Towards Simulating Whole Cells

PIs: Prof. Dr. Tim Conrad (FU/ZIB), Prof. Dr. Martin Lohse, Prof. Dr. Christof Schütte (ZIB)

Background Drug-induced perturbations of the endogenous metabolic network are a potential root cause of cellular toxicity and can lead to severe complications during a drug therapy. A mechanistic understanding of such unwanted side effects is therefore vital for patient safety in terms of dosing, diagnosis, or the design of curative intervention strategies.

An attractive system to study such phenomena are so called genome-scale metabolic models, which contain information about associated biochemical reactions and their stoichiometry ([3]). The release of human genome-scale metabolic models has enabled large-scale metabolic analysis of many diseases, and provides a promising approach for predicting gene-to-phenotype linkages, for example, drug side effect associations. However, modeling the metabolism of a cell is only one piece of a very complex puzzle. To understand comprehensively, what is happening inside a cell, much more components are needed, such as signaling, transport or transcription and translation. Or in the language of omics: not only metabolomics is needed but also proteomics and genomics.

During the last years, many more of these building blocks became available as models, of- ten even including detailed kinetic information, e.g. for the cell cycle or particular signaling pathways. Coupling and integrating all these components leads to very complex *whole cell models*. Whole cell models are believed to be the only way on how to fully understand how a cell functions and thus contributing to our understanding of life and how to use this for designing personalized therapies based on patient-adjusted simulations of drug metabolism. The general feasibility of this idea has been shown by Karr and coworkers [4] on the example of the human pathogen *Mycoplasma genitalium*. This whole cell model includes all molecular components and interactions which are known about this organism and it was demonstrated that a full simulation of this model could reproduce most of the experimental data available for a variety of different conditions.

Despite the immense work needed to generate whole cell models, also strategies about simulations of these systems should be thought of, especially for the case when the models get more complicated than the Mycoplasma genitalium. Simulation environments such as V-Cell, M-Cell and E-Cell are available today to perform detailed simulations of biological systems based on ODEs, PDEs, Monte Carlo and stochastic (Gillespie-type) methods. Obviously, the main bottle- neck here is computational complexity and thus the pure run-time for a simulation on biologically relevant time-scales.

In the suggested project, we are interested in building a new mathematical method to allow drastic reduction of simulation time, inspired by ideas from the molecular dynamics community. It will use current ideas from state-of-the-art model reduction techniques and extend them to be applicable to simulations of whole cells.

Goals of the Project and Methodologies The goal of this project is to develop a method for learning a low-dimensional approximation of a highly non-linear dynamical system on the example of an available whole-cell simulation system. We suggest to use an architecture based on neural networks and auto-encoders to approximate the underlying Koopman operator. The Koopman operator can be used to describe how a dynamical system evolves over time. Usually, the Koopman operator can be approximated from given data (e.g. simulation output of a dynamical system) by methods such as Extended Dynamic Mode Decomposition (EDMD). The problem with EDMD is that it need a dictionary of functions as an input, but no efficient methods is known to choose the dictionary such that the representation of the data in the Koop- man subspace is exact. To circumvent this problem, extensions have been proposed such as Kernel-EDMD, but this only shifts the problem to how to choose the kernel function. Further, the identified number of building blocks that EDMD and Kernel-EDMD identify (e.g. modes or eigenfunctions) grow with the size of the dictionary.

What we are interested in in this project is therefore a method to identify a small sub-set of modes that can be used to build reduced order models. On the one hand, this idea is in principle similar to available approaches such as Optimal Mode Decomposition, sparsity-promoting DMD or SINDy but on the other hand differs from recent approaches that aim at learning an optimal EDMD dictionary through deep learning (see e.g. [1]). However, the main difference to the mentioned available approaches is that those are based on the idea to reconstruct the full state in a *linear* way. We suggest a method in which the full state will be reconstructed in a *non-linear* way using a deep auto-encoder. We believe that this is beneficial since it will allow to use less observables for reconstruction.

Deep learning network models have the ability to efficiently learn complex patterns from data. Similarly, part of this project will be to show that they are also able to represent complex non-linear observables more efficiently than dictionary based methods, given that large amounts of training data is available. This is the case in this project, since we can simulate the full system in arbitrary length even under different conditions. We suggested a model structure based on deep encoder and decoder networks:

Encoder: $z = \Phi_{enc}(x; M_{enc})$ (where M_{enc} contains the model parameters) and Decoder: $x^{i} = \Phi_{dec}(z; M_{dec})$ (where M_{dec} contains the model parameters)

Here, the encoder network will learn the dictionary. The main advantage to use a non-linear decoder is that it can extract more information about the full state from fewer features which are extracted by the encoder. This usually leads to smaller dictionaries compared to linear approaches, if the encoder and decoder are trained simultaneously.

The missing piece is now to learn an evolution operator E describing the time-evolution of z which can also be learned through the training process. We will work on the analysis of E to establish connections to Koopman modes, based on our earlier work [2].

[1] Q Li, F Dietrich, EM Bollt, and IG Kevrekidis. Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the koopman operator. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(10):103111, 2017.

[2] S Klus, Koltai P, and Schütte Ch. On the numerical approximation of the perron-frobenius and koopman operator. *Journal of Computational Dynamics*, 3(2), 2016.

[3] A Bordbar and Bernhard PO Palsson. Using the reconstructed genome-scale human metabolic network to study physiology and pathology. *Journal of internal medicine*, 271(2):131–141, 2012.

[4] JR Karr, JC Sanghvi, DN Macklin, MV Gutschow, JM Jacobs, B Bolival Jr, N Assad- Garcia, JI Glass, and MW Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389-401, 2012.