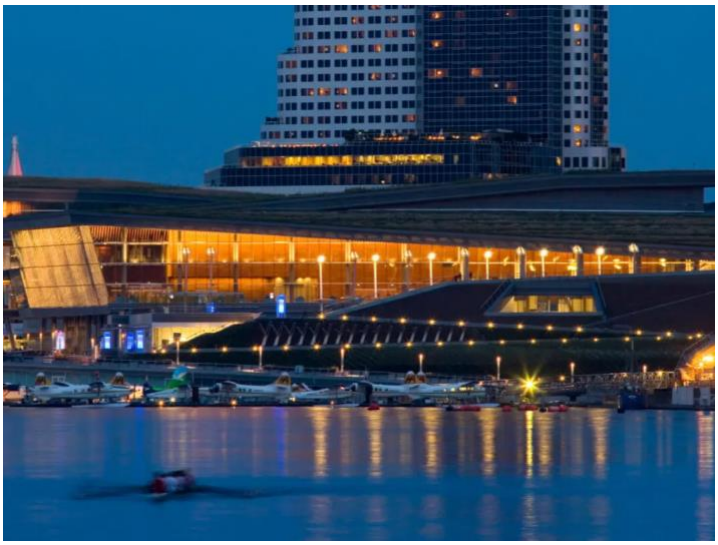
CCAIM



NEURAL INFORMATION PROCESSING SYSTEMS

CCAIM at NeurIPS 2024

Researchers from the Cambridge Centre for AI in Medicine are presenting a record-breaking 17 papers at the thirty-eighth Conference on Neural Information Processing Systems at the Vancouver Convention Center, 10 - 15 December 2024.

Members of CCAIM are represented in *italics*.

CCAIM at NeurIPS 2024

CCAIM Faculty members will present 17 papers when the thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2024) takes place from 10 – 15 December, one of the most globally renowned for presenting and publishing cutting-edge research and the conference includes invited talks, demonstrations, symposia, and oral and poster presentations of refereed papers. A professional exposition focusing on machine learning in practice, a series of tutorials, and topical workshops that provide a less formal setting for the exchange of ideas takes place alongside the conference.

This year's number of 17 papers is well is yet another great showing of the Centre, being consistently strong since its inception. Of these papers, the van der Schaar Lab's paper on Digital Twins has been selected as spotlight – a fantastic representation of the Lab's increasing focus on Digital Twins.

Presenting a variety of topics in its publications, the Cambridge Centre for AI in Medicine (CCAIM) is at the forefront of research in machine learning and AI. The papers – from Prof Mihaela van der Schaar, Prof Adrian Weller, Prof José Miguel Hernández-Lobato, and Prof Pietro Liò – cover a number of the Centre's key research areas, including Digital Twins, interpretable and reliable AI, discovery, reality-centric AI, data-centric AI, Large Language Models, deep learning, neural networks, generative models, and PDEs.

We are especially proud of Tennison Liu, PhD student at CCAIM, who is co-authoring three papers with Prof van der Schaar on LLMs and Digital Twins.

Together, our researchers continuously explore and create revolutionary technologies that will shape the future of healthcare, change clinical work, and improve patient outcomes.

Details for all 17 papers are given below:

DEFT: Efficient Fine-tuning of Diffusion Models by Learning the Generalised *h*-transform

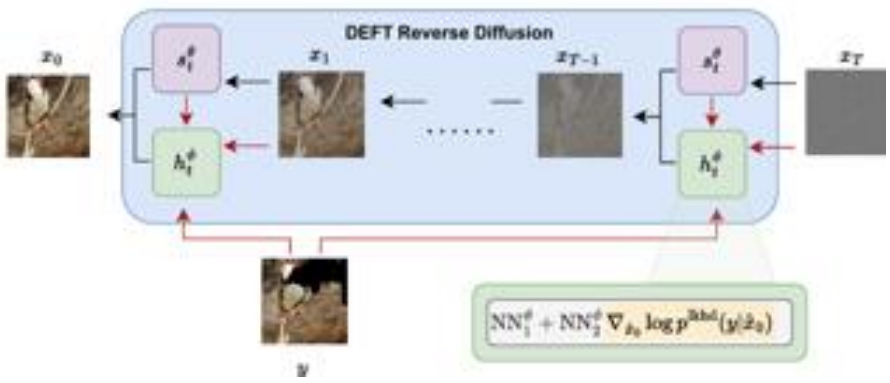
Alexander Denker, Francisco Vargas, Shreyas Padhy, Kieran Didi, Simon Mathis, Riccardo Barbano, Vincent Dutordoir, Emile Mathieu, Urszula Julia Komorowska, *Pietro Liò*

Find the paper here

.

Abstract

Higher-order relations are widespread in nature, with numerous phenomena involving complex interactions that extend beyond simple pairwise connections. As a result, advancements in higher-order processing can accelerate the growth of various fields requiring structured data. Current approaches typically represent these interactions using hypergraphs. We enhance this representation by introducing cellular sheaves for hypergraphs, a mathematical construction that adds extra structure to the conventional hypergraph while maintaining their local, higher-order connectivity. Drawing inspiration from existing Laplacians in the literature, we develop two unique formulations of sheaf hypergraph Laplacians: linear and non-linear. Our theoretical analysis demonstrates that incorporating sheaves into the hypergraph Laplacian provides a more expressive inductive bias than standard hypergraph diffusion, creating a powerful instrument for effectively modelling complex data structures. We employ these sheaf hypergraph Laplacians to design two categories of models: Sheaf Hypergraph Neural Networks and Sheaf Hypergraph Convolutional Networks. These models generalize classical Hypergraph Networks often found in the literature. Through extensive experimentation, we show that this generalization significantly improves performance, achieving top results on multiple benchmark datasets for hypergraph node classification.



Deep Equilibrium Algorithmic Reasoning

Dobrik Georgiev, Joseph Wilson, Davide Buffelli, *Pietro Liò*

Find the paper here

GitHub Page

Abstract

Neural Algorithmic Reasoning (NAR) research has demonstrated that graph neural networks (GNNs) could learn to execute classical algorithms. However, most previous approaches have always used a recurrent architecture, where each iteration of the GNN matches an iteration of the algorithm. In this paper we study neurally solving algorithms from a different perspective: since the algorithm's solution is often an equilibrium, it is possible to find the solution directly by solving an equilibrium equation. Our approach requires no information on the ground-truth number of steps of the algorithm, both during train and test time. Furthermore, the proposed method improves the performance of GNNs on executing algorithms and is a step towards speeding up existing NAR models. Our empirical evidence, leveraging algorithms from the CLRS-30 benchmark, validates that one can train a network to solve algorithmic problems by directly finding the equilibrium. We discuss the practical implementation of such models and propose regularisations to improve the performance of these equilibrium reasoners.

