PAPER

Optimizing ODE-derived Synthetic Data for Transfer Learning in Dynamical Biological Systems

Julian Zabbarov,[1,‡] Simon Witzke,[1,‡] Maximilian Kleissl,[1] Pascal Iversen,[1,2] Bernhard Y. Renard[1,3] and Katharina Baum[1,2,3,*]

[1]Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Potsdam, Germany, [2]Department of Mathematics and Computer Science, Free University Berlin, Berlin, Germany and [3]Windreich Department of Artificial Intelligence and Human Health & Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York City, USA
[‡]Shared authorship, [*]Corresponding author, katharina.baum@fu-berlin.de
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Motivation: Successfully predicting the development of biological systems can lead to advances in various research fields, such as cellular biology and epidemiology. While machine learning has proven its capabilities in generalizing the underlying non-linear dynamics of such systems, unlocking its predictive power is often restrained by the limited availability of large, curated datasets. To supplement real-world data, informing machine learning by transfer learning with data simulated from ordinary differential equations has emerged as a promising solution. However, the success of this approach highly depends on the designed characteristics of the synthetic data.

Results: We optimize dataset characteristics such as size, diversity, and noise of ordinary differential equation-based synthetic time series datasets in three relevant and representative biological systems. To achieve this, we here, for the first time, present a framework to systematically evaluate the influence of such design choices on transfer learning performance in one place. We achieve a performance improvement of up to 92% in mean absolute error for our optimized simulation-based transfer learning compared to non-informed deep learning. We find a strong interdependency between dataset size and diversity effects. The optimal transfer learning setting heavily relies on real-world data characteristics as well as its coherence with the synthetic data's dynamics, emphasizing the relevance of such a framework.

Availability and Implementation: The code is available at `https://github.com/DILiS-lab/opt-synthdata-4tl`.

Key words: informed machine learning, ordinary differential equations, transfer learning, time series, epidemiology

1. Introduction

Forecasting the development of biological systems over time and conditions enables, for example, the design of drugs and the implementation of effective measures against infectious diseases. However, this task is inherently difficult due to highly nonlinear dependencies present in biological systems, high measurement inaccuracies, and biological variability.

Besides traditional statistical approaches [6], machine learning (ML) has emerged to more powerfully address non-linear forecasting settings [37]. However, ML requires large and curated datasets [1] to learn the dynamics of biological systems. Gathering data in these systems is commonly costly [25] or even unethical, for example, if responses to therapeutics are concerned. In other cases, such as early stages of infectious disease outbreaks, by definition, only few time points are available. Therefore, small biological data often poses a major challenge towards leveraging the capabilities of ML in biosciences [34].

Transfer learning (TL) was introduced as a means to transfer knowledge between related datasets for improved ML-based predictions [26]. Usually, a (deep learning) ML model is pre-trained on data related to the task of interest to learn domain-relevant features [42]. The trained model weights then act as the initialization for fine-tuning the model on the target dataset, i.e., the type of data for which predictions should eventually be derived. TL has become a common technique to address small data and fully leverage available datasets in biological domains. For instance, TL approaches enabled improved infection [33, 27, 13] and drug response predictions [35].

In another approach for dealing with small data settings, researchers have begun to inform ML by including domain knowledge [17, 41], e.g., as encoded in mechanistic models [28], such as ordinary differential equations (ODEs). ODEs, which explicitly include dependencies between variables, have been established for various biological systems [22]. As their solutions often describe the evolution of variables over time, they are

© The Authors 2024.

particularly suited for informing ML for time series. This has been exploited for biomedicine [11, 21], cell migration [20], and epidemiology [30, 12]. Addressing the issue of small data, simulation-based TL, i.e., informing ML by pre-training with ODE-derived synthetic data, was successfully used to improve predictions of chemotherapy treatments [38] or metabolism [9]. To make simulation-based TL more broadly applicable, we introduced the open-source Python library SimbaML [19] that simplifies the on-demand generation of diverse ODE-derived synthetic datasets along with ML-based training and inference.

However, the benefits of informing ML with ODE-based simulations depend on the quality and quantity of the generated synthetic data [11]. The quantity, i.e., the size of the training dataset, is known to play a crucial role in successful ML as larger datasets generally improve the prediction performance of ML methods. However, for simulation-based TL, the impact of the synthetic dataset size is less clear as it is an important factor for how strongly the pre-trained ML model is conditioned to the characteristics of the synthetic data. Consequently, the benefit of large training data is related to synthetic data quality. Specifically, it is important how well the synthetic data captures relevant properties of the target, real-world data, i.e., how coherent the latter is with the employed ODE model. For example, Steinacker et al. [38] observed more accurate predictions when the synthetic data show intermediate oscillatory amplitudes coherent with the target data. Further, a potential mechanism for providing more relevant patterns to properly train ML models is learning from a biological system under various conditions. For simulation-based TL, this can be realized by increasing the diversity of the synthetic time series dynamics through variation of the underlying properties of an ODE, such as its kinetic parameters (KPs) and initial conditions (ICs). So far, the impact of this diversity on the subsequent time series forecasting performance is unknown. Moreover, the addition of random noise to training datasets is a known strategy for enhancing the generalization capabilities of neural networks [3]. In a simulation-based TL setting, it so far remains unclear whether and which levels of synthetic noise can improve time series forecasts. Further, all described factors interact with each other. Thus, consistently leveraging the predictive power of simulation-based TL for time series data is bound to a systematic investigation of the required characteristics of synthetic datasets.

