arXiv:2202.03881v2 [cs.LG] 9 Feb 2022

Robust Hybrid Learning With Expert Augmentation

Antoine Wehenkel [1 2] **Jens Behrmann** [3] **Hsiang Hsu** [4 2] **Guillermo Sapiro** [3] **Gilles Louppe** [1]
Jörn-Henrik Jacobsen [3]

Abstract

Hybrid modelling reduces the misspecification of expert models by combining them with machine learning (ML) components learned from data. Like for many ML algorithms, hybrid model performance guarantees are limited to the training distribution. Leveraging the insight that the expert model is usually valid even outside the training domain, we overcome this limitation by introducing a hybrid data augmentation strategy termed *expert augmentation*. Based on a probabilistic formalization of hybrid modelling, we show why expert augmentation improves generalization. Finally, we validate the practical benefits of augmented hybrid models on a set of controlled experiments, modelling dynamical systems described by ordinary and partial differential equations.

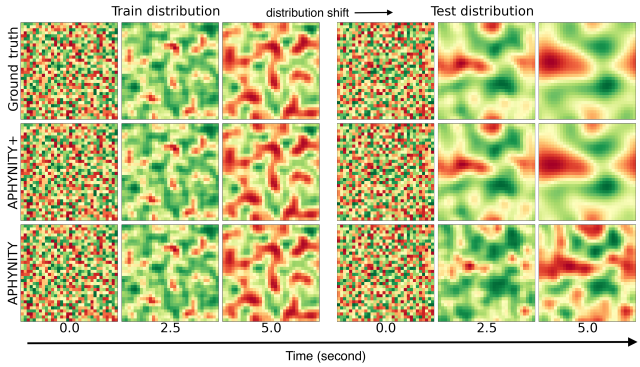

Figure 1. APHYNITY, an existing hybrid modelling strategy, is unable to predict accurately the dynamic of a 2D diffusion reaction for a shifted test distribution although it predicts well configuration that follows the training distribution. On the opposite, APHYNITY+, the same model fine-tuned with our data augmentation, generalizes to shifted distributions as expected from the validity of the underlying physics.

1. Introduction

Generalization to unseen data is a key property of a useful model. When training and test data are independently and identically distributed (IID), one way to check generalization is by evaluating the model on a held out subset of the training data or with k-fold cross validation. Unfortunately, this setting is often unrealistic because the training scenario is rarely fully representative of the test scenario. This has motivated lot of recent research efforts to focus on the robustness of ML models (Gulrajani & Lopez-Paz, 2020; Geirhos et al., 2020; Koh et al., 2021). Common strategies can be broadly grouped in two categories: The first class of methods aims at aligning specific properties of the model (e.g., invariance, equivariance, monotonicity, etc.) with expertise on the problem of interest (Cubuk et al., 2019; Mahmood et al., 2021; Keriven & Peyré, 2019; Silver et al., 2017). The second category is data focused (Sagawa et al., 2019; Arjovsky et al., 2019; Krueger et al., 2021; Creager et al., 2021) and leverages variations present in the training data, e.g. by minimizing worst case sub-group performance, to achieve robustness.

The data oriented methods, which include GroupDRO (Sagawa et al., 2019) and Invariant Risk Minimization (Arjovsky et al., 2019, IRM), can be very appealing because they only require implicit specification of invariances via domains or environments. However, these methods' performance is limited to variations present in the training data and the inductive bias of the ML algorithm. This may be insufficient when the modelling problem is too complex or the variations of interest are not present in the training data. On the other hand, methods based on domain-specific expertise do not suffer from such limitations. Embedding expertise into a model can be done via architectural inductive biases (LeCun et al., 1995; Xu et al., 2018), data augmentation (Cubuk et al., 2019), or a learning objective (Cranmer et al., 2020) that enforces established symmetries of the problem. As an example, simple data augmentation techniques combined with convolutions lead to excellent performance on natural image problems (Cubuk et al., 2019). Another natural approach to embed expertise in ML models, and the one studied in this paper, is called hybrid learning (HyL). HyL combines an expert model (e.g., physics-motivated equations) with a learned component that improves the expert model so that the combination better fits real-world

[*]Equal contribution [1]University of Liege [2]Work done as an intern at Apple [3]Apple [4]Harvard University. Correspondence to: Antoine Wehenkel <antoine.wehenkel@uliege.be>.

Copyright 2022 by the author(s).

data. A particularity of HyL is the central role played by the expertise, which is supposed to provide a simple and well-grounded parametric description of the process considered. HyL usually considers the expert model as an analytical function, or as a set of equations, that relates the expert parameters to the quantity of interest. The expert model is often motivated by the underlying physics of the system considered. Hence, we will use the terms *expert* model and *physical* model interchangeably.

In recent work (Yin et al., 2021; Takeishi & Kalousis, 2021; Qian et al., 2021; Mehta et al., 2020; Lei & Mirams, 2021; Reichstein et al., 2019), HyL demonstrated success in complementing partial physical models and improving the inference of the corresponding parameters. However, contrarily to the common belief that HyL achieves better generalization than black box ML models, we argue that hybrid models do not meet their promise regarding robustness. Although HyL achieves strong performance on IID test distributions by exploiting the inductive bias of the expert models, we show that their performance collapses when the test domain is not included in the training domain. This is unsatisfactory as the expert model is typically well-defined for a range of parameters that can correspond to realistic data far outside of the training distribution. A test distribution not covered by the training data, but for which an expert model exists, happens often in the real world. As an example, Qian et al. (2021) apply HyL to a pharmacological model describing the effect of a COVID-19 treatment for which only a limited quantity of real-world data is available. In this context, although the underlying biochemical dynamic of treatments is well modelled, data is often scarce and biased. Therefore, the hybrid model does not necessarily generalize to configurations that are well modelled by the pharmacological model but unseen during training.

We introduce *expert augmentations* for training augmented hybrid models (AHMs), a procedure that extends the range of validity of hybrid models and improves generalization as pictured by Figure 1. Our contribution is to first formalise the HyL problem as: 1) Learning a probabilistic model partially defined by the expert model; 2) Performing inference over this probabilistic hybrid model. In this context, we show that HyL is vulnerable to distribution shifts for which the expert model is well defined (see Figure 1, bottom row). Motivated by our analysis, we propose to fine-tune the hybrid model on an expert-augmented dataset that includes distribution shifts (see results of augmentation in Figure 1, middle row). These expert augmentations only rely on the hybrid model itself, leveraging that the expert model is also well-defined outside of the training distribution. Our experiments on various controlled HyL problems demonstrate that AHMs achieve multiple orders of magnitude superior generalization in realistic situations and can be applied to any state-of-the-art HyL algorithm.

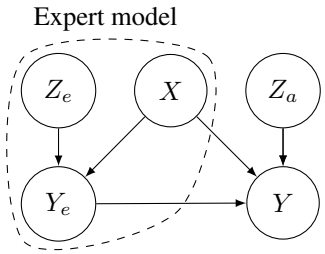

Figure 2. A hybrid probabilistic model which describes the relationship between the input X and the output Y for a configuration of the system as defined by the latent variables Z_e and Z_a. The prescribed expert model defines the conditional density $p(y_e|z_e, x)$, where Y_e is an approximation of Y. Hybrid learning aims at learning the conditional distribution $p(y|z_a, y_e, x)$.

2. Hybrid learning

In order to show that our proposed expert augmentations lead to robust models, we first formalize hybrid learning with the probabilistic model depicted in Figure 2. In this Bayesian network, capital letters denote random variables (e.g., Y) and, in the following, we will use calligraphic letters for the domain of the corresponding realization (e.g., $y \in \mathcal{Y}$). In our formalism, the expert model 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 a parametric description of the system z_e, denoting expert or physical parameters. We augment the expert model with the *interaction model* which is a conditional distribution $p(y|x, y_e, z_a)$ that describes the distribution of the observation Y given the input x, the expert model response y_e, and a parametric description of the interaction model z_a.

