Theory-guided Data Science models

Thesis syntesis - *Simone Monaco*

We are living in an era of unprecedented data generation: petabytes of data are produced globally every year, fueling the rapid advancement of data-driven technologies. This vast availability of data has given rise to powerful machine learning techniques that aim to discover patterns and extract actionable insights from complex datasets [40]. The paradigm of data science, which emerged from this idea, has revolutionized various fields, enabling automated decision-making and predictive modeling at a completely new scale.

However, scientific progress has historically relied on a different paradigm—one grounded in the formulation, validation, and refinement of theories through systematic experimentation and observation. Since the 17th century, the scientific method has been the cornerstone of knowledge acquisition, emphasizing hypothesis-driven discovery. In contrast, modern ML methods, often operate without explicit theoretical foundations, relying purely on statistical correlations rather than causal or mechanistic understanding.

On the theory-driven side, the relationships between input variables (drivers) and response variables in scientific modeling have been captured using process-based models, which are grounded in established scientific principles and equations. These models leverage mechanistic understanding to ensure interpretability and consistency with known physical, chemical, or biological laws. However, despite their theoretical strengths, process-based models face several limitations that hinder their applicability in complex real-world scenarios [70]. Challenges include inherent simplifications or biases in model formulations, inaccuracies in parameter estimation, and high computational demands, particularly when solving high-fidelity simulations.

These limitations make it difficult to scale process-based models for large, high-dimensional, or dynamically changing systems.

In response to these challenges, Machine Learning (ML) has emerged as a promising alternative for capturing intricate relationships between inputs and outputs directly from empirical data. ML models can learn complex patterns without requiring explicit knowledge of underlying governing equations, making them particularly attractive for domains where formulating accurate process-based models is difficult.

While the success of these black-box models has led some to proclaim "the end of theory" [4], this view neglects the importance of domain knowledge in ensuring the robustness, interpretability, and generalizability of predictive models. In fact, purely data-driven models also come with significant drawbacks. One of the primary concerns is their limited generalizability, particularly when faced with Out-of-Distribution (OOD) data [11]. This issue is especially problematic in scientific applications where extrapolation is often necessary. Unlike in mainstream AI applications such as computer vision and natural language processing, where massive datasets have driven the success of deep learning models, many scientific disciplines suffer from data scarcity. This lack of large-scale, high-quality labeled data exacerbates the challenges of training robust ML models that can generalize well beyond their training distribution.

Moreover, black-box ML models often struggle to align with established scientific theories, potentially generating physically inconsistent or non-meaningful predictions. This lack of interpretability and theoretical grounding limits their utility for scientific discovery, as they fail to provide mechanistic insights into the patterns they learn. Without explicit constraints to enforce domain knowledge, these models risk producing spurious correlations rather than capturing causative or physically meaningful relationships. Put in another way, purely data-driven models can lead to unrealistic predictions that violate known physical, biological, or economic principles [69, 134].

These limitations from both theoretical and data-driven sides suggest the possibility of exploring new ways of taking advantages from one world for the other and vice versa. In other words, it would be crucial to investigate wether it is possible to improve ML models generalizability with theoretical laws as well as speed-up scientific discovery with ML.

Looking at the world with these two lenses together gave rise to the emerging paradigm of Theory-Guided Data Science (TGDS) [69], also called Knowledge-Guided Machine Learning (KGML). KGML approaches integrate domain knowledge (that is, physical laws, conservation principles, or structural constraints) into data-driven models. By integrating the strengths of both process-based modeling and ML, hybrid approaches seek to enhance algorithms across multiple dimensions. The anticipated benefits include:

- Improved generalizability to OOD data and domains.
- Enhanced robustness in low-data regimes.
- Increased interpretability of model predictions.
- Ensuring physically consistent and meaningful outputs.

However, rigorous experimentation is necessary to verify that a given solution effectively achieves these benefits for a specific task.

Traditional methods of integrating knowledge into machine learning include feature engineering, domain-aware labeling, and structured regularization. However, more recent techniques introduce deeper forms of knowledge integration, such as logical constraints [42], algebraic formulations [132, 35], and differential equations embedded within neural network architectures [118].

0.1 KGML paradigm

The TGDS paradigm [69] has demonstrated significant success across a wide range of scientific and engineering disciplines by integrating domain knowledge into data-driven models. Its applications span diverse fields, including climate science [51, 110], where it has been used to improve weather forecasting and climate anomaly detection; cyber-physical systems [116], where it enhances predictive maintenance and system reliability; turbulence modeling [98], where it refines fluid dynamics simulations; materials discovery [20, 115], where it accelerates the identification of novel materials with desirable properties; biological sciences [112], where it aids in understanding multi-scale biological processes; quantum chemistry [108], where it assists in predicting molecular properties; and hydrology [35], where it improves the modeling of water systems and hydrodynamic processes.

Von Rueden et al. [134] provided a comprehensive taxonomy of the avaliable techniques to emprove ML models under this paradigm. All approaches are categorized on three key aspects:

1. **Knowledge source** – the origin of the integrated knowledge, whether it stems from established scientific theories, empirical observations, or expert intuition.
2. **Knowledge representation** – the format in which the knowledge is encoded, such as mathematical equations, logical rules, or probabilistic models.
3. **Knowledge integration** – the stage in the machine learning pipeline where domain knowledge is incorporated, such as during data preprocessing, model architecture design, or loss function formulation.

0.1.1 Knowledge Sources

The **source** of domain knowledge plays a critical role in shaping the integration process and can be categorized into three primary types.

Scientific Knowledge

Scientific knowledge is typically formalized and validated through rigorous theoretical derivations, analytical proofs, or experimental observations. This category encompasses principles derived from the natural sciences, engineering, mathematics, and physics, where established laws and governing equations—such as the Navier-Stokes equations in fluid dynamics or Maxwell's equations in electromagnetism—serve as fundamental constraints for ML models. By embedding such theoretical knowledge into data-driven approaches, models can benefit from well-understood mechanistic relationships, leading to increased interpretability, robustness, and generalization beyond training data. This kind of knowledge is generally the easiest to encorporate, as it is well-formalized. In contrast, embedding partial knowledge could limit the output space and then the final model expressivity. In other words, forcing a ML model to follow a certain theoretical behaviour would make it suffer they suffer from research gaps eventually present in the theory itself, limiting its effectiveness in many real-world scientific applications [70]. For this reason, a crucial aspect of KGML falls under the the applications of either *hard* or *soft* constraints, depending on the use case.

