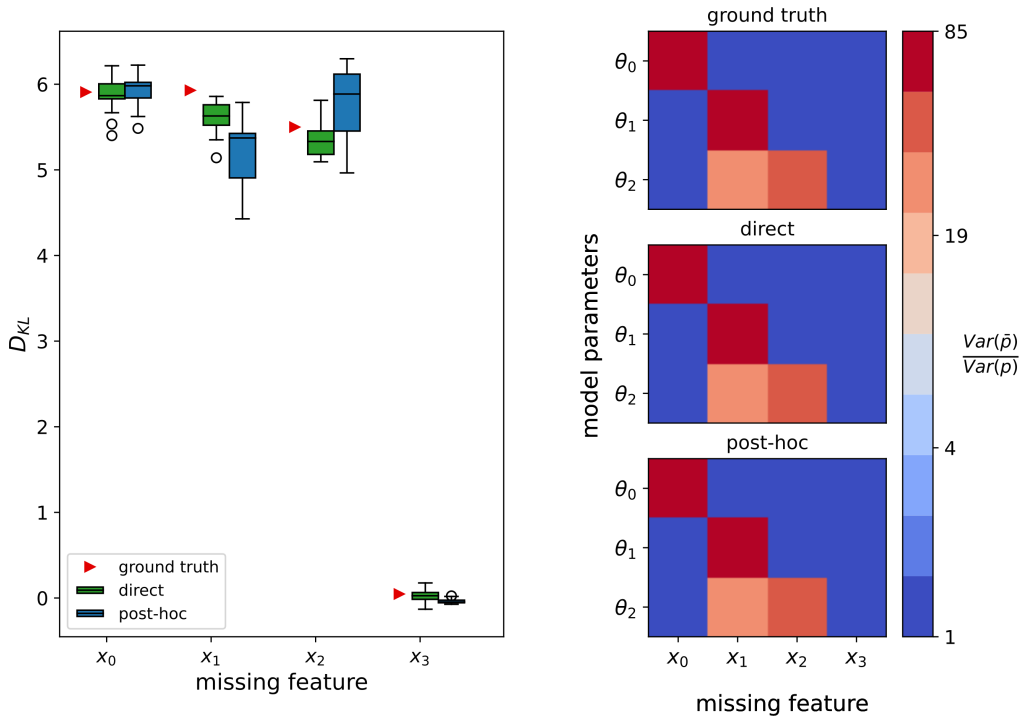

Figure 3.3: Predicted single feature constrainment according to analytic, direct and post-hoc adjusted NPE posterior estimates. left: KL-divergence between the full and partial posterior distributions $\mathcal{D}_{KL}[p(\theta \mid x)||p(\theta \mid \bar{x}_i)]$. right: Median increase in variance of the marginal posterior distributions, by missing feature.

3.1.3 Transfer Learning

Finally, we conduct this analysis for the last proposed method. Firstly, we can observe that the predictions from the pre-trained density estimator match the ground truth very well. Secondly, the dispersion of the KL estimates in the transfer learned case is smaller and more comparable to that of the direct case. Although, this is expected since transfer learning could be considered a form of directly learning partial posteriors (see Fig. [3.4, left]), hence no difference in dispersion should be visible. Additionally and similar as in the two previous cases, we also see good agreement between the median increases of the marginal variances upon dropping features (see Fig. [3.4, right]). More interesting than the predictions of this methods are it's convergence rates however, since the aim of this method is to reduce the acquisition time of partial posterior estimates. To this end we created 10 differently seeded training sets, each with 10,000 examples. Then, one full and four partial posterior estimates were learned on each set. The total training time in epochs and minutes was recorded. The results can be seen in Fig. [3.5] and show that our implementation of transfer learning yields no time-savings compared to the time required to learn a full

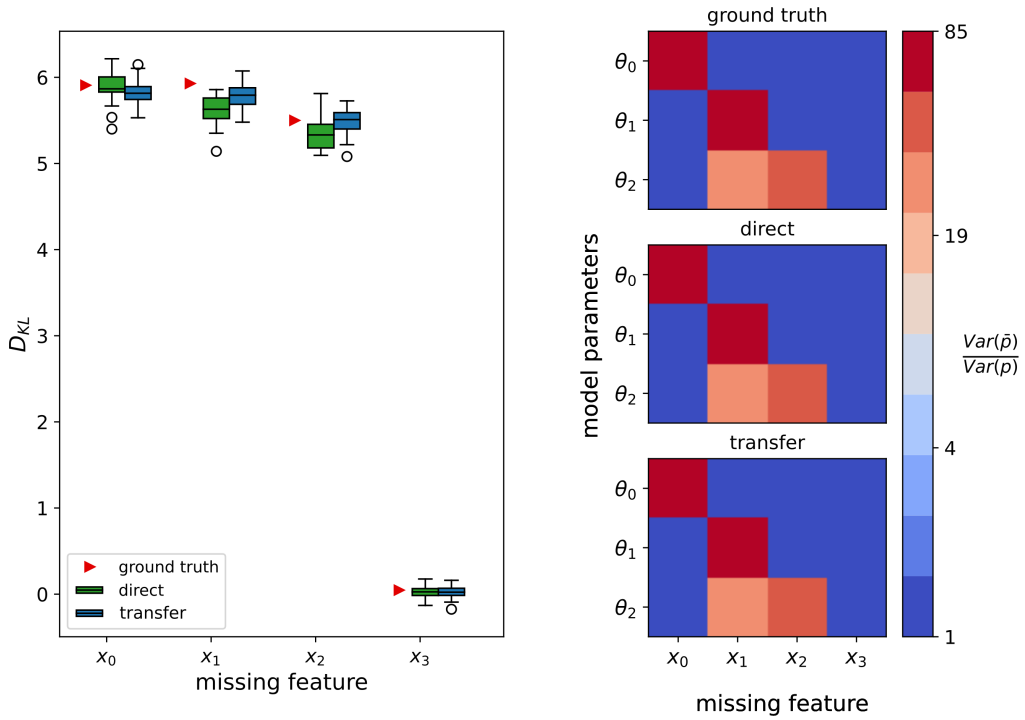

Figure 3.4: Predicted single feature constrainment according to analytic, direct and transfer learned estimates. left: KL-divergence between the full and partial posterior distributions $\mathcal{D}_{KL}[p(\theta \mid x)||p(\theta \mid \bar{x}_i)]$. right: Median increase in variance of the marginal posterior distributions, by missing feature. Samples were obtained across 20 different pairs of training and sampling seeds.

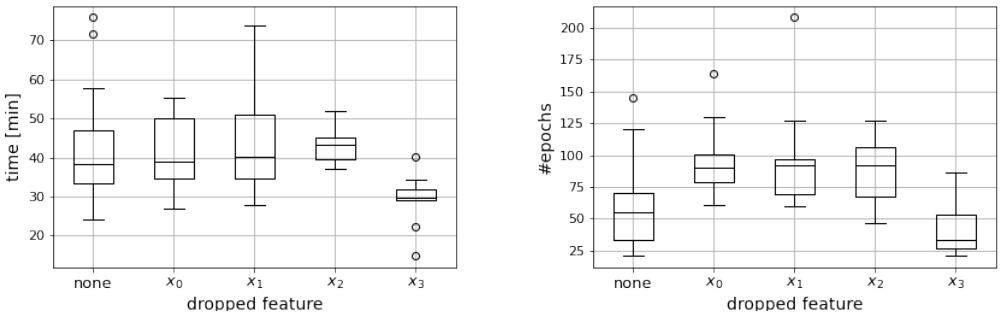

Figure 3.5: Convergence times for partial posterior estimates initialised with pre-trained weights. none refers to full posterior estimate.

posterior estimate (see Fig. [3.5], left most box-plot). The convergence rate of learning a full posterior estimate from random initialisation was the same as learning a partial estimate starting from a pre-trained state.

3.2 Hodgkin-Huxley Model

Having demonstrated the viability of the different approaches in Sec. 3.1, we now apply this methodology to HH-models to identify the features that con-

strain posterior estimates. For this we will only be working with the NLE approach (see Sec. 2.1.1), as it does not rely on any mathematical approximations and requires no retraining. Furthermore, the highly non-Gaussian evidence of the HH-model leads to inaccurate MC estimates compared to the posteriors obtained via post-hoc likelihood marginalisation with NLE.

3.2.1 Synthetic Data

In a first step we fix the reversal potentials $E_{Na} = 53\,mV$, $E_K = -107\,mV$, $E_{Ca} = 131\,mV$ and parameters $k_{T_{adj}}$, r_{SS}, A_{Soma}, since these are mainly used to increase model flexibility to fit experimental observations. This results in a HH model with 9 free parameters. We then create 1 million simulations with input parameters drawn from a uniform prior which covers biologically sensible parameter ranges and summarise the resulting output traces using a set of 5 summary features from Tab. 2.1, namely $\boldsymbol{x} = (APA, AHP, APC, \mu(V_{rest}))$. Before we can learn a density estimate however, we have to deal with so called "bad features" [21]. Unfortunately, many simulated voltage traces do not exhibit any spiking behaviour and thus their summary statistics are nonsensical. In order to deal with this problem, we decide to replace any non-numerical features with numerical ones that lie distinctly outside of the distribution of spiking features. In our testing we found that taking the mean value of a "good feature" and adding about 10-15 standard deviations, provided a good default recoding. Even more important however, was the addition of noise to the recoded values. We discovered that adding about 5 standard deviations of noise created a good amount of overlap and separation between the "good" and "bad" feature distributions. Especially the amount of overlap turned out to be really important for the quality of results.

As a first test, we choose a relatively simple fast spiking voltage trace, which exhibits no adaptation or bursting, to act as an observation (see Fig. [3.6, left]) and compare direct estimates of the partial posteriors to those obtained using post-hoc adjusted estimates.

