Improving Generalization with Physical Equations

Antoine Wehenkel[*]
University of Liège

Jens Behrmann
Apple

Hsiang Hsu
Harvard

Guillermo Sapiro
Apple

Gilles Louppe
University of Liège

Jörn-Henrik Jacobsen
Apple

Abstract

Hybrid modelling reduces the misspecification of expert physical models with a machine learning (ML) component learned from data. Similarly to many ML algorithms, hybrid model performance guarantees are limited to the training distribution. To address this limitation, here we introduce a hybrid data augmentation strategy, termed *expert augmentation*. Based on a probabilistic formalization of hybrid modelling, we demonstrate that expert augmentation improves generalization. We validate the practical benefits of expert augmentation on a set of simulated and real-world systems described by classical mechanics.

1 Introduction

Recently, hybrid learning demonstrated success in complementing partial physical models with a machine learning component, e.g., [1, 2, 3, 4, 5, 6]. These works have shown that hybrid models are more faithful than their stand-alone physical counterparts. Moreover, training hybrid models is more sample-efficient and generalizes better than purely data-driven models. Indeed, the underlying physical models are often valid beyond the training data, and therefore, the corresponding hybrid models shall generalize to some unseen scenarios.

In this work, we first observe that current hybrid learning algorithms are sub-optimal in the amortized inference setting – when we aim to build hybrid models that are valid for various test configurations. Contrary to the common belief that hybrid learning achieves better generalization than black box ML models, we argue that hybrid learning algorithms do not yet meet their promise regarding robustness in amortized settings. Motivated by this observation, we introduce *expert augmentations* to extend the range of validity of hybrid models. Our experiments demonstrate that the proposed augmented hybrid models (AHMs) achieve generalization superior to standard hybrid learning algorithms.

2 Hybrid learning

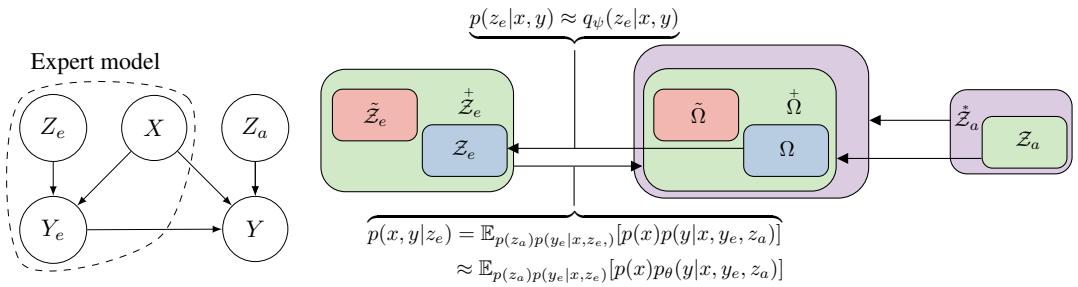
We formalize hybrid learning with the probabilistic model depicted in Figure 1a. The expert model, often derived from first-principles physics, is a conditional density $p(y_e|x, z_e)$ that describes the distribution of the *expert* response Y_e to an input x together with the physical parameters z_e. The *interaction model* is a conditional distribution $p(y|x, y_e, z_a)$ that aims to correct the expert model.

Our goal is to create a robust predictive model $p(y|x, (x_o, y_o))$ of the random variable Y, given the input x and independent observations (x_o, y_o) of the same system. Given an accurate estimation of z_a and z_e, we can predict the distribution of Y for any input x to the same system. In particular, the

[*]Work done as an intern at Apple.

Corresponding author: awehenkel@apple.com

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.



(a) A hybrid probabilistic model relates the output Y to the input X provided the latent variables Z_e and Z_a. The expert model defines the conditional density $p(y_e|x, z_e)$, where Y_e is an approximation of Y. Hybrid learning aims at learning the conditional distribution $p(y|x, y_e, z_a)$.

