How Does Knowledge Injection Help in Informed Machine Learning?

Laura von Rueden[1,2], Jochen Garcke[1,3], Christian Bauckhage[1,2]
[1]*University of Bonn,* [2]*Fraunhofer IAIS,* [3]*Fraunhofer SCAI*
Sankt Augustin, Germany
laura.von.rueden@iais.fraunhofer.de

***Abstract*—Informed machine learning describes the injection of prior knowledge into learning systems. It can help to improve generalization, especially when training data is scarce. However, the field is so application-driven that general analyses about the effect of knowledge injection are rare. This makes it difficult to transfer existing approaches to new applications, or to estimate potential improvements. Therefore, in this paper, we present a framework for quantifying the value of prior knowledge in informed machine learning. Our main contributions are threefold. Firstly, we propose a set of relevant metrics for quantifying the benefits of knowledge injection, comprising in-distribution accuracy, out-of-distribution robustness, and knowledge conformity. We also introduce a metric that combines performance improvement and data reduction. Secondly, we present a theoretical framework that represents prior knowledge in a function space and relates it to data representations and a trained model. This suggests that the distances between knowledge and data influence potential model improvements. Thirdly, we perform a systematic experimental study with controllable toy problems. All in all, this helps to find general answers to the question how knowledge injection helps in informed machine learning.**

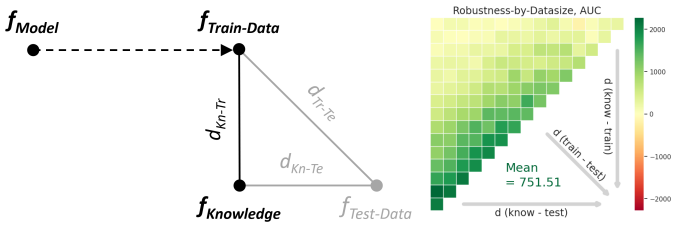
***Index Terms*—Hybrid AI, Informed Machine Learning, Prior Knowledge Injection, Neural Networks**

I. Introduction

Hybrid AI combines data-driven and knowledge-based models [1]–[3]. A particular approach that recently gained a lot of popularity is informed machine learning, which describes the injection of additional prior knowledge into learning systems [4]. This can help to improve model performance, especially when relevant training data is scarce [5]–[8]. Other potential benefits are that it can increase model robustness [9]–[11], help to ensure knowledge conformity [12]–[14], or can even improve explainability [15].

There are many applications where informed machine learning is successfully used – especially in scientific and engineering domains, where data acquisition can be expensive, but lots of prior knowledge is available. Just to give a few examples: In neural networks for climate prediction, physical laws are injected via knowledge-based loss functions [5]. In robotics, simulations are used as an additional source for training data [8]. Or in autonomous driving, spatial prototypes are employed to improve object detection [11].

However, the field is so application-driven that it has led to the development of many and rather specific approaches. In contrast, general analyses about informed machine learning are still missing [4]. This makes it difficult to transfer existing approaches to new applications, or to estimate potential improvements in advance.



(a) Function Space Illustration with Representations of Data and Prior Knowledge. (b) Model Improvement using Prior Knowledge.

Fig. 1: In Informed Machine Learning, prior knowledge is integrated into data-based learning [4], [16]. To better understand its effect, we propose a framework that represents knowledge in a function space (See a). We analyze how the distances between knowledge, train, and test data influence the potential model improvements. E.g., we find that informed learning greatly improves model robustness, especially when the knowledge is close to out-of-distribution test data (See b).

Therefore, in this paper our objective is to find general answers to the research question of how knowledge injection via informed machine learning does help. We further subdivide this into the following subquestions:

- How can knowledge injection improve machine learning?
- What are the requirements for the injected knowledge?
- How should the knowledge be injected?

Our approach is to develop a framework for quantifying the value of prior knowledge in informed machine learning. For this, we first define a set of metrics that quantify potential benefits. Then we propose a theoretical framework that helps to formalize prior knowledge injection. It is a first step towards an informed learning theory. Our main idea is to regard prior knowledge as a function that can be represented in the same space as the model or the training data (See Figure 1a). We conjecture that the distance between data and knowledge determines the potential benefits of informed machine learning. To illustrate the framework, we perform a systematic experimental study with toy problems. As the toy problems we propose a classification task, which allow to vary the knowledge and the injection method in a controllable manner. We vary relevant parameters, such as the distance between data and knowledge (See Figure 1b), and measure the potential improvements through informed machine learning.

In summary, the main contributions of this paper are:

1) We propose a set of metrics for quantifying the benefits of informed machine learning.
2) We present a first theoretical framework for informed machine learning.
3) We perform a systematic experimental study with controllable toy problems.

Each of these contributions helps to answer the above research questions. The paper is structured accordingly.

II. Related Work

Our work is mainly related to hybrid AI and informed machine learning, but also reuses concepts from learning theory.

