Inferring microscale parameters from EEG-data with the help of simulation-based inference

Master thesis by Katharina Anderer

First supervisor: Prof. Jakob Macke

Second supervisor: Prof. Martin Butz

Tübingen, 16.05.2022

Abstract

Attempts trying to understand how the human brain works are often of a reductionist manner. The working of a single neuron is nowadays well understood, but how neurons engage in network dynamics and how macro-level signals emerge, like e.g. from an EEG, remains largely unanswered yet [Einevoll et al., 2019]. In this work, we investigate how to fill the gap between micro-level signals and macro-level signals, with the help of simulation-based inference. A mechanistic model of the brain is used in order to simulate an event-related potential, where 17 micro-parameters are left free to vary in a predefined range. We further propose an adapted version of SNPE [Greenberg et al., 2019] that reduces simulation time and performs comparatively.

With this approach, micro-scale processes of the brain can be recovered by inferring them from an EEG-signal. In comparison to optimization algorithms where point estimates for parameters are derived, SBI aims to recover the full solution space. Degeneracy and compensation mechanisms [Marder and Taylor, 2011] are biologically very plausible mechanisms in the brain, such that we need to have good methods for discovering them. We will show how we can investigate compensation mechanisms between parameters of the mechanistic model that we used.
Further, we will compare parameter settings between different conditions within an experimental paradigm and show that the inferred posteriors can recover and distinguish the observations from these conditions.

Declaration

I hereby declare that I have written this thesis by my own, that I have not used any aids and sources other than those indicated and that I have marked all statements taken verbatim or in spirit from other works as such.

---

date, place, signature

Contents

Introduction

Understanding how macro-scale signals evolve from micro-scale parameters is an interesting question in many domains, e.g. in research about global climate, gene expression, epidemiology or brain phenomena like the signal coming from an electroencephalography (EEG). The later is an example for a macro-scale signal of the brain that evolves through the activation of many neurons that fire in parallel. In more depth, it measures the intracellular current flow in the long and spatially-aligned pyramidal neuron dendrites [Neymotin et al., 2020]. While macro-scale signals are the product out of the combination of many signals, we are often interested in the origins of these signals and the underlying mechanisms. These mechanisms can be described by a mechanistic model that meets assumptions about e.g. the information flow circuits, the morphology of the cells in the brain or the weights between different neurons [Kohl et al., 2022].

During the last years, lots of approaches, grouped under the term 'simulation-based inference' (SBI), were developed with the aim to combine mechanistic models with statistical approaches in order to study the emergence of signals that are composed of micro-scale signals [Goncalves et al., 2018, Gonçalves et al., 2020, Greenberg et al., 2019]. SBI is also known under the term likelihood-free inference as it gives a solution to approximate the likelihood [Papamakarios et al., 2019] or the posterior [Greenberg et al., 2019, Goncalves et al., 2018, Papamakarios and Murray, 2016], in cases where the likelihood is intractable.

The assumptions for a mechanistic model are usually provided through invasive animal studies [Kohl et al., 2021]. Simulating macro-scale signals based on a mechanistic model, is trying to synergise micro-scale dynamics with macro-scale dynamics and to provide information about how macro-scale signals evolve from micro-signal processes [Kohl et al., 2021]. An example of a mechanistic model is the Human Neocortical Neurosolver (HNN), developed by Neymotin et al. [2020]. As we will exclusively work with the HNN model for this work, it will be described in detail in Section 3.2.1.

Given a model like the HNN, signals like an EEG or MEG can be simulated based on it. The model makes assumptions about the architecture and information flow processes in the brain. These simulations can then be used to evaluate which micro-scale parameters have likely caused a certain EEG or MEG signal. As we have many different parameters involved and further, signals like EEG or MEG have a stochastic nature, the likelihood function $p(x|\Theta)$ is intractable as one would have to

trace every possible parameter set and compute the integral of this [Cranmer et al., 2020].

For an approximation of the likelihood or the posterior density, neural networks can be used. One particular neural density estimation technique is called normalizing flows, in which one starts with a simple base distribution which is put into the network and then transformed through multiple inverse transformations that have a tractable Jacobian [Cranmer et al., 2020]. In this work, we will use neural spline flows [Durkan et al., 2019] and masked autoregressive flows [Papamakarios et al., 2017], which are two specific forms of normalizing flows.

Compared to ABC methods, neural density estimation has several advantages. First, we do not only get a representation of the posterior as a set a samples, but instead derive a parameterized posterior that can also be combined with previous calculated posteriors in principle [Papamakarios and Murray, 2016]. Second, ABC methods get more computationally expensive the smaller ϵ is chosen [Papamakarios and Murray, 2016]. ϵ describes the allowed distance between an observation and a simulation. The smaller ϵ, the more exact the approximation, but also the more simulations needed [Papamakarios and Murray, 2016]. Neural density estimation does not suffer from this problem. The simulations are directly used to learn the posterior, by maximizing the probability of parameter vectors under a particular observation [Papamakarios and Murray, 2016].

The motivation for using simulation-based inference in EEG-analysis is mainly driven by the fact that we want to investigate how microscale parameters dynamically interact and how they compose to a certain macroscale signal. Another advantage over other methods lays in the possibility of SBI to investigate the whole parameter space. We therefore not only get single point estimates for the parameters, but instead derive densities that can be of interesting shape. Gonçalves et al. [2020] showed that the joint density of two parameters can reveal a certain relationship. One parameter might compensate the change of another parameter and the end result will still be the same. As we move along the path of high probability, the combination of the two parameters is still highly plausible, even that the parameter values can vary a lot [Gonçalves et al., 2020]. Therefore, a strength of SBI is also its capability to visualise compensation and interaction mechansims. A good reason to investigate compensation mechanisms is given by Marder and Taylor [2011]. They argue that there are multiple solutions for similar outputs in the brain [Marder and Taylor, 2011]. This degeneracy makes biologically sense in that it makes a system capable to react to perturbations [Marder and Taylor, 2011]. Marder and Taylor [2011] further argue that the more (conductance) parameters a model has, the more likely there will be a path along the joint parameter space that allows for homeostatic processes

in the brain.

The HNN simulator is based on a model with many different parameters. If we want to infer the parameter space of all these parameters, we have to deal with the issue of high dimensionality. That is, it gets difficult to infer parameter values that have led to a certain signal because there are exponentially many possible combinations of parameters. We need more simulations, the more parameters we want to infer. To tackle this problem, we propose a new approach, where we divide the inference process in temporally diverse steps. For our model and the parameters that we use, we make the assumption that some parameters do not exert any influence on the signal before a certain point in time. We will explain this approach in more depth in Section 3.1 and show that it helps to reduce simulation time and that it restricts the parameter space to a higher degree than the SNPE approach, proposed by Greenberg et al. [2019].

What is also needed to tackle the problem of high dimensionality are good summary statistics. Usually, summary statistics rather than all of the data is used in order to reduce the resolution of the data. Sufficient statistics must capture all of the main characteristics of the data, while at the same time reduce the high dimensionality of the data [Lueckmann et al., 2021]. To find summary statistics, one could either use so called embedding networks that automatically try to find sufficient features with the help of deep learning, or define hand-crafted features that are transparent and can be tailored to the specific problem at hand.

There are some interesting approaches to automatically learn summary statistics, which has the advantage that there is no need to design them carefully. This, however, suffers from the issue of non-transparency.
There are also methods that try to learn permutation invariant summary statistics with the help of deep neural nets [Radev et al., 2020]. Besides, a method of 'deep signature statistics' was proposed by Dyer et al. [2021] where signature transforms are embedded and summary statistics and posterior estimates are learned concurrently.

Concentrating on hand-crafted summary statistics in this work, we investigate reasonable summary statistics and also embed them in our new incremental approach, such that for certain parameters only a particular set of summary statistics is taken into account.

There are different EEG signals that one can investigate. Among the extensively studied signals are event-related potentials (ERPs) and oscillatory signals like alpha waves [Ziegler et al., 2010]. It has been studied that these signals are typically altered with aging [Ziegler et al., 2010] and through diseases like schizophrenia [Han-

lon et al., 2005, Hamilton and Northoff, 2021]. Therefore, learning how an ERP is composed by underlying brain processes could teach us more about e.g. the development of diseases. This work exclusively studies event-related potentials, focusing on the 200 ms after the occurrence of an event. The term 'event' refers to a sensory stimulus that can be e.g. tactile or visual. The 200 ms time range under study has 3 characteristic peaks, usually referred to as the P50, N100 and P200 as they happen around 50 ms, 100 ms, and 200 ms after the stimulus (the event) that was evoking it. An event-related potential is only derived after averaging many trials in which the same stimulus is presented several times. This reduces the random noise of the brain and makes the potentials measurable. The P50, N100 and P200 components are associated with different steps of stimulus processing. Whereas the P50 component has been associated with e.g. sensory gating [Cadenhead et al., 2000], the N100 component is amongst others associated with selective attention [Thornton et al., 2007].

This work therefore proposes a SBI pipeline in order to find the underlying micro-parameters of an event-related potential. This might open the door to study cognitive processing like selective attention on the micro-scale level, only from an EEG-signal. This might provide 'bridges between levels of understanding' [Dayan and Abbott, 2005], and moves away from purely reductionist attempts to study to brain.
The code and data will be made available here: `https://github.com/mackelab/sbi_for_eeg_data`.

Related work

Simulation-based inference (SBI) methods are currently developing fast [Cranmer et al., 2020]. Yet, there are still few papers that apply SBI to neuroscience research [Lueckmann et al., 2021, Jallais et al., 2021, Goncalves et al., 2018, Schröder et al., 2019].
Approximate Bayesian computation (ABC), that requires a rejection criterion and model-specific algorithms, is still more prevalent among the neuroscience and cognitive science community [West et al., 2021, Kangasrääsiö et al., 2017]. However, these classical ABC methods cannot efficiently scale up to a high parameter space because a great amount of the simulations gets rejected and cannot be used for inference.

The development of Sequential Neural Posterior Estimation (SNPE) [Greenberg et al., 2019, Lueckmann et al., 2017, Papamakarios and Murray, 2016] made these approaches more efficient and applicable to high-dimensional problems. Lueckmann et al. [2017] showed how SBI can be used to study micro-scale neural dynamics and how it can be used to derive a model to predict voltage traces of neurons. A paper by Gonçalves et al. [2020] expanded this idea and showed that SNPE can scale to complex neuronal models such as receptive fields, ion channels, and Hodgkin-Huxley models. Gonçalves et al. [2020] further showed that one can also identify processes such as homeostatic regulation and compensation mechanisms between parameters, through studying the correlations of the derived posteriors. In the case where many parameters are interacting with each other, degeneracy often occurs such that there are multiple parameter settings that can all produce the same outcome [Marder and Taylor, 2011, Gonçalves et al., 2020]. The phenomenon of degeneracy can be studied by SBI methods because one can investigate the paths of highest densities of the posteriors that can reveal if multiple solutions are possible [Gonçalves et al., 2020].

SBI research enters more fields of neuroscience research recently. There has been, e.g., a study by Jallais et al. [2021] that uses simulation-based inference to investigate certain parameters of grey matter in the brain, which could be an interesting application for dementia or Parkinson diagnosis and treatment.

Attempts to model the brain in a mechanistic way started with the well-known work of Hodgkin and Huxley [1952]. They developed a model for predicting the generation of action potentials in the squid giant axon [Hodgkin and Huxley, 1952]. Since then, a couple of biophysical models were developed to model specific neuron types of certain brain areas like the sensory cortex or the thalamus [Einevoll et al., 2019].

More recently, there have been attempts to simulate neurons in the brain on a larger network scale, e.g. the Blue Brain project that is modelling the neocortial columns of the human brain [Schürmann et al., 2007].

Other prestigious projects, that follow an attempt to simulate networks of neurons on a large-scale, include the European Union's Human Brain Project [Markram et al., 2011] and the project MindScope, that is developed by the Allen Brain Institute [Einevoll et al., 2019].

The HNN simulator, proposed by Neymotin et al. [2020], was developed to simulate the electrical activity of the neocortical cells and circuits. We propose a pipeline that connects SBI with the HNN simulator in order to investigate how micro-processes in the brain interact with each other. This enables the study of neurological diseases on the micro-scale level and can therefore provide a powerful tool for clinicians and drug-makers.

Methods

3.1 Neural Incremental Posterior Estimation (NIPE)

As the parameter space gets larger, the number of simulations needed for inference grows exponentially fast. We can approximately calculate the number of simulations that is needed to draw at least some samples close to the true parameter. If we want to infer 1 parameter, using a uniform prior, and we want to sample draws within +-15% around the true parameter, only 30% of our draws will be in this region. We can also calculate the percentage of samples within the target region of +-15% around the true parameters for a larger number of parameters, using the formula $percentage^{\#\ parameters}$, which yields the following percentage for e.g. 18 parameters: $0.3^{18} = 3.87 \cdot 10^{-10}$. We would therefore need around $2.6 \cdot 10^9$ simulations in order to get 1 draws within the target region.

Our approach tries to reduce this number of needed simulations by making the assumption that a time order of parameters exist such that some parameters come into play not before a certain time. Parameters occurring later in this time order, do not exert an influence on parameters occurring earlier in the order. This assumption is well justified for the HNN model where the micro-parameters belong to either a forward- or feedback-signal that is firmly placed in time. We will split the inference process into steps such that not all parameters are inferred together right from the beginning, but such that we incrementally increase the number of parameters that are taken into account for inference.

Let's assume that parameters belonging to the subset Θ_1 mainly play a role in the time up to time t and parameters belonging to subset Θ_2 mostly exert their influence in the time range between t and m, where $t < m$.

We do not work with the whole data x, but instead use summary statistics $s(x)$ that describe the data. We define them such that a time order exists and e.g. $s_{1:t}(x)$ holds only information up to time t.

