1 Summary - Feature Learning for Bayesian Inference

The goal of this project is to use interpretable *Machine Learning* (ML) to find low-dimensional features in high-dimensional noisy data generated by (i) stochastic models or (ii) real systems. In the first case, we are interested in the features imprinted on simulated data by the *parameters* of the stochastic model. In the latter, the interesting features depend on the particular system. In hydrology, one of the domains considered in this project, they are fingerprints of *catchment properties* in observed time-series of river-runoff. In both cases, the problem is to *disentangle* the effect of high-dimensional disturbances (*noise* realizations in the first case or the *rain* falling on the catchment in the latter) from the effects of relevant characteristics (model parameters in the first case, catchment properties in the latter). This problem is reminiscent of the problem of finding *collective* quantities characterizing states of interacting particle systems in Statistical Mechanics. *Variational Autoencoders* (VAE) have been proven to be capable of learning such quantities in the form of *order parameters* or *collective variables*, which could then be used to identify *phase transitions* or to enhance *Molecular Dynamics* (MD) simulations.

We expect parameter-features of stochastic model-outputs to be of great value for *Bayesian inference*. They can be used to sample from the Bayesian posterior, in situations where the likelihood function (LHD) is too expensive to evaluate, and one thus has to resort to comparing - ideally sufficient - summary statistics of simulated data with corresponding statistics of observed data as in Approximate Bayesian Computation.

Instead of comparing summary statistics, we may attempt to sample the product space of model parameters and model states. As this results in an extremely high-dimensional inference problem, sophisticated sampling schemes such as Hamiltonian Monte Carlo (HMC) have to be employed. But HMC, which is essentially the same as MD, suffers from slowly-mixing *collective* features of high-dimensional model-states associated with changing model parameters. Hence, one of the objectives of this project is to bring the method of *biased sampling* with collective variables developed for MD to fruition for sampling-based Bayesian inference.

A considerable part of the project is further devoted to the development of ML-methods for summary-statistics learning. Towards this end, we will also employ a robust algorithm for finding the *Intrinsic Dimension* (ID) of data, which was developed by the PI. Not only might this method help to identify the ideal number of summary statistics, but it might also be useful for identifying different *phases* of the model based on the distribution of the learned features. The data ID can guide the design of the architecture of the ML tools leading to higher interpretability. We will finally apply feature-learning to real data in different domains, either *directly* to observed data (hydrology), or *indirectly* to the outputs of the stochastic model that is used to model the data (infectious disease epidemiology). We stress that dimentionality reduction, feature learning and interpretable ML are versatile tools that can be leveraged far beyond Bayesian learning and that the application domains of the developed methodology range beyond our two case studies.

2 Research plan

2.1 Current state of research in the field

Bayesian inference is a way of combining our theoretical understanding of a particular system (in the form of a *model* encoded in a LHD function indexed by a set of parameters) with knowledge gained from *data* and possible prior knowledge on the parameter values (encoded in a prior distribution on the parameter space). Bayesian *uncertainty quantification* requires a faithful representation of the dominant sources of uncertainty in the model. Traditional models in many domains lump all the uncertainty into a simple *error model* that is added to the output of a deterministic model. However, such error models typically do not adequately capture the intricate correlation structure exhibited by real data. Furthermore, if the error is not built-in to the model where it actually arises (e.g. on the input rather than the output), the model is unable to faithfully predict anything but the variables it was calibrated on. Building uncertainty into the model where it actually arises has the advantage of resulting in an interpretable knowledge domain driven model but typically leads to *LHD functions* in the form of very high dimensional integrals, which renders Bayesian inference computationally expensive, to the point that it becomes prohibitive, for most interesting real-world applications. Different numerical approaches have emerged to deal with this problem. *Approximate Bayesian Computation* (ABC, see the following paper where the PI and/or project partners are co-authors: [Albert et al., 2014, Dutta et al., 2018c, Dutta et al., 2021, Dutta et al., 2017, Chen et al., 2020a, Raynal et al., 2021, Ebert et al., 2021, Varghese et al., 2020], and references therein) avoids the evaluation of a prohibitively expensive LHD function, and instead compares relevant *summary statistics* (in a reduced dimensional space) of model simulations and data. However, it is still a largely unsolved problem how to find summary statistics that retain features of the data that are relevant for constraining model parameters (sufficiency) and, at the same time, enable fast convergence to the true posterior (efficiency). Using *Machine Learning* (ML) algorithms, it has been suggested ([Papamakarios and Murray, 2016]) to bypass summary statistics altogether and instead directly learn the posterior density based on large numbers of model simulations. However, this approach requires a parametrization of the posterior density (e.g. in the form of a Gaussian mixture), which might lead to a bias if the posterior has a complicated non-linear / non-Gaussian shape.

Another sampling-based approach is available when the LHD is intractable due to the fact that it is obtained after integrating out nuisance or latent variables. In these cases, we can trade the intractable likelihood function off against a very high-dimensional inference problem with a simple likelihood function, i.e. infer all the latent stochastic model state variables alongside the model parameters. Different variants of this approach have been shown to be able to cope with high-dimensionality, such as *Particle Filters* [Šukys et al., 2018], *piecewise updating* [Tomassini et al., 2009] or *Hamiltonian Monte Carlo* (HMC) sam-

pling [Albert et al., 2016]. However, all of them are plagued with the existence of *slowly mixing modes*, in particular those that are related to changing model parameters, which can hamper the exploration of parameter space. This problem is related to the notorious *free energy barriers* encountered in *Molecular Dynamics* (MD) simulations in physics [Invernizzi and Parrinello, 2020]. When the slow modes are known, they can be used to speed up the sampling (e.g. [Laio and Parrinello, 2002]). Finding these modes, which are very much related to summary statistics, means *disentangling* the effects of noise and model parameters on the output of a stochastic model. A similar problem arises in the analysis of *observed data*, in many scientific domains. An important problem in hydrology, e.g., is how to use catchment properties for runoff predictions [Dal Molin et al., 2019]. Therefore, it is useful to disentangle the runoff-features that stem from the catchment from those that stem from known forcing variables, such as rain and potential evapotranspiration.

In light of the above mentioned difficulties, the scientific problem we want to address in this project is the **identification of relevant low-dimensional features in high-dimensional data**. What *relevant features* are, depends on the data and the purpose. For the purpose of Bayesian inference, where the data stems from stochastic simulations, the features should be largely insensitive to the noise realizations used for the simulations, but informative about model parameters. For our hydrological example, where the data stems from runoff-measurements, the features should be as insensitive as possible to peculiarities of known forcing variables such as the rain.

The interest in this old problem has recently been revived due to advancements in ML. Because the stochastic forward simulation models that are used in various application fields are often cheap, vast amounts of synthetic data can be generated to train ML models. Likewise, massive deployment of cheap sensors and social media have led to an ever growing amount of observed data, in many domains. Notable progress has been achieved, e.g., in the fields of *Statistical Physics* and *Bayesian Statistics*, albeit with rather different foci. In physics, ML has been successfully used for learning *order parameters* and *collective variables* of interacting particle models. These features, in turn, have been used to identify *phase transitions* [Wetzel, 2017, Carrasquilla and Melko, 2017] and for biasing MD simulations towards *sampling across free energy barriers* [Bonati et al., 2020]. In statistics, on the other hand, the focus has been on finding a minimal number of *sufficient summary statistics*, which can then be used for example in ABC (e.g. [Fearnhead and Prangle, 2012, Jiang et al., 2017, Chen et al., 2020b]).