As we can see in Fig. [3.7], using only 5 features results in posterior estimates that are very broad, even in the case of the full distribution (orange). However, the post-hoc adjusted posterior distributions (blue) match those obtained directly (red) very well. Furthermore, we can see that the posterior distribution when $\mu(V_{rest})$ is dropped becomes less constraint in E_{leak} as well as in g_L. This is also reflected in the estimates of the KL divergence between the full and partial posteriors as well as the median increase in marginal variance

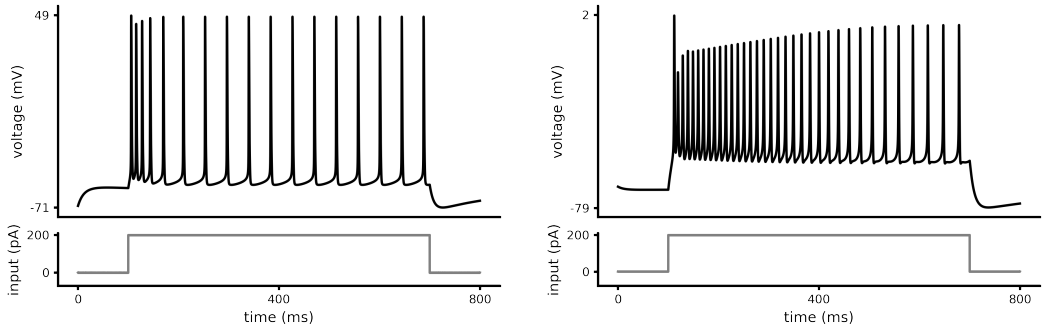

Figure 3.6: Simulated membrane voltage traces from the HH-model, that are used as an observation x_o in the following experiments. left: Simple fast spiking trace with no adaptation currents ($x_{o,1}$). right: Fast spiking trace with Ca adaptation currents ($x_{o,2}$).

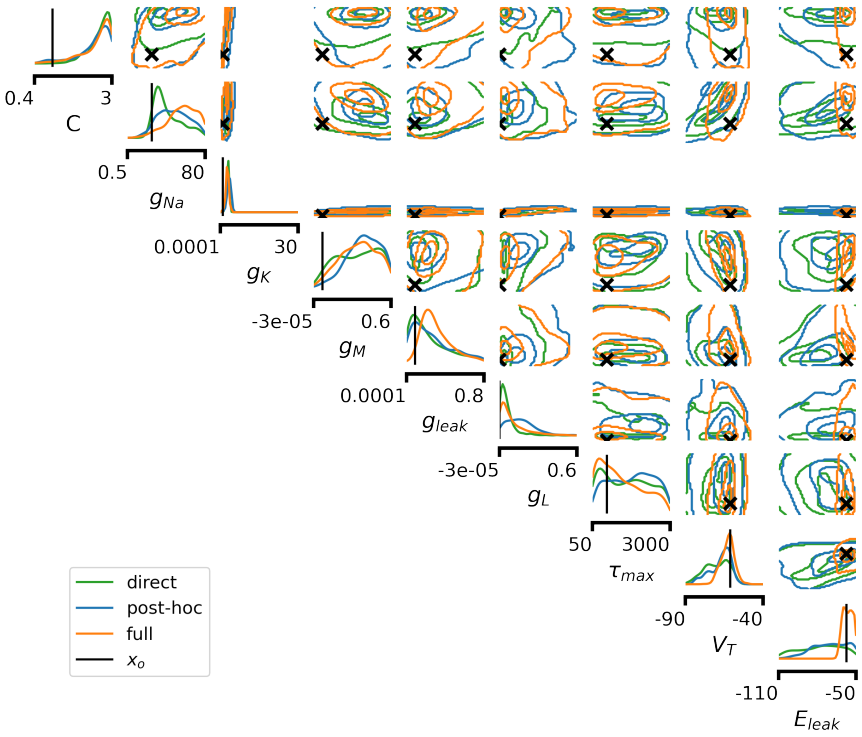

Figure 3.7: Pairwise contour plots (upper triangle) and kernel density estimates of the marginal distributions (diagonal) for analytic ground truth direct and post-hoc estimates of $p(\boldsymbol{\theta} \mid APA, AHP, APC)$. As a reference the full posterior distribution is also included.

(see Fig. [3.8]) for a set of 20 different training and sampling seeds. Here we see that both metrics correctly predict the effect of $\mu(V_{rest})$ on E_{leak}, which was already visible in the posterior distribution in Fig. [3.7]. Furthermore, the metrics obtained from the direct posterior estimates and those from the post-hoc adjusted variants match very well, except for a larger discrepancy in the case of APA. In our testing we found that this behaviour is consistent across different seeds, but not across other feature sets or even across the

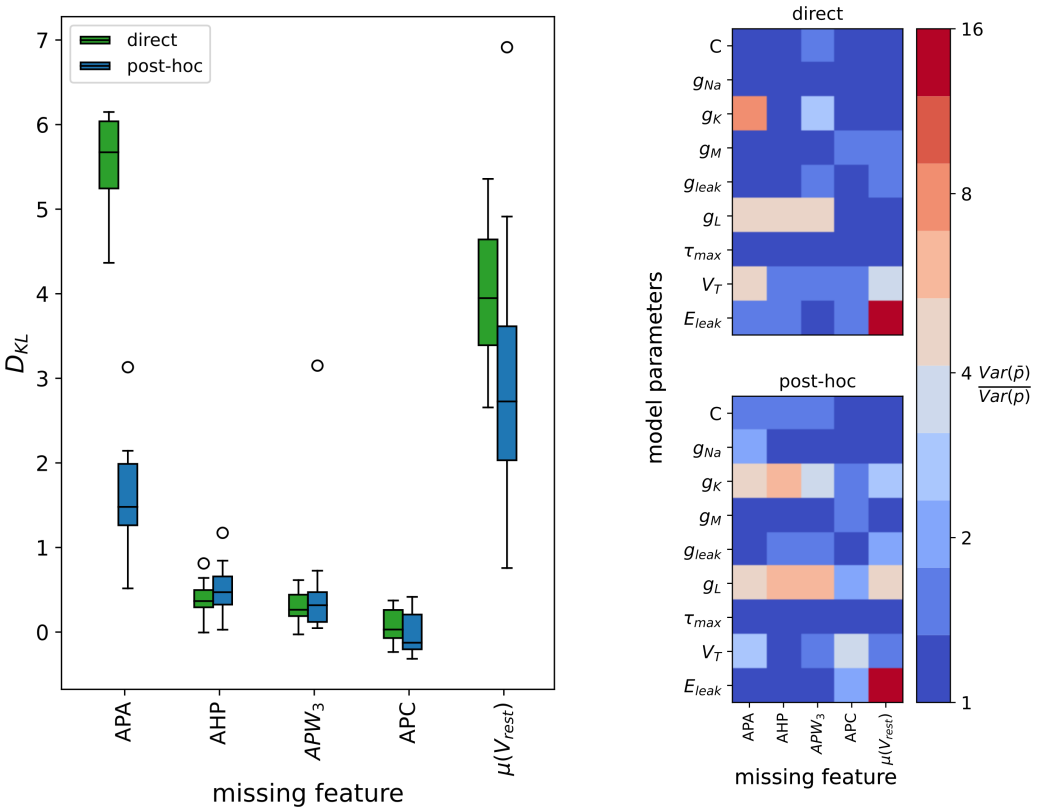

Figure 3.8: Predicted single feature constrainment according to direct and post-hoc adjusted posterior estimates. left: KL-divergence between the full and partial posterior distributions $\mathcal{D}_{KL}[p(\theta \mid x)||p(\theta \mid \bar{x}_i)]$. right: Median increase in variance of the marginal posterior distributions per missing feature.

same feature set and different observations (see Fig. [3.16, left]). Although the post-hoc adjusted match the direct posterior distributions very well in the large majority of cases, this indicates that sometimes both methods converge to different posterior estimates. We hypothesise that some information contained within the features we remove post-hoc leaks into the partial posterior estimate and causes some parameters to be constrained in a different fashion. Another oddity is a large dispersion in the KL estimates of some of the partial distributions, namely $\mu(V_{rest})$. While some part can definitely be attributed to training instabilities, we have found that the main source of error can be trace back to MCMC. If we sample from very flat distributions (see Fig. [3.7], $p(E_{leak})$), small fluctuations in the sampling process can lead to large fluctuations in both metrics. Nonetheless, both metrics agree that $\mu(V_{rest})$ is critical in constraining the posterior, which also becomes apparent when looking at samples from the corresponding partial distributions (see Fig. [3.9]). Looking at Fig. [3.8, right], we can make even more nuanced observations about the feature constrainment in this particular example. Firstly, we observe that

along with E_{leak}, $\mu(V_{rest})$ also has a small impact on g_{leak} in both metrics. Furthermore, both approaches agree that APA and APW_3 all constrain g_L, g_K and V_T, while AHP also constrains g_L. However, again there are some discrepancies in the predictions of both approaches. While the post-hoc approach predicts additional constrainment of g_K by AHP and $\mu(V_{rest})$, this is not corroborated by the direct estimate. Again, we hypothesise that this is due to some information leaking into the post-hoc adjusted posterior distribution.

As an additional analysis, we now look at the effect that skipping features has on the posterior samples. For this purpose, we run the simulator on 500 parameter samples from the full posterior estimate and an exemplary partial posterior $p(\boldsymbol{\theta} \mid \bar{\boldsymbol{x}}_\mu(V_{rest}))$. We then plot the relative errors between the summary statistics of our observation and those of the posterior samples, as well reviewing five representative traces from the distributions.