World Knowledge

Unlike scientific knowledge, world knowledge consists of intuitive, everyday information that is typically acquired through human perception and experience. This knowledge is not necessarily formalized but is often structured through commonsense reasoning. Examples include semantic relationships in language, intuitive physical interactions (e.g., an object falls when dropped), and general facts about the environment (e.g., a cat has two ears and can meow). While world knowledge lacks formal scientific validation, it can still provide valuable priors for machine learning models, particularly in areas such as natural language processing, computer vision, and cognitive AI systems. Linguistic knowledge, including syntactic structures and semantic associations, also falls under this category and can be leveraged to improve language models and AI-driven reasoning systems.

Expert Knowledge

Expert knowledge is specialized and domain-specific, typically held by a restricted group of professionals or practitioners. It may resemble world knowledge expet the fact that it includes all the potentially non-structured knowledge which is however formalized in theoretical or practical fundaments.

Unlike scientific knowledge, it may not always be mathematically formalized, yet it is validated through experience, empirical observations, or consensus within a field. Examples include medical diagnoses by doctors, engineering heuristics used by seasoned professionals, and economic forecasting models based on expert insights. In machine learning, expert knowledge can be integrated through:

- Human-in-the-loop systems, where domain experts guide model training and validation.
- Rule-based systems, where explicit domain heuristics are encoded into ML frameworks.

- Weak supervision, where expert annotations or structured priors refine model learning.

Expert knowledge is particularly useful in fields where data is scarce, expensive to acquire, or highly specialized, such as medicine, finance, and engineering. By integrating expert-driven constraints and insights, ML models can achieve greater reliability, transparency, and trustworthiness, reducing the risks associated with purely data-driven decision-making.

0.1.2 Knowledge Representation

The *Knowledge Representation* category concerns how the prior information is formalized. Depending on the available knowledge for a given task, different representations can be utilized. Below, we summarize the most widely adopted and relevant alternatives for this discussion.

Equations

When differential or algebraic equations are involved, the final solution may exhibit a partially known behavior or be subject to constraints that can be formalized mathematically. Constraints are typically represented using algebraic equations or inequalities. Prominent examples include the energy-mass equivalence ($E = mc^2$) or the mass invariance encapsulated in the Minkowski metric, which has been integrated using a Lorentz layer in [18].

Regarding inequalities, final or intermediate solutions may have physical upper or lower bounds (e.g., the velocity of an object cannot exceed the speed of light). This scenario is explored in [107], where the authors investigate methods for embedding priors such as bounds and monotonicity constraints into learning processes.

Similar considerations apply to differential equations governing the dynamic behavior of state variables, inputs, and outputs. The underlying principles may be fully known but impractical to implement, partially known but not entirely representative of the real solution, or completely unknown [136].

In such cases, machine learning algorithms can be employed to solve differential equations, as demonstrated in [118], to learn the residual dynamics given prior knowledge about the solution's behavior, as shown in [124], or to directly learn the spatiotemporal dynamics itself, as in [71].

Simulation Results

Many physical systems can be modelled using simulators, which numerically solve mathematical models with varying levels of precision. While for some use cases, simulators can be considered a true but often computationally unfeasible solution, some other tasks can be only partially explained by simulation only. In this last class of problem, a simulation, seen as a partial solution, can be incorporated alongside input data, potentially enabling the deep learning model to learn the necessary corrective terms [35].

Domain-Specific Invariances

Certain types of input data exhibit intrinsic invariances due to their structural properties. For example, images may retain key characteristics under translation or rotation. Other types of data may exhibit permutation invariance, time invariance, or periodicity. In each of these cases, the algorithm can benefit from learning these properties while approaching a solution to the main task. This can either improve models performance and avoid unfeasible solutions. One natural and powerful trend to handle invariances generally leads to specialized model architectures that can more effectively capture these properties as *inductive biases* [8].

0.1.3 Knowledge integration

The integration of knowledge into machine learning algorithms can take place at various stages of the pipeline: at the beginning, by embedding it within the training dataset; in the middle, through the design of tailored architectures and learning strategies; or at the end, by influencing the model's output.

0.1.4 Knowledge as additional input information

A traditional method for incorporating prior knowledge into data is feature engineering, where additional features are derived from sampled data to highlight domain-specific insights [150]. Another already mentioned approach involves augmenting the dataset with synthetic information obtained from simulations, as demonstrated in [35, 106], allowing the final algorithm to learn residuals from these approximations.

In a completely different context, recent advancements in Natural Language Processing and the development of Large Language Models (LLM) have enabled these architectures to incorporate prior knowledge through free-text input. Many techniques aimed at improving LLMs performance involve *augmenting* the prompt with additional contextual information, either to provide relevant background knowledge for generating responses or to guide the reasoning process explicitly.

Such approaches fall under the umbrella of In-Context Learning [45], the same principle that underpins the success of Retrieval-Augmented Generation (RAG), which is becoming increasingly popular in various industrial applications. These techniques represent another example of knowledge integration through supplementary input. The key distinction, however, is that in this case, the additional information is provided solely at inference time rather than during training.

0.1.5 Learning with Regularization Terms

Another way to integrate prior knowledge is by constraining the learning process through the introduction of a physics-informed loss function alongside standard supervised objectives. This general approach can be formalized as follows [141]:

$$\mathscr{L} = \mathscr{L}_{\mathrm{SUP}}(Y_{\mathrm{TRUE}}, Y_{\mathrm{PRED}}) + \gamma \mathscr{L}_{\mathrm{PHY}}(Y_{\mathrm{PRED}}) + \lambda R(W) \; ; \tag{1}$$

where $\mathscr{L}_{\mathrm{SUP}}$ represents the supervised loss (e.g., Mean Squared Error, cross-entropy), R is an additional regularization term that limits model complexity, and $\mathscr{L}_{\mathrm{PHY}}$ incorporates physics-informed constraints. The coefficients γ and λ control the relative contributions of these terms. The physics-based term may include algebraic, differential, or logical constraints.

For instance, in predicting lake temperature variations with depth, Willard et al. [141] introduced a penalty term that prevents predictions from violating theoretical water density constraints.

Similarly, Beucler et al. [14] enforced conservation laws in climate modeling by incorporating them as *soft constraints* within the loss function and as *hard constraints* by limiting the degrees of freedom in the neural network's predictions. In this approach, some variables are computed deterministically through fixed layers rather than learned. A comparable strategy is applied to AC optimal power flow in [48].

Architectural biases

A more advanced approach to embedding physical knowledge into deep learning models involves designing architectures that inherently respect domain-specific principles [134]. Deep learning models can be structured to incorporate *relational inductive biases* [8] from the outset, shaping their ability to learn meaningful representations even before training begins.

From a theoretical perspective, this can be seen as a constrained version of the universal approximation theorem for neural networks. A MultiLayer Perceptron (MLP), as the most general neural network architecture, is theoretically capable of approximating any function mapping inputs to outputs. However, while there exists a set of weights that perfectly represents the target function, this does not mean it can be easily found with limited data, computational resources, or training time. Architectures tailored to specific tasks introduce useful properties into the approximating function (i.e., the neural network), thereby simplifying the optimization process and improving convergence. The advantage of certain architectures for particular problems lies in their ability to encode relevant properties or invariances inherent to the solution.