In this work, we provide a systematic approach for optimizing the ODE-based synthetic dataset characteristics for time series forecasting in a simulation-based TL setting. For this purpose, we design a framework that allows selecting characteristics of ODE-derived datasets and their multivariate investigation on the basis of their TL performance on the target dataset. We highlight the capabilities of our framework and derive generally applicable conclusions by thorough investigation of three representative and relevant biological systems, including the prediction of COVID-19 infections and predator-prey dynamics.

2. Methods

2.1. Optimizing ODE-derived Synthetic Data for Transfer Learning

We design a framework to optimize the ODE-based synthetic dataset characteristics for simulation-based TL. Our framework consists of two main components. First, we generate synthetic datasets with different characteristics (Fig. 1A-C): we consider the synthetic dataset size, the diversity of the ODE dynamics represented in the synthetic data, and the synthetic noise applied to the data. Second, we evaluate the benefit of each synthetic dataset configuration for a real-world target dataset in a simulation-based TL approach (Fig. 1D-F): we pre-train deep learning (DL) models with synthetic data and employ the pre-trained models to train and predict observed dynamics in the real-world dataset. In addition, we compare the simulation-based TL results to a baseline DL setting, in which DL models are exclusively trained on the real-world dataset without any pre-training. A second baseline is derived from calibrating the ODE system to the real-world dataset. For further optimization diagnostics, we assess the coherence between the employed dynamics of the ODE system and the real-world dataset.

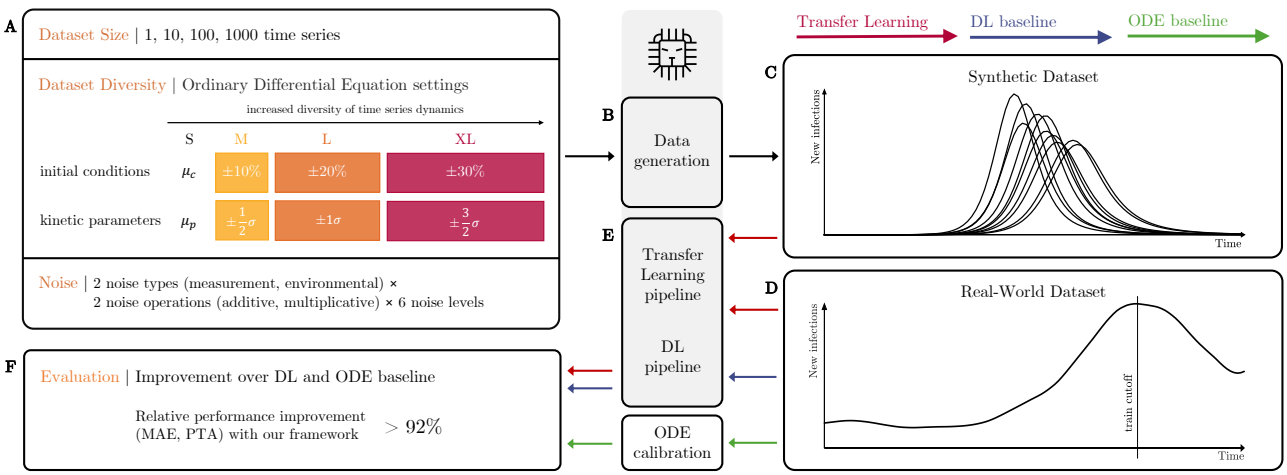

Fig. 1. Overview of our framework. We optimize dataset size, diversity, and noise (A) to obtain suitable synthetic data for pre-training DL models. For an ODE model and a synthetic dataset configuration, we generate a dataset (B) by sampling uniformly from differently sized intervals for initial conditions and kinetic parameters and by adding noise. Thereby, we obtain a dataset of one or more time series following the chosen ODE dynamic (C), which is then used for pre-training models before fine-tuning using (real-world) target data (D, E). We evaluate the performance of transfer learning (TL) with each specific dataset configuration against DL baseline and ODE calibration approaches conditioned only on real-world data (F). We use SimbaML [19] to generate synthetic time series and train DL models.

2.1.1. ODE-based Synthetic Dataset Characteristics

To investigate the impact of synthetic dataset *size* on TL, we consider datasets with 1, 10, 100, and 1000 individual time series.