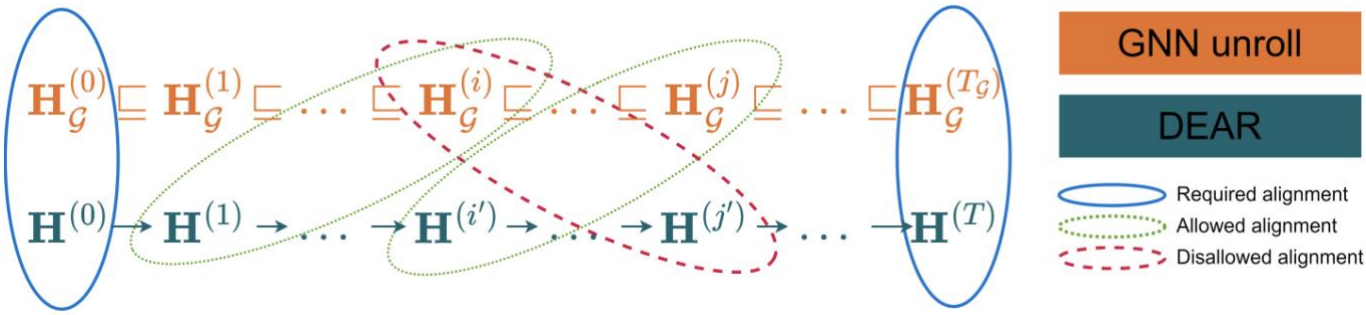


Figure 1: Proposed alignment rule: every state in the DEAR trajectory should "go forward". Alignments to a GNN state that has already been "passed" are disallowed. First and last states must always align. We intentionally use arrows instead of $\sqsubseteq$ for DEAR, as $\sqsubseteq$ may not hold for DEAR's trajectory Interpretable Graph Networks Formulate Universal Algebra Conjectures

Improving Linear System Solvers for Hyperparameter Optimisation in Iterative Gaussian Processes

Jihao Andreas Lin, Shreyas Padhy, Bruno Mlodozeniec, Javier Antorán,
José Miguel Hernández-Lobato

Find the paper here

Abstract

Scaling hyperparameter optimisation to very large datasets remains an open problem in the Gaussian process community. This paper focuses on iterative methods, which use linear system solvers, like conjugate gradients, alternating projections or stochastic gradient descent, to construct an estimate of the marginal likelihood gradient. We discuss three key improvements which are applicable across solvers: (i) a pathwise gradient estimator, which reduces the required number of solver iterations and amortises the computational cost of making predictions, (ii) warm starting linear system solvers with the solution from the previous step, which leads to faster solver convergence at the cost of negligible bias, (iii) early stopping linear system solvers after a limited computational budget, which synergises with warm starting, allowing solver progress to accumulate over multiple marginal likelihood steps. These techniques provide speed-ups of up to 72× when solving to tolerance, and decrease the average residual norm by up to 7× when stopping early.

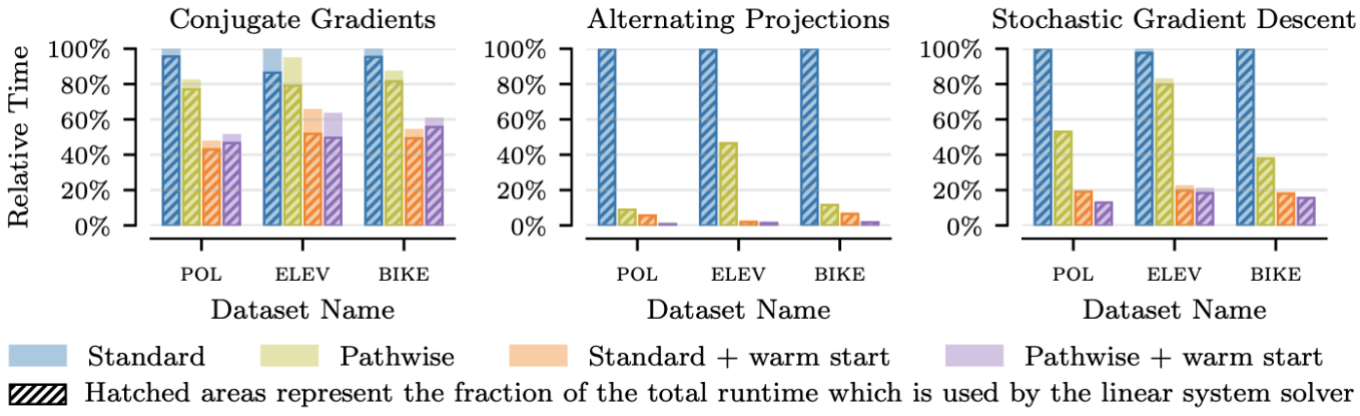


Figure 1: Comparison of relative runtimes for different methods, linear system solvers, and datasets. The linear system solver (hatched areas) dominates the total training time (coloured patches). The pathwise gradient estimator requires less time than the standard estimator. Initialising at the previous solution (warm start) further reduces the runtime of the linear system solver for both estimators.

On conditional diffusion models for PDE simulations

Aliaksandra Shysheya, Cristiana Diaconu, Federico Bergamin, Paris Perdikaris, *José Miguel Hernández-Lobato*, Richard E. Turner, Emile Mathieu

Find the paper here

Abstract

Modelling partial differential equations (PDEs) is of crucial importance in science and engineering, and it includes tasks ranging from forecasting to inverse problems, such as data assimilation. However, most previous numerical and machine learning approaches that target forecasting cannot be applied out-of-the-box for data assimilation. Recently, diffusion models have emerged as a powerful tool for conditional generation, being able to flexibly incorporate observations without retraining. In this work, we perform a comparative study of score-based diffusion models for forecasting and assimilation of sparse observations. In particular, we focus on diffusion models that are either trained in a conditional manner, or conditioned after unconditional training. We address the shortcomings of existing models by proposing 1) an autoregressive sampling approach that significantly improves performance in forecasting, 2) a new training strategy for conditional score-based models that achieves stable performance over a range of history lengths, and 3) a hybrid model which employs flexible pre-training conditioning on initial conditions and flexible post-training conditioning to handle data assimilation. We empirically show that these modifications are crucial for successfully tackling the combination of forecasting and data assimilation, a task commonly encountered in real-world scenarios.

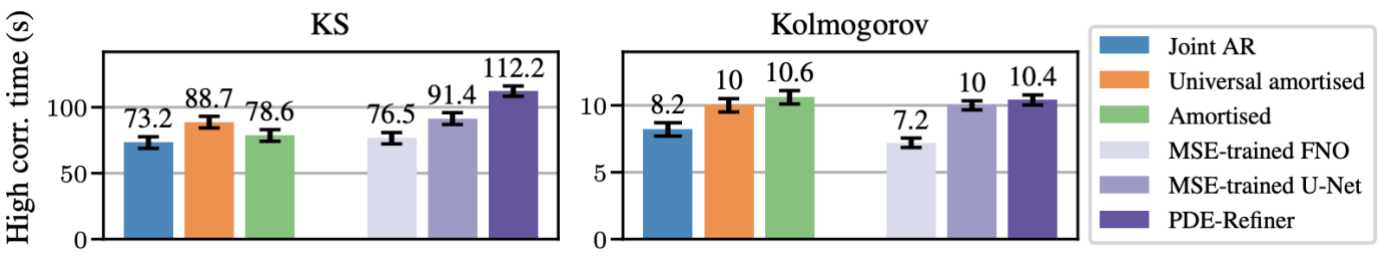


Figure 2: The best joint and amortised models compared against standard ML-based benchmarks for forecasting on the KS (left) and Kolmogorov (right) datasets. High correlation time (↑) is reported, showing the mean ± 3 standard errors, as computed on the test trajectories. We show in Tab. 5 the configurations for the best models.