A. *Informed Machine Learning*

In informed machine learning, pre-given formalized knowledge is injected into data-driven learning systems [4], [16], [17]. It is sometimes also called theory-guided data science [18], or causally-aware machine learning [19]. The taxonomy of informed learning depicts the diversity of applications and methods in terms of knowledge source, representation type, and integration method [4].

However, related work about the general, application-independent, effect of knowledge injection in informed machine learning is rare. We shortly describe the works that go in this direction, ordered by our contributions in terms of 1) metric quantification, 2) theoretical framework and 3) experimental study.

A first work that presents an approach for the quantification of domain knowledge in informed machine learning is given by Yang et al. [20]. They proposed a method based on the Shapley value to quantify the contribution of injected prior knowledge to the model performance improvement. *The main difference to our work is that they consider a set of knowledge pieces and attribute the contribution of the individual pieces, whereas we consider knowledge as an abstract unit and analyze which properties it needs to have.*

A first theoretical study about physics-informed neural networks was presented by Shin et al. [21]. They provide a convergence theory with respect to the number of data samples. Yang et al. have also presented a theoretical study on informed learning by wide neural networks [22]. They especially investigate the trade-off between knowledge and data labels.

A first experimental comparison of informed learning methods is given by Monaco et al. [23]. They consider three application examples and on each they evaluate two informed learning methods. In particular they measure the performance for variations of the training data size. *The difference to our work is that they investigate pre-given applications, whereas we investigate toy problems, which allow us to adapt the experiments and the knowledge in a controllable manner. Moreover, they only measure the prediction error for various data sizes, whereas we develop and measure a total catalogue of metrics.*

B. *Learning Theory*

The foundations of statistical learning theory have been developed already many years ago by Vapnik et al. [24]. Overviews about learning theory can be found in [25], [26]. At the heart of it is the principle of empirical risk minimization, which we shortly recap. The goal of a learning task is to find a model $f : \mathcal{X} \to \mathcal{Y}$, with $f \in \mathcal{F}$, based on some given training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1\ldots n}$, with features $x \in \mathcal{X}$, labels $y \in \mathcal{Y}$ and sample size n. The model can then be approximated by minimizing the empirical risk $R(f)$ with a given loss function l:

$$\hat{f} := \arg\min_{f \in \mathcal{F}} R_{\mathcal{D}}(f), \quad R_{\mathcal{D}}(f) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} l(f(x), y) \quad (1)$$

Recently, an extension of the statistical learning theory was proposed in terms of also taking into account the preservation of invariants [27], also called invariant risk minimization [28], [29]. This approach can be motivated by the goal of out-of-distribution generalization [30]: Assuming training data is collected in various environments, then statistical invariants across them should also hold in novel testing environments [31]. *This idea is similar to our understanding of informed machine learning: Prior knowledge describes causal relationships that are underlying a given data distribution, i.e. invariants. Integrating these into a learning task can thus improve model performance. The main difference is that in invariant risk minimization the invariants still need to be learned, whereas in informed machine learning they are given by prior knowledge.*

III. Metrics for Informed Learning

As described in [4], the main goals of informed machine learning are to train with less data, to achieve a better model performance, to increase knowledge conformity, or to increase

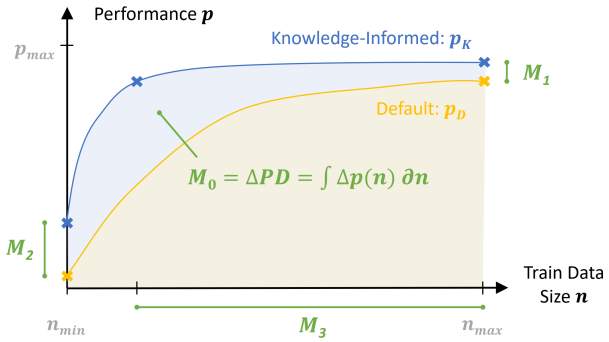


Fig. 2: Illustration of Performance vs. Size of Training Data. Models that are trained with informed machine learning usually achieve a higher performance, e.g. accuracy or robustness, for smaller training data sizes [5], [6], [11]. We propose a new metric that quantifies performance and data need in a single metric in terms of the area under the curve: *Performance-by-Data AUC*, in short PD (see Section III-A). All in all, we suggest to quantify improvements in terms of four metric flavours: Increase in Performance-by-Data AUC (M_0), increase in performance at max. and min. data size (M_1 and M_2), as well as data reduction for a specific performance (M_3).

interpretability. However, most works about informed learning methods present individual metrics to quantify the benefits of their method. For example, [5] reports test error and physical inconsistency for various data sizes, [20] compares test accuracy for full data size, or [11] reports ouf-of-distribution robustness for various data sizes.

Here, we propose a systematic metric catalogue, as well as a new metric that combines performance improvement and data efficiency. These allow a more transparent, and standardized comparison of various methods. Moreover, they provide the basis for future benchmarks of informed learning methods.

