Toward Knowledge-Guided AI for Inverse Design in Manufacturing: A Perspective on Domain, Physics, and **Human-AI Synergy**

Hugon Lee	¹¹ , Hyeonbin	$Moon^{1\dagger}$,	lunhyeong .	Lee ¹⁷ , and	Seunghwa Ryu ¹	

¹Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

Funding: National Research Foundation(NRF) of Korea, Grant/Award Number: RS-2023-00222166 and RS-2023-00247245

Keywords: inverse design, physics-informed machine learning, large language models, design automation, human-AI collaboration, smart manufacturing

[†] Equal contribution.

^{*} Corresponding author: ryush@kaist.ac.kr

Abstract

Artificial intelligence (AI) is reshaping inverse design across manufacturing domain, enabling high-performance discovery in materials, products, and processes. However, purely data-driven approaches often struggle in realistic settings characterized by sparse data, high-dimensional design spaces, and nontrivial physical constraints. This perspective argues for a new generation of design systems that transcend black-box modeling by integrating domain knowledge, physics-informed learning, and intuitive human-AI interfaces. We first demonstrate how expert-guided sampling strategies enhance data efficiency and model generalization. Next, we discuss how physics-informed machine learning enables physically consistent modeling in data-scarce regimes. Finally, we explore how large language models emerge as interactive design agents connecting user intent with simulation tools, optimization pipelines, and collaborative workflows. Through illustrative examples and conceptual frameworks, we advocate that inverse design in manufacturing should evolve into a unified ecosystem, where domain knowledge, physical priors, and adaptive reasoning collectively enable scalable, interpretable, and accessible AI-driven design systems.

1. Introduction

Artificial intelligence (AI) has opened transformative opportunities for inverse design in manufacturing, enabling data-driven models to propose design configurations that meet specific performance criteria. Surrogate modeling techniques—such as neural networks (NNs) and Gaussian processes (GPs)—have shown promise in replacing costly simulations and trial-and-error experiments with fast and flexible approximations. Supported by universal approximation theorems,^{1,2} these models can, in principle, approximate complex physical behaviors with arbitrary precision, making them attractive tools for modeling and optimization in engineering contexts.

In recent years, AI applications have expanded beyond predictive analytics into generative and prescriptive domains, where the objective is to discover novel material formulations,^{3–5} product geometries,^{6–9} or process parameters.^{10–12} This inverse design paradigm has been empowered by advances in generative modeling, surrogate-assisted optimization, and reinforcement learning.^{12–15} However, despite the increasing capacity of machine learning (ML) models, their performance in real-world engineering problems often remains fragile due to limited by sparse data availability, high-dimensional design spaces, physical constraints, and lack of interpretability.^{10,11} These challenges highlight the need for a more robust, knowledge-integrated design framework.

This perspective builds on the insight that future AI-driven inverse design requires integration of three key components: (i) domain knowledge that informs the sampling of meaningful design points considering problem characteristics, (ii) physics-informed modeling that embeds known governing principles into the model to enhance generalizability, and (iii) natural language (NL)-based interfaces that support intuitive and interactive human-AI collaboration. Through this lens, we propose that intelligent inverse design should evolve into

a tightly integrated ecosystem, harmonizing data-driven models, physical priors, and human expertise.

The remainder of this article is organized as follows. Section 2 highlights how expert-guided data acquisition improves model relevance and supports sample-efficient optimization. Building on this, Section 3 introduces physics-informed machine learning (PIML) methods that embed known physical principles into AI models, improving generalization and robustness under data-scarce regimes. Section 4 focuses on more accessible utilization of AI models via large language models (LLMs) and multi-agent frameworks, which are emerging tools for making AI-based design more accessible, explainable, and human-centered. Finally, Section 5 concludes with a broader reflection on how combining domain knowledge, physical models, and human-centered design interfaces can shape the next generation of intelligent, scalable, and interpretable design systems in manufacturing.

2. Importance of expert-guided data sampling

Inverse design in manufacturing inherently involves navigating complex, high-dimensional design spaces under practical constraints such as limited data availability, expensive evaluations, and physical feasibility.^{7,10–12,15} While ML models offer powerful tools to model input-output relationships, their performance is fundamentally tied to how the data is sampled. In real-world settings, simulations or experiments are costly, and resulting datasets are often sparse, noisy, or unevenly distributed. In such cases, naïve sampling strategies often fail to capture meaningful behavior. Instead, expert-guided data acquisition is essential to construct informative datasets that support generalizable and physically consistent surrogate models.