Our final goal is to create a robust predictive model $p(y|x, (x_o, y_o))$ of the random variable Y, given the input x together with independent observations (x_o, y_o) of the same system, where the subscript o denotes an observed quantity. As a concrete example, we consider predicting the evolution of a damped pendulum (described in Section 4.1) given its initial angle and speed ($x = \left[\theta, \dot{\theta}\right]$) and a sequence of observations of the same pendulum. The expert model we assume is able to describe a frictionless pendulum whose dynamic is only characterized by one parameter $z_e := \omega_0$, denoting its fundamental frequency. A perfect description of the system should model the friction with a second parameter $z_a := \alpha$, the damping factor. In this problem, (x_o, y_o) and (x, y) are IID realization of the same pendulum which corresponds, in general terms, to samples from $p(x, y|z_a, z_e)$ for some fixed but unknown values of z_a and z_e. The expert variables z_e (e.g., ω_0) together with z_a (e.g., α) should accurately describe the system that produces Y (e.g., the evolution of the pendulum's angle and speed along time) from X (e.g., the initial pendulum's state). In our setting we assume that we are given a pair (x_o, y_o) (e.g., past observations) from

which we can accurately infer the state of the system (z_a, z_e) as described by the interaction and expert models, and then predict the distribution of Y for a given input x (e.g., forecasting future observations) to the same system. Because the interaction between z_e and y is essentially defined by the expert model, it should be possible, and preferable, to learn an accurate predictive model of Y whose accuracy is independent from the training distribution of the expert variables z_e. Provided all probability distributions in Figure 2 are known, the Bayes optimal hybrid predictor p_B can be written as

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

We observe that the Bayes optimal predictor explicitly depends on the posterior $p(z_a, z_e|(x_o, y_o))$ which is itself a function of the marginal distribution over z_e. This may preclude the existence of a good predictor that is invariant to shift of $p(z_e)$. However, in the following we will consider that the pair (x_o, y_o) contains enough information about the parameters z_a, z_e. As a consequence, the posterior distribution shrinks around the correct parameters value and the effect of the prior becomes negligible.

2.1. Hybrid generative modelling

We consider expert models that are deterministic; that is, for which $p_\theta(y_e|x, z_e)$ is a Dirac distribution. The expert model describes the system as a function $f_e : \mathcal{X} \times \mathcal{Z}_e \to \mathcal{Y}_e$ that computes the response y_e to an input x, parameterized by expert variables z_e. The goal of hybrid modelling is to augment the expert model with a learned component from data as depicted in Figure 2. Formally, 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 but is close to the expert model. For example, we could define closeness via a small L2-distance between expert and hybrid outputs or via a small Kullback-Leibler (KL) divergence between the marginal distributions of Y and Y_e.

Learning a model that is close to the expert model and fits the training data well is a hard problem. However, the APHYNITY algorithm (Yin et al., 2021) and the Hybrid-VAE (Takeishi & Kalousis, 2021, HVAE) are two recent approaches that offer promising solutions to this problem. We now briefly describe these two methods and how they can be used to approximate the Bayes optimal predictor of (1). Our augmentation strategy is compatible (and effective) with both approaches.

APHYNITY. Yin et al. (2021) formulate hybrid learning in a context where the expert model is an ordinary differential equation (ODE). They consider an additive hybrid model that should perfectly fit the data, which is equivalent to assuming the conditional distribution $p_\theta(y|x, y_e, z_a)$ is a Dirac distribution. Formally, they solve the optimization problem

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

where $||\cdot||$ is a norm operator on the function space, $F_a : \mathcal{Y}_t \times \mathcal{Z}_a \to \mathcal{Y}_t$ is a learned function, $F_e : \mathcal{Y}_t \times \mathcal{Z}_e \to \mathcal{Y}_t$ defines the expert model and $\mathcal{D}$ is a dataset of initial states $x := y_0$ and sequences $y \in \mathcal{Y} := (\mathcal{Y}_t)^k$, where k is the number of observed timesteps. APHYNITY solves this problem with Lagrangian optimization and Neural ODEs (Chen et al., 2018) to compute derivatives. In the context of ODEs, the random variable X is the initial state of the system at t_0 and Y is the observed sequence of k states between t_0 and t_1.

This formulation only considers learning a missing dynamic for one realization of the system described by Figure 2, for a single z_a and z_e. However, we are interested in learning a hybrid model that works for the full set of systems described by Figure 2. As suggested in Yin et al. (2021), we use an encoder network $g_\psi(\cdot,\cdot) : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}_a \times \mathcal{Z}_e$ that corresponds to a Dirac distribution located at g_ψ as the approximate posterior $q_\psi(z_a, z_e|x, y)$. The interaction model is a product of Dirac distributions whose locations correspond to the solution of the ODE

$$\frac{dy_t}{dt} = F_e(y_t, z_e) + F_a(y_t, z_a; \theta), \quad y_0 := x. \quad (3)$$

Hence the corresponding approximate Bayes predictor replaces the parameters (z_a, z_e) in (3) with the prediction of g_ψ and predicts a product of Dirac distributions.

Hybrid-VAE (HVAE). In contrast to APHYNITY, the model proposed by Takeishi & Kalousis (2021) is not limited to additive interactions between the expert model and the ML model, nor to ODEs. Instead, their goal is to learn the generative model described by Figure 2. They achieve this with a variational auto-encoder (VAE) where the decoder specifically follows Figure 2. Similarly to the amortized APHYNITY model, the encoder $g_\psi(x, y)$ predicts a posterior distribution over z_a and z_e, and the model is trained with the classical Evidence Lower Bound on the likelihood (ELBO). Takeishi & Kalousis (2021) observe that relying only on an architectural inductive bias and maximum likelihood training is not enough to ground the generative model to the expert equations. They propose to add three regularizers $R_{PPC}, R_{DA,1}$, and $R_{DA,2}$ that encourage the generative model to rely on the expert model. The final objective is

$$\max_{\theta,\psi} \mathbb{E}_{\mathcal{D}}\left[\text{ELBO}((x,y);\psi,\theta)\right] + \alpha R_{PPC} + \beta R_{DA,1} + \gamma R_{DA,2}. \quad (4)$$

The first regularizer, R_{PPC}, encourages the marginal distribution of samples generated by the complete model to be

close to the marginal distribution that would be only generated by the physical model. The two other regularizers specifically require the encoder network for z_e to be made of two sub-networks. The first network filters the observations to keep only what can be generated by the expert model alone, and the second should map the filtered observations to the posterior distribution over z_e. $R_{DA,1}$ penalizes the objective if the observations generated by the expert model are not close to the filtered observations. Finally, $R_{DA,2}$ relies on data augmentation with the expert model to enforce that the second sub-network correctly identifies the expert variables z_e when the observations are correctly filtered. We refer the reader to Takeishi & Kalousis (2021) for more details on HVAE. For HVAE, the approximate 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 $p(y|x, z_e, z_a)$ by the learned hybrid generative model.

3. Robust hybrid learning

We now formalize our definition of out of distribution (OOD) and robustness. In general, a test scenario is OOD if the joint test distribution $\tilde{p}(x, y)$ is different from the training distribution $p(x, y)$, that is $d(\tilde{p}, p) > 0$ for any properly defined divergence or distance d. In the following, we reduce our discussion to a sub-class of distribution shifts for which the marginal train and test distributions over z_e may be different, $d(p(z_e), \tilde{p}(z_e)) > 0$, but the marginals of z_a and x are constant. As a consequence, the joint distribution of (x, y) pairs is also shifted. Formally, the training and test distributions are respectively defined as

$$p(x, y) := \mathbb{E}_{p(z_e)p(z_a)p(y_e|z_e,x)} \left[p(x)p(y|z_a, x, y_e)\right],$$
$$\tilde{p}(x, y) := \mathbb{E}_{\tilde{p}(z_e)p(z_a)p(y_e|z_e,x)} \left[p(x)p(y|z_a, x, y_e)\right].$$

In this context, we demonstrate, theoretically and empirically, that classical hybrid models fail. To address this failure, we introduce *augmented hybrid models* and show that, under some assumptions, they achieve optimal performance on both the train and test distributions.

Our goal is to learn a predictive model

$$p_{\theta,\psi}(y|x, (x_o, y_o)) = \mathbb{E}_{\substack{q_\psi(z_a,z_e|x_o,y_o) \\ p(y_e|z_e,x)}} \left[p_\theta(y|y_e, x, z_a)\right]$$

that is *exact* on both the train and test domains when they follow the aforementioned training and testing distribution shifts. We say that a learned predictive model $\hat{p}(a|b)$ is $\mathcal{E}$-*exact*, or *exact* on the sample space $\mathcal{E}$, if $\hat{p}(a|b) = p(a|b) \quad \forall (a, b) \in \mathcal{E}$. Here we qualify a predictive model as *robust* to a test scenario if its *exactness* on the training domain is sufficient to ensure 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 3, 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 may 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.*

Although strong, **A1** is consistent with the recent literature on hybrid modelling, which assumes that $p(y_e|x, z_e)$ is an accurate description of the system, thereby $p_\theta(y|y_e, x, z_a)$ should not be overly complex. As an example, we consider an additive interaction model in our experiments for which extrapolation to unseen y_e holds if this assumption is correct. That said, we still notice that 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 does not achieve exactness outside Ω.