A. *Performance-by-Data AUC*

We propose to measure performance (e.g., test accuracy) for various train data sizes and summarize the results in a single metric that we call *Performance-by-Data AUC*. As illustrated in Figure 2, the metric quantifies the area under the curve of performance p vs. training data size n.

Definition 1 (Performance-by-Data AUC)**.**

$$PD = \int_{n_{min}}^{n_{max}} p(n)\,\mathrm{d}n \tag{2}$$

This metric can be normalized through dividing by the maximum possible area, i.e. by $p_{max} * (n_{max} - n_{min})$, where p_{max} is the maximum possible performance (e.g., 100% test accuracy). Then $PD \in [0.0, 1.0]$ and the larger the better.

For comparing two models, e.g., a (knowledge-)informed model with performance p_K and a default, data-based model with performance p_D, the difference between the two area integrals can be computed.

Definition 2 (Improvement of Performance-by-Data AUC)**.**

$$\Delta PD = \int_{n_{min}}^{n_{max}} \Delta p(n)\,\mathrm{d}n \tag{3}$$

$$= \int_{n_{min}}^{n_{max}} (p_{\mathrm{K}}(n) - p_{\mathrm{D}}(n))\,\mathrm{d}n \tag{4}$$

The proposed ΔPD metric has the advantage that it encapsulates the performance for all data set sizes in a single metric. This means, one does not need to choose a specific data set size for which to compare the performance, or vice versa.

B. *Metrics Catalogue*

For evaluating informed learning methods, we focus on metrics that are especially relevant for model generalization: In-Distribution Test Accuracy, Out-of-Distribution Robustness, and Knowledge Conformity. The generic performance p from above can be any of these 3 metric types. As indicated in Figure 2, we specifically evaluate each metric in 4 metric flavours: The above described Performance-by-Data AUC (M_0 in Figure 2), but also the performance at max. and min. data size (M_1 and M_2), as well as the data amount that is required to achieve a specific performance (M_3). Further metric types for evaluating informed methods are training time and model size. Also a measurement of the model interpretability is interesting, however, such a quantification is currently still an open research question [15].

In summary, we suggest the metric catalogue that is shown in Box 1 for evaluating informed learning methods.

Box 1: Metrics Catalogue: Improvements through Informed Learning

This catalogue represents the various goals of informed learning and depicts how knowledge injection can improve machine learning.

1) **Increase of In-Distribution (IID) Test Accuracy**
   a) Increase: IID Accuracy-by-Datasize
   b) Increase: IID Accuracy for Max. Datasize
   c) Increase: IID Accuracy for Min. Datasize
   d) Reduction: Training Datasize for specific IID Accuracy
2) **Increase of Out-of-Distribution (OOD) Robustness**
   a) Increase: OOD Robustness-by-Datasize
   b) Increase: OOD Robustness for Max. Datasize
   c) Increase: OOD Robustness for Min. Datasize
   d) Reduction: Training Datasize for specific OOD Robustness
3) **Increase of Knowledge Conformity**
   a) Increase: Knowledge Conf.-by-Datasize
   b) Increase: Knowledge Conf. for Max. Datasize
   c) Increase: Knowledge Conf. for Min. Datasize
   d) Reduction: Training Datasize for specific Knowledge Conf.
4) **Reduction of Training Data ***
5) **Reduction of Training Time**
6) **Reduction of Model Size**
7) **Improvement in Interpretability**

* Please note that the important goal of data reduction is represented below each of first three metric types (See 1a+d, 2a+d, 3a+d).

IV. A Framework for an Informed Learning Theory

We want to better understand what influences the expected performance gains of informed learning. In particular, it is of great interest what the requirements on the injected knowledge are. To investigate this, we employ and extend concepts from statistical learning theory [25], [26]. This way, we hope to make a first step in the direction of an *informed* learning theory.

A. *Knowledge in Function Space*

The question about the requirements on the injected knowledge is non-trivial, because knowledge can be represented in versatile forms. As depicted in the informed learning taxonomy [4], typical representations of prior knowledge are algebraic equations, logic rules, knowledge graphs, simulation results, or human feedback. An investigation on the requirements for each type could already be exhaustive.

Here, we therefore take an abstract view and conjecture:

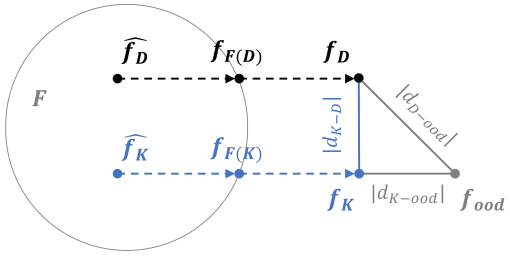


Fig. 3: Function space with representations of prior knowledge f_K, data f_D, and OOD test data f_{ood} (right), and decomposition in generalization error terms (left). The circle illustrates the function space F used by a learning algorithm. Beyond the circle is the space of all possible functions. $\hat{f}_D$ is the empirical best solution of the algorithm (See Equation 1). $f_{F(D)}$ is the best possible solution in F. f_D represents the (*unknown*) data distribution. The blue elements show the respective representations for prior knowledge (our proposed informed learning extension). Particularly, f_K represents the (*known*) prior knowledge.