Figure 1 illustrates this point with a simplified two-variable design space (inputs x₁ and x₂) and a single-objective function.^a The true response surface features multiple local maxima and a global optimum (Figure 1a).^b Line A—A' intersects both a local optimum (white star) and the global optimum (black star), and the predicted values along this line reveal how different sampling strategies impact ML accuracy (Figure 1b). Three training scenarios (grid sampling in Figure 1c, random sampling in Figure 1d, and localized sampling near a local optimum in Figure 1e) use the same number of nine training data points, yet yield markedly different models. Grid sampling, through coarse, correctly captures the global optimum while random sampling misses it entirely. Localized sampling produces accurate predictions within its narrow region but fails to extrapolate, misleadingly favoring the local optimum.

This scenario reflects a common industrial question: "Given a certain number of design variables, how many samples are sufficient for this problem?" While there are rules of thumb, there is no universal answer. A model trained on a million high-fidelity samples narrowly clustered around a suboptimal region may still miss the global optimum if it lacks coverage elsewhere, despite reporting excellent test accuracy (e.g., R^2 around 0.99) within that local domain. This highlights the critical difference between interpolation accuracy and design exploration capability, particularly in high-dimensional spaces where blind sampling becomes exponentially less efficient. Without expert insight to guide exploratory sampling, especially in physically meaningful or uncertain regions, AI models will likely miss critical behaviors,

-

^a In real manufacturing scenarios, the problem usually features many more design variables and objectives that often involve trade-offs, such as minimizing cycle time, reducing defect rates, and achieving desired functional performance.

^b The illustrative two-input one-output problem utilized is the *shifted-rotated Rastrigin function*, ¹⁶ one of common choices in evaluating ML model performances. The ML model utilized is GP regression with radial basis function kernel¹⁷.

leading to suboptimal or even misleading conclusions. Thus, domain knowledge is helpful and often essential in shaping the data acquisition process and ensuring meaningful coverage of the design space.

Building on this idea, a recent review categorizes inverse design challenges into four representative scenarios based on dataset coverage and tractability of the design space, ¹⁴ as depicted in Figure 2. As shown schematically in Figure 2a, surrogate models map design variables to objective values. The categories include: (i) *Interpolation* (Figure 2b): Dense, uniform data across the space enables global optimization using off-the-shelf methods. (ii) *Extrapolation* (Figure 2c): Train data lies in a small subregion of a vast feasible space. (iii) *Small data* (Figure 2d): Limited high-cost evaluations constrain the search space and necessitate strategic sampling. (iv) *Multi-fidelity data* (Figure 2e): Surrogates integrate low-fidelity (e.g., fast approximations) and high-fidelity (e.g., costly experiments) data.

These latter three cases dominate real-world manufacturing problems, each posing specific challenges. In extrapolation regimes, domain knowledge and physical priors are essential for guiding model generalization. In small data scenarios, careful definition of input bounds and constraints is critical. Bayesian optimization and active learning can be powerful, but only when initialized with physically meaningful domains. Poorly chosen design ranges, even with sophisticated algorithms, can yield misleading or narrow Pareto fronts. In multifidelity settings involving heterogeneous data sources, low-fidelity or source-domain data can provide broad coverage, while high-fidelity or target-domain data refine accuracy. Approaches such as GP-based or NN-based multi-fidelity models can integrate both. ^{18–20} However, their effectiveness relies on two key conditions: strong correlation between domains, and sufficient support across the input space from the lower-fidelity or source data. ²¹ Without these, integrating multi-source information can compromise both predictive accuracy and reliability.

Taken together, Figures 1 and 2 emphasize that data sampling is not a passive preprocessing step but an active part of the design process. Domain experts are indispensable in: (i) Prioritizing regions of exploration, (ii) Identifying key design variables, and (iii) Constructing meaningful priors or constraints for surrogate models. In short, inverse design is not just a matter of model training but a co-optimization of data acquisition and model inference. Expert-guided sampling strategies, fidelity-aware learning, and embedded physical reasoning must be tightly integrated to ensure scalable, interpretable, and robust inverse design workflows for manufacturing applications.

3. PIML for data efficiency and reliable generalization

To mitigate the limitations of purely data-driven models, namely their dependence on large, high-quality datasets and their difficulty in enforcing physical consistency, PIML, or scientific ML (SciML), has emerged as a promising strategy. PIML combines traditional scientific computing with machine learning techniques (Figure 3a), aiming to improve data efficiency and enhance model generalizability by embedding physical priors directly into the learning process. 13,22–24 This hybrid approach not only increases interpretability by constraining solutions to obey known physics, but also leverages data to capture hidden patterns not fully explained by theory alone. Recent studies have demonstrated its utility in solving forward and inverse problems, reconstructing hidden states and characteristics from sparse observations, 25,26 and accelerating expensive computational workflows. 27 Two representative PIML frameworks stand out in the literature and practice: physics-informed neural networks (PINNs) and operator learning architectures. Both offer avenues for embedding governing equations into model training, but they differ in structure, scalability, and practical applicability.