(b) Visualization of the distribution shifts considered in this work. The train support (in blue) Ω of (x, y) results from $(z_a, z_e) \in \mathcal{Z}_a \times \mathcal{Z}_e$. The test supports (in red) are denoted with a tilde symbols as $\tilde{\mathcal{Z}}_e$ for z_e and $\tilde{\Omega}$ for (x, y). The augmented support (in green) includes both train and test scenarios and corresponds to $(z_a, z_e) \in \mathcal{Z}_a \times \overset{+}{\mathcal{Z}}_e$. The outer violet domain depicts a situation in which the domain of z_a is also shifted. Hybrid modelling algorithms alone may learn a mapping $p_\theta : \overset{+}{\mathcal{Z}}_e \to \overset{+}{\Omega}$ but augmentation is necessary to learn the inverse mapping $q_\psi : \overset{+}{\Omega} \to \overset{+}{\mathcal{Z}}_e$.

Figure 1

Bayes optimal hybrid predictor p_B is

$$p_B(y|x, (x_o, y_o)) = \mathbb{E}_{p(z_a,z_e|(x_o,y_o))} \left[p(y|x, z_a, z_e)\right]. \tag{1}$$

In the amortized setting, we aim to learn a model of both the predictive model $p(y|x, z_a, z_e)$ and of the posterior over the parameters $p(z_a, z_e|(x_o, y_o))$. We will see that existing hybrid learning algorithms neglect the importance of building a robust encoder $p(z_a, z_e|(x_o, y_o))$ to make predictions in out-of-distribution (OOD) settings.

2.1 Hybrid generative modelling

We consider deterministic expert models $f_e : \mathcal{X} \times \mathcal{Z}_e \to \mathcal{Y}_e$, for which $p_\theta(y_e|x, z_e)$ is a Dirac distribution. Given a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$ of N IID samples, we aim to learn the interaction model $p_\theta(y|x, y_e, z_a)$ that fits the data well and minimally corrects the expert model. Two recent approaches, the APHYNITY algorithm [1] and the Hybrid-VAE [2, HVAE], compete in this setting.

APHYNITY. Yin et al. [1] augment an expert ordinary differential equation (ODE) with an additive correction. They assume the hybrid models can fit the data perfectly. This assumption is equivalent to taking a Dirac distribution for $p_\theta(y|x, y_e, z_a)$. They propose solving the following problem

$$\min_{z_e, F_a} ||F_a|| \quad \text{s.t.} \quad \forall (x, y) \in \mathcal{D}, \forall t, \frac{dy_t}{dt} = F_e(y_t) + F_a(y_t) \quad \text{with} \quad y_0 := x, \tag{2}$$

where $||\cdot||$ is a norm on the function space, $F_a : \mathcal{Y}_t \times \mathcal{Z}_a \to \mathcal{Y}_t$ is a learned function, $F_e : \mathcal{Y}_t \times \mathcal{Z}_e \to \mathcal{Y}_t$ is the expert model, and $\mathcal{D}$ is a training dataset, which contains initial states $x := y_0$ and k-long sequences $y \in \mathcal{Y} := (\mathcal{Y}_t)^k$. We amortise APHYNITY as suggested in the supplementary material of [1]. The encoder network $g_\psi(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}_a \times \mathcal{Z}_e$ corresponds to a Dirac distribution located at g_ψ. The interaction model is the solution of the augmented ODE in (2).

Hybrid-VAE (HVAE). In contrast to APHYNITY, the HVAE is not limited to additive interactions, nor to ODEs. The HVAE is a variational auto-encoder (VAE) in which the decoder specifically follows Figure 1a. The encoder $g_\psi(x, y)$ predicts a posterior distribution over z_a and z_e. The model is trained with the classical Evidence Lower Bound on the likelihood (ELBO) and three regularizers $R_{PPC}, R_{DA,1},$ and $R_{DA,2}$ that encourage the hybrid model to stay close to the expert model. We refer the reader to Takeishi and Kalousis [2] for more details. The predictor takes the form described by (1) where $p(z_a, z_e|(x_o, y_o))$ is approximated by the encoder $q_\psi(z_a, z_e|x, y)$ and $p(y|x, z_a, z_e)$ by the learned hybrid generative model.

3 Robust hybrid learning

We consider the important sub-class of distribution shifts for which the marginal train distribution $p(z_e)$ and test distribution $\tilde{p}(z_e)$ may be different but the marginals of z_a and x are constant. The

encoder of classical hybrid models fails in this context; it is unable to predict accurately z_e. In consequence, the hybrid model derived from (1) does not generalize. We address this failure by introducing *expert augmentation* to fine tune the encoder. Our goal is to learn a predictive model that is *exact* on both the train and shifted test domains. A predictive model $\hat{p}(a|b)$ is $\mathcal{E}$-*exact*, or *exact* on the sample space $\mathcal{E}$, if $\hat{p}(a|b) = p(a|b) \forall (a, b) \in \mathcal{E}$ for a given ground-truth $p(a|b)$. A model is *robust* to a test scenario if its *exactness* on the training domain implies exactness on the test domain.