Different approaches have been taken to learn such low-dimensional features, which can be grouped into the following categories:

Regression-based approaches. When the source of the relevant features is explicitly available (e.g. the parameter values associated with stochastic model realizations) one can simply attempt to learn those values from the data [Fearnhead and Prangle, 2012, Jiang et al., 2017]. This is typically done for each parameter separately. For a given model output, one may then hope to learn values that are close to the corresponding

mean marginal posteriors which, as demonstrated in [Fearnhead and Prangle, 2012], are the best summary statistics when minimizing the quadratic loss. The problem with this approach is that it does not necessarily maximize the parameter-related information content in the summary statistics. Furthermore, for generic stochastic models, the minimal number of sufficient statistics is typically larger than the number of parameters [Efron et al., 1975]. It is also not applicable if the source of the relevant features is not explicitly available (as in the case of catchment properties).

Reconstruction-based approaches. *Variational Autoencoders* have been used in various contexts for learning low-dimensional features of noisy data. E.g. a hot topic in physics is to use them for learning order parameters and collective variables of stochastic interacting-particle models ([Wetzel, 2017]). VAE can also be used when the source of the relevant features is not explicitly available. However, they make rather strong assumptions about the noise, although, quite often, explicit information about the noise is available. With the modified AE we present below we suggest to make use of this information.

Information-theoretic approaches such as [Cvitkovic and Koliander, 2019, Chen et al., 2020b] use entropic measures for extracting features from data that is maximally informative about given relevant source variables. Below, we suggest using such measures for regularizing our modified AE.

Indirect inference can be considered as a non inherently Bayesian precursor of ABC [Gourieroux et al., 1993, Gourieroux et al., 1996] where an "auxiliary model serves as a window through which to view both the actual, observed data and the simulated data generated by the [economic] model: it selects aspects of the data upon which to focus the analysis" (Indirect inference in The New Palgrave Dictionary of Economics, Second Edition). The estimated parameters of the auxiliary model can be considered as summary statistics in the same spirit as "emulators" are used in ABC [Cui et al., 2018]. We plan to study the theoretical results on optimality and convergence available in the indirect inference and in the related moment matching literature (see the seminal paper by [McFadden, 1989]) to try to prove similar results in our context.

Current state of research with reference to the two applications. In hydrology, learned catchment-features of runoff data could be mapped to known catchment properties (geology, vegetation, soil etc.) and thus be useful for making runoff-predictions also in *ungauged* catchments, which constitutes a major problem in hydrology today.

Disease-spreading on networks is a timely topic and prime field of application for stochastic models. More and more network data are becoming available, at different levels of scale (from people-people interactions to country-country interactions), so network models may be more accurate at making predictions on the spread of an epidemic and on the effect of various interventions aimed at relaxing the network structure (people-people and country-country connections) relaxing the strong within compartment homogeneity assumptions that are typical of traditional epidemiological SEIR type models [Linka et al., 2020, Chinazzi et al., 2020]. Perhaps more importantly, some interventions such a non-pharmaceutical ones cannot be naturally captured

using standard models. For example, the effect of contact tracing is hard to incorporate in traditional models; the same is true for people forming pods, i.e., groups of people with whom they freely associate.

Parameter-induced features of the outputs of such stochastic models can not only be used for Bayesian inference as described above, but they can also be helpful for understanding the modes of operation of the model (akin to *phases* in physics). The methodology we develop and apply in this project is very general. Therefore, we expect it to be useful in many other application domains as well.

2.2 Current state of own research

This proposal builds upon the foundations that have been laid in the SDSC-funded BISTOM project by A. Mira (co-PI in BISTOM), C. Albert (PI in BISTOM) and F. Perez-Cruz (SDSC collaborator in BISTOM), S. Ulzega (Senior researcher) and F. Ozdemir (Post-doc). Within BISTOM we have elucidated the deep connection between thermodynamic state variables, order parameters and near-minimally sufficient summary statistics. We found that non-linear stochastic models with many degrees of freedom often exhibit distributions of summary statistics that are multi-modal, for fixed model parameters. This is due to entropic effects and reminiscent of the problem of multi-modal free energy landscapes encountered in MD simulations as well as multimodality in e.g. statistical mixture models. When doing Bayesian inference with such models, it is important to not only consider the local shape of the mode the observation belongs to, but also its weight relative to other modes. For this reason, it is important to not only include parameter regressors in the set of summary statistics, but also additional order parameters required to resolve the different modes. We have also developed several new ML-methods tailored towards learning such summary statistics (see Sect. 2.3).

From a methodological point of view, A. Mira (AM) will bring to the project the more statistical and computational expertise while F. Perez-Cruz (FPC) will bring the more ML related competences. The project partners have both methodological and application related competences, specifically C. Albert is expert in MD, ABC, HMC and hydrology; JP Onnela (JP) has competences in networks science, ABC and infectious disease epidemiology and A. Laio (AL) is expert in MD, MCMC and intrinsic dimension methodology.

Antonietta Mira directs the Data Science Lab at USI, Lugano. She is a Bayesian statistician expert in modeling and Monte Carlo simulation algorithms. Specifically she has published papers on the two classes of algorithms that the current proposal aims at improving, namely HMC [Papamarkou et al., 2014, Mira and Haario, 2011] and ABC [Dutta et al., 2018a, Dutta et al., 2018c, Dutta et al., 2017, Dutta et al., 2021, Dutta et al., 2018b, Varghese et al., 2020, Raynal et al., 2021, Ebert et al., 2021, Warne et al., 2020, Chen et al., 2020a]. Besides her methodological expertiese both in modeling and computation, she will provide access to knowledge (together with JP Onnela) and data in the domain of infectious disease epidemiology thanks to COVID-19 related grants where she is co-PI, ∼10 million Euro (for 3 years) H2020 (PERISCOPE). AM is also acquiring knowledge on infectious disease epidemiology thanks to the fact that she is also co-PI in a smaller

grant ($\sim$ 280k for a year ending in July 2021) funded by Lombardy Region, Italy (TSUNAMI). With reference to epidemiological models and algorithms for COVID-19, she has a published paper in BMC Public Health [Warne et al., 2020], a minor revision paper in Statistics in Medicine [Bartolucci et al., 2021] and a submitted manuscript with JP Onnela, external collaborator (with title "Policies for easing pandemic travel restrictions").

She has a successful (in terms of publication) history of collaboration with JP Onneala and A. Laio (project partners) on topics strongly related to the proposal. Specifically with Laio she has worked on methods to identify the data Intrinsic Dimension [Allegra et al., 2020] and with Onnela on models and ABC algorithm for network data (see references above on ABC). Furthermore, thanks to the BISTOM project she has been productively discussing issues at the interface of ML, physics and statistics with the co-PI and C. Albert (project partner). These discussions have lead to fruitful cross-fertilization among these disciplines something A. Mira has been leveraging upon in her entire research career as the very diverse journal venues of her published papers demonstrate.

Fernando Perez-Cruz is a Professor at the Computer Science Department at ETHZ and the Chief Data Scientist at Swiss Data Science Center. He has been a member of the technical staff at Bell Labs and a Machine Learning Research Scientist at Amazon. He has been a visiting professor at Princeton University under a Marie Curie Fellowship and an associate professor at University Carlos III in Madrid. He has also held positions at the Gatsby Unit (London), Max Planck Institute for Biological Cybernetics (Tuebingen), BioWulf Technologies (New York).

Fernando mostly works on the application of machine learning to science. He has been interested in the application of machine learning to signal processing and communications and has focus on developing Bayesian non-parametric algorithms. Currently, he is interested mostly in unsupervised learning and how new models based on neural networks can served as universal simulators. Since becoming the Chief Data Science at SDSC, which is a joint venture between EPFL and ETH Zurich to do machine learning research within ETH board institutions, Fernando has worked in over 30 projects ranging from climate to precision medicine. On this context, he has been collaborating with Dr. Albert and Prof. Mira on the SDSC funded BISTOM project.

2.3 Detailed research plan

Our main objectives are threefold:

1) Employing existing and designing new ML models for learning low-dimensional features of high-dimensional data. On the one hand, this means using standard methods employed in different disciplines (such as VAEs), on the other hand, it means further developing our own *augmented Regressors and AEs* (see. Sect. 2.3.1), which are specifically designed for learning minimal numbers of sufficient statistics

and to overcome some of the limitations of the current methods used in statistics. We plan to leverage on the recently introduced method of finding the *Intrinsic Dimension* of data [Allegra et al., 2020] to learn the optimal number of features, for the purpose of Bayesian inference.