Further, we assume that the first summary statistics are not dependent on Θ_2, such that

$$p(s_{1:t}|\Theta_1, \Theta_2) = p(s_{1:t}|\Theta_1) \tag{3.1}$$

First, we start by learning $p(\Theta_1|s_{1:t}(x))$. We argue that mainly the summary statistics relevant in the time up to t are informative about Θ_1, although later summary statistics could be informative as well. Even if later summary statistics are

informative, such that $p(\Theta_1|s_{1:m}(x)) \neq p(\Theta_1|s_{1:t}(x))$, we can already gain a good estimate of the parameter posterior that can be further restricted in subsequent rounds.

The entropy H of the full posterior $p(\Theta_1|s_{1:t}(x), s_{t:m}(x))$ is always higher compared to a posterior that is conditioned on particular features:

$$H(X|Y) \leq H(X) \Leftrightarrow p(\Theta_1|s_{1:t}(x)) \leq p(\Theta_1|s_{1:m}(x)) \tag{3.2}$$

, such that supposing that $s_{1:t}(x)$ is most informative for Θ_1, we should already be able to restrict the posterior of the first subset to a certain degree. In later steps, all features will be included, such that they can further restrict the posterior, if they entail any information about the parameters within Θ_1.

To infer $p(\Theta_1|s_t(x))$, parameters are drawn from a proposal and then put into a simulator to gain observations x. Based on the sampled thetas and simulated observations, a neural density network then learns the posterior.

The posterior is derived by the neural net estimator under the minimized negative log-likelihood $\mathcal{L}(\phi) = -\sum_{j=1}^{N} \log q_{F(x_j,\phi}(\theta_j)$ [Greenberg et al., 2019]:

$$\hat{p}(\Theta_1|s_{1:t}) = q_{F(s_{1:t},\phi)}(\Theta_1) \tag{3.3}$$

In a second step, we aim to learn $p(\Theta_1, \Theta_2|s_{1:t}(x), s_{t:m}(x))$. Building up an the idea to use an already restricted proposal, we combine the posterior $p(\Theta_1|s_{1:t}(x))$, gained in the first step, with the prior of the next subset $p(\Theta_2)$. We therefore define a new distribution class that samples the thetas belonging to Θ_1 from $p(\Theta_1|s_{1:t}(x))$ and the thetas belonging to Θ_2 from $p(\Theta_2)$. This new distribution class then acts as a proposal for the second step. See Fig.3.1 for a sketch of this.

The proposal is defined as follows, assuming an independent prior:

$$\hat{p}(\Theta_1, \Theta_2) = p(\Theta_1|s_{1:t}(x)) \cdot p(\Theta_2) \tag{3.4}$$

The log-likelihood of the new distribution class is defined as:

$$\mathcal{L}_{\Theta_1,\Theta_2} = \mathcal{L}_{(\Theta_1|s_{1:i})} + \mathcal{L}_{\Theta_2}$$

The neural net is again trained on all pairs of thetas and simulated observations x. Then, the posterior is derived by the neural net again under the minimized loss $\mathcal{L}$:

$$\hat{p}(\Theta_1, \Theta_2|s_{1:m}(x)) = q_{F(s_{1:m},\Phi)}((\Theta_1|s_{1:t}(x)), \Theta_2) \tag{3.5}$$

Now, we include all summary statistics $s_{1:m}(x)$ to infer the posterior. The parameters within the first subset Θ_1 can now still get informed by $s_{t:m}(x)$.

As we do not include any importance weights for the proposal, $\hat{p}(\Theta_1, \Theta_2|s_{1:m})$ yields only an approximation.

The second step is repeated if the parameters are separated into more than two subsets.

Coming back to the example from above with 18 parameters, we could split e.g. into 3 subsets of 6 parameters. For the first step, we then have a $0.3^6 = 0.000729$ chance to be within the target region for all 6 dimensions.
We then need around 1400 simulations to get 1 draw within the target region of the first subset. Optimally, the parameter space of the first set already gets well restricted such that mainly the parameters of the second set need to get restricted during the second step. In the ideal case where the first step already leads to a parameter restriction such that the posterior spcae does not capture more than the target region, we would again need 1400 simulations in the second step to get a draw within the target region for all dimensions of the first two subsets. This ideal case is very optimistic and only serves an illustrative purpose. We cannot be sure that the parameter space gets already well restricted in step 1. In comparison to the case where all parameters are inferred at the same time, roughly 10^5 times less simulations would be needed in the illustrative ideal case.

As we have also defined a time order of the summary statistics, and we assume that for a particular parameter set only the time series up to a certain time is relevant, we can further use early stopping of the simulator, where we interrupt the simulator after a certain time t. For the first subsets, we then e.g. only simulate 70 ms instead of the full time series of 200 ms. This should increase simulation efficiency.

In total, NIPE should need less simulations because the parameter space is much smaller for a subset that is only a half or a third of the total number of parameters. Starting with a small subset, we aim to restrict the parameter space of the first subset to a significant degree such that later inference with more parameters will be easier. Together with the use of early stopping of simulations, this should increase simulation-efficiency and possibly make inference of larger parameter sets more approachable.

We will investigate efficiency in the context of speed-up and quality of inference, comparing this new approach to the SNPE approach, proposed by Greenberg et al. [2019].

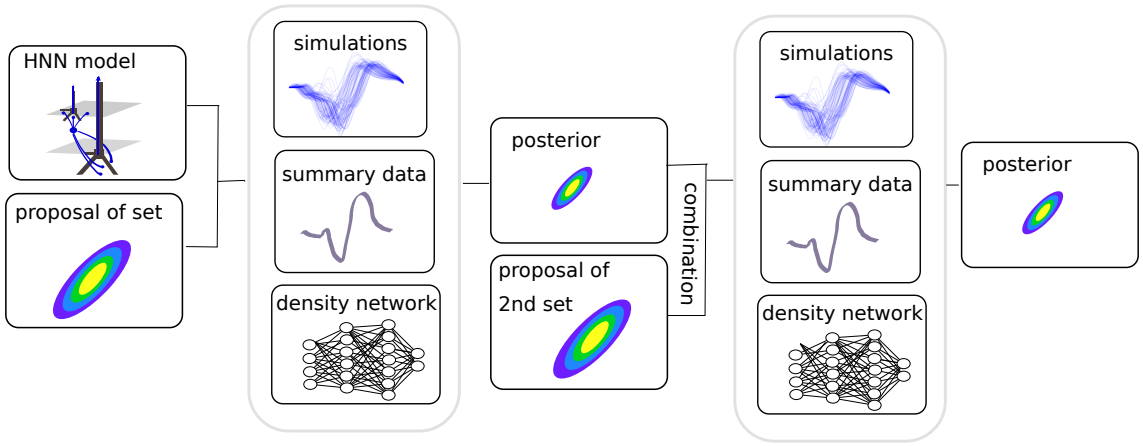

Figure 3.1: **Neural Incremental Posterior Estimation.**
Visualization of the NIPE approach: HNN model meets assumptions about the brain architecture and cell conductances. The model then simulates event-related potentials with parameters that are sampled from the proposal. Summary statistics are calculated from the simulated time series from which the density estimator takes its information to infer the posterior. The inferred posterior is then combined with the prior of the next parameter set. Illustrates the approach for two subsets. There could be more subsets, though.

Algorithm 1: Neural Incremental Posterior Estimation.
M is the number of theta subsets, whereas N is the number of simulations for each step. If i is bigger than 1, all thetas from previous rounds are sampled from the posterior of the last round.

---

for $i = 1{:}\ M$ **do**
  for $j = 1{:}\ N$ **do**
    $\Theta_{ij} \sim p(\Theta_i)$;
    If $i > 1$:
      $\Theta_{\{1:i\}j} \sim \hat{p}(\Theta_{\{1:i\}}|x)$
    $x_j \sim p(x|\Theta_j)$
  end
  train $q_\Phi(\Theta|x = x_0)$ on $\{\Theta_n, x_n\}$
  Set $\hat{p}(\Theta_i|x) = q_{F(x,\phi)}(\Theta_i)$;
end

3.1.1 Simulation budgets

The thetas within the first subset are more likely to be well restricted compared to later subsets within our proposed approach. There are multiple reasons for this. First, inference in the first step is only on a small number of parameters. Therefore

the parameter space is smaller and inference is easier. In the second step, inference is made on a greater number of parameters, which is in general much harder. If we expect interactions and dependencies between parameters, inference is possibly also less confident, because inference on later subsets is dependent on the quality of inference of previous subsets. Further, we go on to restrict the parameter space for the first subset in the second and third step. The restricted space of the first subset is combined with the prior of the next subset and gets again restricted in the next step. This might be the cause of an over-dispersion of the thetas belonging to the last subset (see Section 4.1.2). We therefore introduced different simulation budgets for the different steps. For the first step, we then use e.g. only a $\frac{1}{7}$ fraction of the number of simulations that would have been used with equally distributed budgets and for the last round a $\frac{1}{13}$ fraction, such that we again arrive at the same total number of simulations. The last step, that is the most difficult for inference with the largest number of parameters, gets more emphasis with a reallocation of simulation budgets. We will show that this can help to reduce over-dispersion of the estimated posterior variances.

3.1.2 Comparison to SNPE

For the NIPE approach, inference is not made for all parameters at the same time as for the SNPE approach. Simulations of event-related potentials by the HNN simulator are quite costly. A single 200 ms simulation takes already about 58 seconds on a standard CPU core. We can make the process more efficient by first reducing the number of simulations that are needed for inference, and second by early stopping as explained previously. Therefore, we argue that NIPE should have an advantage with respect to time- and simulation-efficiency, compared to SNPE.

One disadvantage of the approach is that we have to train the neural density estimator from scratch for each step because we expect interactions between parameters. Therefore, we cannot just infer all subsets separately, but have to include all parameters step-by-step. Inference time is comparatively very low with respect to simulation time, as we will show, such that retraining from scratch is not a huge time factor. Another advantage of SNPE is that it can train from data of all rounds, combining the loss terms of different rounds by simply adding them together [Greenberg et al., 2019]. NIPE cannot reuse the simulated data from earlier steps because different numbers of thetas are used in each step.

In the following sections, we first introduce a toy example that is used for comparing the efficiency and accuracy of inference between SNPE and NIPE. Then, we move on to inferring parameters of the HNN model from event-related potentials.

3.1.3 Gaussian toy example

As we have a smaller parameter space for the first two steps of NIPE, we should need less simulations to arrive at an equally good posterior estimation, compared to a multi-round SNPE approach.

To test this hypothesis, we used a simple toy example with 15 Gaussians, all having a mean between 0 and 100 and a standard deviation of 1. For NIPE, we started with the inference of 5 parameters. After approximating the posterior, we then sampled from the posterior for the first 5 parameters and sampled from the prior of the next 5 parameters. For the third step, we then sampled from the posterior derived after the second step, combined with sampling from the prior of the last subset. For the multi-round SNPE approach, all 15 parameters were inferred at the same time.

In order to test the quality of inference, we computed the Kullback-Leibler divergence between the analytic posterior and the inferred posterior for both approaches and plotted this for a different number of simulations. The Kullback-Leibler divergence between two Gaussian distributions is calculated in the following way:

$$KL(q||p) = \log \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x^2 + (\mu_x - \mu_y)^1}{2 \cdot \sigma_y^2} - \frac{1}{2} \tag{3.6}$$

We further tested the NIPE approach with another toy example, a piecewise linear function, and compared it again to the SNPE approach. Method and results for this can be found in the appendix 6.1. The piecewise linear function is more related to time-series data, with which we will proceed in the next section. The Gaussian toy example, however, was easier to investigate in terms of the KL-divergence and the number of needed simulations.

3.1.4 Validation checks

In order to evaluate the quality of inference, we made some checks that are described below.

Posterior predictive checks

Posterior predictive checks simulate observations from samples drawn from the posterior and check if the true observation x_o is laying within the support of the posterior. Further, we simulate observations from samples drawn from the prior and check how the posterior is restricting the area of support.

Histograms of the summary statistics

If the summary statistics do not capture the data well, it is likely that the learned inference network does not mirror the whole complexity of the real world data and thus, one gets a misspecified model [Schmitt et al., 2021].

A simple check for the the quality of the chosen summary statistics is based upon plotting histograms of the single summary statistics. For deriving the histogram plots, several samples are drawn from the posterior and the prior. In a next step, we simulate data from these samples and calculate the summary statistics from the simulated data.

Plotting the range of values for each single summary statistic coming from the posterior versus the ones that are coming from the prior, shows how inference restricts the parameter space.

Ideally, summary statistics should strongly restrict the values, but at the some time capture the statistics that are calculated from the observation on which we conditioned on.

Correlation matrices and density plots

In order to check for compensation mechanisms between parameters, we used (conditional) correlation matrices and density plots. Whereas correlation matrices indicate correlations when all parameters are free to vary, conditional correlation matrices show correlations when all but two parameters are fixed. As the conditional correlation matrices are based on a single sample, one should ideally average over several conditional correlation matrices to have a robust result. We always used an average over 5 correlation matrices.

3.1.5 Calibration check

The 'golden standard' to check how well posteriors are calibrated is simulation-based calibration (SBC).

For SBC, posterior samples have to be drawn according to an observation drawn from the prior [Talts et al., 2018]. Sequential methods are not amortized and the posterior cannot be conditioned on different observations. Applying SBC is therefore not straightforward for sequential methods. Therefore, we could neither use SBC for checking the posteriors derived with SNPE nor for NIPE.

If the true posteriors are known as in our Gaussian toy example, we however have the possibility to check if the variances are underestimated or overestimated by the posteriors. We plotted estimated variances for both approaches and compared them to the analytic variance. In order to check the robustness of results derived by SNPE and NIPE, we inferred 5 different posteriors for each approach and then visualized

the estimated variances with box plots. The box plots show a box for the range from the first to third quartile of the data with extended lines for visualizing + - 1.5 times of this range. The median, as well as data points laying outside the quartiles are visualized as well.