PINNs are among the earliest and most influential frameworks in the development of PIML.²⁸ As shown in Figure 3b, they embed physical laws, typically partial differential

equations (PDEs), into the loss function of a neural network. The model inputs consist of spatial (e.g., x and y) and temporal (t) coordinates, and the output is a physical quantity of interest (f), such as displacement or temperature. In the purely data-driven case, training minimizes only a data loss term, $\mathcal{L}_{\text{data}}$, comparing predicted and labeled outputs. PINNs augment this with a physics-based loss, $\mathcal{L}_{\text{phys}}$, which penalizes violation of governing equations, often calculated using automatic differentiation to obtain PDE residuals concerning input variables.²⁹ This enables training even in semi-supervised or unsupervised regimes where labeled data is scarce.

Despite their conceptual elegance, PINNs face substantial practical hurdles. The incorporation of stiff partial differential equations (PDEs) into the loss function makes training demanding, often leading to vanishing gradients, imbalanced loss scales, and high sensitivity to hyperparameters.³⁰ These problems are particularly severe in high-dimensional or time-dependent problems, where convergence can be slow or unstable. Moreover, for inverse design tasks requiring repeated evaluations under varying conditions (*e.g.*, change in input geometry or boundary conditions), PINNs typically require retraining or fine-tuning. This limits their practical advantage over classical solvers such as finite element methods.

To overcome these issues, operator learning has gained attention as a scalable alternative. Instead of learning mappings between fixed-dimensional vectors, operator learning models approximate mappings between infinite-dimensional function spaces. This allows them to learn solution operators that generalize across varying boundary conditions, geometries, and material properties. As shown in Figure 3c, a typical operator model such as DeepONet consists of two sub-networks: the branch network, which processes the input function **u** (*e.g.*, boundary condition or material property field) and provides coefficients for the basis function, and the trunk network, which processes spatial and temporal coordinates **y** and provides basis function values at the coordinate. The inner product of their outputs yields the final solution

 $G_{\theta}(\mathbf{u})(\mathbf{y})$ where G is the solution operator and θ is model hyperparameters. This architecture, based on universal approximation theorem for operators, enables prediction across diverse problem settings without retraining. Its physics-informed variant, physics-informed DeepONet (PIDON), extends this capability by incorporating a physics loss analogous to PINNs.^{33,34} Other notable operator learning models also include Fourier neural operator (FNO)³⁵ and physics-informed neural operator (PINO).³⁶

Operator models are particularly advantageous in manufacturing design contexts requiring rapid inference. Once trained, they can evaluate thousands of new design configurations at negligible computational cost—far faster than traditional solvers or PINNs. The operator models without physics loss are typically more stable during training than PINNs since they do not rely on dynamic PDE residuals. However, this comes at the cost of requiring large, diverse training datasets, and they often struggle with extrapolation beyond the data distribution.

In practice, PIML models have demonstrated tangible benefits across a range of manufacturing applications. PINNs have been employed for stress or temperature reconstruction from sparse measurements in composite materials or thermal systems, leveraging physics to achieve consistent and interpretable results. ^{37,38} Operator learning models such as DeepONet have enabled high-throughput predictions of mechanical properties from microstructural data and rapid flow or heat field simulations under varying conditions, beneficial in porous material design, additive manufacturing, and thermal management systems. ^{39–41} Together, these approaches address core challenges in inverse design: limited data, physical constraint satisfaction, and interpretability.

Nonetheless, several challenges persist. PINNs often exhibit computational inefficiencies when applied to complex geometries or high-dimensional problems. On the other

hand, operator learning models without physical constraints typically demand large-scale training data and may generalize poorly outside the training distribution. Moreover, both PINN and PIDON assume access to known governing equations, which may be incomplete or partially known in real-world multi-physics or empirical systems. To address these gaps, future research calls for hybrid models that flexibly combine partial physics with learning, scalable training strategies, and tighter integration with human-in-the-loop systems such as natural language-based design agents (see Section 4). Through these advances, physics-informed machine learning can evolve into a practical and generalizable foundation for next-generation AI-driven design systems.

4. Human-centered inverse design through LLM

LLMs are transformer-based neural architectures trained on massive corpora of textual data. ⁴² In recent years, they have emerged as powerful tools for enabling natural language (NL) interaction in complex domains, including engineering design and manufacturing. LLMs offer two key advancements: (i) they allow for intuitive, scalable NL interface for technical workflows, and (ii) they enable human-like reasoning and decision support, helping bridge knowledge gaps among engineers, designers, and operators. With the advent of multimodal LLMs, these models can now interpret not only text but also images, audio, and other data types, opening new possibilities for human-machine collaboration in industrial settings. Building on these strengths, a range of LLM-based frameworks has been explored to tackle challenges in inverse design and process automation. These efforts can be broadly grouped into three approaches (Figure 4): (i) zero- or few-shot inference using pre-trained LLMs, (ii) retrieval-augmented generation (RAG) and fine-tuning for contextual grounding, and (iii) multi-agent systems that coordinate tasks through LLM-driven planning and decision workflows.