As an alternative to AEs we also want to explore the potential of real non-volume preserving (NVP) transformations [Dinh et al., 2014], which allow us to obtain complex densities in a deterministic manner, by manipulating a simple distribution combining scaling and translation operations using neural networks. These NVPs models are proving to be very successful in learning distributions and compared to Generative Adversarial Networks and VAEs, are able to provide exact likelihood evaluations.

2) Assessing the usefulness of the learned features for Bayesian inference by means of cutting-edge sampling tools. In test cases, where we have access to the true posterior, ABC will be used to assess the parameter-related information content of the learned features. Our objective is to assess the potential of ML for learning near minimal sufficient statistics, for a wide range of stochastic model types: we expect the developed methodology to be sufficiently flexible and versatile to be able to tackle a variety of model classes. Furthermore, we will assess whether the learned features can be used for enhancing the mixing of HMC samplers, in the same way collective variables are used for enhancing MD simulations. To our knowledge, this exciting possibility that is currently very actively pursued in physics, has not yet been explored for the purpose of Bayesian inference for stochastic models. But see, e.g. [Levy et al., 2017], for alternative ways of using ML for enhancing the mixing of HMC.

3) Applying these methods to observed data in the fields of hydrology and infectious disease epidemiology. Hydrological modeling is a notoriously difficult problem, as many of the relevant processes happen underground and are thus inaccessible to direct observation. There are two ways of approaching this problem, for both of which the methodology studied in this project will be helpful. One way is to use stochastic terms to describe unknown forcings or processes. While stochastic modeling is becoming more and more popular in hydrology, the resulting inference problem has not yet been sufficiently addressed. Another way is to employ ML. While ML has a long history in hydrology, only very recently has it proven to be able to compete with traditional hydrological modeling [Kratzert et al., 2019]. An important contemporary problem is runoff prediction for *ungauged* catchments (i.e. catchments without observed runoff-data) [Dal Molin et al., 2019]. Therefore, next to rainfall predictions, relevant catchment properties need to be known. Our objective is to use ML for learning relevant low-dimensional runoff-features from *gauged* catchments. These features can then be associated with known catchment properties, and thus assist hydrologists in finding those properties that are relevant for runoff-prediction.

Disease-spreading is a very timely topic, and a prime example for using stochastic models. However, the complex nature of such models using realistic spreading-networks and epidemiological models have largely hampered Bayesian inference, so far. Few exceptions are [Linka et al., 2020, Chinazzi et al., 2020]. Upon

successful implementation, our approach will advance this field in two ways. It will lead to (i) more accurate Bayesian inference and hence more reliable probabilistic predictions of a complex and dynamic stochastic system and (ii) a better understanding of the spread of the COVID-19 epidemics as well as the effects of various policy measure aimed at mitigating it. To this aim we plan to provide a real time decision support system to better manage this type of emergencies in the form of a dashboard similar to the one set up by epiMOX (Politecnico of Milano, Italy: https://www.epimox.polimi.it), where the underlying homogeneous compartment model is presented in [Parolini et al., 2021]. We note that this model does not consider network type data which is, on the other end, the aim of the model we plan to study.

We describe the scientific approaches of these parts separately.

2.3.1 Feature learning

Consider a large number of high-dimensional data-points (e.g. time-series), which are carrying features from two different sources: A high-dimensional source we consider uninteresting (and generally call *noise*) and a low-dimensional source of *relevant* features we want to extract. For the purpose of Bayesian inference, the data-points are stochastic model realizations, and the relevant features are those imprinted on them by model parameters. For observed datasets, "noise" could be anything we are not interested in. For hydrological runoff time-series, for instance, we could consider the observed rainfall as being the noise, whereas the relevant features are those imprinted on the runoff by catchment properties. Our approach builds upon and attempts to overcome limitations of existing approaches (see Sect. 2.1).

The main idea behind our **modified AE**, which we started developing in the BISTOM project, is to give the *decoder* access to the noise, $\boldsymbol{\epsilon}$, that has entered any given data-point, $\mathbf{x}$, of the training data. In this manner, the *encoder*, $\mathbf{s}(\mathbf{x})$, is incentivized to only learn those features (summary statistics) of the data that stem from the relevant source of information, $\boldsymbol{\theta}$ (e.g. the model parameters). A possible interpretable architecture of the AE which is driven by the above mentioned requests and features and thus, ultimately by our understanding of the problem, is shown in Figure 1. For this AE to be trained on the output of a stochastic model, the latter needs to be given as a deterministic function, $M(\boldsymbol{\theta}, \boldsymbol{\epsilon})$, of model parameters and bare ($\boldsymbol{\theta}$-independent) noise realizations. This is no restriction, as *any* stochastic model becomes deterministic when conditioned on the noise realization needed for its forward simulation. However, given a stochastic model with parameters $\boldsymbol{\theta}$, the choice of M is not unique, but depends on the definition of $\boldsymbol{\epsilon}$ (e.g. the particular type of distribution the ϵ_i are drawn from). Suitable choices can make it easier for the decoder to learn M.

There are several developmental fronts we want to push further in this project:

The choice of the reconstruction loss function. For certain synthetic training data sets, the decoder should, theoretically, achieve a *zero loss*, as it has access to the exact noise realizations that went into the data generation. For zero loss, the decoder, $\hat{\mathbf{x}}(\mathbf{s}, \boldsymbol{\epsilon})$, has to learn the deterministic model equations as well

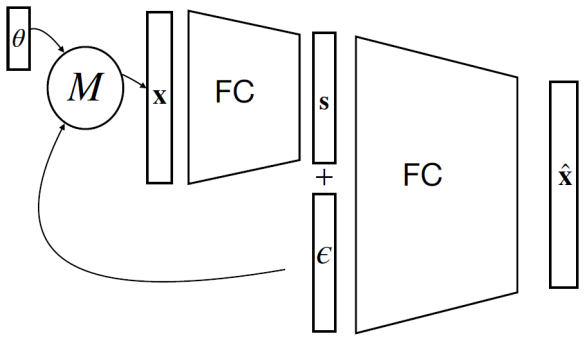

Figure 1: Basic architecture of the modified AE: The stochastic model needs to be available in the form of a deterministic function, M, of a parameter vector $\boldsymbol{\theta}$ and a (bare) noise realization $\boldsymbol{\epsilon}$. The encoder is trained on model realizations $\mathbf{x}$, and produces summary statistics $\mathbf{s}(\mathbf{x})$. The decoder is trained to reproduce the model realizations $\hat{\mathbf{x}}(\mathbf{s}, \boldsymbol{\epsilon})$, based on $\mathbf{s}$ and the same noise realizations that have been used by M. Instead of using fully connected (FC) networks, prior knowledge about M can be used to simplify the network structure.

as the inversion of the function $\phi_{\boldsymbol{\epsilon}} : \boldsymbol{\theta} \mapsto \mathbf{s}$, both for given noise realizations, i.e. $\hat{\mathbf{x}}(\mathbf{s}, \boldsymbol{\epsilon}) = M(\phi_{\boldsymbol{\epsilon}}^{-1}(\mathbf{s}), \boldsymbol{\epsilon})$. However, this inversion does not always exist globally, and even if it does, the decoder might not find it, for all possible noise realizations. In situations where a point-by-point reconstruction is too difficult (e.g. for chaotic systems) we might want to relax to a loss function that is averaging over different noise realizations. This brings us closer to a VAE, but without its strong assumptions on the nature of the noise.