We compare the variances of the 1d marginals of each parameter posterior and investigate if the posteriors derived by SNPE and NIPE are well calibrated or if they show under- or over-dispersion.

3.2 Event-related potentials

For inferring the parameters of an event-related potential, a simulator based on a biophysical model is needed. We next describe the biophysical model that we used and then move on to describe our hand-crafted summary statistics and the parameters of interest. The last part of the section describes the paradigm that we used to compare the micro-parameters of different experimental conditions.

3.2.1 HNN simulator

The Human Neocorical Neurosolver (HNN), developed by Neymotin et al. [2020], is a simulator for macro-scale signals like event-related potentials or oscillatory dynamics.
It is based on a model that tries to represent the neocortical circuits of pyramidal neurons and interneurons. The model has a 3-layered structure with pyramidal neurons and inhibitory interneurons in a 3-to-1 ratio of pyramidal to inhibitory cells [Neymotin et al., 2020]. The 3 layers that are modeled are Layer 2/3 (also referred to as supragranual layer), Layer 4 and Layer 5 (also referred to as infragranular layer). The morphology of the model is based on the cat's visual cortex pyramidal cells, but adapted for the human brain [Kohl et al., 2021].
The HNN model distinguishes between so called proximal drives, coming from the thalamus and signaling to the supragranular layers of the cortex, and so called distal drives, representing cortical-cortical inputs or non-lemniscal thalamic drives that signal directly into the supragranular layers and from there further downwards to the infragranular layers [Neymotin et al., 2020].
Event-related potentials are usually composed of a sequence of proximal and distal drives. For each drive, there are up to 10 parameters that can be tuned for the HNN model. These include the onset of the drive, the number of spikes and the weights of synaptic inputs to the specific layers [Neymotin et al., 2020]. Proximal drives are usually associated with positive peaks in an event-related potential. Regarding the first 200 ms, there are two characteristic positive peaks in the signal - the P50 and the the P200 component. These are related to two different proximal drives.

The N100 component, in contrast, is related to a distal drive and reflects a negative potential.

HNN is based on the NEURON environment. Taken on from NEURON, membrane voltages are based on Hodgkin-Huxley equations and current flow between compartments is modeled by cable theory [Neymotin et al., 2020].
Further, the model captures different ion channels like Na, K, Km, KCa and others and codes the thresholds for these [Neymotin et al., 2020].

Stochasticity. The onset of the 3 evoked drives can be made stochastic. This is done by setting `sigma`> 0 for the `net.add_evoked_drive` function. Increasing `sigma` leads to an increased divergence of the neurons spiking times. This means that the higher `sigma`, the less parallel will be the firing of the neurons within the same layer. `sigma` can be included as a parameter for the inference pipeline. In the optimization tutorial[1], it is claimed that increasing or decreasing sigma is sometimes necessary to fit the ERP curve. A larger sigma of e.g. the first proximal drive leads to a wider shape of the P50 component, whereas a smaller sigma leads to a sharp peak.

If `sigma` should be included as a free parameter is highly disputable because it is not a parameter that can be directly manipulated biologically, in comparison to e.g. NMDA weights. A larger or smaller sigma should ideally be generated by other parameters for which we have a direct, biological relation. Further, as we can directly make the link between how to change sigma in order to get a wider or sharper peak, one somehow gets the output that was given as input, which seems a bit dubious. However, whether to include sigma or not, is not trivial. Sigma can be seen as the 'degree of synchronization' between neurons. If we do not know the biological cause of synchronization, we might need to include a parameter without a biological relation in order to model this. For most our experiments, we did not include sigma as a parameter and set it as a constant. However, we also tested our inference pipeline with the inclusion of sigma as a parameter and will discuss this again later on.

We used a smoothing window of 30 ms for the simulated time series and a scaling factor of 3000.

3.2.2 Summary statistics

Faced with a time series of over 8000 values, it is important to reduce the input space of our data and to find summary statistics that represent the data sufficiently well. A sufficient representation is given when the posterior under the summary statistics $p(\Theta|s(x))$ is equal to the posterior $p(\Theta|x)$ without the summary statistics [Lueckmann et al., 2021].

We embed domain knowledge about the main characteristics of an ERP in the

- time of the P50 peak	- time of the N100 peak
- time of the P200 peak	- amplitude of the P50
- amplitude of the N100	- amplitude of the P200
- mean of time range around P50 (from 10 ms before till 10 ms after)	- mean of time range around N100 (from 10 ms before till 10 ms after)
- variance of time range around P50	- variance of time range around P200
- mean of time range around P200	- variance of time range around N100

Table 3.1: **Summary statistics.** This is an overview of the main summary statistics that were used. Further, we also calculated mean values of some time ranges.

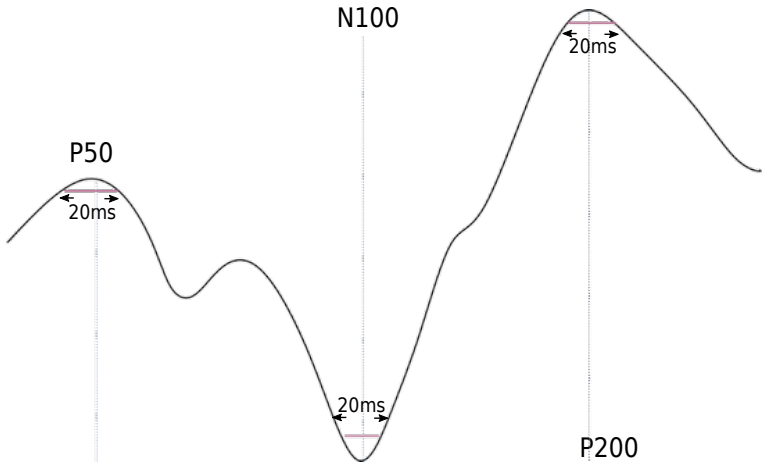

Figure 3.2: **Scheme of summary features.**
Time and amplitude of the P50, N100 and P200 are taken into account. Mean and variance of the curve for the time 10 ms before till 10 ms after P50, N100 and P200 are calculated as well.

process of defining relevant summary statistics and mainly try to capture the important statistics of the before-mentioned peaks of an ERP. An overview of the summary statistics that have been used can be found in Table 3.1.

Arguably, one could also only reduce time resolution by a vast factor and check if this is already enough in order to capture the main characteristics. The summary statistics described above should, however, be more robust to time displacements of signals having similar characteristics and are specific and transparent.

Figure 3.2 illustrates the relevant summary statistics. They aim at capturing the timing of the P50, N100 and P200 components as well as at representing the amplitude and shape of the waveform. Further, mean values were calculated for time ranges where often a dent or local extreme in the signal was present.

step 1	step 2	step 3
prox1 ampa L2 basket	dist ampa L2 pyr	prox2 ampa L2 pyr
prox1 ampa L2 pyr	dist ampa L2 basket	prox2 ampa L5 pyr
prox1 ampa L5 basket	dist nmda L2 pyr	prox2 nmda L2 pyr
prox1 nmda L5 basket	dist nmda L5 pyr	prox nmda L5 pyr
prox1 nmda L5 pyr	dist nmda L2 pyr	time prox2
time prox1	dist nmda L2 basket	
	time distal	

Table 3.2: **Micro-scale parameters**. List of parameters that were inferred. In the first column, all parameters are associated with the first proximal drive, whereas in the second column, all parameters are associated with the distal drive, and in the last column, the parameters are associated with the second proximal drive.

3.2.3 Parameters

As we wanted to have a setting that we could easily compare to the optimization procedure that was used in the tutorial by the Jones Neurolab[1], we took over the most important parameters from there. A list of all parameters can be found in Table 3.2.

For SNPE, all parameters are inferred at the same time and for NIPE, we start with the five parameters in the first column. The following paragraph should emphasize that the parameters are bound to a certain time range and parameters in the second column come into play later in the time sequence, compared to parameters in the first column. Parameters within the same column belong to the same subset. 'Prox1' refers to the first proximal drive, 'dist' to the distal drive and 'prox2' to the second proximal drive. All parameters in the first step are associated with the first proximal drive. The parameters describe the weights/activity of AMPA and NMDA receptors from the supragranual layer (L2/3) and infragranual layer (L5). Further, pyramidal (pyr) and basket cells are differentiated.
For the prior ranges, we took over the prior ranges of the Jones tutorial[1(*Fig.*11)].

In dependence on these 3 parameter subsets, we chose early stopping times for the simulation of the time series. As for the first step, only parameters dependent on the first proximal drive were inferred, we stopped the simulation after 70 ms, arguing that the time range shortly before and after the P50 component should be mainly associated to these parameters, and most informative about these parameters.

For the second step, the parameters that are associated with the distal drive are now also taken into account. We therefore stopped the simulation only after 120 ms, such that the time interval around the N100 component could be assessed as well.

[1] `https://jonescompneurolab.github.io/hnn-tutorials/optimization/optimization`

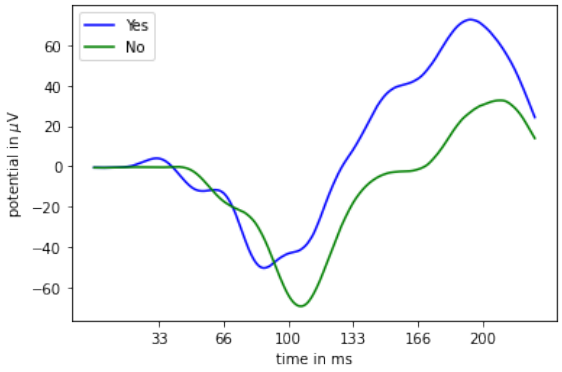

Figure 3.3: **Experimental paradigm** The two different conditions are plotted for the tactile stimulus experiment. The 'Yes' condition is defined in the way that 50% of the stimuli were detected, whereas the 'No' condition is defined such that no stimuli were detected.

For the last round, we simulated up to 200 ms, which incorporates the whole series on which we are interested on.

3.2.4 Experimental paradigm

In order to test how our pipeline can contribute to the study of an experimental or clinical paradigm, we used real, not simulated, ERP data, provided by Jones et al. [2007]. In this experimental paradigm, ERPs, that were source-localized to the somatosensory cortex, were measured for the two following conditions. In the first condition, a perceptual, tactile stimulus was detected in 50% of the cases ('Yes' condition), whereas in the second condition, the stimulus was never detected ('No' condition). The stimuli consisted of brief taps of 100 Hz sine waves that were delivered to the hand [Jones et al., 2007]. The difference of the waveform in the ERPs of the two conditions is visualized in Fig.3.3. In a tutorial by the Jones Neurolab[1], the ERP waveforms coming from different conditions were fitted by the COBYLA algorithm. The algorithm adapts parameter choices by repeating optimization rounds that reduce the overall RMSE between the fitted and the real data. With this approach, one arrives only at a single point solution for each parameter. Besides, it does not allow for uncertainty measures. SBI, instead, allows to investigate the possibly vast or narrow solution space of the parameters.
We tested how SBI, in particular NIPE, is capable of finding differences in microscale processes between these two conditions. The marginals of the posteriors of the two conditions are compared with contour plots of the parameter densities that show the 68% and the 95% percentiles, which means that the area, where 68% and 95% of the posterior samples lay, is visualized. We further show posterior predictive checks in order to see how the posteriors are able to recover the observations and

how confident the simulated predictions are. Beyond, we compared to the optimized values that were derived by the Jones Neurolab[1].

Results

4.1 Gaussian toy example

4.1.1 Less simulations needed with NIPE, compared to SNPE

KL divergence

The KL divergence is indicative for the quality of inference. The lower the KL divergence, the closer are inferred and analytic posterior.

Fig. 4.1 shows the comparison of the KL divergences between the inferred posteriors and the analytic posterior, with an increasing number of simulations per round.

Due to an instability issue when using mdn's with SNPE (issue #669[1]), we could not use the standard SNPE pipeline with mdn's where simulations of previous rounds are taken into account for calculating the new loss. In order to test the performance with the standard SNPE pipeline with the `proposal= proposal` argument being set, we therefore tested the toy example with a masked autoregressive flow [Papamakarios et al., 2017]. The variance for SNPE is higher, as well as the mean KL divergence (Fig. 4.1). This is the case over all numbers of simulations.

4.1.2 Over-dispersion of the posteriors - reallocating simulation budget partially resolves it

In order to check if the variances of the posteriors are under- or overestimated, we looked at the variances of the 15 single Gaussians, inferred with a maf as in the previous section 4.1.1. The variances are plotted in Fig. 4.2, with box plots visualizing the distribution of the variances for the 5 repeated posteriors. We compare between NIPE and SNPE, and further between NIPE with equal simulation budget (lightblue) for the 3 steps and NIPE with reallocated budget (purple). A simulation budget of $\frac{1}{30}$ for the first step and accordingly $\frac{59}{30}$ is used for the last step for the NIPE variant with reallocated budget that we call 'NIPE-BUDGET' here.