The *diversity* of a synthetic dataset, i.e., the variety of the dynamics that are represented within it, potentially influences the generalization ability of the DL model pre-trained on it. We alter the diversity of time series dynamics by two properties of the ODE system: (i) the variety in the ICs of the ODE model, i.e., the starting values of the simulated time-series, and (ii) the variety in the KPs of the ODE model that enter the rate equations of the ODE system. To create a variety in these entities (and thus in the dynamics of the generated time series), their values are not considered fixed, but they are uniformly sampled from intervals during data generation. We translate larger synthetic dataset diversity thus to larger ranges of the employed sampling intervals. For our experiments, we consider four different interval ranges with increasing length (S, M, L, and XL) for ICs and KPs (see Supplement Sec. A.1.2).

For the combination of dataset size and diversity that yields the optimal simulation-based TL prediction performance, we further evaluate the impact of *synthetic noise* that is added to the ODE-derived data. Noise plays an especially relevant role in biological time series, as these are commonly subject to pronounced fluctuations. The appropriate representation of these fluctuations for pre-training may aid in further enhancing simulation-based TL results. Our framework supports two different types of synthetic noise, either measurement or environmental, and two different noise operations, either additive or multiplicative. We always allow only a single combination of synthetic noise type and operation at a time. Measurement noise is the error that is expected to occur due to inaccuracies in the measurement process. We model measurement noise by either adding or multiplying independent and identically distributed noise from a user-defined distribution (see Supplement Sec. A.1.3) to each variable at each time point of a synthetic time series. Environmental noise accounts for time-varying fluctuations in the environment, such as in enzyme activities, nutrient sources, and infectiousness of a virus, that directly impact the temporal development of the considered entities. To model the propagation of such fluctuations to subsequent time steps, we implement environmental noise as an additive term or multiplicative factor in the rate equations of the ODE model. For each of the described noise types and operations, we investigate different noise levels, where larger noise levels correspond to larger introduced perturbations (see Supplement Sec. A.1.3).

2.1.2. Time Series Prediction and Evaluation

We apply four broadly used DL model types representing major concepts of DL in time series forecasting [42] in our simulation-based TL approach: long short-term memory neural network (LSTM), gated recurrent unit neural network (GRU), convolutional neural network (CNN), and dense neural network (DNN). Each model is combined with a two-layer multi-perceptron that produces the final output. To investigate how the impact of size and diversity of synthetic datasets compares for different-large DL models, we evaluate a small and large network architecture for each DL model type.

For a specific synthetic dataset, we leverage SimbaML's TL pipeline to pre-train each DL model on the ODE-based synthetic dataset. We then fine-tune the last two layers for five epochs on the real-world target dataset. Note that 10% of the synthetic and real-world time series in the training data are used as a validation set. To avoid overfitting during pre-training, all models are trained until the validation loss on the pre-training dataset converges. We use checkpointing throughout the complete pipeline to save the best DL model state according to validation loss. See Supplement Secs. A.4 and A.6 for details on the DL model architectures and training setup such as dataset-specific train-test splits, respectively.

Table 1. Overview of the representative biological datasets for which we optimize ODE-based synthetic datasets characteristics in a simulation-based TL approach. Related work (rel. work) captures citation counts of the biological dynamics from PubMed (see Supplement Sec. A.2.4). We indicate the ODE used to generate the synthetic datasets, the source of the real-world dataset, the periodic nature of the biological system, and if synthetic and real-world data in our experiments show coherent time series dynamics.

	COVID-19	**Rotifers-Algae**		**Lynx-Hares**
Dynamics	epid. infections	population dynamics		
Rel. work	ca. 380,800	ca. 16,200		
ODE	SIR [18]	SAR [32]		LV [23, 40]
Dataset	[29]	[4]	[4]	[24]
Periodicity	no	yes	yes	yes
Coherence	yes	yes	no	partial

To evaluate the prediction performance achieved by optimizing the ODE-based synthetic dataset characteristics, TL results are compared to two baselines. First, we consider a non-informed, DL baseline setting in which we train the randomly initialized DL models only on the train set of the real-world data. Note that we compare the TL results of each DL model type to the performance of its small architecture in the baseline DL setting. Second, we assess the performance of predictions via simulations of the ODE model after calibrating it, i.e., fitting its KPs, to the combined train and validation set of the real-world dataset (see Supplement Sec. A.7).

The time series predictions of the simulation-based TL, the DL baseline, and the ODE calibration are evaluated based on two metrics: mean absolute error (MAE) [43] and prediction trend accuracy (PTA), a variant of mean directional accuracy [5] (see Supplement Sec. A.5). While MAE considers the error between the predictions and the observed measurement in the real-world dataset, PTA only evaluates the correctness of the predicted trend, a desirable property for time series. We consider 1-PTA such that lower values are better for both metrics.

2.2. Coherence Quantification

To quantify the coherence of time series dynamics between the investigated real-world datasets and our synthetic datasets, we perform an analysis based on multivariate dynamic time warping (DTW) [2]. We sample time series from a synthetic dataset and compute their DTW distance to the corresponding real-world time series. We compare the obtained distribution to a baseline distribution of DTW distances solely between synthetic time series from the synthetic dataset (see Supplement Sec. A.3).