One of the most direct ways to leverage LLMs in manufacturing is through structured prompting—crafting targeted NL queries to elicit useful outputs from pre-trained models (Figure 4b). 43–45 This approach has shown practical effectiveness across various tasks, such as summarizing technical documents, assisting in code generation for automation workflows, and interpreting unstructured maintenance logs. 42,43 By allowing users to accomplish complex tasks through NL alone, prompting makes it easier for non-experts to access the benefit of AI without needing programming skills or specialized tools.

LLMs also exhibit strong capabilities in structured evaluation. Benchmarking platforms such as MT-Bench and Chatbot Arena have demonstrated that these models can provide consistent assessments or user queries and model responses. In engineering applications, these evaluative skills have been extended to design contexts. Recent studies show that vision-language models (VLMs) can store early-stage engineering sketches against performance criteria with a level of agreement comparable to expert reviewers. These developments suggest that LLMs can serve not only as generators but also as scalable evaluators—supporting rapid feedback in iterative design cycles through their language-based reasoning abilities.

Despite these advantages, pre-trained LLMs are inherently limited by their static training corpus. They may hallucinate facts or lack the domain-specific context required for accurate responses. To overcome this, two complementary strategies have emerged: fine-tuning and RAG. Fine-tuning adapts an LLM to domain-specific corpora such as equipment manuals, production logs, or CAD instructions (Figure 4c). 48–52 For example, CAD-Llama and CAD-Coder were developed by training LLMs and VLMs, respectively, on parametric 3D modeling tasks, enabling them to generate editable CAD code from text or sketches with high accuracy. 53,54 In contrast, RAG pipelines enhance context fidelity without retraining the model,

offering more modular and data-efficient approach. Here, relevant documents—like material datasheets or standard operating procedures—are dynamically retrieved and appended to user queries at inference time.^{48,51,55,56} A recent example is AMGPT, which combines a Llama 2-7B model with a curated technical corpus to support expert-level question answering in additive manufacturing, significantly improving response accuracy.⁵⁶

As LLMs become more versatile, they are increasingly deployed as autonomous agent connected to external tools—sensors, controllers, robotic platforms, or design software (Figure 4d). These tool-integrated agents can automate workflows such as design optimization, process control, and real-time decision-making by interpreting user commands and invoking appropriate tools, interacting with manufacturing environments. In composite manufacturing, for instance, LLM-powered agents have been used to streamline process planning and enhance decision quality based on user intent.⁵⁷

Furthermore, as manufacturing systems evolve toward higher levels of autonomy, multi-agent LLM architectures are increasingly being adopted. In these systems, multiple specialized agents take on distinct yet complementary roles, enabling decentralized task execution, robust inter-agent communication, and collaborative problem-solving (Figure 4d). This modular and decentralized structure mirrors real-world industrial ecosystems and offers a scalable pathway toward intelligent, adaptive manufacturing environments. These developments are transforming the way engineers engage with design tools via frameworks like the "design agents", which integrate VLMs, LLMs, and geometric deep learning to automate sketching, 3D modeling, and simulation in automotive design. By reducing design cycles from weeks to minutes, such systems highlight the potential of AI to transform engineering productivity and creativity.

Collectively, these approaches suggest a paradigm shift in manufacturing—from automation to human-centered interactive design. Pre-trained LLMs, augmented by retrieval systems and agent frameworks, are enabling domain experts and non-exports alike to engage in inverse design, parameter tuning, and defect prediction using NL. Importantly, this does not require training specialized models from scratch. Instead, it leverages the embedded knowledge and reasoning abilities of existing foundational models. As this ecosystem matures, we anticipate that LLMs will become integral to intelligent manufacturing workflows, enabling data-efficient, interpretable, and collaborative design environments.

5. Conclusion and future perspective

Inverse design in manufacturing is undergoing a fundamental transformation. As AI matures, the goal is no longer to automate optimization, but to integrate knowledge, physical reasoning, and human intent into a cohesive, intelligent design process. This perspective has outlined three pillars that are reshaping this landscape: (i) expert-guided data acquisition that enhances efficiency and relevance, (ii) physics-informed learning models that ensure physical consistency and generalization in sparse-date regimes, and (iii) natural language interfaces powered by LLMs that enable intuitive, human-centered interaction with AI systems.

These components are not isolated; they represent complementary layers in an emerging design ecosystem. Expert knowledge informs which regions of the design space should be explored. PIML models embed physical constraints into learning systems to produce reliable predictions. LLMs act as flexible collaborators that bridge users, simulations, and real-world tools—making advanced modeling and design accessible even to non-experts.

