Challenge accepted - Framework for simulation-based evaluation of machine learning models

Machine learning (ML) is the art of deriving predictions from largely unstructured data. There are many different model versions and architectures that can use diverse inputs and generate diverse outputs. However, not all models are equally suitable for all prediction tasks, and models that perform extremely well on the training set might generalize worse than other models.

How good is your machine learning model? Can it really help solving your problem?

These questions are central to ML. Of course, more or less complex train-validation-test split strategies and evaluation metrics such as accuracy allow for an assessment. But especially in domains such as disease prediction and its biomolecular basics, data tends to be sparse and expensive to gather. How do we know whether the data we have is sufficient to obtain the predictions we aim for? And even more general: How can we find out and systematically evaluate whether and which ML models are performing well for the task we plan, with the data we have?

This project aims to develop a computational software framework that addresses exactly these questions. The framework will enable the user to generate synthetic data as ground truth for a multitude of prediction problems in very different scenarios that can then be used to evaluate the ML models.

Simulating nonlinear dynamics to represent reality

Inspired by a class of problems in molecular life sciences, such as the response of a patient to a drug (see box *Application background: Molecular drug response*), we aim for a framework to represent scenarios that also generalizes to other domains. In detail, it relies on simulations using dynamical systems. A dynamical system is a mathematical model of a process of interest that can be used to simulate its development over time. Complex, highly nonlinear dependencies between variables can be represented, and scenarios from a wide range of application domains can be addressed.

Application background: Molecular drug response

Living cells are made up from millions of molecules that interact and form a highly complex network. This network is composed of a multitude of modules, called molecular pathways, that are responsible for certain functions. In case of disease, e.g. cancer, some of the cellular molecules do not act properly. They could be overly active, disturbed in their connections to other molecules, or even missing entirely. This can cause cellular functions to be disrupted. Drugs to cure the disease are then needed that interfere with the molecules in question, or other relevant molecules (the drug's targets), to restore proper function of the cell. But which drug is the right one for which type of disease? This prediction of drug response is a problem where ML-based approaches have been shown to be helpful.

Users of your framework should be able to determine which features of the scenario are made available as input to the ML model, and which remain hidden. Just as in real-life problems, data is sparse and expensive to generate. Therefore, although it is in principle possible to generate an infinite amount of ground truth and thus training data from the framework, you will implement a flexible approach such that users can define the amount of data (as in: number of samples) that is generated and supplied to the prediction model. In addition, the framework should allow representing different facets of more and less domain-specific confounding factors that can arise in the scenarios, such as noise from measurement techniques.

In the project, different ML models will be implemented to employ the evaluation framework. A variety of complex architectures such as graph neural networks or physics-based neural networks that include knowledge of the underlying dependencies between variables can be used. Approaches to assess which types of data would improve the ML model most, as in uncertainty estimation, will help to decide how to move on.

Why should you join this project? – "Challenge accepted"

We plan this project in a challenge-format (Fig. 1): One team, the challengers, is building a software framework to synthesize data for a difficult prediction task. The other team, the opponents, develops and trains ML models that solve the prediction challenge.
Challengers: The synthesized data resembles nature and should represent real-world observations, e.g., biological processes in molecular pathways, using nonlinear dynamical modeling. Can you fool the opponents by hiding features, letting them predict even more complex functions, or adding a little more noise at just the right spots? After implementing a framework for your challenge-generator, you can switch sides and start tackling challenges yourself.
Opponents: The ML models you develop can range from simple to very modern architectures that include graph-based information or even knowledge of the underlying nonlinear dynamics. In reality, data differ in the associated costs to obtain them, so you will be provided with a certain budget to "purchase" data of varying quality from the challengers. Which data would you buy to improve your model and predictions the most? But be careful, don't use up all your credits at once!