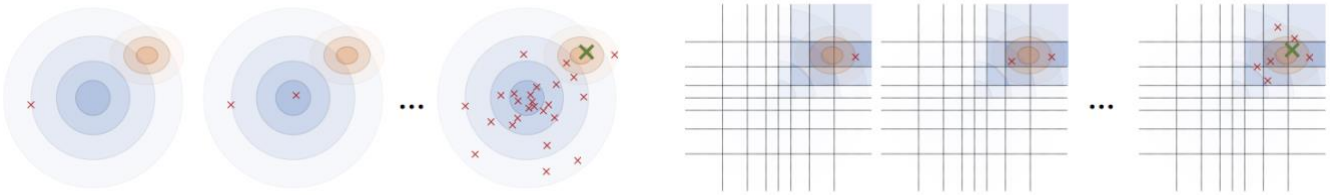
Accelerating Relative Entropy Coding with Space Partitioning

Jiajun He, Gergely Flamich, *José Miguel Hernández-Lobato*

Find the paper here

Abstract

Relative entropy coding (REC) algorithms encode a random sample following a target distribution Q, using a coding distribution P shared between the sender and receiver. Sadly, general REC algorithms suffer from prohibitive encoding times, at least on the order of 2DKL[Q||P], and faster algorithms are limited to very specific settings. This work addresses this issue by introducing a REC scheme utilizing space partitioning to reduce runtime in practical scenarios. We provide theoretical analyses of our method and demonstrate its effectiveness with both toy examples and practical applications. Notably, our method successfully handles REC tasks with DKL[Q||P] about three times greater than what previous methods can manage, and reduces the bitrate by approximately 5-15% in VAE-based lossless compression on MNIST and INR-based lossy compression on CIFAR-10, compared to previous methods, significantly improving the practicality of REC for neural compression.



(a) Procedure of standard REC algorithm. (b) Procedure of REC with space partitioning.

Figure 1: An illustrative comparison between the standard REC algorithm and REC with space partitioning. We illustrate the prior P's density in blue and Q's density in orange. (a) In a standard REC algorithm, we may draw numerous samples (colored in red) before identifying one that aligns well with Q (colored in green). The majority of these samples do not directly contribute to the desired result. (b) In the method we propose, we first divide the search space into smaller grids and then reweight each grid. This amounts to adjusting the prior P to a search heuristic P_1, which can align better with Q. The samples from P_1 will thus be more relevant to Q, potentially reducing the runtime.

A Generative Model of Symmetry Transformations

James Urquhart Allingham, Bruno Kacper Mlodozeniec, Shreyas Padhy, Javier Antorán, David Krueger, Richard E. Turner, Eric Nalisnick, *José Miguel Hernández-Lobato*

Find the paper here

Abstract

Correctly capturing the symmetry transformations of data can lead to efficient models with strong generalization capabilities, though methods incorporating symmetries often require prior knowledge. While recent advancements have been made in learning those symmetries directly from the dataset, most of this work has focused on the discriminative setting. In this paper, we take inspiration from group theoretic ideas to construct a generative model that explicitly aims to capture the data's approximate symmetries. This results in a model that, given a prespecified broad set of possible symmetries, learns to what extent, if at all, those symmetries are actually present. Our model can be seen as a generative process for data augmentation. We provide a simple algorithm for learning our generative model and empirically demonstrate its ability to capture symmetries under affine and color transformations, in an interpretable way. Combining our symmetry model with standard generative models results in higher marginal test-log-likelihoods and improved data efficiency.

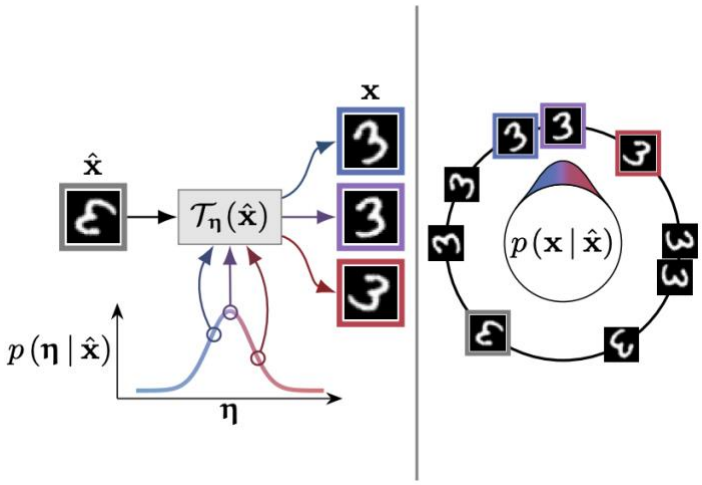


Figure 1: **Left:** An example of a symmetry-aware generative process that we aim to model in this paper. A *prototype* $\hat{\mathbf{x}}$ ( ) is transformed by $\mathcal{T}_\eta$ into an observation $\mathbf{x}$ (, ,). The transformation—e.g., rotation—is parameterized by η—e.g., an angle. **Right:** The corresponding orbit—i.e., the set of all possible instances of $\mathbf{x}$ that can result from applying $\mathcal{T}_\eta$—with a few elements shown. Under this generative process, the prototype is an arbitrary orbit element. Each element in the orbit has a probability $p(\mathbf{x} \mid \hat{\mathbf{x}})$ induced by $p(\eta \mid \hat{\mathbf{x}})$. E.g., for handwritten '3's, we expect digits in an upright orientation with some rotation around, say $\pm 40°$, corresponding to natural variations in handwriting.

Large Language Models Must Be Taught to Know What They Don't Know

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, *Adrian Weller*, Samuel Dooley, Micah Goldblum, Andrew Gordon Wilson

Find the paper here

Abstract

When using large language models (LLMs) in high-stakes applications, we need to know when we can trust their predictions. Some works argue that prompting high-performance LLMs is sufficient to produce calibrated uncertainties, while others introduce sampling methods that can be prohibitively expensive. In this work, we first argue that prompting on its own is insufficient to achieve good calibration and then show that fine-tuning on a small dataset of correct and incorrect answers can create an uncertainty estimate with good generalization and small computational overhead. We show that a thousand graded examples are sufficient to outperform baseline methods and that training through the features of a model is necessary for good performance and tractable for large open-source models when using LoRA. We also investigate the mechanisms that enable reliable LLM uncertainty estimation, finding that many models can be used as general-purpose uncertainty estimators, applicable not just to their own uncertainties but also the uncertainty of other models. Lastly, we show that uncertainty estimates inform human use of LLMs in human-AI collaborative settings through a user study.