We now define an augmented distribution $\overset{+}{p}(z_e)$ over the expert variables whose support $\overset{+}{\mathcal{Z}}_e$ includes the joint support $\mathcal{Z}_e \cup \tilde{\mathcal{Z}}_e$ between the train and test distribution of the physical parameters. As depicted in Figure 1b, we denote the corresponding support over the observation space $\mathcal{X} \times \mathcal{Y}$ as $\overset{+}{\Omega}, \Omega$, and $\tilde{\Omega}$, respectively. In this context, and with **A1**, we demonstrate that even under perfect learning, classical hybrid learning algorithms do not produce an $\tilde{\Omega}$-*exact* predictor while our augmentation strategy does.

Assumption 1 (A1): *Hybrid modelling learns an interaction model $p_\theta(y|y_e, x, z_a)$ that is $\overset{+}{\Omega}$-exact.*

The hypothesis **A1** is consistent with the recent literature on hybrid modelling [1, 2], which assumes that $p_\theta(y|y_e, x, z_a)$ should not be overly complex. For example, extrapolation to unseen y_e should hold for additive defects. That said, the exactness of the interaction model p_θ on $\overset{+}{\Omega}$ is insufficient to prove that the predictive model $p_{\theta,\psi}$ is $\overset{+}{\Omega}$-*exact*. Indeed, the encoder q_ψ is only trained on the training data and cannot rely on a strong inductive bias in contrast to p_θ. Thus, even if the encoder is exact on the training distribution, the corresponding predictive model is not $\overset{+}{\Omega}$-*exact*. While the decoder's performance are not limited to the training scenarios thanks to the broader validity of the expert model, the encoder does not generalize to unseen settings as it is purely data-driven.

3.1 Expert augmentation

We increase the encoder's generalization domain by training it on additional synthetic configurations generated by the hybrid decoder. Once trained, the hybrid model is composed of an encoder q_ψ and an interaction model p_θ that are respectively Ω- and $\overset{+}{\Omega}$-*exact*. We create an augmented training set with support over $\overset{+}{\Omega}$ by sampling physical parameters $z_e \in \overset{+}{\mathcal{Z}}_e$. Then, we finetune the encoder q_ψ on $\overset{+}{\Omega}$. Under perfect training, the predictive model becomes robust; $p_{\theta,\psi}(y|x, (x_o, y_o))$ is $\overset{+}{\Omega}$-*exact*.

After training a hybrid model, we have to decide on a realistic distribution $\overset{+}{p}(z_e)$ in order to create the dataset $\overset{+}{\mathcal{D}}$ by sampling from the hybrid model. The augmented training set $\overset{+}{\mathcal{D}}$ contains ground truth values for the expert's variables z_e, in contrast to the original dataset. We freeze the interaction model, as we assume it is $\overset{+}{\Omega}$-*exact*. We only fine-tune the encoder q_ψ on $\overset{+}{\mathcal{D}}$ with a combination of the loss function ℓ of the original hybrid learning algorithm and supervision on the latent variable objective. We then learn a decoder that solves

$$\psi^\star = \arg\min_\psi \mathbb{E}_{\overset{+}{\mathcal{D}}} \left[\ell(x, y; \theta, \psi) - \log q_\psi(z_e|x, y) \right].$$

4 Experimental results

We assess the benefits of expert augmentation on two synthetic problems described by an ODE

$$\frac{dy_t}{dt} = F_e(y_t; z_e) + F_a(y_t; z_a). \tag{3}$$

The initial state y_0 corresponds to X, and the sequence of states $y_{1:t_1} := [y_{i\Delta t}]_{i=1}^{t_1/\Delta t}$ to Y.