3.1. Expert augmentation

We propose a data augmentation strategy to improve the robustness of hybrid models to unseen test scenarios. 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 may create a new training distribution with a support over $\overset{+}{\Omega}$ by sampling physical parameters z_e from a distribution that covers $\overset{+}{\mathcal{Z}}_e$. We can then train the encoder q_ψ on $\overset{+}{\Omega}$, under perfect training the corresponding predictive model $p_{\theta,\psi}(y|x, (x_o, y_o))$ is $\overset{+}{\Omega}$-*exact*, hence exact on both train and test domains.

Our learning strategy is grounded in existing hybrid modelling algorithms. Here, we focus on APHYNITY and HVAE, but our approach is applicable to other HyL algorithms. We first train an encoder q_ψ and a decoder p_θ with a HyL algorithm. Together with experts we then decide on a realistic distribution $\overset{+}{p}(z_e)$ and create a new dataset $\overset{+}{\mathcal{D}}$ by sampling from the hybrid generative model defined by Figure 2 and the interaction model p_θ. A notable difference between the augmented training set $\overset{+}{\mathcal{D}}$ and the original training set $\mathcal{D}$ is that the former contains ground truth values for the expert's variables z_e. As we assume that the interaction model is $\overset{+}{\Omega}$-*exact*, we freeze it and only fine-tune the encoder q_ψ on $\overset{+}{\mathcal{D}}$. We use a combination of the loss function ℓ of the original HyL algorithm (e.g., (4) for HVAE, and the Lagrangian of (2) for APHYNITY) and a supervision on

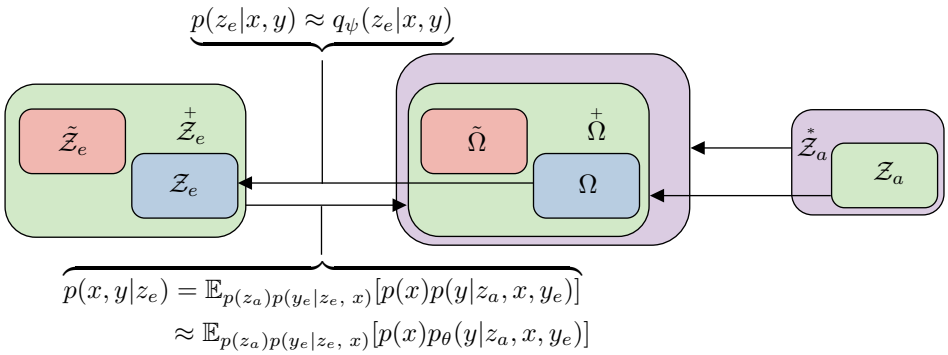

Figure 3. Visualization of the distribution shifts considered in this work. The train support Ω of (x, y) results from $(z_a, z_e) \in \mathcal{Z}_e \times \mathcal{Z}_a$. The test supports (in red) are denoted with a tilde symbols as $\tilde{\mathcal{Z}}_e$ for z_e and $\tilde{\Omega}$ for (x, y). The augmented support $\overset{+}{\Omega}$ (in green) includes both train and test scenarios and corresponds to $(z_a, z_e) \in \overset{+}{\mathcal{Z}}_e \times \mathcal{Z}_a$. The outer violet domain that includes $\overset{+}{\Omega}$ depicts one of our experiment 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 $q_\psi : \overset{+}{\Omega} \to \overset{+}{\mathcal{Z}}_e$.

the latent variable objective to learn a decoder that solves

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

In our experiments we chose a Gaussian model for the posterior, which is equivalent to a mean square error (MSE) loss on the physical parameters. We provide a detailed description of the expert augmentation scheme in Appendix A.

As a side note, we would like to emphasize the difference between the data augmentation proposed in this paper and the one from Takeishi & Kalousis (2021). While HVAE also requires to sample new physical parameters z_e, it is only to ensure that a sub-part of the encoder is able to infer correctly z_e given y_e. This augmentation does not contribute to robustness distribution shifts on y in contrast to ours.

4. Experiments

4.1. Problem description

We assess the benefits of expert augmentation on three controlled problems described and simulated by the ODE

$$\frac{dy_t}{dt} = F_e(y_t; z_e) + F_a(y_t; z_a), \tag{5}$$

where $F_e : \mathcal{Y}_t \times \mathcal{Z}_e \to \mathcal{Y}_t$ is the expert model and $F_a : \mathcal{Y}_t \times \mathcal{Z}_a \to \mathcal{Y}_t$ complements it. In our notation X is the initial state y_0 and the response Y is the sequence of states $y_{1:t_1} := [y_{i\Delta t}]_{i=1}^{t_1/\Delta t}$. For all experiments we train the models to maximize $p_{\theta,\psi}(y = y_{1:t_1}|x = y_0)$ on the training data. We validate and test the models on the predictive distribution $p(y = y_{1:t_2}|x = y_0, x_o = y_0, y_o = y_{1:t_1})$, where $t_2 > t_1$ assesses the generalization over time. A brief description of the different problems is provided below.

The damped pendulum is often used as an example in the hybrid modelling literature (Yin et al., 2021; Takeishi & Kalousis, 2021). 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 (5), where $z_e := \omega$, $z_a = \alpha$ and

$$F_e := \begin{bmatrix}\dot{\theta} & -\omega_0^2 \sin\theta\end{bmatrix}^T \quad \text{and} \quad F_a := \begin{bmatrix}0 & -\alpha\dot{\theta}\end{bmatrix}^T. \tag{6}$$

The corresponding systems are defined by the damping factor α and ω_0, the fundamental frequency of the pendulum.

The RLC series circuits are electrical circuits made of 3 electrical components that may model a large range of transfer functions. These models are often used in biology (e.g., the Hodgkin-Huxley class of models (Hodgkin & Huxley, 1952), in photoplethysmography (Crabtree & Smith, 2003)) and in electrical engineering to model the dynamics of various systems. The system's state at time t is $y_t = \begin{bmatrix}U_t & I_t\end{bmatrix}^T$, where U_t is the voltage around the capacitance and I_t the current in the circuit. The evolution of the state over time is described by (5), where $z_e := \{L, C\}$, $z_a = \{R\}$ and

$$F_e := \begin{bmatrix}\frac{I_t}{C} \\ \frac{1}{L}(V(t) - U_t)\end{bmatrix} \quad \text{and} \quad F_a := \begin{bmatrix}0 \\ -\frac{R}{C} I_t\end{bmatrix}. \tag{7}$$

The dynamics described by the RLC circuit is more diverse than for the pendulum and the system can be hard to identify. This system is characterised by the resistance R, capacitance C, and inductance L, provided $V(t)$ is known.

The 2D reaction diffusion was used by Yin et al. (2021) to assess the quality of APHYNITY. It is a 2D FitzHugh-Nagumo on a 32×32 grid. The system's state at time t is a $2 \times 32 \times 32$ tensor $y_t = \begin{bmatrix}u_t & v_t\end{bmatrix}^T$. The evolution of the state over time is described by (5), where $z_e := \{a, b\}$, $z_a = \{k\}$ and

$$F_e := \begin{bmatrix}a\Delta u_t \\ b\Delta v_t\end{bmatrix} \quad \text{and} \quad F_a := \begin{bmatrix}R_u(u_t, v_t; k) \\ R_v(u_t, v_t)\end{bmatrix}, \tag{8}$$

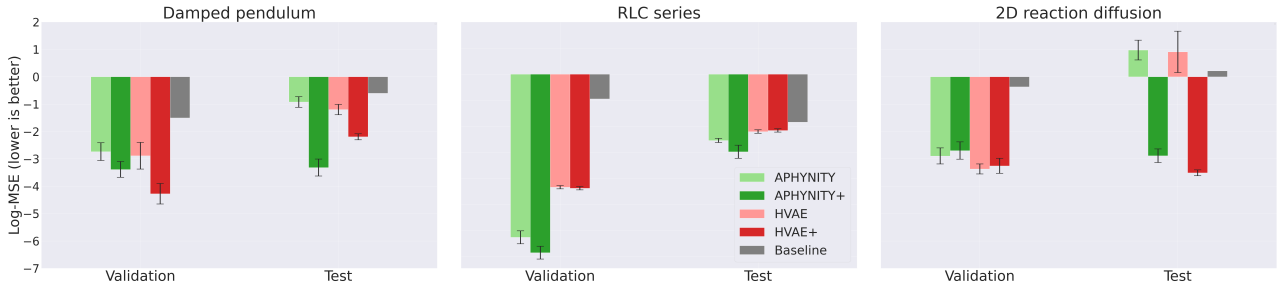

Figure 4. The average log-MSEs over 10 runs for three synthetic problems on the validation and test sets. We compare HVAE (in red) and APHYNITY (in green), in light colours, to their expert augmented versions HVAE+ and APHYNITY+, in darker colours. *On the test sets, AHMs outperform the original models, and by a large margin on the pendulum and diffusion problems. Moreover, augmentation conserves the relatively good performance on the validation set (IID w.r.t. the training set).*

where Δ is the Laplace operator, the local reaction terms are $R_u(u, v; k) = u - u^3 - k - v$ and $R_v(u, v) = u - v$. This model is interesting to study as it considers a state space for which neural architectures may have a real advantage compared to other ML models.