NIPE. For the Gaussians that were grouped into the first subsets, the variances were slightly lower than the analytic variance for NIPE, but well estimated for NIPE-BUDGET, with low dispersion between repetitions. The variances for the second subset are well estimated for both NIPE and NIPE-BUDGET. For the last subset, variances were higher than the true variance, and showed a wider dispersion between repetitions. The variances were a bit lower for the NIPE-BUDGET variant,

[1] https://github.com/mackelab/sbi

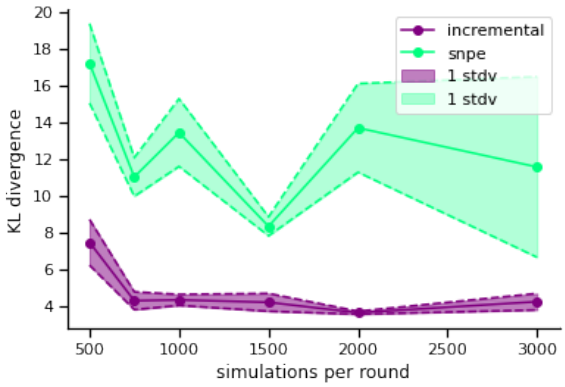

Figure 4.1: **KL divergence** for the SNPE approach (green) and the NIPE approach (purple). For each number of simulations per round, posteriors were calculated 5 times. The solid line describes the mean of the KL divergence between the 5 inferred posteriors and the analytic posterior. (a) inference with mdn. (b) inference with maf.

but showed high dispersion between repetitions.

SNPE. The variances for the SNPE approach were higher compared to the analytic variance, with a mean laying roughly between 2 and 5. Further, there was a higher dispersion between repetitions for the first two subsets, but a lower dispersion between repetitions for the last subset. The single posteriors were all over-dispersed. Further, sampling from the SNPE posterior took a longer time, with a sampling acceptance rate between roughly 0.124 percent for 1000 simulations per round, but even a lower acceptance rate for a higher number of simulations used per round. NIPE, in contrast, always had a sampling acceptance rate of 1.

In total, variances for NIPE were much lower compared to SNPE for the first two subsets, but showed a similar over-dispersion for the last subset. As the results seemed to improve a bit with NIPE-BUDGET, we always used reallocation of budgets for investigating event-related potentials.

4.2 Event-related potentials

Before we tested our pipeline on real data, we conditioned the posterior on a 'fake' observation that was simulated by the HNN simulator from a given parameter set. This has the advantage that we can later compare the posterior marginals to the 'ground truth'. In the following section, we start by comparing SNPE and NIPE on the basis of a simulated observation where we therefore have the ground truth. Later, we go on to test our approach on real data from the experimental setting that

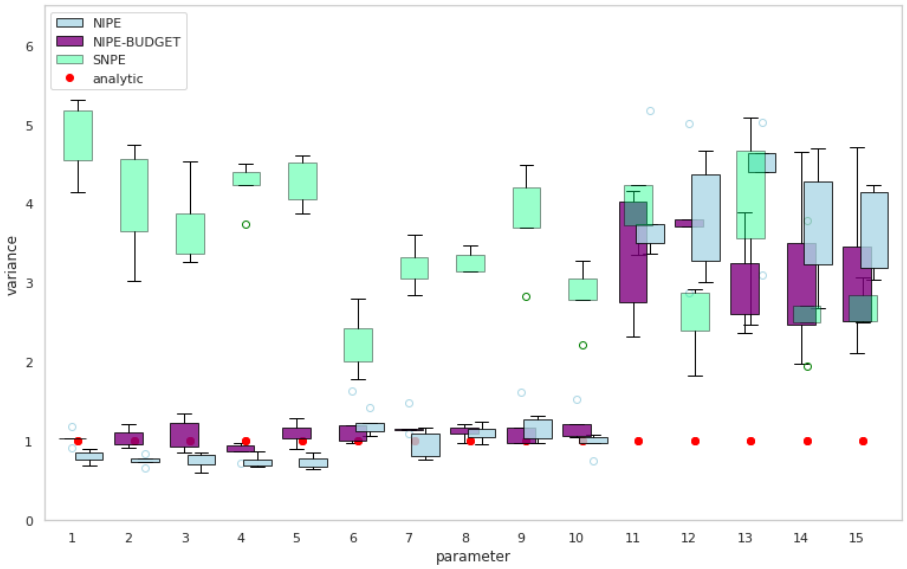

Figure 4.2: **Box plot for the variances of inferred and analytic posteriors** Visualizes the distribution (over 5 iterations) of inferred and analytic variances of the posteriors. The red points describe the ground truth (std of 1) for the analytic posterior. The light green box plots show the distribution for the SNPE approach for every single parameter. The purple box plots show the variance distribution for the NIPE approach with reallocated budget, such that the last step gets more simulations (NIPE-BUDGET). The light blue box plots show the variances for NIPE with equal budgets in each step.

we have described in the methods section.

4.2.1 Summary statistic values are more restricted for the posterior, compared to the proposal distribution

Checking the histograms of the summary statistics of the posterior and proposal, we can see that the histogram distribution of the summary statistics from the prior is much wider, as expected. The true summary statistics are plotted in red. They are calculated from a 'fake observation' that has been simulated from parameters that are seen as the 'ground truth' later on. For all of the summary statistics, we can observe that the values derived from the posterior (in blue) are close to the true value. Four examples are plotted in Fig. 4.3. For all histogram plots, see the appendix 6.3. The values that are derived from the proposal are plotted in orange. The distribution for these is much wider, as expected.

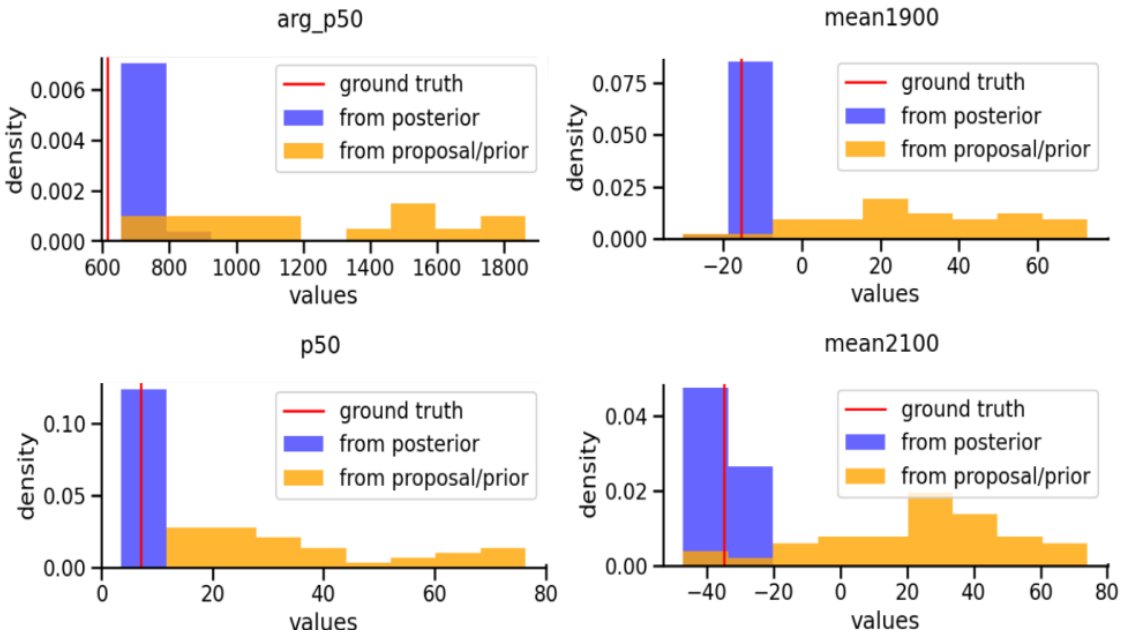

Figure 4.3: **Histograms for summary statistics** The histograms for the following four summary statistics are shown as an example: 'P50' describes the mean around the P50 component (10 ms before to 10 ms after the peak), 'arg_p50' describes the time point of P50, 'mean1900' and 'mean2100' describe the mean for two particular time ranges. The values that are calculated from the prior samples are visualized in orange, while the values that are calculated from the posterior samples are visualized in blue. The red line indicates the ground truth. In this case, the ground truth is calculated from a 'fake observation' where we know the ground truth parameters and simulate this 'fake observation' from these ground truth parameters.

4.2.2 Compensation mechanisms can be recovered by NIPE and SNPE

NIPE approach. The density plots and correlation matrices for the NIPE approach are shown in Fig.4.4. Total simulation and inference time was about 27 hours. Inference time was under a minute for the first step, around 13 minutes for the second step and around 24 minutes for the last step. The posterior was derived with a neural spline flow [Durkan et al., 2019].

The true parameters, plotted in red, mostly lay within the high density regions of the posteriors with some exceptions where the true parameters lay close to the high density regions, but not within. Some relationships between the parameters are indicated, visible via the shapes of the 2d marginals in the density plot and via correlation coefficients in the correlation matrices. There seem to be more correlations within the theta subsets belonging to the same drive, and less between the different subsets.

A positive correlation between NMDA pyramidal cells and the onset of the first proximal drive is indicated in both density plot and correlation matrix. These NMDA cells belonging to the first proximal drive are, in contrast, negatively correlated with the onset of the distal drive. The onset of the first proximal drive and the onset of the distal drive are also negatively correlated, which matches the other findings.

A negative conditional correlation between the AMPA weights of L5 pyramidal cells and NMDA L5 pyramidal cells, belonging to the second proximal drive, is indicated in all three plots of Fig.4.4. There is also a negative correlation between the NMDA weights of L2 and L5 cells belonging to the second proximal drive. In general, lots of parameters belonging to the second proximal drive are negatively correlated with each other.

Further, a negative correlation between NMDA weights of L5 pyramidal cells and the onset of the distal drive can be observed in all of the plots in Fig.4.4.

Interestingly, most of the weights belonging to the first proximal drive have multiple modes. There seem to be no conditional correlations between the weights of the first proximal drive, which could be related to the complicated shape of the marginals.

Looking at the posterior predictive checks (Fig. 4.5a), one can observe that the area, where 95% of the posterior simulations are laying, is partially not within the area where 95% of the prior simulations are laying. The 'true observation', plotted in red, is perfectly covered by the posterior area. The posterior area is more restricted for the early time range up to about 60ms, and then gets a bit broader.

SNPE approach. The density plots for the posteriors derived with SNPE are plotted in Fig.4.6. The true parameters are all laying in sampling regions of the posterior. One can observe some interesting shapes that indicate relationships between the parameters. The density plot indicates a negative correlation between the

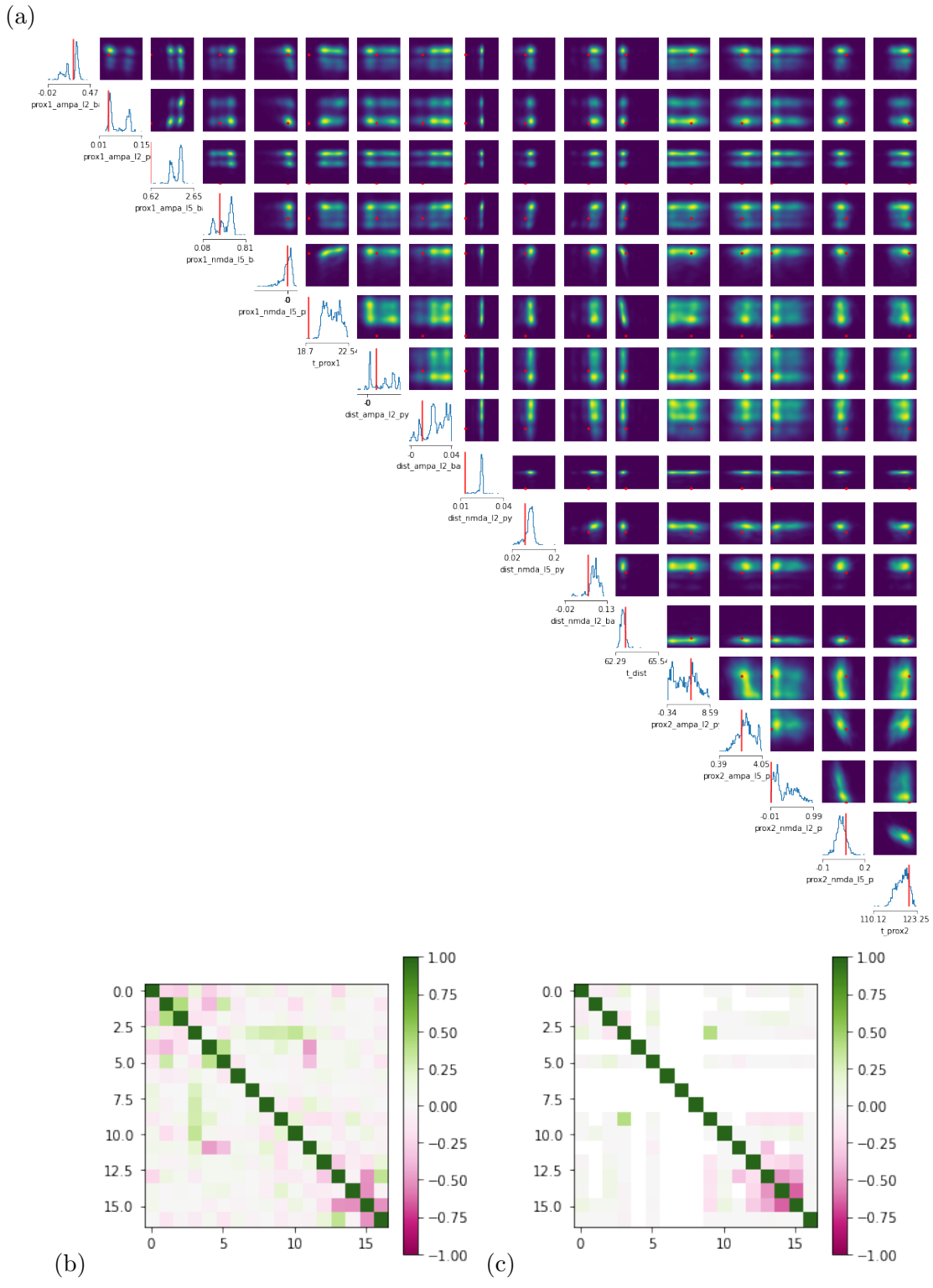

Figure 4.4: **Density plots** Inference on 17 parameters with the NIPE method, using a neural spline flow. (b) correlation matrix for the parameters. (c) conditional correlation matrix (all other parameters are kept fixed, except 2)