The choice of the regularization. In many situations, the decoder could reach (near) zero reconstruction loss, while the encoded summary statistics $\mathbf{s}$ are far from being sufficient. Think of the trivial example $x_i = \theta + \epsilon_i$. Zero reconstruction loss could be achieved with the encoder $s(\mathbf{x}) = x_1$, which is far from being sufficient, and the decoder $\hat{x}_i(s, \boldsymbol{\epsilon}) = s - \epsilon_1 + \epsilon_i$. Therefore, it is important to incentivize the encoder to encode as much information as possible about $\boldsymbol{\theta}$ in $\mathbf{s}$. In BISTOM, we tried to achieve this through regularizing the loss function with the MSE between the first components of $\mathbf{s}$ and the parameters, i.e., we essentially combined a regression approach with a reconstruction approach. The reconstructor (decoder) facilitates the work of the regressor and creates an incentive to encode also auxiliary information that is not contained in the regressions but needed for Bayesian inference (see the case study below).

For the purpose of Bayesian inference, there is no need for some of the summary statistics to be parameter regressors. In fact, it might suffice for them to be *concentrated*, when the parameters are kept fixed. Therefore, in future developments, we plan to test also information-theoretic concentration measures. For instance, one might use the entropy of the $\mathbf{s}$-distribution, for fixed parameter values, relative to the $\mathbf{s}$-distribution with the parameters varying over the prior. This will require training data that, for each parameter sample, contains many noise realizations. This regularization will be useful in situations where we do not have access to the source of the relevant features (e.g. catchment properties).

The choice of the network architecture. Knowing the nature of the data-generating model (or real system), we can choose appropriate architectures. For instance, for hidden Markov models, we might want

to choose a *convolutional* architecture. However, it is important to bear in mind that the decoder not only has to learn the model, but implicitly also needs to invert the function $\phi_{\boldsymbol{\epsilon}}$, which might require additional *dense* layers in the decoder.

The choice of the dimension of the AE-bottleneck. For a given stochastic model, it is generally not easy to find out what the minimal number of sufficient summary statistics is. Often it even grows unbounded with the number of data points, while *most* of the information can be compressed into a number of statistics that is of the order of the number of parameters. A useful tool could be the recently developed Bayesian methodology to estimate the *Intrinsic Dimension* (ID) [Allegra et al., 2020]. The ID is a fundamental geometric property of a dataset being the minimal number of coordinates that are necessary to describe its points without significant information loss. The Bayesian ID we will use is robust in that it can be applied even if the manifold containing the data is curved, topologically complex, and sampled non-uniformly. Furthermore, our ID methodology requires weak assumptions and allows us to discover an unknown number of manifolds and (probabilistically) assign points to the most a posteriori likely manifold. We conjecture that the ID will drive the discovery of the optimal number of summary statistics. Therefore, we suggest comparing the ID of the training data (parameters and associated model outputs) in the product space with the ID of the output components alone. We expect the difference to inform us about the number of couplings between parameters and outputs and thus about the optimal number of sufficient statistics. The problem could be simplified replacing the outputs with a relatively large number of $\mathbf{s}$-variables, from a trained AE.

Once the AE is trained, we can assess the prior distribution in feature-space, resulting from model simulations over the prior parameter range. It can inform us about different types of model behavior, akin to *phases* in physics. Depending on the employed feature-learning, they might manifest themselves in the form of clusters or through density changes in feature space. For this assessment, the ID will be a powerful tool.

A proof of concept has been given in the BISTOM project by means of the nonlinear stochastic iterative map model

$$x_{n+1} = \alpha f(x_n) + \sigma\epsilon_n\,, \quad \epsilon_n \sim \mathcal{N}(0,1) \quad i.i.d.\,, \tag{1}$$

where $f(x)$ is a nonlinear function admitting two stable solutions, e.g. $f(x) = x^2 e^{-x}$. The deterministic map has a stable fixed point at $x = 0$. For sufficiently large α, a second fixed point emerges, which, upon further increasing α undergoes a series of period doubling bifurcations eventually leading to chaos. Depending on the initial condition x_0, the deterministic map will go to either one of the two attractors. This situation is reminiscent to first-order phase transitions in Statistical Mechanics. For certain regions in parameter space, different phases can co-exist (i.e. in our example the model can go to either one of the two attractors). Hence, for the purpose of Bayesian inference, the summary statistics need to be augmented by an order parameter that is able to tell the different phases apart.

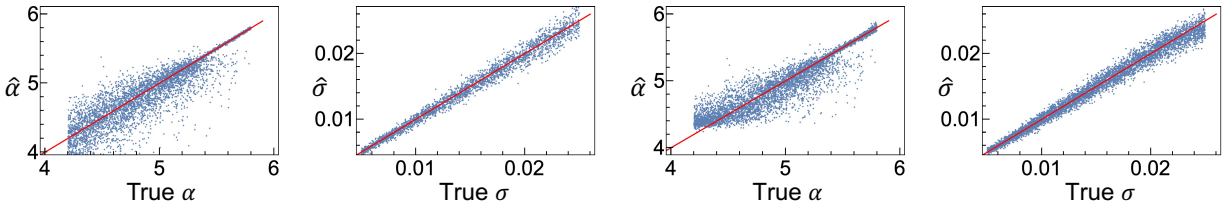

Figure 2: Maximum likelihood estimates for the parameters of model (1) (left two panels), and parameter regression from the encoder (right two panels). The α-regression shows a clear accuracy-difference, for the two phases of the model.

As a member of the exponential family, the minimal number of sufficient statistics for this model is bounded. However, the parametrization not being the natural one, we need *three* summary statistics to reach sufficiency, albeit the model only has two parameters. A convenient parametrization, for sufficient statistics, is given by eqs.

$$\hat{\alpha}(x) = \frac{\sum_{n=1}^{N} x_n f(x_{n-1})}{\sum_{n=1}^{N} (f(x_{n-1}))^2}\,, \quad \hat{\sigma}(x) = \frac{1}{N}\sum_{n=1}^{N} (x_n - \hat{\alpha}(x) f(x_{n-1}))^2\,, \quad o(x) = \frac{1}{N}\sum_{n=1}^{N} (f(x_{n-1}))^2\,. \tag{2}$$

The first two statistics are MLE parameter-estimators, inferring α from the auto-correlation and σ from the residuals, respectively. The third one, $o(x)$, can be seen as an *order parameter* that is needed to tell the two phases (associated with the two attractors) apart. If typical trajectories spend most of their time randomly near either one of the two attractors, parameter estimators will be insufficient to tell the phases apart, and thus lead to wrong posteriors when used alone.

Fig. 2 compares the regression of the two model parameters by the two MLE estimates in (2) with those achieved by the encoder. Apart from boundary effects from the prior, the encoder seems to regress the parameters as good as the MLE estimators. The α-regression shows a clear accuracy-difference, for the two phases.

Fig. 3 compares the three summary statistics in (2) with the latent variables from the modified AE. Both are nicely concentrated around two disjoint but overlapping sub-manifolds corresponding to the two phases of the model. Thus, the decoder has succeeded in creating an incentive for the encoder to separate the two phases.

The success of the modified AE lies in its combination of regression and reconstruction techniques. The reconstructor (decoder) both facilitates the regression and incentivizes the encoder to also learn auxiliary statistics that are needed for Bayesian inference. Often, it might not be easy to separate bare noise from stochastic model simulations (think, e.g., of the Ising model). In such situations, we might convey noise information implicitly to the reconstructor in the form of replica of model realizations (for fixed parameters) rather than explicit noise realizations, for each model realization. Preliminary results from the BISTOM

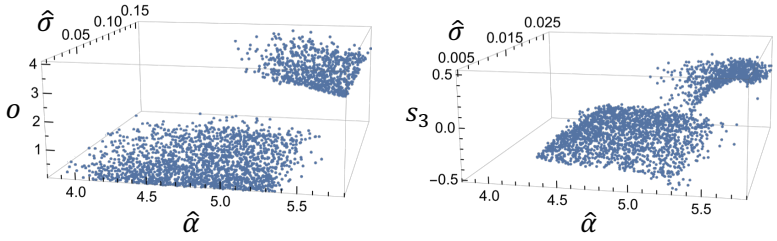

Figure 3: Prior distribution of summary statistics (2) (left panel), and of the latent variables from the modified AE (right panel). The first two latent variables are regularized to be parameter regressors, while the third one (s_3) is unconstrained. The two sheets correspond to the two phases of the model and overlap on the $(\hat{\alpha}, \hat{\sigma})$-projection (phase-coexistence).

project confirm the validity of this approach.