Axiom 3 (Prior Knowledge). Prior knowledge describes relations between concepts and can be represented as a function.

We use this to relate it to given data:

Axiom 4 (Knowledge Representation in Function Space). Prior knowledge can be represented in the same function space as given data representations.

Figure 3 illustrates the knowledge representation in a function space. Here, we also illustrate the distance $|d_{K-D}|$ between the *known* knowledge representation f_K and the *unknown* data representation f_D. In addition to the in-distribution data, we also consider an out-of-distribution data, which is represented by the *unknown* data representation f_{ood}. The illustration in function space depicts how prior knowledge can give hints about the unknown data representations.

B. Knowledge-to-Data Distance

Let us consider the case for in-distribution (IID) generalization. This means that a model is tested on data that follows the same underlying distribution as the training data.

We are interested in the expected performance improvement through informed learning by using the prior knowledge f_K. In the statistical learning theory, maximizing model performance is equivalent to minimizing the empirical risk (see Equation 1). We thus regard the risks $R(\hat{f_D})$ and $R(\hat{f_K})$. The generalization error for the default, data-based model can be decomposed as follows (see *black* drawing in Figure 3):

$$\underbrace{R(\hat{f_D}) - R(f_D)}_{\text{generalization error}} = \underbrace{\Big(R(\hat{f_D}) - R(f_{F(D)})\Big)}_{\text{estimation error}} + \underbrace{\big(R(f_{F(D)}) - R(f_D)\big)}_{\text{approximation error}} \tag{5}$$

We propose to also formalize the error for a purely informed model with respect to generalization to the in-distribution data (see *blue* drawing in Figure 3):

$$\underbrace{R(\hat{f_K}) - R(f_D)}_{\text{know. generalization error}} = \underbrace{\Big(R(\hat{f_K}) - R(f_{F(K)})\Big)}_{\text{know. estimation error}} + \underbrace{\big(R(f_{F(K)}) - R(f_K)\big)}_{\text{know. approximation error}} + \underbrace{\big(R(f_K) - R(f_D)\big)}_{\text{know.-to-data error}} \tag{6}$$

For the model distance in terms of their generalization errors follows then:

$$R(\hat{f_K}) - R(\hat{f_D}) = C + \underbrace{\big(R(f_K) - R(f_D)\big)}_{\text{know.-to-data error}} \propto C + \underbrace{|d_{K-D}|}_{\text{know.-to-data distance}} \tag{7}$$

Conjecture 5 (Informed IID-Generalization Improvement). The smaller the distance between knowledge and data, the larger the improvement through informed learning on in-distribution generalization.

C. Knowledge-to-OOD Distance

Let us consider the case of out-of-distribution generalization. Out-of-distribution generally refers to the evaluation on test data that follows another distribution then the train data [30].

Here, for the model distance in terms of their out-of-distribution generalization errors follows then:

$$R_{ood}(\hat{f_K}) - R_{ood}(\hat{f_D}) \propto C_{ood} + \underbrace{|d_{K-ood}|}_{\text{know.-to-ood dist.}} - \underbrace{|d_{D-ood}|}_{\text{data-to-ood dist.}} \tag{8}$$

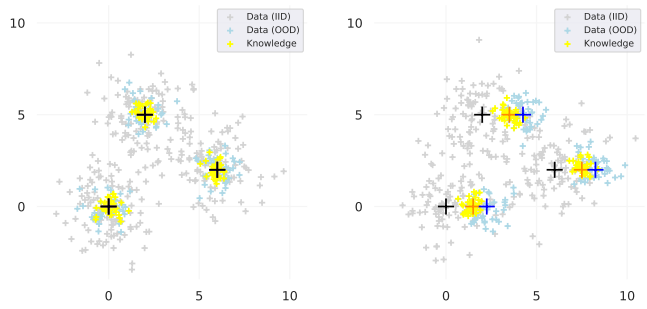
Conjecture 6 (Informed OOD-Generalization Improvement). The smaller the distance between knowledge and the OOD data, and the larger the distance between IID and OOD data, the larger the potential improvement through informed learning on the OOD generalization.

V. Systematic Experimental Analysis

We performed a systematic experimental study of the effect of knowledge injection in informed machine learning. For this, we defined a controllable toy problem. We measured the performance metrics as defined in Section III and employ the theoretical framework from Section IV.