2.3. Biological Data and ODE models

To demonstrate the capabilities of our framework for optimizing the synthetic dataset characteristics for simulation-based TL, we select representative biological time series datasets using the

following criteria: The datasets should (i) be publicly available; (ii) contain periodic and non-periodic time series among them as both are important types of biological dynamics; (iii) have been described to correspond to a specific ODE system; (iv) contain some dynamics fully coherent with the ODE and other dynamics being (at least) partially coherent and incoherent, ideally within the same biological system to minimize confounding factors.

Accordingly, we select four real-world datasets and forecasting problems from three biological systems of two distinct, relevant biological fields (see Tab. 1 for their properties and Supplement Sec. A.2 for details). First, we aim to forecast the number of new infections from the fourth COVID-19 infection wave in Germany [29]. We use synthetic infection waves generated from the susceptible-infected-recovered (SIR) model [18] for pre-training. Second, we predict the predator counts in population dynamics of a predator-prey ecosystem containing rotifers and algae species, and substrate for the algae [4] with synthetic time series generated from the nitrogen-algae-rotifers (SAR) model [32], an extension of the Lotka-Volterra predator-prey model. Third, we aim at forecasting predator counts using real-world predator-prey relations between lynx and snowshoe hares [16] with the original Lotka-Volterra (LV) model [14] that we use for synthetic data generation.

Our coherence computations show coherent dynamics for the COVID-19 dataset and SIR-ODE model. Also, they confirmed findings [4, 14] that the rotifers-algae dynamics and lynx-hares dynamics in the investigated datasets contain predator-prey cycles incoherent with their respective ODE model dynamics (see Supplement Sec. A.2). Thus, we rely on two subsets of the observed rotifer-algae data as separate datasets (rotifer-algae coherent, rotifer-algae incoherent) to contrast dynamics that are coherent and incoherent with synthetic time series and consider the full lynx-hares dataset as partially incoherent.

Note that we use literature-based KPs and ICs for the synthetic data generation with the ODE models. The parameter values, listed in Supplement Sec. A.2, were derived by fitting the ODE models to related datasets [10, 32, 8].

3. Results

We examine ODE-based synthetic dataset characteristics for simulation-based TL with our designed framework for four representative datasets. We first use our framework to evaluate the influence of synthetic dataset size and diversity on simulation-based TL by generating synthetic datasets with different combinations of numbers of time series and sampling intervals for KPs and ICs (Sec. 3.1 to 3.3), without any synthetic noise. We address synthetic noise in a subsequent analysis (Sec. 3.4).

3.1. Optimal Synthetic Dataset Characteristics

Our framework enables the identification of optimal synthetic dataset and DL architecture configurations for simulation-based TL. We outperform all baselines in all four datasets in terms of 1-PTA, and the DL baseline for three out of four datasets and the ODE calibration-based results for two of the datasets in terms of MAE (see Tab. 2 and Supplement Sec. A.8). Thereby, superior performance of the simulation-based TL over a baseline DL approach depends on the coherence between time series dynamics in the real-world and generated datasets. With our optimized synthetic datasets, simulation-based TL shows inferior average performance, by a small margin (5.6% worse MAE), only in settings with pronounced incoherence between the real-world dataset and the synthetic data. In this case, i.e., the incoherent rotifers-algae dataset, our framework recommends pre-training on small synthetic datasets (a single time series), thus approaching the better-performing DL baseline setting. For the COVID-19 dataset, we observe the greatest relative improvements with simulation-based TL over the DL baseline and ODE approach. Notably, any type of pre-training improves the prediction performance of the LSTM and GRU over the baselines for this dataset.

Table 2. Overview of the optimal synthetic dataset characteristics for simulation-based TL that we identified with our framework for each dataset (see Sec. 2.3). We report the number of time series (TS) and diversity of KPs and ICs for the best performing DL model types based on MAE and PTA (see Supplement Sec. A.8.1 for details on the optimization criterion). The 'DL' and 'ODE' columns show the relative difference in MAE and 1-PTA of the best TL setting to the respective DL baseline run (with the same DL architecture as the TL), and the calibrated ODE. See Supplement Tab. S4 and S5 for more details on achieved performances.

Dataset	**Best TL Run**			**Model**	**MAE**		**1-PTA**	
	TS	IC	KP		DL	ODE	DL	ODE
COV-19	1000	XL	XL	LSTM	-92%	-99%	-94%	-97%
R-A (coh.)	100	M	XL	LSTM	-23%	-3.8%	-44%	-47%
R-A (inc.)	1	S	S	LSTM	5.6%	-60%	-25%	-8.3%
L-H	1000	M	S	CNN	-21%	-2.7%	-15%	-52%

3.2. Impact of Synthetic Dataset Size

We can leverage our framework to systematically examine the influence of the dataset size on simulation-based TL performance (see Fig. 2). For coherent time series dynamics between synthetic and real-world datasets (COVID-19, rotifers-algae coherent), the median MAE performance of TL runs can be improved by increasing the size of synthetic datasets. However, pre-training on 100 (and not 1000) time series yields the vast majority of best-performing dataset configurations for the coherent rotifers-algae dataset. This might be due to discrepancies (despite coherence) between synthetic and real-world data. Thus, the DL model is too strongly conditioned on incorrect dynamics within the larger synthetic dataset.