For instance, convolutional layers are well suited for capturing spatial invariances in images, while recurrent layers effectively model sequential dependencies in time-series data. The deep learning counterpart for arbitrary relational structures is the Graph Neural Network (GNN), which inherently respects graph properties such as permutation invariance of nodes. As a result, GNNs have proven highly effective in numerous scientific applications. Graph layers serve as the equivalent of

Convolutional Neural Network (CNN)s for graph-structured data [87], improving representation learning in domains such as knowledge graphs for image processing, natural language processing [113], and various scientific fields [86, 84, 103].

Beyond relational inductive biases, which enforce sample-to-sample constraints, more subtle biases can be introduced to ensure robustness to specific input transformations, such as rotations and translations. By designing the ML architecture to explicitly encode physical properties, such as equivariance in a dynamical particle system, the model can inherently respect domain constraints. While equivariance can also be approximated through data augmentation [103], incorporating it directly into the architecture reduces the number of trainable parameters, helping to mitigate overfitting and improve generalization.

Many existing equivariant models extend the translation equivariance of standard CNNs to broader group equivariances, as seen in architectures like G-CNN [28] and Steerable CNNs [29]. The EGNN model [122] further generalizes this concept to higher-dimensional spaces within a graph structure. Notably, several widely used ML architectures were originally inspired by fundamental physical principles. As already mentioned, the convolution operation in CNNs naturally preserve translational equivariance. Steerable and equivariant neural networks, rather then defining a substantially new architecture, enhance standard DL building blocks (mainli CNNs and GNNs) to preserve these properties.

0.1.6 Main contribution

We have just shown how TGDS techniques can enhance the robustness, generalizability, and reliability of ML models. Unfortunately, not all that glitters is gold. TGDS strategies are inherently task-specific, making it extremely difficult to quantify their benefits at scale. Moreover, it has been shown that domain injection is not always straightforward [99]. This is because imposing additional constraints can complicate the learning process, leading to models that are more generalizable to some extent but weaker in terms of performance. As a result, evaluating the advantages of this paradigm remains an open question.

The goal of this thesis is to address this gap by systematically analyzing and comparing different TGDS techniques across diverse use cases. We aim to identify the key building blocks that define successful knowledge integration and evaluate their impact on model performance, interpretability, and generalizability. By doing so, we hope to contribute to the development of more robust, theory-informed deep learning models that combine the strengths of both scientific reasoning and data-driven discovery.

The works analyzed in this paper have been chosen among the most representative found in the cited surveys for which both data and code are available.

This dissertation focuses on the following research objectives:

- To analyze and categorize different strategies for integrating domain knowledge into machine learning and deep learning models, providing a structured overview of existing approaches across various scientific domains.
- To experimentally evaluate the effectiveness of different domain knowledge injection techniques on both synthetic and real use cases, assessing their impact on model accuracy, generalizability, and interpretability.
- To investigate the Physics-Informed Neural Network (PINN) paradigm, identifying its limitations and proposing novel solutions to enhance robustness and reliability in scientific applications.
- To explore the role of architectural biases in neural networks, demonstrating their effectiveness in two case studies: a CNN-based image segmentation model inspired by human annotation patterns and a GNN-based climate prediction model incorporating climatological priors.
- To provide a comprehensive discussion of future directions in knowledge-guided machine learning, outlining key challenges and opportunities for further research in the field.

4.3 Dissertation outline

This thesis is structured as follows.

Chapter 2 explores different strategies for **incorporating domain knowledge into DL models**. It presents an experimental evaluation of three use cases within the context of climate science, assessing their effectiveness and limitations in improving model performance and generalizability. The experiments introduced in this chapter are based on the work published in [102].

Chapter 3 focuses on the **Physics-Informed Neural network (PINN) paradigm**, analyzing how these models integrate physical laws directly into the learning process. The chapter highlights common failure modes of PINNs, investigates their underlying causes, and proposes alternative strategies to mitigate these issues, enhancing both robustness and accuracy. This work was published in [99]. The chapter concludes then with another use case in which soft constraints -as the one enforced in PINNs- is introduced to limit the output space of neural networks.

Chapter 4 examines **architectural biases in neural networks**, which is another powerful knowledge-injection technique. This chapter explores the design of a CNN for image segmentation, which is specifically tailored to mimic the decision-making process of human annotators. The chapter presents these use cases coming from two published works [104, 105].

Each chapter also discusses future directions in its respective area, identifying both challenges and opportunities for further advancements. Finally, Chapter 5 concludes the dissertation by summarizing the key findings and presenting broader recommendations for future research in knowledge-guided machine learning.

References

[1] Lucie P Aarts and Peter Van Der Veer. Neural network method for solving partial differential equations. *Neural Processing Letters*, 14(3):261–271, 2001.

[2] Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J Zico Kolter. Differentiable convex optimization layers. *Advances in neural information processing systems*, 32, 2019.

[3] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024.

[4] Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

[5] Rosita R. Asawa, Carina Danchik, Alexey Zakharov, Yuchi Chen, Ty Voss, Ajit Jadhav, Darren P. Wallace, Josephine F. Trott, Robert H. Weiss, Anton Simeonov, and Natalia J. Martinez. A high-throughput screening platform for polycystic kidney disease (pkd) drug repurposing utilizing murine and human adpkd cells. *Scientific Reports*, 10:4203, 2020. doi: https://doi.org/10.1038/s41598-020-61082-3.

[6] Ambarish M. Athavale, Peter D. Hart, Mathew Itteera, David Cimbaluk, Tushar Patel, Anas Alabkaa, Jose Arruda, Ashok Singh, Avi Rosenberg, and Hemant Kulkarni. Development and Validation of a Deep Learning Model to Quantify Interstitial Fibrosis and Tubular Atrophy From Kidney Ultrasonography Images. *JAMA Network Open*, 4(5):e2111176, 2021.

[7] Kyongtae T. Bae, Wen Zhou, Chengli Shen, Douglas P. Landsittel, Zhiyuan Wu, Cheng Tao, Arlene B. Chapman, Vicente E. Torres, Alan S.L. Yu, Michal Mrug, William M. Bennett, Peter C. Harris, and the Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP). Growth Pattern of Kidney Cyst Number and Volume in Autosomal Dominant Polycystic Kidney Disease. *Clinical Journal of the American Society of Nephrology*, 14(6): 823–833, 2019.

[8] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David

Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261 [cs, stat]*, October 2018. URL http://arxiv.org/abs/1806.01261. 2651 citations (Semantic Scholar/arXiv) [2024-04-18] arXiv: 1806.01261.