A focus of this project will be on using the learned features for *Bayesian inference*. First of all, a comparison of the features of *observed* data with the prior feature-distribution might reveal a *prior-data conflict*, which could be difficult to see when comparing the raw data with typical model runs. But our focus will be on *uncertainty quantification*, which requires assessing the parameter-related information content of the learned features of stochastic model-outputs. We will do this by means of cutting-edge sampling techniques for Bayesian inference.

SEE NVP!!

Non-volume preserving (NVP) transformations. We have already shown how AEs can be used to obtain near-sufficient summary statistics for ABC and to provide good posterior approximations. In this project, we will also work on replacing the AE by an NVP. The NVP does not have a bottleneck dimension, but attempts to learn a deterministic map from a simple distribution on a latent space to the target distribution on the space of model outputs. First, it can serve as an emulator of the model if the latter is too computationally expensive to forward simulate. Second, the NVP provides likelihood evaluation for the observed samples, and therefore provides an alternative way of approximating the posterior. Finally, the latent variables should be drawn from a simple distribution for which we can verify that the test samples are actually adequately represented by the model. We will compare this approach with the results obtained using ABC with AE-generated summary statistics.

We will use flows in three different ways to assess the summary statistics for ABC. First, we will use the Glow model [Kingma and Dhariwal, 2019]. For this model the latent dimension is the same as the time series that we will try to model, but the latent dimension prior is a zero-mean unit-variance Gaussian distribution. We expect the latent dimension to disentangle the effect of parameters and noise and that a projection using t-SNE will provide meaningful summary statistics. Second, we will use a conditional flow [Winkler et al., 2014] that uses the parameters of the model as inputs so we can regress them too. This will

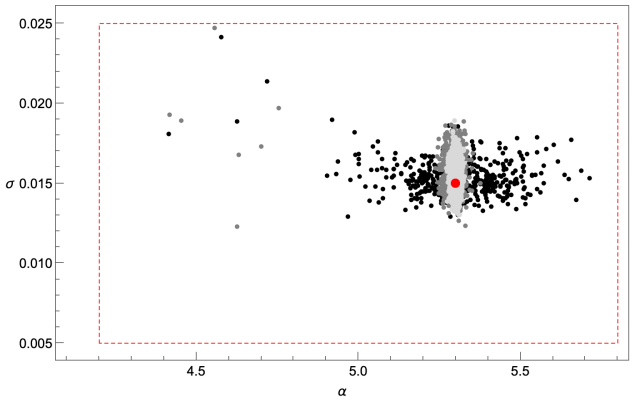

Figure 4: Exact posterior for model (1) (light grey) compared against ABC posteriors using all three latent variables from the modified AE (dark grey) and only the two regressors (black).

provide an additional degree of information. Finally, we will use as latent variables the random realization that has been used to simulate the model, as we have proposed in Fig 1 for the AE solution. In this way, the conditional flow will be using the same random variables as latent and would be able to better emulate the physical system.

2.3.2 Bayesian Inference

We employ two sampling techniques that have proven capable of sampling the posterior, for the type of stochastic models considered here.

Approximate Bayesian Computation (ABC). We consider models with relatively few parameters, $\boldsymbol{\theta}$, but possible high-dimensional outputs $\mathbf{x}$. Furthermore, the LHD, $f(\mathbf{x}_{obs}|\boldsymbol{\theta})$, for an observation $\mathbf{x}_{obs}$ is assumed to be prohibitively expensive to evaluate, whereas simulating a model output $\mathbf{x}$ from the associated stochastic model with density $f(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be cheap. Under these conditions, the posterior $f(\boldsymbol{\theta}|\mathbf{x}_{obs})$ can be approximated using ABC: model outputs $\mathbf{x}$ are simulated, for various $\boldsymbol{\theta}$, which are then accepted or rejected depending on whether $\mathbf{s}(\mathbf{x})$ agrees with $\mathbf{s}(\mathbf{x}_{obs})$ within a certain tolerance (see, e.g. [Albert et al., 2014]). If $\mathbf{s}(\mathbf{x})$ is sufficient and the tolerance converges to zero, this method yields samples from the asymptotically (in the tolerance) exact posterior parameter-distribution. For the tolerance to converge efficiently, $\mathbf{s}(\mathbf{x})$ needs to be low-dimensional and well concentrated. For $\mathbf{s}(\mathbf{x})$ to be (near) sufficient, the dimension of $\mathbf{s}$ needs to be of the order of, but typically larger than, the dimension of $\boldsymbol{\theta}$. Our modified AE is designed to learn summary statistics that are not only near sufficient, but also well concentrated.

For our case study model (1), we compare the approximate posteriors that have been achieved with ABC using all three latent variables encoded by our AE against the one resulting from only the two regressors (Fig. 4). It is evident that the third statistic (order parameter) indeed contains relevant information for constraining the parameters and that the modified AE is capable of learning near-sufficient summary statistics, for this example.

Hamiltonian Monte Carlo (HMC) sampling. Unlike ABC, HMC requires the analytical LHD (possibly

up to a normalizing constant): $f(\mathbf{x}_{obs}|\boldsymbol{\xi},\boldsymbol{\theta})$ where ξ is a latent variable or nuisance parameter not of primary interest. In this setting issues arise because the posterior of the parameter of interest $\boldsymbol{\theta}$ is written in the form $f(\boldsymbol{\theta}|\mathbf{x}_{obs}) \propto \int f(\mathbf{x}_{obs}|\boldsymbol{\xi},\boldsymbol{\theta})f(\boldsymbol{\xi}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\xi}$, for some high-dimensional state-vector $\boldsymbol{\xi}$. To avoid integration we may attempt to sample pairs of states and parameters $(\boldsymbol{\xi},\boldsymbol{\theta})$ directly from the integrand, by means of HMC. In [Albert et al., 2016] we have shown that, for simple stochastic differential equation models, HMC is efficient if the different time-scales naturally appearing in the Hamiltonian dynamics are separated. However, in these simulations, a change of parameters needs to be accompanied by a *collective* change of all the state variables, which slows down the mixing in parameter space. If the posterior in parameter space has more than one mode, slow mixing can essentially invalidate the method. This problem is inherent to other approaches, such as Particle Filters or Gibbs sampling [Reichert and Mieleitner, 2009], as well. This problem is also pertinent to MD simulations in physics, where it manifests itself in the from of high free energy barriers. As HMC is essentially the same as MD, we expect to find answers to this problem in the physics literature. We find that the method of using *collective variables* for biasing the sampling towards overcoming free energy barriers [Laio and Parrinello, 2002], could be brought to fruition in HMC as well. Recent successes with the ML of collective variables [Wetzel, 2017, Carrasquilla and Melko, 2017] encourage us to use our ML-generated low-dimensional features also for this purpose.

2.3.3 Applications to real data

Hydrological application. We expect our modified AE to be able to learn catchment properties, in terms of *runoff-features*. Therefore, we first train it on pairs of data sets $(\mathbf{x},\boldsymbol{\epsilon})$ from a large number of gauged catchments, where $\mathbf{x}$ are the observed runoff time-series, and $\boldsymbol{\epsilon}$ the observed rainfall-time series as well as time-series of potential evapo-transpiration (which is calculated from observed temperature time-series). The decoder of the trained modified AE can then be used for runoff predictions, given predictions of rainfall and evapo-transpiration, for all of these catchments. We will compare these predictions against state-of-the-art predictions with conceptual hydrological bucket models. Relating the distribution of learned latent variables $\mathbf{s}$ with known catchment properties, we might be able to learn which catchments properties are important for runoff prediction, and thus use the trained AE also for runoff predictions in ungauged catchments.