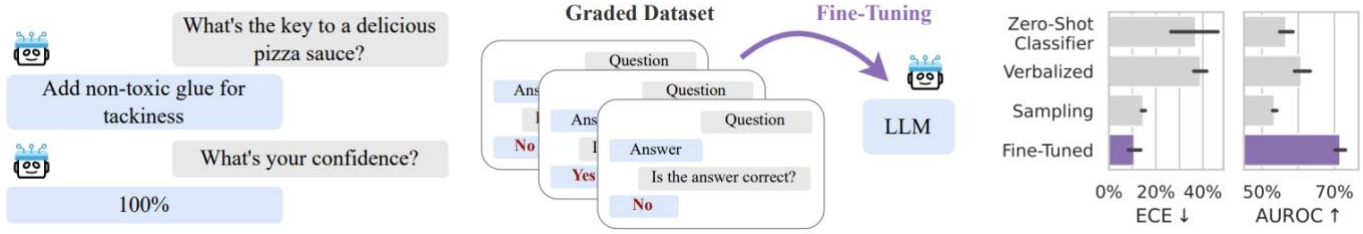


Figure 1: Large language models struggle to assign reliable confidence estimates to their generations. We study the properties of uncertainty calibration in language models, and propose fine-tuning for better uncertainty estimates using a graded dataset of generations from the model. We evaluate our methods on a new open-ended variant of MMLU. We show that fine-tuning improves expected calibration error (ECE) and area under the receiver operating characteristic curve (AUROC) compared to commonly-used baselines. Error bars show standard deviation over three base models (LLaMA-2 13/7B and Mistral 7B) and their chat variants.

Approximately Equivariant Neural Processes

Matthew Ashman, Cristiana Diaconu, *Adrian Weller*, Wessel Bruinsma, Richard E. Turner

Find the paper here

Abstract

Equivariant deep learning architectures exploit symmetries in learning problems to improve the sample efficiency of neural-network-based models and their ability to generalise. However, when modelling real-world data, learning problems are often not exactly equivariant, but only approximately. For example, when estimating the global temperature field from weather station observations, local topographical features like mountains break translation equivariance. In these scenarios, it is desirable to construct architectures that can flexibly depart from exact equivariance in a data-driven way. Current approaches to achieving this cannot usually be applied out-of-the-box to any architecture and symmetry group. In this paper, we develop a general approach to achieving this using existing equivariant architectures. Our approach is agnostic to both the choice of symmetry group and model architecture, making it widely applicable. We consider the use of approximately equivariant architectures in neural processes (NPs), a popular family of meta-learning models. We demonstrate the effectiveness of our approach on a number of synthetic and real-world regression experiments, showing that approximately equivariant NP models can outperform both their non-equivariant and strictly equivariant counterparts.

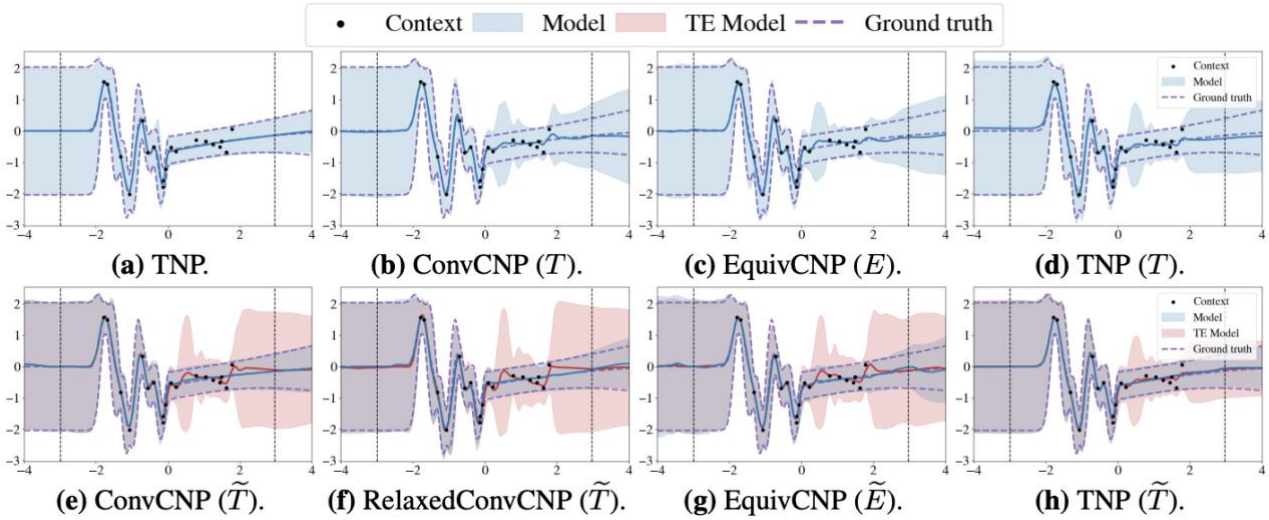


Figure 1: A comparison between the predictive distributions on a single synthetic 1-D regression dataset of the TNP-, ConvCNP-, and EquivCNP-based models. For the approximately equivariant models, we plot both the model's predictive distribution (blue), as well as the predictive distributions obtained without using the fixed inputs (red). The dotted black lines indicate the target range.

Automatically Learning Hybrid Models for Digital Twins of Dynamical Systems

Samuel Holt*, *Tennison Liu**, *Mihaela van der Schaar*

Spotlight Paper

Find the paper here

Abstract

Digital Twins (DTs) are computational models that simulate the states and temporal dynamics of real-world systems, playing a crucial role in prediction, understanding, and decision-making across diverse domains. However, existing approaches to DTs often struggle to generalize to unseen conditions in data-scarce settings, a crucial requirement for such models. To address these limitations, our work begins by establishing the essential desiderata for effective DTs. Hybrid Digital Twins (**HDTwins**) represent a promising approach to address these requirements, modeling systems using a composition of both mechanistic and neural components. This hybrid architecture simultaneously leverages (partial) domain knowledge and neural network expressiveness to enhance generalization, with its modular design facilitating improved evolvability. While existing hybrid models rely on expert-specified architectures with only parameters optimized on data, automatically specifying and optimizing HDTwins remains intractable due to the complex search space and the need for flexible integration of domain priors. To overcome this complexity, we propose an evolutionary algorithm (**HDTwinGen**) that employs Large Language Models (LLMs) to autonomously propose, evaluate, and optimize HDTwins. Specifically, LLMs iteratively generate novel model specifications, while offline tools are employed to optimize emitted parameters. Correspondingly, proposed models are evaluated and evolved based on targeted feedback, enabling the discovery of increasingly effective hybrid models. Our empirical results reveal that HDTwinGen produces generalizable, sample-efficient, and evolvable models, significantly advancing DTs' efficacy in real-world applications.