A. Experimental Setup

1) Toy Datasets: Let us consider a toy problem for the task of classification, as illustrated in Figure 4b. We have also investigated a toy problem for regression, which shows similar results. Since the effects of knowledge injection, especially the influence of the distances between knowledge and data, can



(a) No Distances. (b) With Distances.

Fig. 4: Toy dataset for classification with 3 classes. We use distinct sets for in-distribution data (grey), out-of-distribution data (blue), and prior knowledge (yellow). The distance can be varied between the sets (as motivated by the theoretical framework in Section IV). (a) shows the case when the centers of IID data, OOD data, and knowledge overlap, i.e. for $|d_{K-D}| = 0$, $|d_{K-OOD}| = 0$. (b) shows an example with distances between them (Here: $|d_{K-D}| = 1.5$, $|d_{K-OOD}| = 0.75$). In our experimental study, we measure the effect of informed machine learning for various distance setups.

be more clearly with the classification problem, we consider this in the following.

The toy dataset contains three classes. Each blob in Figure 4b represents another class (i.e. the top blob, lower left blob, and middle right blob). The number of samples is 288, with 96 samples per class.

In addition to the main (IID) data, we also consider a smaller sets of OOD data, and of prior knowledge representations. Here, the original prior knowledge representation can be understood as class prototypes, similar as in [11]. In applications, such prototypes can, e.g., be structural templates (e.g. traffic sign templates for image recognition). Such knowledge can be transformed into a data format by rendering. Since prior knowledge is more concise than data, we consciously chose smaller standard deviations for the knowledge set.

The distances between the main (IID) data, the OOD data, and the prior knowledge can be controlled and varied. An example for a distance setup is shown in Figure 4b.

2) Systematic Analysis: In our systematic study, we vary several parameters: 1) Distances between knowledge and data, 2) Amount of training data, 3) (informed) learning method. For each setup, we measure the metrics from our metrics catalogue. Especially, we focus on IID Test Accuracy, OOD Robustness, and Knowledge Conformity (i.e., accuracy on the IID data set, accuracy on the OOD data set, and accuracy on knowledge samples).

We investigate a range of distance setups, as illustrated in Figure 5. For this, we keep the position of the IID data set fixed and move the OOD data set and/or the knowledge to the side. In particular, we consider a maximum distance of 3.5 with a step size of 0.25, i.e. a total of 15 positions. We combine distances of know-data with distances of know-ood, resulting in the illustrated position triangles. For every position we perform separate trainings.

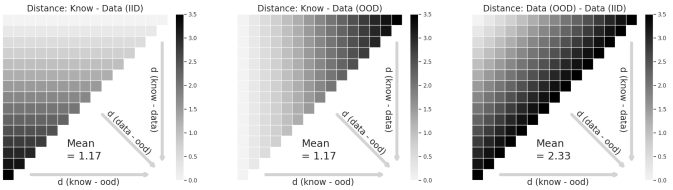


Fig. 5: Illustration of distance variation in systematic experimental study: Every position represents a unique experimental set up in terms of distances between prior knowledge, (IID) data, and OOD data. For every setup, we perform a default neural network training and informed trainings in order to measure the gained performance improvements. In addition, we train every setup for a range of training data sizes.

Furthermore, we vary the size of the training data. We consider 6 unique sizes from 10 to 300 data samples with an exponential growth. By taking into account various training data sizes, we can measure the metric flavours, as described in Section III: Performance-by-Data AUC, performance at max. data, performance at min. data, and data need to reach a specific performance.

As informed machine learning, we apply two methods, similar as in [11]: Combining training data and knowledge samples in terms of 1) Concurrent Training, 2) Informed Pre-Training.

3) Learning Setup: We apply a neural network with 1 hidden layer with 100 neurons. We use stochastic gradient descent and cross entropy loss for the learning algorithm. As the hyperparameters we use: batch size = 18, learning rate = 0.01, momentum = 0.9, early stopping after 3 stagnating epochs, regularization with weight decay = 0.2. Each experiment is repeated 10 times. For every run the data samples are generated randomly.

B. Results

The complete results in terms of improvements of informed learning over the default setup can be found in Figure 7. Results for Informed Pre-Training are shown in Figure 8. Both informed learning methods show that our distance theorems from Section IV are confirmed. We also nicely see, that our introduced metric of Performance-by-Data AUC (Definition 2) is a good summary of the other metrics. In general, we see that informed learning can greatly improve OOD robustness.

A subset of the results is shown for a closer look in Figure 6. The left subfigure shows the improvement in IID generalization. We observe that the smaller the distance between knowledge and training data (upper pixel rows) the larger the improvement. This confirms our Conjecture 5 from above. The right subfigure shows the improvement in OOD robustness. Here, we can see that the the improvement is largest when the distance between knowledge and training data is large (lower pixel rows) and the distance between knowledge and OOD test data is small (closer to diagonal). This confirms our Conjecture 6 from above.