In the following experiments we analyze the effect of our data augmentation strategy on APHYNITY and HVAE. All models explicitly use the assumption that the interaction model follows the structure of (5). For each problem the validation and test sets are respectively IID and OOD with respect to the training distribution. The best models are always selected based on validation performance, that is with samples from Ω. We provide additional details on the different expert models, dataset creation, and neural networks architectures in Appendix B.

4.2. Results

Performance gain from augmentation. *This experiment demonstrates that HVAE and APHYNITY are not robust to OOD test scenarios in opposition to the corresponding AHMs, as shown in Figure 1* for the 2D diffusion problem and in Appendix C for the two other problems. We emphasize that our intention is not to declare a winner between HVAE and APHYNITY. Indeed, both algorithms have already demonstrated superior performance than black box ML models. Hence, we only report a very simple baseline that is the mean value of the signals. We want to compare performance in OOD settings and empirically validate the benefit of AHMs. We compare the predictive performance in Figure 4 (see Table 2 for the raw numbers). Although classical hybrid learning strategies do very well on the IID validation set, they exhibit poor generalization on OOD test sets for all three problems. We also observe some disparity between APHYNITY and HVAE. In addition to different learning strategies, this is probably due to differences in the networks' architectures as they were respectively inspired from the corresponding pendulum experiment in each paper. However, even if one method may outperform the other for some problems, they both benefit from our augmentation strategy (APHYNITY+, HVAE+). Overall, the effect of augmentation goes up to dividing the test error by a factor of $e^{4.6} \approx 100$ in some cases.

Dataset		APH.	HVAE	APH.+	HVAE+
Pendulum	Valid.	$6_{\pm 2}$	$3_{\pm 1}$	$6_{\pm 2}$	$2_{\pm 1}$
	Test	$66_{\pm 9}$	$117_{\pm 10}$	$10_{\pm 4}$	$11_{\pm 2}$
RLC	Valid.	$6_{\pm 3}$	$38_{\pm 2}$	$7_{\pm 5}$	$28_{\pm 1}$
	Test	$17_{\pm 3}$	$25_{\pm 2}$	$5_{\pm 2}$	$12_{\pm 1}$
Diffusion	Valid.	$2_{\pm 0}$	$2_{\pm 0}$	$2_{\pm 0}$	$2_{\pm 0}$
	Test	$27_{\pm 2}$	$32_{\pm 10}$	$3_{\pm 1}$	$2_{\pm 0}$

Table 1. Comparison of mean relative precision (in %, $\pm$ indicates one standard deviation) over 10 runs of predicted physical parameters of different hybrid modelling strategies in validation and OOD test settings. Augmented versions are denoted with a $+$. *While the accuracy of APHYNITY and HVAE is good on the validation set, it collapses on the OOD test set. On the opposite, the augmented versions perform well on both validation and test sets.*

Stability for non-exact models. The empirical results from Figure 4 are very important as they show that even when the decoder is not Ω-*exact* (and hence not $\overset{+}{\Omega}$-*exact*), augmentation is still useful. In particular, Table 1 shows that the encoder does not predict the physical parameters perfectly. This indicates that the encoder is not Ω-*exact* and neither should be the decoder. This table shows the relative error on the physical parameters computed as $\sum_{i=1}^{k} \frac{1}{k} \left| \frac{z_e^i - \mu_\theta^i}{z_e^i} \right|$, where μ_θ^i is the estimated most likely value of the i^{th} component of the physical parameters. We first notice that APHYNITY and HVAE perform differently and their performance depends on the specific problem. While APHYNITY accurately estimates the physical parameters on the IID validation set for the 3 problems, HVAE's performance are mixed on the RLC problem as it makes prediction that are 38% away from the nominal parameter value on average whereas APHYNITY reduces this error to 6%. Interestingly, we observe that the proposed augmentation strategies improve the encoder such that it accurately estimates the physical parameters also on the OOD test set even for HVAE on the RLC problem. This confirms that the augmentation strategy is helpful even when the hybrid

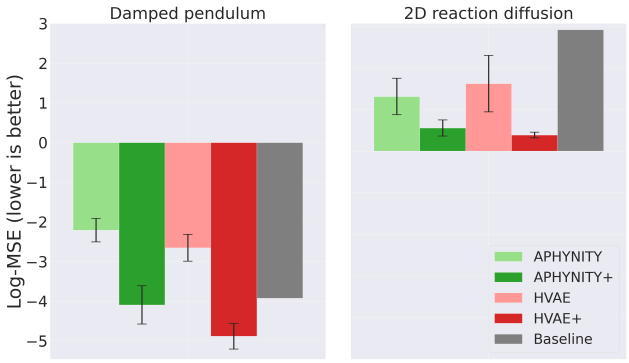

Figure 5. The average log-MSEs over 10 runs for the *damped pendulum* and *2D reaction diffusion* problems on a test distribution for which z_a, in addition to z_e, is also shifted. *AHM achieves better peformance than stand HyL algorithms even when the test distribution support z_a differs from the training.*

model is not Ω-*exact*. As a conclusion, augmented hybrid learning outperforms classical hybrid learning both on the predictive accuracy and at inferring the expert variables.

Effect of out of expertise shift. *This experiment shows that our augmentation strategy may remain beneficial even when the train and test supports of z_a are not identical.* This scenario corresponds to samples (x, y) generated by $(z_a, z_e) \in (\overset{*}{\mathcal{Z}}_a \setminus \mathcal{Z}_a) \times \tilde{\mathcal{Z}}_e$ depicted by the violet domains in Figure 3. In Figure 5 we observe the log-MSE of augmented and non-augmented hybrid models trained for $(z_a, z_e) \in \mathcal{Z}_a \times \mathcal{Z}_e$ on test data that are generated with $(z_a, z_e) \in \tilde{\mathcal{Z}}_a \times \tilde{\mathcal{Z}}_e$. For the pendulum, the support over $z_a = \alpha$ is $[0, 0.3]$ in train and $[0.3, 0.6]$ in test; For the 2D reaction diffusion, $z_a = k$ is $[0.003, 0.005]$ in train and $[0.005, 0.008]$ in test. We observe that augmented models outperform the original models by a large margin. These results suggest that augmentation could be very valuable in practice, even when the distribution shift is also caused by non expert variables. However, if the shift on z_a becomes the dominant effect, augmented models also eventually becomes vulnerable to shifts on z_e as demonstrated by supplementary experiments in Appendix B.

5. Related work

5.1. Hybrid modelling

Hybrid Learning (HyL), or gray box modelling as called in its early days in the 90's (Psichogios & Ungar, 1992; Rico-Martinez et al., 1994; Thompson & Kramer, 1994; Rivera-Sampayo & Vélez-Reyes, 2001; Braun & Chaturvedi, 2002), has been an appropriate method to learn models that are both expressive and interpretable, while also allowing them to be learnt on fewer data. The interest for HyL (Mehta et al., 2020; Lei & Mirams, 2021; Reichstein et al., 2019; Saha et al., 2020; Guen & Thome, 2020; Levine & Stuart, 2021; Espeholt et al., 2021) has greatly renewed since the outbreak of recent neural network architectures that simplify the combination of physical equations within ML models. As an example, Neural ODE (Chen et al., 2018) and convolutional neural networks (LeCun et al., 1995, CNN) are privileged architectures to work with dynamical systems described by ODEs or PDEs. While most of the HyL's literature focus on the predictive performance of hybrid models, recent work have also showed that HyL may help to infer the physical parameter accurately (Yin et al., 2021; Takeishi & Kalousis, 2021). This is aligned with Zyla et al. (2020) (see Section 40.2.2.2) which observe that inference on incomplete models results in a *systematic bias*. Similar to HyL, they extend the model with *nuisance* parameters in order to improve its fidelity, and to reduce the systematic bias.