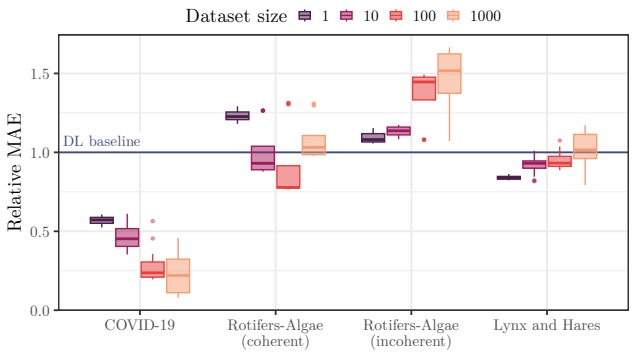

Fig. 2. Univariate analysis of the performance impact of synthetic dataset size on simulation-based TL. The boxplots show the performance of TL over various dataset diversity configurations for each dataset size relative to the DL baseline (blue) for the best performing DL model (see Tab. 2). Larger synthetic datasets improve MAE performance if time series dynamics in synthetic and real-world datasets are coherent (COVID-19, coherent rotifers-algae) and reduce it for incoherent datasets. We obtain similar results for larger DL architectures and PTA (see Supplement Sec. A.9, including ODE baselines that we do not show here due to their different scale).

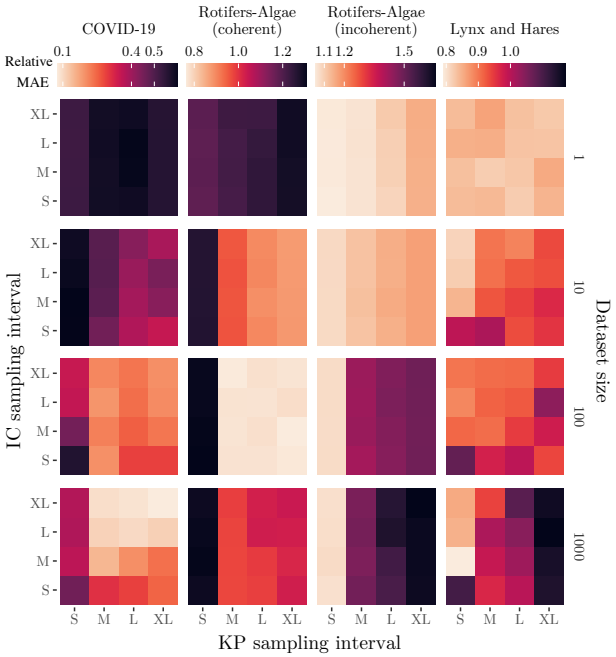

Fig. 3. Multivariate analysis of the impact of dataset size, IC-based and KP-based dataset diversity on simulation-based TL performance. For each dataset, we show the MAE for the best-performing DL model type (see Tab. 2, 'Model' column) relative to its performance in the DL baseline. We observe that the optimal choice of the characteristics of synthetic datasets highly varies between real-world target datasets. This underlines the necessity of optimizing ODE-derived synthetic datasets to fully leverage the predictive power of simulation-based TL. See Supplement Sec. A.10 for the univariate analyses and PTA results.

For the incoherent rotifers-algae and lynx-hares datasets, our framework detects declining (median) performances for increasing dataset size and thus optimizes TL towards the DL baseline setting. Nonetheless, the variance of prediction performances for larger synthetic datasets across different diversity settings and DL architectures is substantially increased. Therefore, the largest dataset size setting (1000 time series) can yield both the best and worst performance results for a dataset, e.g., for the lynx and hares dataset, and optimization, as provided by our framework, is required.

3.3. Impact of Synthetic Dataset Diversity

The driver for the observed increase in variances in prediction performance with larger datasets is revealed in the multivariate analysis of dataset size and diversity (Fig. 3). We find that changing the dataset diversity by generating time series with a differing variety of KP has a particularly strong effect on prediction performance for larger dataset sizes. Univariate analysis shows that the absolute difference in MAE between the best and worst TL performance associated with each KP sampling interval correlates with its length (see Supplement Fig. S8). MAE scores of TL are improved by increased diversity if the system dynamics between synthetic and real-world datasets are coherent (COVID-19 and coherent rotifers-algae). For these datasets, we observe the most substantial improvement in MAE scores when increasing the dataset diversity by changing from the small (S) to medium (M) KP sampling interval. The opposite effect is observed in the incoherent rotifers-algae and lynx-hares datasets, where simulation-based TL performance decreases with increased dataset diversity, especially for larger synthetic dataset sizes. Due to the partial and non-coherence between synthetic and observed time series dynamics in these cases, the increased dataset diversity supports the conditioning of DL models on system behavior that is not contained in the target dataset. Both trends are also observed with the large DL model architectures and model types across all datasets (see Supplement Fig. S9 and Tab. S9).