[9] Valentina Benedetti, Valerio Brizi, Patrizia Guida, Susanna Tomasoni, Osele Ciampi, Elena Angeli, Ugo Valbusa, Ariela Benigni, Giuseppe Remuzzi, and Christodoulos Xinaris. Engineered kidney tubules for modeling patient-specific diseases and drug discovery. *EBioMedicine*, 33:253–268, 2018.

[10] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research*, 290(2), 2021.

[11] David Berend, Xiaofei Xie, Lei Ma, Lingjun Zhou, Yang Liu, Chi Xu, and Jianjun Zhao. Cats are not fish: Deep learning testing calls for out-of-distribution awareness. In *Proceedings of the 35th IEEE/ACM international conference on automated software engineering*, pages 1041–1052, 2020.

[12] Carsten Bergmann, Lisa M Guay-Woodford, Peter C Harris, Shigeo Horie, Dorien JM Peters, and Vicente E Torres. Polycystic kidney disease. *Nature reviews Disease primers*, 4(1):1–24, 2018.

[13] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[14] Tom Beucler, Michael Pritchard, Stephan Rasp, Jordan Ott, Pierre Baldi, and Pierre Gentine. Enforcing Analytic Constraints in Neural Networks Emulating Physical Systems. *Physical Review Letters*, 126(9):098302, March 2021. doi: 10.1103/PhysRevLett.126.098302. URL https://link.aps.org/doi/10.1103/PhysRevLett.126.098302. 185 citations (Semantic Scholar/DOI) [2024-04-18] GSCC: 0000000 Publisher: American Physical Society.

[15] Simon Bohlender, Ilkay Oksuz, and Anirban Mukhopadhyay. A survey on shape-constraint deep learning for medical image segmentation. *IEEE Reviews in Biomedical Engineering*, 16:225–240, 2021.

[16] David M Bossens and Nicholas Bishop. Explicit explore, exploit, or escape (e 4): Near-optimal safety-constrained reinforcement learning in polynomial time. *Machine Learning*, 112(3):817–858, 2023.

[17] Valerio Brizi, Valentina Benedetti, Angelo Michele Lavecchia, and Christodoulos Xinaris. Engineering kidney tissues for polycystic kidney

disease modeling and drug discovery. In *Methods in Cell Biology*, volume 153, pages 113–132. Elsevier, 2019.

[18] Anja Butter, Gregor Kasieczka, Tilman Plehn, and Michael Russell. Deep-learned Top Tagging with a Lorentz Layer. *SciPost Physics*, 5(3):028, September 2018. ISSN 2542-4653. doi: 10.21468/SciPostPhys.5.3.028. URL http://arxiv.org/abs/1707.08966. 145 citations (Semantic Scholar/arXiv) [2024-04-18] 145 citations (Semantic Scholar/DOI) [2024-04-18] arXiv:1707.08966 [hep-ex, physics:hep-ph].

[19] Shengze Cai, Zhiping Mao, Zhicheng Wang, Minglang Yin, and George Em Karniadakis. Physics-informed neural networks (pinns) for fluid mechanics: A review. *Acta Mechanica Sinica*, pages 1–12, 2022.

[20] Ruijin Cang, Hechao Li, Hope Yao, Yang Jiao, and Yi Ren. Improving direct physical properties prediction of heterogeneous materials from imaging data via convolutional neural network and a morphology-aware generative model. *Computational Materials Science*, 150:212–221, 2018.

[21] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. Hardnet: A low memory traffic network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3552–3561, 2019.

[22] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016.

[23] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html. 4589 citations.

[24] Xinyun Chen and Yuandong Tian. Learning to Perform Local Rewriting for Combinatorial Optimization. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/131f383b434fdf48079bff1e44e2d9a5-Abstract.html. GSCC: 0000000 00112.

[25] Yuyao Chen, Lu Lu, George Em Karniadakis, and Luca Dal Negro. Physics-informed neural networks for inverse problems in nano-optics and metamaterials. *Optics express*, 28(8):11618–11633, 2020.

[26] Charlotte Chorley. Healx and the pkd charity collaborate to apply ai to the discovery of novel treatments for rare kidney diseases, Dec 2021.

[27] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted*

Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, pages 424–432. Springer, 2016.

[28] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*. PMLR, 2016.

[29] Taco S Cohen and Max Welling. Steerable cnns. In *International Conference on Learning Representations*, 2017.

[30] Luca Colomba, Alessandro Farasin, Simone Monaco, Salvatore Greco, Paolo Garza, Daniele Apiletti, Elena Baralis, and Tania Cerquitelli. A Dataset for Burned Area Delineation and Severity Estimation from Satellite Imagery. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3893–3897, Atlanta GA USA, October 2022. ACM. ISBN 978-1-4503-9236-5. doi: 10.1145/3511808.3557528. URL https://dl.acm.org/doi/10.1145/3511808.3557528. 11 citations (Semantic Scholar/DOI) [2025-02-25].

[31] Adrián Cordido, Eva Cernadas, Manuel Fernández-Delgado, and Miguel A. García-González. CystAnalyser: A new software tool for the automatic detection and quantification of cysts in Polycystic Kidney and Liver Disease, and other cystic disorders. *PLoS computational biology*, 16(10):e1008337, 2020.

[32] Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020.

[33] Tirtharaj Dash, Ashwin Srinivasan, and Lovekesh Vig. Incorporating symbolic domain knowledge into graph neural networks. *Machine Learning*, 110(7): 1609–1636, 2021.

[34] Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12(1):1040, January 2022. ISSN 2045-2322. doi: 10.1038/s41598-021-04590-0. URL https://www.nature.com/articles/s41598-021-04590-0. 79 citations (Semantic Scholar/DOI) [2024-04-18] GSCC: 0000000 Number: 1 Publisher: Nature Publishing Group.

[35] Arka Daw, Anuj Karpatne, William Watkins, Jordan Read, and Vipin Kumar. Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling. *arXiv:1710.11431 [physics, stat]*, 2017. URL http://arxiv.org/abs/1710.11431. 427 citations (Semantic Scholar/arXiv) [2024-04-18] arXiv: 1710.11431.

[36] Arka Daw, R. Quinn Thomas, Cayelan C. Carey, Jordan S. Read, Alison P. Appling, and Anuj Karpatne. Physics-Guided Architecture (PGA) of Neural Networks for Quantifying Uncertainty in Lake Temperature Modeling. *arXiv:1911.02682 [physics, stat]*, November 2019. URL http:

//arxiv.org/abs/1911.02682. 92 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv: 1911.02682.

[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[38] Shujian Deng, Xin Zhang, Wen Yan, Eric I-Chao Chang, Yubo Fan, Maode Lai, and Yan Xu. Deep learning in digital pathology image analysis: a survey. *Frontiers of Medicine*, 14(4):470–487, 2020. doi: 10.1007/s11684-020-0782-9.