In this work, we decided to study Yin et al. (2021) and Takeishi & Kalousis (2021) for two reasons that distinguish them from the rest of the HyL literature. First, these are notable examples of HyL algorithms that can be applied to a broad class of problems in contrast to papers that focus on specific applications (Lei & Mirams, 2021; Reichstein et al., 2019). Second, those methods also learn a reliable inference model for the physical parameters, suggesting that the expert model is used properly in the generative model, which is a key assumption for our augmentation. While Takeishi & Kalousis (2021) claim to achieve robustness with HyL, we argue that this statement is incomplete as HVAE fails in OOD settings. In particular, their approach is only able to generalize with respect to unseen time or initial state if the model correctly identifies the latent variables z_a, z_e.

5.2. Combining hybrid modelling and data augmentation

Close to our idea is the one proposed in Shrivastava et al. (2017) where they train a GAN model that improves the realism of a simulated image while conserving its semantic content (e.g. eyes colour) as modeled by the simulation parameters. The generated data with their annotations may then be used for a downstream task, such as inferring the properties of real images that corresponds to simulation parameters. The GAN objective from Shrivastava et al. (2017) requires that the two distributions induced by the semantic content of real and simulated data are identical. On the opposite, we consider training data that corresponds to expert parameters with limited diversity, and overcome this scarcity with expert augmentation. Another line of work similar to ours is Sim2Real, which considers the task of transferring a model trained on simulated data to real world (Doersch & Zisserman, 2019; Sadeghi et al., 2018; 2017). Robust HyL, as a way to enhance simulations, could be used for Sim2Real.

5.3. Robust ML and Invariant Learning

Various statistical methods have been introduced to ensure models generalize under distribution shift. Domain-adversarial objectives aimed at learning (conditionally) invariant predictors (Ganin et al., 2016; Zhang et al., 2017; Li et al., 2018), GroupDRO (Sagawa et al., 2019) optimizing for worst-case loss over multiple domains and IRM (Arjovsky et al., 2019) as well as sub-group calibration (Wald et al., 2021) aiming to satisfy calibration or sufficiency constraints to learn features invariant across domains. Extensions, able to infer domain labels from training data have been proposed as well (Lahoti et al., 2020; Creager et al., 2021), partially inspired by fairness objectives (Hébert-Johnson et al., 2018; Kim et al., 2019). In contrast to AHM, all of these methods rely on the variation of interest being present in the training data.

6. Discussion

We now examine the assumptions we made to derive our augmentation strategy and discuss potential limitations.

Erroneous interaction model. The exactness of the hybrid component $p_\theta(y|x, y_e, z_a)$ is a critical assumption underlying our expert-based augmentation strategy. Unfortunately, this component is learned from training data only, hence we cannot prove its exactness on the test domain, which corresponds to a different domain $\mathcal{Y}_e$. However, we argue that soft assumptions on the class of interaction model may alleviate this problem. As an example, when we consider an additive hybrid model, as in APHYNITY (Yin et al., 2021), and embed this hypothesis into the interaction model, generalization to unseen y_e follows. When this assumption is too strong, we could still expect generalization of $p_\theta(y|x, y_e, z_a)$ because HyL drives y samples from p_θ to be close to y_e. It implies that the corresponding function approximator is smooth, which helps generalization to unseen scenarios. This contrasts with the encoder q_ψ for which a a good inductive bias usually is not available.

Diagnostic. While crucial, we cannot guarantee the exactness of the decoder p_θ in general because we only evaluate the encoder and the decoder jointly on data points (x, y, x_o, y_o). However, in some cases we can detect model misspecification by observing that the predictive model $p_{\theta,\psi}(y|x, x_o, y_o)$ is imperfect. Making this observation is not always simple as it requires prior knowledge on the expected accuracy of an exact model. However, when the system is deterministically identifiable, we may argue that the accuracy should be only limited by the intrinsic noise between x and y given z_a and z_e.

Relaxing exactness. Even with a strong inductive bias on the decoder, achieving exactness is hopeless in practical settings. However, our experiments demonstrate that expert-augmentation works in practice. We can explain this by taking a look at Figure 3. If the generative model that maps x and (z_a, z_e) is incorrect, the mapping from $\mathcal{Z}_a$ and $\mathcal{Z}_e$ could be slightly off from $\overset{+}{\Omega}$. However, this does not preclude the set of augmented samples to be closer to $\overset{+}{\Omega}$ than Ω and to induce a better predictive model on $\overset{+}{\Omega}$ than the original model trained only on Ω.

Limitations We considered expert models that are parameterized by a small number of parameters, which can be covered densely via sampling. Covering densely a higher dimensional parameter space with the augmentation strategy becomes quickly impossible, hence a smarter sampling strategy would be required, such as worst-case sampling. Another difficulty is to choose a plausible range of parameters that contains both the train and the test support, this will often require a human expert in the loop. Finally, we assume that the train distribution of z_a should be representative of the test distribution, we empirically observed that a softer version of this assumption could be enough. However, performance will eventually decline as the support of the test distribution for z_a is far from the training domain.

7. Conclusion

In this work, we describe HyL with a probabilistic model in which one component of the latent process, denoted the expert model, is known. In this context, we establish that state-of-the-art HyL algorithms are vulnerable to distribution shifts even when the expert model is well defined for such configurations. Grounded in this formalisation, we derive that expert augmentations induce robustness to OOD settings. We discuss how our assumptions can transfer to real-world settings and describe how to diagnose potential shortcomings. Finally, empirical evidence asserts that expert augmentations may be beneficial even when one of our assumptions on the class of distribution shift is violated.

Our augmentation is applicable to a large class of hybrid models, hence it should benefit from future progress in HyL. Thus, we believe research in HyL and formally defining its targeted objectives is an important direction for further improving the robustness of hybrid models. As an example, the minimal description length principle (Grünwald, 2007) could be a great resource to investigate the balance between the model's capacity and robustness. Finally, robust ML models must eventually translate to real-world applications, hence a next step would be to apply AHMs to real-wold data. Paving the way to future research combining AHM with robust ML methods.

Acknowledgements

We would like to acknowledge Andy Miller, Dan Busbridge, Jason Ramapuram, Joe Futoma, and Mark Goldstein for providing useful feedback on this manuscript or an earlier version of it.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Braun, J. E. and Chaturvedi, N. An inverse gray-box model for transient building load prediction. *HVAC&R Research*, 8(1):73–99, 2002.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 6572–6583, 2018.

Crabtree, V. P. and Smith, P. R. Physiological models of the human vasculature and photoplethysmography. *Electronic Systems and Control Division Research, Department of Electronic and Electrical Engineering, Loughborough University*, pp. 60–63, 2003.

Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., and Ho, S. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

Doersch, C. and Zisserman, A. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *Advances in Neural Information Processing Systems*, 32: 12949–12961, 2019.

Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C., Gazen, C., Hickey, J., Bell, A., and Kalchbrenner, N. Skillful twelve hour precipitation forecasts using large context neural networks. *arXiv preprint arXiv:2111.07470*, 2021.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Grünwald, P. D. *The minimum description length principle*. MIT press, 2007.

Guen, V. L. and Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11474–11484, 2020.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Hodgkin, A. L. and Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117 (4):500–544, 1952.

Keriven, N. and Peyré, G. Universal invariant and equivariant graph neural networks. *Advances in Neural Information Processing Systems*, 32:7092–7101, 2019.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. H. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

LeCun, Y., Bengio, Y., et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Lei, C. L. and Mirams, G. R. Neural network differential equations for ion channel modelling. *Frontiers in Physiology*, pp. 1166, 2021.

Levine, M. E. and Stuart, A. M. A framework for machine learning of model error in dynamical systems. *arXiv preprint arXiv:2107.06658*, 2021.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018.

Mahmood, O., Mansimov, E., Bonneau, R., and Cho, K. Masked graph modeling for molecule generation. *Nature communications*, 12(1):1–12, 2021.