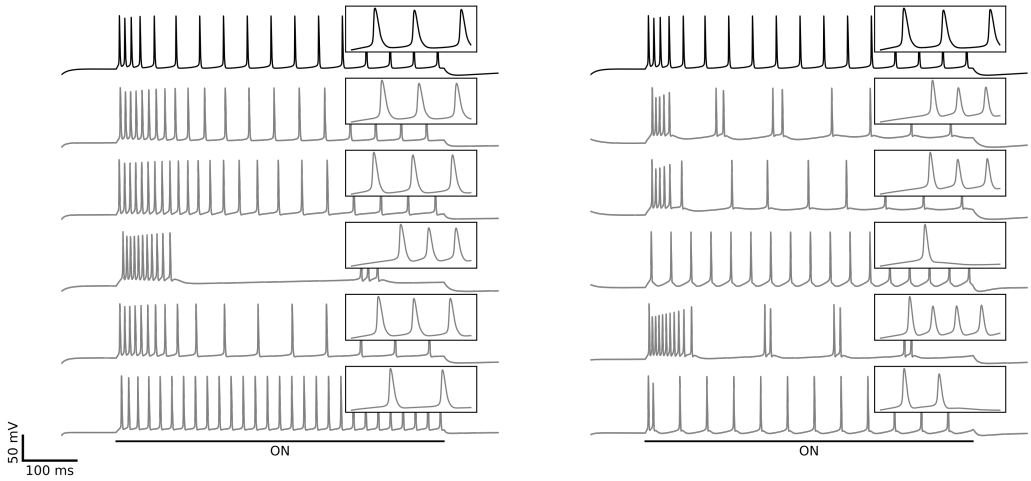

Figure 3.9: left: $V_{o,1}(t)$ (black) compared to posterior samples from $p(\boldsymbol{\theta} \mid APA, AHP, APW_3, APC, \mu(V_{rest}))$. right: $V_{o,1}(t)$ (black) compared to posterior samples from $p(\boldsymbol{\theta} \mid APA, AHP, APW_3, APC)$. The insets show the first 30 ms after stimulus onset.

While training a posterior estimate on all 5 parameters (Fig. [3.9, left]) yields remarkably similar traces to the observation, we can see that removing parameters indicated as important in Fig. [3.8] significantly diminishes the quality of posterior samples (Fig. [3.9, right]). In particular, dropping the membrane's mean resting potential leads to traces that deviate strongly from the observed resting potential (Fig. [3.9, right]). More specifically, larger variations in the membrane's resting potential and the AP hyperpolarisation currents, which both are characteristics associated with the leak currents. This

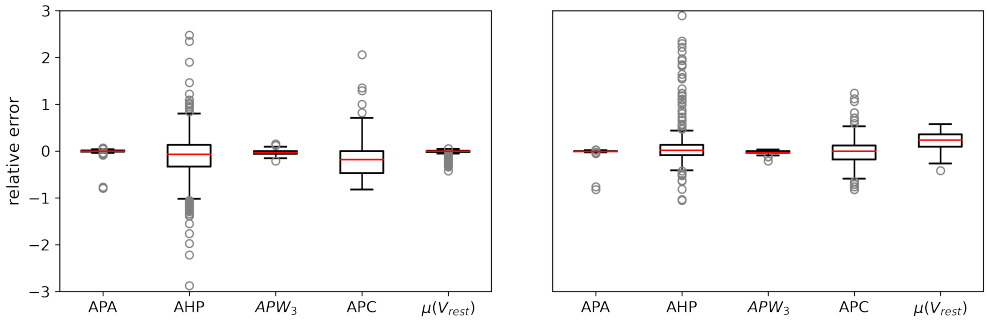

Figure 3.10: Comparison of the relative discrepancies between features of the observation and of the posterior samples. left: between $\boldsymbol{x}_o$ and $\boldsymbol{x}$. right: between $\boldsymbol{x}_o$ and $\bar{\boldsymbol{x}}_{\mu(V_{rest})}$.

was correctly predicted in Fig. [3.2], as removing $\mu(V_{rest})$ goes along with increased uncertainty in the parameters of the leak currents E_{leak} and g_{leak}. Additionally, we can observe all features are fit well, when considering the mean error across all traces, however AHP and APC show a very large spread in values, while APA, APW_3 and $\mu(V_{rest})$ are very tightly bound (see Fig. [3.10]). Large variations in AHP and APC are also somewhat visible in Fig. [3.9]. Though, this can be explained by the fact that the posterior estimate $p(\boldsymbol{\theta} \mid APA, AHP, APW_3, APC, \mu(V_{rest}))$ is still very broad for the parameters that are associated with both features (see Fig. [3.7]) and that both features are hardly constraining the posterior distribution as can be seen in Fig. [3.16, left]. However, more interesting is that the removal of $\mu(V_{rest})$ not only leads to an increased spread of values, but also an upwards bias in the relative error (see Fig. [3.10]), reconfirming what was already visible in Fig. [3.9]).

Having shown that the post-hoc adjusted posterior distributions correctly predict strongly constraining features for a simple fast spiking trace on a set of five summary features, we now expand our analysis to all 23 features and include an additional, more complex trace (see Fig. [3.6, right]). Unfortunately, scaling up is not as straightforward as just adding more features, as we our method has become increasingly unstable in cases with more than five data features. In order to scale, we therefore have to resort to aggregating data over a large variety of smaller subsets and combine them afterwards. Therefore we train a set of density estimators on 100 randomly sampled combinations of five features. We then drop out every feature on every partial posterior estimate one at a time and draw 500 samples. With these samples we can then calculate the increases in variance for each distribution similar to Fig. [3.8] and aggregate the data by taking the median values for each dropped feature and parameter. This allows us to estimate the contribution of each feature and even take into account a

limited amount of pairwise feature correlations under the assumption that each feature shares a subset with every other feature enough times. Furthermore, aggregation over multiple samples increases robustness of the method to both the high variance encountered previously and the discrepancies observed for some post-hoc adjusted features (see Fig. [3.8]). We note, that this scheme is very extensive for a result that is arguably much easier obtained by training 23 separate density estimators, one for each dropped feature. However, this is only plausible as long as features are removed one at a time. If pairs or a greater number of features are to be removed at once, an exponentially growing amount of separately trained density estimators would be required. This issue only becomes more problematic for larger sets of summary statistics. We therefore argue that the use of such a scheme is justified, especially for larger studies that do not rely on large sample draws and also generate accurate predictions as we will demonstrate in the following. Fig. [3.11] shows the results of this process for the left trace in Fig. [3.6]. On first glance, we see that the

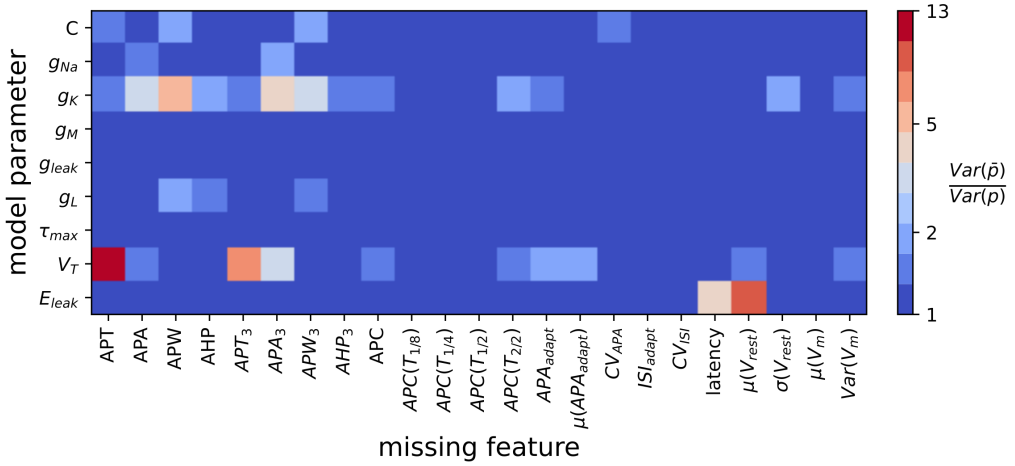

Figure 3.11: Median increase in variance of the marginal posterior distributions compared to the full posterior skipping one feature at a time and for each of the 23 summary features of $x_{o,1}$.

plot looks incredibly sparse. Only very few features cause a noticeable increase in the marginal variance if removed and three parameters are not being constrained at all. This is interesting as it either indicates a lot of redundancy in the provided statistics or combined effects, that cannot be identified by removing features one at a time. Nonetheless, we can clearly identify some strong single feature contributions. The first main effect is an increase in marginal variance of the threshold voltage V_T that is caused mainly by the removal of APT, but also APT_3 and APA_3. Some minor involvement also comes from APA, APC, APA_{adapt} and $\mu(APA_{adapt})$. This is in line with the expectation that variations in V_T directly lead to traces with varying AP thresholds [7].

The second parameter that is being affected by a lot of features is g_K. Not only is it strongly constrained by APW and APA_3, but also by APA, APW_3, $APC(T_{2/2})$ and $\sigma(V_{rest})$. Since the potassium currents are mainly associated with repolarisation of the membrane, it makes sense that the constrainment of the potassium conductance depends on features that are associated with or derive from the shape of APs. Finally, an increase in marginal variance of the leak reversal potential E_{leak} upon removal of $\mu(V_{rest})$ or *latency* can also be observed. The effect of $\mu(V_{rest})$ on E_{leak} could already be seen in Fig. [3.9] and is also expected, as changes to the leak reversal potential lead to voltage traces with varying resting potentials. However, it is still surprising that it is being strongly and exclusively constrained, by just these 2 features. Some additional and minor constrainment to C and g_{Na} is also visible. While both APW features affect C, both APA features constrain g_{Na}. Meanwhile, g_L is constrained by APW.