However, realizing the full potential of this ecosystem demands addressing several challenges. These include: (i) Developing robust frameworks for incorporating partial or uncertain physical knowledge into learning systems. (ii) Enhancing extrapolation performance

under sparse or distribution-shifted conditions. (iii) Ensuring trust, interpretability, and control when integrating LLMs into critical engineering workflows. (iv) Creating modular, extensible toolchains that allow seamless collaboration between human designers and AI agents.

The future of inverse design is not defined by replacing humans, but by augmenting their reasoning. We envision next-generation systems that not only search optimal solutions, but also help frame better design questions, reason through trade-offs, and uncover new physical insights. This shift—from black-box prediction to knowledge-guided interaction, from automation to co-creation—will define the trajectory of intelligent manufacturing in the future.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grand funded by the Korea government(MSIT) (No. RS-2023-00222166 and No. RS-2023-00247245).

Conflicts of interest

The authors declare no conflict of interest.

Data availability statement

No new data were created or analyzed in this study.

References

- Micchelli, C. A., Xu, Y. & Zhang, H. Universal Kernels. Journal of Machine Learning Research vol. 7 (2006).
- 2. Hornik, K., Stinchcombe, M. & White, H. *Multilayer Feedforward Networks Are Universal Approximators*. *Neural networks* vol. 2 359–366 (Elsevier, 1989).

- 3. Kang, Y. & Kim, J. ChatMOF: An Artificial Intelligence System for Predicting and Generating Metal-Organic Frameworks Using Large Language Models. Nature communications vol. 15 4705 (Nature Publishing Group UK London, 2024).
- 4. Lee, I. *et al.* Uncovering the Relationship between Metal Elements and Mechanical Stability for Metal–Organic Frameworks. *ACS Appl. Mater. Interfaces* **16**, 52162–52178 (2024).
- 5. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
- 6. Serles, P. et al. Ultrahigh Specific Strength by Bayesian Optimization of Carbon Nanolattices. Adv. Mater. 37, 2410651 (2025).
- 7. Lee, H. *et al.* Bayesian optimization of tailgate rib structures enhancing structural stiffness under manufacturing constraints of injection molding. *J. Manuf. Process.* **134**, 739–748 (2025).
- 8. Oh, S., Jung, Y., Kim, S., Lee, I. & Kang, N. Deep Generative Design: Integration of Topology Optimization and Generative Models. *J. Mech. Des.* **141**, 111405 (2019).
- 9. Cho, H., Lee, H. & Ryu, S. Sequential multi-objective Bayesian optimization of air-cooled battery thermal management system with spoiler integration. *J. Energy Storage* **114**, 115586 (2025).
- Jung, J., Park, K., Lee, H., Cho, B. & Ryu, S. Comparative Study of Multi-objective Bayesian Optimization and NSGA-III based Approaches for Injection Molding Process. *Adv. Theory Simul.* 7, 2400135 (2024).
- 11. Lee, H., Lee, M., Jung, J., Lee, I. & Ryu, S. Enhancing Injection Molding Optimization for SFRPs Through Multi-Fidelity Data-Driven Approaches Incorporating Prior Information in Limited Data Environments. *Adv. Theory Simul.* 7, 2400130 (2024).

- Kim, J.-Y., Kim, H., Nam, K., Kang, D. & Ryu, S. Development of an injection molding production condition inference system based on diffusion model. *J. Manuf. Syst.* 79, 162–178 (2025).
- 13. Brunton, S. L. & Kutz, J. N. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control.* (Cambridge University Press, Cambridge, 2019). doi:10.1017/9781108380690.
- 14. Lee, J. *et al.* Machine learning-based inverse design methods considering data characteristics and design space size in materials design and manufacturing: a review. *Mater. Horiz.* **10**, 5436–5456 (2023).
- 15. Yu, J. & Ryu, S. Deep Reinforcement Learning-Based Optimization for the Shape of an Adhesive Pillar With Enhanced Adhesion Strength. *J. Mech. Des.* 1–12 (2025) doi:10.1115/1.4068747.
- 16. Mainini, L. *et al.* Analytical Benchmark Problems for Multifidelity Optimization Methods. Preprint at https://doi.org/10.48550/arXiv.2204.07867 (2022).
- 17. Williams, C. & Rasmussen, C. Gaussian Processes for Regression. in *Advances in Neural Information Processing Systems* vol. 8 (MIT Press, 1995).
- 18. Kennedy, M. & O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **87**, 1–13 (2000).
- 19. Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N. D. & Karniadakis, G. E. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proc. R. Soc. Math. Phys. Eng. Sci.* **473**, 20160751 (2017).
- 20. Meng, X. & Karniadakis, G. E. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems. *J. Comput. Phys.* **401**, 109020 (2020).