Mehta, V., Char, I., Neiswanger, W., Chung, Y., Nelson, A. O., Boyer, M. D., Kolemen, E., and Schneider, J. Neural dynamical systems: Balancing structure and flexibility in physical prediction. *arXiv preprint arXiv:2006.12682*, 2020.

Psichogios, D. C. and Ungar, L. H. A hybrid neural network-first principles approach to process modeling. *AIChE Journal*, 38(10):1499–1511, 1992.

Qian, Z., Zame, W. R., van der Schaar, M., Fleuren, L. M., and Elbers, P. Integrating expert odes into neural odes: Pharmacology and disease progression. *arXiv preprint arXiv:2106.02875*, 2021.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

Rico-Martinez, R., Anderson, J., and Kevrekidis, I. Continuous-time nonlinear signal processing: a neural network based approach for gray box identification. In *Proceedings of IEEE Workshop on Neural Networks for Signal Processing*, pp. 596–605. IEEE, 1994.

Rivera-Sampayo, R. and Vélez-Reyes, M. Gray-box modeling of electric drive systems using neural networks. In *Proceedings of the 2001 IEEE International Conference on Control Applications (CCA'01)(Cat. No. 01CH37204)*, pp. 146–151. IEEE, 2001.

Sadeghi, F., Toshev, A., Jang, E., and Levine, S. Sim2real view invariant visual servoing by recurrent control. *arXiv preprint arXiv:1712.07642*, 2017.

Sadeghi, F., Toshev, A., Jang, E., and Levine, S. Sim2real viewpoint invariant visual servoing by recurrent control. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4691–4699, 2018.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2019.

Saha, P., Dash, S., and Mukhopadhyay, S. Physics-incorporated convolutional recurrent neural networks for source identification and forecasting of dynamical systems. *arXiv preprint arXiv:2004.06243*, 2020.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2107–2116, 2017.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Takeishi, N. and Kalousis, A. Physics-integrated variational autoencoders for robust and interpretable generative modeling. *Advances in Neural Information Processing Systems*, 34, 2021.

Thompson, M. L. and Kramer, M. A. Modeling chemical processes using prior knowledge and neural networks. *AIChE Journal*, 40(8):1328–1340, 1994.

Wald, Y., Feder, A., Greenfeld, D., and Shalit, U. On calibration and out-of-domain generalization. *arXiv preprint arXiv:2102.10395*, 2021.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Yin, Y., Le Guen, V., Dona, J., de Bézenac, E., Ayed, I., Thome, N., and Gallinari, P. Augmenting physical models with deep networks for complex dynamics forecasting. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124012, 2021.

Zhang, Y., Barzilay, R., and Jaakkola, T. Aspect-augmented adversarial networks for domain adaptation. *Transactions of the Association for Computational Linguistics*, 5:515–528, 2017.

Zyla, P. et al. Review of Particle Physics. *PTEP*, 2020(8): 083C01, 2020. doi: 10.1093/ptep/ptaa104.

A. Additional description of expert augmentation

We provide the procedure to do expert augmentation for robust HyL as the sequence of steps below.

1. Train both the encoder $q_\psi(z_a, z_e|x, y)$ and the interaction model $p_\theta(y|x_o, z_a, y_e)$ with a HyL algorithm, by minimizing the corresponding loss $\mathcal{L}(\psi, \theta) = \mathbb{E}_{\mathcal{D}}[\ell(x, y; \theta, \psi)]$ on the training set $\mathcal{D}$;
2. Decide on an augmented distribution $\overset{+}{p}(z_e)$ for z_e that contains both train and test scenarios;
3. Reproduce the following steps to generate a dataset $\overset{+}{\mathcal{D}}$ of observations and expert variables $(x, y, z_e) \sim \mathbb{E}_{p(z_a)p(y_e|z_e,x,y)}[p(z_e)p(x)p_\theta(y|y_e, z_a, x)]$:
   (a) Sample (x_o, y_o) from the data;
   (b) Sample z_a from the posterior $q_\psi(z_a|x_o, y_o)$;
   (c) Sample z_e from $\overset{+}{p}(z_e)$;
   (d) Push forward x, z_a and z_e in the generative model as $y_e \sim p(y_e|x_o, z_e)$ and $y \sim p_\theta(y|x_o, z_a, y_e)$;
   (e) Add the triplet (x_o, y, z_e) to the augmented training set $\overset{+}{\mathcal{D}}$.
4. Freeze the interaction model, and fine-tune the encoder $q_\psi(z_a, z_e|x, y)$ on the augmented dataset $\overset{+}{\mathcal{D}}$ by minimizing $\overset{+}{\mathcal{L}}(\psi, \theta) = \mathbb{E}_{\overset{+}{\mathcal{D}}}[\ell(x, y; \theta, \psi) - \log q_\psi(z_e|x, y)]$.

B. Additional details on experiments

B.1. Damped pendulum

Datasets. We use Neural Ordinary Differential Equations (NODE) (Chen et al., 2018) to solve the ODE ruling the damped pendulum. Each sample is simulated for $t_0 = 0s$, $t_1 = 5s$, and $t_2 = 20s$, with a time resolution equal to 0.1 second. The models are trained with only the realizations between t_0 and t_1. At test and validation time, the pair $(x_o, y_o) = (y_0, [y_{i\Delta t}]_{i=1}^{t_1/\Delta t})$, $x = y_{t_1}$ and the model predicts $y = [y_{i\Delta t}]_{i=t_2/\Delta t+1}^{t_2/\Delta t}$. The initial angular speed is always 0 and $\theta_0 \sim \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$.

The training set is made of 1000 samples and the validation set of 100 samples. They are both generated by sampling uniformly $z_a := \alpha$ from $\mathcal{Z}_a := [0, 0.6]$ and $z_e := \omega_0$ from $\mathcal{Z}_e := [1.5, 3.1]$. The shifted test set contains 100 samples generated by sampling uniformly z_a in $\mathcal{Z}_a$ and z_e in $\tilde{\mathcal{Z}}_e := [0.5, 1.5]$.

APHYNITY. Our model is composed of a 1-layer RNN with 128 units that encodes the input signal $y_{0:t_1}$ as $h(y_{0:t_1}) \in \mathbb{R}^{128}$. An MLP with 3 layers of 150 units and ReLU activations maps h to $\mathbb{R}_+$ to predict ω_0. The function $f_a : \mathbb{R}^{128} \times \mathbb{R}^2$ is an MLP with 3 layers of 50 units and ReLU activations (no activation for the last layer). The models are trained for 50 epochs with Adam with no weight decay and a learning rate equal to 0.0005. For the Lagrangian optimization we use $N_{iter} = 5, \lambda_0 = 10, \tau_2 = 5$ (see (Yin et al., 2021). The augmented data are generated by sampling uniformly $z_e \in \overset{+}{\mathcal{Z}}_e := [0.5, 3.5]$ and z_a from the marginal predictive prediction of the model, that is we use the training dataset to infer values of z_a and use these as samples. The batch size is 100.

HVAE. We use the notations from Takeishi & Kalousis (2021) to describe the architecture of the VAE. The network $g_{p,1} : \mathbb{R}^2 \times \mathbb{R}^{d_a}$, where $d_a = 1$ is the size of the latent space for the interaction model, is supposed to filter the observations so that they can be generated by the expert model. It has 2 hidden layers with 128 units, $g_{p,2}$ is an MLP with the following hidden layers $[128, 128, 256, 64, 32]$ and takes the full sequence of filtered states and predicts the mean and variance of a normal distribution that parameterize the posterior $p_\theta(z_e|x, y, z_a)$. Another network, g_a takes the sequence of observations and predict the posterior distribution of z_a as a normal distribution. This network has the following hidden layers $[256, 256, 128, 32]$. All networks have SeLU activations. In general the decoder of HVAE can be anything that combines the expert model in order to produce samples in the observation space, as we made the hypothesis that the ODE is just missing an additive term, the decoder is a NODE where the function is the sum of f_e and f_a a two hidden layers MLP with 64 units and SeLU activation (except for the last layer that has no activation). The likelihood model is also Gaussian

with the mean being predicted by the NODE and the variance learned but shared for all observations. For additional details on our architecture and implementation details we encourage the interested reader to check our code.