[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[40] Vasant Dhar. Data science and prediction. *Communications of the ACM*, 56 (12):64–73, 2013.

[41] Michelangelo Diligenti, Marco Gori, and Claudio Sacca. Semantic-based regularization for learning and inference. *Artificial Intelligence*, 244, 2017.

[42] Michelangelo Diligenti, Soumali Roychowdhury, and Marco Gori. Integrating prior knowledge into deep learning. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 920–923. IEEE, 2017.

[43] Alden A. Dima, John T. Elliott, James J. Filliben, Michael Halter, Adele Peskin, Javier Bernal, Marcin Kociolek, Mary C. Brady, Hai C. Tang, and Anne L. Plant. Comparison of segmentation algorithms for fluorescence microscopy images of cells. *Cytometry Part A*, 79A(7):545–559, 2011. doi: 10.1002/cyto.a.21079.

[44] Ke Dong, Chao Zhang, Xin Tian, Daniel Coman, Fahmeed Hyder, Ming Ma, and Stefan Somlo. Renal plasticity revealed through reversal of polycystic kidney disease in mice. *Nature Genetics*, Online:11 Oct 2021, 2021.

[45] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, 2024.

[46] Suchuan Dong and Naxian Ni. A method for representing periodic functions and enforcing exactly periodic boundary conditions with deep neural networks. *Journal of Computational Physics*, 435:110242, 2021.

[47] Priya L Donti, David Rolnick, and J Zico Kolter. Dc3: A learning method for optimization with hard constraints. In *International Conference on Learning Representations*, 2020.

[48] Priya L. Donti, David Rolnick, and J. Zico Kolter. DC3: A learning method for optimization with hard constraints. *ICLR 21 + arXiv:2104.12225 [cs, math, stat]*, April 2021. URL http://arxiv.org/abs/2104.12225. 115 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv: 2104.12225.

[49] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[50] Hamidreza Eivazi, Mojtaba Tahani, Philipp Schlatter, and Ricardo Vinuesa. Physics-informed neural networks for solving reynolds-averaged navier–stokes equations. *Physics of Fluids*, 34(7), 2022.

[51] James H Faghmous and Vipin Kumar. A big data guide to understanding climate change: The case for theory-guided data science. *Big data*, 2(3): 155–163, 2014.

[52] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. *arXiv:2006.11392 [cs, eess]*, July 2020. URL http://arxiv.org/abs/2006.11392. 689 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv: 2006.11392.

[53] Ferdinando Fioretto, Terrence WK Mak, and Pascal Van Hentenryck. Predicting ac optimal power flows: Combining deep learning and lagrangian dual methods. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 630–637, 2020.

[54] Ferdinando Fioretto, Pascal Van Hentenryck, Terrence WK Mak, Cuong Tran, Federico Baldo, and Michele Lombardi. Lagrangian duality for constrained deep learning. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*. Springer, 2021.

[55] Shiqi Gong, Qi Meng, Jue Zhang, Huilin Qu, Congqiao Li, Sitian Qian, Weitao Du, Zhi-Ming Ma, and Tie-Yan Liu. An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging. *Journal of High Energy Physics*, 2022 (7):30, July 2022. ISSN 1029-8479. doi: 10.1007/JHEP07(2022)030. URL http://arxiv.org/abs/2201.08187. 58 citations (Semantic Scholar/arXiv) [2024-04-18] 58 citations (Semantic Scholar/DOI) [2024-04-18] GSCC: 0000000 arXiv:2201.08187 [hep-ex, physics:hep-ph].

[56] Yu Gordienko, Peng Gang, Jiang Hui, Wei Zeng, Yu Kochura, Oleg Alienin, Oleksandr Rokovyi, and Sergii Stirenko. Deep learning with lung segmentation and bone shadow exclusion techniques for chest x-ray analysis of lung cancer. In *Advances in Computer Science for Engineering and Education 13*, pages 638–647. Springer, 2019.

[57] Adriana V. Gregory, Deema A. Anaam, Andrew J. Vercnocke, Marie E. Edwards, Vicente E. Torres, Peter C. Harris, Bradley J. Erickson, and Timothy L. Kline. Semantic Instance Segmentation of Kidney Cysts in MR Images: A Fully Automated 3D Approach Developed Through Active Learning. *Journal of Digital Imaging*, 34(4):773–787, 2021.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[59] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32:582–596, 2019.

[60] Magnus R Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5), 1969.

[61] Chien-Hsiang Huang, Hung-Yu Wu, and Youn-Long Lin. HarDNet-MSEG: A Simple Encoder-Decoder Polyp Segmentation Neural Network that Achieves over 0.9 Mean Dice and 86 FPS. *arXiv:2101.07172 [cs]*, January 2021. URL http://arxiv.org/abs/2101.07172. 131 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv: 2101.07172.

[62] Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*, 2023.

[63] Ameya D. Jagtap and George Em Karniadakis. Extended Physics-Informed Neural Networks (XPINNs): A Generalized Space-Time Domain Decomposition Based Deep Learning Framework for Nonlinear Partial Differential Equations. *Communications in Computational Physics*, 28(5):2002–2041, June 2020.

[64] Ameya D. Jagtap, Kenji Kawaguchi, and George Em Karniadakis. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *Journal of Computational Physics*, 404:109136, March 2020. ISSN 0021-9991. doi: 10.1016/j.jcp.2019.109136. URL https://www.sciencedirect.com/science/article/pii/S0021999119308411. 502 citations (Semantic Scholar/DOI) [2024-04-18] GSCC: 0000000.

[65] Andrey Kan. Machine learning applications in cell image analysis. *Immunology & Cell Biology*, 95(6):525–530, 2017. doi: https://doi.org/10.1038/icb.2017.16.

[66] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

[67] Anuradha Kar, Manuel Petit, Yassin Refahi, Guillaume Cerutti, Christophe Godin, and Jan Traas. Assessment of deep learning algorithms for 3d instance segmentation of confocal image datasets. *bioRxiv*, 2021. doi: 10.1101/2021.06.09.447748.

[68] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, June 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00314-5. URL https://www.nature.com/articles/s42254-021-00314-5. 1504 citations (Semantic Scholar/DOI) [2024-04-18] GSCC: 0000000 Number: 6 Publisher: Nature Publishing Group.

[69] Anuj Karpatne, Gowtham Atluri, James Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, October 2017. ISSN 1041-4347.

[70] Anuj Karpatne, Xiaowei Jia, and Vipin Kumar. Knowledge-guided Machine Learning: Current Trends and Future Prospects, May 2024. URL http://arxiv.org/abs/2403.15989. 6 citations (Semantic Scholar/arXiv) [2025-02-13] 6 citations (Semantic Scholar/DOI) [2025-02-13] arXiv:2403.15989 [cs].