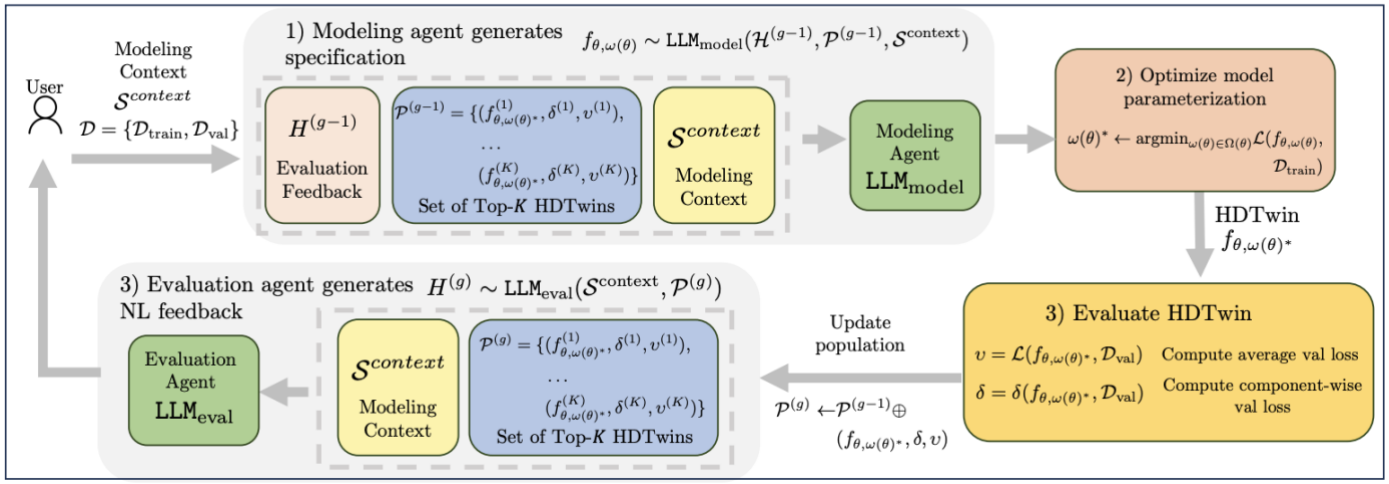


Figure 1: HDTwinGen: evolutionary framework. The process begins with user-provided modeling context S context and D = {Dtrain, Dval}. 1) In iteration g, the modeling agent generates model specification as a Python program fθ,ω(θ) . 2) Parameters are optimized using the offline optimization tool to yield fθ,ω∗(θ) . 3) The HDTwin is evaluated based on model loss υ and component-wise loss δ. Subsequently, the model pool P (g) is updated with top-K models. 4) The evaluation agent provides targeted feedback for model improvement H(g) by analyzing models in P (g) using performance metrics requirements outlined in S context. This iterative loop repeats for G iterations.

Self-Healing Machine Learning: A Framework for Autonomous Adaptation in Real-World Environments

Paulius Rauba, Nabeel Seedat, Krzysztof Kacprzyk, *Mihaela van der Schaar*

Find the paper here

Abstract

Real-world machine learning systems often encounter model performance degradation due to distributional shifts in the underlying data generating process (DGP). Existing approaches to addressing shifts, such as concept drift adaptation, are limited by their reason-agnostic nature. By choosing from a pre-defined set of actions, such methods implicitly assume that the causes of model degradation are irrelevant to what actions should be taken, limiting their ability to select appropriate adaptations. In this paper, we propose an alternative paradigm to overcome these limitations, called self-healing machine learning (SHML). Contrary to previous approaches, SHML autonomously diagnoses the reason for degradation and proposes diagnosis-based corrective actions. We formalize SHML as an optimization problem over a space of adaptation actions to minimize the expected risk under the shifted DGP. We introduce a theoretical framework for self-healing systems and build an agentic self-healing solution H-LLM which uses large language models to perform self-diagnosis by reasoning about the structure underlying the DGP, and self-adaptation by proposing and evaluating corrective actions. Empirically, we analyze different components of H-LLM to understand why and when it works, demonstrating the potential of self-healing ML.

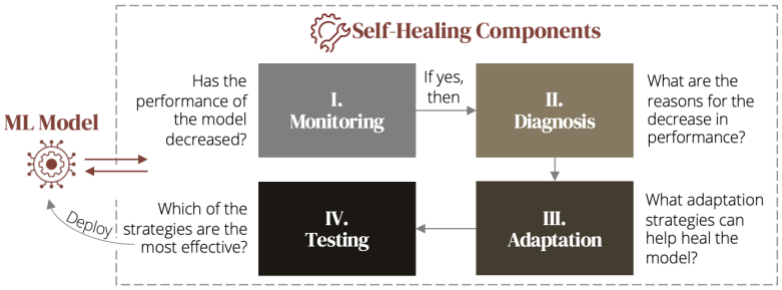


Figure 2: Our work introduces self-healing machine learning. A healing mechanism $\mathcal{H}$ interacts with a deployed model f. $\mathcal{H}$ contains four *components*: monitoring, diagnosis, adaptation, and testing. The overall goal of SHML is to find optimal adaptation actions to maximize the predictive performance of a model f.

A theoretical design of concept sets: improving the predictability of Concept Bottleneck Models

Max Ruiz Luyten, *Mihaela van der Schaar*

Find the paper here

Abstract

Concept-based learning, a promising approach in machine learning, emphasizes the value of high-level representations called concepts. However, despite growing interest in concept-bottleneck models (CBMs), there is a lack of clear understanding regarding the properties of concept sets and their impact on model performance. In this work, we define concepts within the machine learning context, highlighting their core properties: 'expressiveness' and 'model-aware inductive bias', and we make explicit the underlying assumption of CBMs. We establish theoretical results for concept-bottleneck models (CBMs), revealing how these properties guide the design of concept sets that optimize model performance. Specifically, we demonstrate that well-chosen concept sets can improve sample efficiency and out-of-distribution robustness in the appropriate regimes. Based on these insights, we propose a method to effectively identify informative and non-redundant concepts. We validate our approach with experiments on CIFAR-10 and MetaShift, showing that concept-bottleneck models outperform the foundational embedding counterpart, particularly in low-data regimes and under distribution shifts. We also examine failure modes and discuss how they can be tackled.

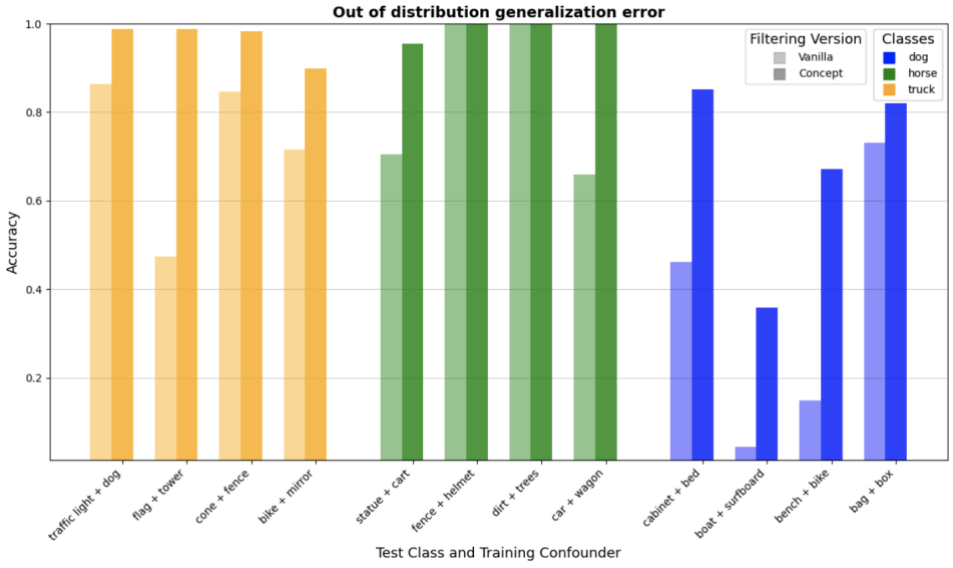


Figure 3: Concept representations consistently improve the out-of-distribution accuracy compared to directly using the CLIP embedding across different classes and distribution shifts. For three of the horse datasets, the CBM correctly identifies the presence of a horse.