The effects described by Fig. [3.12] can also be confirmed qualitatively, by looking at samples from a posterior estimate trained on the specific subsets of interest, i.e. all features except for *latency* and $\mu(V_{rest})$ in Fig. [3.12, centre]. Overall it can be noted that training on a large subset of features produces

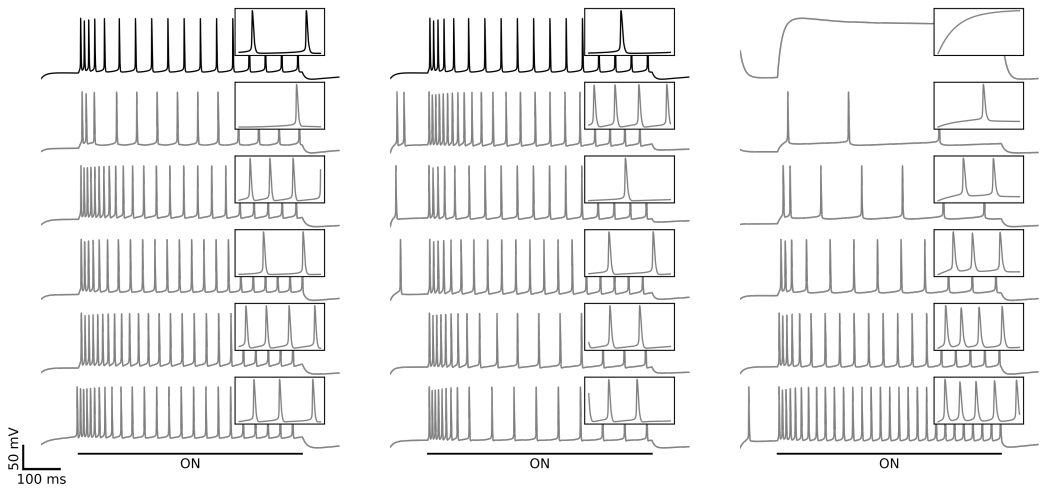

Figure 3.12: left: $V_{o,1}(t)$ (black) compared to posterior samples from $p(\boldsymbol{\theta} \mid \boldsymbol{x})$. centre: $V_o(t)$ (black) compared to posterior samples from $p(\boldsymbol{\theta} \mid \bar{\boldsymbol{x}}_{\mu(V_{rest}),latency})$. right: Effects on $\boldsymbol{x}_{o,1}$ if E_{leak} is varied between -80mV to -50mV, while the remaining parameters are kept constant. The insets show the first 50 ms after stimulus onset.

traces (grey) that provide very good fits for the observational trace (black), arguably superior to those derived from conditioning on just five features (see Fig. [3.9, left]). Moreover, we can observe that the removal of $\mu(V_{rest})$ and

latency (Fig. [3.12, centre]), the two features most constraining E_{leak}, causes the voltage traces sampled from the resulting partial posterior to show much more varying interspike intervals and pre-stimulus spiking. Both of these properties are linked to changes in the leak currents. It is therefore interesting that the voltage traces obtained from the partial posterior very closely resemble those obtained when E_{leak} is just varied in $\boldsymbol{\theta}_o$ (Fig. [3.12, right]). This shows that the aggregated increases in marginal variance in Fig. [3.11] correctly predict the effects of removing strongly constraining features from $\boldsymbol{x}$.

Due to the simplicity of the previous trace, we will now consider a second trace that exhibits spike amplitude adaptation (see Fig [3.6, right]). As before, we look at the median increases in marginal variance aggregated over 100 feature combinations, for all 23 features. Similarly as before, we can observe

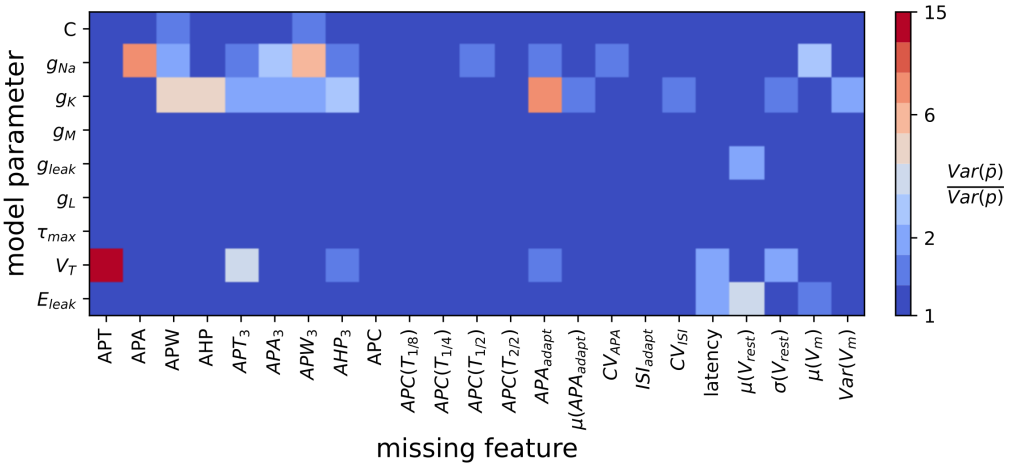

Figure 3.13: Median increase in variance of the marginal posterior distributions compared to the full posterior skipping one feature at a time and for each of the 23 summary features of $x_{o,2}$.

a strong variance increase for V_T that is associated with skipping APT and again *latency* as well as $\mu(V_{rest})$ constraining E_{leak}. Furthermore, $\mu(V_{rest})$ additionally constrains g_{leak} for this observation. In contrast to Fig. [3.11], we can observe that several features now prominently constrain g_{Na}, mainly APA and APW_3, along with APA_3 and $\mu(V_m)$. Meanwhile, g_K is again strongly constrained by APW, AHP and AHP_3. Additionally and contrary to Fig. [3.11], g_{Na} is also being constrained by APA_{adapt}. Since the observation displays spike amplitude adaptation, it makes sense that the spike depolarisation currents are affected. Overall though, the contributions remain relatively sparse.

The effects of removing the features most constraining a certain parameter can again also be demonstrated qualitatively via samples from a posterior es-

timate trained directly on a feature subset deemed of interest based on Fig. [3.13]. As an example we can look at the effect that features have on g_K, by drawing samples from a posterior estimate that ignores the features that most constrain it. The first thing we see can in Fig. [3.14] is that the traces

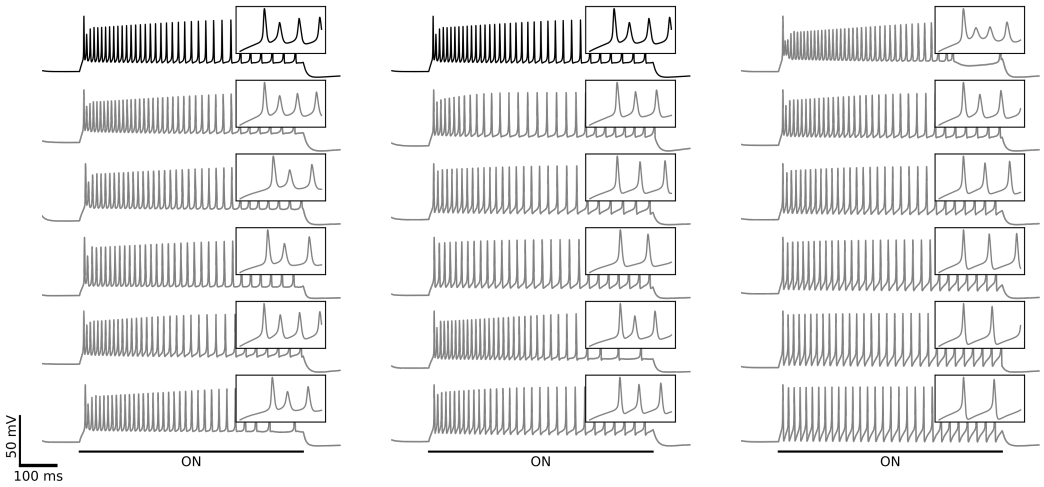

Figure 3.14: left: $V_{o,2}(t)$ (black) and posterior samples from $p(\boldsymbol{\theta} \mid \boldsymbol{x})$. centre: $V_o(t)$ (black) and posterior samples from $p(\boldsymbol{\theta} \mid \bar{\boldsymbol{x}}_{APW,AHP,APW,AHP_3,APA_{adapt}})$. right: Effects on $\boldsymbol{x}_{o,2}$ if g_K is varied between 3mV and 10mV, while the remaining parameters are kept constant. The insets show the first 40 ms after stimulus onset.

sampled from the full posterior distribution generally match our observational trace very well (Fig. [3.14, right]). This is the case especially at the onset of stimulation. Comparing these to the samples in Fig. [3.14, centre], which come from $p(\boldsymbol{\theta} \mid \bar{\boldsymbol{x}}_{APW,AHP,APW,AHP_3,APA_{adapt}})$, reveals a greater variety in the first few spikes after stimulation and in amplitude adaptation. Furthermore, a small amount of additional variation can be found at the end of the repolarisation phase of the AP. If we compare these to what we would expect from changing the potassium conductance manually in $\boldsymbol{\theta}_o$, than we can observe in Fig. [3.14, right], that these traces are very similar to those sampled from $p(\boldsymbol{\theta} \mid \bar{\boldsymbol{x}}_{APW,AHP,APW,AHP_3,APA_{adapt}})$.

Overall, it is noteworthy that in both examples and contrary to our prior belief spike count related features did not contribute towards constraining posterior uncertainty, even if provided along with just four other features (see Fig. [3.8]). Although small increases in variance can be seen for APC related features in Fig. [3.11], this could also be due to random fluctuations during training or sampling and would need further investigation. Furthermore, we can observe that no single feature included in our selection of 23 constrains τ_{max} to a visible

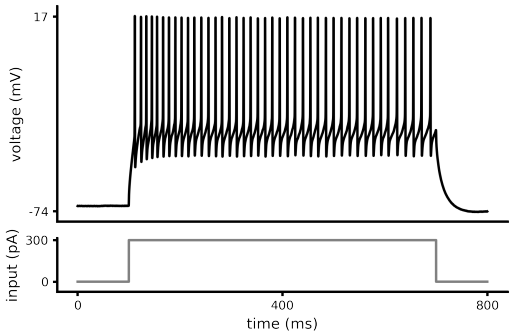

Figure 3.15: Experimentally recorded membrane voltage response of a M1 neuron in the mouse motor cortex to a square wave depolarising current.

degree. This is however in line with previous experience where fitting this set of summary statistics to HH models did not produce sharp estimates of τ_{max}. Additionally, we could observe that most of the features almost exclusively constrain a single parameter. This is interesting as it means that just based on the choice of these features it is possible to target single parameters, while leaving room for others to vary.