[71] K. Kashinath, M. Mustafa, A. Albert, J-L. Wu, C. Jiang, S. Esmaeilzadeh, K. Azizzadenesheli, R. Wang, A. Chattopadhyay, A. Singh, A. Manepalli, D. Chirila, R. Yu, R. Walters, B. White, H. Xiao, H. A. Tchelepi, P. Marcus, A. Anandkumar, P. Hassanzadeh, and null Prabhat. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200093, February 2021. doi: 10.1098/rsta.2020.0093. URL https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0093. 272 citations (Semantic Scholar/DOI) [2024-04-18] GSCC: 0000000 Publisher: Royal Society.

[72] Marat F Khairoutdinov and David A Randall. Cloud resolving modeling of the arm summer 1997 iop: Model formulation, results, uncertainties, and sensitivities. *Journal of the Atmospheric Sciences*, 60(4):607–625, 2003.

[73] Taehun Kim, Hyemin Lee, and Daijin Kim. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2167–2175, 2021.

[74] Georgios Kissas, Yibo Yang, Eileen Hwuang, Walter R Witschey, John A Detre, and Paris Perdikaris. Machine learning in cardiovascular flows modeling: Predicting arterial blood pressure from non-invasive 4d flow mri data using

physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 358:112623, 2020.

[75] Timothy L. Kline, Marie E. Edwards, Jeffrey Fetzer, Adriana V. Gregory, Deema Anaam, Andrew J. Metzger, and Bradley J. Erickson. Automatic semantic segmentation of kidney cysts in MR images of patients affected by autosomal-dominant polycystic kidney disease. *Abdominal Radiology*, 46(3): 1053–1061, 2021.

[76] Samuel Kolb, Stefano Teso, Anton Dries, and Luc De Raedt. Predictive spreadsheet autocompletion with constraints. *Machine Learning*, 109:307–325, 2020.

[77] A. V. Konstantinov and L. V. Utkin. A New Computationally Simple Approach for Implementing Neural Networks with Output Hard Constraints. *Doklady Mathematics*, 108(2):S233–S241, December 2023. ISSN 1531-8362. doi: 10.1134/S1064562423701077. URL https://doi.org/10.1134/S1064562423701077.

[78] Alkis Koudounas, Flavio Giobergia, Irene Benedetto, Simone Monaco, Luca Cagliero, Daniele Apiletti, Elena Baralis, et al. baρtti at geolingit: Beyond boundaries, enhancing geolocation prediction and dialect classification on social media in italy. In *CEUR Workshop Proceedings*, 2023.

[79] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural Operator: Learning Maps Between Function Spaces, April 2023. URL http://arxiv.org/abs/2108.08481. 332 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv:2108.08481 [cs, math].

[80] Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 26548–26560. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/df438e5206f31600e6ae4af72f2725f1-Abstract.html. GSCC: 0000000.

[81] Florian Kromp, Lukas Fischer, Eva Bozsaky, Inge M Ambros, Wolfgang Dörr, Klaus Beiske, Peter F Ambros, Allan Hanbury, and Sabine Taschner-Mandl. Evaluation of deep learning architectures for complex immunofluorescence nuclear image segmentation. *IEEE Transactions on Medical Imaging*, 40(7): 1934–1949, 2021.

[82] Manoj Kumar and Neha Yadav. Multilayer perceptrons and radial basis function neural network methods for the solution of differential equations: a survey. *Computers & Mathematics with Applications*, 62(10):3796–3811, 2011.

[83] I.E. Lagaris, A. Likas, and D.I. Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, 9(5):987–1000, September 1998. ISSN 1941-0093. doi: 10.1109/72.712178. 1757 citations (Semantic Scholar/DOI) [2024-04-18] Conference Name: IEEE Transactions on Neural Networks.

[84] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336.

[85] Matthew B Lanktree, Amirreza Haghighi, Elsa Guiard, Ioan-Andrei Iliuta, Xuewen Song, Peter C Harris, Andrew D Paterson, and York Pei. Prevalence estimates of polycystic kidney and liver disease by population sequencing. *Journal of the American Society of Nephrology*, 29(10):2593–2600, 2018.

[86] Shuangli Li, Jingbo Zhou, Tong Xu, Liang Huang, Fan Wang, Haoyi Xiong, Weili Huang, Dejing Dou, and Hui Xiong. Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity, July 2021. URL http://arxiv.org/abs/2107.10670. 130 citations (Semantic Scholar/arXiv) [2024-10-14] arXiv:2107.10670.

[87] Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P Xing. Symbolic graph reasoning meets convolutions. *Advances in Neural Information Processing Systems*, 31, 2018.

[88] Bokai Liu, Yizheng Wang, Timon Rabczuk, Thomas Olofsson, and Weizhuo Lu. Multi-scale modeling in thermal conductivity of polyurethane incorporated with phase change materials using physics-informed neural networks. *Renewable Energy*, 220:119565, 2024.

[89] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378. Springer, 2024.

[90] Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13 (3):1224, 2021.

[91] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[92] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[93] Lu Lu, Raphael Pestourie, Wenjie Yao, Zhicheng Wang, Francesc Verdugo, and Steven G Johnson. Physics-informed neural networks with hard constraints for inverse design. *SIAM Journal on Scientific Computing*, 43(6), 2021.

[94] Riccardo Magistroni, Cristiana Corsi, Teresa Martí, and Roser Torra. A Review of the Imaging Techniques for Measuring Kidney and Cyst Volume in Establishing Autosomal Dominant Polycystic Kidney Disease Progression. *Nephrology*, 48(1):67–78, 2018.

[95] Levi McClenny and Ulisses Braga-Neto. Self-Adaptive Physics-Informed Neural Networks using a Soft Attention Mechanism, April 2022. URL http://arxiv.org/abs/2009.04544. 225 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv:2009.04544 [cs, stat].

[96] Sanket Vaibhav Mehta, Jay Yoon Lee, and Jaime Carbonell. Towards semi-supervised learning for deep semantic role labeling. *arXiv preprint arXiv:1808.09543*, 2018.

[97] Erick Moen, Dylan Bannon, Takamasa Kudo, William Graf, Markus Covert, and David Van Valen. Deep learning for cellular image analysis. *Nature Methods*, 16(12):1233–1246, 2019. doi: 10.1038/s41592-019-0403-1.

[98] Arvind T Mohan and Datta V Gaitonde. A deep learning based approach to reduced order modeling for turbulent flow control using lstm neural networks. *arXiv preprint arXiv:1804.09269*, 2018.