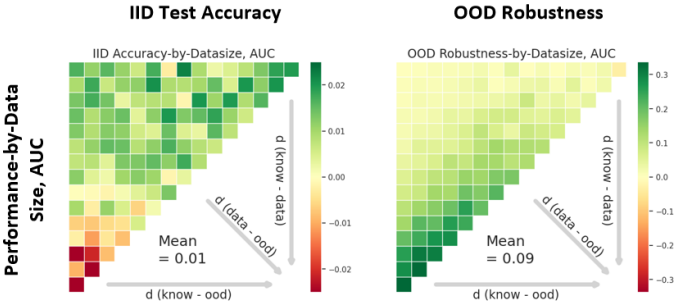


Fig. 6: Experimental Results: Improvements in IID-Generalization and OOD-Robustness through Informed Training. The left plot confirms our Conjecture 5, and the right our Conjecture 6. (Complete Results can be found in Figures 7 and 8.)

VI. Conclusion

In this paper, we presented a framework for quantifying the value of prior knowledge in informed machine learning. We first proposed a set of relevant metrics for quantifying the benefits of knowledge injection, comprising in-distribution accuracy, out-of-distribution robustness, and knowledge conformity. We also introduced a metric that combines performance improvement and data reduction, called performance-by-data AUC. Secondly, we presented a theoretical framework that represents prior knowledge in a function space and relates it to data representations and a trained model. Thirdly, we performed a systematic experimental study with controllable toy problems. These confirmed our theories about the influence of the distances between knowledge and data on potential model improvements. All in all, our contributions hopefully help to find general answers to the question how knowledge injection helps. In particular they form the basis for potential benchmarks of informed machine learning.

Acknowledgment

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence.

References

[1] G. Marcus, "The next decade in ai: four steps towards robust artificial intelligence," *arXiv preprint arXiv:2002.06177*, 2020.

[2] L. von Rueden, S. Mayer, R. Sifa, C. Bauckhage, and J. Garcke, "Combining machine learning and simulation to a hybrid modelling approach: Current and future directions," in *Advances in Intelligent Data Analysis(IDA)*. Springer, 2020.

[3] M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler, "Neuro-symbolic artificial intelligence," *AI Communications*, 2021.

[4] L. Von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, M. Walczak, J. Garcke, C. Bauckhage, and J. Schuecker, "Informed machine learning – a taxonomy and survey of integrating prior knowledge into learning systems," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[5] A. Karpatne, W. Watkins, J. Read, and V. Kumar, "Physics-guided neural networks (pgnn): An application in lake temperature modeling," *arXiv preprint arXiv:1710.11431*, 2017.

[6] R. Stewart and S. Ermon, "Label-free supervision of neural networks with physics and domain knowledge," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[7] T. M. Deist, A. Patti, Z. Wang, D. Krane, T. Sorenson, and D. Craft, "Simulation-assisted machine learning," *Bioinformatics*, 2019.

[8] A. Rai, R. Antonova, F. Meier, and C. G. Atkeson, "Using simulation to improve sample-efficiency of bayesian optimization for bipedal robots," *The Journal of Machine Learning Research*, 2019.

[9] T. Kyono and M. van der Schaar, "Improving model robustness using causal knowledge," *arXiv preprint arXiv:1911.12441*, 2019.

[10] N. M. Gürel, X. Qi, L. Rimanic, C. Zhang, and B. Li, "Knowledge enhanced machine learning pipeline against diverse adversarial attacks," in *International Conference on Machine Learning*. PMLR, 2021.

[11] L. Von Rueden, S. Houben, K. Cvejoski, C. Bauckhage, and N. Piatkowski, "Informed pre-training on prior knowledge," *arXiv preprint arXiv:2205.11433*, 2022.

[12] M. Bahari, I. Nejjar, and A. Alahi, "Injecting knowledge in data-driven vehicle trajectory predictors," *Transportation research part C: emerging technologies*, 2021.

[13] L. von Rueden, T. Wirtz, F. Hueger, J. D. Schneider, N. Piatkowski, and C. Bauckhage, "Street-map based validation of semantic segmentation in autonomous driving," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021.

[14] J. Wörmann, D. Bogdoll, E. Bührle, H. Chen, E. F. Chuo, K. Cvejoski, L. van Elst, T. Gleißner, P. Gottschall, S. Griesche *et al.*, "Knowledge augmented machine learning with applications in autonomous driving: A survey," *arXiv preprint arXiv:2205.04712*, 2022.

[15] K. Beckh, S. Müller, M. Jakobs, V. Toborek, H. Tan, R. Fischer, P. Welke, S. Houben, and L. von Rueden, "Explainable machine learning with prior knowledge: An overview," *arXiv preprint arXiv:2105.10172*, 2021.

[16] L. Von Rueden, S. Mayer, J. Garcke, C. Bauckhage, and J. Schuecker, "Informed machine learning – towards a taxonomy of explicit integration of knowledge into machine learning," *arXiv preprint arXiv:1903.12394*, 2019.