3.2.2 Experimental Data

As a final analysis, we will now use an experimentally recorded membrane voltage trace (see Fig. [3.15]) in place of our observation $\boldsymbol{x}_o$. The recorded trace is part of a larger Patch-seq [46] data-set [47] for which neurons were recorded across all layers of the adult mouse motor cortex, M1 (mostly post-natal day P50+, median age P75). To record the neurons, they were patch-clamped in cute slices and stimulated with brief current impulses to record their electrophysiological activity [47]. In order to show that our method can also be reliably used for experimentally obtained recordings, we replicate the experiments that we have conducted for the synthetic data in Sec. 3.2. Similar to Fig. [3.8], Fig. [3.16] also displays the constrainment for $p(\boldsymbol{\theta} \mid \boldsymbol{x}_o)$ for $\boldsymbol{x} = APA, AHP, APC, \mu(V_{rest})$. We observe that not only are the KL divergences in Fig. [3.16, left] very similar for the direct and post-hoc adjusted posterior estimates, but also the median predicted increases in variance show good agreement. Compared to Fig. [3.8] the KL estimates show an increased amount of dispersion. However, since this is true of both approaches, it is most likely due to the increased difficulty of fitting the experimental data. Comparing Fig. [3.8] and Fig. [3.16] further yields more interesting insights. Firstly, no discrepancy in the $\mathcal{D}_{KL}$ on removal of APA can be observed, despite of using the same features as in the synthetic case, indicating that this behaviour is

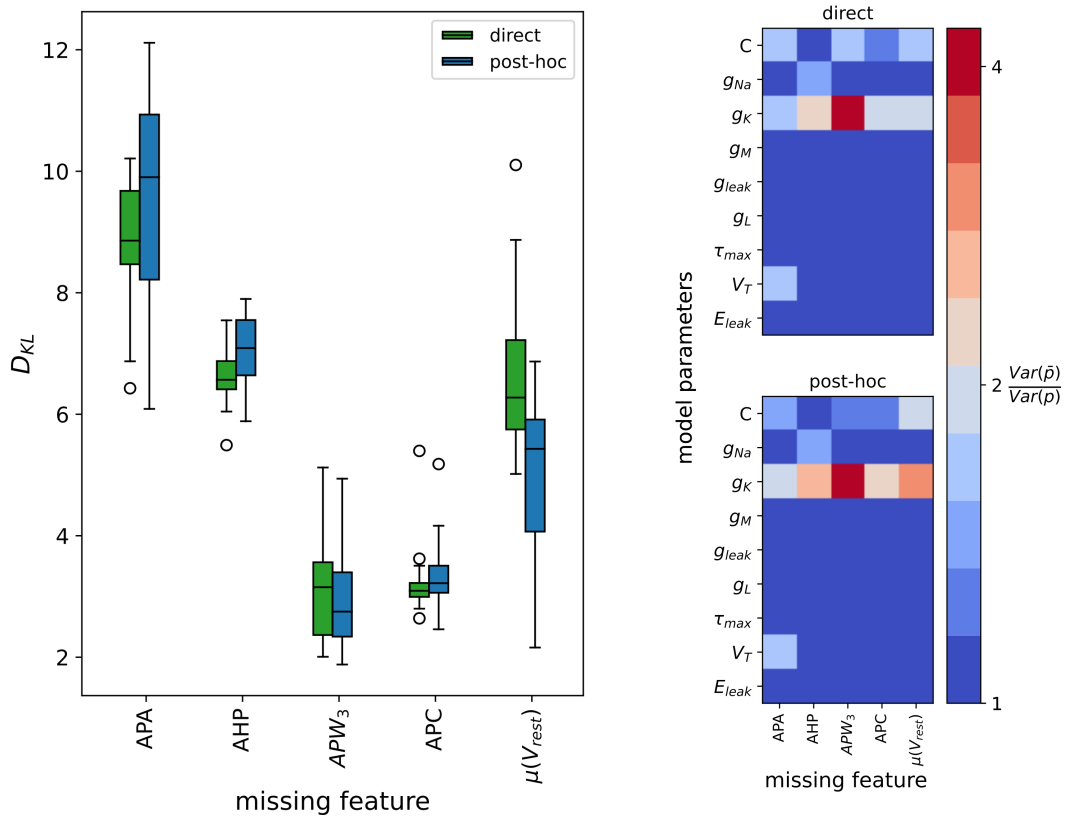

Figure 3.16: Predicted single feature constrainment according to direct and post-hoc adjusted posterior estimates. left: KL-divergence between the full and partial posterior distributions $\mathcal{D}_{KL}[p(\theta \mid x)||p(\theta \mid \bar{x}_i)]$. right: Median increase in variance of the marginal posterior distributions per missing feature.

not feature set dependent. Secondly, the importance of features is very different even though the traces could be considered qualitatively very similar and the same features were used during training. While in the synthetic case, the posterior was mainly constrained by $\mu(V_{rest})$ and APA, for the experimental trace APA and AHP are the most prominent features overall according to the KL estimates. In contrast to this, the strongest effect on a single parameter is that of APW_3 on g_K, as predicted by an increase in its marginal variance. Moreover, we can observe that almost all constrainment is directed towards this single parameter, while other marginals are affected very little.

In order to investigate if these contributions still persist in presence of all other features we now look at the increase in marginal variances aggregated over 100 different combinations of five features. Doing this we obtain Fig. [3.17]. Here we can see that several of the contributions seen in Fig. [3.16] remain strong in Fig. [3.17]. Additionally, several more features can be observed to constrain the sodium conductance g_{Na}, namely the 1st and 2nd moment of V_m,

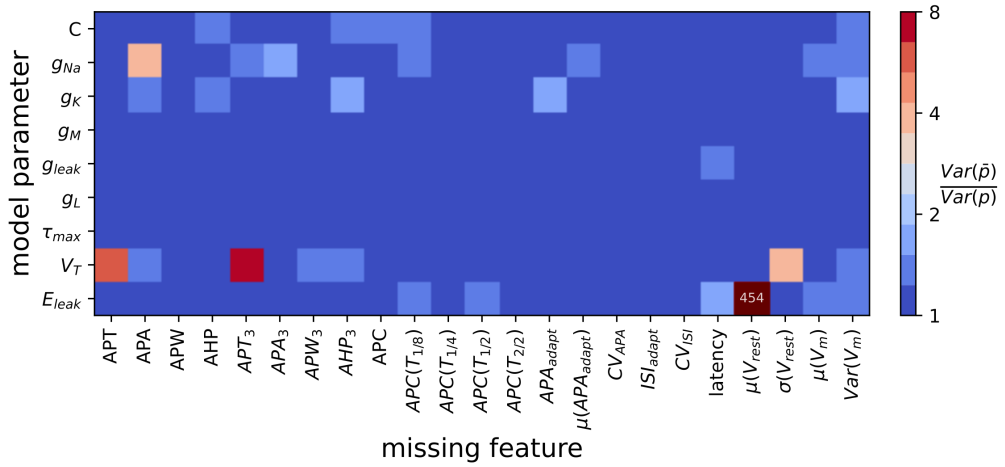

Figure 3.17: Median increase in variance of the marginal posterior distributions compared to the full posterior skipping one feature at a time and for each of the 23 summary features of x_o. The increase in Var for $E_{leak}, \mu(V_{rest})$ far exceeds the other increases, which is why it has been marked separately.

as well as $APC_{1/8}$ and APA_3. Furthermore, while in Fig. [3.16] APW_3 was very important in determining g_K, it ceased constraining g_K in Fig. [3.17]. Instead APT_3 and APA are the most prominent features, affecting mainly V_T and g_{Na} respectively. As a final qualitative assessment of this we can now look at posterior samples from a direct partial posterior estimate in Fig. [3.18]. While the samples from the full posterior distribution in Fig. [3.18, left] re-

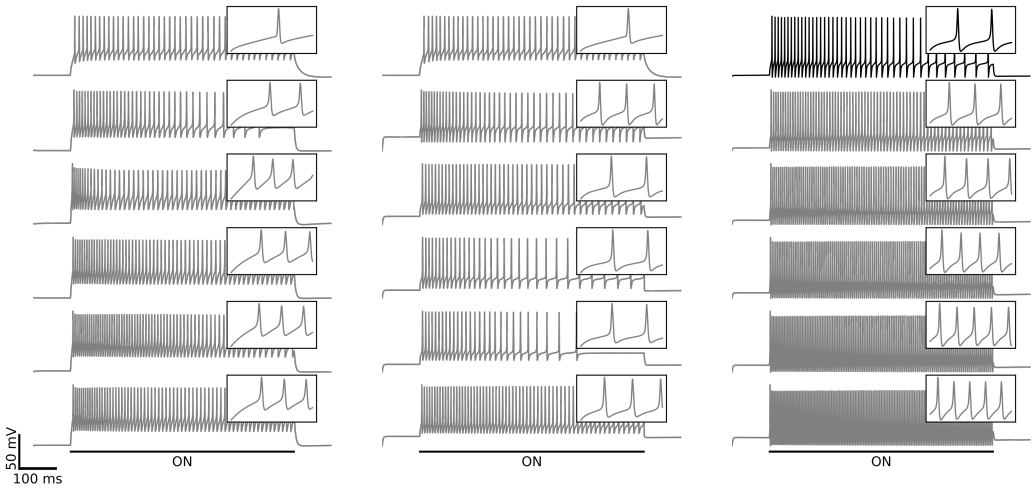

Figure 3.18: left: $V_o(t)$ (black) and posterior samples from $p(\boldsymbol{\theta} \mid \boldsymbol{x})$. centre: $V_o(t)$ (black) and posterior samples from $p(\boldsymbol{\theta} \mid \bar{\boldsymbol{x}}_{\mu(V_{rest})})$. right: Effects on $\boldsymbol{x}_o$ if E_{leak} is varied between -70mV to -60mV, while the remaining parameters are kept constant. The insets show the first 20 ms after stimulus onset.

ally closely resemble the experimental observation, just the removal of $\mu(V_{rest})$ from the set of features causes large variations in the quality of samples. The

traces in Fig. [3.18, centre] display not only an increased resting potential variance, which is in line with previous results of dropping $\mu(V_{rest})$, but also large differences in the APs during the refractory period are visible. Both are characteristics associated with the leak currents. It is therefore interesting that this manifests itself not only in the features of the posterior samples, but also in the predicted constrainment of E_{leak} and also g_{leak} to a lesser degree in Fig. [3.17]. Furthermore, taking the best parameter estimate for x_o and varying E_{leak} manually (see Fig. [3.18, right]), yields voltage traces that look remarkably similar to those in Fig. [3.18, centre].