POCA: Partially Observable Cost-Aware Active-Learning

Nicolás Astorga, Tennison Liu, Nabeel Seedat, *Mihaela van der Schaar*

Find the paper here

Abstract

Conducting experiments and gathering data for machine learning models is a complex and expensive endeavor, particularly when confronted with limited information. Typically, extensive *experiments* to obtain features and labels come with a significant acquisition cost, making it impractical to carry out all of them. Therefore, it becomes crucial to strategically determine what to acquire to maximize the predictive performance while minimizing costs. To perform this task, existing data acquisition methods assume the availability of an initial dataset that is both fully-observed and labeled, crucially overlooking the **partial observability** of features characteristic of many real-world scenarios. In response to this challenge, we present Partially Observable Cost-Aware Active-Learning (POCA), a new learning approach aimed at improving model generalization in data-scarce and data-costly scenarios through label and/or feature acquisition. Introducing µPOCA as an instantiation, we maximise the uncertainty reduction in the predictive model when obtaining labels and features, considering associated costs. µPOCA enhance traditional Active Learning metrics based solely on the observed features by generating the unobserved features through Generative Surrogate Models, particularly Large Language Models (LLMs). We empirically validate µPOCA across diverse tabular datasets, varying data availability, acquisition costs, and LLMs.

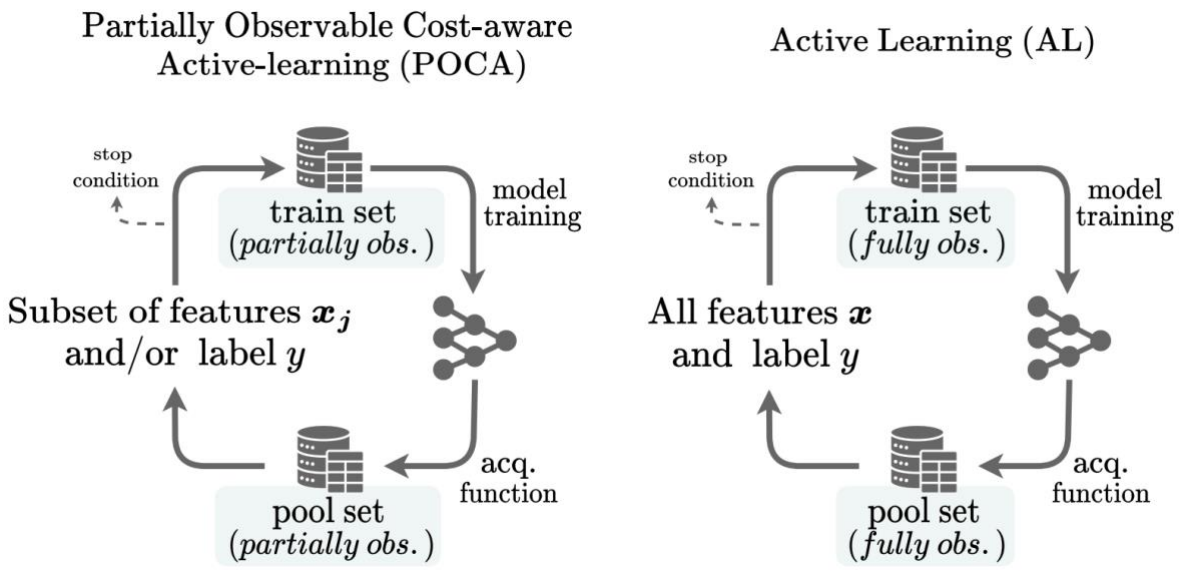


Figure 1: Overview of data acquisition methods. POCA acquires features and/or labels from a partially-observed pool incorporating them into a partially-observed training set. In contrast, AL targets label acquisition assuming a fully-observed pool set and training set.

Context-aware testing: A new paradigm for tabular model testing

Paulius Rauba*, Nabeel Seedat*, Max Ruiz Luyten, *Mihaela van der Schaar*

Find the paper here

Abstract

The predominant de facto paradigm of testing ML models relies on either using only held-out data to compute aggregate evaluation metrics or by assessing the performance on different subgroups. However, such data-only testing methods operate under the restrictive assumption that the available empirical data is the sole input for testing ML models, disregarding valuable contextual information that could guide model testing. In this paper, we challenge the go-to approach of data-only testing and introduce context-aware testing (CAT) which uses context as an inductive bias to guide the search for meaningful model failures. We instantiate the first CAT system, SMART Testing, which employs large language models to hypothesize relevant and likely failures, which are evaluated on data using a self-falsification mechanism. Through empirical evaluations in diverse settings, we show that SMART automatically identifies more relevant and impactful failures than alternatives, demonstrating the potential of CAT as a testing paradigm.

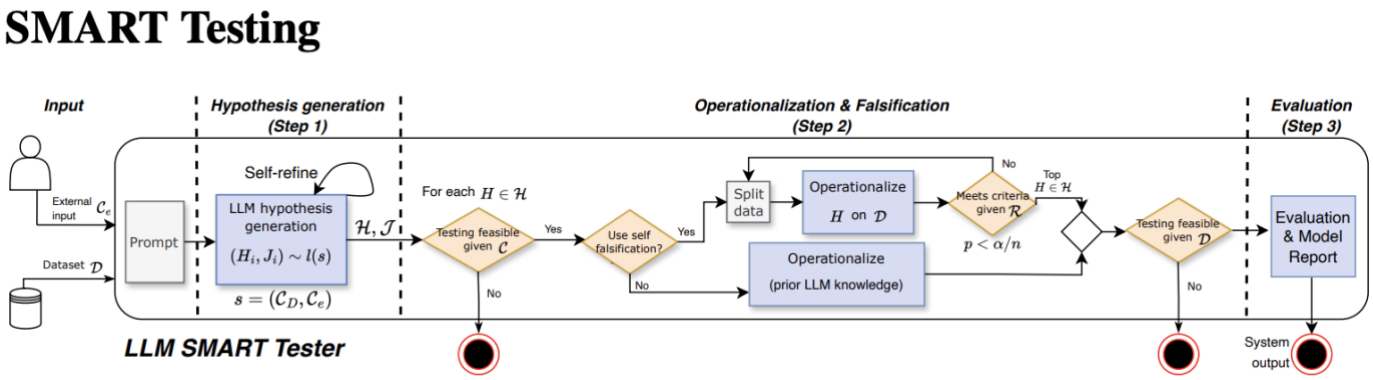


Figure 1: Overview of the SMART Testing Framework showing the four steps. All steps are automatically executed by an LLM.