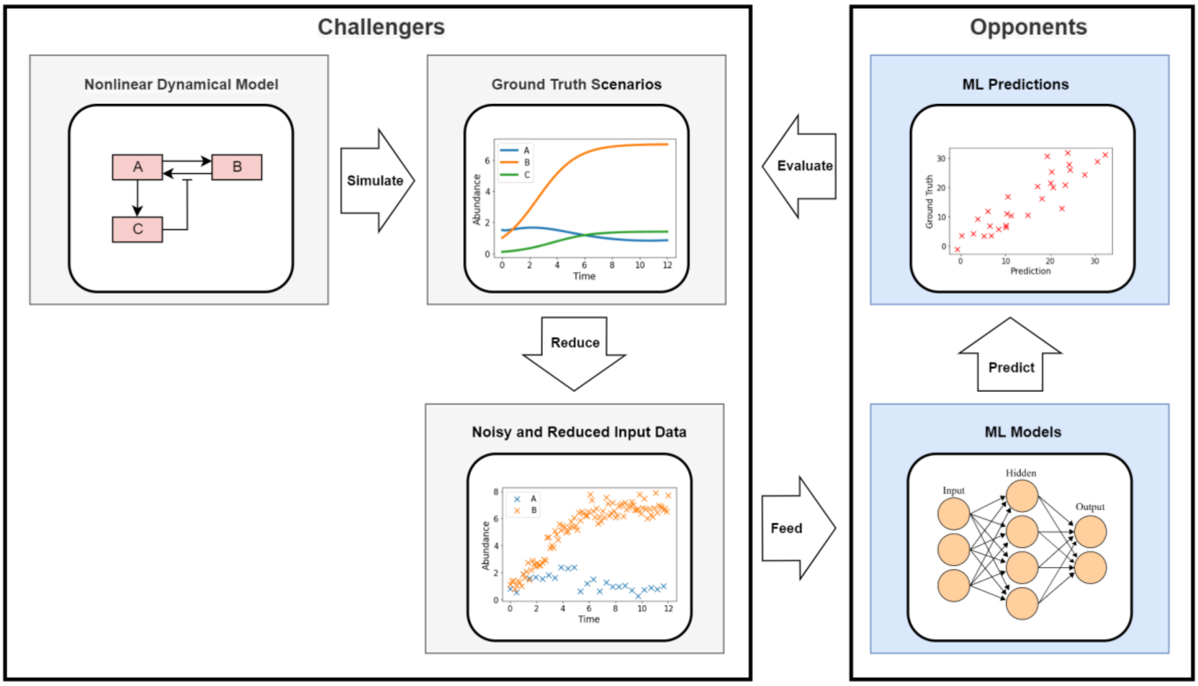

Fig. 1. **Nonlinear dynamical models and predictions by machine learning frameworks.** Challengers construct a framework for simulating dynamical systems and produce multiple scenarios of the systems' evolution. The synthetic data are obfuscated to reflect real-world situations, where noise is omnipresent and variables may be missing. Opponents build a variety of machine learning models to predict quantities of interest using the incomplete data. The predictions are evaluated against the ground truth behavior of the system.

In short and requirements:

- You will practically implement ML models with modern Python programming frameworks such as PyTorch for multiple prediction tasks.
- You can familiarize yourself with nonlinear dynamics and simulations, and with synthetic data generation techniques.
- You will work scientifically and contribute to publishing a paper.
- If you are proficient with git, Python (at least data processing in pandas, NumPy), and ideally already have first or advanced experience with PyTorch or TensorFlow libraries, this project is for you.

Look forward to engaged supervision by Pascal Iversen (pascal.iversen@hpi.de) and Dr. Katharina Baum (katharina.baum@hpi.de) in your upcoming master project! Please do not hesitate to contact us in case of questions.

Further reading

related benchmarking framework for ML (using other dynamical models): Otness, K. *et al.* (2021) *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* Vol. 1 (eds J. Vanschoren, S. Yeung). https://openreview.net/pdf?id=pY9MHwmrymR

graph neural networks (GNNs): informal intro at https://distill.pub/2021/gnn-intro/

physics-informed neural networks (PINN): reviewed in Karniadakis, G. E. *et al.* (2021) Physics-informed machine learning. *Nature Reviews Physics* **3**, 422-440, https://doi.org/10.1038/s42254-021-00314-5.

ML-based prediction of molecular drug response: Umarov, R., Li, Y. & Arner, E. (2021) DeepCellState: An autoencoder-based framework for predicting cell type specific transcriptional states induced by drug treatment. *Plos Comput Biol* 17, e1009465, https://doi.org/10.1371/journal.pcbi.1009465.

Generic nonlinear dynamical modeling for larger networks: Harush, U. & Barzel, B. (2017) Dynamic patterns of information flow in complex networks. *Nature Communications* 8, 2181, https://doi.org/10.1038/s41467-017-01916-3.