- 21. Park, C., Haftka, R. T. & Kim, N. H. Remarks on multi-fidelity surrogates. *Struct. Multidiscip. Optim.* **55**, 1029–1050 (2017).
- 22. Karniadakis, G. E. *et al.* Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
- 23. Kashinath, K. *et al.* Physics-informed machine learning: case studies for weather and climate modelling. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **379**, 20200093 (2021).
- 24. Hao, Z. *et al.* Physics-Informed Machine Learning: A Survey on Problems, Methods and Applications. Preprint at https://doi.org/10.48550/arXiv.2211.08064 (2023).
- 25. Raissi, M., Yazdani, A. & Karniadakis, G. E. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science* **367**, 1026–1030 (2020).
- 26. Moon, H. et al. Physics-Informed Neural Network-Based Discovery of Hyperelastic Constitutive Models from Extremely Scarce Data. Preprint at https://doi.org/10.48550/arXiv.2504.19494 (2025).
- 27. Haghighat, E., Raissi, M., Moure, A., Gomez, H. & Juanes, R. *A Physics-Informed Deep Learning Framework for Inversion and Surrogate Modeling in Solid Mechanics.*Computer Methods in Applied Mechanics and Engineering vol. 379 113741 (Elsevier, 2021).
- 28. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).
- 29. Baydin, A. G., Pearlmutter, B. A., Radul, A. A. & Siskind, J. M. Automatic differentiation in machine learning: a survey. Preprint at https://doi.org/10.48550/arXiv.1502.05767 (2018).
- 30. Krishnapriyan, A., Gholami, A., Zhe, S., Kirby, R. & Mahoney, M. W. Characterizing possible failure modes in physics-informed neural networks. *Adv. Neural Inf. Process. Syst.* **34**, 26548–26560 (2021).

- 31. Lu, L., Jin, P., Pang, G., Zhang, Z. & Karniadakis, G. E. Learning Nonlinear Operators via DeepONet Based on the Universal Approximation Theorem of Operators. Nature machine intelligence vol. 3 218–229 (Nature Publishing Group UK London, 2021).
- 32. Kovachki, N. *et al.* Neural operator: Learning maps between function spaces with applications to pdes. *J. Mach. Learn. Res.* **24**, 1–97 (2023).
- 33. Goswami, S., Bora, A., Yu, Y. & Karniadakis, G. E. Physics-Informed Deep Neural Operator Networks. Preprint at https://doi.org/10.48550/arXiv.2207.05748 (2022).
- 34. Wang, S., Wang, H. & Perdikaris, P. Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Sci. Adv.* 7, eabi8605 (2021).
- 35. Li, Z., Kovachki, N. B. & Azizzadenesheli, K. Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. in *International Conference on Learning Representations* vol. 2 4 (2021).
- 36. Li, Z. *et al.* Physics-Informed Neural Operator for Learning Partial Differential Equations. *ACM IMS J. Data Sci.* **1**, 1–27 (2024).
- 37. Go, M.-S., Lim, J. H. & Lee, S. *Physics-Informed Neural Network-Based Surrogate Model for a Virtual Thermal Sensor with Real-Time Simulation. International Journal of Heat and Mass Transfer* vol. 214 124392 (Elsevier, 2023).
- 38. Sabathiel, S., Sanchis-Alepuz, H., Wilson, A. S., Reynvaan, J. & Stipsitz, M. Neural Network-Based Reconstruction of Steady-State Temperature Systems with Unknown Material Composition. Scientific Reports vol. 14 22265 (Nature Publishing Group UK London, 2024).

- 39. Jin, H. *et al.* Characterization and Inverse Design of Stochastic Mechanical Metamaterials Using Neural Operators. *Adv. Mater.* 2420063 (2025) doi:10.1002/adma.202420063.
- 40. Pieressa, A., Baruffa, G., Sorgato, M. & Lucchetta, G. Enhancing weld line visibility prediction in injection molding using physics-informed neural networks. *J. Intell. Manuf.* (2024) doi:10.1007/s10845-024-02460-w.
- 41. Yang, S., Peng, S., Guo, J. & Wang, F. A review on physics-informed machine learning for monitoring metal additive manufacturing process. *Adv Manuf* 1, 1–28 (2024).
- 42. Minaee, S. *et al.* Large Language Models: A Survey. Preprint at https://doi.org/10.48550/arXiv.2402.06196 (2025).
- 43. Matthes, M., Guhr, O., Krockert, M. & Munkelt, T. Leveraging LLMs for Information Extraction in Manufacturing. in *Advances in Production Management Systems. Production Management Systems for Volatile, Uncertain, Complex, and Ambiguous Environments* (eds. Thürer, M., Riedel, R., Von Cieminski, G. & Romero, D.) vol. 732 355–366 (Springer Nature Switzerland, Cham, 2024).
- 44. Jignasu, A. *et al.* Towards Foundational AI Models for Additive Manufacturing: Language Models for G-Code Debugging, Manipulation, and Comprehension. Preprint at https://doi.org/10.48550/arXiv.2309.02465 (2023).
- 45. Kernan Freire, S., Foosherian, M., Wang, C. & Niforatos, E. Harnessing Large Language Models for Cognitive Assistants in Factories. in *Proceedings of the 5th International Conference on Conversational User Interfaces* 1–6 (ACM, Eindhoven Netherlands, 2023). doi:10.1145/3571884.3604313.
- 46. Zheng, L. *et al.* Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Preprint at https://doi.org/10.48550/arXiv.2306.05685 (2023).