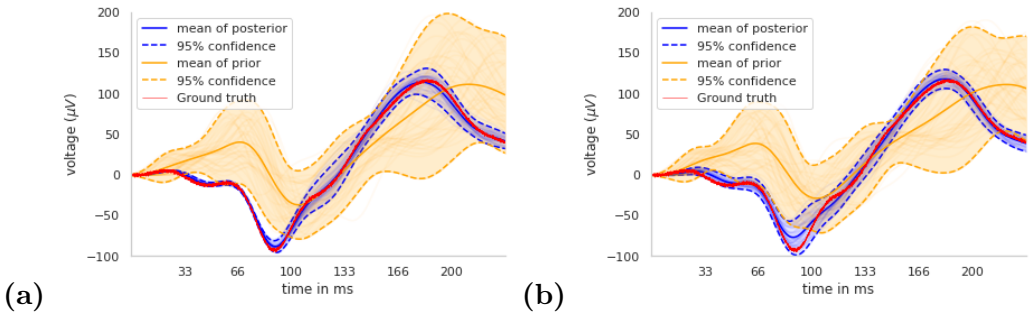

Figure 4.5: **PPC** Posterior predictive checks. The orange area describes the 95% confidence interval area for the simulations from the prior. The blue area the same for the simulations from the posterior. (a) PPC for the NIPE approach. (b) PPC for the SNPE approach.

first proximal drive and the onset of the distal drive. According to the density plot, a later onset of the proximal drive would therefore correlate with an earlier onset of the distal drive. Further, the high density area of the 2d marginal between the AMPA L5 pyramidal cells and AMPA L2 pyramidal cells has a banana-like shape. According to the shape, a higher AMPA pyramidal weight in L2 cells corresponds to a lower AMPA pyramidal weight in L5 cells. The (conditional) correlation matrices indicate strong positive correlations between some parameters belonging to the first proximal drive, and some negative correlations between parameters belonging to the second proximal drive. This is similar to the results for NIPE, except that there were no positive correlations between parameters of the first proximal drive indicated for NIPE. Besides, the 1d marginals for SNPE do not have multiple modes.

The posterior predictive checks (Fig. 4.5b), show, similarly to NIPE that the posterior simulations are sometimes laying outside the area that is spanned by the 95% confidence interval of the 100 prior simulations. Further, the 'true observation', plotted in red, is again perfectly covered by the posterior area. Concerning the negative trough of the signal, the confidence range for NIPE seems narrower, whereas the confidence range afterwards seems to be similarly broad for both PPCs.

Comparison between SNPE and NIPE. In order to evaluate differences in the inference process in one single plot, we plotted the 2d marginals of the NIPE approach in purple contours with 68% and 95% contour levels (Fig.4.8). The 2d marginals of the SNPE approach are plotted in green. This allows to investigate how the contours for the SNPE and the NIPE approach differ.

The parameters of the first two subsets are restricted to a higher degree for NIPE, in comparison to SNPE. While for SNPE, the true parameters always lay within the contour plots, for NIPE there are some exceptions. The parameters of the last subset are restricted a bit more by SNPE.

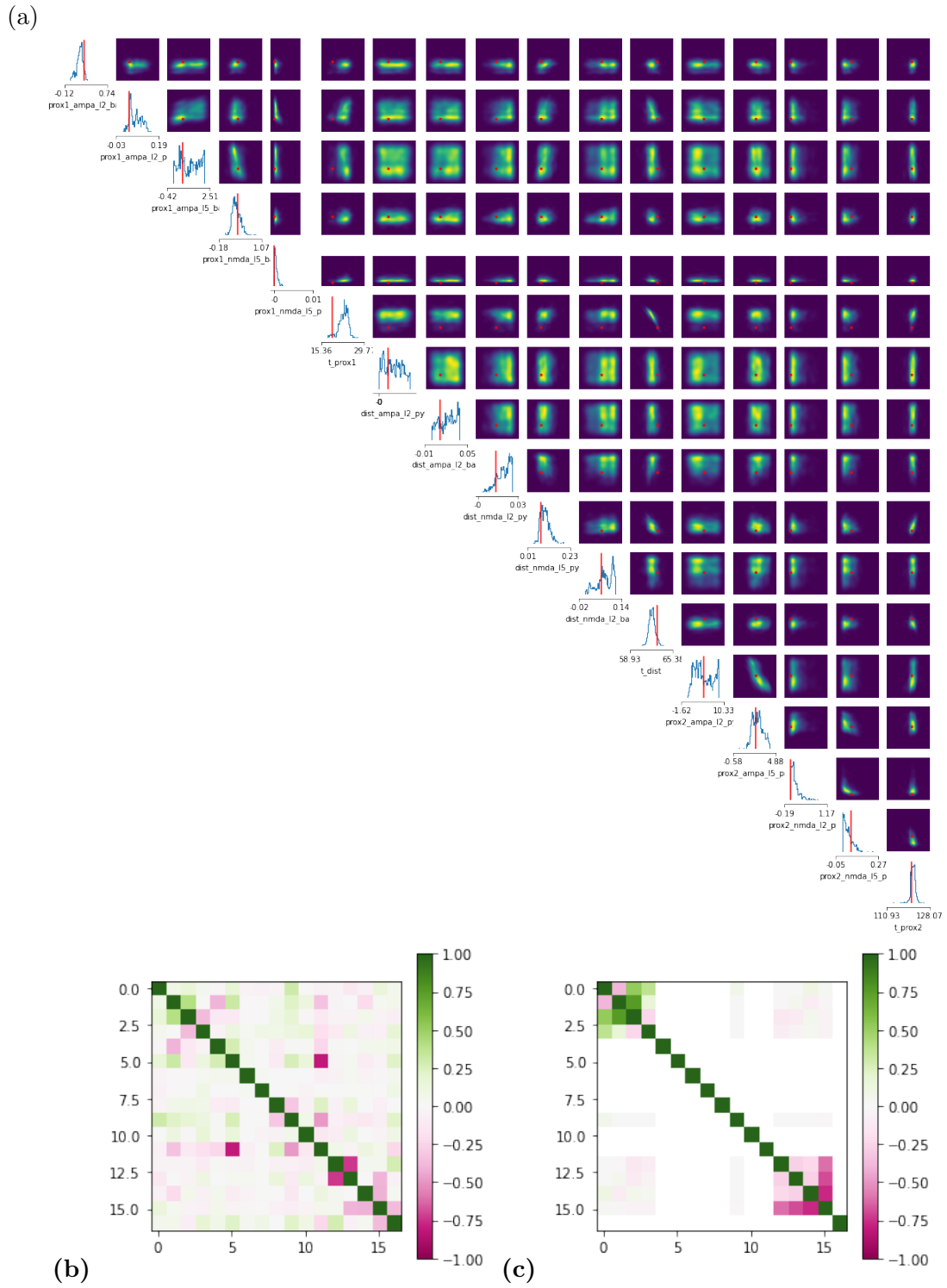

Figure 4.6: **Density plots** (a) Inference on 17 parameters with the multi-round approach (nsf used). The red lines/points indicate the true parameters. The 1d marginals are plotted on the diagonal, while the 2d marginals are plotted off-diagonal. (b) correlation matrix. (c) conditional correlation matrix (all other parameters are fixes, except 2)

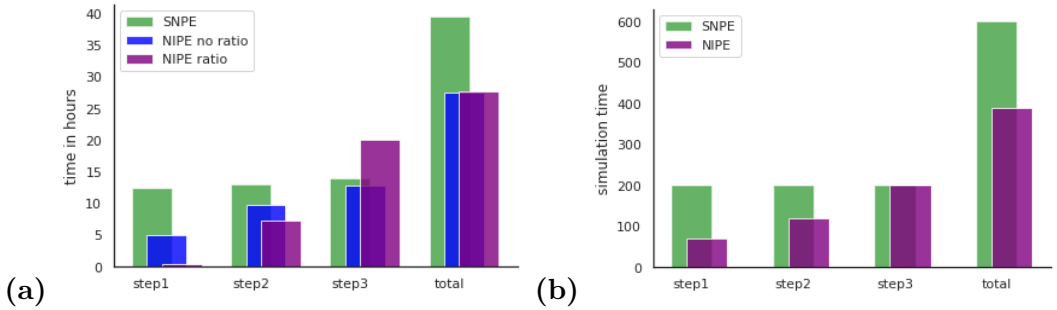

Figure 4.7: **Time for each step - comparison between SNPE and NIPE** (a) The plot compares how many hours each step/round took for SNPE (green), for 'NIPE ratio' (purple) and 'NIPE no ratio' (blue). The y-axis shows the time in hours (64 CPUs were used in parallel). (b) Plot visualizes how much time of the time series is plotted in each step/round. For NIPE, early stopping was used such that not the whole time series was simulated for the first two steps.

While drawing 1000 samples from the posterior took less than a second for the posterior inferred with NIPE, it took long time for SNPE with an acceptance rate of only around $2e^{-05}$. If prior ranges were expanded, the acceptance rate got even worse, such that it took more than 2 hours to only draw around 7 samples. We investigated whether this was due to a leakage issue such that most of the samples where not within the prior. Therefore, we plotted the densities for all samples, taking into account also the samples that were laying outside the prior support. The result of this can be found in the appendix (Fig.6.4). Plotting the densities without excluding the samples laying outside the prior support revealed that the leakage was mainly due to negative values for parameter weights.

While total time for SNPE was around 40 hours, it took around 27 hours for both NIPE approaches (Fig.4.7a).
Fig.4.7b visualizes how much milliseconds of the time series were simulated for each step/round. While for NIPE only 70 ms were simulated for the first step, for SNPE we needed to simulate the whole time series of 200 ms in each round. The differences between the approaches are similar for CPU time and simulation time.

4.2.3 Parameter differences found for the two conditions within the experimental paradigm

Posterior Predictive checks. Fig 4.9 shows the posterior predictive checks for the conditions of the experimental paradigm. Plotted in blue, one can see the simulations sampled from the threshold posterior and plotted in green, the simulations are sampled from the 'No' posterior. The thick lines indicate the true observations on which the posteriors were conditioned on. We added some Gaussian noise to these.

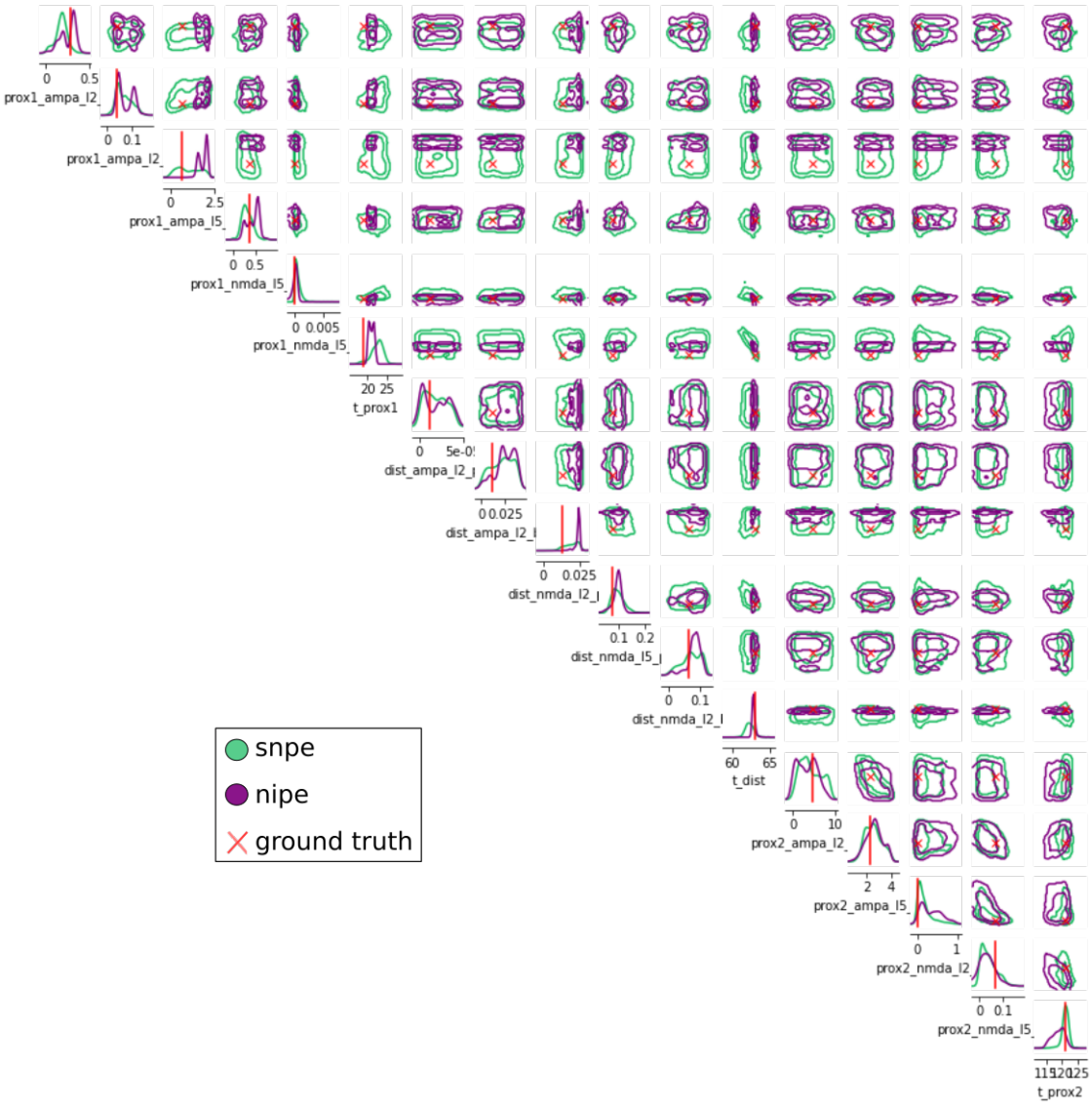

Figure 4.8: **Contour density plots - comparison between SNPE and NIPE** A nsf density estimator was used for both. 68% and 95% percentiles of the posterior densities are shown as contour lines. True parameter values are plotted in red. The true parameters 'exist' here because simulated the observation on which we conditioned with these parameters. The 1d marginals are on the diagonal, the 2d marginals off-diagonal. The results for SNPE are shown in green and the results for NIPE in purple.