Compared to KPs, manipulating dataset diversity by altering the IC sampling intervals shows decisively less influence on simulation-based TL prediction performance (see Supplement Fig. S10). ICs often result in similar time series dynamics, thus exerting a smaller influence on the overall diversity of the dataset. Nonetheless, in some cases, such as the COVID-19 and lynx-hares dataset, identifying a suitable IC sampling interval is still required to fully optimize the prediction performance in a simulation-based TL setting (see Fig. 3). For an investigation of how the synthetic dataset's diversity impacts the PTA, see Supplement Sec. A.10.

3.4. Impact of Synthetic Noise in Synthetic Datasets

Further, we assess the impact of measurement and environmental synthetic noise on simulation-based TL performance, for each real-world dataset. We use our framework to optimize the synthetic noise settings to improve the simulation-based TL results previously optimized for synthetic dataset size, diversity, and DL architecture (see Tab. 2). To make perturbations introduced by noise comparable between datasets, we focus exemplarily on multiplicative noise here, i.e., noise is introduced as a factor drawn from a lognormal distribution around 1. We perform pre-training on synthetic datasets including either synthetic measurement noise or synthetic environmental noise. For each noise type, we consider six different noise levels corresponding to the degree of influence of noise on the data (see Supplement Sec. A.1.3 and Sec. A.12, also for additive noise results).

Our framework optimizes synthetic noise levels for simulation-based TL towards small values for the datasets with coherent dynamics in synthetic and real-world target data (i.e., the COVID-19 and coherent rotifers-algae datasets, see Fig. 4).

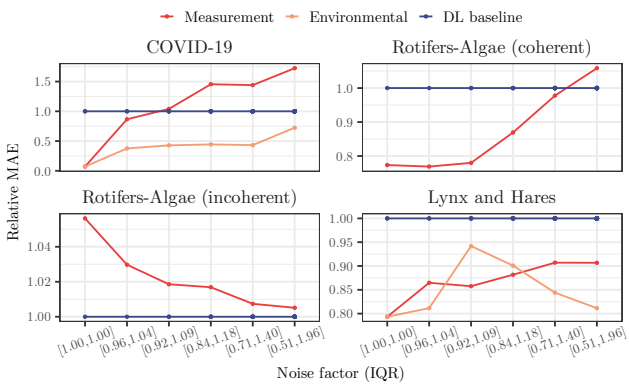

Fig. 4. Univariate analysis of the impact of multiplicative synthetic noise on simulation-based TL performance. Lines show the mean MAE of simulation-based TL performance relative to the DL baseline (blue) for each noise level indicated by the interquartile range (IQR) of the lognormal distribution used for sampling the noise factor. Environmental noise was not used for the rotifers-algae datasets due to numerical instabilities making simulations computationally infeasible. ODE baselines are omitted here due to their different scale, see Supplement Sec. A.12. Low noise levels are optimal for real-world datasets coherent with synthetic data, while larger noise levels can be beneficial for simulation-based TL in incoherent settings.

In contrast, for the incoherent datasets, larger synthetic noise levels can increase performance up to baseline DL performance levels (for the rotifers-algae incoherent dataset) or can yield close to optimal TL performances (for the environmental noise and the lynx-hare dataset). For the lynx and hares dataset, increasing noise levels of multiplicative measurement noise reduces the prediction performance.

The observation of optimal simulation-based TL performance for low synthetic noise levels for the COVID-19 dataset is in line with additional insights that we gained from synthetic-only experiments with our framework (see Supplement Sec. A.13). In these, we find that, in coherent settings, synthetic noise can especially improve performance when its levels are similar to those in the target dataset. Thus, the low noise levels in the COVID-19 dataset that arise due to the smoothing preprocessing steps (see Supplement Sec. A.2.1) are well matched with low synthetic noise levels to achieve optimal simulation-based TL performance.

4. Discussion and Conclusion

We introduce a framework for optimizing ODE-generated datasets for simulation-based TL and apply it to a representative selection of relevant biological time series datasets. We identify the coherence in time series dynamics between synthetic and real-world datasets as a dominant factor that can determine the benefit of simulation-based TL. Pre-training on larger or more diverse synthetic datasets, i.e., using longer IC and KP sampling intervals, together with low synthetic noise levels, can improve the performance of simulation-based TL for coherent settings. Deteriorating effects of dataset diversity and size are observed when incoherence prevails, and increased synthetic noise levels can become beneficial. Overall, the prediction performance in a simulation-based TL setting can vary substantially depending on the chosen ODE-based synthetic dataset characteristics. Our framework enables the optimization of these characteristics such that simulation-based TL regularly outperforms baseline DL and ODE-calibration-based predictions. If time series dynamics in synthetic and real-world datasets are incoherent and baseline performance is not reached, our framework naturally moves towards the optimal non-TL setting.