- 47. Edwards, K. M., Tehranchi, F., Miller, S. R. & Ahmed, F. AI Judges in Design: Statistical Perspectives on Achieving Human Expert Equivalence With Vision-Language Models. Preprint at https://doi.org/10.48550/arXiv.2504.00938 (2025).
- 48. Liu, X., Erkoyuncu, J. A., Fuh, J. Y. H., Lu, W. F. & Li, B. *Knowledge Extraction for Additive Manufacturing Process via Named Entity Recognition with LLMs. Robotics and Computer-Integrated Manufacturing* vol. 93 102900 (Elsevier, 2025).
- 49. Naqvi, S. M. R. et al. Unlocking Maintenance Insights in Industrial Text through Semantic Search. Computers in Industry vol. 157 104083 (Elsevier, 2024).
- 50. Wang, P., Karigiannis, J. & Gao, R. X. Ontology-Integrated Tuning of Large Language Model for Intelligent Maintenance. CIRP annals vol. 73 361–364 (Elsevier, 2024).
- 51. Xu, Q. et al. Generative AI and DT integrated intelligent process planning: a conceptual framework. *Int. J. Adv. Manuf. Technol.* **133**, 2461–2485 (2024).
- 52. Zhou, B. et al. CausalKGPT: Industrial Structure Causal Knowledge-Enhanced Large

 Language Model for Cause Analysis of Quality Problems in Aerospace Product

 Manufacturing. Advanced Engineering Informatics vol. 59 102333 (Elsevier, 2024).
- 53. Li, J. *et al.* CAD-Llama: Leveraging Large Language Models for Computer-Aided Design Parametric 3D Model Generation. Preprint at https://doi.org/10.48550/arXiv.2505.04481 (2025).
- 54. Doris, A. C., Alam, M. F., Nobari, A. H. & Ahmed, F. CAD-Coder: An Open-Source Vision-Language Model for Computer-Aided Design Code Generation. Preprint at https://doi.org/10.48550/arXiv.2505.14646 (2025).
- 55. Freire, S. K. *et al.* Knowledge Sharing in Manufacturing using Large Language Models: User Evaluation and Model Benchmarking. Preprint at https://doi.org/10.48550/arXiv.2401.05200 (2024).

- 56. Chandrasekhar, A., Chan, J., Ogoke, F., Ajenifujah, O. & Farimani, A. B. *AMGPT: A Large Language Model for Contextual Querying in Additive Manufacturing. Additive Manufacturing Letters* vol. 11 100232 (Elsevier, 2024).
- 57. Holland, M. & Chaudhari, K. Large Language Model Based Agent for Process Planning of Fiber Composite Structures. Manufacturing Letters vol. 40 100–103 (Elsevier, 2024).
- 58. Fan, H., Liu, X., Fuh, J. Y. H., Lu, W. F. & Li, B. Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics. *J. Intell. Manuf.* **36**, 1141–1157 (2025).
- 59. Fan, H. et al. AutoMEX: Streamlining Material Extrusion with AI Agents Powered by Large Language Models and Knowledge Graphs. Materials & Design 113644 (Elsevier, 2025).
- 60. Xia, Y., Shenoy, M., Jazdi, N. & Weyrich, M. Towards autonomous system: flexible modular production system enhanced with large language model agents. in 2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA) 1–8 (IEEE, 2023).
- 61. Vyas, J. & Mercangöz, M. Autonomous Industrial Control using an Agentic Framework with Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2411.05904 (2024).
- 62. Lim, J., Vogel-Heuser, B. & Kovalenko, I. Large language model-enabled multi-agent manufacturing systems. in 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE) 3940–3946 (IEEE, 2024).
- Jadhav, Y., Pak, P. & Farimani, A. B. LLM-3D Print: Large Language Models To Monitor and Control 3D Printing. Preprint at https://doi.org/10.48550/arXiv.2408.14307 (2025).