Overall these results show that SBI can reveal how information from multiple data features imposes collective constraints on channel and membrane properties in the HH model. Furthermore, the predictions made by the post-hoc marginalised likelihood distributions obtained via NLE and are being supported by the qualitative assessment of posterior samples.

4 Discussion

As we have seen, the ability of NPE and NLE to obtain full posterior estimates can not just be leveraged to search for mechanistic models which can quantitatively capture experimentally observed behaviour, but also to gauge how these models are constrained by the available data. This makes these approaches not just particularly suited for experiments that aim to reproduce specific behaviours, i.e. realistic firing rates, rate-correlations and response nonlinearities [48, 49], but also to identify multiple sets of parameters in line with a particular data feature [8, 12, 15]. Insights can be gained on how specification, preparation and selection of appropriate data features or summary statistics influences model constrainment and how features can carve out admissible parameter ranges in which compensatory behaviour can take place (a weakly constraining feature also leaves a lot of room for compensation to occur). Ultimately, this improves our understanding of how neuromodulatory [50], physiological [51] or phenotypical [52] variability shape the electrophysiological properties of neurons and allow to identify correlates between microscopic mechanisms and macroscopic behaviour.

4.1 Related Work

Due to a large number of non-linear voltage- and time-dependent currents that contribute to a cell's electrophysiological behaviour, it can be difficult to determine how each current influences the firing dynamics. Identifying suitable models consistent with a given stereotypical behaviour has therefore attracted a lot of interest from a methodological, as well as a physiological standpoint. Additionally, thanks to biological variability in some of the parameters (such as the conductances) within and across species not just one, but many different plausible versions of such models are of interest [15, 52]. Several different approaches have therefore been developed in order to find suitable models of neural activity, while capturing their variability as effectively as possible. One idea to achieve this was first put forward by Prinz et al. [12]. They build a database of various model solutions over a search domain and then searched this database for solutions. A similar method was adopted by Ori et al. [51], who studied how parameter variations in HH models lead to different levels

of membrane excitability and how the parameter space can be partitioned accordingly. They do this by simulating a large set of randomly sampled voltage traces, whose parameter sets are then classified according to the activity they produce. In a related approach Goldman et al. [50] vary the maximal conductance parameters of a model neuron and characterise the activity pattern generated by the cell as silent, tonically firing, or bursting. They use this to identify directions in parameter space along which these basic patterns of neural activity are conserved. However, in all three cases large amounts of parameter combinations are tried and this can become prohibitively expensive for an increased number of parameters or more fine grained / larger parameters sweeps. In order to increase the efficiency of parameter sweeps Druckmann et al. [53], Ben-Shalom et al. [54], Achard and De Schutter [55] therefore propose to use target functions that weigh models with regards to a target solution and then optimise these functions. This approach has the advantage that it can be scaled to include large numbers of parameters. However, it provides no notion of how likely a given set of parameters within a search range is to produce a certain activity and therefore no way of reliably measuring how strongly correlated parameters and electrophysiological features are. For a long time this shortcoming was approached by means of sensitivity analysis [56–58]. After identifying a "good" set of parameters, the effect on a given electrophysiological property of varying one parameter at a time could be observed. However, this approach has two limitations: Firstly, it is based on examining a single model. Secondly, only one parameter is varied at a time. It therefore does not yield any information about simultaneously changing multiple parameters [52]. Alternatively, the same analysis can also be conducted on a population of models that has been randomly or systematically instantiated [25, 50, 55, 59]. A third approach, by Taylor et al. [52], treats data features as scalar functions of the parameters and fits cubic functions to a population of LP models one feature at a time. This allowed them to ask which conductances predominantly influenced electrophysiological properties. Paired with novel visualisation procedures such as Taylor et al. [60], Alonso and Marder [61], these data can then be interpreted or analysed, although their resulting conclusions are often only qualitative. Similar to our analysis of a single compartment HH model, they found that each property of an identified neuron with well defined behaviour is affected by a different subset of several maximal conductances. In our case, this concerned especially the sodium and potassium conductances. However, by fitting these functions one feature at a time what Taylor et al. [52] fail to capture is potential correlations of features. Furthermore, similar to all other methods mentioned, they rely on extensive sweeps of the model or parameter

space, which can be problematic for more complex models.

In stark contrast to the previously mentioned methods, Gonçalves et al. [7] propose the use of Sequential Neural Posterior Estimation (SNPE) for the discovery of large sets of models constrained by available data and demonstrate the power of their approach by identifying suitable HH models for a set of synthetic and experimental voltage traces. This does not just allow for amortised model inference, but also takes into account possible feature and parameter correlations. Furthermore, they show how the full posterior estimates they obtain can be used to conduct a limited analysis of feature constrainment. They qualitatively compare SNPE-estimated posteriors based on 3 different subsets derived from 7 data features in total. The number of spikes, the mean resting potential, the standard deviation of the resting potential, and the first four voltage moments of the membrane voltage. In their experiments, Gonçalves et al. [7] found that providing more data features decreased the uncertainty of the potassium conductance, which we can also corroborate. However, we can make more nuanced observations than that. Firstly, we can also observe that $Var(V_m)$ and APC constrain the potassium conductance. However, compared to the other statistics, we do not consider them to have a strong effect. Rather, as can be seen in both Fig. [3.11, 3.13], the potassium conductance is mainly being constrained by AP related statistics such as APW, APA and AHP. This aligns well with the observation that the spike shape is known to constrain sodium and potassium conductances [26, 53, 62]. Secondly, this effect depends on the observational trace that is picked, which is also something Gonçalves et al. [7] acknowledge and which can be seen in Fig. [3.17], where the potassium conductance was only weakly constrained by a hand full of features. They also note that some parameters (the threshold voltage, the density of sodium channels, the membrane reversal potential and the density of potassium channels) were consistently much stronger constrained than others (the adaptation time constant, the density of slow voltage-dependent channels and the leak conductance), which we can attest to as well. More specifically, we do not only see strong constrainment in Fig. [3.11, 3.13, 3.17], but we further observe that only very few features are necessary to constrain these. Finally, Gonçalves et al. [7] also note that posterior simulations matched the observed data only in those features that had been used for inference, which we can also confirm from our experiments.

4.2 Applicability and Limitations

The ability to conduct inference on all models that can be formulated in terms of a simulator, while not putting any constraints on their design allows NLE and NPE to scale to complex models that have previously been inaccessible to Bayesian inference [21]. By hooking into these frameworks, the methods we devised in this thesis for the purpose of quantifying feature constrainment will be applicable to the same range of problems that NLE and NPE may be applied to. Particularly the method outlined in Sec. 2.1.1 is a powerful example of this, since it does not rely on any approximations. Any simulator model whose output can be summarised in terms of data features can hence be studied to understand how its features constrain its parameters using the methods and metrics presented. For this analysis of the HH model, we have focused on an exemplary list of commonly used summary statistics, which is by no means exhaustive and was selected based on previous experience. However, this method can be easily extended to include additional summary statistics or even automatically obtained features [34–36]. This would also open the door for future studies to compare the constrainment of learned and hand-crafted features.

Nonetheless, we note that a few limitations also come with both the methods and the metrics introduced. For one, the inherent dependence on MDNs for the conditional neural density estimation limits overall accuracy for complex posterior or likelihood distributions. This is not the case for both NPE related methods, since the posterior density estimator can be chosen freely. However, even though the method in Sec. 2.1.2 is not constrained by the neural posterior estimator, it is limited by the selected density estimator to approximate $p(x)$. In order to avoid retraining on every feature subset, an estimator needs to be chosen from which conditioned samples $\boldsymbol{x}_1 \sim p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1)$ can be easily obtained. Otherwise, this would just shift the problem of retraining density estimators from the partial posterior to the conditional evidence distributions. For this reason we selected a Gaussian Mixture Model in our implementation, which struggles with highly non-Gaussian evidence distributions, like those of the HH model. As a benchmark for the other two approaches, we also proposed to add transfer learning capabilities to the neural posterior estimator. As we saw in Sec. 3.1, this did not yield any speed up, suggesting that the density estimator behaved similarly to when initialised randomly. We hypothesise that lowering the learning rate, fixing weights in later layers or adding

more embedding layers could address these issues.

Although most of our analysis was conducted using NLE, it is not always the best approach and its use depends largely on the goal and circumstances of a given study. Nonetheless, whether a method is a good fit can be broadly assessed based on set of two factors. Firstly, the amount of resources available - time or compute - in extreme cases can mean that a brute force approach to obtaining partial posterior estimates might even be preferable to post-hoc adjustment. This is due to the opportunity cost associated with trading off training and sampling time. While NPE obtained posteriors, i.e. via transfer learning, might require significantly more training time, the total time to acquire a set amount of samples can still be less than for post-hoc adjusted likelihood estimates. Secondly, the accuracy depends to a large degree on the density estimator that is used, which in some cases could disqualify the use of post-hoc likelihood marginalisation, due to its reliance on MDNs as opposed to Normalising Flows [63]. The same counts for MC approximations in the case Gaussian Mixture Models are used to approximate the evidence.