The networks are trained jointly for 1000 epochs with Adam optimizer, with a learning rate equal to 0.0005, weight decay equal to 0.000001 and batch size 200. The other parameters are set to $\gamma = 1$, $\alpha = 0.01$ and $\beta = 0.01$. The HVAE also relies on some augmentation during training and in order to compare fairly our model to theirs we use the same distribution for our augmentation and theirs that is $z_a \sim \mathcal{N}(0, I)$ and $z_e \sim \mathcal{U}(0.5, 3.5)$.

B.2. RLC series

Datasets. Similar to the damped pendulum, we use NODE to solve the ODE ruling the RLC circuit. Each sample is simulated for $t_0 = 0s$, $t_1 = 5s$, and $t_2 = 20s$, with a time resolution equal to 0.1 second. The models are trained with only the realizations between t_0 and t_1. At test and validation time, the pair $(x_o, y_o) = (y_0, [y_{i\Delta t}]_{i=1}^{t_1/\Delta t})$, $x = y_{t_1}$ and the model predicts $y = [y_{i\Delta t}]_{i=t_2/\Delta t+1}^{t_2/\Delta t}$. In all experiments, the initial value for $U_0 \sim \mathcal{N}(0, 1)$ and $I_0 = 0$, the voltage source delivers a AC + DC tension $V(t) = 2.5\sin(4\pi t) + 1$.

The training set is made of 2000 samples and the validation set of 100 samples. They are both generated by sampling uniformly $z_a := R$ from $\mathcal{Z}_a := [1, 3]$ and $z_e := [L, C]$ from $\mathcal{Z}_e := [1, 3] \times [0.5, 1.5]$. The shifted test set contains 100 samples and is generated by sampling uniformly z_a in $\mathcal{Z}_a$ and z_e in $\tilde{\mathcal{Z}}_e := [3, 5] \times [1., 2.5]$.

APHYNITY. Our model is composed of a 1-layer RNN with 128 units that encodes the input signal $y_{0:t_1}$ as $h(y_{0:t_1}) \in \mathbb{R}^{128}$. An MLP with 3 layers of 200 units and ReLU activations maps h to $\mathbb{R}_+^2$ that predicts L and C. The function $f_a : \mathbb{R}^{128} \times \mathbb{R}^2$ is an MLP with 3 layers of 150 units and ReLU activations (no activation for the last layer). The models are trained for 50 epochs with Adam with no weight decay and a learning rate equal to 0.0005. For the Lagrangian optimization we use $N_{iter} = 5, \lambda_0 = 10, \tau_2 = 5$ (see (Yin et al., 2021)). The augmented data are generated by sampling uniformly $z_e \in \overset{+}{\mathcal{Z}}_e := [1, 5] \times [0.5, 2.5]$ and z_a from the marginal predictive prediction of the model, that is we use the training dataset to infer values of z_a and use these as samples. The batch size is 100.

HVAE. We use the same networks' architectures than for the damped pendulum experiment. Except that $g_{p,1}$ is has 3 hidden layers with 100 units.

The networks are trained jointly for 1000 epochs with Adam optimizer, with a learning rate equal to 0.0005, weight decay equal to 0.000001 and batch size 100. The other parameters are set to $\gamma = 1$, $\alpha = 0.01$ and $\beta = 0.01$. The HVAE also relies on some augmentation during training and in order to compare fairly our model to theirs we use the same distribution for our augmentation and theirs that is $z_a \sim \mathcal{N}(0, I)$ and $z_e \sim \mathcal{U}(1, 5) \times \mathcal{U}(0.5, 2.5)$.

B.3. 2D reaction diffusion

Datasets. Similar to the damped pendulum, we use NODE to solve the PDE ruling the reaction diffusion. We closely follow the experimental setting described in Yin et al. (2021) and approximate the Laplace operator with a 3×3 discrete version of the operator. Each sample is simulated for $t_0 = 0s$, $t_1 = 1s$, and $t_2 = 5s$, with a time resolution equal to 0.1 second. The models are trained with only the realizations between t_0 and t_1. At test and validation time, the pair $(x_o, y_o) = (y_0, [y_{i\Delta t}]_{i=1}^{t_1/\Delta t})$, $x = y_{t_1}$ and the model predicts $y = [y_{i\Delta t}]_{i=t_2/\Delta t+1}^{t_2/\Delta t}$. The initial state is sampled from a uniform distribution in $[0, 1]$.

The training set is made of 2000 samples and the validation set of 100 samples. They are both generated by sampling uniformly $z_a := k$ from $\mathcal{Z}_a := [0.003, 0.005]$ and $z_e := [a, b]$ from $\mathcal{Z}_e := [0.001, 0.002] \times [0.003, 0.007]$. The shifted test set contains 100 samples and is generated by sampling uniformly z_a in $\mathcal{Z}_a$ and z_e in $\tilde{\mathcal{Z}}_e := [0.002, 0.004] \times [0.001, 0.1]$.

APHYNITY. Our model is composed of a deep CNN that encodes the input sequence of 10 images. The exact architecture can be found in the code. The dimension of z_a is equal to 10. Similarly to Yin et al. (2021) the function f_a is a 3-layers CNN with ReLU activations. The models are trained for 500 epochs with Adam with no weight decay and a learning rate equal to 0.0005. For the Lagrangian optimization we use $N_{iter} = 1, \lambda_0 = 10, \tau_2 = 5.$. The augmented data are generated by sampling uniformly $z_e \in \overset{+}{\mathcal{Z}}_e := [0.001, 0.004] \times [0.001, 0.01]$ and z_a from the marginal predictive prediction of the model, that is we use the training dataset to infer values of z_a and use these as samples. The batch size is 100.

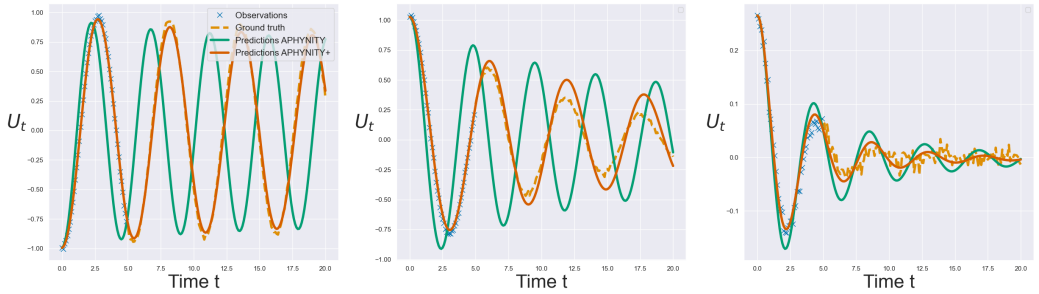

Figure 6. Comparison of the predictions made by APHYNITY and APHYNITY+ on the damped pendulum problem for 3 diverse test examples. It is important to mention that the support of the test distribution is disjoint from the training support. *We clearly observe the beneficial effect of augmentation which lead to more accurate predictions.*

B.3.1. HVAE

We use the notations from Takeishi & Kalousis (2021) to describe the architecture of the VAE. The network $g_{p,1} : \mathbb{R}^{2\times32\times32} \times \mathbb{R}^{d_a}$ is a conditional U-net, where $d_a = 10$ is the size of the latent space for the interaction model, is supposed to filter the observation so that they can be generated by the expert model. The networks $g_{p,1}$ and g_a share a common backbone CNN and are, in addition, respectively parameterized by 2 3-layers MLPs. All networks have ReLU activations. In general the decoder of HVAE can be anything that combines the expert model in order to produce samples in the observation space, as we made the hypothesis that the ODE is just missing an additive term, the decoder is a NODE where the function is the sum of f_e and f_a a 3-layers CNN. The likelihood model is also Gaussian with the mean being predicted by the NODE and the variance learned but shared for all observations. For additional details on our architecture and implementation details we encourage the interested reader to check our code.

The networks are trained jointly for 1000 epochs with Adam optimizer, with a learning rate equal to 0.0005, weight decay equal to 0.00001 and batch size 100. The other parameters are set to $\gamma = 1$, $\alpha = 0.01$ and $\beta = 0.01$. The HVAE also relies on some augmentation during training and in order to compare fairly our model to theirs we use the same distribution for our augmentation and theirs that is $z_a \sim \mathcal{N}(0, I)$ and $z_e \sim \mathcal{U}(0.001, 0.004) \times \mathcal{U}(0.001, 0.01)$.

C. Supplementary results

We now provide additional results for AHM versus standard HyL models.