A Closer Look at Deep Learning Phenomena Through A Telescoping Lens

Alan Jeffares*, Alicia Curth*, *Mihaela van der Schaar*

Find the paper here

Abstract

Deep learning sometimes appears to work in unexpected ways. In pursuit of a deeper understanding of its surprising behaviors, we investigate the utility of a simple yet accurate model of a trained neural network consisting of a sequence of first-order approximations telescoping out into a single empirically operational tool for practical analysis. Across three case studies, we illustrate how it can be applied to derive new empirical insights on a diverse range of prominent phenomena in the literature -- including double descent, grokking, linear mode connectivity, and the challenges of applying deep learning on tabular data -- highlighting that this model allows us to construct and extract metrics that help predict and understand the a priori unexpected performance of neural networks. We also demonstrate that this model presents a pedagogical formalism allowing us to isolate components of the training process even in complex contemporary settings, providing a lens to reason about the effects of design choices such as architecture & optimization strategy, and reveals surprising parallels between neural network learning and gradient boosting.

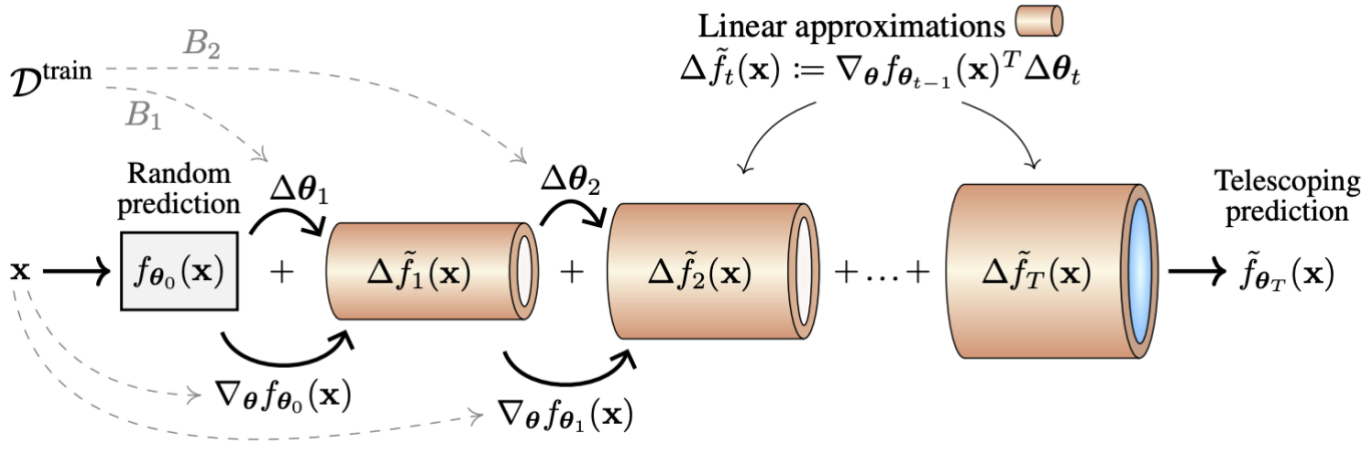


Figure 1: Illustration of the telescoping model of a trained neural network. Unlike the more standard framing of a neural network in terms of an iteratively learned set of parameters, the telescoping model takes a functional perspective on training a neural network in which an arbitrary test example's initially random prediction, fθo (x), is additively updated by a linearized adjustment Δ ˜ft(x) at each step t as in Eq. (5).

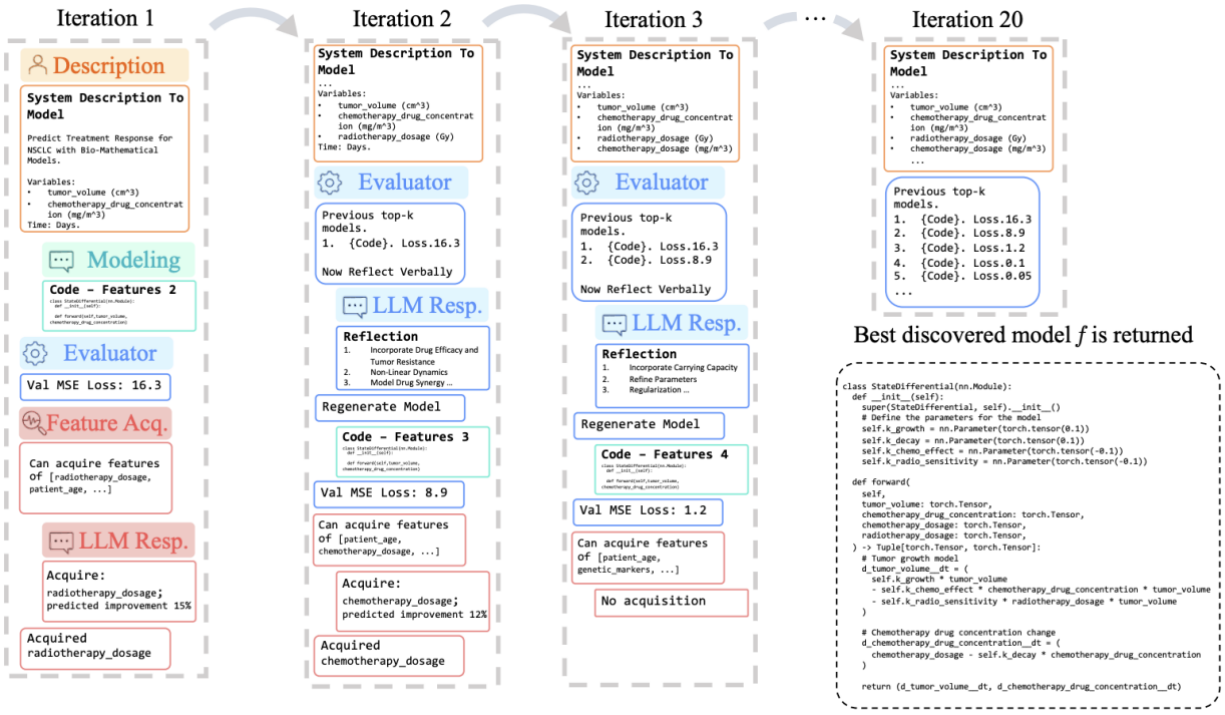
Data-Driven Discovery of Dynamical Systems in Pharmacology using Large Language Models

Samuel Holt*, Zhaozhi Qian*, Tennison Liu, Jim Weatherall, *Mihaela van der Schaar*

Find the paper here

Abstract

The discovery of dynamical systems is crucial across a range of fields, including pharmacology, epidemiology, and physical sciences. *Accurate* and *interpretable* modeling of these systems is essential for understanding complex temporal processes, optimizing interventions, and minimizing adverse effects. In pharmacology, for example, precise modeling of drug dynamics is vital to maximize therapeutic efficacy while minimizing patient harm, as in chemotherapy. However, current models, often developed by human experts, are limited by high cost, lack of scalability, and restriction to existing human knowledge. In this paper, we present the **Data-Driven Discovery (D3)** framework, a novel approach leveraging Large Language Models (LLMs) to iteratively discover and refine interpretable models of dynamical systems, demonstrated here with pharmacological applications. Unlike traditional methods, D3 enables the LLM to propose, acquire, and integrate new features, validate, and compare dynamical systems models, uncovering new insights into pharmacokinetics. Experiments on a pharmacokinetic Warfarin dataset reveal that D3 identifies a new plausible model that is well-fitting, highlighting its potential for precision dosing in clinical applications.