Another point we want to raise and that was already touched upon in Sec. 3.2, is that in some instances, the post-hoc adjusted partial posteriors seem to be differently constrained to what we would expect from obtaining them directly. This indicates, that even though a feature is removed from the full posterior estimate, some residual information of its former presence is still present in the remaining parameters of the post-hoc adjusted posterior estimate, leading to excess constrainment. This violates our assumption that the neural likelihood estimator learns a correct approximation of the full likelihood distribution, such that marginalisation leads to the same marginal likelihood as when it is trained on subsets directly. We have observed that this behaviour is not specific to a given feature or observation and the post-hoc adjusted distributions generally match those obtained directly very well in most cases. When we aggregate the data over a large number of feature subsets, these effects should therefore average out, i.e. in the case of Fig. [3.11]. However, one disadvantage of this procedure is its disregard for feature correlations. This is because we expect feature importance to vary depending on which other features are part of a subset along with it, i.e. the importance of APT with or without APT_3. When interpreting the results, such as Fig. [3.11], we can only make a limited amount of deductions about the interdependence of different features in constraining model parameters.

While aggregating the data in this way does mitigate the misalignment between some post-hoc and direct partial posterior distributions, what we had initially intended, was to analyse feature contributions for all 23 features at once, which unfortunately, we did not manage to do with our method. While we had no problems to add more parameters to the model, increasing the number of features beyond about seven, leads to increasingly unstable results. The more constrained the posterior distributions became in general, the larger were the discrepancies between the direct and post-hoc partial posterior estimates. We hypothesise that more features also lead to increased "leakage of feature information", as there is more opportunities for leakage when the posterior estimate becomes more complex. We also believe that some variability in the results can be attributed to training, more specifically to the handling of "bad features". During our testing we noticed a strong sensitivity of our results to the exact recoding parameters we employed. In particular, just changing the amount of noise or the separation from the distribution of "good features" slightly had a huge impact on the final posterior estimates. Dealing with "bad features" in a more involved way, for example as used in Deistler et al. [8], Lueckmann et al. [21], therefore has the potential to improve the reliability of this method significantly.

Before we conclude, we also want to mention some limitations that arise with trying to quantify informative features. Firstly, due to the relative nature of the increase in variance of the marginal posterior distributions, its magnitude does not only depend on the magnitude of change in posterior uncertainty, but also on the amount of uncertainty in the full posterior estimate. For very narrow distributions a small increase in variance can already lead to a large increase in this metric, while it might not be noticeable for broad distributions. This can lead to constrainment of features being weighted differently in the presence of other strongly constraining features than in their absence and also leads to greater variability for narrow, as opposed to wide distributions. Secondly, we consider features to be "informative" if they are good at constraining posterior estimates. However, by assuming that "informative features" reduce uncertainty, no distinction is made between for example informative handcrafted features and features with little noise. While we can draw conclusions about the reduction in posterior uncertainty that these features provide, we therefore cannot deduce if these features are very informative in the classical sense of the word [64]. In order to investigate the true information content of each feature, a more information theoretic approach would be needed.

5 Conclusion

In this work we devised a set of methods, that makes a detailed analysis of feature constrainment in SBI not only more practical, but also seamlessly integrates NLE [19] and NPE [22]. We suggested two complementary metrics to quantify feature constrainment and demonstrated their viability on a simple toy problem. These experiments particularly demonstrated the potential of post-hoc likelihood marginalisation in conjunction with NLE. We then analysed 2 sets of synthetic and one experimental voltage trace for posterior feature constrainment applying post-hoc likelihood marginalisation to a diverse set of 23 commonly used summary statistics. During our experiments we were not only able to confirm that this approach matches the predictions obtained from training density estimators for each feature subset of interest, but also corroborate the findings of Gonçalves et al. [7], Pospischil et al. [26], Druckmann et al. [53] and Hay et al. [62]. This suggests that our method produces correct predictions for feature constrainment of HH models without necessitating a separate density estimate for each feature subset of interest. This capability does not just open the door for large scale and extensive studies of summary statistics in HH models, such as comparing the constrainment of learned and hand-crafted features, but also can help in the identification of boundaries for compensatory behaviours. By adopting more involved methods to better handle "bad features", for example as shown in Gonçalves et al. [7], Lueckmann et al. [21], we believe that even studies of correlated constraining effects are possible.

6 Acknowledgements

This thesis was supervised by Prof. Dr. Jakob Macke and Prof. Dr. Philipp Berens. I thank them and their groups for their hospitality, valuable input and fruitful discussions, which have lead to numerous amendments to this thesis. In particular I want to give a huge thank you to Michael Deistler and Yves Bernaerts, who have counselled, advised and accompanied me from start to finish of this thesis. Their investment of time and effort into this project has been very generous and was incredibly valuable to me. I have appreciated their support immensely and I have enjoyed working with them on this project. It would not have become what it is without their expertise and advice.

Bibliography

[1] P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, 2001. ISBN 978-0-262-54185-5.

[2] W. Gerstner, H. Sprekeler, and G. Deco. Theory and Simulation in Neuroscience. *Science*, 338(6103):60–65, Oct. 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1227356.

[3] T. O'Leary, A. C. Sutton, and E. Marder. Computational models in the age of large datasets. *Current Opinion in Neurobiology*, 32:87–94, June 2015. ISSN 09594388. doi: 10.1016/j.conb.2015.01.006.

[4] K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences of the United States of America*, 117(48):30055–30062, Dec. 2020. ISSN 1091-6490. doi: 10.1073/pnas.1912789117.

[5] R. E. Baker, J.-M. Peña, J. Jayamohan, and A. Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters*, 14(5):20170660, May 2018. ISSN 1744-9561, 1744-957X. doi: 10.1098/rsbl.2017.0660.

[6] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[7] P. J. Gonçalves, J.-M. Lueckmann, M. Deistler, M. Nonnenmacher, K. Öcal, G. Bassetto, C. Chintaluri, W. F. Podlaski, S. A. Haddad, T. P. Vogels, D. S. Greenberg, and J. H. Macke. Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife*, 9: e56261, Sept. 2020. ISSN 2050-084X. doi: 10.7554/eLife.56261.

[8] M. Deistler, J. H. Macke, and P. J. Gonçalves. Disparate energy consumption despite similar network activity. *bioRxiv*, page 2021.07.30.454484, Aug. 2021. doi: 10.1101/2021.07.30.454484.

[9] A. Hasenstaub, S. Otte, E. Callaway, and T. J. Sejnowski. Metabolic cost as a unifying principle governing neuronal biophysics. *Proceedings of the National Academy of Sciences*, 107(27):12329–12334, July 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0914886107.

[10] B. Sengupta, A. A. Faisal, S. B. Laughlin, and J. E. Niven. The Effect of Cell Size and Channel Density on Neuronal Information Encoding and Energy Efficiency. *Journal of Cerebral Blood Flow & Metabolism*, 33(9): 1465–1473, Sept. 2013. ISSN 0271-678X, 1559-7016. doi: 10.1038/jcbfm.2013.103.

[11] S. B. Laughlin. Communication in Neuronal Networks. *Science*, 301 (5641):1870–1874, Sept. 2003. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1089662.

[12] A. A. Prinz, D. Bucher, and E. Marder. Similar network activity from disparate circuit parameters. *Nature Neuroscience*, 7(12):1345–1352, Dec. 2004. ISSN 1546-1726. doi: 10.1038/nn1352.

[13] G. M. Edelman and J. A. Gally. Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, 98(24): 13763–13768, Nov. 2001. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.231499798.

[14] G. Hu, J. Li, and G.-Z. Wang. Significant Evolutionary Constraints on Neuron Cells Revealed by Single-Cell Transcriptomics. *Genome Biology and Evolution*, 12(4):300–308, Apr. 2020. ISSN 1759-6653. doi: 10.1093/gbe/evaa054.

[15] E. Marder and A. L. Taylor. Multiple models to capture the variability in biological neurons and networks. *Nature Neuroscience*, 14(2):133–138, Feb. 2011. ISSN 1546-1726. doi: 10.1038/nn.2735.

[16] J.-M. Lueckmann, J. Boelts, D. Greenberg, P. Goncalves, and J. Macke. Benchmarking Simulation-Based Inference. In *International Conference on Artificial Intelligence and Statistics*, pages 343–351. PMLR, 2021.

[17] S. N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, Aug. 2010. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature09319.

[18] J.-M. Lueckmann, G. Bassetto, T. Karaletsos, and J. H. Macke. Likelihood-free inference with emulator networks. In *Symposium on Advances in Approximate Bayesian Inference*, pages 32–53. PMLR, 2019.

[19] G. Papamakarios, D. Sterratt, and I. Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

[20] G. Papamakarios and I. Murray. Fast e-free inference of simulation models with bayesian conditional density estimation. In *Advances in neural information processing systems*, pages 1028–1036, 2016.

[21] J.-M. Lueckmann, P. J. Goncalves, G. Bassetto, K. Öcal, M. Nonnenmacher, and J. H. Macke. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in Neural Information Processing Systems*, 30, 2017.

[22] D. Greenberg, M. Nonnenmacher, and J. Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pages 2404–2414. PMLR, 2019.

[23] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, Aug. 1952. ISSN 0022-3751, 1469-7793. doi: 10.1113/jphysiol.1952.sp004764.

[24] W. A. Catterall, I. M. Raman, H. P. C. Robinson, T. J. Sejnowski, and O. Paulsen. The Hodgkin-Huxley Heritage: From Channels to Circuits. *Journal of Neuroscience*, 32(41):14064–14073, Oct. 2012. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3403-12.2012.