The single simulations are plotted in transparent colors. For each condition, 100 simulations were calculated. The areas that are bounded with the fine, dashed lines show the 95% confidence intervals of the 100 simulations. In the threshold condition, there is a local maximum bump around 120 ms, whereas in the 'No' condition, the negative trough is prolonged and the P200 amplitude is lower compared to the threshold condition. The true observations are mostly covered well by the 95% confidence area of the 100 simulations drawn from each posterior. The posterior simulations are more restricted for the time range up to 100 ms and afterwards gets a bit broader.

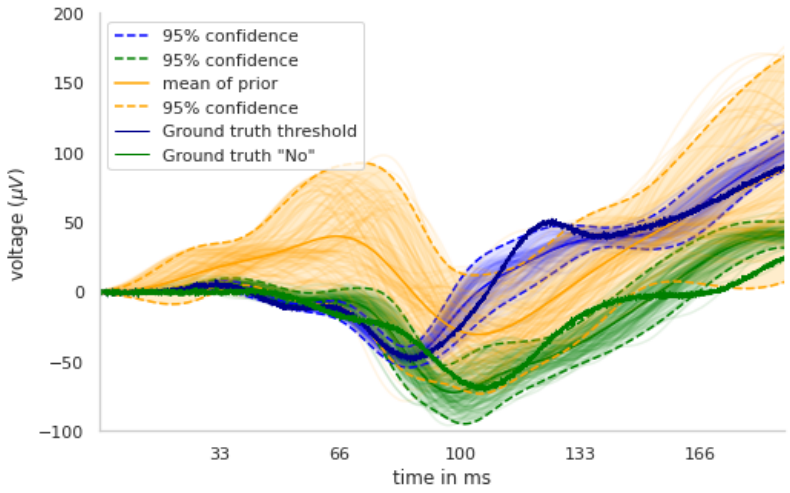

Figure 4.9: **PPC - derived with NIPE (nsf)** The simulations drawn from the prior are visualized in orange (95% confidence intervals), while simulations from the threshold samples are visualized in blue and simulations from the 'No' samples are visualized in green. The true observations are plotted in the same colors, but with strong, solid lines.

Comparison to the optimization process of the Jones Neurolab[1].

Fig.4.10 shows a contour plot of the 68% and 95% percentiles of the posterior densities for the two experimental conditions. In blue, one can see the densities for the threshold condition, whereas the densities for the 'No' condition are visualized in green. On the diagonal, one can see the 1d marginals, and the 2d marginals are plotted off-diagonal.

We compared our posteriors to the optimization values that were derived by the Jones Neurolab[1], plotting the 1d and 2d marginals of the posterior densities together with the point estimates derived by the Jones Neurolab. Fig.4.10 visualizes the densities for the threshold condition in blue, where one can see the point estimates also plotted in blue. For the 'No' condition the point estimates are plotted in green.

The NMDA weights of L5 basket cells belonging to the first proximal drive are estimated to be higher by NIPE, compared to the optimized values. There is also a clear difference between the 1d marginals of the two conditions. The distribution for

the 'No' condition is shifted to the left, such that values for the 'No' condition are estimated to be lower.

Whereas the optimized onset for the first proximal drive was estimated to be '26.61' for the threshold condition and '40.6' for the 'No' condition, the posteriors derived by NIPE lay close together, with the highest density a bit under 30.

The optimized values for the two conditions for the onset of the distal drive are laying within the high density region. The onset of the 'No' condition is estimated to be later, compared to the threshold condition.

The second proximal drive's onset is estimated to be later for the 'No' condition, both from NIPE and from the optimization. The posterior peaks are laying further away, even tough the optimized values are still within the outer contour borders. The other weights belonging to the second proximal drive are stronger restricted for the 'No' condition, compared to the threshold condition.

In the tutorial, it was mentioned that 'changing the standard deviation was necessary for matching the minimum and spread of the experimental dipole data at 75 ms'[1]. If we included the standard deviation as a parameter, however, our SBI pipeline did not work well anymore and was not able to predict observations from the posterior anymore. Posterior predictive checks also got worse when we widened the prior for the AMPA weights belonging to the distal drive, which we tried out because the optimized values for the 'No' condition were not included in our initially chosen prior range. We therefore stuck to our initial prior. In Fig.4.10, one cannot find a green cross for the two AMPA weights belonging to the distal drive, for this reason. We will discuss this later on.

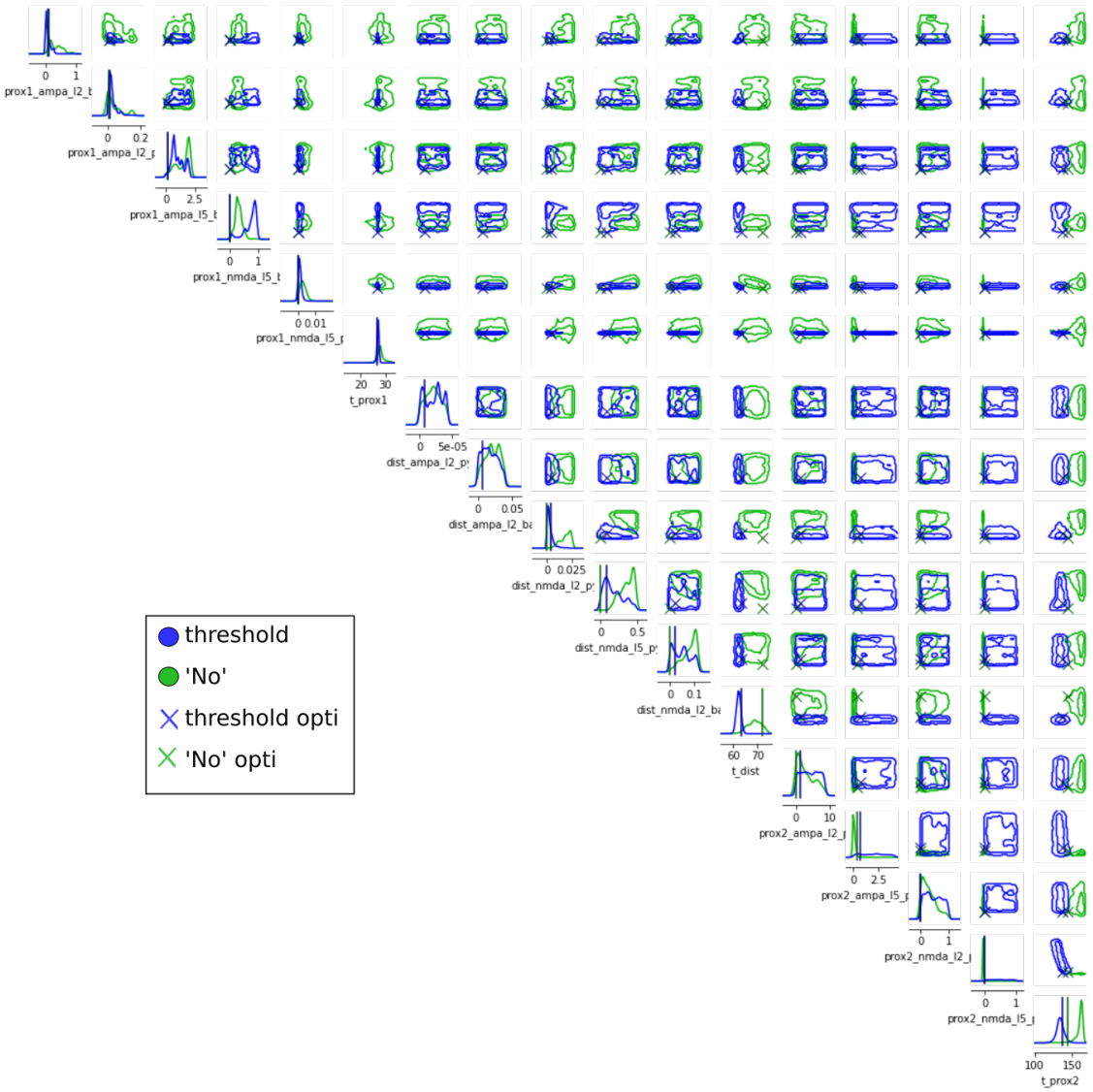

Figure 4.10: **Density plot - Comparison to Optimization tutorial** (a) shows the 1d (diagonal) and 2d marginals (off-diagonal) for the threshold condition, derived with NIPE. The dark green crosses and lines indicate the optimized values of the 'No' condition from the tutorial, while the dark blue crosses/lines indicate the same for the threshold condition.

Discussion

Toy example. We showed that NIPE performs well with respect to the KL divergence and the quality of the density plots. Further, the results indicate that NIPE is more time- and sampling-efficient, compared to SNPE. The calibration check, however, indicated that there is some bias regarding the estimated variances of the single parameter posteriors. Parameters that were inferred later in the inference pipeline showed a higher variance and were over-dispersed, while parameters that were inferred in the first step showed a bit of under-dispersion. If we used NIPE-BUDGET as a variant, where we reallocated the simulation budget towards the last subset, we could reduce the issue of under- and over-dispersion to a certain degree, but did not resolve it completely. How posteriors can be well calibrated should be addressed in future research. There might be a way to include simulations from former steps and to add loss terms and importance weights like it is done for SNPE [Greenberg et al., 2019]. Unfortunately, there is no straightforward way for NIPE because in each step we use a different number of parameters, such that it is not clear how to combine simulated theta vectors of different length.

Nonetheless, in a time-series setting where we assume time-dependencies of the parameter subsets, it is even realistic that parameter estimations for later subsets have a higher uncertainty because they depend on parameter estimations from earlier subsets. For the Gaussian toy example, there are no dependencies between parameters and therefore, the conclusion for time-series problems is limited.

Summary statistics. We were able to show that the designed summary statistics for ERPs restricted the prior space. It is difficult, however, to estimate the sufficiency of these summary statistics, without knowing the ground truth. If the ground truth is known, one could, e.g., calculate the KL divergence with and without a certain summary feature, in order to see how and if the feature restricts the posterior. One could then add features, dependent on how they contribute to a lower KL divergence. Adding more summary statistics, makes inference more complex and slower, such that we have to limit the number of summary statistics to a reasonable amount. Finding the most contributing summary features was beyond the scope of this work and could be an interesting target for future research.

Validation checks. Regarding ERPs, we showed similar density plots with respect to density shapes and covering of the ground truth of the parameters for SNPE and NIPE. As we do not have experimentally derived ground truth parameters, the interpretation of inference quality is limited. Nevertheless, we can evaluate if the true parameters are being recovered by the posterior densities. This was the case for both

approaches, even tough the true parameters were a bit better recovered by SNPE. Looking at the posterior predictive checks, the true observation was well recovered by simulating from the posterior samples. The first half of the PPC was restricted to a higher degree by NIPE. Altogether, we demonstrated that both approaches were able to restrict the parameter space and make good predictions.

Compensation mechanisms. Certain shapes in the 2d marginals of the posteriors indicated compensation mechanisms between parameters. Stronger AMPA weights of L5 pyramidal cells, belonging to the second proximal drive, can probably compensate for weaker NMDA weights of L5 pyramidal cells. Further, a negative correlation between the onset of the first proximal drive and the distal drive was indicated. If the onset of the proximal drive is later in time, the distal drive has to start earlier for deriving a similar ERP curve. This was weakly indicated by NIPE and strongly indicated by SNPE. The correlation seems plausible because the distal drive as a counterpart for the proximal drive, has to exhibit its drives earlier on in order to push against the forces of the proximal drive.

Another possible compensation mechanism was observed between the AMPA weights in L2 pyramidal cells and the AMPA weights of L5 pyramidal cells. A higher AMPA weight in L2 pyramidal cells corresponds to a lower AMPA weight in L5 pyramidal cells and the other way round. This seems logical because a strong signal that is propagated forward does not need strong weights in a higher layer to result in the same output. Increasing the AMPA weight in L5 pyramidal cells leads to an increased amplitude of the P200 component. Increasing the AMPA weight in L2 pyramidal cells also leads to an increased amplitude, such that one of them can be increased and the other one can be decreased for arriving at the same output as before. We verified this by sampling from the default parameters and by only changing one of these two, while holding the other parameters fixed.

We observed a positive, conditional correlation between AMPA weights of the L2 pyramidal cells belonging to the first proximal drive and the ones belonging to the distal drive, indicated by NIPE. If one of them increases, the other one has to be increased as well in order to derive the same result. This again can be explained by the fact that a stronger positive potential can be balanced by a stronger negative potential such that the AMPA weights belonging to the distal drive weaken the positive potential. Overall, the compensation mechanisms that were detected, seem biologically plausible and makes it possible to investigate the whole solution space.

Comparison of SNPE and NIPE. We showed a comparison of the 1d and 2d marginals of the NIPE and the SNPE approach by plotting the transparent contour lines on top of each other, with different colors. The marginals of the NIPE posterior were restricted to a larger extend. As argued in the methods section 3.1, NIPE might need less simulations in order to restrict the parameter space to a high degree.