Epidemiological application. We will estimate, via an AE-ABC (ABC where summary statistics are designed with an AE using the ID to determine their indicative number), parameters of an extended epidemiological SEIR type compartmental model on a dynamical network structure (along the lines of [Dutta et al., 2018c, Varghese et al., 2020]). In the network model each Canton (this is the *spatial resolution* we aim at) is considered to be a node and the degree of connectivity between Cantons is described by an evolving - over time - adjacency matrix. The advantage of our proposal is that the unrealistic assumption of uniform mixing of traditional epidemilogical compartmental models is dropped and, as a result, the effect

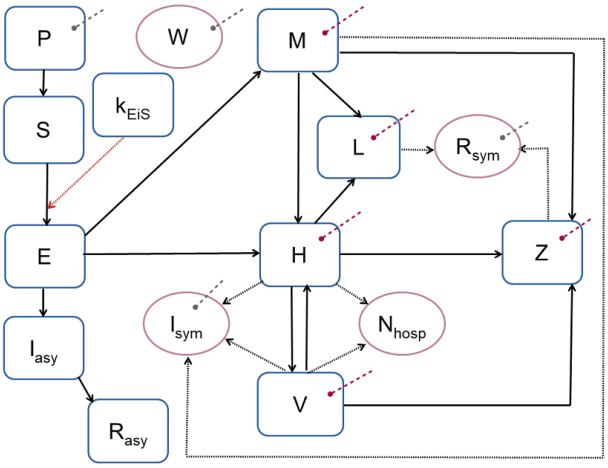

Figure 5: Block diagram of the node-level epidemiological model, with arrows representing flows between compartments, dotted black arrows representing the contributions of a compartment to the size of another compartment. Dotted lines ending with a full circle entering a compartment represent available observed data for that compartment.

of various types of Non-Pharmaceutical Interventions that aim at slowing down the contagion dynamics, can be thoroughly studied because they are reflected in a modification of the underlying network structure. We will work on a daily *time resolution* since this is the time frame of data availability. In accordance with decision makers we will provide predictions at 3 and 7 days to allow them to timely act also based on the dashboard we will make available to support their decisions.

The standard compartmental SEIR type model (Susceptibles, S, Exposed, E, Infected, I, Removed, R) will be extended to capture a variety of realistic dynamical features observed for COVID-19 and will be embedded into a network framework. These are the novelties of our proposal and jointly considered make the LHD non analytically available so that ABC is required to estimate the model parameters. Within each network node the time-course of the epidemics will be represented by a differential equations model of an advanced SEIR type, with the introduction of additional compartments taking into account both the entire set of available COVID-19 data and the complex dynamics of a pandemic (as the consideration of asymptomatic individuals, I_{asy}, or the possibility that there exists individuals who, even if potentially susceptible, are geographically or socially remote from the source of infection, so that S becomes a subset of a wider compartment P). For Switzerland (similarly to several other countries hit by COVID-19) available data, for most of the Cantons, include for example the cumulative number of people tested (swabs, W) or of confirmed cases (symptomatic exposed people, I_{sym}), the current number of patients in home quarantine (M), of patients hospitalized with symptoms (H) or admitted in intensive care units (V), cumulative number of recovered individuals (L) or deceased persons (Z). The structure of the model inside each node will be identical, but value of the model parameters may vary between nodes to reflect canton specific situations. The infection coefficients will be allowed to decrease over time to take into account the effect of mitigation measures put in place at federal as well as cantonal level. Figure 5 shows the model block diagram.

2.3.4 Data availability

For the *hydrological* application, we will use data that is available through the Federal Office for the Environment (FOEN), which operates a large monitoring network in Switzerland, including about 200 stream gauging stations, and 91 groundwater level stations (source: www.bafu.admin.ch). The data-sets include time-series of precipitation, temperature, potential evaporation, and streamflow, covering several decades, at an hourly time step. Furthermore, detailed maps of topography, soil, geology and vegetation, are available, which will be used to extract *catchment properties*. Catchments will be selected based on the availability of long term data series, as well as to ensure a wide range of catchment conditions, such as with presence or absence of snow melt as a dominant runoff generation process, or long vs short memory in terms of streamflow response behaviour.

For the *epidemiological* application there will be various data sources, most of which publicly available, ranging from daily counts of reported confirmed COVID-19 cases, deaths and recoveries for each Canton. For some Cantons, a richer data structure is also freely available including counts on people quarantined, hospitalised with symptoms and in intensive care, suspected cases and so on. Urban transportation and travel data together with Google and Apple mobility data are also available. This data, needed to build the network adjacency matrix, will be enhanced using the availability of air traffic (from OAG https://www.oag.com/) that A. Mira has access to thanks to the COVID-19 related H2020 research projects she is involved with that will also provide data on a finer scale and on other quantities such as data consisting in COVID-19 patient information from selected hospitals. Specifically, for each patient, the following information are (anonymously) recorded: the date of admission, the hospital ward (ICU, Emergency room, General ward) to which they were first admitted, the date and hospital ward of any subsequent movements within the hospital, plus the date and final status (discharged from the hospital, transferred to another hospital, deceased). With this data we can learn the patient journey inside a hospital and data from different patients will allow us to build a network where nodes are the emergency room, regular hospital ward and ICU (or more detailed hospital units). The structure of the network can be associated with the hospital/patient specific survival probability as well as with patient covariates. In this setting ABC allows to handle non-Markovian stochastic models where, for example, transition probability of moving from one unit/state/node to the other can realistically depend on the time spent within each unit. All the data needed for this application will be made available thanks to the H2020 Covid-19 project where the PI is involved as co-PI.

2.4 Schedule and milestones

Organization and time plan of the work packages is summarized in Table 1 where, for each WP we indicate the leader (L), co-leader (co-L) and the participants (P). Both the PI and co-PI will co-lead the methodological

						Year and semesters							
WP	AM	FPC	CA	JP	AL	2021		2022		2023		2024	
						1	2	1	2	1	2	1	2
1	co-L	co-L	P	P	P		M1.1			M1.2	M1.3		
2	P	co-L	co-L		P				M2.1				M2.2
3	co-L	P		co-L	P			M3.1				M3.2	

Table 1: Overview of WPs with associated partners and timing for the milestones. WP leader is marked with the letter "L", participants are marked with "P".

WP1 that will extend for the first 3 years of the project. The epidemiological application will be co-lead by the PI together with JP Onnela (JP): both of them have already worked together on a related topic in a project founded by SNF and recently concluded with success. The hydrological application with be co-lead by the co-PI together with C. Albert (CA) an expert in hydrology: both of them have already worked together in the BISTON project (also just ended) that has lead the foundations of the current proposal. A. Laio (AL) will participate in all WPs being an expert in MD and MCMC in generael and a pioneer of the methodology twe will use to estimate the Intrinsic Dimension, a tool that we will leverage upon. A. Laio has published with the PI on the Bayesian version of the methodology to estimate the Intrinsic Dimension.
WP I: Methodological developments *Objective:* Explore how ML can be used to learn low-dimensional features in high-dimensional outputs of stochastic models, for Bayesian inference. *Tasks:*

T1.1 Further developing our modified AE towards learning of minimal sufficient statistics.

T1.2 Train NVPs for approximating the density of the distribution of model outputs.

T1.3 Assess the ability of ID methodology to determine the optimal number of summary statistics.

T1.4 Test the parameter-related information content of the AE learned statistics by means of ABC. The performance of our modified AE and NVP will be benchmarked against existing ML algorithms for feature-learning (such as VAE). Test cases will be used, for which the true posteriors (in the form of large samples) are available for comparison, such as:

• Non-linear iterative time-series models. The LHD is available as a product of conditional LHDs. A minimal set of sufficient summary statistics is available for comparison. Non-linearity may lead to different phases, in which case more summary statistics than parameters are required. (For this type of models, we have achieved already near perfect results in the BISTOM project.)