[25] A. A. Prinz, C. P. Billimoria, and E. Marder. Alternative to Hand-Tuning Conductance-Based Models: Construction and Analysis of Databases of Model Neurons. *Journal of Neurophysiology*, 90(6):3998–4015, Dec. 2003. ISSN 0022-3077, 1522-1598. doi: 10.1152/jn.00641.2003.

[26] M. Pospischil, M. Toledo-Rodriguez, C. Monier, Z. Piwkowska, T. Bal, Y. Frégnac, H. Markram, and A. Destexhe. Minimal Hodgkin–Huxley type models for different classes of cortical and thalamic neurons. *Biological Cybernetics*, 99(4-5):427–441, Nov. 2008. ISSN 0340-1200, 1432-0770. doi: 10.1007/s00422-008-0263-8.

[27] Q. J. M. Huys, M. B. Ahrens, and L. Paninski. Efficient Estimation of Detailed Single-Neuron Models. *Journal of Neurophysiology*, 96(2):872–890, Aug. 2006. ISSN 0022-3077. doi: 10.1152/jn.00079.2006.

[28] C. Rossant, D. F. Goodman, B. Fontaine, J. Platkiewicz, A. Magnusson, and R. Brette. Fitting Neuron Models to Spike Trains. *Frontiers in Neuroscience*, 5:9, 2011. ISSN 1662-453X. doi: 10.3389/fnins.2011.00009.

[29] R. Ben-Shalom, J. Balewski, A. Siththaranjan, V. Baratham, H. Kyoung, K. G. Kim, K. J. Bender, and K. E. Bouchard. Inferring neuronal ionic conductances from membrane potentials using cnns. *bioRxiv*, page 727974, 2019.

[30] A. C. Daly, D. J. Gavaghan, C. Holmes, and J. Cooper. Hodgkin–Huxley revisited: reparametrization and identifiability analysis of the classic action potential model with approximate Bayesian methods. *Royal Society Open Science*, 2(12):150499, Dec. 2015. ISSN 2054-5703. doi: 10.1098/rsos.150499.

[31] J. Oesterle, C. Behrens, C. Schröder, T. Hermann, T. Euler, K. Franke, R. G. Smith, G. Zeck, and P. Berens. Bayesian inference for biophysical neuron models enables stimulus optimization for retinal neuroprosthetics. *eLife*, 9:e54997, Oct. 2020. ISSN 2050-084X. doi: 10.7554/eLife.54997.

[32] M. G. Blum and O. François. Non-linear regression models for Approximate Bayesian Computation. *Statistics and computing*, 20(1):63–73, 2010.

[33] P. J. Diggle and R. J. Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):193–212, 1984.

[34] M. G. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208, 2013.

[35] B. Jiang, T.-y. Wu, C. Zheng, and W. H. Wong. Learning Summary Statistic for Approximate Bayesian Computation via Deep Neural Network. *Statistica Sinica*, 2018. ISSN 10170405. doi: 10.5705/ss.202015.0340.

[36] R. Izbicki, A. B. Lee, and T. Pospisil. ABC-CDE: Towards Approximate Bayesian Computation with Complex High-Dimensional Data and Limited Simulations. *Journal of Computational and Graphical Statistics*, 28(3):481–492, July 2019. ISSN 1061-8600, 1537-2715. doi: 10.1080/10618600.2018.1546594.

[37] C. M. Bishop. Mixture density networks. *Technical Report*, 1994. ISSN NCRG/94/004. Neural Computing Research Group, Aston University.

[38] C. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, New York, 2006. ISBN 978-0-387-31073-2.

[39] R. M. Neal. Slice Sampling. *The Annals of Statistics*, 31(3):705–741, 2003. ISSN 0090-5364.

[40] A. Tejero-Cantero, J. Boelts, M. Deistler, J.-M. Lueckmann, C. Durkan, P. J. Gonçalves, D. S. Greenberg, and J. H. Macke. SBI – A toolkit for simulation-based inference. *arXiv:2007.09114 [cs, q-bio, stat]*, July 2020.

[41] L. Torrey and J. Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[42] B. Jiang. Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *International conference on artificial intelligence and statistics*, pages 1711–1721. PMLR, 2018.

[43] F. Perez-Cruz. Kullback-Leibler divergence estimation of continuous distributions. In *2008 IEEE International Symposium on Information Theory*, pages 1666–1670, Toronto, ON, Canada, July 2008. IEEE. ISBN 978-1-4244-2256-2. doi: 10.1109/ISIT.2008.4595271.

[44] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[45] J. C. Butcher and N. Goodwin. *Numerical methods for ordinary differential equations*. John Wiley \& Sons, 2016.

[46] C. R. Cadwell, A. Palasantza, X. Jiang, P. Berens, Q. Deng, M. Yilmaz, J. Reimer, S. Shen, M. Bethge, K. F. Tolias, R. Sandberg, and A. S. Tolias. Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nature Biotechnology*, 34(2):199–203, Feb. 2016. ISSN 1546-1696. doi: 10.1038/nbt.3445.

[47] F. Scala, D. Kobak, M. Bernabucci, Y. Bernaerts, C. R. Cadwell, J. R. Castro, L. Hartmanis, X. Jiang, S. Laturnus, E. Miranda, S. Mulherkar, Z. H. Tan, Z. Yao, H. Zeng, R. Sandberg, P. Berens, and A. S. Tolias. Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature*, Nov. 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2907-3.

[48] S. R. Bittner, A. Palmigiano, A. T. Piet, C. A. Duan, C. D. Brody, K. D. Miller, and J. P. Cunningham. Interrogating theoretical models of neural computation with emergent property inference. *bioRxiv*, page 837567, 2021.

[49] D. B. Rubin, S. D. Van Hooser, and K. D. Miller. The Stabilized Supralinear Network: A Unifying Circuit Motif Underlying Multi-Input Integration in Sensory Cortex. *Neuron*, 85(2):402–417, Jan. 2015. ISSN 0896-6273. doi: 10.1016/j.neuron.2014.12.026.

[50] M. S. Goldman, J. Golowasch, E. Marder, and L. F. Abbott. Global Structure, Robustness, and Modulation of Neuronal Models. *The Journal of Neuroscience*, 21(14):5229–5238, July 2001. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.21-14-05229.2001.

[51] H. Ori, E. Marder, and S. Marom. Cellular function given parametric variation in the Hodgkin and Huxley model of excitability. *Proceedings of the National Academy of Sciences*, 115(35):E8211–E8218, Aug. 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1808552115.

[52] A. L. Taylor, J.-M. Goaillard, and E. Marder. How Multiple Conductances Determine Electrophysiological Properties in a Multicompartment Model. *Journal of Neuroscience*, 29(17):5573–5586, Apr. 2009. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.4438-08.2009.

[53] S. Druckmann, Y. Banitt, A. Gidon, F. Schürmann, H. Markram, and I. Segev. A novel multiple objective optimization framework for constraining conductance-based neuron models by experimental data. *Frontiers in Neuroscience*, 1:1, 2007. ISSN 1662-453X. doi: 10.3389/neuro.01.1.1.001.2007.

[54] R. Ben-Shalom, A. Aviv, B. Razon, and A. Korngreen. Optimizing ion channel models using a parallel genetic algorithm on graphical processors. *Journal of Neuroscience Methods*, 206(2):183–194, 2012. Publisher: Elsevier.

[55] P. Achard and E. De Schutter. Complex parameter landscape for a complex neuron model. *PLoS computational biology*, 2(7):e94, 2006. Publisher: Public Library of Science San Francisco, USA.

[56] R. J. Butera Jr, J. Rinzel, and J. C. Smith. Models of respiratory rhythm generation in the pre-Botzinger complex. II. Populations of coupled pacemaker neurons. *Journal of neurophysiology*, 82(1):398–415, 1999. Publisher: American Physiological Society Bethesda, MD.

[57] A. A. Hill, J. Lu, M. A. Masino, O. H. Olsen, and R. L. Calabrese. A model of a segmental oscillator in the leech heartbeat neuronal network. *Journal of computational neuroscience*, 10(3):281–302, 2001. Publisher: Springer.

[58] S. H. Jezzini, A. A. Hill, P. Kuzyk, and R. L. Calabrese. Detailed model of intersegmental coordination in the timing network of the leech heartbeat central pattern generator. *Journal of neurophysiology*, 91(2):958–977, 2004. Publisher: American Physiological Society.

[59] W. R. Foster, L. H. Ungar, and J. S. Schwaber. Significance of conductances in Hodgkin-Huxley models. *Journal of Neurophysiology*, 70(6): 2502–2518, Dec. 1993. ISSN 0022-3077. doi: 10.1152/jn.1993.70.6.2502.

[60] A. L. Taylor, T. J. Hickey, A. A. Prinz, and E. Marder. Structure and Visualization of High-Dimensional Conductance Spaces. *Journal of Neurophysiology*, 96(2):891–905, Aug. 2006. ISSN 0022-3077, 1522-1598. doi: 10.1152/jn.00367.2006.

[61] L. M. Alonso and E. Marder. Visualization of currents in neural models with similar behavior and different conductance densities. *eLife*, 8:e42722, Jan. 2019. ISSN 2050-084X. doi: 10.7554/eLife.42722.

[62] E. Hay, S. Hill, F. Schürmann, H. Markram, and I. Segev. Models of Neocortical Layer 5b Pyramidal Cells Capturing a Wide Range of Dendritic and Perisomatic Active Properties. *PLOS Computational Biology*, 7(7): e1002107, July 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002107.

[63] D. J. Rezende, G. Papamakarios, S. Racanière, M. S. Albergo, G. Kanwar, P. E. Shanahan, and K. Cranmer. Normalizing Flows on Tori and Spheres. *arXiv:2002.02428 [cs, stat]*, July 2020. arXiv: 2002.02428.

[64] O. Lombardi, F. Holik, and L. Vanni. What is Shannon information? *Synthese*, 193(7):1983–2012, July 2016. ISSN 0039-7857, 1573-0964. doi: 10.1007/s11229-015-0824-z.