[99] Simone Monaco and Daniele Apiletti. Training physics-informed neural networks: One learning to rule them all? *Results in Engineering*, 18:101023, June 2023. ISSN 2590-1230. doi: 10.1016/j.rineng.2023.101023. URL https://www.sciencedirect.com/science/article/pii/S2590123023001500. 12 citations (Semantic Scholar/DOI) [2025-02-19].

[100] Simone Monaco, Nicole Bussola, Sara Buttò, Diego Sona, Daniele Apiletti, Giuseppe Jurman, Elisa Viola, Marco Chierici, Christodoulos Xinaris, and Vincenzo Viola. Cyst segmentation on kidney tubules by means of U-Net deep-learning models. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3923–3926, December 2021. doi: 10.1109/BigData52589.2021.9671669. 1 citations (Semantic Scholar/DOI) [2024-04-18] GSCC: 0000000.

[101] Simone Monaco, Salvatore Greco, Alessandro Farasin, Luca Colomba, Daniele Apiletti, Paolo Garza, Tania Cerquitelli, and Elena Baralis. Attention to Fires: Multi-Channel Deep Learning Models for Wildfire Severity Prediction. *Applied Sciences*, 11(22):11060, November 2021. ISSN 2076-3417. doi: 10.3390/app112211060. URL https://www.mdpi.com/2076-3417/11/22/11060. 11 citations (Semantic Scholar/DOI) [2024-04-18] GSCC: 0000000.

[102] Simone Monaco, Daniele Apiletti, and Giovanni Malnati. Theory-Guided Deep Learning Algorithms: An Experimental Evaluation. *Electronics*, 11(18): 2850, January 2022. ISSN 2079-9292. doi: 10.3390/electronics11182850. URL https://www.mdpi.com/2079-9292/11/18/2850. 4 citations (Semantic Scholar/DOI) [2025-02-19] Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.

[103] Simone Monaco, Sebastiano Barresi, and Daniele Apiletti. Lorentz-invariant augmentation for high-energy physics. In *Communications in Computer and Information Science*. Springer, 2023. URL https://iris.polito.it/handle/11583/2981734.

[104] Simone Monaco, Nicole Bussola, Sara Buttò, Diego Sona, Flavio Giobergia, Giuseppe Jurman, Christodoulos Xinaris, and Daniele Apiletti. AI models for automated segmentation of engineered polycystic kidney tubules. *Scientific Reports*, 14(1):2847, February 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-52677-1. URL https://www.nature.com/articles/s41598-024-52677-1. 1 citations (Semantic Scholar/DOI) [2025-02-19] Publisher: Nature Publishing Group.

[105] Simone Monaco, Lorenzo Petrosino, Christodoulos Xinaris, and Daniele Apiletti. Tandem: a confidence-based approach for precise medical image segmentation. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*, 2024.

[106] Simone Monaco, Daniele Apiletti, Andrea Francica, and Tania Cerquitelli. Quantify production planning efficiency through predictive modeling in manufacturing systems. *Computers & Industrial Engineering*, page 110919, 2025. URL https://www.sciencedirect.com/science/article/pii/S0360835225000658. Publisher: Elsevier.

[107] Nikhil Muralidhar, Mohammad Raihanul Islam, Manish Marwah, Anuj Karpatne, and Naren Ramakrishnan. Incorporating Prior Domain Knowledge into Deep Neural Networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 36–45, December 2018. doi: 10.1109/BigData.2018.8621955. 122 citations (Semantic Scholar/DOI) [2024-04-18].

[108] Nikhil Muralidhar, Jie Bu, Ze Cao, Long He, Naren Ramakrishnan, Danesh Tafti, and Anuj Karpatne. Phynet: Physics guided neural networks for particle drag force prediction in assembly. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 559–567. SIAM, 2020.

[109] Yatin Nandwani, Abhishek Pathak, Mausam, and Parag Singla. A Primal Dual Formulation For Deep Learning With Constraints. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/cf708fc1decf0337aded484f8f4519ae-Abstract.html. GSCC: 0000000.

[110] Paul A O'Gorman and John G Dwyer. Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, 10(10): 2548–2563, 2018.

[111] Sihyung Park, Bong Soo Park, Yoo Jin Lee, Il Hwan Kim, Jin Han Park, Junghae Ko, Yang Wook Kim, and Kang Min Park. Artificial intelligence with kidney disease: A scoping review with bibliometric analysis, prisma-scr. *Medicine*, 100(14):e25422, 2021.

[112] Grace CY Peng, Mark Alber, Adrian Buganza Tepole, William R Cannon, Suvranu De, Savador Dura-Bernal, Krishna Garikipati, George Karniadakis, William W Lytton, Paris Perdikaris, et al. Multiscale modeling meets machine learning: What can we learn? *Archives of Computational Methods in Engineering*, 28(3):1017–1037, 2021.

[113] Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*, 2019.

[114] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. GenCast: Diffusion-based ensemble forecasting for medium-range weather, May 2024. URL http://arxiv.org/abs/2312.15796. 84 citations (Semantic Scholar/arXiv) [2025-02-18] 84 citations (Semantic Scholar/DOI) [2025-02-18] arXiv:2312.15796 [cs].

[115] Paul Raccuglia, Katherine C Elbert, Philip DF Adler, Casey Falk, Malia B Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A Friedler, Joshua Schrier, and Alexander J Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73–76, 2016.

[116] Rahul Rai and Chandan K. Sahu. Driven by Data or Derived Through Physics? A Review of Hybrid Physics Guided Machine Learning Techniques With Cyber-Physical System (CPS). *IEEE Access*, 8:71050–71073, 2020. ISSN 2169-3536. GSCC: 0000000 Conference Name: IEEE Access.

[117] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, February 2019. ISSN 0021-9991. doi: 10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125. 6441 citations (Semantic Scholar/DOI) [2024-04-18] GSCC: 0000000.

[118] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations. *arXiv:1711.10561 [cs, math, stat]*, November 2017.

URL http://arxiv.org/abs/1711.10561. 705 citations (Semantic Scholar/arXiv) [2024-04-18] arXiv: 1711.10561 version: 1.

[119] Maziar Raissi, Alireza Yazdani, and George Em Karniadakis. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science (New York, N.Y.)*, 367(6481):1026–1030, February 2020. ISSN 0036-8075. doi: 10.1126/science.aaw4741. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7219083/. 1056 citations (Semantic Scholar/DOI) [2024-04-18] GSCC: 0000000.

[120] Alan Ramponi and Camilla Casula. Diatopit: A corpus of social media posts for the study of diatopic language variation in italy. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, 2023.

[121] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[122] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) Equivariant Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 9323–9332. PMLR, July 2021. URL https://proceedings.mlr.press/v139/satorras21a.html. ISSN: 2640-3498.

[123] Stefan Schweter. Italian bert and electra models, November 2020. URL https://doi.org/10.5281/zenodo.4263142.