[17] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, 2021.

[18] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Transactions on knowledge and data engineering*, 2017.

[19] T. Kyono, "Towards causally-aware machine learning," *PhD Thesis, University of California*, 2021.

[20] J. Yang and S. Ren, "A quantitative perspective on values of domain knowledge for machine learning," *arXiv preprint arXiv:2011.08450*, 2020.

[21] Y. Shin, J. Darbon, and G. E. Karniadakis, "On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type pdes," *arXiv preprint arXiv:2004.01806*, 2020.

[22] J. Yang and S. Ren, "Informed learning by wide neural networks: Convergence, generalization and sampling complexity," *arXiv preprint arXiv:2207.00751*, 2022.

[23] S. Monaco, D. Apiletti, and G. Malnati, "Theory-guided deep learning algorithms: An experimental evaluation," *Electronics*, 2022.

[24] V. Vapnik, "Principles of risk minimization for learning theory," *Advances in neural information processing systems*, 1991.

[25] U. Von Luxburg and B. Schölkopf, "Statistical learning theory: Models, concepts, and results," in *Handbook of the History of Logic*. Elsevier, 2011.

[26] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[27] V. Vapnik and R. Izmailov, "Rethinking statistical learning theory: learning using statistical invariants," *Machine Learning*, 2019.

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

[29] K. Ahuja, J. Wang, A. Dhurandhar, K. Shanmugam, and K. R. Varshney, "Empirical or invariant risk minimization? a sample complexity perspective," *arXiv preprint arXiv:2010.16412*, 2020.

[30] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," *arXiv preprint arXiv:2108.13624*, 2021.

[31] M. Arjovsky, "Out of distribution generalization in machine learning," Ph.D. dissertation, New York University, 2020.