Discovering Preference Optimization Algorithms with and for Large Language Models

Chris Lu*, Samuel Holt*, Claudio Fanconi*, Alex James Chan, Jakob Nicolaus Foerster, Mihaela van der Schaar, Robert Tjarko Lange

Find the paper here

Abstract

Offline preference optimization is a key method for enhancing and controlling the quality of Large Language Model (LLM) outputs. Typically, preference optimization is approached as an offline supervised learning task using manually-crafted convex loss functions. While these methods are based on theoretical insights, they are inherently constrained by human creativity, so the large search space of possible loss functions remains under explored. We address this by performing LLM-driven objective discovery to automatically discover new state-of-the-art preference optimization algorithms without (expert) human intervention. Specifically, we iteratively prompt an LLM to propose and implement new preference optimization loss functions based on previously-evaluated performance metrics. This process leads to the discovery of previously-unknown and performant preference optimization algorithms. The best performing of these we call Discovered Preference Optimization (DiscoPOP), a novel algorithm that adaptively blends logistic and exponential losses. Experiments demonstrate the state-of-the-art performance of DiscoPOP and its successful transfer to held-out tasks.

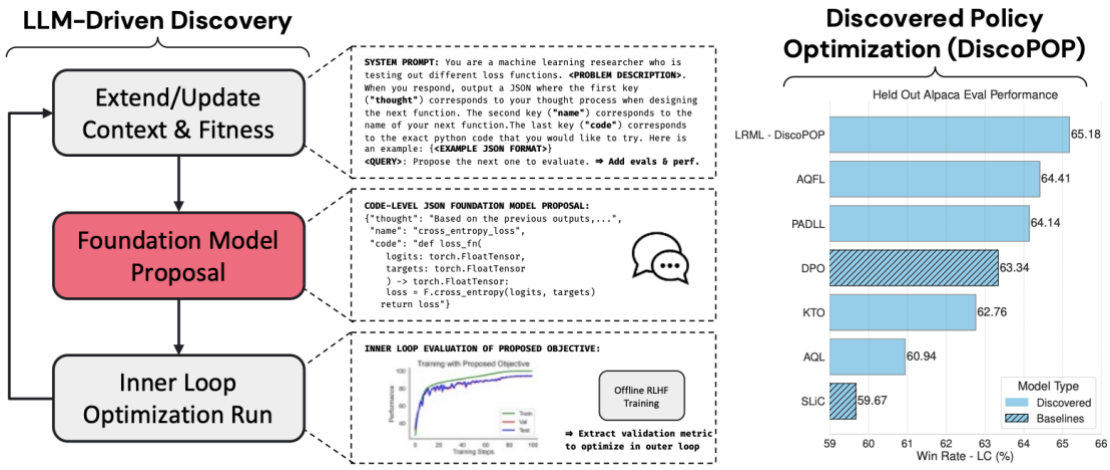


Figure 1: Left. Conceptual illustration of LLM-driven discovery of objective functions. We prompt an LLM to output new code-level implementations of offline preference optimization losses E(yw,yl ,x)∼D [f (βρ)] as a function of the policy (πθ) and reference model's (πref) likelihoods of the chosen (yw) and rejected (yl) completions. Afterwards, we run an inner loop training procedure and evaluate the resulting model on MT-Bench. The corresponding performance is fed back to the language model, and we query it for the next candidate. Right. Performance of discovered objective functions on Alpaca Eval.

Quantifying Aleatoric Uncertainty of the Treatment Effect: A Novel Orthogonal Learner

Valentyn Melnychuk, Stefan Feuerriegel, *Mihaela van der Schaar*

Find the paper here

Abstract

Estimating causal quantities from observational data is crucial for understanding the safety and effectiveness of medical treatments. However, to make reliable inferences, medical practitioners require not only estimating averaged causal quantities, such as the conditional average treatment effect, but also understanding the randomness of the treatment effect as a random variable. This randomness is referred to as aleatoric uncertainty and is necessary for understanding the probability of benefit from treatment or quantiles of the treatment effect. Yet, the aleatoric uncertainty of the treatment effect has received surprisingly little attention in the causal machine learning community. To fill this gap, we aim to quantify the aleatoric uncertainty of the treatment effect at the covariate-conditional level, namely, the conditional distribution of the treatment effect (CDTE). Unlike average causal quantities, the CDTE is not point identifiable without strong additional assumptions. As a remedy, we employ partial identification to obtain sharp bounds on the CDTE and thereby quantify the aleatoric uncertainty of the treatment effect. We then develop a novel, orthogonal learner for the bounds on the CDTE, which we call AU-learner. We further show that our AU-learner has several strengths in that it satisfies Neyman-orthogonality and is doubly robust. Finally, we propose a fully-parametric deep learning instantiation of our AU-learner.

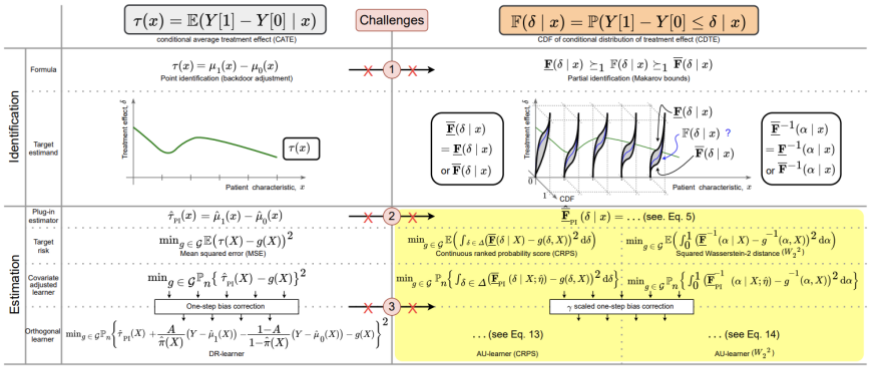


Figure 1: Identification and estimation of the conditional distribution of the treatment effect (CDTE) (=our setting) compared to the (well-studied) identification and estimation of the CATE. In this paper, we focus specifically on the CDF of the CDTE, $\mathbb{P}(Y[1] - Y[0] \leq \delta \mid x)$, shown in orange. Our main contribution relates to the estimation, shown in yellow. However, moving from CATE identification and estimation to our setting comes with important challenges: ① CATE (shown in **green**) is point identifiable but the CDTE is *not* (shown in **blue**); ② there is *no* closed-form expression of the target estimand in terms of nuisance functions and, because of that, CATE learners cannot be directly adapted for estimation; and ③ CATE is an unconstrained target estimand whereas Makarov bounds (shown in gray) are monotonous and contained in the interval $[0, 1]$.