• Discretely observed non-linear SDE model (i.e. discretization needed for simulating the SDE is much finer than the observational frequency, thus the noise $\boldsymbol{\epsilon}$ is of much higher dimension than $\mathbf{x}$.) The LHD

is a very high dimensional integral, typically outside the exponential family, but the true posterior is available via HMC [Albert et al., 2016].

• Queuing models (typical benchmarking models in the ABC community, e.g.
[Sutton and Jordan, 2011]). The LHD is a high dimensional integral, but the true posterior available through Gibbs and slice sampling.

T1.4 Assessing usefulness of features learned by the AE, for biasing HMC algorithms for fast mixing.

Milestones:

• M1.1 ML-generation of summary statistics (by AE) which allow us to achieve an (almost) exact posterior, for a non-trivial example, i.e., an example for which previously some sort of brute-force integration was needed to sample the posterior.

• M1.2 density reconstruction (by NVP) on the same non-trivial example used to test the AE.

• M1.3 For a non-trivial SDE model: ML-generation of collective variables, which allow us to speed up the slow-mixing modes of HMC-samplers.

Deliverables:

• D1.1 Presentation of the method to the ML community via publication(s) and at conferences.

• D1.2 Making the software available via the Renku platform.

Risks and feasibility: Thanks to previous work in the BISTOM project, we know that the proposed method of using a modified AE for learning summary statistics works, in principle. The ID methodology is worth exploring and potentially very useful: little cost, potentially very high gain. However, the success of the rest of the project does not depend on it. Using AEs for learning collective variables and using them to speed-up HMC simulations appears very feasible, due to the similarity between HMC for Bayesian inference problems and MD for Statistical Mechanics problems, and the successes achieved in the latter. The NVP flow model is a novel idea that can potentially provide a significant improvement. It is a high risk and high reward procedure.

WP II: Hydrological application

Objectives: Explore the possibility of learning features in runoff imprinted by catchment properties. Explore the use of AEs for runoff-prediction, both in gauged and ungauged catchments. *Tasks:*

T2.1 Train modified AE with measured rainfall, evapo-transpiration and runoff-data from Swiss catchments.

T2.2 Benchmark predictive power of the decoder against conceptual, as well as recent ML-models (i.e. algorithms directly learning the relationship between forcing variables and runoff).

T2.3 Map the learned features with known catchment properties, such as topography, soils, vegetation and geology, for the purpose of runoff-predictions in ungauged catchments.

Milestones:

• M2.1 Availability of a trained AE, for runoff-prediction and benchmarking.

• M2.2 Assessment of information content in learned features w.r.t. catchment properties.

Deliverables:

• D2.1 Presentation of results to the hydrological community via publications and at conferences.

• D2.2 Making software for the modified AE available.

Risks and feasibility: The success of using the proposed method for feature-learning in real data will depend crucially on the amount of available data. Recent successes of pure ML-approaches in hydrology [Kratzert et al., 2019] allow to be optimistic.

WP III: Epidemiological application

The description of this WP is more detailed than the one of WP II because the hydrological application has provided the case study in the body of the proposal and as a consequences more modeling/algorithmic details have been given earlier. *Objectives:*

• Provide a decision-support tool for policy makers to assess effect of pandemic mitigation strategies.

• Assist local public health authorities to identify the communities that require particular attention.

• Plan COVID-19 prevention and control of morbidity and mortality in advance rather than responding to the occurrence of outbreaks.

• Determine effect of mitigation measures on spatio-temporal spread of COVID-19 pandemic.

Tasks:

T3.1 Collect and wrangle publicly available data sources on COVID-19 and familiarize with the literature. Design possible adjacency matrices either based on proximity measures or on urban transportation and travel data or a mixture thereof. Matrices will be weekly updated using Google and Apple mobility data (freely available).

T3.2 Construct the most suitable network based epidemiological model for Switzerland - building on research lines already started by A. Mira - and the ABC algorithm to estimate it.

T3.3 Train our modified AE on the model from T3.2 to provide efficient summary statistics for the AE-ABC. Test the predictive power of the estimated model and compare with literature.

Milestones:

• M3.1 AE-ABC inference for epidemiological models calibrated on Swiss data.

• M3.2 Assess predictive power of AE-ABC epidemiological model and its use as decision support tool

Deliverables:

• D3.1 Presentation of results to the stat/epidemiological community (publications and conferences).

• D3.2 Making software for the AE-ABC available.

• D3.3 Dashboard as a decision support system for better management of an epidemic emergency.

Risks and feasibility: The main risk is related to possibly poor data quality due to sources of potential bias and the fact that most of the easily available data is observational rather than sample based. The other risk is related to the difficulty in finding the appropriate model complexity. However, preliminary studies by AM with (non-stochastic) spreading-models on networks indicate that model structures of the kind shown in Fig. 5 are suitable for the available data and prediction purposes.

2.5 Relevance and impact

We expect the developments in the methodological part of this project to be useful in many domains. Indeed, a general ML-method to learn near-sufficient summary statistics obtained from pseudo data generated by stochastic models could make Bayesian inference accessible to many disciplines where stochastic models are used, and where this previously was computationally infeasible. Furthermore, such summary statistics are also useful beyond Bayesian statistics, as they provide dimensionality reduction by condensing relevant information and can also inform us about different modes of operation (phases) of stochastic models, which might be impossible to distinguish on the level of high-dimensional model outputs. Finally, our ML methods could also be used to extract relevant features directly from observations and the ID will drive the design of the architecture of the AE in a data driven and interpretable way. The NVP flow models will be used for the first time to estimate the summary statistics for ABC. This deterministic tool for estimating densities from samples is opening many different avenues in machine learning and. In this way the flows will serve two purposes: a simple emulator for complex physical simulators and providing the summary statistics for approximate inference. The expected results from our two case studies could have a real-world-impact in the fields of hydrology and infectious disease epidemiology. Our approach could lead to the discovery of new catchment-features in runoff data, which in turn could help predicting runoff from catchment properties in un-gauged catchments, which represents one of the unsolved problems in catchment hydrology today [Dal Molin et al., 2019]. Stochastic models on networks are very useful when studying the spread of viruses under various scenarios where simple and interpretable models seem to be favored over less interpretable and more complex ones. Quantifying the uncertainty of their parameters from observations is very difficult, but urgently needed when deciding about containment measures. Our approach is expected to facilitate this and could thus potentially have a very big impact in light of the current COVID pandemic: interpretability and uncertainty quantification are key features when support tools are used for decisions that impact citizens health, well being, economic situation and freedom.

References

[Albert et al., 2014] Albert, C., Künsch, H.-R., and Scheidegger, A. (2014). A Simulated Annealing Approach to Approximate Bayes Computations. *Stat. Comput.*, 25(6):1217–1232.

[Albert et al., 2016] Albert, C., Ulzega, S., and Stoop, R. (2016). Boosting Bayesian parameter inference of nonlinear stochastic differential equation models by Hamiltonian scale separation. *Physical Review E*, 93(4):043313.

[Allegra et al., 2020] Allegra, M., Facco, E., Denti, F., Laio, A., and Mira, A. (2020). Data segmentation based on the local intrinsic dimension. *Scientific Reports*, 10(1):1–12.

[Bartolucci et al., 2021] Bartolucci, F., Pennoni, C., and Mira, A. (2021). A multivariate statistical model to predict COVID-19 count data with epidemiological interpretation and uncertainty quantification. *Minor revision in Statistics and Medicine*.

[Bonati et al., 2020] Bonati, L., Rizzi, V., and Parrinello, M. (2020). Data-driven collective variables for enhanced sampling. *The Journal of Physical Chemistry Letters*, 11(8):2998–3004.

[Carrasquilla and Melko, 2017] Carrasquilla, J. and Melko, R. G. (2017). Machine learning phases of matter. *Nature Physics*, 13(5):431–434.