C.1. Log-mses on the 3 synthetic problems

Dataset		APH.	HVAE	APH.+	HVAE+
Pendulum	Val.	$-2.7_{\pm0.3}$	$-2.9_{\pm0.5}$	$-3.4_{\pm0.3}$	$-2.9_{\pm0.6}$
	Test	$-0.9_{\pm0.2}$	$-1.2_{\pm0.2}$	$-3.3_{\pm0.3}$	$-3.1_{\pm0.3}$
RLC	Val.	$-6.3_{\pm0.2}$	$-4.3_{\pm0.1}$	$-6.8_{\pm0.2}$	$-3.8_{\pm1.5}$
	Test	$-2.5_{\pm0.1}$	$-2.2_{\pm0.1}$	$-3.0_{\pm0.3}$	$-2.1_{\pm0.3}$
Diffusion	Val.	$-2.9_{\pm0.3}$	$-3.4_{\pm0.2}$	$-2.7_{\pm0.3}$	$-3.3_{\pm0.3}$
	Test	$1.0_{\pm0.4}$	$0.9_{\pm0.8}$	$-2.9_{\pm0.2}$	$-3.5_{\pm0.1}$

Table 2. Comparison of the log-mse of different hybrid modelling strategies in validation and OOD test settings. *Except on RLC, AHMs always outperform the corresponding HyL models on the test sets. Good performance on the validation set are conserved with augmentation.*

C.2. Distribution shift visualization

Similar to Figure 1, Figure 6 and Figure 7 showcase the behaviour of APHYNITY and APHYNITY+ for OOD test samples.

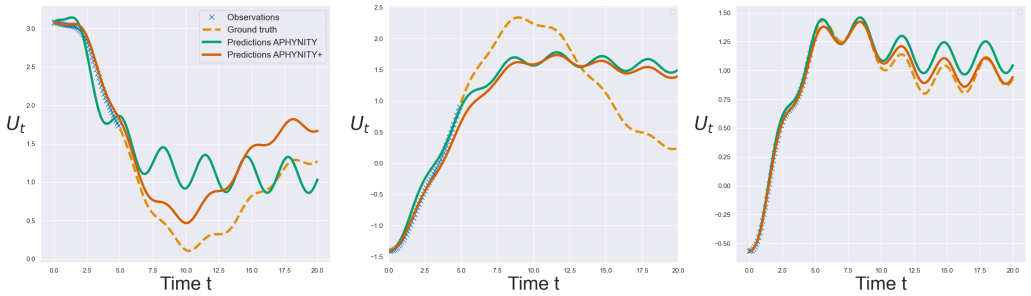

Figure 7. Comparison of the predictions made by APHYNITY and APHYNITY+ on the RLC series problem for 3 diverse test examples. It is important to mention that the support of the test distribution is disjoint from the training support. *We can perceive the beneficial effect of augmentation which lead to more accurate predictions in some cases. However both models are inaccurate. This indicates that the RLC series parameters are not easily identifiable, hence the generative model is not exact and augmentation is not as useful as for the diffusion and the pendulum.*

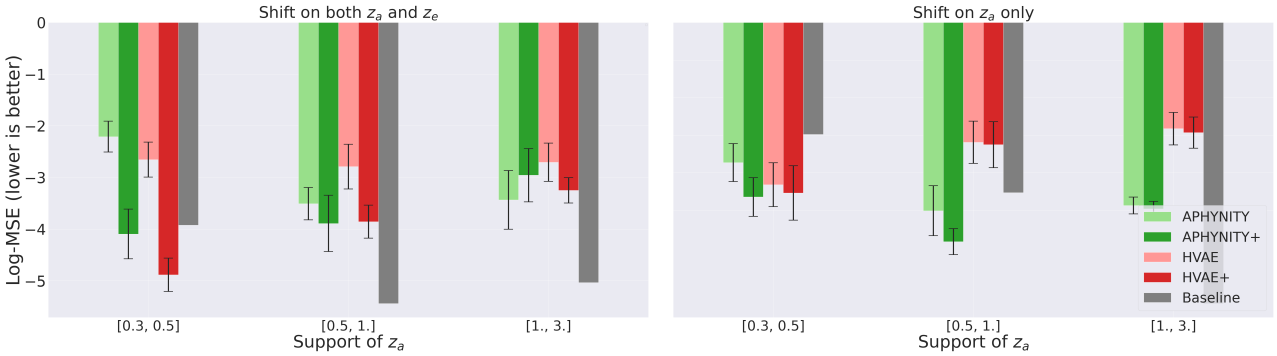

Figure 8. **Damped pendulum.** Effect of a distribution shift on the latent variable z_a of the interaction model. When the shift of z_a is reasonable (less than 1), the augmented models outperforms standard HyL even when the shift is only on z_a.

C.3. On the effect of out of expertise shift

The additional results in Figure 8, Figure 9 and Figure 10 demonstrate that our augmentations is mostly always beneficial. Although the benefit of augmentation decreases with the gap between the support of the distributions of z_a and train and test times, it still performs either better or on par with non-augmented HyL models.

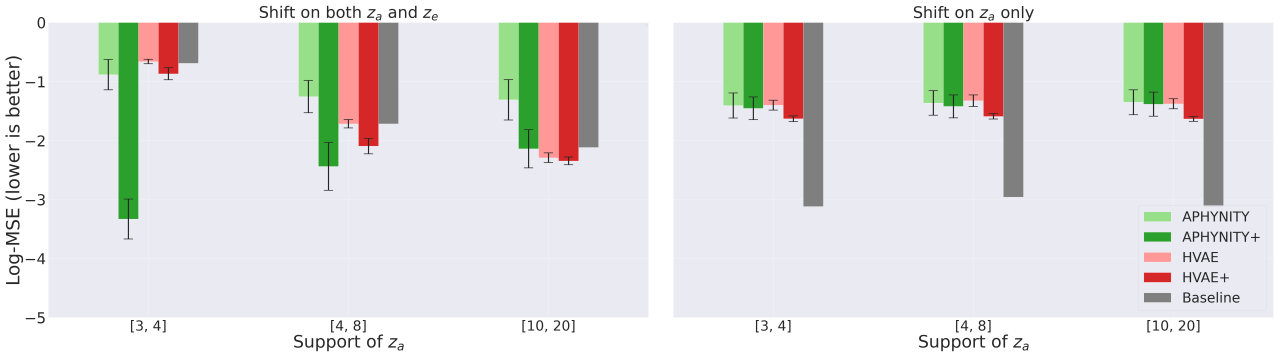

Figure 9. **RLC series.** Effect of a distribution shift on the latent variable z_a of the interaction model. We observe that augmentation is always beneficial, even when the shift is only on z_a. As the dynamics of the RLC series systems depends on the values of all 3 parameters R, L, C, we observe that some distribution shift can even lead to improved performance for the augmented models as for APHYNITY+ when $R \in [3, 4]$

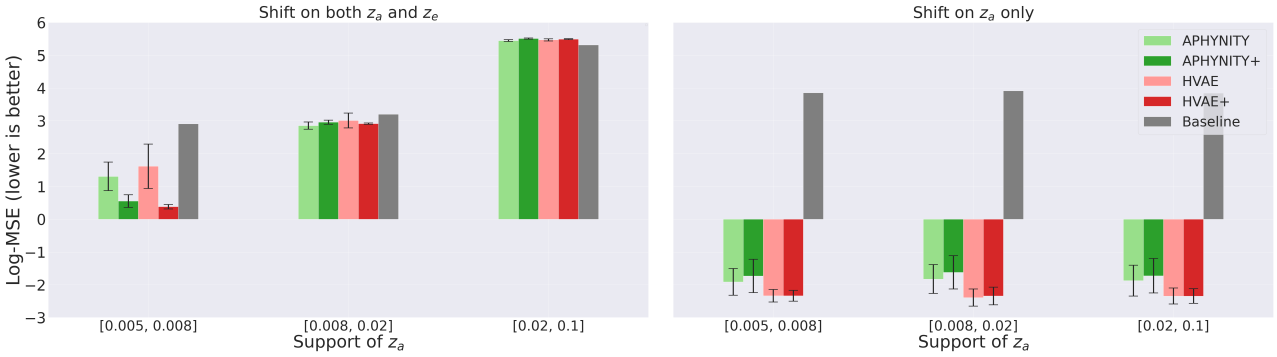

Figure 10. **2D diffusion reaction.** Effect of a distribution shift on the latent variable z_a of the interaction model. When the shift of z_a is reasonable ($k < 0.008$), the augmented models outperforms standard HyL even when the shift is only on z_a.