However, especially the parameter posteriors belonging to the first subset, might possibly be under-dispersed, such that they do not show the full solution space anymore. The 1d marginals belonging to the first subset had a more complex shape, in comparison to SNPE. If this indicates under-dispersion or a better restriction of the parameter space, remains an open question. As both SNPE and NIPE are sequential methods, we can not apply simulation-based calibration, unfortunately. How and if simulation-based calibration could be made applicable for sequential methods, needs to be discovered. When using NIPE, one has to be aware of possible biases of the posterior variance. If simulation budget is not an issue, SNPE might be the better choice.

Sampling from SNPE posteriors, on the other hand, was often very slow. We showed that this was due to a big sampling leakage such that most of the samples were outside the prior support. If efficient sampling is needed, NIPE has a huge advantage over SNPE as it does not suffer from this sampling issue. Especially when wider prior ranges were chosen, SNPE was not applicable anymore. This is a well-known problem for SNPE by [Greenberg et al., 2019], when using maf or nsf estimators. It could help to project the theta samples to an unbounded space for inference and afterwards project samples drawn from the posterior to constrained space again, as explained by Gonçalves et al. [2020]. In order to make multi-round SNPE more sampling-efficient, this could be interesting to investigate in further research.

HNN model and Stochasticity. As already mentioned, it is not an easy question which parameters to include in the inference pipeline. Not only whether or not to include the standard deviation of the drive onsets is a crucial question. It is also disputable whether it is better to use a small subset of parameters that we want to investigate while leaving other parameters fixed, or instead use all available parameters of a model. The later makes inference a lot harder, the former possibly prevents the exploration of the whole parameter space. If one is only interested in compensation mechanisms of a small subset, it might not be necessary to include all parameters of the model, but one could carefully select the parameters of interest.

Ensuring stochasticity of a simulator, is a crucial part for SBI. This is because inference should be robust to small perturbations of the processes. There is evidence that there are many adaptive and compensatory mechanisms going on in the brain that can cope with perturbations [Marder and Taylor, 2011]. Inducing stochasticity into a simulator should therefore make inference robust, but at the same time make it more biologically realistic.

If adding observation noise and choosing `sigma`> 0 is inducing an appropriate degree of stochasticity, has to be tested in real experimental studies. To our knowledge, there does not exist experimental (animal) ERP studies that have tested mechanistic models like the HNN and proofed that the model predicts parameter differences

well. The HNN model has been compared to animal studies in the case of beta bursts in the motor cortex, which is one of the best studied movement signals [Bonaiuto et al., 2021], but this has not be done for ERPs, yet. We made the experience that, when including `sigma` as parameter, inference was either not possible, or it took a long time and even then did not restrict the parameter space well. Hypothetically, including `sigma` as a parameter makes the model too flexible. Fine-tuning as in the Jones tutorial[1] might be possible, but does not seem appropriate for our case.

We used a smoothing window of 30 ms for the time series, which might be too wide to capture interesting local bumps, but might make inference more robust. In the Jones optimization tutorial[1(Fig.11)], it is mentioned that compared to the threshold condition, the supra-threshold condition has very sharp features and therefore needs a shorter time window. We did not vary time windows for different conditions. This might have arguably improved our results. Nevertheless, this again is some form of fine-tuning that should optimally not be needed.

Experimental paradigm. For the experimental paradigm, we observed mainly differences between the conditions for the onset of the distal drive and the onset of the second proximal drive. Some conductance weights were more restricted by one of the conditions, which implies that the weights might only play a role for a certain condition, while it does not get restricted a lot for another one.

The post predictive checks indicated that the simulations from the posterior samples matched the observations well and the simulations from the two conditions were well differentiated. As simulations are quite costly, the PPCs were only based on 100 simulations. Nonetheless, the posterior predictive checks look quite well and showed that NIPE was able to restrict the parameter space and to differentiate the conditions via parameter posteriors.

Further limitations. Our approach is not amortized, such that a trained neural density network cannot be conditioned on new observations. It could be very practical to make the approach amortizable, such that inference is less costly and time-consuming. Unfortunately, this is not straightforward for sequential approaches where the parameter space gets restricted by an observation after each round and is therefore conditioned on a particular observation.

Our new approach is based on specific assumptions, e.g. that there exists a time order for parameter subsets such that earlier subsets of parameters have an effect on later subsets, but not the other way. This assumption makes it possible to separate the inference process in smaller parts where we have to deal with a much smaller dimensionality. The approach is limited to problems where we can make similar assumptions. Nevertheless, we think that this assumption is met by many problems that are investigated in neuroscience research as time series like EEGs or spike

trains always have a time order, naturally. There exist other approaches to increase simulation efficiency (See e.g. [Prangle, 2016]) where simulations of time series are interrupted if they do not seem promising. In our approach, we can use early stopping for the first subsets and therefore increase simulation efficiency in a different way.

Further, it would be interesting if one could not only condition on a single observation, but instead on many observations, belonging to the same condition. This would be very helpful for investigating clinical conditions where there might be a certain variance between e.g. patients with schizophrenia, but where one would like to investigate what the commonalities within a condition are.

5.1 Conclusion

Overall, our approach offers an opportunity to investigate high-dimensional problems that usually need a huge amount of simulations with other methods often used for model-inversion of the brain like MCMC [Hashemi et al.]. Further research should examine how NIPE can be better calibrated such that neither later inferred posteriors are over-dispersed, nor earlier inferred posteriors are under-dispersed.

We showed several compensation mechanisms that suggest that it is sometimes possible to increase one parameter and decrease another one, without changing the final output. Including more parameters increases the chance for degeneracy to occur. We also made the observation that including `sigma` as a parameter introduces too much flexibility, such that inference is not possible anymore. Whereas point estimates provide only one single solution, we visualized that there is a broader solution space. While for the threshold condition certain parameters got well restricted, for the 'No' condition the parameters that got well restricted differed a lot. We therefore offer an approach that is able to show parameter uncertainties and differences between conditions.

Applying NIPE on an experimental paradigm, interesting differences of the two conditions on the micro-level were shown. Revealing micro-scale differences between conditions can contribute to precision medicine applications by providing patient-specific suggestions that make clinical interventions more effective. Our method has a huge potential in filling the gap between micro- and macro-dimensions through studying the interaction of micro-processes and the emergence of complex macro-signals. They can be used to test hypotheses about biophysical processes and to discover medical treatment alternatives.

Their is evidence that the feed-forward and feedback information flow through the layers, and also the laminar structure, is shared across different sensory regions [Kohl

et al., 2021, Atencio and Schreiner, 2010]. Applying SBI in combination with the HNN model, can therefore be applied for many different sensory tasks, e.g. tactile, auditory or visual tasks, and opens the door for many different scientific questions and interesting clinical applications.

In future experiments, it would be interesting to find experimental designs that test how well the biophysical model is specified. Here, we suppose that the assumptions made by the model are well justified and can reveal e.g. clinical or cognitive conditions. How these assumptions can be validated, however, remains an open research question.

Acknowledgements

I want to thank Prof. Jakob Macke, who gave me the possibility to combine my interests in neuroscience and machine learning methods and who gave ideas, interesting paper proposals and feedback during the process of my thesis. I also want to thank Prof. Martin Butz, who agreed to be my second supervisor.

My special thanks goes to Cornelius Schröder, who always gave new ideas, valuable feedback and time to ask him questions and for feedback. We were exploring lots of ideas and different directions and he gave great guidance on where to focus and where to start.

In general, I want to thank the Macke group for their hospitality, valuable feedback and interesting discussions during lunch time and coffee breaks.

Bibliography

Craig A Atencio and Christoph E Schreiner. Columnar connectivity and laminar processing in cat primary auditory cortex. *PLoS One*, 5(3):e9521, 2010.

James J Bonaiuto, Simon Little, Samuel A Neymotin, Stephanie R Jones, Gareth R Barnes, and Sven Bestmann. Laminar dynamics of high amplitude beta bursts in human motor cortex. *NeuroImage*, 242:118479, 2021.

Kristin S Cadenhead, Gregory A Light, Mark A Geyer, and David L Braff. Sensory gating deficits assessed by the p50 event-related potential in subjects with schizotypal personality disorder. *American Journal of Psychiatry*, 157(1):55–59, 2000.

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

Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Joel Dyer, Patrick W Cannon, and Sebastian M Schmon. Deep signature statistics for likelihood-free time-series models. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

Gaute T Einevoll, Alain Destexhe, Markus Diesmann, Sonja Grün, Viktor Jirsa, Marc de Kamps, Michele Migliore, Torbjørn V Ness, Hans E Plesser, and Felix Schürmann. The scientific case for brain simulations. *Neuron*, 102(4):735–744, 2019.

Pedro Goncalves, Jan-Matthis Lueckmann, Giacomo Bassetto, Kaan Oecal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. In *Bonn Brain 3 Conference 2018, Bonn, Germany*, 2018.

Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife*, 9:e56261, 2020.

David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2404–2414. PMLR, 2019. URL `https://proceedings.mlr.press/v97/greenberg19a.html`. ISSN: 2640-3498.

Arthur Hamilton and Georg Northoff. Abnormal erps and brain dynamics mediate basic self disturbance in schizophrenia: A review of eeg and meg studies. *Frontiers in psychiatry*, 12:438, 2021.

Faith M Hanlon, Gregory A Miller, Robert J Thoma, Jessica Irwin, Aaron Jones, Sandra N Moses, Mingxiong Huang, Michael P Weisend, Kim M Paulson, J Christopher Edgar, et al. Distinct m50 and m100 auditory gating deficits in schizophrenia. *Psychophysiology*, 42(4):417–427, 2005.

M. Hashemi, A. N. Vattikonda, V. Sip, M. Guye, F. Bartolomei, M. M. Woodman, and V. K. Jirsa. The bayesian virtual epileptic patient: A probabilistic framework designed to infer the spatial map of epileptogenicity in a personalized large-scale brain model of epilepsy spread. 217:116839. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.116839. URL `https://www.sciencedirect.com/science/article/pii/S1053811920303268`.

Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500, 1952.

Maëliss Jallais, Pedro LC Rodrigues, Alexandre Gramfort, and Demian Wassermann. Cytoarchitecture measurements in brain gray matter using likelihood-free inference. In *International Conference on Information Processing in Medical Imaging*, pages 191–202. Springer, 2021.

Stephanie R Jones, Dominique L Pritchett, Steven M Stufflebeam, Matti Hämäläinen, and Christopher I Moore. Neural correlates of tactile detection: a combined magnetoencephalography and biophysically based computational modeling study. *Journal of Neuroscience*, 27(40):10751–10764, 2007.

Antti Kangasrääsiö, Kumaripaba Athukorala, Andrew Howes, Jukka Corander, Samuel Kaski, and Antti Oulasvirta. Inferring cognitive models from data using approximate bayesian computation. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 1295–1306, 2017.

Carmen Kohl, Tiina Parviainen, and Stephanie R. Jones. Neural mechanisms underlying human auditory evoked responses revealed by human neocortical neurosolver. 2021. ISSN 1573-6792. doi: 10.1007/s10548-021-00838-0. URL https://doi.org/10.1007/s10548-021-00838-0.

Carmen Kohl, Tiina Parviainen, and Stephanie R Jones. Neural mechanisms underlying human auditory evoked responses revealed by human neocortical neurosolver. *Brain topography*, 35(1):19–35, 2022.

Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30, 2017.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pages 343–351. PMLR, 2021.

Eve Marder and Adam L Taylor. Multiple models to capture the variability in biological neurons and networks. *Nature neuroscience*, 14(2):133–138, 2011.

Henry Markram, Karlheinz Meier, Thomas Lippert, Sten Grillner, Richard Frackowiak, Stanislas Dehaene, Alois Knoll, Haim Sompolinsky, Kris Verstreken, Javier DeFelipe, et al. Introducing the human brain project. *Procedia Computer Science*, 7:39–42, 2011.

Samuel A Neymotin, Dylan S Daniels, Blake Caldwell, Robert A McDougal, Nicholas T Carnevale, Mainak Jas, Christopher I Moore, Michael L Hines, Matti Hämäläinen, and Stephanie R Jones. Human neocortical neurosolver (hnn), a new software tool for interpreting the cellular and network origin of human meg/eeg data. *Elife*, 9:e51214, 2020.

George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. In *Advances in neural information processing systems*, pages 1028–1036, 2016.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

Dennis Prangle. Lazy abc. *Statistics and Computing*, 26(1):171–185, 2016.

Stefan T Radev, Andreas Voss, Eva Marie Wieschen, and Paul-Christian Bürkner. Amortized bayesian inference for models of cognition. *arXiv preprint arXiv:2005.03899*, 2020.

Marvin Schmitt, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T Radev. Bayesflow can reliably detect model misspecification and posterior errors in amortized bayesian inference. *arXiv preprint arXiv:2112.08866*, 2021.

Cornelius Schröder, Ben James, Leon Lagnado, and Philipp Berens. Approximate bayesian inference for a mechanistic model of vesicle release at a ribbon synapse. *Advances in Neural Information Processing Systems*, 32, 2019.

Felix Schürmann, Sean Hill, and Henry Markram. The blue brain project: building the neocortical column. *BMC Neuroscience*, 8(2):1–1, 2007.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

A Roger D Thornton, Matthew Harmer, and Brigitte A Lavoie. Selective attention increases the temporal precision of the auditory n100 event-related potential. *Hearing Research*, 230(1-2):73–79, 2007.

Timothy O West, Luc Berthouze, Simon F Farmer, Hayriye Cagnan, and Vladimir Litvak. Inference of brain networks with approximate bayesian computation–assessing face validity with an example application in parkinsonism. *NeuroImage*, 236:118020, 2021.

David A Ziegler, Dominique L Pritchett, Paymon Hosseini-Varnamkhasti, Suzanne Corkin, Matti Hämäläinen, Christopher I Moore, and Stephanie R Jones. Transformations in oscillatory activity and evoked responses in primary somatosensory cortex in middle age: a combined computational neural modeling and meg study. *Neuroimage*, 52(3):897–912, 2010.