- 64. Jeon, J. et al. ChatCNC: Conversational Machine Monitoring via Large Language Model and Real-Time Data Retrieval Augmented Generation. Journal of Manufacturing Systems vol. 79 504–514 (Elsevier, 2025).
- 65. Elrefaie, M. *et al.* AI Agents in Engineering Design: A Multi-Agent Framework for Aesthetic and Aerodynamic Car Design. Preprint at https://doi.org/10.48550/arXiv.2503.23315 (2025).

Figure sets

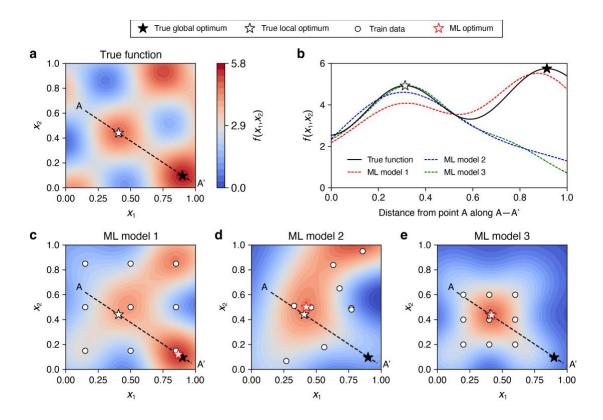


Figure 1. Illustration of surrogate modeling of two inputs-single output problem with different train data sampling techniques. (a) True function and corresponding true optimal points. (b) True function value and ML model predictions along the line A—A' passing through true local and global optimum points. ML model predictions constructed by train data sampled with (c) grid sampling, (d) random sampling and (e) localized grid sampling around true local optimum point.

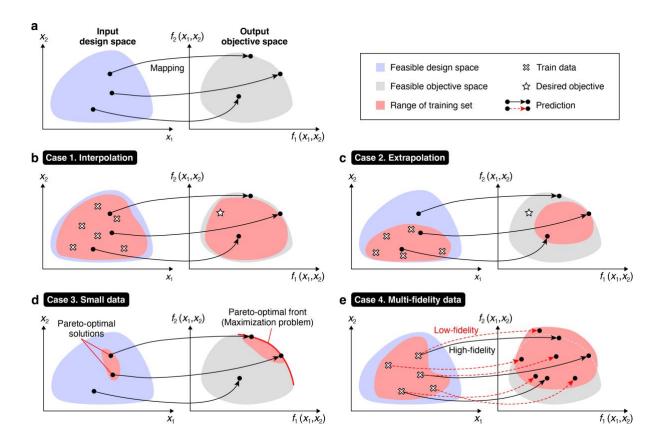


Figure 2. Four categories of typical ML scenarios in engineering problems. (a) ML model as a mapping from input designs in feasible design space to output objective values in feasible objective space. The ML problems are classified into four scenarios of (b) Interpolation, (c) Extrapolation, (d) Small data, and (e) Multi-fidelity data.

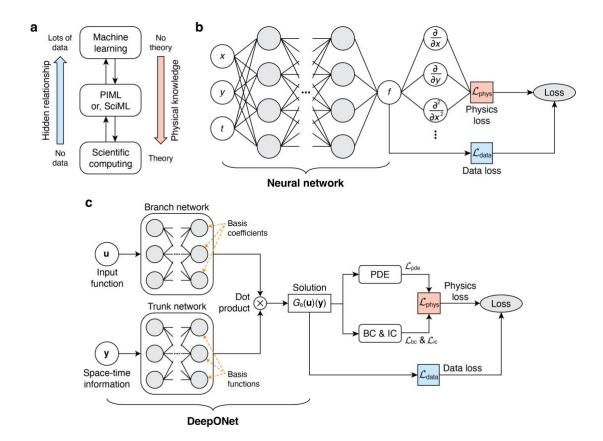


Figure 3. Schematics of representative PIML models. (a) PIML, or SciML, utilizes both physics and data to construct an efficient and generalizable ML models. (b) PINN model architecture enriching NN with physics loss. (c) PIDON model architecture enriching DeepONet with physics loss.

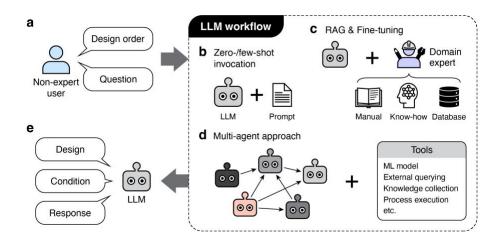


Figure 4. Schematics of leveraging LLMs in manufacturing problems for non-expert users. (a) The user provides design order or questions to the LLM. LLMs are mainly utilized in three ways: by (b) zero-/few-shot invocation, by (c) RAG and fine-tuning, or by (d) multi-agentic approach with utilizing versatile auxiliary tools. (e) After the selected workflow, the LLM provide designs, conditions, or any other responses which satisfy the user's needs.