[124] Sungyong Seo and Yan Liu. Differentiable Physics-informed Graph Networks. *arXiv preprint arXiv:1902.02950*, February 2019. URL https://ui.adsabs.harvard.edu/abs/2019arXiv190202950S. GSCC: 0000000 Publication Title: arXiv e-prints ADS Bibcode: 2019arXiv190202950S Type: article.

[125] Seonho Park and Pascal Van Hentenryck. Self-Supervised Primal-Dual Learning for Constrained Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. URL https://ojs.aaai.org/index.php/AAAI/article/view/25520. GSCC: 0000000.

[126] Tae Young Shin, Hyunsuk Kim, Joong-Hyup Lee, Jong-Suk Choi, Hyun-Seok Min, Hyungjoo Cho, Kyungwook Kim, Geon Kang, Jungkyu Kim, Sieun Yoon, et al. Expert-level segmentation using deep learning for volumetry of polycystic kidney and liver. *Investigative and clinical urology*, 61(6):555, 2020.

[127] Boris Shirokikh, Alexey Shevtsov, Anvar Kurmukov, Alexandra Dalechina, Egor Krivov, Valery Kostjuchenko, Andrey Golanov, and Mikhail Belyaev. Universal Loss Reweighting to Balance Lesion Size Inequality in 3D Medical Image Segmentation, July 2020. URL http://arxiv.org/abs/2007.10033. 15 citations (Semantic Scholar/arXiv) [2024-05-07] arXiv:2007.10033 [cs, eess].

[128] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[129] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit Neural Representations with Periodic Activation Functions, June 2020. URL http://arxiv.org/abs/2006.09661. 1711 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv:2006.09661 [cs, eess].

[130] Federico Siviero, Roberta Arcidiacono, Nicolò Cartiglia, Marco Costa, Marco Ferrero, F Legger, M Mandurrino, V Sola, A Staiano, and M Tornago. First application of machine learning algorithms to the position reconstruction in resistive silicon detectors. *Journal of Instrumentation*, 16(03), 2021.

[131] Federico Siviero, Flavio Giobergia, Luca Menzio, Filippo Miserocchi, Marta Tornago, Roberta Arcidiacono, Nicolo' Cartiglia, Marco Costa, Marco Ferrero, G Gioachin, et al. First experimental results of the spatial resolution of rsd pad arrays read out with a 16-ch board. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 1041, 2022.

[132] Russell Stewart and Stefano Ermon. Label-free supervision of neural networks with physics and domain knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[133] Marta Tornago, Flavio Giobergia, Luca Menzio, Federico Siviero, Roberta Arcidiacono, Nicolò Cartiglia, Marco Costa, Marco Ferrero, Giulia Gioachin, Marco Mandurrino, et al. Silicon sensors with resistive read-out: Machine learning techniques for ultimate spatial resolution. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 1047, 2023.

[134] Laura Von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Michal Walczak, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed Machine Learning – A Taxonomy and Survey of Integrating Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021. ISSN 1041-4347, 1558-2191, 2326-3865. GSCC: 0000000 arXiv: 1903.12394.

[135] Kentaro Wada. Labelme: Image polygonal annotation with python. https://github.com/wkentaro/labelme, 2025. URL https://github.com/wkentaro/labelme.

[136] Rui Wang and Rose Yu. Physics-Guided Deep Learning for Dynamical Systems: A Survey. *arXiv:2107.01272 [cs]*, March 2022. URL http://arxiv.org/abs/2107.01272. 41 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv: 2107.01272.

[137] Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *arXiv:2007.14527 [cs, math, stat]*, July 2020. URL http://arxiv.org/abs/2007.14527. 499 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv: 2007.14527.

[138] Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and Mitigating Gradient Flow Pathologies in Physics-Informed Neural Networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, January 2021. ISSN 1064-8275, 1095-7197. doi: 10.1137/20M1318043. URL https://epubs.siam.org/doi/10.1137/20M1318043. 343 citations (Semantic Scholar/DOI) [2024-04-18] GSCC: 0000000.

[139] Sifan Wang, Shyam Sankaran, and Paris Perdikaris. Respecting causality is all you need for training physics-informed neural networks, March 2022. URL http://arxiv.org/abs/2203.07404. 133 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv:2203.07404 [nlin, physics:physics, stat].

[140] Colby L. Wight and Jia Zhao. Solving Allen-Cahn and Cahn-Hilliard Equations using the Adaptive Physics Informed Neural Networks, July 2020. URL http://arxiv.org/abs/2007.04542. 140 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv:2007.04542 [cs, math].

[141] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating Physics-Based Modeling With Machine Learning: A Survey. *arXiv:2003.04919 [physics, stat]*, March 2022. URL http://arxiv.org/abs/2003.04919. 224 citations (Semantic Scholar/arXiv) [2024-04-18] GSCC: 0000000 arXiv: 2003.04919.

[142] Cynthia J Willey, Jaime D Blais, Anthony K Hall, Holly B Krasa, Andrew J Makin, and Frank S Czerwiec. Prevalence of autosomal dominant polycystic kidney disease in the european union. *Nephrology Dialysis Transplantation*, 32(8):1356–1363, 2017.

[143] Tongtong Wu, Xuekai Li, Yuan-Fang Li, Gholamreza Haffari, Guilin Qi, Yujin Zhu, and Guoqiang Xu. Curriculum-meta learning for order-robust continual relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10363–10369, 2021.

[144] Youshen Xia and Jun Wang. A recurrent neural network for solving nonlinear convex programs subject to linear constraints. *IEEE Transactions on Neural Networks*, 16(2), 2005.

[145] Dejia Xu, Peihao Wang, Yifan Jiang, Zhiwen Fan, and Zhangyang Wang. Signal processing for implicit neural representations. *Advances in Neural Information Processing Systems*, 35, 2022.

[146] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A Semantic Loss Function for Deep Learning with Symbolic Knowledge. In *Proceedings of the 35th International Conference on Machine Learning*, pages

5502–5511. PMLR, July 2018. URL https://proceedings.mlr.press/v80/xu18h.html. ISSN: 2640-3498.

[147] Lijing Yao, Hengyuan Zhang, Mengqin Zhang, Xing Chen, Jun Zhang, Jiyi Huang, and Lu Zhang. Application of artificial intelligence in renal disease. *Clinical eHealth*, 4:54–61, 2021. doi: https://doi.org/10.1016/j.ceh.2021.11.003.

[148] Hanjie Zhang, Max Botler, and Jeroen P Kooman. Deep learning for image analysis in kidney care. *Advances in kidney disease and health*, 30(1):25–32, 2023.

[149] Ji Zhang, Jingkuan Song, Lianli Gao, Ye Liu, and Heng Tao Shen. Progressive meta-learning with curriculum. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5916–5930, 2022.

[150] Alice Zheng and Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018.

[151] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.