The damped pendulum: the system's state at time t is $y_t = \begin{bmatrix} \theta_t & \dot{\theta}_t \end{bmatrix}^T$, where θ_t is the angle of the pendulum at time t and $\dot{\theta}_t$ its angular speed. The evolution of the state over time is described by (3), where $z_e := \omega$, $z_a = \alpha$, $F_e := \begin{bmatrix} \dot{\theta} & -\omega_0^2 \sin\theta \end{bmatrix}^T$, and $F_a := \begin{bmatrix} 0 & -\alpha\dot{\theta} \end{bmatrix}^T$. The train domain is generated from $\mathcal{Z}_e := [1.5, 3.1]$ and $\mathcal{Z}_a := [0, 0.6]$ and at testing $\tilde{\mathcal{Z}}_e := [0.5, 1.5]$. **The 2D reaction**

diffusion: this is a 2D FitzHugh-Nagumo model on a 32×32 grid. The system's state at time t is a $2 \times 32 \times 32$ tensor $y_t = [u_t \quad v_t]^T$. The evolution of the state over time is described by $F_e := [a\Delta u_t \quad b\Delta v_t]^T$ and $F_a := [R_u(u_t, v_t; k) \quad R_v(u_t, v_t)]^T$, where $z_e := \{a, b\}$, $z_a = \{k\}$, Δ is the Laplace operator, and the reaction terms are $R_u(u, v; k) = u - u^3 - k - v$ and $R_v(u, v) = u - v$. The train domain is generated from $\mathcal{Z}_e := [1\text{e}{-3}, 2\text{e}{-3}] \times [3\text{e}{-3}, 7\text{e}{-3}]$ and $\mathcal{Z}_a := [3\text{e}{-3}, 5\text{e}{-3}]$ and at testing $\tilde{\mathcal{Z}}_e := [2\text{e}{-3}, 4\text{e}{-3}] \times [1\text{e}{-3}, 1\text{e}{-1}]$. **The double pendulum:** we use the real-world dataset of a double pendulum introduced by Asseman et al. [7] and consider a friction-less system to derive an expert model; see, e.g., [8] for the details. This idealized model misses non-negligible effects such as errors in extracting the pendulum's positions, frictions, or vibrations. This experiment allows us to validate the effectiveness of expert augmentation when the exact nature of the mispecification is unknown. We use low-energy configurations for the training set and high-energy for the test to enforce a distribution shift. The validation set contains samples whose energy is between the train and test examples. In our experiment, the initial positions of the pendulum are given. The encoder network must predict the initial angular speed given a partial observation of the pendulum position for a few consecutive measurements.

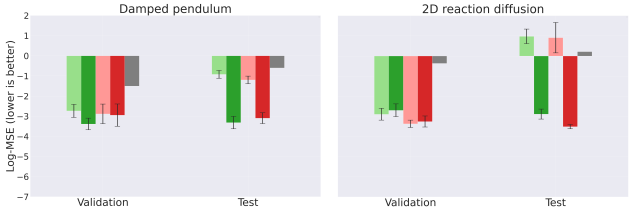


Figure 2: The average log-MSEs over ten runs. We compare HVAE (in red) and APHYNITY (in green), in light colours, to their expert augmented versions HVAE+ and APHYNITY+, in darker colours, and to a simple baseline (in grey). *On the test sets, the proposed AHMs outperform the original ones. Augmentation conserves good performance on the validation set.*

Dataset	Pendulum		RLC	
	Val.	Test	Val.	Test
APH.	$6_{\pm 2}$	$66_{\pm 9}$	$2_{\pm 0}$	$27_{\pm 2}$
APH.+	$6_{\pm 2}$	$10_{\pm 4}$	$2_{\pm 0}$	$3_{\pm 1}$
HVAE.	$3_{\pm 1}$	$117_{\pm 10}$	$2_{\pm 0}$	$32_{\pm 10}$
HVAE+	$2_{\pm 1}$	$11_{\pm 2}$	$2_{\pm 0}$	$2_{\pm 0}$

Table 1: The relative errors on z_e (in %) over ten runs. The + denotes augmented versions. *The accuracy of APHYNITY and HVAE collapses on the OOD test set while the augmented versions' performance is stable.*