Thereby, the greatest relative performance improvement is achieved for the COVID-19 dataset. This is likely due to the small data setting where only parts of the non-periodic COVID-19 infection wave are available. As a consequence, baseline DL models are not able to infer non-observed dynamics and ODE calibration overfits to initial, non-representative system behavior. With simulation-based TL, the DL model is pre-trained on expected but yet unobserved system behavior, e.g., a decline in infection numbers after the peak of the wave, which allows all TL runs to outperform the DL baseline, to a varying degree. This substantial advantage over the DL baseline setting is less present in our experiments with the periodic lynx-hares and rotifers-algae data where complete oscillation cycles are already contained in the train set of the real-world data (see Supplement Sec. A.2). Therefore, we conclude that using our framework to optimize simulation-based TL seems particularly beneficial in small data scenarios when dynamic characteristics are yet incompletely observed.

The phenomenon that increasing the diversity of datasets can improve DL model performance, and, in particular, its generalization abilities, has already been observed in other fields, such as computer vision [36]. Synthetic dataset diversity in simulation-based TL as we employ it here is not only driven by the length of IC and KP sampling intervals that, in principle, tune the potential space of different time series dynamics and shapes that is sampled from, but it also depends on the dataset size: The effective variance of sampled values only increases by sampling from broader intervals if the sample size is sufficiently large to represent the interval. Similarly, interrelations may exist between synthetic dataset diversity and its coherence to the real-world data: The more diverse a synthetic dataset, the less tailored it might be to the time series dynamics of interest, thus potentially negatively impacting simulation-based TL. Moreover, to fully leverage synthetic dataset diversity, the employed DL models in simulation-based TL have to be sufficiently large to capture (and reproduce) the distinct presented patterns during pre-training. Thus, the DL model sizes have to be taken carefully into account. Using our framework, we find consistent results also when increasing to larger DL architectures (see Supplement Sec. A.9, Sec. A.10).

Overall, we observe that the diversity introduced by the variety of ICs has a smaller influence on TL performance compared to those of the KPs, which reflects the overall lower influence of IC than KP on an ODE system's dynamics. However, ICs can have decisively more impact in more complex systems, e.g., containing conserved moieties or multiple steady states [39], and thus may still need to be considered carefully.

We also observe dependencies between dataset coherence and the optimal synthetic noise levels: If the synthetic time series captures the real-world characteristics well, large synthetic noise levels may obstruct learning informative patterns. In contrast, if synthetic dynamics are incoherent with the target dataset, increasing noise dilutes the (unsuitable) patterns in the provided dynamics, thus approaching an uninformed initialization with pre-training. Detecting these cases, our framework reverts to DL baseline performance with optimized dataset characteristics in incoherent settings.

Our investigations shown here are limited to specified configurations for dataset size, diversity, and noise. Depending on the system of interest, they could be extended to examine broader or more fine-grained selections of values. Furthermore, it could be of interest to study alternative ODEs for the selected datasets. In particular, for the noise types, more complex settings could be more appropriate to represent real-world errors, e.g., combinations of measurement and environmental noise and different noise levels for different parts of the system. Similarly, other and more complex DL model architectures, such as GNNs [27, 44] may yield further benefits in a simulation-based TL setting. All these different variations could be readily investigated with the help of our presented framework.

Of note, a successful application of our framework for simulation-based TL highly depends on the availability of an ODE model that provides suitable dynamics for a real-world dataset. Even though it requires additional work to develop suitable ODE models or adapt published ones [22], we found that optimized simulation-based TL with such models performed better than non-informed predictions.

Our presented framework could also help to benchmark further datasets with ODE systems [15], and be extended to optimally leverage existing domain knowledge of more complex simulations, such as the geographical spread of infections [7] or climate models [31]. In these complex settings, it could be beneficial to combine simulation-based TL with other informed ML approaches, such as customized loss functions to enforce modeled constraints more strongly [17], or investigate further alternatives to parameter-based TL [42].

Competing interests

No competing interest is declared.

Author contributions

JZ, SW, KB with the help of PI, MK, BYR conceived the study design. JZ performed the experiments and analyzed the data with the help of SW who also contributed the ODE calibration and coherence computation (together with PI) and MK (contributing the noise experiments). JZ, SW, PI, BYR, KB interpreted the data. SW, KB, BYR supervised the work. JZ drafted the manuscript with the help of MK, KB; JZ, SW, PI, BYR, KB substantially revised the draft. All authors agree to the final version.