Appendix

6.1 Toy example - Piecewise linear function

6.1.1 Method procedure

To get a first impression how the NIPE approach performs, we tested it with a simple toy example - a piecewise linear function where one can vary offset and slope parameters. The model was defined with 3 pieces in the following way:

$$y[x < cp_1] = b + a_1 \cdot x + \epsilon$$
$$y[cp_2 \leq x \leq cp_2] = y_{cp_1} + a_2 \cdot x + \epsilon$$
$$y[x \geq cp_1] = y_{cp_2} + a_3 \cdot x + \epsilon$$

, where a_1: first slope, b: offset, a_2: second slope, a_3: third slope, cp_1 and cp_2: changing points, ϵ: noise. y_{cp_1} and y_{cp_2} are the function values at the changing points.

We varied 4 parameters - the offset and the 3 slopes. The ground truth was arbitrarily set at the beginning, such that the posteriors were conditioned on the observation under the ground truth after each step (for NIPE)/round (for SNPE).

For the NIPE approach, we defined 3 inference steps and set the number of simulations to 300 for each of these steps. In the first step, we simulated only the first piece up to cp_1, and then inferred the posteriors for b and a_1. The posteriors were combined with the prior for a_3, such that the next thetas for b and a_1 were sampled from the posterior and the thetas for a_2 were sampled from the prior. In the second step, we then simulated up to cp_2 and inferred the posteriors for b, a_1 and a_2. For the third step, a_3 was inferred as well.

For the SNPE approach we used a multi-round approach (3 rounds) and used 300 simulations for each round.

We then used posterior predictive checks (further explained in Section 3.1.4) and density plots of the inferred posteriors to compare the two approaches.

6.1.2 NIPE restricts piecewise linear model to a higher extend

Testing the NIPE approach on the piecewise linear toy example shows that it performed really well. With the same amount of simulations (900 in total), the NIPE

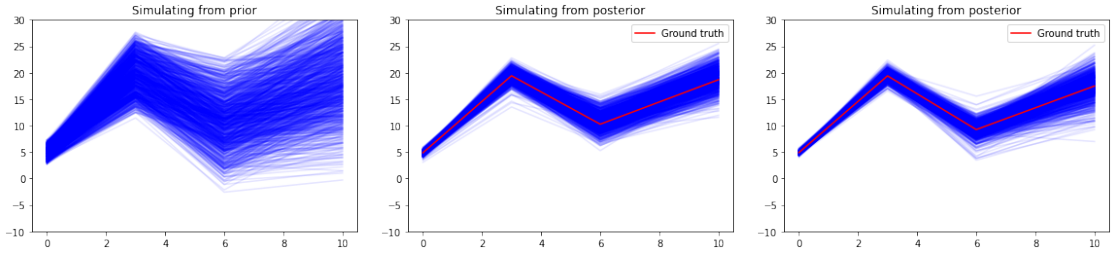

Figure 6.1: **Post-predictive checks: Piecewise linear model** Inferring the parameters (slopes and changing points) with the SNPE approach (middle) and the NIPE approach (right). The ground truth is plotted in red. On the left side, one can observe simulations from samples drawn from the prior. The plot visualizes how the posteriors restrict the simulation space.

approach makes equally good predictions, compared to SNPE, according to the posterior predictive check (Fig.6.1). It restricts the parameter space to a higher extend, which is observable from the density plot (Fig. 6.2).
The whole simulation and inference process took about 30 seconds for both approaches.

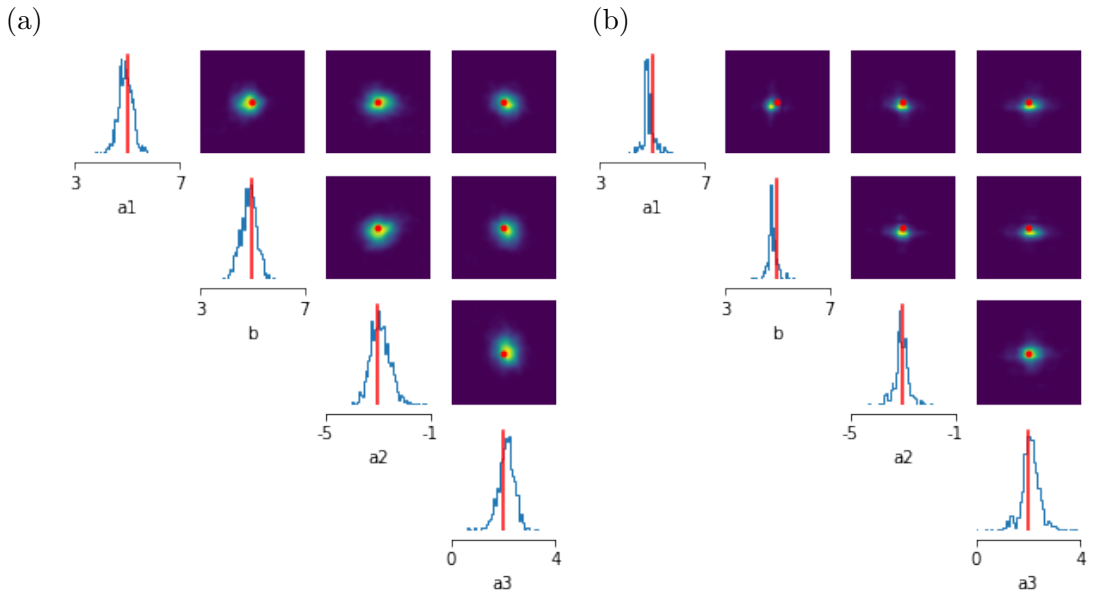

Figure 6.2: **Piecewise linear model** Density plots for (a) the SNPE approach (b) the NIPE approach. The 1d marginals of the posteriors are plotted on the diagonal, while the 2d marginals are represented off-diagonal. The red lines/points are indicating the ground truth of the parameters. a_1 to a_4 are the slope parameters of a piecewise linear function while b describes the offset.

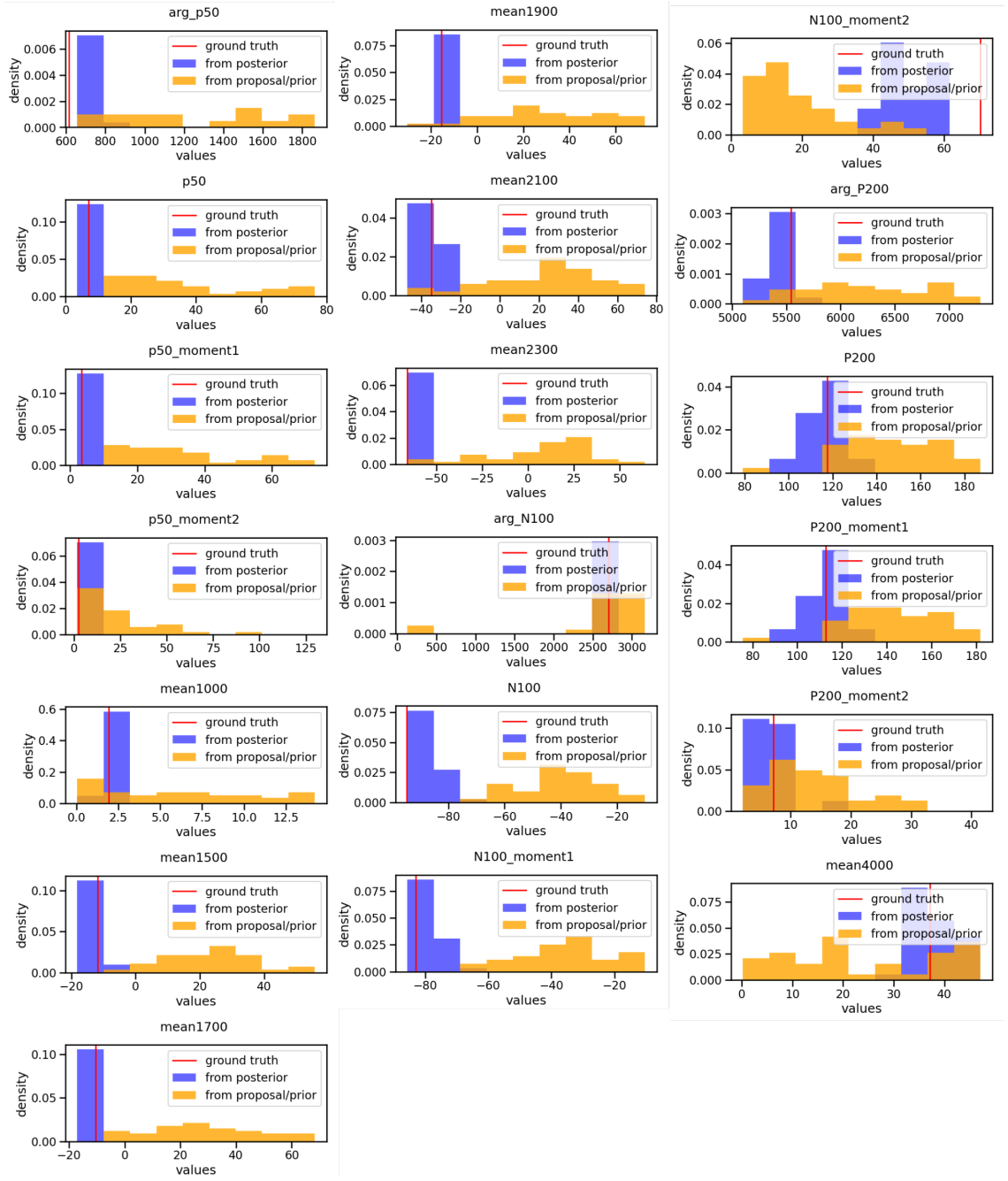

Figure 6.3: **Histograms of the summary statistics.** All hand-crafted summary statistics are shown. The posterior always includes the ground truth value (shown in red) and restricts the space of values. The posterior was conditioned on a simulated observation here.

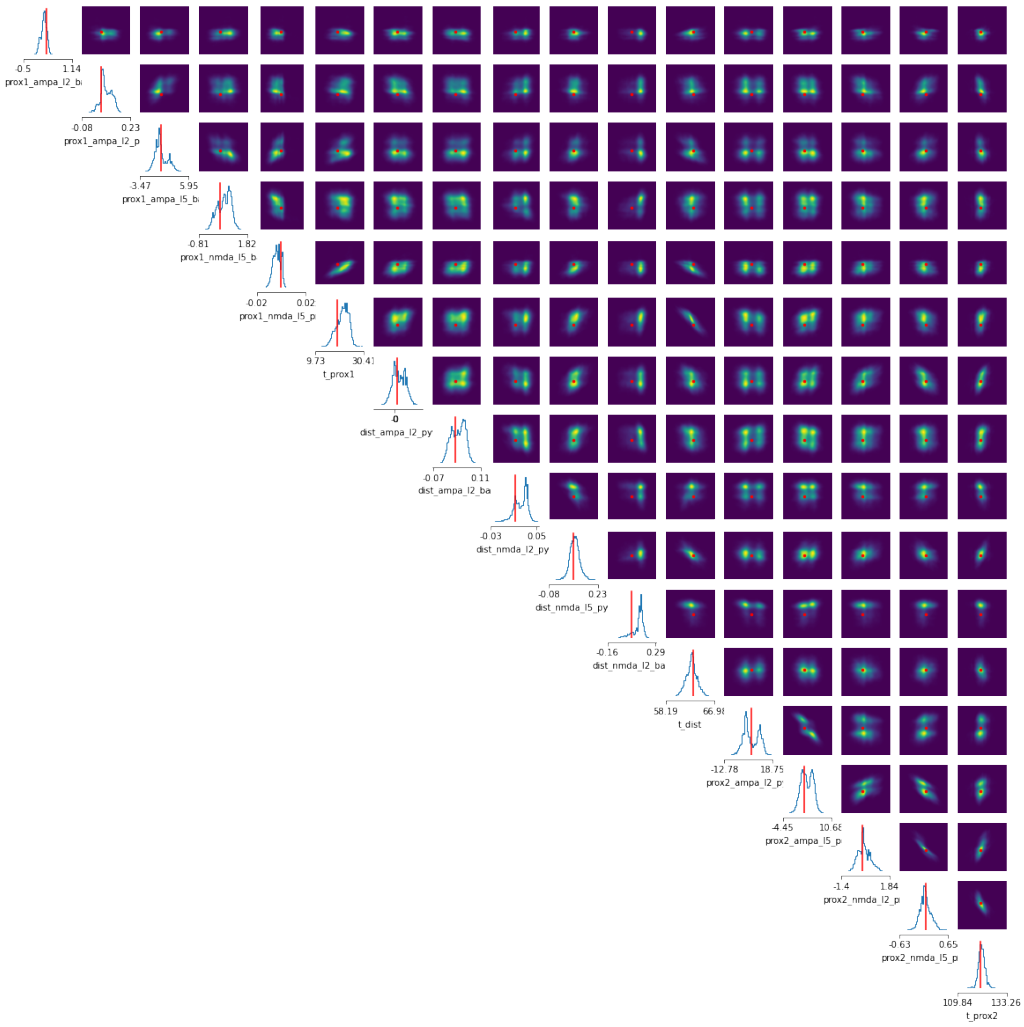

Figure 6.4: **Density plot for SNPE - leakage taken into account** Sampling from the posterior inferred by SNPE had an acceptance rate of only around $7e^{-5}$. Accepting all samples, also taking into account the ones not within the prior, revealed these 1d and 2d marginals. One can see that zero weights are seen as plausible by the inferred distribution. It is not clear how the HNN simulator deals with these negative values and if it turns it into positive ones.