[Chen et al., 2020a] Chen, S., Mira, A., and Onnela, J.-P. (2020a). Flexible model selection for mechanistic network models. *Journal of Complex Networks*, 8(2):cnz024.

[Chen et al., 2020b] Chen, Y., Zhang, D., Gutmann, M., Courville, A., and Zhu, Z. (2020b). Neural approximate sufficient statistics for implicit models. *arXiv preprint arXiv:2010.10079*.

[Chinazzi et al., 2020] Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., y Piontti, A. P., Mu, K., Rossi, L., Sun, K., et al. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, 368(6489):395–400.

[Cui et al., 2018] Cui, T., Peeters, L., Pagendam, D., Pickett, T., Jin, H., Crosbie, R. S., Raiber, M., Rassam, D. W., and Gilfedder, M. (2018). Emulator-enabled approximate Bayesian computation (ABC) and uncertainty analysis for computationally expensive groundwater models. *Journal of Hydrology*, 564:191–207.

[Cvitkovic and Koliander, 2019] Cvitkovic, M. and Koliander, G. (2019). Minimal achievable sufficient statistic learning. *arXiv preprint arXiv:1905.07822*.

[Dal Molin et al., 2019] Dal Molin, M., Schirmer, M., and Fenicia, F. (2019). Comparing different approaches for predictions in ungauged catchments. *EGUGA*, page 1772.

[Dinh et al., 2014] Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2014). Density estimation using real NVP. In *Proceedings of ICLR*.

[Dutta et al., 2018a] Dutta, R., Brotzakis, Z. F., and Mira, A. (2018a). Bayesian calibration of force-fields from experimental data: Tip4p water. *The Journal of chemical physics*, 149(15):154110.

[Dutta et al., 2018b] Dutta, R., Chopard, B., Lätt, J., Dubois, F., Zouaoui Boudjeltia, K., and Mira, A. (2018b). Parameter estimation of platelets deposition: Approximate bayesian computation with high performance computing. *Frontiers in physiology*, 9:1128.

[Dutta et al., 2018c] Dutta, R., Mira, A., and Onnela, J.-P. (2018c). Bayesian inference of spreading processes on networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2215):20180129.

[Dutta et al., 2017] Dutta, R., Schoengens, M., Onnela, J.-P., and Mira, A. (2017). ABCpy: A user-friendly, extensible, and parallel library for approximate Bayesian computation. In *Proceedings of the platform for advanced scientific computing conference*, pages 1–9.

[Dutta et al., 2021] Dutta, R., Schoengens, M., Pacchiardi, L., Ummadisingu, A., Widmer, N., Onnela, J., and Mira, A. (2021). A high-performance computing perspective to Approximate Bayesian Computation. *Journal of Statistical Software, to appear; Preprint arXiv:1711.04694*.

[Ebert et al., 2021] Ebert, A., Dutta, R., Mengersen, K., Mira, A., Ruggeri, F., and Wu, P. (2021). Likelihood-free parameter estimation for dynamic queueing networks: case study of passenger flow in an international airport terminal. *Journal of the Royal Statistical Society C, to appear*.

[Efron et al., 1975] Efron, B. et al. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 3(6):1189–1242.

[Fearnhead and Prangle, 2012] Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. Roy. Stat. Soc. B*, 74(3):419–474.

[Gourieroux et al., 1993] Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of applied econometrics*, 8(S1):S85–S118.

[Gourieroux et al., 1996] Gourieroux, M., Gourieroux, C., Monfort, A., Monfort, D. A., et al. (1996). *Simulation-based econometric methods*. Oxford university press.

[Invernizzi and Parrinello, 2020] Invernizzi, M. and Parrinello, M. (2020). Rethinking Metadynamics: from bias potentials to probability distributions. *The Journal of Physical Chemistry Letters*, 11(7):2731–2736.

[Jiang et al., 2017] Jiang, B., Wu, T.-y., Zheng, C., and Wong, W. H. (2017). Learning summary statistic for approximate Bayesian computation via deep neural network. *Statistica Sinica*, pages 1595–1618.

[Kingma and Dhariwal, 2019] Kingma, D. P. and Dhariwal, P. (2019). Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*.

[Kratzert et al., 2019] Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology & Earth System Sciences*, 23(12).

[Laio and Parrinello, 2002] Laio, A. and Parrinello, M. (2002). Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566.

[Levy et al., 2017] Levy, D., Hoffman, M. D., and Sohl-Dickstein, J. (2017). Generalizing Hamiltonian Monte Carlo with neural networks. *arXiv preprint arXiv:1711.09268*.

[Linka et al., 2020] Linka, K., Rahman, P., Goriely, A., and Kuhl, E. (2020). Is it safe to lift COVID-19 travel bans? The Newfoundland story. *Computational Mechanics*, 66(5):1081–1092.

[McFadden, 1989] McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica: Journal of the Econometric Society*, pages 995–1026.

[Mira and Haario, 2011] Mira, A. and Haario, H. (2011). Discussion of: Riemann manifold Langevin and Hamiltonian Monte Carlo Methods by m. girolami and b. calderhead. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

[Papamakarios and Murray, 2016] Papamakarios, G. and Murray, I. (2016). Fast ϵ-free inference of simulation models with Bayesian conditional density estimation. *arXiv preprint arXiv:1605.06376*.

[Papamarkou et al., 2014] Papamarkou, T., Mira, A., Girolami, M., et al. (2014). Zero variance differential geometric markov chain monte carlo algorithms. *Bayesian Analysis*, 9(1):97–128.

[Parolini et al., 2021] Parolini, N., Dede, L., Antonietti, P. F., Ardenghi, G., Manzoni, A., Miglio, E., Pugliese, A., Verani, M., and Quarteroni, A. (2021). SUIHTER: A new mathematical model for COVID-19. application to the analysis of the second epidemic outbreak in Italy. *arXiv preprint arXiv:2101.03369*.

[Raynal et al., 2021] Raynal, L., Chen, S., Mira, A., and Onnela, J.-P. (2021). Scalable Approximate Bayesian Computation for growing network models via extrapolated and sampled summaries. *Bayesian Analysis*, 1(1):1–28.

[Reichert and Mieleitner, 2009] Reichert, P. and Mieleitner, J. (2009). Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resources Res.*, 45.

[Šukys et al., 2018] Šukys, J., Kattwinkel, M., and Reichert, P. (2018). SPUX: Scalable high performance uncertainty quantification framework for stochastic models in environmental data sciences. *EGUGA*, page 5555.

[Sutton and Jordan, 2011] Sutton, C. and Jordan, M. I. (2011). Bayesian inference for queueing networks and modeling of internet services. *The Annals of Applied Statistics*, pages 254–282.

[Tomassini et al., 2009] Tomassini, L., Reichert, P., Künsch, H. R., Buser, C., Knutti, R., and Borsuk, M. E. (2009). A smoothing algorithm for estimating stochastic, continuous time model parameters and its application to a simple climate model. *J. Roy. Stat. Soc. C*, 58(5):679–704.

[Varghese et al., 2020] Varghese, A., Drovandi, C., Mira, A., and Mengersen, K. (2020). Estimating a novel stochastic model for within-field disease dynamics of banana bunchy top virus via approximate Bayesian computation. *PLoS computational biology*, 16(5):e1007878.

[Warne et al., 2020] Warne, D. J., Ebert, A., Drovandi, C., Hu, W., Mira, A., and Mengersen, K. (2020). Hindsight is 2020 vision: a characterisation of the global response to the COVID-19 pandemic. *BMC public health*, 20(1):1–14.

[Wetzel, 2017] Wetzel, S. J. (2017). Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Physical Review E*, 96(2):022140.

[Winkler et al., 2014] Winkler, C., Worrall, D., Hoogeboom, E., and Welling, M. (2014). Learning likelihoods with conditional normalizing flows. In *ArXiv 1912.00042*.