References

1. Benzekry, S. Artificial intelligence and mechanistic modeling for clinical decision making in oncology. *Clin. Pharmacol. Ther.*, 108(3):471–486, 2020.
2. Berndt, D.J. et al. Using dynamic time warping to find patterns in time series. In *Proc. 3rd KDD*, AAAIWS'94, pages 359–370. AAAI Press, 1994.
3. Bishop, C.M. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
4. Blasius, B. et al. Long-term cyclic persistence in an experimental predator–prey system. *Nature*, 577(7789): 226–230, 2019.
5. Blaskowitz, O. et al. On economic evaluation of directional forecasts. *Int. J. Forecast.*, 27(4):1058–1065, 2011.
6. Box, G.E.P. et al. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015.
7. Brockmann, D. et al. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342(6164): 1337–1342, December 2013.
8. Carpenter, B. Predator-Prey population dynamics: the Lotka-Volterra model in stan. `https://mc-stan.org/users/documentation/case-studies/lotka-volterra-predator-prey.html`, January 2018. Accessed: 2023-7-5.
9. Costello, Z. et al. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *NPJ Syst Biol Appl*, 4:19, 2018.
10. Dehning, J. et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*, 369(6500), July 2020.
11. Deist, T.M. et al. Simulation-assisted machine learning. *Bioinformatics*, 35(20):4072–4080, 2019.
12. Gao, J. et al. STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. *J. Am. Med. Inform. Assoc.*, 28(4):733–743, 2021.
13. Gautam, Y. Transfer learning for COVID-19 cases and deaths forecast using LSTM network. *ISA Trans.*, 124: 41–56, 2022.
14. Gilpin, M.E. Do hares eat lynx? *Am. Nat.*, 107(957): 727–730, 1973.
15. Hass, H. et al. Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics*, 35(17):3073–3082, 2019.
16. Hewitt, C.G. *The conservation of the wild life of Canada*. New York: C. Scribner, 1921.
17. Karniadakis, G.E. et al. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.
18. Kermack, W.O. et al. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A Math. Phys. Sci.*, 115(772):700–721, 1927.
19. Kleissl, M. et al. SimbaML: Connecting mechanistic models and machine learning with augmented data. *ICLR Tiny Paper*, 2023.
20. Lagergren, J.H. et al. Biologically-informed neural networks guide mechanistic modeling from sparse experimental data. *PLoS Comput. Biol.*, 16(12):e1008462, December 2020.
21. Lancaster, M.C. et al. Improved prediction of drug-induced torsades de pointes through simulations of dynamics and machine learning algorithms. *Clin. Pharmacol. Ther.*, 100 (4):371–379, 2016.
22. Le Novère, N. et al. BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.*, 34(Database issue):D689–91, January 2006.
23. Lotka, A.J. Analytical note on certain rhythmic relations in organic systems. *Proc. Natl. Acad. Sci. U. S. A.*, 6(7): 410–415, 1920.
24. MacLulich, D.A. *Fluctuations in the numbers of the varying hare (Lepus americanus)*. University of Toronto Press, Toronto, 1937.
25. Martin, L. et al. How much do clinical trials cost? *Nat. Rev. Drug Discov.*, 16(6):381–382, 2017.
26. Pan, S.J. et al. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010.
27. Panagopoulos, G. et al. Transfer graph neural networks for pandemic forecasting. *AAAI*, 35(6):4838–4845, May 2021.
28. Procopio, A. et al. Combined mechanistic modeling and machine-learning approaches in systems biology - a systematic literature review. *Comput. Methods Programs Biomed.*, 240:107681, 2023.
29. Robert Koch-Institut. SARS-CoV-2 Infektionen in Deutschland, 2022.
30. Rodríguez, A. et al. EINNs: Epidemiologically-Informed neural networks. *AAAI*, 37(12):14453–14460, 2023.
31. Rolnick, D. et al. Tackling climate change with machine learning. *ACM Comput. Surv.*, 55(2):1–96, 2022.
32. Rosenbaum, B. et al. Estimating parameters from multiple time series of population dynamics using bayesian inference. *Frontiers in Ecology and Evolution*, 6, 2019.
33. Roster, K. et al. Forecasting new diseases in low-data settings using transfer learning. *Chaos Solitons Fractals*, 161:112306, 2022.
34. Sapoval, N. et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun.*, 13(1):1728, 2022.
35. Sharifi-Noghabi, H. et al. AITL: Adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics. *Bioinformatics*, 36:i380–i388, 2020.
36. Shorten, C. et al. A survey on image data augmentation for deep learning. *J. Big Data*, 6(1), 2019.
37. Siami-Namini, S. et al. A comparison of ARIMA and LSTM in forecasting time series. In *17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, pages 1394–1401. IEEE, 2018.
38. Steinacker, M. et al. Individual modelling of haematotoxicity with NARX neural networks: A knowledge transfer approach. *Heliyon*, 9(7):e17890, 2023.
39. Tyson, J.J. et al. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell Biol.*, 15(2):221–231, 2003.
40. Volterra, V. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118:558–560, 1926.

41. Von Rueden, L. et al. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE TKDE*, 35(1):614–633, 2021.
42. Weber, M. et al. Transfer learning with time series data: A systematic mapping study. *IEEE Access*, 9:165409–165432, 2021.
43. Willmott, C.J. et al. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.*, 2005.
44. Witzke, S. et al. Mobility data improve forecasting of COVID-19 incidence trends using graph neural networks. In *epiDAMIK 6.0 at KDD*, 2023.