Metric	Log-MSE			Rel. Err.		
Dataset	Train	Valid	Test	Train	Valid	Test
Exp.	$-4.4_{\pm 0}$	$-3.9_{\pm .4}$	$2.8_{\pm .4}$	$324_{\pm 3}$	$186_{\pm 1}$	$222_{\pm 3}$
APH.	$\mathbf{-6.7_{\pm 1.2}}$	$-5.2_{\pm .8}$	$-3.4_{\pm .4}$	$211_{\pm 90}$	$106_{\pm 41}$	$147_{\pm 61}$
APH.+	$-6.1_{\pm .6}$	$\mathbf{-5.3_{\pm .5}}$	$\mathbf{-4.4_{\pm .3}}$	$\mathbf{157_{\pm 27}}$	$\mathbf{71_{\pm 10}}$	$\mathbf{72_{\pm 10}}$

Table 2: The average log-MSEs and average relative errors on the initial angular speeds of the double pendulum over three runs for three predictive models: the expert model only (Exp.), APHYNITY (APH.) and APHYNITY followed by expert augmentation (APH.+). *Expert augmentation outperforms other models except for predicting the state evolution on the training set.*

Our synthetic experiments show the effect of expert augmentation on APHYNITY and HVAE. All models explicitly assume that the interaction model follows the structure of (3) where F_a is a small neural network. We select the best models from validation performance.

Figure 2 and Table 1 demonstrate that HVAE and APHYNITY are not robust to OOD test scenarios in comparison to the corresponding proposed AHMs. We only compare performance in OOD settings in Figure 2 as hybrid models have already demonstrated greater performance than non-hybrid models in the literature [1, 2]. Both algorithms strongly benefit from expert augmentation. Table 1 shows that the encoder does not predict the physical parameters perfectly. This failure indicates that the encoder is not Ω-*exact*, and neither should the decoder. However, this departure from an ideal setting does not preclude the effectiveness of the proposed expert augmentation. Expert augmentation also improves the parameter estimation in the OOD setting. The augmented encoder accurately estimates the physical parameters in the IID and OOD settings. Finally, Table 2 show that the proposed expert augmentations is also effective in practical settings. *Overall, AHMs outperform classical hybrid learning both in the predictive accuracy and in inferring the expert variables.*

5 Conclusions

We have presented a simple augmentation strategy, termed *expert augmentation*, to improve the generalization capabilities of hybrid learning models to scenarios where the expert model is valid.

Broader Impact Statement

This work simplifies the combinations of domain knowledge within machine learning models. As with most machine learning algorithms, the real-world impact may be positive or negative. However, improving robustness and interpretability of machine learning models may eventually help prevent harms caused by the failure and unpredictability of purely data-driven models under real-world distribution shifts.

References

[1] Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel de Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12): 124012, 2021.

[2] Naoya Takeishi and Alexandros Kalousis. Physics-integrated variational autoencoders for robust and interpretable generative modeling. *Advances in Neural Information Processing Systems*, 34, 2021.

[3] Zhaozhi Qian, William R Zame, Mihaela van der Schaar, Lucas M Fleuren, and Paul Elbers. Integrating expert odes into neural odes: Pharmacology and disease progression. *arXiv preprint arXiv:2106.02875*, 2021.

[4] Viraj Mehta, Ian Char, Willie Neiswanger, Youngseog Chung, Andrew Oakleigh Nelson, Mark D Boyer, Egemen Kolemen, and Jeff Schneider. Neural dynamical systems: Balancing structure and flexibility in physical prediction. *arXiv preprint arXiv:2006.12682*, 2020.

[5] Chon Lok Lei and Gary R Mirams. Neural network differential equations for ion channel modelling. *Frontiers in Physiology*, page 1166, 2021.

[6] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

[7] Alexis Asseman, Tomasz Kornuta, and Ahmet Ozcan. Learning beyond simulated physics. In *Modeling and Decision-making in the Spatiotemporal Domain Workshop*, 2018. URL `https://openreview.net/pdf?id=HylajWsRF7`.

[8] Tomasz Stachowiak and Toshio Okada. A numerical analysis of chaos in the double pendulum. *Chaos, Solitons & Fractals*, 29(2):417–422, 2006.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No] Due to the 4-page limit. However, we will do it in the corresponding long version.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the broader impact statement.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We will do it for the corresponding long version of this paper.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [No] Due to space limit and strong discouragement of appendices, it was difficult to provide the necessary details to reproduce our experiment exactly. However, the corresponding long version will describe the experiments thorougly in the appendices.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] Final results were generated on a 32GB MacBook Pro from 2020.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [No]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The dataset is publicly available and we cite properly the authors.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]