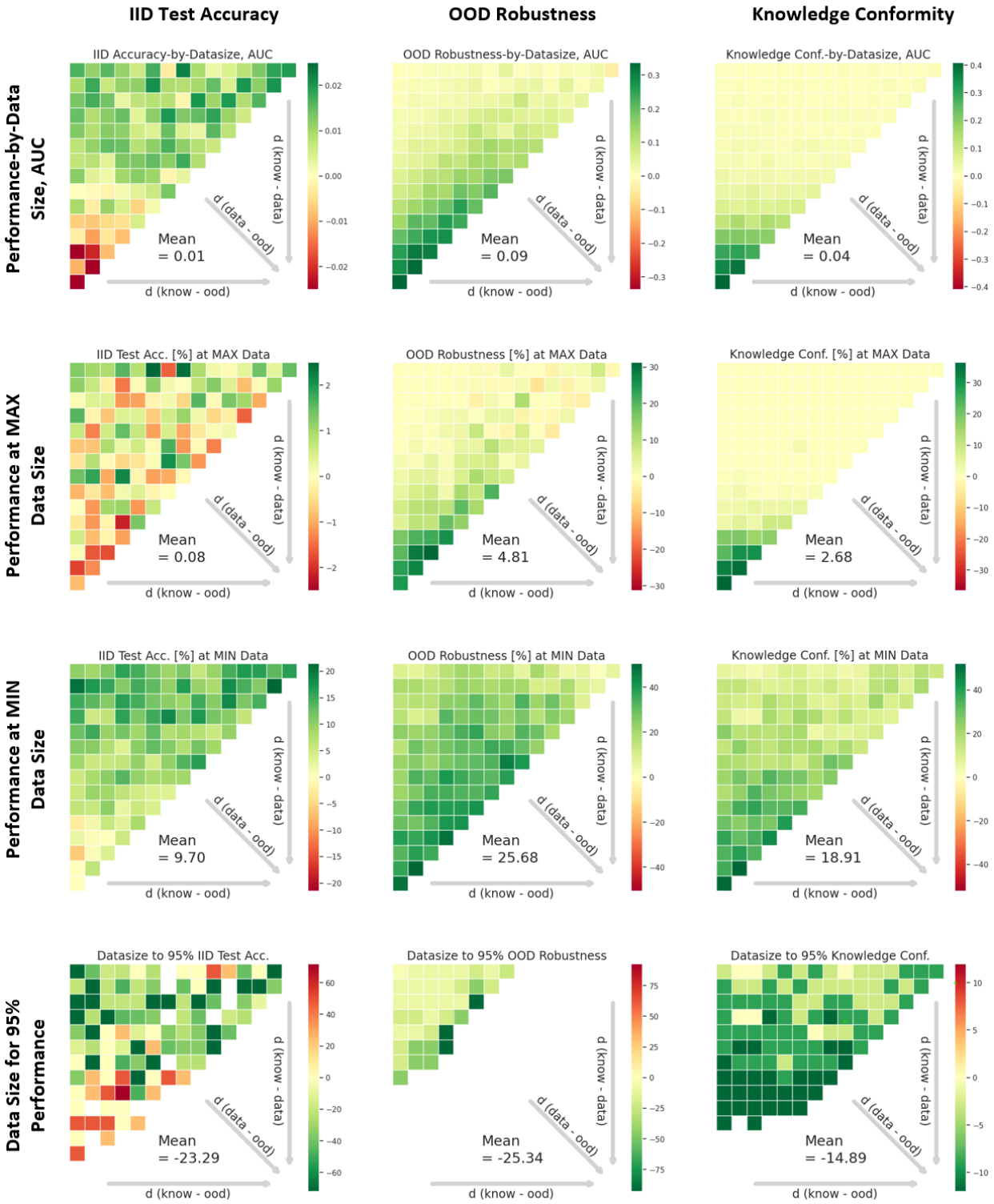


Fig. 7: Experimental Results: Improvements through Informed Training.

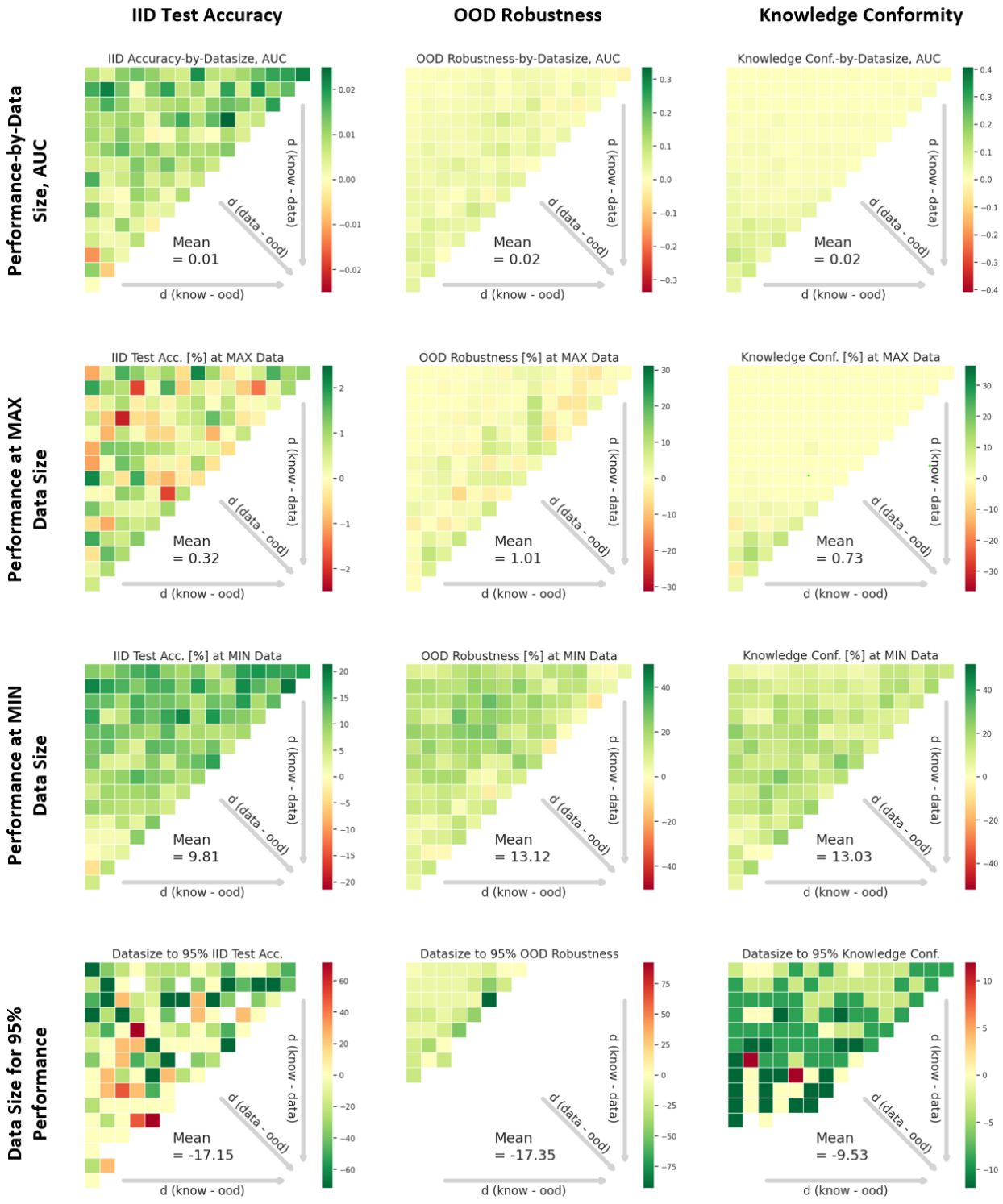


Fig. 8: Experimental Results: Improvements through Informed Pre-Training.